Dear Reader,

It is a great pleasure and privilege to welcome you to the first volume of the SIAM Journal on Financial Mathematics. This journal, with the help of its editorial board, contributors, and readers and the SIAM staff, follows in the outstanding tradition of high-quality SIAM journals and, we hope, will become a leading journal in mathematical modeling and analysis of financial problems.

The journal began accepting submissions in October 2008. You will find the first papers accepted for this all-electronic publication. These papers provide a representative glance at the diversity of topics covered by the journal, something clearly observable from the outstanding editorial board itself. A number of papers currently in the pipeline further stress the broad vision of the journal.

The SIAM Journal on Financial Mathematics came to exist thanks to the tremendous effort of numerous people. The impetus came from the SIAM Activity Group on Financial Mathematics and Engineering (SIAG/FME) at its inaugural meeting in Boston in July 2006. While the idea of a new journal focused on computational finance had been discussed for a number of years, the enthusiasm at this meeting persuaded us, and other officers of the SIAG, to work toward this end. We received enthusiastic support from the publications committee and SIAM's board and council, and the journal was approved in July 2008. We are grateful to Tim Kelley and David Marshall for their hard work and advice in establishing the journal.

Preparing for the creation of the journal was arduous but rewarding. We aimed high, and at times, we wondered whether our goals were realistic. After agreeing on a charter for the journal, we made a wish list of associate editors, and it was no surprise to be asked by members of the council, "What makes you think that these distinguished scholars will accept your invitation to work for a new journal?" Except for two special cases, we were delighted that all of them accepted. In hindsight, the warm reception by the applied mathematics community at large has made the experience rewarding.

The SIAM staff is a wonderful asset. In particular, Mitch Chernoff, Brian Fauth, and Heather Blythe have been a critical source of support.

The current editorial board is working very hard to guarantee timely and high-quality reviews. The whole cycle of reviewing and production has been fast on average. Please enjoy your reading, and we look forward to receiving your high-quality submissions in the future.

René Carmona and Ronnie Sircar
Editors-in-Chief,
SIAM Journal on Financial Mathematics
January 21, 2010

# Local Volatility Enhanced by a Jump to Default* 

Peter Carr ${ }^{\dagger}$ and Dilip B. Madan ${ }^{\ddagger}$


#### Abstract

A local volatility model is enhanced by the possibility of a single jump to default. The jump has a hazard rate that is the product of the stock price raised to a prespecified negative power and a deterministic function of time. The empirical work uses a power of -1.5 . It is shown how one may simultaneously recover from the prices of credit default swap contracts and equity option prices both the deterministic component of the hazard rate function and revised local volatility. The procedure is implemented on prices of credit default swaps and equity options for General Motors and the Ford Motor Company over the period October 2004 to September 2007.


Key words. recovering default free option prices, truncated power prices, Weibull distribution, default adjusted drifts

AMS subject classifications. 60G99, 60J60, 60J75
DOI. 10.1137/090750731

1. Introduction. The development of liquid markets in credit default swaps (CDSs) has made it clear that some of the value of deep downside put options must be due to the probability of default rendering equity worthless. Stock price models that assume a strictly positive price for equity potentially overstate downside volatilities to compensate for the assumed absence of the default event in the model. To rectify this situation one needs to recognize the default possibility in the stock price model. This could be done in a variety of parametric models [9], [2], [1], [14], [3], [4], and one could then seek to infer the parameters of the model from traded option prices. Having done so, one would extract in principle a risk neutral default probability from option prices that could then be compared to a similar probability obtained from the prices of CDSs.

Alternatively one could recognize that option prices constitute an indirect assessment of default, while the CDS is clearly more directly focused on this event. This suggests that we jointly employ data on CDSs and equity options to simultaneously infer the risk neutral stock dynamics in the presence of default as a likely event. Such an approach is called for all the more in the context of local volatility stock price dynamics [12], [11] that introduce a two dimensional local volatility surface describing the instantaneous volatility of the innovations in the stock price as a deterministic function of the stock price and calendar time. All the equity option prices are then needed to infer the local volatility surface, and one needs other traded assets to infer the parameters related to the default likelihood. Given the popularity of

[^0]local volatility in assessing equity risks, we consider enhancing this model to include a default event. For other enhancements that have recently been considered we refer the reader to Ren, Madan, and Qian [17].

One could seek to perform such an enhancement by allowing the stock to diffuse to zero and to then be absorbed there. There are such models referenced above that lack a local volatility structure, and one may seek to enhance these with a local volatility formulation. However, it has long been recognized that a purely diffusive formulation has difficulties with credit spreads especially at the shorter maturities. Hence the popularity of reduced form credit models [13] that incorporate jumps in the asset price. In the interests of parsimony we introduce just a single jump to default with a hazard rate or instantaneous default probability that is just a function of the stock price and time. Unlike local volatility with its eye on the two dimensional surface of implied volatilities, we recognize that our additional assets are at best the one dimensional continuum of CDS quotes and so we model the hazard rate as a product of two functions, one a deterministic function of calendar time, while the other is just given by the stock price raised to a negative power. The latter recognizes the intuition that default gets more likely as equity values drop.

With this enhancement we show how one may use CDS and equity option prices jointly to infer the deterministic function of calendar time in the hazard rate and the local volatility surface of this default extended local volatility model for a prespecified level of the elasticity of hazard rates to the stock price. The procedures developed are implemented on CDS and equity option prices for the Ford Motor Company and General Motors (GM) covering the period October 2004 to September 2007.

It is observed that on average the elasticity of the proportion of traditional local volatility attributed to the default component declines with respect to strike and is -.4768 and -.3967 for GM and Ford, respectively. This proportion is positively related to maturity with elasticities of .0646 and .0504 for GM and Ford, respectively.

Additionally the deterministic component of the hazard function is generally increasing with maturity and has semielasticities and elasticities of .1306 and .2163 for GM and .0975 and .2641 for Ford.

The outline of the rest of the paper is as follows. Section 2 presents the formal model for the stock price evolution. In section 3 we develop the equations for the deterministic component of the hazard rate and the default enhanced local volatility surface. Section 4 presents the methods employed to evaluate truncated power prices in the embedded default free dynamics that are needed for the local volatility construction. Section 5 presents the application of the methods to Ford and GM and summarizes the results. Section 6 concludes the paper.
2. Local volatility enhanced by a single jump to default. We enhance a local volatility formulation for the risk neutral evolution of the stock price by including a single jump to default. The jump occurs at random time given by the first time a counting process $N(t)$ jumps by unity. Thereafter we set the arrival rate of jumps to zero, and the process $N(t)$ remains frozen at unity. Prior to the jump in the process $N(t)$ it has an instantaneous arrival rate $\lambda(t)$ of a jump given by the product of a deterministic function of time $f(t)$ and the stock price raised to the power $p$, which will be a negative number in our applications:

$$
\lambda(t)=\left(1-N\left(t_{-}\right)\right) f(t) S\left(t_{-}\right)^{p}
$$

This model for the default intensity in the context of a constant volatility was previously considered and solved in closed form by Linetsky [14]. An extension to the context of a constant elasticity of variance for the volatility specification was solved by Carr and Linetsky [8]. We generalize here to the local volatility context.

Let $\sigma(S, t)$ denote the volatility in the stock price when the stock price is at level $S$ at time $t$ and default has not yet occurred. The stock price dynamics may then be written as

$$
\begin{align*}
d S= & (r(t)-q(t)) S\left(t_{-}\right) d t+\sigma\left(S\left(t_{-}\right), t\right) S\left(t_{-}\right) d W(t)  \tag{1}\\
& -S\left(t_{-}\right)\left[d N(t)-\left(1-N\left(t_{-}\right)\right) f(t) S\left(t_{-}\right)^{p} d t\right],
\end{align*}
$$

where $W(t)$ is a Brownian motion and $r(t), q(t)$ are deterministic continuously compounded interest rates and dividend yields, respectively.

Equation (1) may be explicitly solved in terms of the product of two martingales, one continuous, $M(t)$, and the other, $J(t)$, that is continuous until it jumps to zero and then stays there.

The stock price may also be written as

$$
\begin{aligned}
S(t) & =A(t) M(t) J(t) \\
A(t) & =S(0) \exp \left(\int_{0}^{t}(r(u)-q(u)) d u\right) \\
M(t) & =\exp \left(\int_{0}^{t} \sigma\left(S\left(u_{-}, u\right)\right) S\left(u_{-}\right) d W(u)-\frac{1}{2} \int_{0}^{t} \sigma^{2}\left(S\left(u_{-}, u\right)\right) S\left(u_{-}\right)^{2} d u\right), \\
J(t) & =\exp \left(\int_{0}^{t}\left(1-N\left(u_{-}\right)\right) f(u) S\left(u_{-}\right)^{p} d u\right)(1-N(t)) .
\end{aligned}
$$

The only path to default or a zero equity value is the jump in the process $N(t)$ to the level 1 . The dynamics are risk neutral, and the stock growth rate is the interest rate less the dividend yield.

Embedded in the stock evolution subject to the possibility of default is the default free stock price model, or the law of the stock price conditioned on no default or, equivalently, on being positive. This is a useful process, and we shall use it in reconstructing the local volatility functions, in particular from quoted option prices. Hence we now proceed to describe this process to which we henceforth refer as the default free process.

We begin by noting that the probability of surviving $t$ units of time is given by

$$
\begin{equation*}
P(\text { no default to } t)=V(t)=E\left[\exp \left(-\int_{0}^{t} f(u)\left(1-N\left(u_{-}\right)\right) S\left(u_{-}\right)^{p} d u\right)\right] \tag{2}
\end{equation*}
$$

Consider now any path dependent claim paying at $t: F(S(u), u \leq t)$ if there is no default until time $t$, and zero otherwise. The time $t$ forward price, $w$, of this claim is

$$
\begin{equation*}
w=E\left[\exp \left(-\int_{0}^{t} \lambda(u) d u\right) F(S(u), u \leq t)\right] . \tag{3}
\end{equation*}
$$

Let $\widetilde{Q}$ be the default free law or the law of the stock conditional on no default to time $t$. It follows by definition that

$$
\begin{equation*}
w=V(t) E^{\widetilde{Q}}[F(S(u), u \leq t)] \tag{4}
\end{equation*}
$$

Combining (3) and (4) we observe that

$$
\begin{equation*}
\frac{d \widetilde{Q}}{d Q}=\frac{\exp \left(-\int_{0}^{t} \lambda(u) d u\right)}{E\left[\exp \left(-\int_{0}^{t} \lambda(u) d u\right)\right]} \tag{5}
\end{equation*}
$$

Define by $p(S, t)$ the density of the stock at time $t$ under the measure $\widetilde{Q}$. This density can in principle be recovered from call option prices conditioned on no default via the methods of Breeden and Litzenberger [5]. The prices conditional on default may be recovered from market prices and the CDS curve using (7) and (8) as explained later. By construction this density integrates to unity over the positive half-line, and it may be extracted from data on default free prices or option prices under the law $\widetilde{Q}$.

Of particular interest is the growth rate of the stock under the law $\widetilde{Q}$. This is determined by evaluating the expectation of $S(t)$ under $\widetilde{Q}$. This is given by

$$
\left.\left.\begin{array}{rl}
E^{\tilde{Q}}[S(t)]= & E^{Q}\left[\frac{\exp \left(-\int_{0}^{t} \lambda(u) d u\right)}{E[ } \exp \left(-\int_{0}^{t} \lambda(u) d u\right)\right] \\
\hline
\end{array}\right)\right] \quad \begin{aligned}
= & S(0) E^{Q}\left[\frac{\exp \left(-\int_{0}^{t} \lambda(u) d u\right)}{E\left[\exp \left(-\int_{0}^{t} \lambda(u) d u\right)\right]} \exp \left(\int_{0}^{t}(r(u)-q(u)) d u\right)\right. \\
& \quad \times \exp \left(\int_{0}^{t} \sigma\left(S\left(u_{-}, u\right)\right) S\left(u_{-}\right) d W(u)-\frac{1}{2} \int_{0}^{1} \sigma^{2}\left(S\left(u_{-}, u\right)\right) S\left(u_{-}\right)^{2} d u\right) \\
& \left.\quad \times \exp \left(\int_{0}^{t} \lambda(u) d u\right)\right] \\
= & S(0) \frac{E^{Q}\left[\exp \left(\int_{0}^{t}(r(u)-q(u)) d u\right)\right]}{V(t)}
\end{aligned}
$$

as under $\widetilde{Q}, N(t)=0$.
Let us now write

$$
V(t)=\exp \left(-\int_{0}^{t} h(u) d u\right)
$$

where by construction

$$
h(t)=-\frac{\partial \log V(t)}{\partial t} .
$$

We then obtain that

$$
E^{\widetilde{Q}}[S(t)]=S(0) \exp \left(\int_{0}^{t}(r(u)-(q(u)-h(u))) d u\right) .
$$

We then see that under the $\widetilde{Q}$ measure we have a dividend yield adjustment to

$$
\begin{equation*}
q_{a}(t)=q(t)-h(t) \tag{6}
\end{equation*}
$$

The exposure to the default hazard appears in the default free model as a negative dividend yield of $h(t)$. This is quite intuitive, as an operational way to get default free is to buy insurance
against it at the premium flow $h(t)$ that is an expense to be paid out of the dividend stream. Hence the reduced dividend flow.

Default free option prices or prices of options under the measure $\widetilde{Q}$ may be constructed by a simple transformation. Suppose we have estimated $V(t)$, the survival probability curve, from CDS prices. We can then construct prices of default free call and put options by

$$
\begin{align*}
& \widetilde{C}(K, t)=\frac{C(K, t)}{V(t)},  \tag{7}\\
& \widetilde{P}(K, t)=\frac{P(K, t)-K e^{-\int_{0}^{t} r(u) d u}(1-V(t))}{V(t)} . \tag{8}
\end{align*}
$$

Equation (7) reflects the computation that the defaultable call price is the probability of no default times the conditional expectation of the call payoff given no default. For (8) we must recognize in addition that the defaultable put price also includes the receipt of the strike in default.

We also have by definition of $p(S, t)$ that these prices satisfy the equations

$$
\begin{aligned}
& \widetilde{C}(K, t)=e^{-\int_{0}^{t} r(u) d u} \int_{K}^{\infty}(S-K) p(S, t) d S \\
& \widetilde{P}(K, t)=e^{-\int_{0}^{t} r(u) d u} \int_{0}^{K}(K-S) p(S, t) d S
\end{aligned}
$$

We may evaluate the put call parity condition for these prices and observe that

$$
\begin{aligned}
\widetilde{C}(K, t)-\widetilde{P}(K, t) & =\frac{C}{V(t)}-\frac{P-K e^{-\int_{0}^{t} r(u) d u}(1-V(t))}{V(t)} \\
& =\frac{S(0) e^{-\int_{0}^{t} q(u) d u}}{V(t)}-K e^{-\int_{0}^{t} r(u) d u} .
\end{aligned}
$$

We see again that for these default free prices the value of the forward stock is as shown earlier,

$$
\widetilde{E}[S(t)]=\frac{S(0) e^{-\int_{0}^{t} q(u) d u}}{V(t)}
$$

We may define term dividend yields and hazard rates by

$$
\begin{aligned}
& \widetilde{q}(t)=\frac{\int_{0}^{t} q(u) d u}{t} \\
& \eta(t)=\frac{\int_{0}^{t} h(u) d u}{t}
\end{aligned}
$$

and write

$$
\begin{equation*}
\widetilde{E}[S(t)]=S(0) e^{-(\widetilde{q}(t)-\eta(t)) t} . \tag{9}
\end{equation*}
$$

We may now apply any standard default free model to the prices $\widetilde{C}(K, t), \widetilde{P}(K, t)$ with the dividend yield adjusted to $q_{a}(t)=\widetilde{q}(t)-\eta(t)$ to recover the default free densities $p(S, t)$, and
we shall shortly show how we use this density to build the deterministic component of the hazard rate $f(t)$ and the local volatility function.

We note here that the actual default free law given by the density (5) has a path dependent hazard rate $\lambda(t)$ that depends on the path of the stock price. The pricing of path dependent claims under $\widetilde{Q}$ would require the use of this hazard rate. However, for mere functions of the final stock price we may proceed from the prices (7), (8) to extract the densities $p(S, t)$, and for this we make the appropriate adjustment to the forward price given by the default free forward price (9). As these are all the functions we need, we do not work with the more involved path dependent hazard rates.
3. Recovering hazard and volatility functions from CDS and option markets. In this section we describe how to recover the deterministic component of the hazard rate and the local volatility function from prices of CDSs and equity options. A similar analysis was conducted in Carr and Javaheri [7] in a slightly different context. For completeness and the specificity of our context we provide a comparable derivation.

The first step is to recover the survival function $V(t)$ from CDS quotes, and for this we employ the Weibull model for the life curve and the methods described in Madan, Konikov, and Marinescu [15].

We next observe on differentiating $\ln V(t)$ with respect to $t$ that

$$
\begin{equation*}
\frac{1}{V(t)} \frac{\partial V(t)}{\partial t}=-f(t) E^{\widetilde{Q}}\left[S(t)^{p}\right] . \tag{10}
\end{equation*}
$$

Hence the function $f(t)$ may be recovered from $V(t)$ and the prices of options under $\widetilde{Q}$. We see immediately how we will use the law of $S(t)$ under $\widetilde{Q}$ or the density $p(S, t)$ to recover the function deterministic component of hazard rates or the function $f(t)$. In fact the characteristic function of $\ln S(t)$ under $\widetilde{Q}$ evaluated at -ip gives us directly the default free power price $E^{\widetilde{Q}}\left[S(t)^{p}\right]$ from which we may construct $f(t)$ in accordance with (10).

For the recovery of the local volatility function with deterministic interest rates we proceed as follows. By definition market call prices are given by

$$
C(K, t)=\exp \left(-\int_{0}^{t} r(u) d u\right) E\left[\exp \left(-\int_{0}^{t} \lambda(u) d u\right)(S(t)-K)^{+}\right]
$$

Differentiation with respect to $K$ yields

$$
\begin{aligned}
C_{K} & =-e^{-\int_{0}^{t} r(u) d u} E\left[\exp \left(-\int_{0}^{t} \lambda(u) d u\right) \mathbf{1}_{S(t)>K}\right] \\
C_{K K} & =C_{K K}=e^{-\int_{0}^{t} r(u) d u} E\left[\exp \left(-\int_{0}^{t} \lambda(u) d u\right) \mathbf{1}_{S\left(t_{-}\right)=K}\right] \\
C-K C_{K} & =E\left[\exp \left(-\int_{0}^{t} \lambda(u) d u\right) S(t) \mathbf{1}_{S(t)>K}\right]
\end{aligned}
$$

Applying the Meyer-Tanaka formula [16], [10], [18] to the call price payoff yields

$$
\begin{aligned}
(S(t)-K)^{+}= & (S(0)-K)^{+}+\int_{0}^{t}\left(1-N\left(u_{-}\right)\right) \mathbf{1}_{S\left(u_{-}\right)>K} d S(u) \\
& +\frac{1}{2} \int_{0}^{t}\left(1-N\left(u_{-}\right)\right) \mathbf{1}_{S(u-)=K} \sigma^{2}(S(u), u) S(u)^{2} d u \\
& +K \int_{0}^{t} \mathbf{1}_{S\left(u_{-}\right)>K}\left(1-N\left(u_{-}\right)\right) d N(u) .
\end{aligned}
$$

Taking expectations on the left and expectations of time $u$ conditional expectations of the integrands on the right, we get

$$
\begin{aligned}
e^{\int_{0}^{t} r(u) d u} C(K, t)= & (S(0)-K)^{+}+\int_{0}^{t} E\left[\exp \left(-\int_{0}^{u} \lambda(v) d v\right) S\left(u_{-}\right)(r(u)-q(u))\right] d u \\
& +\frac{1}{2} \int_{0}^{t} E\left[\exp \left(-\int_{0}^{u} \lambda(v) d v\right) \mathbf{1}_{S\left(u_{-}\right)=K}\right] \sigma^{2}(K, u) K^{2} d u \\
& +K \int_{0}^{t} f(u) E\left[\exp \left(-\int_{0}^{u} \lambda(v)\right) S\left(u_{-}\right)^{p} \mathbf{1}_{S\left(u_{-}\right)>K}\right] d u
\end{aligned}
$$

Differentiating with respect to $t$ and multiplying by the discount factor, we get that

$$
\begin{aligned}
r(t) C(K, t)+C_{t}= & e^{-\int_{0}^{t} r(u) d u} E\left[\exp \left(-\int_{0}^{t} \lambda(u) d v\right) S\left(t_{-}\right)(r(t)-q(t)) \mathbf{1}_{S\left(t_{-}\right)>K}\right] \\
& +\frac{1}{2} e^{-\int_{0}^{t} r(u) d u} E\left[\exp \left(-\int_{0}^{t} \lambda(u) d u\right) \mathbf{1}_{S\left(t_{-}\right)=K}\right] \sigma^{2}(K, t) K^{2} \\
& +K e^{-\int_{0}^{t} r(u) d u} f(t) E\left[\exp \left(-\int_{0}^{t} \lambda(u) d u\right) S\left(t_{-}\right)^{p} \mathbf{1}_{S\left(t_{-}\right)>K}\right] .
\end{aligned}
$$

It follows that

$$
\begin{aligned}
r(t) C(K, t)+C_{t}= & (r(t)-q(t))\left(C-K C_{K}\right)+\frac{1}{2} K^{2} C_{K K} \sigma^{2}(K, t) \\
& +K f(t) V(t) e^{-\int_{0}^{t} r(u) d u} E^{\widetilde{Q}}\left[S\left(t_{-}\right)^{p} \mathbf{1}_{S\left(t_{-}\right)>K}\right] .
\end{aligned}
$$

We define the truncated power price under $\widetilde{Q}$ as $\zeta(K, t)$ :

$$
\begin{equation*}
\zeta(K, t)=e^{-\int_{0}^{t} r(u) d u} E^{\widetilde{Q}}\left[S\left(t_{-}\right)^{p} \mathbf{1}_{S\left(t_{-}\right)>K}\right] . \tag{11}
\end{equation*}
$$

Hence we construct

$$
\begin{equation*}
\sigma^{2}(K, t)=2 \frac{C_{t}+q(t) C+(r(t)-q(t)) K C_{K}-K f(t) V(t) \zeta(K, t)}{K^{2} C_{K K}} . \tag{12}
\end{equation*}
$$

Both the functions $f(t)$ and $\sigma(K, t)$ may then be recovered from the prices of options under $\widetilde{Q}$. We consider in the next section the procedures for constructing the truncated power prices (11) under $\widetilde{Q}$.

We may usefully rewrite (12) in the form

$$
\begin{align*}
K^{2} \sigma^{2}(K, t)+\frac{K f(t) V(t) \zeta(K, t)}{C_{K K}} & =2 \frac{C_{t}+q(t) C+(r(t)-q(t)) K C_{K}}{C_{K K}}  \tag{13}\\
& =\sigma_{L V}^{2}(K, T),
\end{align*}
$$

where the right-hand side of (13) is the traditional local dollar variance formulation of Dupire [12] and Derman and Kani [11]. We then see that the new local variance is reduced by the survival probability times the expected hazard in the region $S(t)>K$ from where we can jump to zero and earn $K$ dollars on the call, relativized by the density at $K$. We shall report on the relative magnitudes of the diffusion component and the jump component in the partitioning of Dupire local variance.
4. Truncated power prices. We use CDS quotes to construct Weibull density parameters for the survival function $V(t)$. We then transform market prices to default free prices using (7), (8). The dividend yields are adjusted using (6). One may now estimate on this data with these adjusted dividend yields any default free model of option prices. We employ for this purpose the VGSSD model reported in Carr et al. [6]. We thereby estimate the characteristic function of the logarithm of $S(t)$ for each $t$ under the law $\widetilde{Q}$. Our use of this model here is merely in its capacity as an interpolator permitting smooth access to prices of all strikes and maturities as synthesized in the relevant marginal distribution. We subsequently use these prices in (12) for the recovery of the local volatility dynamics.

We now describe the explicit construction of truncated power prices of (11) from the estimated VGSSD default free parameters. We seek the value of

$$
W(K)=\exp \left(-\int_{0}^{t} r(u) d u\right) E^{\widetilde{Q}}\left[S(t)^{p} \mathbf{1}_{S(t)>K}\right] .
$$

We have the characteristic function of

$$
x=\ln (S(t)),
$$

which we denote by $\phi_{x}(u)$. We are interested in

$$
w(k)=e^{-r t} \int_{k}^{\infty}\left(e^{x p}-e^{k}\right) f(x) d x
$$

where $r$ is now the term discount rate.
Consider the Fourier transform

$$
\begin{align*}
\gamma(u) & =e^{-r t} \int_{-\infty}^{\infty} e^{(\alpha+i u) k} \int_{k}^{\infty} e^{x p} f(x) d x d k \\
& =e^{-r t} \int_{-\infty}^{\infty} d x f(x) \int_{-\infty}^{x} e^{x p+(\alpha+i u) k} d k \\
& =e^{-r t} \int_{-\infty}^{\infty} d x f(x) \frac{e^{(\alpha+p+i u) x}}{\alpha+i u} \\
& =e^{-r t} \frac{\phi_{x}(u-i(\alpha+p))}{\alpha+i u} \tag{14}
\end{align*}
$$

Table 1
Weibull parameters from CDS curves.

|  | GM |  | Ford |  |
| :--- | ---: | ---: | ---: | ---: |
|  | $c$ | $a$ | $c$ | $a$ |
| mean | 8.2788 | 1.2610 | 8.5832 | 1.2889 |
| std | 3.0418 | 0.1962 | 3.0655 | 0.1633 |
| max | 15.6194 | 1.5579 | 16.2712 | 1.5469 |
| $\min$ | 3.2041 | 0.7798 | 5.0914 | 0.8949 |

We obtain the truncated power prices by Fourier inversion as

$$
\begin{equation*}
w(k)=\frac{e^{-\alpha k}}{2 \pi} \int_{-\infty}^{\infty} e^{-i u k} \gamma(u) d u \tag{15}
\end{equation*}
$$

These may then be substituted into (12) to obtain the local volatility surface. We compute these at a grid of strikes and maturities for which we seek the local volatilities $\sigma(K, T)$.

Once we have estimated the parameters for the no default process we may compute $\widetilde{C}(K, T)$ at a grid of strikes and maturities. We then use our estimated survival function to define

$$
C(K, T)=V(T) \widetilde{C}(K, T)
$$

These prices are then used in the expression (12) along with the no default truncated power prices computed as per section 4 to find the local volatility surface $\sigma(K, T)$. For the function $f(t)$ we use (10) and the characteristic function for $\log$ prices evaluated at $-i p$.
5. Sample computations for GM and Ford. We now illustrate these procedures for data on GM and Ford from October 2004 to September 2007. We estimated the Weibull parameters for 795 and 784 days, respectively, out of 1085 and 1078 days for GM and Ford, respectively. The results are summarized in Table 1.

The price of a CDS is analytically expressed in terms of the survival function and hence in terms of the parameters $c, a$. A least squares minimization between market and model CDS prices across all maturities results in the estimates for $c, a$ for each company on each day.

We then used the implied Weibull survival functions to adjust market prices to default free prices along with adjusting dividend yields to get the right default free forwards. This is done in line with $(7),(8)$, and (6) with $V(t)$ being analytically specified in the Weibull form given $c, a$.

The resulting candidates for default free prices and dividend yields are employed to fit the interpolating VGSSD model to get the parameters for a smooth version of the default free option prices. These are summarized in Tables 2 and 3 for GM and Ford, respectively.

The next step is the computation of truncated power prices with the no default dynamics and the adjusted dividend yields. These are constructed for an expanding strike range as we raise maturity from 0.05 to 1 . We used the strike range

$$
\begin{aligned}
& k d(t)=p / 1.1-.2 * p * \operatorname{sqrt}\left(t-t_{\min }\right) \\
& k u(t)=p * 1.1+.2 * p * \operatorname{sqrt}\left(t-t_{\min }\right)
\end{aligned}
$$

Table 2
GM default free surface.

|  | $\sigma$ | $\nu$ | $\theta$ | $\gamma$ |
| :--- | :---: | :---: | ---: | :---: |
| mean | 0.2796 | 0.3471 | -0.3579 | 0.3628 |
| std | 0.0743 | 0.3764 | 0.3584 | 0.1014 |
| max | 0.5036 | 2.2549 | 0.6492 | 0.5266 |
| min | 0.0035 | 0 | -2.5097 | 0.0339 |

Table 3
Ford default free surface.

|  | $\sigma$ | $\nu$ | $\theta$ | $\gamma$ |
| :--- | :---: | :---: | ---: | :---: |
| mean | 0.2821 | 0.3155 | -0.1732 | 0.4101 |
| std | 0.0837 | 0.3601 | 0.2949 | 0.0717 |
| max | 0.5381 | 2.4358 | 3.1131 | 0.6925 |
| min | 0.0099 | 0 | -1.2926 | 0.1827 |

where $k d(t)$ is the lowest strike in the grid at maturity $t$, while $k u(t)$ is the highest strike in the grid at maturity $t$, and $p$ here is the initial spot price.

On this grid we construct truncated power prices under the $\widetilde{Q}$ measure by the Fourier inversion described in (15) and (14). We use a power of -1.5 . We now have the power price under $\widetilde{Q}$ as well as the truncated power prices and we can use the Weibull c.d.f. and (10) for $f(t)$ to build the function $f(t)$.

In the next step we construct $\widetilde{Q}$ call prices using the adjusted dividend yields (6) and the smooth interpolation, VGSSD parameters. These prices are then transformed back to market call prices that are defaultable prices via

$$
C(K, t)=V(t) \widetilde{C}(K, t),
$$

where $V(t)$ comes from the Weibull model. We now have all the ingredients to implement the revised local volatility construction of (12). The final output consists of a local volatility function and the function $f(t)$.

These are graphed in Figure 1 for GM and Ford over four to five subsets partitioning levels of the parameter $c$ in the survival function and the level of aggregate volatility of the default free surface as measured by $\sigma^{2}+\theta^{2} \nu$. We observe that when $c$ is high and the mean lifetime is large the deterministic component of the hazard rate is fairly flat. For low mean lifetimes and hence $c$, the deterministic component rises sharply when default free volatility is low but rises more slowly for a high default free volatility. Given a fixed observed volatility in the defaultable market volatilities, there is a trade-off in how volatility splits between the hazard rate and the implied default free structure. The higher the volatility in the hazard rate, the lower it is in the default free options. This trade-off is observed in comparing the local volatility graphs and the hazard rate graphs. The kinks occurring in the tails are numerical consequences of being deep out of the money when the volatility has dropped to a low level.

For Ford we had four subsets, while for GM we had five subsets.
With a view toward summarizing the results we computed the proportion of the traditional


Figure 1. Deterministic components of hazard rate functions for GM and Ford.
local volatility that is allocated to the default component, or the fraction

$$
\sigma_{d e f}^{2}(K, t)=\frac{K f(t) V(t) \zeta(K, t)}{C_{K K} \sigma_{L V}^{2}(K, t)}
$$

This proportion of volatility allocated to the default component is graphed in Figure 2 for GM and Ford. It generally varies with strike and maturity, and we computed the elasticities of this proportion by regressing the logarithm of this ratio on the logarithm of the strike and maturity. Additionally we also regressed the logarithm of $f(t)$ on maturity and the logarithm of maturity. To avoid the effects of cases where these summary functional forms did not fit well, we report summary statistics of these regressions only when the $R^{2}$ exceeded $90 \%$ along with the proportion of times that this criterion was met. Table 4 provides the results for the default proportion of local volatility, while Table 5 presents the results for the deterministic hazard function $f(t)$.

We see from Table 4 that the elasticity with respect to the strike of the default proportion is -0.4768 and -0.3967 for GM and Ford, respectively. The corresponding elasticities with respect to maturity are 0.0646 and 0.0504 , respectively.

We see from Table 5 that the hazard function is generally increasing with maturity with semielasticities and elasticities of 0.1306 and 0.2163 for GM and 0.0975 and 0.2641 being the corresponding values for Ford.


Figure 2. Implied local volatilities for GM and Ford.
Table 4
$\log$ default proportion regressions.

| GM $\left(R^{2}\right.$ criterion satisfaction $\left.79.89 \%\right)$ |  |  |  |
| :--- | :---: | :---: | :---: |
|  | constant | $\log (\mathrm{K} / 100)$ | $\log (t)$ |
| mean | 2.3585 | -0.4768 | 0.0646 |
| std | 1.7332 | 0.3487 | 0.0537 |
| min | 0.7516 | -1.7387 | 0.0042 |
| $\max$ | 8.5069 | -0.1531 | 0.2555 |
| Ford $\left(R^{2}\right.$ criterion | satisfaction | $71.43 \%)$ |  |
| mean | 1.9454 | -0.3967 | 0.0504 |
| std | 1.0498 | 0.2145 | 0.0275 |
| $\min$ | 0.8528 | -1.5003 | 0.0097 |
| $\max$ | 7.3322 | -0.1731 | 0.1937 |

6. Conclusion. We enhance a local volatility model by the addition of the possibility of a single jump to default with a hazard rate that is a deterministic function of time scaled by the stock price raised to a prespecified negative power. Our empirical work uses the prespecified power of -1.5 . We show in this context how one may simultaneously recover from prices of CDS contracts and the equity option prices both the deterministic component of the hazard rate function and revised local volatility. The procedure requires one to construct, after estimating the survival probabilities to various maturities, the prices of default free options to which one fits a standard default free model with revised dividend yields to account for the

## Table 5

$\log f(t)$.

| GM $\left(R^{2}\right.$ criterion satisfaction $\left.88.77 \%\right)$ |  |  |  |
| :--- | :---: | ---: | :---: |
|  | constant | maturity | $\log$ (maturity) |
| mean | 4.3717 | 0.1306 | 0.2163 |
| std | 0.6799 | 0.1654 | 0.1999 |
| min | 3.2663 | -0.4932 | -0.1839 |
| max | 5.7113 | 0.4656 | 0.6203 |
| Ford $\left(R^{2}\right.$ criterion satisfaction $\left.94.33 \%\right)$ |  |  |  |
| mean | 4.1925 | 0.0975 | 0.2641 |
| std | 0.5754 | 0.1234 | 0.1533 |
| min | 3.1920 | -0.5760 | -0.0122 |
| max | 5.4697 | 0.3426 | 0.5740 |

payment of premia necessary to get default free in a defaultable world. This default free model is critically used to infer the prices of powers of the stock price truncated to be above strike levels for a variety of maturities. These truncated power prices are needed in constructing the revised local volatility function from a grid of defaultable call prices that may be inferred from the default free model coupled with the survival function.

The entire procedure was implemented on prices of CDSs and equity options for GM and Ford over the period October 2004 to September 2007. We found that the revised local volatility must be reduced to accommodate the possibility of default by a proportion that is dependent on both the strike and the maturity. On average the elasticity of the default proportion of local volatility is -.4768 and -.3967 for GM and Ford, respectively. The corresponding elasticities with respect to maturity are positive at 0.0646 and 0.0504 . The deterministic component of the hazard function is generally increasing with respect to maturity with semielasticities and elasticities of 0.1306 and 0.2163 for GM and 0.0975 and 0.2641 for Ford.

Acknowledgment. We acknowledge helpful discussions on the subject matter of this paper with Bruno Dupire, Alireza Javaheri, and Liuren Wu. Any remaining errors are our responsibility.

## REFERENCES

[1] C. Albanese and O. Chen, Pricing equity default swaps, Risk, 6 (2005), pp. 83-87.
[2] L. Andersen and D. Buffum, Calibration and implementation of convertible bond models, J. Comput. Finance, 7 (2003/04), pp. 1-34.
[3] M. Atlan and B. Leblanc, Hybrid equity-credit modelling, Risk, 8 (2005), pp. 61-66.
[4] E. Ayache, P. A. Forsyth, and K. R. Vetzal, The valuation of convertible bonds with credit risk, J. Derivatives, 9 (2003), pp. 11-29.
[5] D. Breeden and R. L. Litzenberger, Pricing of state-contingent claims implicit in option prices, J. Business, 51 (1978), pp. 621-651.
[6] P. Carr, H. Geman, D. Madan, and M. Yor, Self-decomposability and option pricing, Math. Finance, 17 (2007), pp. 31-57.
[7] P. Carr and A. Javaheri, The forward PDE for European options on stocks with fixed fractional jumps, Internat. J. Theoret. Appl. Finance, 8 (2005), pp. 239-253.
[8] P. Carr and V. Linetsky, A jump to default extended CEV model: An application of Bessel processes, Finance Stoch., 10 (2006), pp. 303-330.
[9] M. Davis and F. Lischka, Convertible bonds with market risk and credit risk, in Applied Probability, AMS/IP Stud. Adv. Math. 26, AMS, Providence, RI, 2002, pp. 45-58.
[10] C. Dellacherie and P. Meyer, Probabilités et Potentiel, Theorie des Martingales, Hermann, Paris, 1980.
[11] E. Derman and I. Kani, Riding on a smile, Risk, 7 (1994), pp. 32-39.
[12] B. Dupire, Pricing with a smile, Risk 7 (1994), pp. 18-20.
[13] R. A. Jarrow and S. Turnbull, Pricing options on financial securities subject to default risk, J. Finance, 50 (1995), pp. 53-86.
[14] V. Linetsky, Pricing equity derivatives subject to bankruptcy, Math. Finance, 16 (2006), pp. 255-282.
[15] D. Madan, M. Konikov, and M. Marinescu, Credit and basket default swaps, J. Credit Risk, 2 (2006), pp. 67-87.
[16] P. Meyer, Un Cours sur les Intégrales stochastiques, in Séminaire de Probabilités X, Lecture Notes in Math. 511, Springer-Verlag, Berlin, 1976, pp. 245-400.
[17] Y. Ren, D. Madan, and M. Qian, Calibrating and pricing with embedded local volatility models, Risk, 9 (2007), pp. 138-143.
[18] M. Yor, Rappels et Préliminaires Généraux, in Temps Locaux, Astérique 52/53, J. Azéma and M. Yor, eds., Société Mathématique de France, Paris, 1978, pp. 17-22.

# Minimizing the Expected Market Time to Reach a Certain Wealth Level* 

Constantinos Kardaras ${ }^{\dagger}$ and Eckhard Platen ${ }^{\ddagger}$


#### Abstract

In a financial market model, we consider variations of the problem of minimizing the expected time to upcross a certain wealth level. For exponential Lévy markets, we show the asymptotic optimality of the growth-optimal portfolio for the above problem and obtain tight bounds for the value function for any wealth level. In an Itô market, we employ the concept of market time, which is a clock that runs according to the underlying market growth. We show the optimality of the growth-optimal portfolio for minimizing the expected market time to reach any wealth level. This reveals a general definition of market time which can be useful from an investor's point of view. We utilize this last definition to extend the previous results in a general semimartingale setting.


Key words. numéraire portfolio, growth-optimal portfolio, market time, upcrossing, overshoot, exponential Lévy markets, Itô markets, semimartingale markets

AMS subject classifications. 60H99, 60G44, 91B28, 91B70
DOI. 10.1137/080741124

1. Introduction. The problem of quickly reaching certain goals in wealth management is one of the most fundamental tasks in the theory and practice of finance. However, making this idea mathematically precise has been a challenge. In particular, this would require a quantification of what is meant by achieving goals "quickly" in a model-independent manner, or, even better, coming endogenously from the description of the market as perceived by its participants. Such a mathematically precise description of the flow of time, as well as the corresponding optimal investment strategy, is clearly valuable. If a robust, model-independent answer to the previous questions can be given, it would go a long way towards a better understanding of the problem, as its statement should provide a deep insight into key quantitative characteristics of the market. Our aim in this paper is to present a way of addressing the aforementioned issues.

We proceed with a more thorough description of the problem. Imagine an investor holding some minute capital-in-hand, aiming to reach as quickly as possible a substantial wealth level by optimally choosing an investment opportunity in an active market. No matter what the mathematical formalization of the objective is, as long as it reasonably describes the above informal setting, intuition suggests that the investor should pick an aggressive strategy that provides ample wealth growth. The most famous wealth-optimizing strategy that could

[^1]potentially achieve this is the growth-optimal strategy, which is sometimes also called the Kelly strategy, as the latter was introduced in [16]. Therefore, the portfolio generated by the growth-optimal strategy is a strong candidate for solving the aforementioned problem, at least in an approximate sense. This last point is augmented by the long line of research on the importance and optimality properties of the growth-optimal portfolio; we mention, for example, the following very incomplete list: [17], [1], [3], [18], [7], [12]. Note also that minimizing expected time to reach a wealth level is not the only interesting objective that one can seek. For example, maximizing the probability that a wealth level will be reached before some future time is also interesting; in this respect, see [6], [9].

Here, we shall identify a variant of the "quickest goal reach" problem for continuous-time models where the growth-optimal portfolio is indeed the best. The problem we consider then is that of minimizing the expected market time that it will take to reach a certain wealth level. Market time will be defined as a natural time scale which runs fast when the compensation for taking risk in the market is high and vice versa. In a market with continuous asset prices, this will be achieved by setting the slope of the market time equal to half the squared risk premium. In this case, it equals the growth rate of the corresponding growth-optimal portfolio, which leads to the interpretation of market time as integrated maximum growth rate.

The first attempt to minimize the expected upcrossing time in a discrete-time gamblingsystem model was described in [5], where indeed the near optimal wealth process was found to be characterized by Kelly's growth-optimal strategy. Models of gambling systems, as considered in [5], could be interpreted as discrete-time financial markets where the log-asset-price processes are random walks with a finite number of possible values for the increment of each step. The natural continuous-time generalization of the above setting is to consider exponential Lévy markets, i.e., markets where the log-asset-price processes have independent and stationary increments. For these markets, we establish here the exact analogues of the results in [5].

A continuous-time problem in the context of a Black-Scholes market was treated in [11], and then as an application of a more abstract problem in [10], essentially using methods of dynamic programming. In this case, the numéraire portfolio of the market, which was introduced in [17] and is also called the growth-optimal portfolio, as it is generated by the analogue of Kelly's growth-optimal strategy, is truly optimal for minimizing the expected calendar time to reach any wealth level. Unfortunately, the moment that one considers more complex Itô-process models, for example ones that are modelling feedback effects, such as the leverage effect in [4], the growth-optimal portfolio is no longer optimal for the problem of minimizing expected calendar time for upcrossing a certain wealth level. In fact, for general non-Markovian models there does not seem to be any hope in identifying what the optimal strategy and wealth process are when minimizing expected calendar time. We note, however, that for Markovian models one can still characterize the optimal strategy and portfolio in terms of a Hamilton-Jacobi-Bellman equation, which will most likely then have to be solved numerically.

We introduce in this paper a market clock which does not count time according to the natural calendar flow but rather according to the overall market growth. Under the objective that one minimizes expected market time, we show here that the solution again yields the growth-optimal portfolio as nearly optimal. There is a slight problem that results in the
nonoptimality of the growth-optimal portfolio, if for finite wealth levels some overshoot is possible over the targeted wealth level at the time of the upcrossing. If there is no overshoot, which happens in particular in models with continuous asset prices, then the growth-optimal portfolio is indeed optimal. In [2], the author considers a ramification of the problem by offering a rebate for the overshoot that results in the growth-optimal portfolio again being optimal. Of course, we could do this even in the most general case. Since this rebate inclusion is somewhat arbitrary, we shall refrain from using it in our own analysis.

The optimality of the growth-optimal portfolio for minimizing expected time according to a clock counting time according to the overall market growth sounds a bit like a tautological statement. However, we shall make a conscious effort to convey that the concept of market time is very natural, by taking a stepwise approach in the model generality that we consider. The exponential Lévy process case is considered first. There, the market-time flow coincides with the calendar-time flow up to a multiplicative constant, since the model coefficients remain constant through time. As soon as the model coefficients are allowed to randomly change, one can regard the passage of time in terms of the opportunities for profit that are available. We first discuss this in the realm of markets where asset-prices are modeled via Itô processes, where the arguments are more intuitive. As soon as the natural candidate for the market time is understood, we proceed to discuss the results in the very general semimartingale model.

The results presented in this work are generalizations of the constant-coefficient result in [11]. The use of martingale methods and a natural definition of market time that we utilize make the proof of our claims more transparent and widens the scope and validity of the corresponding statements.

The structure of the paper is as follows. In section 2 we introduce the general financial market model, we define the problem of minimizing expected market time, and we present the standing assumptions, which are basically the existence of the numéraire portfolio. In section 3 we specialize in the case of exponential Lévy market models, where market time and calendar time coincide up to a multiplicative constant. Our first main result gives tight bounds for the near optimal performance of the growth-optimal portfolio for any wealth level that also result in its asymptotic optimality for increasing wealth levels. In section 4 we use Itô processes to model the market. After some discussion on the concept of market time, our second main result also shows here the optimality of the growth-optimal portfolio. In section 5 , the concept of market time in a general semimartingale setting is introduced and a general result that covers all previous cases is presented. Finally, section 6 contains the proofs of the results in the previous sections.
2. Description of the problem. In the following general remarks we fix some notation that will be used throughout.

By $\mathbb{R}_{+}$we shall denote the positive real line, $\mathbb{R}^{d}$ the $d$-dimensional Euclidean space, and $\mathbb{N}$ the set of natural numbers $\{1,2, \ldots\}$. Superscripts will be used to indicate coordinates, both for vectors and for processes; for example $z \in \mathbb{R}^{d}$ is written $z=\left(z^{1}, \ldots, z^{d}\right)$. On $\mathbb{R}^{d},\langle\cdot, \cdot\rangle$ will denote the usual inner product: $\langle y, z\rangle:=\sum_{i=1}^{d} y^{i} z^{i}$ for $y$ and $z$ in $\mathbb{R}^{d}$. Also $|\cdot|$ will denote the usual norm: $|z|:=\sqrt{\langle z, z\rangle}$ for $z \in \mathbb{R}^{d}$.

On $\mathbb{R}_{+}$equipped with the Borel $\sigma$-field $\mathcal{B}\left(\mathbb{R}_{+}\right)$, Leb will denote the Lebesgue measure.
All stochastic processes appearing in what follows are defined on a filtered probability
space $(\Omega, \mathcal{F}, \mathbf{F}, \mathbb{P})$. Here, $\mathbb{P}$ is a probability on $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is a $\sigma$-algebra that will make all involved random variables measurable. The filtration $\mathbf{F}=\left(\mathcal{F}_{t}\right)_{t \in \mathbb{R}_{+}}$is assumed to satisfy the usual hypotheses of right-continuity and saturation by $\mathbb{P}$-null sets. It will be assumed throughout that $\mathcal{F}_{0}$ is trivial modulo $\mathbb{P}$.

For a càdlàg (right-continuous with left limits) stochastic process $X=\left(X_{t}\right)_{t \in \mathbb{R}_{+}}$, define $X_{t-}:=\lim _{s \uparrow t} X_{s}$ for $t>0$ and $X_{0-}:=0$. The process $X_{-}$will denote this last left-continuous version of $X$, and $\Delta X:=X-X_{-}$will be the jump process of $X$.
2.1. Assets and wealth processes. The $d$-dimensional semimartingale $S=\left(S^{1}, \ldots, S^{d}\right)$ will be denoting the discounted, with respect to the savings account, price process of $d$ financial assets.

Starting with initial capital $x \in \mathbb{R}_{+}$, and investing according to some predictable and $S$-integrable strategy $\vartheta$, an investor's discounted total wealth process is given by

$$
\begin{equation*}
X^{x, \vartheta}:=x+\int_{0}\left\langle\vartheta_{t}, \mathrm{~d} S_{t}\right\rangle \tag{2.1}
\end{equation*}
$$

Reflecting the investor's ability to hold only a portfolio of nonnegative total tradeable wealth, we then define the set of all nonnegative wealth processes starting from initial capital $x \in \mathbb{R}_{+}$:

$$
\mathcal{X}(x):=\left\{X^{x, \vartheta} \text { as in (2.1) } \mid \vartheta \text { is predictable and } S \text {-integrable, and } X^{x, \vartheta} \geq 0\right\} .
$$

It is straightforward that $\mathcal{X}(x)=x \mathcal{X}(1)$ and that $x \in \mathcal{X}(x)$ for all $x \in \mathbb{R}_{+}$. We also set $\mathcal{X}:=\bigcup_{x \in \mathbb{R}_{+}} \mathcal{X}(x)$.
2.2. The problem. We shall be concerned with the problem of quickly reaching a wealth level $\ell$ starting from capital $x$. This, of course, is nontrivial only when $x<\ell$, which will be tacitly assumed throughout. The challenge is now to rigorously define what is meant by "quickly." Take $\mathcal{O}=\left(\mathcal{O}_{t}\right)_{t \in \mathbb{R}_{+}}$to be an increasing and adapted process such that, $\mathbb{P}$-a.s., $\mathcal{O}_{0}=0$ and $\mathcal{O}_{\infty}=+\infty$. $\mathcal{O}$ will be representing some kind of internal clock of the market, which we shall call market time. In the following sections we shall be more precise on choosing $\mathcal{O}$, guided by what we shall learn when identifying the consequences of applying the growthoptimal strategy.

For any càdlàg process $X$ and $\ell \in \mathbb{R}_{+}$, define the first upcrossing market time of $X$ at level $\ell$ :

$$
\begin{equation*}
\mathcal{T}(X ; \ell):=\inf \left\{\mathcal{O}_{t} \in \mathbb{R}_{+} \mid X_{t} \geq \ell\right\} \tag{2.2}
\end{equation*}
$$

Of course, if $\ell \leq x$, then $\mathcal{T}(X ; \ell)=0$ for all $X \in \mathcal{X}(x)$. With the aforementioned inputs, define for all $x<\ell$ the value function

$$
\begin{equation*}
v(x ; \ell):=\inf _{X \in \mathcal{X}(x)} \mathbb{E}[\mathcal{T}(X ; \ell)] \tag{2.3}
\end{equation*}
$$

Our aims in this work are to

- identify a natural definition for the market time $\mathcal{O}$,
- obtain an explicit formula, or at least some useful tight bounds, for the value function $v(x ; \ell)$ of (2.3), and
- find the optimal, or perhaps near optimal, portfolio for the above problem.
2.3. Standing assumptions. In order to make headway with the problem described in section 2.2 , we shall make two natural and indispensable assumptions regarding the financial market that will be in force throughout.

Assumptions 2.1. In our financial market model, we assume the following:
(1) There exists $\widehat{X} \in \mathcal{X}(1)$ such that $X / \widehat{X}$ is a supermartingale for all $X \in \mathcal{X}$.
(2) For every $\ell \in \mathbb{R}_{+}$, there exists $X \in \mathcal{X}(1)$, possibly depending on $\ell$, such that, $\mathbb{P}$-a.s., $\mathcal{T}(X ; \ell)<+\infty$.
A process $\widehat{X}$ with the properties described in Assumption 2.1(1) is unique and is called the numéraire portfolio. Existence of the numéraire portfolio is a minimal assumption for the viability of the financial market. It is essentially equivalent to the boundedness in probability of the set $\left\{X_{T} \mid X \in \mathcal{X}(1)\right\}$ of all possible discounted wealth starting from unit capital and observed at any time $T \in \mathbb{R}_{+}$. We refer the interested reader to [7], [12], and [15] for more information in this direction. We shall frequently refer to the numéraire portfolio as the growth-optimal portfolio, as the two notions coincide.

Assumption 2.1(2) constitutes what has been coined a "favorable game" in [5], and it is necessary in order for the problem described in (2.3) to have finite value and therefore to be well-posed. Under Assumption 2.1(2), and in view of the property $\mathcal{X}(x)=x \mathcal{X}(1)$ for $x \in \mathbb{R}_{+}$, it is obvious that for all $x \in \mathbb{R}_{+}$and $\ell \in \mathbb{R}_{+}$there exists $X \in \mathcal{X}(x)$ such that $\mathbb{P}[\mathcal{T}(X ; \ell)<+\infty]=1$.

Actually, if Assumption 2.1(1) is in force, Assumption 2.1(2) has a convenient equivalent.
Proposition 2.2. Under Assumption 2.1(1), Assumption 2.1(2) is equivalent to
(2') $\lim _{t \rightarrow+\infty} \widehat{X}_{t}=+\infty, \mathbb{P}$-a.s.
This last result enables one to easily check the validity of Assumptions 2.1 by looking only at the numéraire portfolio. In each of the specific cases we shall consider in what follows, equivalent characterizations of Assumptions 2.1 will be given in terms of the model under consideration.

## 3. Exponential Lévy markets.

3.1. The setup. For this section we assume that the discounted asset-price processes satisfy $\mathrm{d} S_{t}^{i}=S_{t-}^{i} \mathrm{~d} R_{t}^{i}$ for $t \in \mathbb{R}_{+}$, where, for all $i=1, \ldots, d, R^{i}$ is a Lévy process on $(\Omega, \mathcal{F}, \mathbf{F}, \mathbb{P})$. Each $R^{i}$ for $i=1, \ldots, d$ is the total returns process associated with $S^{i}$.

In order to make sure that the asset-price processes remain nonnegative, it is necessary and sufficient that $\Delta R^{i} \geq-1$ for all $i=1, \ldots, d$. We shall actually impose a further restriction on the structure of the jumps of the returns processes, also bounding them from above. This is mostly done in order to obtain later in Theorem 3.3 a statement which parallels the result in [5]. For the asymptotic result that will be presented in section 3.6 this bounded-jump assumption will be dropped.

Assumption 3.1. For all $i=1, \ldots, d$ we have $-1 \leq \Delta R^{i} \leq \kappa$ for some $\kappa \in \mathbb{R}_{+}$.
Denote by $R$ the $d$-dimensional Lévy process $\left(R^{1}, \ldots, R^{d}\right)$. In view of the boundedness of the jumps of $R$, as stated in Assumption 3.1, we can write

$$
\begin{equation*}
R_{T}=a T+\sigma W_{T}+\int_{[0, T] \times \mathbb{R}^{d}} z(\mu(\mathrm{~d} z, \mathrm{~d} t)-\nu(\mathrm{d} z) \mathrm{d} t) \tag{3.1}
\end{equation*}
$$

for all $T \in \mathbb{R}_{+}$. In view of Assumption 3.1, the elements in the above representation satisfy the following:

- $a \in \mathbb{R}^{d}$.
- $\sigma$ is a $(d \times m)$-matrix, where $m \in \mathbb{N}$.
- $W$ is a standard $m$-dimensional Brownian motion on $(\Omega, \mathcal{F}, \mathbf{F}, \mathbb{P})$.
- $\mu$ is the jump measure of $R$, i.e., the random counting measure on $\mathbb{R}_{+} \times \mathbb{R}^{d}$ defined via $\mu([0, T] \times E):=\sum_{0 \leq t \leq T} \mathbb{I}_{E \backslash\{0\}}\left(\Delta R_{t}\right)$ for $T \in \mathbb{R}_{+}$and $E \subseteq \mathbb{R}^{d}$.
- $\nu$, the compensator of $\mu$, is a Lévy measure on $\left(\mathbb{R}^{d}, \mathcal{B}\left(\mathbb{R}^{d}\right)\right)$, where $\mathcal{B}\left(\mathbb{R}^{d}\right)$ is the Borel $\sigma$-field on $\mathbb{R}^{d}$. More precisely, $\nu$ is a measure with $\nu[\{0\}]=0, \nu\left[\mathbb{R}^{d} \backslash[-1, \kappa]\right]=0$, and $\int_{\mathbb{R}^{d}}|x|^{2} \nu[\mathrm{~d} x]<+\infty$.
For more information on Lévy processes one can check, for example, [19].
Define the $(d \times d)$ matrix $c:=\sigma \sigma^{\top}$, where "丁" denotes matrix transposition. The triplet $(a, c, \nu)$ will play a crucial role in the discussion below.

In the notation of (2.1), let $X^{x, \vartheta} \in \mathcal{X}(x)$. The nonnegativity requirement $X^{x, \vartheta} \geq 0$ is equivalent to $\Delta X^{x, \vartheta} \geq X_{-}^{x, \vartheta}$, or further to $\langle\vartheta, \Delta S\rangle \geq X_{-}^{x, \vartheta}$. Since $\Delta S^{i}=S_{-}^{i} \Delta R^{i}$ for each $i=1, \ldots, d$, and recalling that $\nu$ is the Lévy measure of $R$, we conclude that $X^{x, \vartheta} \geq 0$ if and only if

$$
\left(\vartheta_{t}^{i}(\omega) S_{t-}^{i}(\omega)\right)_{i=1, \ldots, d} \in X_{t-}^{x, \vartheta}(\omega) \mathfrak{C} \quad \text { for all }(\omega, t) \in \Omega \times \mathbb{R}_{+},
$$

where $\mathfrak{C}$ is the set of natural constraints defined via

$$
\mathfrak{C}:=\left\{\eta \in \mathbb{R}^{d} \mid \nu\left[z \in \mathbb{R}^{d} \mid\langle\eta, z\rangle<-1\right]=0\right\} .
$$

It is easy to see that $\mathfrak{C}$ is convex; it is also closed, as follows from Fatou's lemma.
3.2. Growth rate. For any $\pi \in \mathfrak{C}$, define

$$
\begin{equation*}
\mathfrak{g}(\pi):=\langle\pi, a\rangle-\frac{1}{2}\langle\pi, c \pi\rangle-\int_{\mathbb{R}^{d}}[\langle\pi, z\rangle-\log (1+\langle\pi, z\rangle)] \nu[\mathrm{d} z] . \tag{3.2}
\end{equation*}
$$

For $\pi \in \mathfrak{C}, \mathfrak{g}(\pi)$ is the drift rate of the logarithm of the wealth process $X \in \mathcal{X}(1)$ that satisfies $\mathrm{d} X_{t}=X_{t-}\left\langle\pi, \mathrm{d} R_{t}\right\rangle=X_{t-} \mathrm{d}\left\langle\pi, R_{t}\right\rangle$ for all $t \in \mathbb{R}_{+}$; for this reason, $\mathfrak{g}(\pi)$ is also called the growth rate of the last wealth process.

Define $\mathfrak{g}^{*}:=\sup _{\pi \in \mathfrak{C}} \mathfrak{g}(\pi)$ to be the maximum growth rate. Since $0 \in \mathfrak{C}$, we certainly have $\mathfrak{g}^{*} \geq \mathfrak{g}(0)=0$. Actually, under the bounded-jump Assumption 3.1, the standing Assumptions 2.1 are equivalent to $0<\mathfrak{g}^{*}<\infty$. In order to achieve this last claim, we shall connect the viability of the market with the concept of immediate arbitrage opportunities, as will now be introduced.
3.3. Market viability. Define the set $\mathfrak{I}$ of immediate arbitrage opportunities to consist of all vectors $\xi \in \mathbb{R}^{d}$ such that $c \xi=0, \nu\left[z \in \mathbb{R}^{d} \mid\langle\xi, z\rangle<0\right]=0$, and $\langle\xi, a\rangle \geq 0$ and where further at least one of $\nu\left[z \in \mathbb{R}^{d} \mid\langle\xi, z\rangle>0\right]>0$ or $\langle\xi, a\rangle>0$ holds. As part of the next result, we get that the previously described exponential Lévy market is viable if and only if the intersection of $\mathfrak{I}$ with the recession cone of $\mathfrak{C}$, defined as $\check{\mathfrak{C}}:=\bigcap_{u>0} u \mathfrak{C}$, is empty.

Proposition 3.2. Assumptions 2.1 are equivalent to requiring both $\mathfrak{I} \cap \check{\mathfrak{C}}=\emptyset$ and $\mathfrak{g}^{*}>0$.
Suppose now that the above is true, as well as that Assumption 3.1 is in force. Then, $\mathfrak{g}^{*}<\infty$ and there exists $\rho \in \mathfrak{C}$ such that $\mathfrak{g}(\rho)=\mathfrak{g}^{*}$. Furthermore, the numéraire portfolio $\widehat{X}$ satisfies the dynamics $\mathrm{d} \widehat{X}_{t}=\widehat{X}_{t-}\left\langle\rho, \mathrm{d} R_{t}\right\rangle=\widehat{X}_{t-} \mathrm{d}\left\langle\rho, R_{t}\right\rangle$. In other words, for $T \in \mathbb{R}_{+}$,

$$
\begin{equation*}
\log \left(\widehat{X}_{T}\right)=\left\langle\rho, R_{T}\right\rangle-\frac{1}{2}\langle\rho, c \rho\rangle T-\sum_{0 \leq t \leq T}\left(\left\langle\rho, \Delta R_{t}\right\rangle-\log \left(1+\left\langle\rho, \Delta R_{t}\right\rangle\right)\right) . \tag{3.3}
\end{equation*}
$$

Instead of using the general Assumptions 2.1 in this section, we shall use the equivalent conditions $\mathfrak{I} \cap \check{\mathfrak{C}}=\emptyset$ and $\mathfrak{g}^{*}>0$. We also note that the vector $\rho \in \mathfrak{C}$ in the statement of Proposition 3.2 that leads to the numéraire portfolio is essentially unique, modulo any degeneracies that might be present in the market and lead to nonzero portfolios having zero returns.
3.4. The main result. Since Lévy processes have stationary and independent increments, the natural candidate for market time is to consider calendar time up to a multiplicative constant $\gamma>0$, i.e., to set $\mathcal{O}_{t}=\gamma t$ for $t \in \mathbb{R}_{+}$. In Theorem 3.3, we shall actually choose $\gamma=\mathfrak{g}^{*}$. This turns out to be the appropriate choice of market velocity that reflects a universal characteristic of the market and will result in the bounds (3.4) for the optimal upcrossing time in Theorem 3.3 not depending on the actual model under consideration.

Theorem 3.3. We work under Assumption 3.1 and also assume that $\mathfrak{I} \cap \check{\mathfrak{C}}=\emptyset$ and $\mathfrak{g}^{*}>0$. Define the finite nonnegative constant $\alpha:=\inf \left\{\beta \in \mathbb{R}_{+} \mid \nu\left[z \in \mathbb{R}^{d} \mid\langle\rho, z\rangle>\beta\right]=0\right\}$. Let the market time $\mathcal{O}$ be defined via $\mathcal{O}_{t}=\mathfrak{g}^{*} t$ for all $t \in \mathbb{R}_{+}$. With $\widehat{X}(x):=x \widehat{X}$, we have the inequalities

$$
\begin{equation*}
\log \left(\frac{\ell}{x}\right) \leq v(x ; \ell) \leq \mathbb{E}[\mathcal{T}(\widehat{X}(x) ; \ell)] \leq \log \left(\frac{\ell}{x}\right)+\log (1+\alpha) \tag{3.4}
\end{equation*}
$$

Actually, Theorem 3.3 is an instance of a more general statement that will be presented in section 5 . We note that the bounds (3.4) are in complete accordance with the discrete-time result in [5] and that the nonnegative constant $\log (1+\alpha)$ does not involve $x$ or $\ell$.

Remark 3.4. Under a mild condition, namely that the marginal one-dimensional distributions of $\log (\widehat{X})$ are nonlattice, the overshoot of $\log (\widehat{X})$ over the level $\log (\ell)$ actually has a limiting distribution as $\ell \rightarrow \infty$ that is supported on $[0, \log (1+\alpha)]$. In that case,

$$
\lim _{\ell \rightarrow \infty}\left(\mathbb{E}[\mathcal{T}(\widehat{X}(x) ; \ell)]-\log \left(\frac{\ell}{x}\right)\right)
$$

exists and is exactly equal to the mean of that limiting distribution.
3.5. True optimality. There is a special case when the growth-optimal portfolio is indeed optimal for all levels $\ell$, which covers in particular the Black-Scholes market result in [11]. The following result directly stems out of the statement of Theorem 3.3.

Corollary 3.5. Suppose that the numéraire portfolio $\widehat{X}$ of (3.3) has no positive jumps: $\langle\rho, \Delta R\rangle \leq 0$. Then,

$$
v(x ; \ell)=\log \left(\frac{\ell}{x}\right)=\mathbb{E}[\mathcal{T}(\widehat{X}(x) ; \ell)]
$$

For an easy example where the last equality occurs, consider in (3.1) the case where $d=1$, $\kappa=0$, and $a=a^{1}>0$. This is a reasonable model where the excess rate of return is strictly positive and only negative jumps are present in the dynamics of the discounted asset-price process.
3.6. Asymptotic optimality without the bounded-jump assumption. Theorem 3.3 gives the asymptotic (for large $\ell$ ) optimality of the growth-optimal portfolio, since, by (3.4),

$$
\begin{equation*}
\lim _{\ell \rightarrow \infty} \frac{v(x ; \ell)}{\log (\ell)}=1=\lim _{\ell \rightarrow \infty} \frac{\mathbb{E}[\mathcal{T}(\widehat{X}(x) ; \ell)]}{\log (\ell)} . \tag{3.5}
\end{equation*}
$$

The validity of the asymptotic optimality in (3.5) goes well beyond the bounded-jump Assumption 3.1, as we shall describe now. For the total returns process $R=\left(R^{1}, \ldots, R^{d}\right)$, we can write the canonical representation (3.1) if and only if the Lévy measure $\nu$ is such that $\int_{\mathbb{R}^{d}}\left(|x| \wedge|x|^{2}\right) \nu[\mathrm{d} x]<+\infty$. In that case, the definition in (3.2) of the growth rate is still the same, even without the validity of Assumption 3.1. We then have the following result.

Proposition 3.6. Suppose that the canonical representation (3.1) is valid. Then, if $\mathfrak{I} \cap \check{\mathfrak{C}}=\emptyset$ and $\mathfrak{g}^{*}>0$ hold, we have $\mathfrak{g}^{*}<\infty$ and there exists $\rho \in \mathfrak{C}$ such that $\mathfrak{g}(\rho)=\mathfrak{g}^{*}$. One can then define the growth-optimal portfolio $\widehat{X}$ using (3.3). Defining $\mathcal{O}$ via $\mathcal{O}_{t}=\mathfrak{g}^{*}$ t, and with $\widehat{X}(x):=x \widehat{X}$, the asymptotics (3.5) hold.
4. Itô markets and market time. As already mentioned in the introduction, the growthoptimal portfolio is not optimal for the problem of minimizing the expected calendar time to reach a wealth level when considering models where the coefficients may change randomly through time. If the objective is somewhat altered into minimizing expected market time, as we shall define below, then the growth-optimal portfolio is indeed optimal. It is our belief that the notion of market time, as it naturally emerges in our paper, has a very clear and natural interpretation and makes deep sense, and is therefore worth studying beyond the context of the questions raised.

To keep the technical details simple, in this section we assume that $S$ is an Itô process. Later, in section 5 , we shall see how to relax this assumption to more complex models and still keep the main result holding.
4.1. The setup. The dynamics of the discounted asset-prices are

$$
\begin{equation*}
\mathrm{d} S_{t}^{i}=S_{t}^{i}\left(a_{t}^{i} \mathrm{~d} t+\sum_{j=1}^{m} \sigma_{t}^{i j} \mathrm{~d} W_{t}^{j}\right) \tag{4.1}
\end{equation*}
$$

for each $i=1, \ldots, d$ and $t \in \mathbb{R}_{+}$. Here $a=\left(a^{i}\right)_{i=1, \ldots, d}$ is the predictable $d$-dimensional process of excess appreciation rates, $\sigma=\left(\sigma^{i j}\right)_{i=1, \ldots, d, j=1, \ldots, m}$ is a predictable $(d \times m)$-matrix-valued process of volatilities, and $W=\left(W^{j}\right)_{j=1, \ldots, m}$ is a standard $m$-dimensional Brownian motion on $(\Omega, \mathcal{F}, \mathbf{F}, \mathbb{P})$. We let $c:=\sigma \sigma^{\top}$ denote the $(d \times d)$-matrix-valued process of local covariances.
4.2. Assumptions. The general Assumptions 2.1 have a well-described equivalent for the Itô market we are considering.

Proposition 4.1. Assumptions 2.1 are equivalent to the following:
(1) There exists a d-dimensional predictable process $\rho$ such that, $(\mathbb{P} \otimes$ Leb $)$-a.e., $c \rho=a$. (In that case, $\rho=c^{\dagger} a$, where $c^{\dagger}$ is the Moore-Penrose pseudoinverse of $c$.)
(2) $\int_{0}^{T}\left|\lambda_{t}\right|^{2} \mathrm{~d} t<\infty$ for all $T \in \mathbb{R}_{+}$, where $\lambda:=\sigma^{\top} c^{\dagger} a$ is the $m$-dimensional risk premium process. (Then, $|\lambda|^{2}=\left\langle a, c^{\dagger} a\right\rangle=\langle\rho, c \rho\rangle$.)
(3) $\int_{0}^{\infty}\left|\lambda_{t}\right|^{2} \mathrm{~d} t=\infty, \mathbb{P}$-a.s.

In this case, it follows that the logarithm of the numéraire portfolio $\widehat{X}$ is given by

$$
\begin{equation*}
\log (\widehat{X})=\frac{1}{2} \int_{0}\left|\lambda_{t}\right|^{2} \mathrm{~d} t+\int_{0} \lambda_{t} \mathrm{~d} W_{t} . \tag{4.2}
\end{equation*}
$$

It follows from (4.2) that $\mathfrak{g}_{t}^{*}:=(1 / 2)\left|\lambda_{t}\right|^{2}$ equals the maximum growth rate at time $t \in \mathbb{R}_{+}$ in the given Itô market.

As we did in the case of exponential Lévy markets, we shall use statements (1), (2), and (3) of Proposition 4.1 in place of the general Assumptions 2.1 in what follows.
4.3. Market time. With the above notation define now, similar to the previous section, the market time process $\mathcal{O}=\left(\mathcal{O}_{t}\right)_{t \in \mathbb{R}_{+}}$by setting it equal to the integral over the maximum growth rate, i.e.,

$$
\mathcal{O}_{t}:=\int_{0}^{t} \mathfrak{g}_{s}^{*} \mathrm{~d} s=\frac{1}{2} \int_{0}^{t}\left|\lambda_{s}\right|^{2} \mathrm{~d} s
$$

for $t \in \mathbb{R}_{+}$. Observe that, under the validity of statements (1), (2), and (3) of Proposition 4.1, we have $\mathbb{P}\left[\mathcal{O}_{\infty}=\infty\right]=1$ as follows from Proposition 4.1(3). As explained in section 2.2, for given $x<\ell$, our aim is to find the wealth process $X \in \mathcal{X}(x)$ that minimizes $\mathbb{E}[\mathcal{T}(X ; \ell)]$.

We briefly explain why the problem of minimizing expected market time to reach a wealth level using such a random clock and not calendar time is natural and worth studying. Consider for simplicity the one-asset case $d=1$. Then, at any time $t \in \mathbb{R}_{+},\left|\lambda_{t}\right|^{2}=\left|a_{t} / \sigma_{t}\right|^{2}$ is the "squared signal to noise ratio" of the asset-price process or more precisely the squared risk premium. When this quantity is small, the opportunities for making profits over those obtainable from the savings account are rather small; on the other hand, when $\left|\lambda_{t}\right|^{2}$ is large, at time $t \in \mathbb{R}_{+}$an investor has a lot of opportunities to use the favorable fact that the premium for taking risk is high. Stalling to reach the wealth level $\ell$ when opportunities are favorable should be punished more severely, especially for fund managers, and this is exactly what the market time $\mathcal{O}$ does. From an economic point of view, market time simply conforms with the underlying growth of the market.
4.4. The main result. We are ready to present the solution to the optimization problem of section 2.2, both giving an expression for the value function $v$ and again showing that the growth-optimal portfolio is optimal.

Theorem 4.2. Under the validity of statements (1), (2), and (3) of Proposition 4.1 for an Itô market, and with $\widehat{X}(x):=x \widehat{X} \in \mathcal{X}(x)$, for $x<\ell$ we have

$$
v(x ; \ell)=\log \left(\frac{\ell}{x}\right)=\mathbb{E}[\mathcal{T}(\widehat{X}(x) ; \ell)] .
$$

Once again, this last result is a special case of Theorem 5.3 that will be presented in the next section.
5. Market time in general semimartingale markets. The purpose of this section is to give a wide-encompassing definition of market time for semimartingale financial markets and to present a general result on the expected market time to reach a given wealth level, of which both Theorems 3.3 and 4.2 are special cases. We are now in the very general market model described in section 2.
5.1. Market time. Guided by the discussions and results in both the exponential Lévy market case of section 3 and the Itô market case of section 4, it makes sense to define market time as the underlying optimal growth of the market, i.e., the drift part of the logarithm of the growth-optimal portfolio. We shall have to make minimal assumptions for market time to be well defined, namely, that the drift part of the logarithm of the growth-optimal portfolio does exist. The following result, which is a refined version of Proposition 2.2, ensures that the discussions that follow make sense.

Proposition 5.1. Under the validity of Assumption 2.1(1), further assume that the logarithm of the numéraire portfolio $\widehat{X}$ is a special semimartingale and write $\log (\widehat{X})=\mathcal{O}+M$ for its canonical decomposition, where $\mathcal{O}$ is a predictable nondecreasing process and $M$ is a local martingale. Then, Assumption 2.1(2) is equivalent to
$\left(2^{\prime \prime}\right) \lim _{t \rightarrow+\infty} \mathcal{O}_{t}=+\infty, \mathbb{P}$-a.s.
The following slightly strengthened version of Assumptions 2.1 will enable us to state our general result in Theorem 5.3.

Assumption 5.2. With Assumptions 2.1 in force, we further postulate that the logarithm of the numéraire portfolio $\widehat{X}$ is a special semimartingale.

Under Assumption 5.2, we can write $\log (\widehat{X})=\mathcal{O}+M$, where $\mathcal{O}$ is a predictable nondecreasing process and $M$ is a local martingale. We then define market time to be the nondecreasing predictable process $\mathcal{O}$. According to Proposition 5.1, we have, $\mathbb{P}$-a.s., $\mathcal{O}_{0}=0$ and $\mathcal{O}_{\infty}=\infty$. This makes $\mathcal{O}$ a bona fide clock.
5.2. A general result. In what follows, $\alpha$ will denote a nonnegative, possibly infinitevalued random variable such that

$$
\begin{equation*}
\frac{\Delta \widehat{X}}{\widehat{X}_{-}} \leq \alpha \tag{5.1}
\end{equation*}
$$

Of course, $\alpha$ can be chosen in a minimal way as $\alpha:=\sup _{t \in \mathbb{R}_{+}}\left(\Delta \widehat{X}_{t} / \widehat{X}_{t-}\right)$.
Theorem 5.3. Let Assumption 5.2 be in force. With the above definition of the market time $\mathcal{O}$ and a random variable $\alpha$ satisfying (5.1), we have

$$
\begin{equation*}
\log \left(\frac{\ell}{x}\right) \leq v(x ; \ell) \leq \mathbb{E}[\mathcal{T}(\widehat{X}(x) ; \ell)] \leq \log \left(\frac{\ell}{x}\right)+\mathbb{E}[\log (1+\alpha)] . \tag{5.2}
\end{equation*}
$$

It is straightforward that Theorem 5.3 covers both Theorem 3.3 and Theorem 4.2 as special cases. For Theorem 3.3, $\alpha$ is the constant defined in its statement, while for Theorem 4.2 we have $\alpha=0$.

Dividing the inequalities (5.2) with $\log (\ell)$ throughout, we get the following corollary of Theorem 5.3.

Corollary 5.4.In the setting of Theorem 5.3, suppose that $\mathbb{E}[\log (1+\alpha)]<\infty$. Then,

$$
\lim _{\ell \rightarrow \infty} \frac{v(x ; \ell)}{\log (\ell)}=1=\lim _{\ell \rightarrow \infty} \frac{\mathbb{E}[\mathcal{T}(\widehat{X}(x) ; \ell)]}{\log (\ell)}
$$

This last result shows that, under some integrability condition on the possible size of the jumps of the logarithm of the growth-optimal portfolio, the problem of possible overshoots vanishes asymptotically when considering increasing wealth levels $\ell$.
6. Proofs. Before we embark on proving all the results of the previous sections, we define, in accordance to (2.2), for any càdlàg process $X$ and $\ell \in \mathbb{R}_{+}$,

$$
\tau(X ; \ell):=\inf \left\{t \in \mathbb{R}_{+} \mid X_{t} \geq \ell\right\}
$$

to be the first upcrossing calendar time of $X$ at level $\ell$. It is clear that $\tau(X ; \ell)$ is a stopping time and that $\mathcal{O}_{\tau(X ; \ell)}=\mathcal{T}(X ; \ell)$ for all càdlàg processes $X$ and $\ell \in \mathbb{R}_{+}$.
6.1. Proof of Proposition 2.2. Recall that the clock $\mathcal{O}$ satisfies $\mathbb{P}\left[\mathcal{O}_{\infty}=\infty\right]=1$. Therefore, for any $X \in \mathcal{X}$ and $\ell \in \mathbb{R}_{+}, \mathbb{P}[\tau(X ; \ell)<\infty]=1$ is equivalent to $\mathbb{P}[\mathcal{T}(X ; \ell)<\infty]=1$.

Condition (2') of Proposition 2.2 obviously implies Assumption 2.1(2). Conversely, assume that Assumptions 2.1 are in force. For any $n \in \mathbb{N}$, pick $X \in \mathcal{X}(1)$ such that $\mathbb{P}\left[\tau^{n}<\infty\right]=1$, where $\tau^{n}:=\tau(X ; n)$. Since $X / \widehat{X}$ is a nonnegative supermartingale, the optional sampling theorem (see, for example, section 1.3.C of [13]) gives

$$
1 \geq \mathbb{E}\left[\frac{X_{\tau^{n}}}{\widehat{X}_{\tau^{n}}}\right] \geq n \mathbb{E}\left[\frac{1}{\widehat{X}_{\tau^{n}}}\right]
$$

It follows that $\left(1 / \widehat{X}_{\tau^{n}}\right)_{n \in \mathbb{N}}$ converges to zero in probability. As $1 / \widehat{X}$ is a nonnegative supermartingale, this implies that $\lim _{t \rightarrow \infty}\left(1 / \widehat{X}_{t}\right)=0, \mathbb{P}$-a.s., which establishes the result.

### 6.2. Proof of Proposition 5.1. Under the assumption that the numéraire portfolio $\widehat{X}$ is

 a special semimartingale with canonical decomposition $\widehat{X}=\mathcal{O}+M$, the event equality$$
\left\{\lim _{t \rightarrow \infty} \widehat{X}_{t}=+\infty\right\}=\left\{\lim _{t \rightarrow \infty} \mathcal{O}_{t}=+\infty\right\}
$$

which is to be understood in a modulo $\mathbb{P}$ sense, is a consequence of Proposition 3.21 in [12]. Then, the result of Proposition 5.1 readily follows in view of Proposition 2.2.
6.3. Proof of Proposition 3.2. The fact that $\mathfrak{I} \cap \check{\mathfrak{C}}=\emptyset$ is equivalent to the existence of $\rho \in \mathfrak{C}$ such that $\mathfrak{g}(\rho)=\mathfrak{g}^{*}<\infty$, as well as that $\widehat{X}$ as defined in (3.3) is the numéraire portfolio, is a consequence of Lemma 4.1 in [14], as soon as one also uses the bounded-jump Assumption 3.1.

Now, it is straightforward to check that $\mathfrak{g}^{*}=0$ is equivalent to $\widehat{X}$ being a positive local martingale, in which case we have that, $\mathbb{P}$-a.s., $\lim _{t \rightarrow \infty} \widehat{X}_{t}<\infty$. On the other hand, if $\mathfrak{g}^{*}>0$, then the Lévy process $\log (\widehat{X})$ is integrable and has strictly positive drift $\mathfrak{g}^{*}$; therefore, $\mathbb{P}$-a.s., $\lim _{t \rightarrow \infty} \widehat{X}_{t}=\infty$. In view of Proposition 2.2, the result follows.
6.4. Proof of Proposition 4.1. The fact that (1) and (2) of Proposition 4.1 are equivalent to the existence of the numéraire portfolio $\widehat{X}$, as well as that $\widehat{X}$ is given by (4.2), is a special case of Theorem 3.15 in [12]-see also [8]. Under the validity of (1) and (2) of Proposition 4.1, it is straightforward to see that (3) of Proposition 4.1 is equivalent to $\lim _{t \rightarrow \infty} \widehat{X}_{t}=\infty$. Using Proposition 2.2, the result follows.
6.5. Proof of Theorem 5.3. Let $\widehat{L}(x):=\log (\widehat{X}(x))$. Observe that, since $\Delta \widehat{X} \leq \alpha \widehat{X}_{-}$,

$$
\begin{equation*}
\Delta \widehat{L}(x)=\log \left(1+\frac{\Delta \widehat{X}}{\widehat{X}_{-}}\right) \leq \log (1+\alpha) \tag{6.1}
\end{equation*}
$$

Write $\widehat{L}(x)=\log (x)+\mathcal{O}+M$, where $M$ is a local martingale. Let $\left(\tau^{n}\right)_{n \in \mathbb{N}}$ be a localizing sequence for $M$. The estimate (6.1) gives, for all $n \in \mathbb{N}$,

$$
\log (x)+\mathbb{E}\left[\mathcal{O}_{\tau^{n} \wedge \tau(\widehat{X}(x) ; \ell)}\right]=\mathbb{E}\left[\widehat{L}_{\tau^{n} \wedge \tau(\widehat{X}(x) ; \ell)}(x)\right] \leq \log (\ell)+\mathbb{E}[\log (1+\alpha)] .
$$

Now letting $n$ tend to infinity and using the monotone convergence theorem, we get

$$
\begin{equation*}
\mathbb{E}[\mathcal{T}(\widehat{X}(x) ; \ell)] \leq \log (\ell / x)+\mathbb{E}[\log (1+\alpha)] \tag{6.2}
\end{equation*}
$$

Now take any $X \in \mathcal{X}(x)$. If $\mathbb{P}[\mathcal{T}(X, \ell)=\infty]>0$, we have $\mathbb{E}[\mathcal{T}(X, \ell)]=\infty$ and $\log (\ell / x) \leq$ $\mathbb{E}[\mathcal{T}(X, \ell)]$ is trivial. It remains to consider the case $\mathbb{P}[\mathcal{T}(X, \ell)<\infty]=1$, or equivalently $\mathbb{P}[\tau(X, \ell)<\infty]=1$.

For all $\epsilon \in(0,1)$, define $X^{\epsilon}:=(1-\epsilon) X+\epsilon x$. Then, $X^{\epsilon} \in \mathcal{X}(x)$ and $\tau\left(X^{\epsilon}, \epsilon x+(1-\epsilon) \ell\right)=$ $\tau(X, \ell)$. The drift part of the process $L^{\epsilon}:=\log \left(X^{\epsilon}\right)$ is bounded above by $\mathcal{O}$. Therefore,

$$
L^{\epsilon} \leq \log (x)+\mathcal{O}+M^{\epsilon}
$$

for some local martingale $M^{\epsilon}$. Let $\left(\tau^{\epsilon, n}\right)_{n \in \mathbb{N}}$ be a localizing sequence for $M^{\epsilon}$. Since the stopped process $M_{\tau(X, \ell) \wedge \tau^{\epsilon, n} \wedge \text {. }}^{\epsilon}$ is a martingale, we have that

$$
\mathbb{E}\left[L_{\tau(X, \ell) \wedge \tau^{\epsilon}, n}^{\epsilon}\right] \leq \log (x)+\mathbb{E}\left[\mathcal{O}_{\tau(X, \ell) \wedge \tau^{\epsilon, n}}\right]=\log (x)+\mathbb{E}\left[\mathcal{T}(X, \ell) \wedge \mathcal{O}_{\tau^{\epsilon, n}}\right]
$$

Now, $L^{\epsilon}$ is uniformly bounded from below by $\log (\epsilon x)$. Furthermore, $\uparrow \lim _{n \rightarrow \infty} \mathcal{O}_{\tau^{n}}=\infty$ holds in a $\mathbb{P}$-a.s. sense. Therefore, applications of Fatou's lemma and the monotone convergence theorem will give

$$
\begin{aligned}
\log (\ell)+\log (1-\epsilon) \leq \mathbb{E}\left[L_{\tau(X, \ell)}^{\epsilon}\right] & \leq \liminf _{n \rightarrow \infty} \mathbb{E}\left[L_{\tau(X, \ell) \wedge \tau^{n}}^{\epsilon}\right] \\
& \leq \log (x)+\liminf _{n \rightarrow \infty} \mathbb{E}\left[\mathcal{T}(X, \ell) \wedge \mathcal{O}_{\tau^{n}}\right] \\
& =\log (x)+\mathbb{E}[\mathcal{T}(X, \ell)]
\end{aligned}
$$

Now sending $\epsilon$ to zero, we also get $\log (\ell / x) \leq \mathbb{E}[\mathcal{T}(X, \ell)]$ for all $X \in \mathcal{X}(x)$ that satisfy $\mathbb{P}[\mathcal{T}(X, \ell)<\infty]=1$. This, coupled with (6.2), finishes the proof.
6.6. Proof of Proposition 3.6. The existence of $\rho \in \mathfrak{C}$ such that $\mathfrak{g}(\rho)=\mathfrak{g}^{*}<\infty$ follows from Lemma 4.1 in [14] in view of $\mathfrak{I} \cap \mathfrak{C} \neq \emptyset$. Note that the finiteness of $\mathfrak{g}^{*}$ is straightforward from the defining equation (3.2) for $\mathfrak{g}$.

Call $\widehat{L}:=\log (\widehat{X})$. For each $n \in \mathbb{N}$, let

$$
\widehat{L}^{n}:=\widehat{L}-\sum_{t \leq \cdot}\left(\Delta \widehat{L}_{t}\right) \mathbb{I}_{\left\{\Delta \widehat{L}_{t}>n\right\}} .
$$

Then, $\widehat{L}^{n}$ is a Lévy process and we can write

$$
\widehat{L}_{t}^{n}=\mathfrak{g}^{n} t+M_{t}^{n}
$$

for all $t \in \mathbb{R}_{+}$, where $M^{n}$ is a Lévy martingale and $\uparrow \lim _{n \rightarrow \infty} \mathfrak{g}^{n}=\mathfrak{g}^{*}>0$. Then,

$$
\mathbb{E}[\mathcal{T}(\widehat{X}(x) ; \ell)]=\mathfrak{g}^{*} \mathbb{E}[\tau(\widehat{X}(x) ; \ell)] \leq \mathfrak{g}^{*} \mathbb{E}\left[\tau\left(\widehat{L}^{n}(x) ; \log (\ell)\right)\right] \leq \frac{\mathfrak{g}^{*}}{\mathfrak{g}^{n}}\left(\log \left(\frac{\ell}{x}\right)+\log (1+n)\right)
$$

holds for all $n \in \mathbb{N}$ such that $\mathfrak{g}^{n}>0$, where the last inequality follows along the same lines of the proof of (6.2). It then follows that

$$
\limsup _{\ell \rightarrow \infty} \frac{\mathbb{E}[\mathcal{T}(\widehat{X}(x) ; \ell)]}{\log (\ell)} \leq \frac{\mathfrak{g}^{*}}{\mathfrak{g}^{n}}
$$

holds for all $n \in \mathbb{N}$ such that $\mathfrak{g}^{n}>0$. Since $\uparrow \lim _{n \rightarrow \infty} \mathfrak{g}^{n}=\mathfrak{g}^{*}>0$, sending $n$ to infinity in the last inequality we get

$$
\limsup _{\ell \rightarrow \infty} \frac{\mathbb{E}[\mathcal{T}(\widehat{X}(x) ; \ell)]}{\log (\ell)} \leq 1
$$

Of course, in view of the bounds (5.2) of Theorem 5.3, we always have

$$
1=\lim _{\ell \rightarrow \infty} \frac{v(x ; \ell)}{\log (\ell)} \leq \liminf _{\ell \rightarrow \infty} \frac{\mathbb{E}[\mathcal{T}(\widehat{X}(x) ; \ell)]}{\log (\ell)}
$$

which completes the proof.

## REFERENCES

[1] P. H. Algoet and T. M. Cover, Asymptotic optimality and asymptotic equipartition properties of log-optimum investment, Ann. Probab., 16 (1988), pp. 876-898.
[2] D. C. Aucamp, An investment strategy with overshoot rebates which minimizes the time to attain a specified goal, Management Sci., 23 (1976/77), pp. 1234-1241.
[3] D. Becherer, The numeraire portfolio for unbounded semimartingales, Finance Stoch., 5 (2001), pp. 327341.
[4] F. Black, Studies of stock price volatility changes, in Proceedings of the 1976 Meetings of the Business and Economics Statistics Section, American Statistical Association, Berkeley, CA, 1976, pp. 177-181.
[5] L. Breiman, Optimal gambling systems for favorable games, in Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probablility, Vol. I, University of California Press, Berkeley, CA, 1961, pp. 65-78.
[6] S. Browne, Reaching goals by a deadline: Digital options and continuous-time active portfolio management, Adv. in Appl. Probab., 31 (1999), pp. 551-577.
[7] M. M. Christensen and K. Larsen, No arbitrage and the growth optimal portfolio, Stoch. Anal. Appl., 25 (2007), pp. 255-280.
[8] F. Delbaen and W. Schachermayer, The existence of absolutely continuous local martingale measures, Ann. Appl. Probab., 5 (1995), pp. 926-945.
[9] H. Föllmer and P. Leukert, Quantile hedging, Finance Stoch., 3 (1999), pp. 251-273.
[10] D. Heath, S. Orey, V. Pestien, and W. Sudderth, Minimizing or maximizing the expected time to reach zero, SIAM J. Control Optim., 25 (1987), pp. 195-205.
[11] D. Heath and W. Sudderth, Continuous-Time Portfolio Management: Minimizing the Expected Time to Reach a Goal, manuscript, 1984.
[12] I. Karatzas and C. Kardaras, The numéraire portfolio in semimartingale financial models, Finance Stoch., 11 (2007), pp. 447-493.
[13] I. Karatzas and S. E. Shreve, Brownian Motion and Stochastic Calculus, 2nd ed., Grad. Texts in Math. 113, Springer-Verlag, New York, 1991.
[14] C. Kardaras, No-free-lunch equivalences for exponential Lévy models under convex constraints on investment, Math. Finance, 19 (2009), pp. 161-187.
[15] C. Kardaras and E. Platen, On the Semimartingale Property of Discounted Asset-Price Processes, preprint, http://arxiv.org/abs/0803.1890.
[16] J. L. Kelly, Jr., A new interpretation of information rate, Bell System Tech. J., 35 (1956), pp. 917-926.
[17] J. B. J. Long, The numéraire portfolio, J. Financial Economics, 26 (1990), pp. 29-69.
[18] E. Platen, A benchmark approach to finance, Math. Finance, 16 (2006), pp. 131-151.
[19] K.-I. Sato, Lévy Processes and Infinitely Divisible Distributions, Cambridge Stud. Adv. Math. 68, Cambridge University Press, Cambridge, UK, 1999.

# Optimal Convergence Rate of the Binomial Tree Scheme for American Options with Jump Diffusion and Their Free Boundaries* 

Jin Liang ${ }^{\dagger}$, Bei $\mathrm{Hu}^{\ddagger}$, and Lishang Jiang ${ }^{\dagger}$

Abstract. An American put option with jump diffusion can be modeled as an integro-variational inequality. With a penalization approximation and under the stability condition $\frac{\sigma^{2} \Delta t}{\Delta x^{2}} \leqslant 1$, where $\Delta x=\ln \frac{S_{n+1}}{S_{n}}$ ( $S_{t}$-underlying asset price), we obtain the optimal convergence rate $O\left((\Delta x)+(\Delta t)^{1 / 2}\right.$ ) of the binomial tree scheme for this variational inequality. Moreover, we define an approximate optimal exercise boundary within the framework of the binomial tree scheme and derive the convergence rate estimate $O\left((\Delta t)^{1 / 4}\right)$ to the actual free boundary.

Key words. convergence rate, jump diffusion, explicit difference scheme, American options, free boundary, binomial tree method

AMS subject classifications. 91G80, 91G60, 65M06, 35R35
DOI. 10.1137/090746239

1. Introduction. An American put option is a contract which gives a holder a right to sell the underlying asset at any time before the expiration date for a certain price. In the Black-Scholes framework [7], the valuation of an American put option is a free boundary problem.

Assume that the underlying asset price $\left\{S_{t}\right\}$ is a random process, which can be modelled by the following stochastic differential equation in a given probability space $(\Omega, \mathcal{F}, \mathcal{P})$ (see also [22] by Kou):

$$
\begin{equation*}
\frac{d S_{\tau}}{S_{\tau}}=\mu d \tau+\sigma d W_{\tau}+d\left(\sum_{j=1}^{\hat{N}_{\tau}} U_{j}\right) \tag{1.1}
\end{equation*}
$$

where $\mu, \sigma$ are positive constants; $\left\{W_{\tau}\right\}_{\tau>0}$ is the standard Brownian motion with

$$
\mathrm{E}\left(d W_{\tau}\right)=0, \quad \operatorname{Var}\left(d W_{\tau}\right)=d \tau ;
$$

$\left\{\hat{N}_{\tau}\right\}_{\tau>0}$ is a Poisson process with positive constant intensity $\lambda$; and the sequence $\left\{U_{j}\right\}_{j \geqslant 1}$ consists of square integrable, independent, identically distributed random variables taking values in $(-1, \infty)$.

[^2]In this paper, the jumps are assumed to be lognormally distributed; however, our work extends to other types (e.g., described as Kou's models in [22]) of jumps satisfying (2.7)-(2.8). If (1.1) is used to describe the dynamics under the pricing measure of the stock market, then for the corresponding American put option pricing at time $\tau$, given by

$$
V_{\tau}=V\left(S_{\tau}, \tau\right),
$$

the function $V(s, \tau)$ satisfies the following variational inequality (see, e.g., [32], [33], [34], [36], [38]):

$$
\begin{array}{lc}
\min \left(r^{-1} \widetilde{L}[V], V(S, \tau)-\widetilde{g}(S)\right)=0, & (S, \tau) \in \mathbb{R}^{+} \times[0, T], \\
V(S, T)=\widetilde{g}(S), & S \in \mathbb{R}^{1}, \tag{1.3}
\end{array}
$$

where

$$
\begin{align*}
\widetilde{L}[V]= & \frac{\partial V}{\partial \tau}+\frac{1}{2} \sigma^{2} S^{2} \frac{\partial^{2} V}{\partial S^{2}}+(r-\lambda \omega) S \frac{\partial V}{\partial S}  \tag{1.4}\\
& -(r+\lambda) V+\lambda \int_{-1}^{\infty} V(S(1+z), t) d \widetilde{N}(z), \\
\widetilde{g}(S)= & \max (K-S, 0), \tag{1.5}
\end{align*}
$$

with $\widetilde{N}(z)$ being the distribution of $U_{1}$ and $r, K, \sigma>0$ being the interest rate, striking price, and volatility of the option, respectively. Also $\mu=r-\lambda \omega, \omega=\mathrm{E} U_{1}$. For simplicity, throughout the paper, we suppose that $K=1$. We remark here that the factor $r^{-1}$ in (1.2) is not essential; in fact, the problem is equivalent to the one without this factor. However, we use this factor in the problem for the normalization purpose for the later proof.

An intuitive explanation about (1.1) is given as follows: the stock price $S_{\tau}$ may have finitely many discontinuities on each interval $[0, \tau]$; i.e., it may jump at some random times controlled by the Poisson process $\left\{\hat{N}_{\tau}\right\}_{\tau>0}$ with $\lambda$ and $U_{j}$ describing the frequency and the relative amplitude of jump, respectively. Between two jumps, it follows a continuous lognormal random walk modeled by Brownian motion $\left\{W_{\tau}\right\}_{\tau>0}$.

American option pricing in Merton's jump-diffusion model was developed by Zhang [38], [39] as a variational inequality problem with a free boundary. Using the results of free boundary and probability theories, Pham [33] studied the behavior of the optimal stopping boundary and proved its continuity with some restriction on the size of jump risks. By using the penalization technique to approximate the free boundary problem of American options with jump diffusion, Yang, Jiang, and Bian [37] obtained many properties of the free boundary.

It is well known that no closed-form solutions are available even for the American vanilla option. In 1979, Cox and Rubinstein (see [11]) were the first to propose the binomial tree method (BTM), which is a discrete time model to value vanilla options. Since then, it has been a very popular choice for computing the pricing of many options including American options. The method is widely used.

Now, let us consider an American put option with a jump diffusion. Suppose that the lifetime of the option $[0, T]$ is divided into equally spaced nodes with a distance $h$ apart, $N$ is the number of discrete time points, and $h=T / N$. Let $\mathbb{Z}=\{l: l=0, \pm 1, \pm 2, \ldots\}$. Amin
[4] developed a simple, discrete model to value options when the underlying asset follows a jump-diffusion process. It is assumed that $S$ is the stochastic stock price at time $t$ and $S_{\tau}$ takes values in a discrete set $\left\{S_{l}=S u^{l}: l \in \mathbb{Z}\right\}$ after a small time interval $h$. Here the Brownian motion $\left\{W_{\tau}\right\}_{\tau>0}$ in the continuous time case is responsible for "local" change of the stock price: $S u$ and $S u^{-1}$. When a Poisson jump occurs, the stock price "jumps" to potentially any state on the space grid at the next time.

Let $\hat{\lambda}$ be the probability that a Poisson jump happens in small time interval $h$; i.e., approximately, $\hat{\lambda}=h \lambda$. Then the American put option with jump diffusion price $V_{i}^{n}=$ $V\left(S_{i}^{n}, \tau_{n}\right), i \in \mathbb{Z}$, can be defined by the following inverse induction process [4]:

$$
\left\{\begin{array}{l}
V_{i}^{N}=g\left(S_{i}^{N}\right) ; \\
\text { if } V_{i}^{n+1} \text { is already known, then } V_{i}^{n} \text { is defined as } \\
V_{i}^{n}=\max \left(\frac{1}{\rho}\left((1-\hat{\lambda})\left(q V_{i+1}^{n+1}+(1-q) V_{i-1}^{n+1}\right)+\hat{\lambda} \sum_{l \in \mathbb{Z}} V_{l+1}^{n+1} \hat{p}_{l}\right), g\left(S_{i}^{n}\right)\right),
\end{array}\right.
$$

where $0 \leqslant n \leqslant N-1, \rho=e^{r h}, u=e^{\sigma \sqrt{h}}$, and

$$
\begin{align*}
q & =\frac{(\rho-\hat{\lambda}(\omega+1)) /(1-\hat{\lambda})-u^{-1}}{u-u^{-1}},  \tag{1.6}\\
\hat{p}_{l} & =\operatorname{Prob}\left(\ln \left(1+U_{1}\right) \in\left[\left(l-\frac{1}{2}\right) \sigma \sqrt{h},\left(l+\frac{1}{2}\right) \sigma \sqrt{h}\right)\right) \\
& =\widetilde{N}\left(\left(l+\frac{1}{2}\right) \sigma \sqrt{h}\right)-\widetilde{N}\left(\left(l-\frac{1}{2}\right) \sigma \sqrt{h}\right), \tag{1.7}
\end{align*}
$$

where, as mentioned before, $\widetilde{N}(x)$ is the distribution of jump amplitude $U_{1}$, and $\widetilde{N}(x)=$ $N\left(e^{x}-1\right)$ takes the form of norm distribution (for some $\mu, \hat{\sigma}>0$ ), i.e.,

$$
\begin{equation*}
N(x)=\frac{1}{\hat{\sigma} \sqrt{2 \pi}} \int_{-\infty}^{x} e^{-\frac{(\eta-\mu)^{2}}{2 \hat{\sigma}^{2}}} d \eta . \tag{1.8}
\end{equation*}
$$

Xu, Qian, and Jiang [36] proved that the explicit finite difference scheme (EFDS) of a European option with jump diffusion is stable and convergent (with rate $O(\Delta x+\sqrt{\Delta t})$ ) under the stability condition

$$
\begin{equation*}
\sigma^{2} \frac{\Delta t}{\Delta x^{2}} \leqslant 1 \tag{1.9}
\end{equation*}
$$

They also showed that the BTM described as above is equivalent to an EFDS for a European option with jump diffusion under the condition $(\Delta x=\ln u, \Delta t=h)$ when the equality holds in (1.9).

In a recent paper, Qian et al. [34] extended the proof of equivalence of an EFDS and an explicit difference scheme to an American option with jump diffusion when the equality holds in (1.9). They also proved the uniform convergence of the EFDS under (1.9) for American options in the sense of viscosity solution.

An option pricing model with jump diffusion is a PDE with a nonlocal integral term. There has been much interest in the model with jump diffusions [2], [3], [4], [5], [6], [8], [10],
[14], [18], [22], [30], [33], [34], [36], [37], [39]. An explicit scheme like the BTM is easy to use. However, as an implicit scheme has an advantage on the stability, many studies have been carried out. In [39], Zhang first applied an implicit-explicit mixed scheme on a jump-diffusion model; the scheme was explicit on the integral term and implicit on the partial derivative terms in the equation of the problem. An error estimate was obtained by using an energy estimate. In [14], [8], and [2], implicit/mixed schemes were used to calculate a jump-diffusion model, where convergence was discussed. However, the convergence rate of these schemes remains an open interesting question.

In this paper, the penalization technique is used to study the convergence rate of the EFDS for the American option with jump diffusion model, which includes the BTM as a special case of the EFDS. We derive the optimal convergence rate $O\left((\Delta x)+(\Delta t)^{1 / 2}\right)$. This work extends the results in [29] and [30], where the convergence rate $O\left((\Delta x)^{2 / 3}+(\Delta t)^{1 / 3}\right)$ was obtained for the American vanilla options and the American option with jump diffusion, respectively. We also establish the rate of convergence of the free boundary under certain additional assumptions. The convergence study for the American vanilla options was also done in [20], [21] in a PDE framework without using probability theory, where uniform convergence was obtained.

Lamberton [26] in 1998 obtained the BTM convergence rate between $O\left((\Delta t)^{2 / 3}\right)$ and $O\left((\Delta t)^{3 / 4}\right)$ for an American put option at a fixed time; the estimate is improved to $O((\Delta t)$ $\left.(\sqrt{|\log (\Delta t)|})^{4 / 5}\right)$ under additional conditions. At the fixed time level this estimate is better than ours. Our work is different in the sense that we look for the estimate on the entire time interval including the initial time where the data is not smooth. To our knowledge, the idea of using an approximating smooth solution as a bridge first came from Krylov [23], [24], who worked on Hamilton-Jacobi-Bellman (HJB) equations arising from stochastic control problems. For the FDS on these equations, in [23], [24] the convergence rates of $O\left((\Delta x)^{1 / 3}\right)$ for constant coefficients equations and of $O\left((\Delta x)^{1 / 27}\right)$ for variable coefficients were established, using a combination of analytic and probabilistic methods. Jakobsen studied general Bellman equations [17]. Later, Jakobsen, Karlsen, and La Chioma worked on an HJB equation associated with jump diffusion [18], which was similar to our equation; they got the error estimates to $O\left((\Delta x)^{1 / 2}\right)$ for certain situations, using purely analytic methods. In [15], we improved PDE estimates and obtained the optimal convergence rate of the BTM for the American vanilla option. The current work not only extends the error estimates to the current model with jump diffusion but also gives the error estimates of the free boundary.

When we work on maximum-norm error estimates of the EFDS for the variational inequality problem (1.2)-(1.3), there are three intrinsic difficulties to this discrete scheme: First, the optimal regularity of the solution $V(S, \tau)$ of the variational inequality (1.2) is only of $W_{\infty}^{2,1}$; in fact, $V_{S S}$ is bounded but not continuous at the free boundary; i.e., the truncation error near the free boundary is more difficult to estimate. Second, the initial datum $\widetilde{g}$ is only in $W_{\infty}^{1}$; the derivative $\widetilde{g}^{\prime}$ is bounded but not continuous. Third, the problem has a nonlocal jump-diffusion term, which can be reduced to an integro-partial differential problem. The method to overcome the first difficulty is partially motivated by the work [24], which deals with Bellman's equations. We introduce a regularized solution $V^{\varepsilon}$ to a penalized problem with a penalty $\beta_{\varepsilon}(\cdot)$. The error $\left|V-V_{h}\right|$ between the original solution and the finite difference solution can be estimated by the sum of two parts: $\left|V-V^{\varepsilon}\right|$ and $\left|V_{h}-V^{\varepsilon}\right|$; then we find a
balance between $\varepsilon$ and $h$ to obtain the optimal convergence rate. For the second difficulty, we eliminate the singularity of the initial datum $\widetilde{g}$ by subtracting the corresponding European option price $U(S, \tau)$. Combining these two techniques, we obtain the error estimate $O(\Delta x)$ for $\left|V-V_{h}\right|$. The general method dealing with the integro-differential equation is used to solve the third difficulty. We believe that the method is also suitable for deriving error estimates of the BTM for other American style options.

It is well known that an American option has an optimal exercise boundary, which is called a free boundary in PDE theory. Explicitly locating this free boundary is always significant in applications. This boundary can be estimated within the BTM scheme; however, the convergence rate remains open. In this paper, we give a definition of an approximate boundary within the BTM scheme and prove that this approximate boundary is convergent to the actual free boundary by the rate $O\left(h^{1 / 4}\right)$ under certain conditions.

This paper is organized as follows: In section 2, we state the variational problem and its properties. In section 3, the FDS of the problem is presented. A comparison theorem for the FDS is shown in this section. In section 4, we consider a penalized problem and obtain a regularized solution. The estimates to this regularized solution are derived in section 5. In section 6 , the estimates of the truncation error of the regularized solution are obtained. These estimates are used to derive the main theorem in section 7. The estimates for errors between the finite difference free boundary and the actual free boundary are derived in section 8. Some remarks and some numerical results are collected in section 9.
2. Variational inequality problem. As stated in the introduction, the pricing of an American option can be reduced to a variational problem (1.2)-(1.3). By changing variables

$$
\begin{align*}
& x=\log S, \quad t=T-\tau, \quad y=\log (1+z),  \tag{2.1}\\
& u(x, t) \equiv V(S, \tau), \quad g(x) \equiv \widetilde{g}(S), \quad N(y)=\widetilde{N}(z), \tag{2.2}
\end{align*}
$$

the problem is reduced to

$$
\begin{align*}
& F[u]:=\min \left(r^{-1} L[u], u-g(x)\right)=0, \quad(x, t) \in \mathbb{R}^{1} \times(0, T),  \tag{2.3}\\
& u(x, 0)=g(x), \quad x \in \mathbb{R}^{1}, \tag{2.4}
\end{align*}
$$

where

$$
\begin{align*}
& L[u]=u_{t}-\frac{\sigma^{2}}{2} u_{x x}-\left(r-\lambda \omega-\frac{\sigma^{2}}{2}\right) u_{x}+(r+\lambda) u \\
& \quad-\lambda \int_{\mathbb{R}^{1}} u(x+y, t) d N(y),  \tag{2.5}\\
& g(x)=\left(1-e^{x}\right)^{+}, \quad x \in \mathbb{R}^{1}, \tag{2.6}
\end{align*}
$$

and

$$
\begin{equation*}
\int_{\mathbb{R}^{1}} d N(y)=1, \quad \int_{\mathbb{R}^{1}}\left(e^{y}-1\right) d N(y)=\omega<\infty . \tag{2.7}
\end{equation*}
$$

Note that if $N$ is given by (1.8), then it also satisfies

$$
\begin{equation*}
\int_{\mathbb{R}^{1}} e^{|y|} d N(y)<\infty . \tag{2.8}
\end{equation*}
$$

The discussion of this paper actually applies to general $N$ satisfying (2.7)-(2.8).
For this variational problem, it is well known that there exists a unique solution in $C\left(\mathbb{R}^{1} \times\right.$ $[0, T]) \cap W_{p, l o c}^{2,1}\left(\mathbb{R}^{1} \times(0, T]\right)$ (e.g., see [12, section 1.6]), where $W_{p, l o c}^{2,1}\left(\mathbb{R}^{1} \times(0, T]\right)$ consists of functions $u(x, t)$ which belong to $W_{p}^{2,1}\left(\left(x_{0}-1, x_{0}+1\right) \times[\tau, T]\right)$ for any $x_{0} \in \mathbb{R}^{1}$ and any $0<\tau<T$.

Throughout this paper, we discuss the problem in spaces $W_{p}^{2,1}(Q)$ and $C^{m+\alpha,(m+\alpha) / 2}(Q)$ for $Q=\Omega \times[0, T], \Omega \subset \mathbb{R}^{1}, p \geqslant 1,0 \leqslant \alpha, \beta<1$, and $m$ is any nonnegative integer. The norms of these spaces are defined as follows:

$$
\begin{aligned}
& \|f\|_{L_{p}(Q)}=\left(\int_{Q}|f|^{p} d x d t\right)^{1 / p}, \\
& \|f\|_{W_{p}^{2,1}(Q)}=\|f\|_{L_{p}(Q)}+\left\|D_{x}^{2} f\right\|_{L_{p}(Q)}+\left\|D_{t}^{1} f\right\|_{L_{p}(Q)}, \\
& {[f]_{C^{\alpha, \beta}(Q)}=\sup _{(x, t),(y, \tau) \in Q,(x, t) \neq(y, \tau)} \frac{|f(x, t)-f(y, \tau)|}{|x-y|^{\alpha}+|t-\tau|^{\beta}},} \\
& \|f\|_{C^{\alpha, \beta}(Q)}=\sup _{(x, t) \in Q}|f|+[f]_{C^{\alpha, \beta}(Q)}, \\
& \|f\|_{C^{2 m+\alpha, m+\alpha / 2}(Q)}=\sup _{(x, t) \in Q}|f|+\left[D_{x}^{2 m} f\right]_{C^{\alpha, \alpha / 2}(Q)}+\left[D_{t}^{m} f\right]_{C^{\alpha, \alpha / 2}(Q)}, \\
& {[f]_{C_{t}^{\beta}(Q)}=\sup _{(x, t),(x, \tau) \in Q, t \neq \tau} \frac{|f(x, t)-f(x, \tau)|}{|t-\tau|^{\beta}},} \\
& \|f\|_{C_{t}^{\beta}(Q)}=\|f\|_{L^{\infty}(Q)}+[f]_{C_{t}^{\beta}(Q)} \\
& {[f]_{C_{x}^{\alpha}(Q)}=\sup _{(x, t),(y, t) \in Q, x \neq y} \frac{|f(x, t)-f(y, t)|}{|x-y|^{\alpha}},} \\
& \|f\|_{C_{x}^{\alpha}(Q)}=\|f\|_{L^{\infty}(Q)}+[f]_{C_{x}^{\alpha}(Q)},
\end{aligned}
$$

where $D_{x}^{l} f=\frac{\partial^{l} f}{\partial x^{l}}, D_{t}^{l} f=\frac{\partial^{l} f}{\partial t^{l}}$ for any positive integer $l$.
Furthermore, the following comparison principle holds.
Lemma 2.1 (comparison principle 1). If $u_{j} \in C\left(\mathbb{R}^{1} \times[0, T]\right) \cap W_{p, l o c}^{2,1}\left(\mathbb{R}^{1} \times(0, T]\right)(j=1,2, p \geqslant$ 2) are functions such that $u_{1}-u_{2}$ is bounded from below by a constant, $\int_{\mathbb{R}^{1}} u_{j}(x+y, t) d N(y)$ $(j=1,2)$ are finite for all $(x, t)$, and

$$
\begin{aligned}
& L\left[u_{1}\right]+c(x, t) u_{1} \geqslant L\left[u_{2}\right]+c(x, t) u_{2} \quad \text { for } \quad(x, t) \in \mathbb{R}^{1} \times(0, T), \\
& u_{1}(x, 0) \geqslant u_{2}(x, 0) \quad \text { for } x \in \mathbb{R}^{1},
\end{aligned}
$$

where $c(x, t) \geqslant 0$, then

$$
\begin{equation*}
u_{1}(x, t) \geqslant u_{2}(x, t) \quad \text { for }(x, t) \in \mathbb{R}^{1} \times[0, T] . \tag{2.9}
\end{equation*}
$$

Proof. Using (2.7) we can take $c_{0}$ to be sufficiently large so that

$$
\begin{align*}
L\left[c_{0}+x^{2}\right]= & -\sigma^{2}-2\left(r-\lambda \omega-\frac{\sigma^{2}}{2}\right) x+r\left(c_{0}+x^{2}\right) \\
& -\lambda \int_{\mathbb{R}^{1}}\left(2 x y+y^{2}\right) d N(y) \\
\geqslant & 1 \quad \text { for } x \in \mathbb{R}^{1} . \tag{2.10}
\end{align*}
$$

Thus for any small $\delta>0$, we have, for $u_{1}^{\delta}=u_{1}+\delta\left(c_{0}+x^{2}\right)$,

$$
\begin{equation*}
L\left[u_{1}^{\delta}-u_{2}\right]+c(x, t)\left(u_{1}^{\delta}-u_{2}\right) \geqslant \delta \quad \text { for } \quad(x, t) \in \mathbb{R}^{1} \times(0, T) . \tag{2.11}
\end{equation*}
$$

We want to show that, for any $\delta>0, u_{1}^{\delta} \geqslant u_{2}$ for $0 \leqslant t \leqslant T$. Then the lemma follows by letting $\delta \rightarrow 0$.

The introduction of $c_{0}+x^{2}$ reduces the problem to a finite domain which also excludes the initial time $t=0$. In fact, using the boundedness of $u_{j}$, the assumption on the initial value, and the continuity of $u_{j}$, we conclude that there exist $t_{\delta}>0$ and $R_{\delta}>0$ such that

$$
\begin{gathered}
u_{1}^{\delta}(x, t)>u_{2}(x, t) \quad \text { on }\left\{(x, t) ;|x| \geqslant R_{\delta}, 0 \leqslant t \leqslant T\right\} \cup\left\{\left(x, t_{\delta}\right) ;|x| \leqslant R_{\delta}\right\}, \\
R_{\delta} \rightarrow \infty, \quad t_{\delta} \rightarrow 0 \quad \text { as } \delta \rightarrow 0 .
\end{gathered}
$$

If $\lambda=0$, we can immediately conclude that $u_{1}^{\delta} \geqslant u_{2}$ for $|x| \leqslant R_{\delta}, t_{\delta} \leqslant t \leqslant T$ by the classical weak maximum principle for parabolic equations.

If $\lambda>0$, we write $L[u]=L_{0}[u]-\lambda \int_{\mathbb{R}^{1}} u(x+y) d N(y)$. For each $\delta>0$, we take $u_{1}^{\delta}$ and $t_{\delta}, R_{\delta}$ as above. Let $\left[t_{\delta}, T_{1}\right]\left(T_{1} \leqslant T\right)$ be the maximal interval such that

$$
\begin{equation*}
\int_{\mathbb{R}^{1}}\left\{u_{1}^{\delta}(x+y, t)-u_{2}(x+y, t)\right\} d N(y)>-\frac{\delta}{\lambda} \quad \text { for } \quad x \in \mathbb{R}^{1}, t_{\delta} \leqslant t<T_{1} \tag{2.12}
\end{equation*}
$$

It follows from (2.11) that $L_{0}\left[u_{1}^{\delta}-u_{2}\right]+c(x, t)\left(u_{1}^{\delta}-u_{2}\right)>0$ for $t_{\delta}<t<T_{1}$. Applying the weak maximum principle to the operator $L_{0}$ we obtain $u_{1}^{\delta} \geqslant u_{2}$ for $t_{\delta} \leqslant t \leqslant T_{1}$. In particular, this implies, for $t=T_{1}$, the integral in (2.12) satisfies $\int_{\mathbb{R}^{1}}\left\{u_{1}^{\delta}\left(x+y, T_{1}\right)-u_{2}\left(x+y, T_{1}\right)\right\} d N(y) \geqslant 0$. If $T_{1}<T$, then by the continuity and boundedness of the functions $u_{j}$, the inequality (2.12) can be extended beyond $t=T_{1}$, which is a contradiction to the maximality of the interval $\left[t_{\delta}, T_{1}\right]$. We conclude that $T_{1}=T$, and this concludes the proof.

We introduce the regularized solution $u^{\varepsilon}$ to the penalized problem

$$
\begin{align*}
& L\left[u^{\varepsilon}\right]=\beta_{\varepsilon}\left(g-u^{\varepsilon}\right), \quad(x, t) \in \mathbb{R}^{1} \times(0, T),  \tag{2.13}\\
& u^{\varepsilon}(x, 0)=g(x)=\left(1-e^{x}\right)^{+}, \quad x \in \mathbb{R}^{1}, \tag{2.14}
\end{align*}
$$

where $\beta_{\varepsilon}(z)=\beta(z / \varepsilon)$, and $\beta$ satisfies

$$
\begin{gathered}
\beta(z)=0 \quad \text { for } z<0, \quad \beta(z)=z-1 \quad \text { for } z>2, \\
\beta \in C^{\infty}(\mathbb{R}), \quad \beta^{\prime}(z) \geqslant 0, \quad \beta^{\prime \prime}(z) \geqslant 0, \\
\beta^{\prime}(z)>0, \quad \beta^{\prime \prime}(z)>0 \quad \text { for } 0<z<2 .
\end{gathered}
$$

By using Lemma 2.1, the existence of a classical solution $u^{\varepsilon} \in C^{\infty}$ to (2.13)-(2.14) for all $\varepsilon>0$, and its convergence to the solution of (2.3), (2.4) can be derived in the same way as in [12, Chapter 1].

Lemma 2.2 (comparison principle 2). If $u_{j} \in C\left(\mathbb{R}^{1} \times[0, T]\right) \cap W_{p, \text { loc }}^{2,1}\left(\mathbb{R}^{1} \times(0, T]\right)(j=1,2, p \geqslant$ 2) are functions such that $u_{1}-u_{2}$ is bounded from below by a constant, $\int_{\mathbb{R}^{1}} u_{j}(x+y, t) d N(y)$ $(j=1,2)$ are finite for all $(x, t)$, and

$$
\begin{aligned}
& L\left[u_{1}\right]-\beta_{\varepsilon}\left(g-u_{1}\right) \geqslant L\left[u_{2}\right]-\beta_{\varepsilon}\left(g-u_{2}\right) \quad \text { for }(x, t) \in \mathbb{R}^{1} \times(0, T), \\
& u_{1}(x, 0) \geqslant u_{2}(x, 0) \quad \text { for } x \in \mathbb{R}^{1},
\end{aligned}
$$

then

$$
\begin{equation*}
u_{1}(x, t) \geqslant u_{2}(x, t) \quad \text { for }(x, t) \in \mathbb{R}^{1} \times[0, T] \tag{2.15}
\end{equation*}
$$

Proof. If we set $w=u_{1}-u_{2}$, then $w$ satisfies

$$
\begin{aligned}
& L[w]+\beta_{\varepsilon}^{\prime}(\xi(x, t)) w \geqslant 0 \quad \text { for } \quad(x, t) \in \mathbb{R}^{1} \times(0, T) \\
& w(x, 0) \geqslant 0 \text { for } x \in \mathbb{R}^{1}
\end{aligned}
$$

where $\xi(x, t)$ is a number between $g(x)-u_{2}(x, t)$ and $g(x)-u_{2}(x, t)$. By Lemma 2.1 and the definition of $\beta_{\varepsilon}$, the lemma follows.

Lemma 2.3 (comparison principle 3). If $u_{j} \in C\left(\mathbb{R}^{1} \times[0, T]\right) \cap W_{p, l o c}^{2,1}\left(\mathbb{R}^{1} \times(0, T]\right)(j=1,2, p \geqslant$ 2) are functions such that $u_{1}-u_{2}$ is bounded from below by a constant, $\int_{\mathbb{R}^{1}} u_{j}(x+y, t) d N(y)$ $(j=1,2)$ are finite, and

$$
\begin{aligned}
& F\left[u_{1}\right] \geqslant F\left[u_{2}\right] \quad \text { for }(x, t) \in \mathbb{R}^{1} \times(0, T) \\
& u_{1}(x, 0) \geqslant u_{2}(x, 0) \quad \text { for } x \in \mathbb{R}^{1}
\end{aligned}
$$

then

$$
\begin{equation*}
u_{1}(x, t) \geqslant u_{2}(x, t) \quad \text { for }(x, t) \in \mathbb{R}^{1} \times[0, T] \tag{2.16}
\end{equation*}
$$

Proof. Notice that on the set $\Omega_{1}^{T}:=\left\{(x, t) ; u_{2}(x, t)-g(x) \leqslant r^{-1} L\left[u_{2}\right]\right\}$ we automatically have $u_{1}(x, t)-g(x) \geqslant F\left[u_{1}(x, t)\right] \geqslant F\left[u_{2}(x, t)\right]=u_{2}(x, t)-g(x)$, so that $u_{1}(x, t) \geqslant u_{2}(x, t)$ on $\Omega_{1}^{T}$.

On the set $\Omega_{2}^{T}:=\left\{(x, t) ; u_{2}(x, t)-g(x)>r^{-1} L\left[u_{2}\right]\right\}$, we have $r^{-1} L\left[u_{1}\right] \geqslant F\left[u_{1}\right] \geqslant F\left[u_{2}\right]=$ $r^{-1} L\left[u_{2}\right]$.

We are now in a situation where

$$
\begin{aligned}
& L\left[u_{1}\right] \geqslant L\left[u_{2}\right] \quad \text { for } \quad(x, t) \in \Omega_{2}^{T} \\
& u_{1}(x, t) \geqslant u_{2}(x, t) \quad \text { for } \quad(x, t) \in \mathbb{R}^{1} \times(0, T) \backslash \Omega_{2}^{T} .
\end{aligned}
$$

If $\lambda=0$, then we can apply the maximum principle on $\Omega_{2}^{T}$ to conclude $u_{1}(x, t) \geqslant u_{2}(x, t)$ on $\Omega_{2}^{T}$. On the other hand, if $\lambda>0$, we can repeat the proof for Lemma 2.1 by defining $u_{1}^{\delta}$ and letting $\left[t_{\delta}, T_{1}\right]\left(T_{1} \leqslant T\right)$ be the maximal interval such that (2.12) holds. Following the maximum principle, we conclude that $u_{1}^{\delta} \geqslant u_{2}$ for $(x, t) \in \Omega_{2}^{T} \cap\left(\mathbb{R}^{1} \times\left[0, T_{1}\right]\right)$. However, since we already proved $u_{1}^{\delta} \geqslant u_{2}$ for $(x, t) \in \Omega_{1}^{T} \cap\left(\mathbb{R}^{1} \times\left[0, T_{1}\right]\right)$, we actually have $u_{1}^{\delta} \geqslant u_{2}$ for $(x, t) \in \mathbb{R}^{1} \times\left[0, T_{1}\right]$. In particular, this implies, for $t=T_{1}$, the integral in (2.12) satisfies $\int_{\mathbb{R}^{1}}\left\{u_{1}^{\delta}\left(x+y, T_{1}\right)-u_{2}\left(x+y, T_{1}\right)\right\} d N(y) \geqslant 0$. If $T_{1}<T$, then the inequality (2.12) can be extended beyond $t=T_{1}$ as before, which is a contradiction to the maximality of the interval $\left[t_{\delta}, T_{1}\right]$. We conclude that $T_{1}=T$, and this concludes the proof.

An American option price without jump diffusion and its free boundary is well studied. Recently, Yang, Jiang, and Bian [37] developed results into the one with jump diffusion. We write these results as follows.

Lemma 2.4. The American put option price $u$ with jump diffusion is the solution of the variational inequality (2.1), (2.3), which can be decomposed into two parts: the first is a European option price $U$ and the second is a so-called early exercise premium $(u-U)$. The optimal
exercise boundary is a free boundary $x=s(t)$ which is continuous and strictly decreasing. The free boundary $x=s(t)$ divides the domain into two regions: the continuation region $x>s(t)$ and stopping region $x<s(t)$. In the stopping region $u \equiv g$, and in the continuous region $u>g$ and $L[u]=0$. Moreover, the solution $u$ belongs to $W_{\infty}^{2,1}\left(\mathbb{R}^{1} \times[\delta, T]\right)$ for any $\delta>0$. For $t>0, u, u_{x}$ are continuous across the free boundary.

Remark 2.1. Further regularity has been discussed in [37]. Note that our system (2.3)-(2.7) corresponds to the system (1.1) in [37] with $q=0$ there and $N(y)$ here as $N\left(e^{y}-1\right)$ there. By setting $q=0$ in their system and using a transform $y \rightarrow e^{y}-1$, [37, Theorem 5.6] implies that if

$$
\begin{equation*}
\nu:=r-\lambda \int_{0}^{\infty}\left(e^{y}-1\right) d N(y) \geqslant 0 \tag{2.17}
\end{equation*}
$$

for our system (2.3)-(2.7), then $s \in C^{1}(0, T]$. The condition (2.17) was recently removed in Bayraktar and Hao [5, Theorem 4.1]. Higher order regularity was also obtained in [5, Theorem 5.1] under the assumption that $N$ has a density which is smooth.

Since our initial datum contains a cusp, it is convenient to use a European option price $U$ with jump diffusion as a comparison function, which is a solution to the problem

$$
\begin{align*}
& L[U]=0 \quad \text { for } 0<t<T, x \in \mathbb{R}^{1}  \tag{2.18}\\
& U(x, 0)=g(x) \text { for } x \in \mathbb{R}^{1} \tag{2.19}
\end{align*}
$$

The properties of $U$ are shown in the following lemma.
Lemma 2.5. Under the transformation of (2.1)-(2.2), the European option price with jump diffusion $U$ is the solution of the problem (2.18), (2.19), which satisfies

$$
\begin{align*}
& \|U\|_{L^{\infty}\left(\mathbb{R}^{1} \times[0, T]\right)}+[U]_{C_{t}^{1 / 2}\left(\mathbb{R}^{1} \times[0, T]\right)}+\left\|U_{x}\right\|_{L^{\infty}\left(\mathbb{R}^{1} \times[0, T]\right)} \leqslant C,  \tag{2.20}\\
& \left\|\left[-U_{t}\right]^{+}\right\|_{L^{\infty}\left(\mathbb{R}^{1} \times[0, T]\right)}+\left\|\left[-U_{x x}\right]^{+}\right\|_{L^{\infty}\left(\mathbb{R}^{1} \times[0, T]\right)} \leqslant C,  \tag{2.21}\\
& U_{x x}(x, t) \leqslant \frac{C}{\sqrt{t}} \exp \left(-\frac{x^{2}}{C t}\right), \quad U_{t}(x, t) \leqslant \frac{C}{\sqrt{t}} \exp \left(-\frac{x^{2}}{C t}\right), \tag{2.22}
\end{align*}
$$

and for $m \geqslant 1,0 \leqslant \alpha<1$,

$$
\begin{equation*}
\|U\|_{C^{m+\alpha,(m+\alpha) / 2}\left(\mathbb{R}^{1} \times[\tau, T]\right)} \leqslant C \tau^{-(m-1+\alpha) / 2} \tag{2.23}
\end{equation*}
$$

where we use the notation $\left\|[f]^{+}\right\|_{L_{\infty}}=\|\max \{f, 0\}\|_{L_{\infty}}$.
Proof. Define $\hat{L}[u]=u_{t}-\frac{\sigma^{2}}{2} u_{x x}-\left(\hat{r}-\frac{\sigma^{2}}{2}\right) u_{x}+\hat{r} u$, where $\hat{r}=r-\lambda \omega$. In order to prove (2.20)-(2.23), we decompose $U(x, t)$ into the following:

$$
U(x, t)=U_{1}(x, t)+U_{2}(x, t),
$$

where $U_{1}(x, t)$ is the solution to the standard Black-Scholes equation associated with $\hat{L}$ (i.e., $\hat{L}\left[U_{1}\right]=0$ ) with initial condition $g(x)$, given by an explicit formula, which can be found in many text books, e.g., [16, Chapter 13] and [19, section 5.3]. It can be verified directly from the explicit formula that $U_{1}$ satisfies (2.20)-(2.23).

Clearly, $U_{2}(x, t)$ satisfies that

$$
\begin{align*}
& \hat{L} U_{2}=\lambda\left\{(\omega-1)\left[U_{1}+U_{2}\right]+\int_{\mathbb{R}^{1}}\left[U_{1}+U_{2}\right](x+y, t) d N(y)\right\},  \tag{2.24}\\
& U_{2}(x, 0)=0 . \tag{2.25}
\end{align*}
$$

Thus, by Schauder estimates (cf. [31, Theorem 4.9, p. 59]; see also [13], [25]) modified to include the jump-diffusion term (this is a lower order term), $\left(U_{2}\right)_{t}$ and $\left(U_{2}\right)_{x x}$ are both bounded (and Hölder continuous). It follows that $U(x, t)$ has the same regularities as $U_{1}(x, t)$; i.e., (2.20)(2.22) hold true for $U(x, t)$ as well. We can then proceed to derive (2.23) by differentiating in $x$ and $t$, respectively, and applying the parabolic estimates.

The proof can also be found in [36, Theorem 3.1, Proof in Appendix] for a European call option with jump diffusion.

Remark 2.2. Actually, the problem (2.18), (2.19) has an explicit solution which has an infinite series form. See [32, Chapter 9].

Remark 2.3. In this section we use $W_{p}^{2,1}$ solutions. These solutions are called strong solutions since they satisfy only the equation excluding a zero measure set. However, the solution to the penalized problem is smooth and therefore is classical. The differentiations, although formal, are still rigorous under appropriate function spaces.
3. Explicit difference scheme. The corresponding explicit finite difference scheme (EFDS) for variational problem (1.2)-(1.3) is defined by

$$
\begin{align*}
& F_{h}\left[u_{h}\right]:=\min \left(r^{-1} L_{h}\left[u_{h}\right], u_{h}-g(x)\right)=0 \text { for }(x, t) \in \mathbb{R}^{1} \times[0, T],  \tag{3.1}\\
& u_{h}(x, t)=g(x) \text { for }-h \leqslant t \leqslant 0, x \in \mathbb{R}^{1}, \tag{3.2}
\end{align*}
$$

where

$$
\begin{aligned}
& L_{h}[u]= \frac{u(x, t)-u(x, t-h)}{h} \\
&-\frac{\sigma^{2}}{2} \frac{u(x+\sigma k, t-h)+u(x-\sigma k, t-h)-2 u(x, t-h)}{(\sigma k)^{2}} \\
&+\left(\frac{\sigma^{2}}{2}+\lambda \omega-r\right) \frac{u(x+\sigma k, t-h)-u(x-\sigma k, t-h)}{2 \sigma k} \\
&+(r+\lambda) u(x, t)-\lambda \sum_{l=-\infty}^{\infty} u(x+l \sigma k, t-h) p_{l} \\
&=\frac{u(x, t)}{h}+(r+\lambda) u(x, t)+\left(\frac{1}{k^{2}}-\frac{1}{h}\right) u(x, t-h) \\
& \quad-\left\{\left[\frac{1}{2 k^{2}}-\frac{1}{2 \sigma k}\left(\frac{\sigma^{2}}{2}+\lambda \omega-r\right)\right] u(x+\sigma k, t-h)\right. \\
&\left.+\left[\frac{1}{2 k^{2}}+\frac{1}{2 \sigma k}\left(\frac{\sigma^{2}}{2}+\lambda \omega-r\right)\right] u(x-\sigma k, t-h)\right\} \\
&-\lambda \sum_{l=-\infty}^{\infty} u(x+l \sigma k, t-h) p_{l},
\end{aligned}
$$

where $k, h>0$, and

$$
\begin{equation*}
p_{l}=\int_{(l-1 / 2) \sigma k}^{(l+1 / 2) \sigma k} d N(y)=N\left(\left(l+\frac{1}{2}\right) \sigma k\right)-N\left(\left(l-\frac{1}{2}\right) \sigma k\right) . \tag{3.3}
\end{equation*}
$$

We assume that the following Courant-Friedrichs-Lax condition holds:

$$
\begin{equation*}
0<\sqrt{h} \leqslant k \tag{3.4}
\end{equation*}
$$

Note that if $\sigma>0$ and $k$ is small enough, then

$$
\begin{equation*}
\left|\frac{\sigma^{2}}{2}+\lambda \omega-r\right| k \leqslant \sigma \tag{3.5}
\end{equation*}
$$

Remark 3.1. The EFDS (3.1)-(3.2) indeed defines an algorithm for computing $u_{h}$. Indeed, given $u_{h}(\cdot, t-h)$, we can rewrite (3.1) as

$$
\min \left\{r^{-1}\left(\frac{1}{h}+(r+\lambda)\right) u_{h}(x, t)+f(x, t), u_{h}(x, t)-g(x)\right\}=0
$$

where the function $f(x, t)$ denotes all those terms involving only $u_{h}(\cdot, t-h)$. From this equation we find that $u_{h}(x, t)$ is uniquely determined by

$$
u_{h}(x, t)=\max \left(\frac{-r f(x, t)}{h^{-1}+r+\lambda}, g(x)\right) .
$$

Lemma 3.1. Under the conditions (3.4) and (3.5), the EFDS (3.1) and (3.2) is uniformly convergent with respect to $(x, t) \in \mathbb{R}^{1} \times[0, T]$. Moreover, when equality holds in (3.4), the scheme reduces to the BTM for the American option with jump diffusion.

This result is proved by Qian et al. in [34].
Lemma 3.2 (comparison principle 4). Let $F_{h}$ be defined by (3.1) satisfying conditions (3.4) and (3.5). If $u$ and $v$ are piecewise continuous and satisfy

$$
\begin{align*}
& F_{h}[u] \geqslant F_{h}[v] \quad \text { for }(x, t) \in \mathbb{R}^{1} \times(0, T],  \tag{3.6}\\
& u(x, t) \geqslant v(x, t) \quad \text { for }-h<t \leqslant 0, \quad x \in \mathbb{R}^{1}, \tag{3.7}
\end{align*}
$$

then

$$
\begin{equation*}
u(x, t) \geqslant v(x, t) \quad \text { for }(x, t) \in \mathbb{R}^{1} \times(-h, T] . \tag{3.8}
\end{equation*}
$$

Proof. We shall prove by induction

$$
\begin{equation*}
u(x, t) \geqslant v(x, t) \quad \text { for } \quad(x, t) \in \mathbb{R}^{1} \times((n-1) h, n h] \tag{3.9}
\end{equation*}
$$

for $n=0,1,2, \ldots, N$, where $N h=T$. When $n=0$, (3.9) coincides with the assumption (3.7). Assuming that (3.9) holds for $n=m$, we shall prove that (3.9) also holds for $n=m+1$.

We set

$$
a=\frac{1}{h}-\frac{1}{k^{2}}, \quad a_{ \pm}=\frac{1}{k^{2}}\left[\frac{1}{2} \pm \frac{k}{2 \sigma}\left(\frac{\sigma^{2}}{2}+\omega \lambda-r\right)\right]
$$

From (3.4), (3.5),

$$
a \geqslant 0, \quad a_{ \pm} \geqslant 0
$$

For any $(x, t)$ with $m h<t \leqslant(m+1) h$, if

$$
v(x, t)-g(x)=\min \left(r^{-1} L_{h}[v], v(x, t)-g(x)\right)=F_{h}[v]
$$

then clearly

$$
u(x, t)-g(x) \geqslant F_{h}[u] \geqslant F_{h}[v]=v(x, t)-g(x)
$$

which implies $u(x, t) \geqslant v(x, t)$ in this case. On the other hand, if

$$
r^{-1} L_{h}[v]=\min \left(r^{-1} L_{h}[v], v(x, t)-g(x)\right)=F_{h}[v]
$$

then

$$
r^{-1} L_{h}[u] \geqslant F_{h}[u] \geqslant F_{h}[v]=r^{-1} L_{h}[v] .
$$

It follows from (3.9) for $n=m$ and the above inequality that

$$
\begin{aligned}
& \left(\begin{array}{l}
\left.\frac{1}{h}+r+\lambda\right) u(x, t) \\
=L_{h}[u]
\end{array}+a u(x, t-h)+a_{-} u(x+\sigma k, t-h)+a_{+} u(x-\sigma k, t-h)\right. \\
& \quad+\lambda \sum_{l=-\infty}^{\infty} u(x+l \sigma k, t-h) p_{l} \\
& \geqslant L_{h}[v]+a v(x, t-h)+a_{-} v(x+\sigma k, t-h)+a_{+} v(x-\sigma k, t-h) \\
& \quad+\lambda \sum_{l=-\infty}^{\infty} v(x+l \sigma k, t-h) p_{l} \\
& =\left(\frac{1}{h}+r+\lambda\right) v(x, t)
\end{aligned}
$$

and thus we also have $u(x, t) \geqslant v(x, t)$ in this case. The lemma is established.
To simplify notation, we shall take without loss of generality $k=\sqrt{h}$ throughout the rest of this paper.
4. Penalized problems and regularized solution. In this section we shall derive some basic estimates for the penalized problem (2.13), (2.14). We begin by a simple lemma which removes the singularity of the obstacle.

Lemma 4.1. The solution $u^{\varepsilon}$ to the problem (2.13), (2.14) also satisfies

$$
\begin{align*}
& L\left[u^{\varepsilon}\right]=\beta_{\varepsilon}\left(1-e^{x}-u^{\varepsilon}\right), \quad(x, t) \in \mathbb{R}^{1} \times(0, T),  \tag{4.1}\\
& u^{\varepsilon}(x, 0)=g(x)=\left(1-e^{x}\right)^{+}, \quad x \in \mathbb{R}^{1} \tag{4.2}
\end{align*}
$$

Proof. We shall establish

$$
\begin{equation*}
\beta_{\varepsilon}\left(1-e^{x}-u^{\varepsilon}(x, t)\right) \equiv \beta_{\varepsilon}\left(g(x)-u^{\varepsilon}(x, t)\right) . \tag{4.3}
\end{equation*}
$$

If $x \leqslant 0$, the above equality is obvious.
By comparison principle 2 (Lemma 2.2), $u^{\varepsilon} \geqslant 0$ everywhere. Thus for $x \geqslant 0$, we have $1-e^{x}-u^{\varepsilon}(x, t) \leqslant 0$ and $g(x)-u^{\varepsilon}(x, t) \leqslant 0$. Therefore from the definition of $\beta_{\varepsilon}$, we get $\beta_{\varepsilon}\left(1-e^{x}-u^{\varepsilon}(x, t)\right) \equiv \beta_{\varepsilon}\left(g(x)-u^{\varepsilon}(x, t)\right) \equiv 0$ for $x \geqslant 0$.

Remark 4.1. Since $\beta_{\varepsilon}$ is a smooth function with bounded first order derivative (although the bound depends on $\varepsilon$ ), the solution $u^{\varepsilon}$ is smooth in $\left(\mathbb{R}^{1} \times[0, T]\right) \backslash\left(B_{\delta}(0) \times[0, \delta]\right)$, for any $\delta>0$, by the parabolic theory. In particular, $u^{\varepsilon}$ is continuous and bounded on $\left(\mathbb{R}^{1} \times[0, T]\right)$, and $\left(u^{\varepsilon}\right)_{x},\left(u^{\varepsilon}\right)_{x x},\left(u^{\varepsilon}\right)_{t}$ are bounded and continuous on $\left(\mathbb{R}^{1} \times[0, T]\right) \backslash\left(B_{\delta}(0) \times[0, \delta]\right)$ for any $\delta>0$.

Lemma 4.2. There hold

$$
\begin{align*}
& \frac{\partial}{\partial x}\left(u^{\varepsilon}(x, t)-\left(1-e^{x}\right)\right) \geqslant 0 \quad \text { for }(x, t) \in \mathbb{R}^{1} \times(0, T),  \tag{4.4}\\
& \frac{\partial u^{\varepsilon}(x, t)}{\partial x} \geqslant-1 \quad \text { for }(x, t) \in \mathbb{R}^{1} \times(0, T),  \tag{4.5}\\
& \frac{\partial u^{\varepsilon}(x, t)}{\partial x} \leqslant 0 \quad \text { for } \quad(x, t) \in \mathbb{R}^{1} \times(0, T) . \tag{4.6}
\end{align*}
$$

Proof. First, by a direct computation using (2.7), we derive

$$
L\left[1-e^{x}\right]=r .
$$

Differentiating (4.1) in $x$ and using the above equality, we obtain

$$
\begin{equation*}
L\left[\left(u^{\varepsilon}-\left(1-e^{x}\right)\right)_{x}\right]=-\beta_{\varepsilon}^{\prime}\left(1-e^{x}-u^{\varepsilon}\right) \cdot\left(u^{\varepsilon}-\left(1-e^{x}\right)\right)_{x} . \tag{4.7}
\end{equation*}
$$

From the initial condition we can easily verify that when $t=0$, (4.4) holds. Note that $u_{x}^{\varepsilon}$ has a discontinuity at $(x, t)=(0,0)$. But we can always approximate the initial data with smooth functions and take the limit. Since the solution to system (4.1)-(4.2) is unique (for each fixed $\varepsilon$ ), we can approximate the initial data in any way we wish while yielding the same limit. To prove (4.4), we shall approximate the initial data with smooth function $g^{\delta}$ such that $\left(g^{\delta}-\left(1-e^{x}\right)\right)_{x} \geqslant 0$. Applying comparison principle 1 (Lemma 2.1) to the approximated system and then taking the limit, we obtain (4.4).

Similarly, differentiating (4.1) in $x$,

$$
L\left[u_{x}^{\varepsilon}\right]+\beta_{\varepsilon}^{\prime}\left(1-e^{x}-u^{\varepsilon}\right) \cdot u_{x}^{\varepsilon}=-\beta_{\varepsilon}^{\prime}\left(1-e^{x}-u^{\varepsilon}\right) \cdot e^{x} \leqslant 0,
$$

so that we can apply comparison principle 1 (Lemma 2.1) in a manner similar to the approximated system to derive (4.6).

Finally

$$
\begin{aligned}
& L\left[u_{x}^{\varepsilon}+1\right]+\beta_{\varepsilon}^{\prime}\left(1-e^{x}-u^{\varepsilon}\right) \cdot\left(u_{x}^{\varepsilon}+1\right) \\
& \quad=r+\beta_{\varepsilon}^{\prime}\left(1-e^{x}-u^{\varepsilon}\right) \cdot\left(1-e^{x}\right)>\beta_{\varepsilon}^{\prime}\left(1-e^{x}-u^{\varepsilon}\right) \cdot\left(1-e^{x}\right) .
\end{aligned}
$$

Recalling, from the end of the proof of Lemma 4.1, that $\beta_{\varepsilon}\left(1-e^{x}-u^{\varepsilon}\right) \equiv 0$ for $x \geqslant 0$, we conclude that

$$
\beta_{\varepsilon}^{\prime}\left(1-e^{x}-u^{\varepsilon}\right) \cdot\left(1-e^{x}\right) \begin{cases}=0 & \text { if } x \geqslant 0 \\ >0 & \text { if } x<0\end{cases}
$$

Again, we can apply comparison principle 1 (Lemma 2.1) in a manner similar to the approximated system to derive (4.5).

Lemma 4.3. There holds

$$
\begin{equation*}
\beta_{\varepsilon}\left(g-u^{\varepsilon}\right) \leqslant r, \tag{4.8}
\end{equation*}
$$

and therefore

$$
\begin{equation*}
g-u^{\varepsilon} \leqslant(r+1) \varepsilon \tag{4.9}
\end{equation*}
$$

Proof. For any $\delta>0$, let $z=1-e^{x}-u^{\varepsilon}, z_{\delta}=z-\delta\left(c_{0}+x^{2}\right)$, where $c_{0}$ is the positive constant defined in (2.10). By (2.10) we also have

$$
L\left[z_{\delta}\right]=\left[1-e^{x}\right]-L\left[u^{\varepsilon}\right]-\delta L\left[c_{0}+x^{2}\right] \leqslant r-\beta_{\varepsilon}(z)-\delta .
$$

Clearly, $z$ and $z_{\delta}$ are bounded from above. At the time $t=0,\left.z_{\delta}\right|_{t=0}<0$. For fixed $t>0$, $\lim _{x \rightarrow \pm \infty} z_{\delta}(x, t)=-\infty$. Thus the maximum of $z_{\delta}$ on $\mathbb{R}^{1} \times[0, T]$, if positive, must be attained at an interior point $\left(x_{0}, t_{0}\right)$, with

$$
\left.L\left[z_{\delta}\right]\right|_{x=x_{0}, t=t_{0}} \geqslant 0
$$

for any $\delta>0$. Since $\beta_{\varepsilon}^{\prime} \geqslant 0$, we obtain

$$
\begin{aligned}
& \max _{x \in \mathbb{R}^{1}, 0 \leqslant t \leqslant T} \beta_{\varepsilon}\left(z_{\delta}(x, t)\right) \leqslant \beta_{\varepsilon}\left(z_{\delta}\left(x_{0}, t_{0}\right)\right) \\
& \leqslant \beta_{\varepsilon}\left(z\left(x_{0}, t_{0}\right)\right)=\left.\left(r-L\left[z_{\delta}\right]-\delta L\left[c_{0}+x^{2}\right]\right)\right|_{x=x_{0}, t=t_{0}} \leqslant r-\delta .
\end{aligned}
$$

If the maximum of $z_{\delta}$ on $\mathbb{R}^{1} \times[0, T]$ is negative, then

$$
\max _{x \in \mathbb{R}^{1}, 0 \leqslant t \leqslant T} \beta_{\varepsilon}\left(z_{\delta}(x, t)\right) \leqslant \beta_{\varepsilon}\left(z_{\delta}\left(x_{0}, t_{0}\right)\right)=0<r .
$$

Letting $\delta \rightarrow 0$, in both cases, we obtain

$$
\beta_{\varepsilon}(z(x, t)) \leqslant r .
$$

Also using (4.3) we derive (4.8).
Since $\beta^{\prime \prime} \geqslant 0$, we have $\beta(z) \geqslant z-1$ for all $z$, and thus (4.9) is now an immediate consequence from the definitions of $\beta$ and $\beta_{\varepsilon}$.

By using the bounds on $\beta_{\varepsilon}, u_{x}^{\varepsilon}$ and the equation, we immediately obtain the following lemma.

Lemma 4.4.

$$
\begin{equation*}
\left[u^{\varepsilon}\right]_{C_{t}^{1 / 2}\left(\mathbb{R}^{1} \times[0, T]\right)} \leqslant C . \tag{4.10}
\end{equation*}
$$

Proof. For any function $\varphi$ with continuous first order $t$-derivative and continuous second order $x$-derivative in the interior of the domain, and for any $x$ and $t_{2}>t_{1}>0$, take $x_{1}<x<$ $x_{2}$. Then

$$
\begin{aligned}
\varphi\left(x, t_{2}\right)-\varphi\left(x, t_{1}\right)= & \int_{x_{1}}^{x_{2}}\left\{\frac{\varphi\left(x, t_{2}\right)-\varphi\left(x, t_{1}\right)}{x_{2}-x_{1}}-\frac{\varphi\left(\xi, t_{2}\right)-\varphi\left(\xi, t_{1}\right)}{x_{2}-x_{1}}\right\} d \xi \\
& +\frac{1}{x_{2}-x_{1}} \int_{x_{1}}^{x_{2}} \int_{t_{1}}^{t_{2}}\left\{\varphi_{t}(\xi, \tau)-\frac{\sigma^{2}}{2} \varphi_{x x}(\xi, \tau)\right\} d \tau d \xi \\
& +\frac{\sigma^{2}}{2\left(x_{2}-x_{1}\right)} \int_{t_{1}}^{t_{2}}\left(\varphi_{x}\left(x_{2}, \tau\right)-\varphi_{x}\left(x_{1}, \tau\right)\right) d \tau
\end{aligned}
$$

Applying the mean value theorem to the first term, we find that

$$
\begin{aligned}
\left|\varphi\left(x, t_{2}\right)-\varphi\left(x, t_{1}\right)\right| \leqslant & 2\left\|\varphi_{x}\right\|_{L^{\infty}}\left(x_{2}-x_{1}\right)+\left\|\varphi_{t}-\frac{\sigma^{2}}{2} \varphi_{x x}\right\|_{L^{\infty}} \frac{t_{2}-t_{1}}{x_{2}-x_{1}} \\
& +\frac{\sigma^{2}}{2} 2\left\|\varphi_{x}\right\|_{L^{\infty}} \frac{t_{2}-t_{1}}{x_{2}-x_{1}} .
\end{aligned}
$$

Upon taking $x_{2}-x_{1}=\sqrt{t_{2}-t_{1}}$ and taking "sup" over $x \in \mathbb{R}^{1}$ and $0<t_{1}<t_{2}<T$, we obtain

$$
[\varphi]_{C_{t}^{1 / 2}\left(\mathbb{R}^{1} \times[0, T]\right)} \leqslant(2+\sigma)\left\|\varphi_{x}\right\|_{L^{\infty}}+\left\|\varphi_{t}-\frac{\sigma^{2}}{2} \varphi_{x x}^{\varepsilon}\right\|_{L^{\infty}}
$$

Applying this inequality and Lemmas 4.2 and 4.3 to $u^{\varepsilon}$, we obtain

$$
\left[u^{\varepsilon}\right]_{C_{t}^{1 / 2}\left(\mathbb{R}^{1} \times[0, T]\right)} \leqslant C\left(\left\|u_{x}^{\varepsilon}\right\|_{L^{\infty}}+\left\|u_{t}^{\varepsilon}-\frac{\sigma^{2}}{2} u_{x x}^{\varepsilon}\right\|_{L^{\infty}}\right) \leqslant C
$$

where the norms of the spaces are defined before Lemma 2.1.
Lemma 4.5. The solution $u^{\varepsilon}$ of the problem (2.13), (2.14) satisfies, for $0<\alpha<1$,

$$
\begin{equation*}
\left\|u^{\varepsilon}\right\|_{C^{1+\alpha,(1+\alpha) / 2}\left(\mathbb{R}^{1} \times[\tau, T]\right)} \leqslant C \tau^{-\alpha / 2} \tag{4.11}
\end{equation*}
$$

where $C=C_{\alpha}$ is independent of $\varepsilon$ and $\tau$.
Proof. Let $U$ be the European option defined in (2.18)-(2.19). We decompose $u^{\varepsilon}$ into two terms:

$$
\begin{equation*}
u^{\varepsilon}=U+v^{\varepsilon} . \tag{4.12}
\end{equation*}
$$

Then $v^{\varepsilon}$ does not have any singularities at $t=0$, and

$$
\begin{align*}
& L\left[v^{\varepsilon}\right]=\beta_{\varepsilon}\left(1-e^{x}-U-v^{\varepsilon}\right) \quad \text { for } \quad(x, t) \in \mathbb{R}^{1} \times(0, T),  \tag{4.13}\\
& \left.v^{\varepsilon}\right|_{t=0} \equiv 0 \tag{4.14}
\end{align*}
$$

Since $\beta_{\varepsilon}\left(1-e^{x}-U-v^{\varepsilon}\right)$ is uniformly bounded, $v^{\varepsilon}$ is uniformly bounded. Thus, by [31, Theorem 4.30, p. 79], we obtain

$$
\begin{equation*}
\left\|v^{\varepsilon}\right\|_{C^{1+\alpha,(1+\alpha) / 2}\left(\mathbb{R}^{1} \times[0, T]\right)} \leqslant C \tag{4.15}
\end{equation*}
$$

for some uniform constant $C$. Hence $u^{\varepsilon}=U+v^{\varepsilon}$ satisfies

$$
\begin{equation*}
\left\|u^{\varepsilon}\right\|_{C^{1+\alpha,(1+\alpha) / 2}\left(\mathbb{R}^{1} \times[\tau, T]\right)} \leqslant\|U\|_{C^{1+\alpha,(1+\alpha) / 2}\left(\mathbb{R}^{1} \times[\tau, T]\right)}+C . \tag{4.16}
\end{equation*}
$$

The lemma now follows from the estimate for $U$ (cf. (2.23)).
Lemma 4.6. The solution $u^{\varepsilon}$ of the problem (2.13), (2.14) satisfies

$$
\begin{align*}
& -(r+1) \varepsilon \leqslant F\left[u^{\varepsilon}\right] \leqslant 0 \quad \text { for }(x, t) \in \mathbb{R}^{1} \times(0, T),  \tag{4.17}\\
& 0 \leqslant u-u^{\varepsilon} \leqslant(r+1) \varepsilon \quad \text { for }(x, t) \in \mathbb{R}^{1} \times[0, T] . \tag{4.18}
\end{align*}
$$

Proof. From the definition of $\beta_{\varepsilon}$, it is clear that

$$
\min \left(r^{-1} L\left[u^{\varepsilon}\right], u^{\varepsilon}-g\right)=\min \left(r^{-1} \beta_{\varepsilon}\left(g-u^{\varepsilon}\right), u^{\varepsilon}-g\right) \leqslant 0 .
$$

It follows from Lemma 4.3 that

$$
\begin{aligned}
\min \left(r^{-1} L\left[u^{\varepsilon}\right], u^{\varepsilon}-g\right) & =\min \left(r^{-1} \beta_{\varepsilon}\left(g-u^{\varepsilon}\right), u^{\varepsilon}-g\right) \\
& \geqslant \min (0,-(r+1) \varepsilon) \\
& \geqslant-(r+1) \varepsilon .
\end{aligned}
$$

Now we have proved (4.17). For (4.18), since

$$
\begin{align*}
& F[u]=0 \geqslant F\left[u^{\varepsilon}\right], \quad(x, t) \in \mathbb{R}^{1} \times(0, T),  \tag{4.19}\\
& u(x, 0) \equiv u^{\varepsilon}(x, 0), \tag{4.20}
\end{align*}
$$

we conclude from Lemma 2.3 that $u(x, t) \geqslant u^{\varepsilon}(x, t)$ for $(x, t) \in \mathbb{R}^{1} \times[0, T]$.
Similarly, using (4.17), we obtain

$$
\begin{aligned}
F[u-(r+1) \varepsilon] & =-(r+1) \varepsilon+F[u] \\
& =-(r+1) \varepsilon \leqslant F\left[u^{\varepsilon}\right], \quad(x, t) \in \mathbb{R}^{1} \times(0, T) .
\end{aligned}
$$

By Lemma 2.3 again, we derive $u(x, t)-(r+1) \varepsilon \leqslant u^{\varepsilon}(x, t)$ for $(x, t) \in \mathbb{R}^{1} \times[0, T]$.
5. Higher order derivative estimates for regularized solutions. In order to get estimates for higher order derivatives of the regularized solution of the penalized problem, we have to deal with the penalized term $\beta_{\varepsilon}\left(g-v_{\varepsilon}\right)$ and its derivatives. We start with an interpolation inequality (cf. [31, inequality (4.51), p. 73 with $b=d=0]$ ).

Lemma 5.1 (interpolation inequality). Suppose that

$$
\begin{gathered}
0 \leqslant \alpha, \beta, \gamma \leqslant 1, \quad l+\gamma<k+\alpha<m+\beta, \\
0<\sigma<1, \quad \sigma(m+\beta)+(1-\sigma)(l+\gamma)=k+\alpha .
\end{gathered}
$$

If the domain satisfies the uniform interior cone condition, then

$$
\begin{equation*}
\|u\|_{C^{k+\alpha,(k+\alpha) / 2}} \leqslant C\|u\|_{C^{m+\beta,(m+\beta) / 2}}^{\sigma}\|u\|_{C^{l+\gamma,(l+\gamma) / 2}}^{1-\sigma} . \tag{5.1}
\end{equation*}
$$

We also need the following lemma.

Lemma 5.2.

$$
\begin{align*}
& {[u \cdot v]_{C^{\alpha, \alpha / 2}} \leqslant\|u\|_{L^{\infty}}[v]_{C^{\alpha, \alpha / 2}}+[u]_{C^{\alpha, \alpha / 2}}\|v\|_{L^{\infty}},}  \tag{5.2}\\
& \|u \cdot v\|_{C^{\alpha, \alpha / 2}} \leqslant\|u\|_{L^{\infty}}\|v\|_{C^{\alpha, \alpha / 2}}+\|u\|_{C^{\alpha, \alpha / 2}}\|v\|_{L^{\infty}} . \tag{5.3}
\end{align*}
$$

Inequality (5.2) can be found in [9, Lemma 1.1, p. 19], and inequality (5.3) is a corollary of (5.2) and the definition.

We next derive estimates for $v_{t}^{\varepsilon}$ and $v_{x x}^{\varepsilon}$, where $v^{\varepsilon}$ is defined by (4.12).
Lemma 5.3. For $(x, t) \in \mathbb{R}^{1} \times(0, T)$, there hold

$$
\begin{gather*}
v_{t}^{\varepsilon}(x, t) \leqslant C  \tag{5.4}\\
v_{x x}^{\varepsilon}(x, t) \leqslant C \tag{5.5}
\end{gather*}
$$

where the constant $C$ is independent of $\varepsilon$ (but may depend on $T$ ).
Proof. For each fixed $\varepsilon$, the solution to system (4.13)-(4.14) is unique. Thus we can approximate the right-hand side of (4.13) in any way we wish while yielding the same limit. Using (2.21), we can take $C>0$ such that $C+U_{t}>0$. We shall approximate $U(x, t)$ in (4.13) by smooth functions $U^{\delta}(x, t)$ such that $C+U_{t}^{\delta}(x, t)>0$ and $\beta_{\varepsilon}\left(1-e^{x}+U^{\delta}(x, 0)\right)=0$. Then the approximated solution $v^{\varepsilon, \delta}$ is smooth with $\left(v^{\varepsilon, \delta}\right)_{t}$ bounded and continuous on $\mathbb{R}^{1} \times[0, T]$.

Differentiating the approximated system of (4.13) in $t$, we obtain

$$
\begin{aligned}
& L\left[v_{t}^{\varepsilon, \delta}-C\right]+\beta_{\varepsilon}^{\prime}\left(1-e^{x}-U^{\delta}-v^{\varepsilon, \delta}\right) \cdot\left(v_{t}^{\varepsilon, \delta}-C\right) \\
& \quad=-C r-\beta_{\varepsilon}^{\prime}\left(1-e^{x}-U^{\delta}-v^{\varepsilon, \delta}\right) \cdot\left(C+U_{t}^{\delta}\right) \leqslant 0,(x, t) \in \mathbb{R}^{1} \times(0, T)
\end{aligned}
$$

It is clear that $v_{t}^{\varepsilon, \delta}-\left.C\right|_{t=0}=-C \leqslant 0$, so that, by comparison principle 1 (Lemma 2.1),

$$
v_{t}^{\varepsilon, \delta} \leqslant C \quad \text { for } \quad(x, t) \in \mathbb{R}^{1} \times(0, T) .
$$

Taking the limit as $\delta \rightarrow 0$, we obtain (5.4).
Inequality (5.5) now follows from (5.4), (4.15), (4.8), and (4.13).
Corollary 5.4. For $(x, t) \in \mathbb{R}^{1} \times(0, T)$, there hold

$$
\begin{align*}
& -C \leqslant u_{t}^{\varepsilon}(x, t) \leqslant C+\frac{C}{\sqrt{t}} \exp \left(-\frac{x^{2}}{C t}\right),  \tag{5.6}\\
& -C \leqslant u_{x x}^{\varepsilon}(x, t) \leqslant C+\frac{C}{\sqrt{t}} \exp \left(-\frac{x^{2}}{C t}\right),  \tag{5.7}\\
& u_{x x}^{\varepsilon}(x, t) \geqslant-e^{x} \tag{5.8}
\end{align*}
$$

where the constant $C$ is independent of $\varepsilon$ (but may depend on $T$ ).
Proof. The upper bounds in (5.6)-(5.7) follow from Lemma 5.3 and (2.22). It is clear that for any positive constant $C_{0}, u_{t}^{\varepsilon}+C_{0}$ satisfies

$$
\begin{aligned}
& L\left[u_{t}^{\varepsilon}+C_{0}\right]+\beta_{\varepsilon}^{\prime}\left(1-e^{x}-u^{\varepsilon}\right) \cdot\left(u_{t}^{\varepsilon}+C_{0}\right) \\
& \quad=C_{0}\left(r+\beta_{\varepsilon}^{\prime}\left(1-e^{x}-u^{\varepsilon}\right)\right) \geqslant 0 \quad \text { for }(x, t) \in \mathbb{R}^{1} \times(0, T) .
\end{aligned}
$$

Note that the initial datum for $u_{t}^{\varepsilon}+C_{0}$ is smooth except $x=0$. At $x=0, u^{\varepsilon}(x, 0)=g(x)$ is only Lipschitz continuous and the second derivative $g_{x x}$ produces a delta function at $x=0$ (Dirac measure, or point mass) which is nonnegative. Again, since the solution $u^{\varepsilon}$ to the system (4.1)-(4.2) is unique, we can approximate the system in any way we wish while yielding the same limit. We shall approximate the initial data in (4.2) with smooth function $g^{\delta}(x)$ such that $\left(g^{\delta}\right)_{x x}(x) \geqslant-1$. Then the approximated solution $u^{\varepsilon, \delta}$ satisfies

$$
\lim _{t \rightarrow 0+} u_{t}^{\varepsilon, \delta}(x, t) \geqslant-C_{0}
$$

where some constant $C_{0}$ is independent of $\varepsilon, \delta$. Applying comparison principle 1 (Lemma 2.1) to $u_{t}^{\varepsilon, \delta}$, we conclude that

$$
u_{t}^{\varepsilon, \delta}(x, t) \geqslant-C_{0} \quad \text { for } \quad(x, t) \in \mathbb{R}^{1} \times[0, T]
$$

Taking the limit as $\delta \rightarrow 0$, we obtain the lower bound for $u_{t}^{\varepsilon}$ in (5.6).
Using (4.8), (4.15), and (4.13), we also obtain $u_{x x}^{\varepsilon} \geqslant-C$.
To establish (5.8), we differentiate (4.7) in $x$ and use the fact that $\beta_{\varepsilon}^{\prime \prime} \geqslant 0$ to obtain, for $\varphi^{\varepsilon}=u^{\varepsilon}-1+e^{x}$, the equation

$$
L\left[\varphi_{x x}^{\varepsilon}\right]+\beta_{\varepsilon}^{\prime}\left(-\varphi^{\varepsilon}\right) \cdot \varphi_{x x}^{\varepsilon}=\beta_{\varepsilon}^{\prime \prime}\left(-\varphi^{\varepsilon}\right) \cdot\left(\varphi_{x}^{\varepsilon}\right)^{2} \geqslant 0
$$

Again, approximating the initial data for $u^{\varepsilon}$ with smooth functions such that $\varphi_{x x}^{\varepsilon, \delta}(x, 0) \geqslant 0$, applying comparison principle 1 (Lemma 2.1) to the approximated system then taking the limit, we obtain $\varphi_{x x}^{\varepsilon}(x, t) \geqslant 0$, and this implies (5.8).

To derive further quantitative regularity of the approximating system we want to study the function

$$
\begin{equation*}
f^{\varepsilon}:=\beta_{\varepsilon}\left(g-u^{\varepsilon}\right)+\lambda \int_{\mathbb{R}^{1}} u^{\varepsilon}(x+y, t) d N(y) \tag{5.9}
\end{equation*}
$$

Lemma 5.5. There hold

$$
\begin{align*}
& \left|f_{t}^{\varepsilon}(x, t)\right| \leqslant C \varepsilon^{-1} t^{-1 / 2}  \tag{5.10}\\
& \left|f_{x}^{\varepsilon}(x, t)\right| \leqslant C \varepsilon^{-1 / 2} t^{-1 / 4}, \quad\left|f_{x}^{\varepsilon}(x, t)\right| \leqslant C \varepsilon^{-1},  \tag{5.11}\\
& \left|f_{x x}^{\varepsilon}(x, t)\right| \leqslant C \varepsilon^{-1} t^{-1 / 2} \tag{5.12}
\end{align*}
$$

Proof. Differentiating (5.9) in $t$, we obtain

$$
\begin{equation*}
f_{t}^{\varepsilon}=-\beta_{\varepsilon}^{\prime} \cdot u_{t}^{\varepsilon}+\lambda \int_{\mathbb{R}^{1}} u_{t}^{\varepsilon}(x+y, t) d N(y) \tag{5.13}
\end{equation*}
$$

Using (5.6) and the definition of $\beta_{\varepsilon}$, we immediately obtain (5.10).
Differentiating (5.9) in $x$, we obtain

$$
\begin{align*}
f_{x}^{\varepsilon} & =\beta_{\varepsilon}^{\prime} \cdot\left(g_{x}-u_{x}^{\varepsilon}\right)+\lambda \int_{\mathbb{R}^{1}} u_{x}^{\varepsilon}(x+y, t) d N(y)  \tag{5.14}\\
f_{x x}^{\varepsilon} & =\beta_{\varepsilon}^{\prime} \cdot\left(g_{x x}-u_{x x}^{\varepsilon}\right)+\beta_{\varepsilon}^{\prime \prime} \cdot\left(g_{x}-u_{x}^{\varepsilon}\right)^{2}+\lambda \int_{\mathbb{R}^{1}} u_{x x}^{\varepsilon}(x+y, t) d N(y) \tag{5.15}
\end{align*}
$$

By Lemma 4.5 and (5.7), it is not difficult to see that the integral terms in (5.14) and (5.15) satisfy, when $0<\varepsilon<1,0<t<T$,

$$
\begin{align*}
& \lambda \int_{\mathbb{R}^{1}} u_{x}^{\varepsilon}(x+y, t) d N(y) \leqslant C \leqslant \min \left\{\varepsilon^{-1 / 2} t^{-1 / 4}, \varepsilon^{-1}\right\},  \tag{5.16}\\
& \lambda \int_{\mathbb{R}^{1}} u_{x x}^{\varepsilon}(x+y, t) d N(y) \leqslant C t^{-1 / 2} \leqslant C \varepsilon^{-1} t^{-1 / 2} . \tag{5.17}
\end{align*}
$$

This gives the estimates for the integral terms in (5.14) and (5.15).
To estimate the terms in (5.14) and (5.15) involving $\beta_{\varepsilon}$, we need only consider the region $\left\{(x, t) ; g-u^{\varepsilon}>0\right\}$, since $\beta_{\varepsilon}\left(g-u^{\varepsilon}\right) \equiv 0$ outside this region.

For $x \geqslant 0, t>0$, we have $u^{\varepsilon}>0=g(x)$. For each $t>0$, we can derive from (4.4) that there exists $s_{\varepsilon}(t)<0$ such that

$$
\left\{x ; g(x)-u^{\varepsilon}(x, t)>0\right\}=\left\{x ; x<s_{\varepsilon}(t)\right\} .
$$

Thus, for each fixed $t$, we can use (4.9), (5.7), and the interpolation inequality (Lemma 5.1) on the interval $\left(-\infty, s_{\varepsilon}(t)\right)$ to obtain (the domain condition for interpolation is satisfied)

$$
\begin{align*}
& \left\|g_{x}-u_{x}^{\varepsilon}(\cdot, t)\right\|_{L^{\infty}\left(-\infty, s_{\varepsilon}(t)\right)} \\
& \quad \leqslant C\left\|g-u^{\varepsilon}\right\|_{W^{2, \infty}\left(-\infty, s_{\varepsilon}(t)\right)}^{1 / 2}\left\|g-u^{\varepsilon}\right\|_{L^{\infty}\left(-\infty, s_{\varepsilon}(t)\right)}^{1 / 2} \leqslant C \varepsilon^{1 / 2} t^{-1 / 4} . \tag{5.18}
\end{align*}
$$

Substituting this estimate and (5.16) into (5.14) we establish the first inequality in (5.11). The second inequality in (5.11) is a direct consequence of the first order derivative estimates and the definition of $\beta_{\varepsilon}$.

To establish (5.12) in the region $x<0$, we note that in this region $g_{x x}=-e^{x}$ is bounded and $u_{x x}^{\varepsilon}$ is estimated by (5.7). Thus the absolute value of the first term in (5.15) is estimated by $C \varepsilon^{-1} t^{-1 / 2}$. The second term in (5.15) can be estimated in a similar way, using the definition of $\beta_{\varepsilon}$ and (5.18). The third estimate comes from (5.17).

Lemma 5.6. The solution $u^{\varepsilon}$ of the problem (2.13), (2.14) satisfies, for $0<\alpha<1$,

$$
\begin{equation*}
\left\|u^{\varepsilon}\right\|_{C^{2+\alpha,(2+\alpha) / 2}\left(\mathbb{R}^{1} \times[\tau, T]\right)} \leqslant C \tau^{-(1+\alpha) / 2}+C(\varepsilon \sqrt{\tau})^{-\alpha / 2} \tag{5.19}
\end{equation*}
$$

where $C=C_{\alpha}$ is independent of $\varepsilon$ and $\tau$.
Proof. We write $u^{\varepsilon}=U+v^{\varepsilon}$ as in (4.12). Then $v^{\varepsilon}$ satisfies (4.13)-(4.14). We take a cut-off function $\zeta \in C^{\infty}\left(\mathbb{R}^{1}\right)$ ( $\zeta$ depends on $\tau$ ) such that

$$
\begin{gather*}
\zeta(t)=0 \quad \text { for } t<\tau / 2, \quad \zeta(t)=1 \quad \text { for } t>\tau,  \tag{5.20}\\
0 \leqslant \zeta(t) \leqslant 1, \quad 0 \leqslant \zeta^{\prime}(t) \leqslant C \tau^{-1}, \quad\left|\zeta^{\prime \prime}(t)\right| \leqslant C \tau^{-2} . \tag{5.21}
\end{gather*}
$$

By the interpolation inequality (Lemma 5.1) and (5.21), we immediately derive

$$
\begin{align*}
& \|\zeta\|_{C^{\alpha / 2}} \leqslant C\|\zeta\|_{C^{0}}^{(2-\alpha) / 2}\|\zeta\|_{C^{1}}^{\alpha / 2} \leqslant C \tau^{-\alpha / 2}  \tag{5.22}\\
& \left\|\zeta^{\prime}\right\|_{C^{\alpha / 2}} \leqslant C\left\|\zeta^{\prime}\right\|_{C^{0}}^{(2-\alpha) / 2}\left\|\zeta^{\prime}\right\|_{C^{1}}^{\alpha / 2} \leqslant C \tau^{-1-\alpha / 2} .
\end{align*}
$$

Clearly, $\zeta v^{\varepsilon}$ satisfies

$$
\begin{equation*}
L_{0}\left[\zeta v^{\varepsilon}\right]=\zeta \cdot f^{\varepsilon}+\zeta^{\prime} \cdot v^{\varepsilon}:=F^{\varepsilon}, \tag{5.23}
\end{equation*}
$$

where $L_{0}$ is a linear parabolic differential operator defined by

$$
L_{0}[u]=u_{t}-\frac{\sigma^{2}}{2} u_{x x}-\left(r-\lambda \omega-\frac{\sigma^{2}}{2}\right) u_{x}+(r+\lambda) u .
$$

By the interpolation inequality (Lemma 5.1) and Lemma 5.5,

$$
\left\|f^{\varepsilon}\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[\tau / 2, T]\right)} \leqslant C\left\|f^{\varepsilon}\right\|_{C^{2,1}\left(\mathbb{R}^{1} \times[\tau / 2, T]\right)}^{\alpha / 2}\left\|f^{\varepsilon}\right\|_{L^{\infty}}^{1-\alpha / 2} \leqslant C(\varepsilon \sqrt{\tau})^{-\alpha / 2}
$$

so that (notice that $\zeta(t) \equiv 0$ for $t \leqslant \tau / 2$ )

$$
\begin{align*}
& \left\|\zeta \cdot f^{\varepsilon}\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[0, T]\right)} \\
& \leqslant\|\zeta\|_{L^{\infty}}\left\|f^{\varepsilon}\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[\tau / 2, T]\right)}+\|\zeta\|_{C^{\alpha / 2}([\tau / 2, T])}\left\|f^{\varepsilon}\right\|_{L^{\infty}}  \tag{5.24}\\
& \leqslant C(\varepsilon \sqrt{\tau})^{-\alpha / 2}+C \tau^{-\alpha / 2} .
\end{align*}
$$

Next, since $\left.v^{\varepsilon}\right|_{t=0} \equiv 0$, we can apply Lemma 4.4 and (2.20) to obtain

$$
\begin{equation*}
\left\|v^{\varepsilon}\right\|_{L^{\infty}\left(\mathbb{R}^{1} \times[0, \tau]\right)} \leqslant\left[v^{\varepsilon}\right]_{C_{t}^{1 / 2}\left(\mathbb{R}^{1} \times[0, \tau]\right)} \tau^{1 / 2} \leqslant C \tau^{1 / 2} . \tag{5.25}
\end{equation*}
$$

By the interpolation inequality (Lemma 5.1), (5.25), and (4.15), we obtain

$$
\begin{equation*}
\left\|v^{\varepsilon}\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[0, \tau]\right)} \leqslant C\left\|v^{\varepsilon}\right\|_{C^{1,1 / 2}\left(\mathbb{R}^{1} \times[0, \tau]\right)}^{\alpha}\left\|v^{\varepsilon}\right\|_{L^{\infty}\left(\mathbb{R}^{1} \times[0, \tau]\right)}^{1-\alpha} \leqslant C \tau^{(1-\alpha) / 2} . \tag{5.26}
\end{equation*}
$$

Since $\zeta^{\prime}(t) \equiv 0$ for $t \leqslant \tau / 2$ or $t \geqslant \tau$, we derive (using (5.3))

$$
\begin{align*}
& \left\|\zeta^{\prime} \cdot v^{\varepsilon}\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[0, T]\right)} \\
& \leqslant\left\|\zeta^{\prime}\right\|_{L^{\infty}[\tau / 2, \tau]}\left\|v^{\varepsilon}\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[\tau / 2, \tau]\right)}+\left\|\zeta^{\prime}\right\|_{C^{\alpha / 2}([\tau / 2, \tau])}\left\|v^{\varepsilon}\right\|_{L^{\infty}[\tau / 2, \tau]} \\
& \leqslant C \tau^{-1} \tau^{(1-\alpha) / 2}+C \tau^{-1-\alpha / 2} \tau^{1 / 2} \leqslant C \tau^{-(1+\alpha) / 2} . \tag{5.27}
\end{align*}
$$

Combining (5.24) and (5.27) we find that the $F^{\varepsilon}$ defined in (5.23) satisfies

$$
\begin{equation*}
\left\|F^{\varepsilon}\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[0, T]\right)} \leqslant C \tau^{-(1+\alpha) / 2}+C(\varepsilon \sqrt{\tau})^{-\alpha / 2} . \tag{5.28}
\end{equation*}
$$

Thus by the Schauder estimate (cf. [31, Theorem 4.9, p. 59]) on (5.23), we have

$$
\begin{equation*}
\left\|\zeta \cdot v^{\varepsilon}\right\|_{C^{2+\alpha,(2+\alpha) / 2}\left(\mathbb{R}^{1} \times[0, T]\right)} \leqslant C \tau^{-(1+\alpha) / 2}+C(\varepsilon \sqrt{\tau})^{-\alpha / 2} . \tag{5.29}
\end{equation*}
$$

Restricting the norm on the left-hand side of the above inequality to $\mathbb{R}^{1} \times[\tau, T]$ (where $\zeta \equiv 1$ ), we obtain the estimates for $v^{\varepsilon}$. Combining it with the estimates for the European option $U$, the lemma follows.

In order to estimate the truncation error between differential operator $L$ and finite difference operator $L_{h}$ for regularized solutions of penalized problem, we need the following main estimate.

Lemma 5.7 (main estimates). For any $1 / 3<\eta<1 / 2$, there holds

$$
\begin{equation*}
\left|u_{t t}^{\varepsilon}\right|+\left|u_{x x x x}^{\varepsilon}\right|+\left|u_{x x t}^{\varepsilon}\right| \leqslant C t^{-3 / 2}+C \varepsilon^{-1} t^{-1+\eta}, \tag{5.30}
\end{equation*}
$$

where $C=C_{\eta}$ is independent of $\varepsilon$.
Proof. First, we shall use the same cut-off function $\zeta(t)$ as defined in (5.20)-(5.22) and multiply it to $u_{t}^{\varepsilon}$; then $\zeta u_{t}^{\varepsilon}$ satisfies

$$
\begin{equation*}
L_{0}\left[\zeta u_{t}^{\varepsilon}\right]=\zeta \cdot f_{t}^{\varepsilon}+\zeta^{\prime} \cdot u_{t}^{\varepsilon}:=\widetilde{F}^{\varepsilon} . \tag{5.31}
\end{equation*}
$$

From (5.13), (5.6), and Lemma 5.6 we find that

$$
\begin{gather*}
\left\|f_{t}^{\varepsilon}\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[\tau / 2, T]\right)} \leqslant\left(\left\|\beta_{\varepsilon}^{\prime}\left(g-u^{\varepsilon}\right)\right\|_{L^{\infty}}+\lambda\right)\left\|u_{t}^{\varepsilon}\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[\tau / 2, T]\right)} \\
\left.\quad+\left\|\beta_{\varepsilon}^{\prime}\left(g-u^{\varepsilon}\right)\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[\tau / 2, T]\right)}\right) u_{t}^{\varepsilon} \|_{L^{\infty}\left(\mathbb{R}^{1} \times[\tau / 2, T]\right)} \\
\leqslant C \varepsilon^{-1}\left\{\tau^{-(1+\alpha) / 2}+(\varepsilon \sqrt{\tau})^{-\alpha / 2}\right\} \\
+C \tau^{-1 / 2}\left\|\beta_{\varepsilon}^{\prime}\left(g-u^{\varepsilon}\right)\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[\tau / 2, T]\right)} \\
\leqslant C \varepsilon^{-1-\alpha / 2} \tau^{-\alpha / 4}+C \varepsilon^{-1} \tau^{-(1+\alpha) / 2} \\
+C \tau^{-1 / 2}\left\|\beta_{\varepsilon}^{\prime}\left(g-u^{\varepsilon}\right)\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[\tau / 2, T]\right)} . \tag{5.32}
\end{gather*}
$$

To estimate $\left\|\beta_{\varepsilon}^{\prime}\left(g-u^{\varepsilon}\right)\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[\tau, T]\right)}$, we note that $\psi(x, t):=\beta_{\varepsilon}^{\prime}\left(g-u^{\varepsilon}\right)$ satisfies

$$
\psi_{x}=\beta_{\varepsilon}^{\prime \prime}\left(g-u^{\varepsilon}\right) \cdot\left(g_{x}-u_{x}^{\varepsilon}\right), \quad \psi_{t}=-\beta_{\varepsilon}^{\prime \prime}\left(g-u^{\varepsilon}\right) \cdot u_{t}^{\varepsilon},
$$

so that the same techniques in the proof of Lemma 5.5 can be used to obtain

$$
\left|\psi_{x}\right| \leqslant C \varepsilon^{-2}\left(\varepsilon^{1 / 2} t^{-1 / 4}\right) \leqslant C \varepsilon^{-3 / 2} t^{-1 / 4}, \quad\left|\psi_{t}\right| \leqslant C \varepsilon^{-2} t^{-1 / 2} .
$$

By applying interpolation separately in $x$ and $t$ (i.e., fixing $x$ or $t$ ), respectively, we obtain

$$
\begin{aligned}
\|\psi(\cdot, t)\|_{C_{x}^{\alpha}\left(\mathbb{R}^{1}\right)} & \leqslant C\|\psi(\cdot, t)\|_{W_{x}^{1, \infty}\left(\mathbb{R}^{1}\right)}^{\alpha}\|\psi\|_{L^{\infty}}^{1-\alpha} \\
& \leqslant C\left(\varepsilon^{-3 \alpha / 2} t^{-\alpha / 4}\right) \varepsilon^{-(1-\alpha)} \leqslant C \varepsilon^{-1-\alpha / 2} t^{-\alpha / 4}
\end{aligned}
$$

and

$$
\begin{aligned}
\|\psi(x, \cdot)\|_{C_{t}^{\alpha / 2}([\tau / 2, T])} & \leqslant C\|\psi(x, \cdot)\|_{W_{t}^{1, \infty}([\tau / 2, T])}^{\alpha / 2}\|\psi\|_{L^{\infty}}^{1-\alpha / 2} \\
& \leqslant C\left(\varepsilon^{-\alpha} \tau^{-\alpha / 4}\right) \varepsilon^{-(1-\alpha / 2)} \leqslant C \varepsilon^{-1-\alpha / 2} \tau^{-\alpha / 4}
\end{aligned}
$$

so that

$$
\begin{equation*}
\left\|\beta_{\varepsilon}^{\prime}\left(g-u^{\varepsilon}\right)\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[\tau, T]\right)} \leqslant C \varepsilon^{-1-\alpha / 2} \tau^{-\alpha / 4} . \tag{5.33}
\end{equation*}
$$

Substituting these estimates into (5.32), we obtain

$$
\begin{align*}
& \left\|f_{t}^{\varepsilon}\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[\tau / 2, T]\right)} \\
& \quad \leqslant C \varepsilon^{-1-\alpha / 2} \tau^{-\alpha / 4}+C \varepsilon^{-1} \tau^{-(1+\alpha) / 2}+C \varepsilon^{-1-\alpha / 2} \tau^{-1 / 2-\alpha / 4} \\
& \quad \leqslant C \varepsilon^{-1} \tau^{-(1+\alpha) / 2}+C \varepsilon^{-(2+\alpha) / 2} \tau^{-(2+\alpha) / 4} \tag{5.34}
\end{align*}
$$

Since, by Hölder's inequality,

$$
\begin{align*}
\varepsilon^{-1} \tau^{-(1+\alpha) / 2} & =(\varepsilon \sqrt{\tau})^{-1} \tau^{-\alpha / 2} \\
& \leqslant \frac{2}{2+\alpha}\left(\varepsilon^{-1} \tau^{-1 / 2}\right)^{(2+\alpha) / 2}+\frac{\alpha}{2+\alpha}\left(\tau^{-\alpha / 2}\right)^{(2+\alpha) / \alpha} \\
& \leqslant \varepsilon^{-(2+\alpha) / 2} \tau^{-(2+\alpha) / 4}+\tau^{-(2+\alpha) / 2} \tag{5.35}
\end{align*}
$$

the estimate (5.34) implies that

$$
\begin{equation*}
\left\|f_{t}^{\varepsilon}\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[\tau / 2, T]\right)} \leqslant C \varepsilon^{-(2+\alpha) / 2} \tau^{-(2+\alpha) / 4}+C \tau^{-(2+\alpha) / 2} \tag{5.36}
\end{equation*}
$$

A similar argument as in the previous lemma shows (using (5.10) and (5.36))

$$
\begin{align*}
\| \zeta & \cdot f_{t}^{\varepsilon} \|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[0, T]\right)} \\
& \leqslant\|\zeta\|_{L^{\infty}}\left\|f_{t}^{\varepsilon}\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[\tau / 2, T]\right)}+\|\zeta\|_{C^{\alpha / 2}([\tau / 2, T])}\left\|f_{t}^{\varepsilon}\right\|_{L^{\infty}} \\
& \leqslant C\left\{\varepsilon^{-(2+\alpha) / 2} \tau^{-(2+\alpha) / 4}+\tau^{-(2+\alpha) / 2}\right\}+C \tau^{-\alpha / 2}\left(\varepsilon^{-1} \tau^{-1 / 2}\right) \\
& \leqslant C\left\{(\varepsilon \sqrt{\tau})^{-(2+\alpha) / 2}+\tau^{-(2+\alpha) / 2}\right\} \tag{5.37}
\end{align*}
$$

where (5.35) is again used in deriving the last inequality. From (5.3), Lemma 5.6, and (5.6), we derive

$$
\begin{align*}
& \left\|\zeta^{\prime} \cdot u_{t}^{\varepsilon}\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[0, T]\right)} \\
& \quad \leqslant\left\|\zeta^{\prime}\right\|_{L^{\infty}[\tau / 2, \tau]}\left\|u_{t}^{\varepsilon}\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[\tau / 2, \tau]\right)}+\left\|\zeta^{\prime}\right\|_{C^{\alpha / 2}([\tau / 2, \tau])}\left\|u_{t}^{\varepsilon}\right\|_{L^{\infty}[\tau / 2, \tau]} \\
& \quad \leqslant C \tau^{-1}\left\{\tau^{-(1+\alpha) / 2}+(\varepsilon \sqrt{\tau})^{-\alpha / 2}\right\}+C \tau^{-1-\alpha / 2} \tau^{-1 / 2} \\
& \quad \leqslant C \tau^{-(3+\alpha) / 2}+C \tau^{-1}(\varepsilon \sqrt{\tau})^{-\alpha / 2} \tag{5.38}
\end{align*}
$$

By Hölder's inequality again,

$$
\begin{align*}
\tau^{-1}(\varepsilon \sqrt{\tau})^{-\alpha / 2} & \leqslant \frac{2}{2+\alpha}\left(\tau^{-1}\right)^{(2+\alpha) / 2}+\frac{\alpha}{2+\alpha}\left((\varepsilon \sqrt{\tau})^{-\alpha / 2}\right)^{(2+\alpha) / \alpha} \\
& \leqslant \tau^{-(2+\alpha) / 2}+(\varepsilon \sqrt{\tau})^{-(2+\alpha) / 2} \tag{5.39}
\end{align*}
$$

Combining (5.37)-(5.39), we obtain

$$
\begin{equation*}
\left\|\widetilde{F}^{\varepsilon}\right\|_{C^{\alpha, \alpha / 2}\left(\mathbb{R}^{1} \times[0, T]\right)} \leqslant C \tau^{-(3+\alpha) / 2}+C(\varepsilon \sqrt{\tau})^{-(2+\alpha) / 2} \tag{5.40}
\end{equation*}
$$

Applying the Schauder estimate (cf. [31, Theorem 4.9, p. 59]) to $\zeta \cdot u_{t}^{\mathcal{\varepsilon}}$ on (5.31), we obtain

$$
\begin{align*}
\left\|u_{t}^{\varepsilon}\right\|_{C^{2+\alpha,(2+\alpha) / 2}\left(\mathbb{R}^{1} \times[\tau, T]\right)} & \leqslant\left\|\zeta \cdot u_{t}^{\varepsilon}\right\|_{C^{2+\alpha,(2+\alpha) / 2}\left(\mathbb{R}^{1} \times[0, T]\right)} \\
& \leqslant C \tau^{-(3+\alpha) / 2}+C(\varepsilon \sqrt{\tau})^{-(2+\alpha) / 2} . \tag{5.41}
\end{align*}
$$

By the interpolation inequality (Lemma 5.1) again,

$$
\begin{aligned}
\left\|u_{t}^{\varepsilon}\right\|_{W_{\infty}^{2,1}\left(\mathbb{R}^{1} \times[\tau, T]\right)} & \leqslant C\left\|u_{t}^{\varepsilon}\right\|_{C^{2+\alpha(2+\alpha) / 2}\left(\mathbb{R}^{1} \times[\tau, T]\right)}^{2 /(2+\alpha)}\left\|u_{t}^{\varepsilon}\right\|_{L^{\infty}\left(\mathbb{R}^{1} \times[\tau, T]\right)}^{\alpha /(2+\alpha)} \\
& \leqslant C\left(\tau^{-(3+\alpha) /(2+\alpha)}+\varepsilon^{-1} \tau^{-1 / 2}\right) \tau^{-\alpha /[2(2+\alpha)]} \\
& \leqslant C \tau^{-3 / 2}+C \varepsilon^{-1} \tau^{-1 / 2-\alpha /[2(2+\alpha)] .} .
\end{aligned}
$$

This establishes (5.30) for $u_{t t}^{\varepsilon}$ and $u_{x x t}^{\varepsilon}$ if we take $\alpha$ such that $\eta=1 /(2+\alpha)$. Since $\alpha$ can be any value in $(0,1)$, we have $1 / 3<\eta<1 / 2$.

To establish the bounds for $u_{x x x x}^{\varepsilon}$, we can first differentiate the equation in $x$ to get the estimate for $u_{x x x}^{\varepsilon}$ in terms of the quantities already estimated. Then we again differentiate the equation in $x$ and solve $u_{x x x x}^{\varepsilon}$ in terms of the other quantities, and then we use (5.12).
6. Estimate of truncation error for regularized solutions. To avoid possible difficulties near $t=0$, we let $w^{\varepsilon}(x, t)=u^{\varepsilon}(x, t+3 h)$. For $w^{\varepsilon}(x, t)$, we have the estimate of truncation error between differential operator $L$ and finite difference operator $L_{h}$ as follows.

Lemma 6.1 (truncation error estimate).

$$
\begin{equation*}
\left|L_{h}\left[w^{\varepsilon}\right]-L\left[w^{\varepsilon}\right]\right|_{t \geqslant \tau} \leqslant C\left((\tau+2 h)^{-3 / 2}+\varepsilon^{-1}(\tau+2 h)^{-1+\eta}\right) h . \tag{6.1}
\end{equation*}
$$

Proof. By the Taylor expansion, for any smooth function $u$,

$$
\begin{aligned}
& \frac{u(x+\sigma k, t-h)+u(x-\sigma k, t-h)-2 u(x, t-h)}{(\sigma k)^{2}}-u_{x x}(x, t-h) \\
& =\frac{1}{6(\sigma k)^{2}}\left\{\int_{0}^{\sigma k} u_{x x x x}(x+\xi, t-h)(\sigma k-\xi)^{3} d \xi\right. \\
& \\
& \left.\quad+\int_{-\sigma k}^{0} u_{x x x x}(x+\xi, t-h)(\sigma k+\xi)^{3} d \xi\right\}
\end{aligned}
$$

so that, for $t \geqslant \tau+h$,

$$
\begin{aligned}
& \left|\frac{u(x+\sigma k, t-h)+u(x-\sigma k, t-h)-2 u(x, t-h)}{(\sigma k)^{2}}-u_{x x}(x, t)\right| \\
& \quad \leqslant\left|u_{x x}(x, t-h)-u_{x x}(x, t)\right|+C k^{2}\left\|u_{x x x x}\right\|_{L^{\infty}} \\
& \quad \leqslant h\left\|u_{x x t}\right\|_{L^{\infty}\left(\mathbb{R}^{1} \times[\tau, T]\right)}+C k^{2}\left\|u_{x x x x}\right\|_{L^{\infty}\left(\mathbb{R}^{1} \times[\tau, T]\right)} \\
& \quad \leqslant C\left(k^{2}+h\right),
\end{aligned}
$$

where the last inequality is from Lemma 5.7. Other derivative terms in the finite difference operator $L_{h}$ can be similarly estimated.

For the integral term of $L_{h}$, by the definition of $p_{l}$ in (3.3),

$$
\begin{aligned}
&\left|\sum_{l \in \mathcal{L}} u(x+l \sigma k, t-h) p_{l}-\int_{\mathbb{R}^{1}} u(x+y, t) d N(y)\right| \\
&=\left|\sum_{l \in \mathcal{L}} u(x+l \sigma k, t-h) \int_{\left(l-\frac{1}{2}\right) \sigma k}^{\left(l+\frac{1}{2}\right) \sigma k} d N(y)-\sum_{l \in \mathcal{L}} \int_{\left(l-\frac{1}{2}\right) \sigma k}^{\left(l+\frac{1}{2}\right) \sigma k} u(x+y, t) d N(y)\right| \\
& \leqslant \sum_{l \in \mathcal{L}} \int_{\left(l-\frac{1}{2}\right) \sigma k}^{\left(l+\frac{1}{2}\right) \sigma k}(|u(x+y, t-h)-u(x+y, t)| \\
&\quad+|u(x+l \sigma k, t-h)-u(x+y, t-h)|) d N(y) \\
& \leqslant C\left(\left\|u_{x x}\right\|_{L^{\infty}\left(\mathbb{R}^{1} \times[\tau, T]\right)} k^{2}+\left\|u_{t}\right\|_{L^{\infty}\left(\mathbb{R}^{1} \times[\tau, T]\right)} h\right)
\end{aligned}
$$

where Taylor's expansion of $u(x+y)$ up to second order around the midpoint is used.
By Lemma 5.7 and Sobolev imbedding theorems (see Chapter 5 of [1]), recalling that we take $k=\sqrt{h}$, for any $\tau \geqslant 0$,

$$
\begin{aligned}
\left|L_{h}\left[w^{\varepsilon}\right]-L\left[w^{\varepsilon}\right]\right|_{t \geqslant \tau} & \leqslant C\left\|w^{\varepsilon}\right\|_{C^{4,2}\left(\mathbb{R}^{1} \times[\tau, T-3 h]\right)} h \leqslant C\left\|u^{\varepsilon}\right\|_{C^{4,2}\left(\mathbb{R}^{1} \times[\tau+3 h, T]\right)} h \\
& \leqslant C\left((\tau+2 h)^{-3 / 2}+\varepsilon^{-1}(\tau+2 h)^{-1+\eta}\right) h .
\end{aligned}
$$

7. Optimal convergence rate of the BTM. Our main theorem is the following.

Theorem 7.1 (main theorem 1). Under the assumption (3.4), the solution of the problem (3.1), (3.2) is convergent to the solution of the problem (2.3), (2.4) with error rate estimates of $O\left(h^{1 / 2}\right)$.

In order to establish this theorem, we shall apply the discrete comparison lemma presented in section 3 to the following auxiliary functions:

$$
\begin{align*}
\Phi_{h}^{ \pm}(x, t)=w^{\varepsilon} \pm & {\left[C \varepsilon^{-1} h \int_{-h}^{t} \frac{d \xi}{(\xi+2 h)^{1-\eta}}\right.} \\
& \left.+C h \int_{-h}^{t} \frac{d \xi}{(\xi+2 h)^{3 / 2}}+\hat{C} \varepsilon+\hat{C} \sqrt{h}\right] \tag{7.1}
\end{align*}
$$

where the constant $C$ is defined in (6.1) and the constant $\hat{C}$ will be defined below.
Lemma 7.2. Let $u_{h}$ be the solution of the problem (3.1)-(3.2) and $\Phi_{h}^{ \pm}(x, t)$ be defined in (7.1); then we have

$$
\begin{equation*}
\Phi_{h}^{-}(x, t) \leqslant u_{h}(x, t) \leqslant \Phi_{h}^{+}(x, t), \quad \mathbb{R}^{1} \times[0, T] \tag{7.2}
\end{equation*}
$$

Proof. First, we claim

$$
\begin{equation*}
F_{h}\left[\Phi^{-}\right] \leqslant 0 \leqslant F_{h}\left[\Phi^{+}\right], \quad \mathbb{R}^{1} \times[0, T] \tag{7.3}
\end{equation*}
$$

In fact, from truncation error estimate (6.1), a direct computation shows that, for $t \geqslant 0$,

$$
\begin{aligned}
& L_{h}\left[C \varepsilon^{-1} h \int_{-h}^{t} \frac{d \xi}{(\xi+2 h)^{1-\eta}}+C h \int_{-h}^{t} \frac{d \xi}{(\xi+2 h)^{3 / 2}}\right] \\
& \quad \geqslant C \varepsilon^{-1} h(t+2 h)^{-1+\eta}+C h(t+2 h)^{-3 / 2} \\
& \quad \geqslant\left|L_{h}\left[w^{\varepsilon}\right]-L\left[w^{\varepsilon}\right]\right|
\end{aligned}
$$

which implies that, for $t \geqslant 0$,

$$
\begin{equation*}
L\left[w^{\varepsilon}\right] \leqslant L_{h}\left[w^{\varepsilon}+C \varepsilon^{-1} h \int_{-h}^{t} \frac{d \xi}{(\xi+2 h)^{1-\eta}}+C h \int_{-h}^{t} \frac{d \xi}{(\xi+2 h)^{3 / 2}}\right] \tag{7.4}
\end{equation*}
$$

and

$$
\begin{equation*}
L\left[w^{\varepsilon}\right] \geqslant L_{h}\left[w^{\varepsilon}-C \varepsilon^{-1} h \int_{-h}^{t} \frac{d \xi}{(\xi+2 h)^{1-\eta}}-C h \int_{-h}^{t} \frac{d \xi}{(\xi+2 h)^{3 / 2}}\right] . \tag{7.5}
\end{equation*}
$$

The estimates (7.4) and (7.5) can be rewritten as follows:

$$
\begin{equation*}
L\left[w^{\varepsilon}\right] \leqslant L_{h}\left[\Phi_{h}^{+}-\hat{C} \varepsilon-\hat{C} \sqrt{h}\right]=L_{h}\left[\Phi_{h}^{+}\right]-\hat{C} \varepsilon-\hat{C} \sqrt{h} \tag{7.6}
\end{equation*}
$$

and

$$
\begin{equation*}
L\left[w^{\varepsilon}\right] \geqslant L_{h}\left[\Phi_{h}^{-}+\hat{C} \varepsilon+\hat{C} \sqrt{h}\right]=L_{h}\left[\Phi_{h}^{-}\right]+\hat{C} \varepsilon+\hat{C} \sqrt{h} \tag{7.7}
\end{equation*}
$$

so that

$$
\begin{align*}
& F\left[w^{\varepsilon}\right]+\hat{C} \varepsilon+\hat{C} \sqrt{h} \\
& \quad \leqslant \min \left\{L_{h}\left[\Phi_{h}^{+}\right], w^{\varepsilon}-g+\hat{C} \varepsilon+\hat{C} \sqrt{h}\right\} \\
& \quad \leqslant \min \left\{L_{h}\left[\Phi_{h}^{+}\right], \Phi_{h}^{+}-g\right\}=F_{h}\left[\Phi_{h}^{+}\right], \quad t>0 \tag{7.8}
\end{align*}
$$

and

$$
\begin{align*}
& F\left[w^{\varepsilon}\right]-\hat{C} \varepsilon-\hat{C} \sqrt{h} \\
& \quad \geqslant \min \left\{L_{h}\left[\Phi_{h}^{-}\right], w^{\varepsilon}-g-\hat{C} \varepsilon-\hat{C} \sqrt{h}\right\} \\
& \quad \geqslant \min \left\{L_{h}\left[\Phi_{h}^{-}\right], \Phi_{h}^{-}-g\right\}=F_{h}\left[\Phi_{h}^{-}\right], \quad t>0 . \tag{7.9}
\end{align*}
$$

Hence, by Lemma 4.6, for $\hat{C}>r+1$,

$$
\begin{aligned}
F_{h}\left[\Phi_{h}^{-}\right] & \leqslant F\left[w^{\varepsilon}\right]-\hat{C} \varepsilon-\hat{C} \sqrt{h} \\
& \leqslant F\left[w^{\varepsilon}\right] \leqslant 0 \leqslant F\left[w^{\varepsilon}\right]+\hat{C} \varepsilon+\hat{C} \sqrt{h} \leqslant F_{h}\left[\Phi_{h}^{+}\right], \quad t>0 .
\end{aligned}
$$

Thus (7.3) is proved.

Next, we verify the following inequalities in the strip $\{-h<t \leqslant 0, x \in \mathbb{R}\}$ :

$$
\Phi_{h}^{-}(x, t) \leqslant g(x) \leqslant \Phi_{h}^{+}(x, t)
$$

For $-h<t \leqslant 0$, we use Lemma 4.4 to derive, for $\hat{C}$ sufficiently large,

$$
\begin{aligned}
\Phi_{h}^{+}(x, t) & \geqslant g(x)+\hat{C} \varepsilon+\hat{C} \sqrt{h}+u^{\varepsilon}(x, t+3 h)-g(x) \\
& \geqslant g(x)+\hat{C} \sqrt{h}-\left[u^{\varepsilon}\right]_{C_{t}^{1 / 2}\left(\mathbb{R}^{1} \times[0, T]\right)} \sqrt{3 h} \\
& \geqslant g(x), \quad-h<t \leqslant 0 .
\end{aligned}
$$

In a similar way,

$$
\Phi_{h}^{-}(x, t) \leqslant g(x), \quad-h<t \leqslant 0
$$

Because $u_{h}$ is the solution of the problem (3.1), (3.2), we have

$$
F_{h}\left[u_{h}\right]=0, \quad u_{h}(x, 0)=g(x)
$$

Using the comparison principle for finite difference operators (Lemma 3.2), we derive

$$
\Phi_{h}^{+}(x, t) \geqslant u_{h}(x, t) \geqslant \Phi_{h}^{-}(x, t)
$$

everywhere.
Lemma 7.3 (main lemma). Let $u^{\varepsilon}$ be the solution of the problem (2.13), (2.14) and $u_{h}$ be the solution of the problem (3.1), (3.2). Then

$$
\left|u_{h}-u^{\varepsilon}\right| \leqslant C h \varepsilon^{-1}+C \varepsilon+C \sqrt{h}, \quad(x, t) \in \mathbb{R}^{1} \times[0, T],
$$

where $C$ is independent of $h$ and $\varepsilon$.
Proof. Combining Lemmas 6.1 and 7.3, we obtain, for $t>0$,

$$
\begin{aligned}
& \left|u^{\varepsilon}(x, t+3 h)-u_{h}(x, t)\right| \\
& \quad \leqslant C \varepsilon^{-1} h \int_{-h}^{t} \frac{d \xi}{(\xi+2 h)^{1-\eta}}+C h \int_{-h}^{t} \frac{d \xi}{(\xi+2 h)^{3 / 2}}+\hat{C} \varepsilon+\hat{C} \sqrt{h} \\
& \quad \leqslant C \varepsilon^{-1} h+\hat{C} \varepsilon+\hat{C} \sqrt{h} .
\end{aligned}
$$

Now using the $C^{1 / 2}$ regularity (uniformly in $\varepsilon$ and $h$, Lemma 5.3), we get

$$
\left|u^{\varepsilon}(x, t+3 h)-u^{\varepsilon}(x, t)\right| \leqslant\left[u^{\varepsilon}\right]_{C_{t}^{1 / 2}\left(\mathbb{R}^{1} \times[0, T]\right)} \sqrt{3 h} \leqslant C \sqrt{h} .
$$

Combining the above two estimates, we conclude our lemma.
Proof of Theorem 7.1. Combining all the estimates, we get

$$
\begin{aligned}
\left|u-u_{h}\right| & \leqslant\left|u-u^{\varepsilon}\right|+\left|u^{\varepsilon}-u_{h}\right| \\
& <C \varepsilon+\left\{C h \varepsilon^{-1}+C \varepsilon+C \sqrt{h}\right\},
\end{aligned}
$$

where $C$ is independent of $\varepsilon$ and $h$. Upon taking $\varepsilon=\sqrt{h}$, we derive the conclusion of the main theorem. That is, we have

$$
\begin{equation*}
\left|u-u_{h}\right|<C_{0} h^{1 / 2} . \tag{7.10}
\end{equation*}
$$

This completes the proof of this theorem.
The $O\left(h^{1 / 2}\right)$ rate is the optimal convergence rate. To show this, we will give a simple example in Remark 9.1 by comparing the BTM solution with the exact solution of the American option pricing model. It can clearly be verified that the error rate of the approximation process by the BTM cannot be better than $O\left(h^{1 / 2}\right)$.
8. Convergence rate of the free boundary. In this section, we shall define a free boundary from the EFDS and prove the convergence rate of this free boundary to the actual free boundary.

Definition 8.1. For any $h>0$, let $u_{h}$ be the solution of the problem (3.1), (3.2). We define

$$
\begin{equation*}
s_{h}(t)=\min \left(0, s_{h}^{*}(t)\right), \tag{8.1}
\end{equation*}
$$

where $s_{h}^{*}(t)$ is a solution to the equation

$$
\begin{equation*}
u_{h}\left(s_{h}^{*}(t), t\right)-g\left(s_{h}^{*}(t)\right)=C_{0} h^{1 / 2} \tag{8.2}
\end{equation*}
$$

where $C_{0}$ is defined in (7.10), and $g(\cdot)$ is defined in (2.6).
Remark 8.1. For every $t>0, u_{h}(-\infty, t)-g(-\infty)=0$, and $u_{h}(+\infty, t)-g(+\infty)>0$. Thus for sufficiently small $h$, we can take $s_{h}^{*}(t)$ to be the smallest number satisfying (8.2). In practice we should use the smallest possible $C_{0}$ such that (7.10) holds, and this $C_{0}$ can actually be estimated in the simulations.

Theorem 8.1 (main theorem 2). Let $s_{h}(t)$ be defined as above and $s(t)$ be the free boundary of the problem (2.3), (2.4). If

$$
\begin{equation*}
\nu:=r-\lambda \int_{0}^{\infty}\left(e^{y}-1\right) d N(y)>0 \tag{8.3}
\end{equation*}
$$

then

$$
\begin{equation*}
\left|s_{h}(t)-s(t)\right| \leqslant C h^{1 / 4} \tag{8.4}
\end{equation*}
$$

where $C$ is some positive constant.
The rest of this section is devoted to the proof of this theorem. We begin with a lemma.
Lemma 8.2. Under the assumption (8.3), we have

$$
\begin{equation*}
\left(u-1+e^{x}\right)_{x x} \geqslant \min \left(1,2 \nu / \sigma^{2}\right) e^{-(r+\lambda) T} \quad \text { for } s(t)<x<\infty, \quad 0<t<T . \tag{8.5}
\end{equation*}
$$

Proof. It is known from Lemma 2.4 (cf. [37]) that for $t>0, u, u_{x}$, and $u_{t}$ are continuous across the free boundary but $u_{x x}$ has a jump.

Since $u$ is monotonically decreasing in $x$, we have

$$
\begin{aligned}
& u(x, t)=1-e^{x} \quad \text { for } \quad x \leqslant s(t) \\
& u(x, t)<u(s(t), t)=1-e^{s(t)} \text { for } x>s(t) .
\end{aligned}
$$

It follows that

$$
\int_{\mathbb{R}^{1}} u(x+y, t) d N(y) \leqslant 1-\int_{-\infty}^{s(t)-x} e^{x+y} d N(y)-\int_{s(t)-x}^{\infty} e^{s(t)} d N(y) .
$$

In particular, for $x=s(t)$,

$$
\begin{aligned}
\int_{\mathbb{R}^{1}} u(s(t)+y, t) d N(y) & \leqslant 1-e^{s(t)}\left\{\int_{-\infty}^{0} e^{y} d N(y)+\int_{0}^{\infty} d N(y)\right\} \\
& =1-e^{s(t)}\left\{\int_{-\infty}^{\infty} e^{y} d N(y)+\int_{0}^{\infty}\left(1-e^{y}\right) d N(y)\right\} \\
& =1-e^{s(t)}(1+\omega)+e^{s(t)} \int_{0}^{\infty}\left(e^{y}-1\right) d N(y) .
\end{aligned}
$$

Using the equation $L[u]=0$ and taking the limit as $x \rightarrow s(t)+0$, we obtain

$$
\begin{aligned}
\frac{\sigma^{2}}{2} & {\left[u_{x x}(x, t)\right]_{x=s(t)+0} } \\
= & \left\{-\left(r-\lambda \omega-\frac{\sigma^{2}}{2}\right) u_{x}(x, t)\right. \\
& \left.+(r+\lambda) u(x, t)-\lambda \int_{\mathbb{R}^{1}} u(x+y, t) d N(y)\right\}_{x=s(t)+0} \\
\geqslant & \left\{\left(r-\lambda \omega-\frac{\sigma^{2}}{2}\right) e^{s(t)}\right. \\
& \left.+(r+\lambda)\left(1-e^{s(t)}\right)-\lambda\left[1-e^{s(t)}(1+\omega)+e^{s(t)} \int_{0}^{\infty}\left(e^{y}-1\right) d N(y)\right]\right\} \\
= & -\frac{\sigma^{2}}{2} e^{s(t)}+r-\lambda e^{s(t)} \int_{0}^{\infty}\left(e^{y}-1\right) d N(y) \geqslant-\frac{\sigma^{2}}{2} e^{s(t)}+\nu .
\end{aligned}
$$

In deriving the last inequality, we made use of the fact that $s(t) \leqslant 0$ (cf. Lemma 2.4). It follows that the function

$$
\begin{equation*}
\varphi(x, t)=u-1+e^{x} \tag{8.6}
\end{equation*}
$$

satisfies

$$
\begin{equation*}
\varphi_{x x}(s(t)+0, t) \geqslant \frac{2 \nu}{\sigma^{2}} \quad \text { for } t>0 \tag{8.7}
\end{equation*}
$$

Taking the limit in (5.8) as $\varepsilon \rightarrow 0$, we also have

$$
\begin{equation*}
\varphi_{x x}(x, t) \geqslant 0 \quad \text { for } \quad x \in \mathbb{R}^{1}, t>0 \tag{8.8}
\end{equation*}
$$

A direct computation also shows that

$$
\begin{aligned}
& L\left[\varphi_{x x}\right]=0 \quad \text { for } \quad x>s(t), 0<t<T, \\
& \varphi_{x x}(x, 0)=e^{x}>1 \text { for } 0<x<\infty .
\end{aligned}
$$

Notice that the comparison principle is established only for the case where the domain is the whole $\mathbb{R}^{1}$. To work in the domain $\{(x, t) ; x>s(t), 0<t<T\}$, we define an operator $L_{1}$ by

$$
L_{1}[v]=v_{t}-\frac{\sigma^{2}}{2} v_{x x}-\left(r-\lambda \omega-\frac{\sigma^{2}}{2}\right) v_{x}+(r+\lambda) v-\lambda \int_{s(t)-x}^{\infty} v(x+y) d N(y)
$$

In our special case, we have

$$
\varphi(x, t) \equiv 0 \quad \text { for } x<s(t),
$$

which implies that $L\left[\varphi_{x x}\right] \equiv L_{1}\left[\varphi_{x x}\right]$. Although $\varphi_{x x}$ has a singularity at $(x, t)=(0,0), \varphi_{x x}$ is uniformly bounded from below by Corollary 5.4.

To overcome the possible discontinuity of $\varphi_{x x}$ at $(0,0)$, we define the operator $L_{0}$ as $L_{0}[v]=v_{t}-\frac{\sigma^{2}}{2} v_{x x}-\left(r-\lambda \omega-\frac{\sigma^{2}}{2}\right) v_{x}+(r+\lambda) v$. Then, by (8.8),

$$
\begin{aligned}
& L_{0}\left[\varphi_{x x}-\min \left(1,2 \nu / \sigma^{2}\right) e^{-(r+\lambda) t}\right] \geqslant \lambda \int_{s(t)-x}^{\infty} \varphi_{x x}(x+y) d N(y) \geqslant 0, \\
& \varphi_{x x}(s(t)+0,0)-\min \left(1,2 \nu / \sigma^{2}\right) e^{-(r+\lambda) t}>0 \quad \text { for } t>0, \\
& \varphi_{x x}(0, x)-\min \left(1,2 \nu / \sigma^{2}\right) \geqslant 0 \quad \text { for } x>0 .
\end{aligned}
$$

Let $G(x, t)$ be the fundamental solution of $L_{0}$, i.e.,

$$
\begin{aligned}
& L_{0}[G]=0 \text { for }(x, t) \in \mathbb{R}^{1} \times(0, T), \quad \lim _{t \rightarrow 0+} G(x, t)=\delta(x), \\
& G(x, t)>0 \text { for }(x, t) \in \mathbb{R}^{1} \times(0, T), \\
& \sup _{0<t<T} \int_{-\infty}^{\infty} G(x, t) d x<\infty, \\
& \lim _{(x, t) \rightarrow(0,0)} G(x, t)=+\infty .
\end{aligned}
$$

This fundamental solution is given explicitly as

$$
G(x, t)=\frac{1}{\sigma \sqrt{2 \pi t}} \exp \left\{-r t-\frac{\left[x+\left(r-\frac{\sigma^{2}}{2}\right) t\right]^{2}}{2 \sigma^{2} t}\right\}
$$

Take the auxiliary function $\delta\left(c_{0}+x^{2}\right)$ in a way similar to Lemma 2.3 such that $L_{0}\left[c_{0}+x^{2}\right]>0$.
We claim that, for any sufficiently small $\eta>0$,

$$
\begin{equation*}
\varphi_{x x}-\min \left(1,2 \nu / \sigma^{2}\right) e^{-(r+\lambda) t}+\eta G(x, t)+\eta\left(c_{0}+x^{2}\right)>0 \quad \text { for } \quad x>s(t), 0<t<T \tag{8.9}
\end{equation*}
$$

Since $\varphi_{x x}-\min \left(1,2 \nu / \sigma^{2}\right) e^{-(r+\lambda) t}$ is bounded from below while $\eta G(x, t)$ goes to $\infty$ as $(x, t) \rightarrow$ $(0,0),(8.9)$ is obviously valid in a small neighborhood of $(0,0)$. Similarly, since $x^{2}$ goes to $\infty$ as $x \rightarrow \infty$, (8.9) is valid for $|x|>R_{\eta}$ if we choose $R_{\eta} \gg 1$. Thus, by the maximum principle
applied to the domain $\left\{(x, t) ; x>s(t),|x|+t>\eta,|x|<R_{\eta}\right\}$ (this is a bounded domain where $\varphi_{x x}$ is continuous), we conclude that (8.9) holds. Letting $\eta \rightarrow 0$ we obtain (8.5).

Proof of Theorem 8.1. Using (7.10) and (8.2) we find that

$$
\begin{aligned}
C_{0} h^{1 / 2} & =u_{h}\left(s_{h}^{*}(t), t\right)-g\left(s_{h}^{*}(t)\right) \\
& <u\left(s_{h}^{*}(t), t\right)+C_{0} h^{1 / 2}-g\left(s_{h}^{*}(t)\right) .
\end{aligned}
$$

This implies that $u\left(s_{h}^{*}(t), t\right)-g\left(s_{h}^{*}(t)\right)>0$ and hence $s_{h}^{*}(t)>s(t)$. Since we also know that $s(t) \leqslant 0$, we conclude that $s_{h}(t) \geqslant s(t)$.

By the definition of $s_{h}(t)$ we have $s_{h}(t) \leqslant 0$ so that

$$
u\left(s_{h}(t), t\right)-g\left(s_{h}(t)\right)=\varphi\left(s_{h}(t), t\right),
$$

where the function $\varphi$ is defined by (8.6). Thus

$$
\begin{aligned}
2 C_{0} h^{1 / 2} & =C_{0} h^{1 / 2}+u_{h}\left(s_{h}(t), t\right)-g\left(s_{h}(t)\right) \\
& \geqslant u\left(s_{h}(t), t\right)-g\left(s_{h}(t)\right)=\varphi\left(s_{h}(t), t\right) \\
& =\varphi(s(t), t)+\varphi_{x}(s(t), t)\left(s_{h}(t)-s(t)\right)+\frac{1}{2} \varphi_{x x}(\xi, t)\left(s_{h}(t)-s(t)\right)^{2} \\
& \geqslant \frac{1}{2} c_{0}\left(s_{h}(t)-s(t)\right)^{2}, \quad \text { where } c_{0}=\min \left(1, \frac{2 \nu}{\sigma^{2}}\right) e^{-(r+\lambda) T},
\end{aligned}
$$

which implies the conclusion of the theorem.
Remark 8.2. In the particular case of the pure diffusion Black-Scholes model (i.e., when $\lambda=0$ ), assumption (8.3) is automatically satisfied.

Remark 8.3. In Definition 8.1, there is a constant $C_{0}$ which comes from (7.10) in Theorem 7.1. Since this constant comes from different parts of the proof, it is not easy to be obtained explicitly. However, for a particular problem, as $C_{0}$ is required only in the numerical computation of the free boundary and represents the error rate of the numerical computation of the solution $u_{h}$, we can estimate it from the numerical computation of the solution $u_{h}$. In the numerical example in the next section, we will estimate the constant $C_{0}$ and use it to calculate the free boundary.

## 9. Remarks and a numerical example.

Remark 9.1. The European option admits an explicit solution. We can use the FDS on this explicit solution to calculate the convergence rate.

Particularly, when $x=0, t=h$, and $\lambda=\omega=0$, the problem (2.18), (2.19) turns to the Black-Scholes equation problem without jumps and gives

$$
d_{1}=\frac{2 r+\sigma^{2}}{2 \sigma} \sqrt{h}, \quad d_{2}=\frac{2 r-\sigma^{2}}{2 \sigma} \sqrt{h} .
$$

Therefore

$$
\begin{aligned}
U(0, h) & =e^{-r h} N\left(-d_{2}\right)-N\left(-d_{1}\right) \\
& =\frac{1}{\sqrt{2 \pi}} \int_{d_{2}}^{d_{1}} e^{-u^{2} / 2} d u+\left(e^{-r h}-1\right) N\left(-d_{2}\right) \\
& =\frac{1}{\sqrt{2 \pi}} \int_{d_{2}}^{d_{1}}\left(1-\frac{u^{2}}{2}+o\left(u^{2}\right)\right) d u+o(\sqrt{h}) \\
& =\frac{1}{\sqrt{2 \pi}}\left(d_{1}-d_{2}\right)+o(\sqrt{h})=\frac{\sigma}{\sqrt{2 \pi}} \sqrt{h}+o(\sqrt{h}) .
\end{aligned}
$$

On the other hand, by the BTM, from (1.6), it gives

$$
q=\frac{1}{2}+\frac{2 r-\sigma^{2}}{4 \sigma} \sqrt{h}+o(\sqrt{h}), \quad 1-q=\frac{1}{2}-\frac{2 r-\sigma^{2}}{4 \sigma} \sqrt{h}+o(\sqrt{h})
$$

so that

$$
\begin{aligned}
U_{h}(0, h) & =\frac{1}{1+r h}\left[q U_{1}^{0}+(1-q) U_{-1}^{0}\right] \\
& =\frac{1}{1+r h}\left[q \cdot 0+\left(\frac{1}{2}-\frac{2 r-\sigma^{2}}{4 \sigma} \sqrt{h}+o(\sqrt{h})\right)\left(1-e^{-\sigma \sqrt{h}}\right)\right] \\
& =(1-(r h)+o(h))\left[\left(\frac{1}{2}-\frac{2 r-\sigma^{2}}{4 \sigma} \sqrt{h}+o(\sqrt{h})\right)(\sigma \sqrt{h}+o(\sqrt{h}))\right] \\
& =\frac{\sigma}{2} \sqrt{h}+o(\sqrt{h}) .
\end{aligned}
$$

Thus,

$$
U_{h}(0, h)-U(0, h)=\left(\frac{1}{2}-\frac{1}{\sqrt{2 \pi}}\right) \sigma \sqrt{h}+o(\sqrt{h}) .
$$

This means that the FDS on a European option cannot have a better convergence rate than $O(\sqrt{h})$, unless $\sigma \equiv 0$. We conclude that the convergence rate of the FDS on a European option cannot be better than $O(\sqrt{h})$.

Remark 9.2. For an American option price $u$ and a European option $U$, there holds

$$
u-u_{h}=(u-U)+\left(U-u_{h}\right) .
$$

Let us check the error at the point $(0, h)$. First,

$$
\begin{aligned}
& U(0, h)-u_{h}(0, h)=U(0, h)-\min \left\{U_{h}(0, h), 0\right\} \\
= & U(0, h)=\frac{\sigma}{\sqrt{2 \pi}} \sqrt{h}+o(\sqrt{h}) .
\end{aligned}
$$

An American option can be decomposed into a European option and a so-called early exercise premium (see section 6.3 in [19]). In other words, let $V(S, t)$ be the American put option price; then

$$
V(S, t)=\widetilde{U}(S, t)+e(S, t)
$$

where $\widetilde{U}(S, t)$ is the European put option price corresponding to (2.18), (2.19) under the original ( $S, t$ ) variables, and $e(S, t)$ is the early exercise premium,

$$
e(S, t)=\int_{t}^{T} d \eta \int_{0}^{S(\eta)}(K r-q \xi) G(S, t ; \xi, \eta) d \xi,
$$

where $r, K$ are defined as above, $q$ is the dividend rate of the underlying asset $S$, and $G(S, t ; \xi, \eta)$ is the fundamental solution of the Black-Scholes equation.

Under our notation, we have

$$
u(x, t)=U(x, t)+e(x, t),
$$

where $e(x, t)$ is called an early exercise premium, which, in our case, can be written as

$$
e(x, t)=\int_{0}^{t} d \eta \int_{-\infty}^{x(\eta)} r G(x, t ; \xi, \eta) d \xi,
$$

where $x(t)$ is the free boundary of the solution and $G(x, t ; \xi, \eta)$ is the fundamental solution of the lognormal Black-Scholes equation:

$$
G(x, t ; \xi, \eta)=\frac{1}{\sigma \sqrt{2 \pi(t-\eta)}} \exp \left\{-r(t-\eta)-\frac{\left[x-\xi+\left(r-\frac{\sigma^{2}}{2}\right)(t-\eta)\right]^{2}}{2 \sigma^{2}(t-\eta)}\right\} .
$$

Thus

$$
\begin{align*}
|e(x, t)| & \leqslant \int_{0}^{t} d \eta \int_{-\infty}^{x(\eta)} r|G(x, t ; \xi, \eta)| d \xi \\
& \leqslant C \int_{0}^{t} e^{r(\eta-t)} d \eta=\frac{C}{r}\left(1-e^{-r t}\right)=O(t)=o(\sqrt{t}) . \tag{9.1}
\end{align*}
$$

The linear regress of the error $R(\sqrt{h})$ is

$$
R(\sqrt{h})=0.03970255974177 \sqrt{h}+0.00001112311713
$$

So the estimate constant $C_{0}$ in Theorem 7.1 in this case is 0.03970255974177 .
Therefore, at the first step $t=h$, we also have $|e(x, h)|=o(\sqrt{h})$.
That means the difference of $u-u_{h}$ will not be better than $O(\sqrt{h})$.
Remark 9.3. The limitation to the convergence rate comes from the option pricing initial datum $g(x)=\left[1-e^{x}\right]^{+}$, which belongs only to $W_{\infty}^{1}$. For the smooth initial datum, the convergence rate is better. We refer the reader to [27], [28].

Remark 9.4. The estimate $\left|u_{( }(x, t)-u(x, t)\right| \leqslant C \sqrt{h}$ is optimal when $t$ is very close to 0 (as the example in Remark 9.1 shows). If $t$ is fixed, some calculations show that there is a better convergence rate. In particular, say when $t=T$, [26] shows that $\left|u_{h}(x, T)-u(x, T)\right|=$ $O\left((h \sqrt{|\log h|})^{4 / 5}\right)$ for the case of the put option and $O(h \sqrt{|\log h|})$ for more regular payoffs.

Table 1
Comparison of errors at the grid points $(x, t)=(0, \Delta t)$.

| $h=\Delta t$ | 0.01 | $2.5 \times 10^{-3}$ | $6.25 \times 10^{-4}$ | $1.563 \times 10^{-4}$ | $3.906 \times 10^{-5}$ | $9.766 \times 10^{-6}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Error | 0.0039756 | 0.0020071 | 0.0010075 | 0.0005054 | 0.0002550 | 0.0001326 |



Figure 1. Maximum error of the solution vs. length of the numerical step.
Table 2
Comparison of the free boundary errors at fixed time $t=0.1$ when "exact solution" $s(0.1)=-0.1981272$.

| $h=\Delta t$ | $2.5 \times 10^{-3}$ | $6.25 \times 10^{-4}$ | $1.563 \times 10^{-4}$ | $3.906 \times 10^{-5}$ | $9.766 \times 10^{-6}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| $s_{h}(0.1)$ | -0.16 | -0.17 | -0.18 | -0.1875 | -0.1925 |
| Error | 0.0381272 | 0.0281272 | 0.0181272 | 0.0106272 | 0.0056272 |

Calculation example. In the case $r=0.1, \lambda=0.001, \sigma=0.4, N(x)=0$ for $x<0.1$, and $N(x)=1$ for $x>0.1, \omega=\exp (0.1)-1$, Table 1 shows the errors of the solution by use of the scheme discussed in this paper. Here we use the approximation solution when $h=\Delta t=6.103515625 \times 10^{-7}$ as the "exact solution" $s(t)$ for comparing to the others. Since the singular point is at $(0,0)$, we expect the maximum error to appear at the position nearest to $(0,0)$. First, we compare the solution at the first step of the calculation at $(0,0)$. The result is shown in Figure 1.

Now use this $C_{0}$ to calculate the free boundary. Table 2 shows the errors of the free boundary, i.e., the last grid point $x_{h}$ when $u\left(x_{h}\right) \leqslant 1-e^{x_{h}}+0.0397 \sqrt{h}$. The calculation shows

Table 3
Comparison of the free boundary errors at the first step of the scheme.

| $h=\Delta t$ | $2.5 \times 10^{-3}$ | $6.25 \times 10^{-4}$ | $1.563 \times 10^{-4}$ | $3.906 \times 10^{-5}$ | $9.766 \times 10^{-6}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| $s(h)$ | -0.0425023 | -0.0215648 | -0.0103148 | -0.0046898 | -0.0021898 |
| $s_{h}(h)$ | -0.02 | -0.01 | -0.005 | -0.0025 | -0.00125 |
| Error | 0.0225023 | 0.0115648 | 0.0053148 | 0.0021898 | 0.0009398 |

Table 4
Convergence rate of the fixed point and the first step.

| $h=\Delta t$ | $2.5 \times 10^{-3}$ | $6.25 \times 10^{-4}$ | $1.563 \times 10^{-4}$ | $3.906 \times 10^{-5}$ | $9.766 \times 10^{-6}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Rate at $t=0.1$ | - | 0.2194 | 0.3169 | 0.3852 | 0.4586 |
| Rate at the 1st step | - | 0.4802 | 0.5608 | 0.6396 | 0.6102 |

that the error of $s_{h}^{*}$, which is defined in Definition 8.1, is lower than the theoretical estimate $O\left(h^{1 / 4}\right)$, which implies that the $O\left(h^{1 / 4}\right)$ is suboptimal. Again, $h=\Delta t=6.103515625 \times 10^{-7}$ is used as the benchmark.

As our estimate rate result is about "maximum error," as the comparison of the solution error, we also compare the free boundary at the first step of the calculation. Table 3 shows the free boundary error compared to the "exact solution" at the first time step.

Now let us estimate the numerical convergence rate of the free boundary. Table 4 shows the convergence rate $\gamma_{i}$, which is computed from the relation $\operatorname{err}_{i} / \operatorname{err}_{i-1}=h_{i}^{\gamma_{i}} / h_{i}^{\gamma_{i-1}}$, where $h_{i}, i=1, \ldots, 5$, are the values of $h$ from Tables 2 and 3 , and err $r_{i}$ are the corresponding errors. As $h_{i}=h_{i-1} / 4$, we have

$$
\gamma_{i}=\frac{\ln e r r_{i-1}-\ln e r r_{i}}{\ln 4}
$$

Acknowledgment. We are very grateful to the anonymous referees for many helpful suggestions.

## REFERENCES

[1] R. A. Adams, Sobolev Spaces, Academic Press, New York, 1975.
[2] A. Almendral and C. O. Oosterlee, Numerical valuation of options with jumps in the underlying, Appl. Numer. Math., 53 (2005), pp. 1-18.
[3] A. L. Amadori, Obstacle problem for nonlinear integro-differential equations arising in option pricing, Ric. Mat., 56 (2007), pp. 1-17.
[4] K. Amin, Jump diffusion option valuation in discrete time, J. Finan., 48 (1993), pp. 1833-1863.
[5] E. Bayraktar and X. Hao, Analysis of the Optimal Exercise Boundary of American Options for Jump Diffusions, manuscript, 2008.
[6] E. Bayraktar and H. Xing, Analysis of the optimal exercise boundary of American options for jump diffusions, SIAM J. Math. Anal., 41 (2009), pp. 825-860.
[7] F. Black and M. Scholes, The price of options and corporate liabilities, J. Polit. Econ., 81 (1973), pp. 637-654.
[8] M. Briani, C. La Chioma, and R. Natalini, Convergence of numerical schemes for viscosity solutions to integro-differential degenerate parabolic problems arising in financial theory, Numer. Math., 98 (2004), pp. 607-646.
[9] Y.-Z. Chen and L.-C. Wu, Second Order Elliptic Equations and Elliptic Systems, Transl. Math. Monogr. 174, AMS, Providence, RI, 1998.
[10] R. Cont and E. Voltchkova, A finite difference scheme for option pricing in jump diffusion and exponential Lévy models, SIAM J. Numer. Anal., 43 (2005), pp. 1596-1626.
[11] J. Cox and M. Rubinstein, Option pricing: A simplified approach, J. Finan. Econ., 7 (1979), pp. 229263.
[12] A. Friedman, Variational Principles and Free Boundary Problems, John Wiley \& Sons, New York, 1982.
[13] A. Friedman, Partial Differential Equations of Parabolic Type, Prentice-Hall, Englewood Cliffs, NJ, 1964.
[14] Y. d'Halluin, P. A. Forsyth, and G. Labahn, Penalty method for American options with jump diffusion processes, Numer. Math., 97 (2004), pp. 321-352.
[15] B. Hu, L. Jiang, and J. Liang, Optimal rate of convergence of the binomial tree scheme for American options, J. Comput. Appl. Math., 230 (2009), pp. 583-599.
[16] J. Hull, Options, Futures and Other Derivatives, 7th ed., Prentice-Hall, Upper Saddle River, NJ, 2008.
[17] E. R. Jakobsen, Error bounds for monotone approximation schemes for parabolic Hamilton-JacobiBellman equations, Math. Models Methods Appl. Sci., 13 (2003), pp. 613-644.
[18] E. R. Jakobsen, K. H. Karlsen, and C. La Chioma, Error estimates for approximate solutions to Bellman equations associated with controlled jump-diffusions, Numer. Math., 110 (2008), pp. 221-255.
[19] L. Jiang, Mathematical Modeling and Methods for Option Pricing, World Scientific, River Edge, NJ, 2005.
[20] L. Jiang and M. Dai, Convergence of the explicit difference scheme and the binomial tree method for American options, J. Comput. Math., 22 (2004), pp. 371-380.
[21] L. Jiang and M. Dai, Convergence of binomial tree methods for European/American options pathdependent options, SIAM J. Numer. Anal., 42 (2004), pp. 1094-1109.
[22] S. G. Kou, A jump-diffusion model for option pricing, Management Sci., 48 (2002), pp. 1086-1101.
[23] N. V. Krylov, On the rate of convergence of finite-difference approximations for Bellman's equations, St. Petersburg Math. J., 9 (1998), pp. 639-650.
[24] N. V. Krylov, On the rate of convergence of finite-difference approximations for Bellman's equations with variable coefficients, Probab. Theory Related Fields, 117 (2000), pp. 1-16.
[25] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, Linear and Quasilinear Equations of Parabolic Type, Nauka, Moscow, 1968.
[26] D. Lamberton, Error estimates for the binomial approximation of American put options, Ann. Appl. Probab., 8 (1998), pp. 206-233.
[27] D. Lamberton and C. Rogers, Optimal stopping and embedding, J. Appl. Probab., 4 (2000), pp. 11431148.
[28] D. Lamberton, Brownian optional stopping and random walks, Appl. Math. Optim., 45 (2002), pp. 283324.
[29] J. Liang, B. Hu, L. Jiang, and B. Bian, On the rate of convergence of the binomial tree scheme for American options, Numer. Math., 107 (2007), pp. 333-352.
[30] J. LiAng, On the convergence rate of the binomial tree scheme for an American option with jump-diffusion, Numerical Math. J. Chinese Univ., 30 (2008), pp. 76-96.
[31] G. M. Lieberman, Second Order Parabolic Differential Equations, World Scientific, River Edge, NJ, 1996.
[32] R. Merton, Option pricing when the underlying stock returns are discontinuous, J. Finan. Econ., 3 (1976), pp. 125-144.
[33] H. Pham, Optimal stopping, free boundary and American option in a jump-diffusion model, Appl. Math. Optim., 35 (1995), pp. 145-164.
[34] X.-S. Qian, C.-L. Xu, L.-S. Jiang, and B.-J. Bian, Convergence of the binomial tree method for American options in a jump-diffusion model, SIAM J. Numer. Anal., 42 (2005), pp. 1899-1913.
[35] W. Wilmott, S. Howison, and J. Dewyne, The Mathematics of Financial Derivatives, Cambridge University Press, Cambridge, UK, 1995.
[36] C. Xu, X. Qian, and L. Jiang, Numerical analysis on binomial tree methods for a jump-diffusion model, J. Comput. Appl. Math., 156 (2003), pp. 23-45.
[37] C. Yang, L. Jiang, and B. Bian, Free boundary and American options in a jump-diffusion model, European J. Appl. Math., 17 (2006), pp. 95-127.
[38] X. Zhang, Analyse Numérique des Options Américaines dans um Modéle de Diffusion avec des Sauts, Doctoral dissertation, École Nationale des Ponts et Chaussées, Paris, 1994.
[39] X. Zhang, Numerical analysis of American option pricing in a jump-diffusion model, Math. Oper. Res., 22 (1997), pp. 668-690.

# Duality for Set-Valued Measures of Risk* 

Andreas H. Hamel ${ }^{\dagger}$ and Frank Heyde ${ }^{\ddagger}$


#### Abstract

Extending the approach of Jouini, Meddeb, and Touzi [Finance Stoch., 8 (2004), pp. 531-552] we define set-valued (convex) measures of risk and their acceptance sets, and we give dual representation theorems. A scalarization concept is introduced that has a meaning in terms of internal prices of portfolios of reference instruments. Using primal and dual descriptions, we introduce new examples for set-valued measures of risk, e.g., set-valued upper expectations, value at risk, average value at risk, and entropic risk measure.


Key words. set-valued risk measures, coherent risk measures, Legendre-Fenchel transform, convex duality, biconjugation, value at risk, scalarization

AMS subject classifications. 91B30, 46A20, 46N10, 26E25
DOI. 10.1137/080743494

1. Introduction. The concept of set-valued coherent measures of risk has been introduced recently by Jouini, Meddeb, and Touzi [16]. The basic question is how to evaluate the (financial) risk of a multivariate random outcome in terms of more than one reference instrument, for example if the regulator accepts deposits in more than one currency. This is of particular importance if transaction costs have to be paid for each transaction between assets including the reference instruments.

This question became unexpectedly topical as the European Central Bank decreed [7] that temporarily, until the end of 2009, "the list of assets eligible as collateral in Eurosystem credit operations will be expanded" by "marketable debt instruments denominated in other currencies than the euro, namely the US dollar, the British pound and the Japanese yen, and issued in the euro area. These instruments will be subject to a uniform haircut add-on of $8 \%$." See also [8].

The following exposition is based on the model used in [16]; in particular, we assume the presence of proportional transaction costs modeled via a closed convex cone which goes back to [17]. A special feature of this model is that the risk of a $d$-dimensional random variable is evaluated in terms of $m$ reference instruments with $1 \leq m \leq d$. Usually $m \ll d$ will be true, but the case $m=d$ is not excluded (and also, of course, neither $m=1, d>1$, nor $m=d=1$ ).

We shall make a few generalizations compared to [16]. First, we do not assume a "substitutability condition"; i.e., we do not assume that everything could be transferred into one distinguished currency. Second, we consider a general subspace $M$ of $\mathbb{R}^{d}$ as the collection

[^3]of risk evaluating reference portfolios rather than $\mathbb{R}^{m} \times\left\{0^{d-m}\right\}$. This was suggested by a referee of an earlier version [12] of this paper, and it also has a financial interpretation: The regulator could accept only reference portfolios where two (for example) components must be in a certain proportion, e.g., one third euro, two thirds dollar. Third, we consider risk measures on $L^{p}$-spaces with $1 \leq p \leq+\infty$ rather than on $L^{\infty}$ only, and we allow for a value which corresponds to $+\infty$ in the scalar case. Finally, we will start by introducing certain subsets of the power set of $M$ as image spaces for set-valued (convex) risk measures. This methodology goes somewhat the opposite way compared to [16] and popular references about set-valued mappings like [1], where the starting point is a relation or correspondence between two vector spaces, and a property like convexity of such a correspondence has consequences for the image sets (compare Property 3.1 in [16]). Going the other way round results, e.g., in a condition for the closedness of the values of a set-valued risk measure defined via an acceptance set (compare Definition 2.2 below).

The main accomplishment of the paper is a complete duality theory for set-valued convex risk measures which is parallel to the scalar case (and includes it as a special case). This is an extension of previous results: In [16] and [4], only the coherent case could be dealt with. Moreover, in contrast to the mentioned references, our dual representations of setvalued closed convex risk measures are formulated in terms of set-valued upper expectations involving two types of dual variables which can be given a financial interpretation. One is a set of (vector) probability measures that are absolutely continuous with respect to the given physical measure (slightly deviating even in the sublinear case from [16], where a set of general vector measures is used); the other is finite dimensional, reflects the possible preferences of the investor among the reference instruments, and is closely related to linear scalarizations. A coupling condition for the two types of dual variables (see Lemma 4.1 below) indicates that the investor has a certain but not total freedom regarding the relationship between her risk attitude in the individual markets, on the one hand, and her preference between the reference instruments, on the other hand.

It will turn out that set-valued convex risk measures admit a scalar representation (alongside a "primal" via acceptance sets and a "dual" via set-valued upper expectations) by a family of linear scalarizations which can be seen as generalizations of scalar risk measures for multivariate random variables, as recently defined by Burgert and Rüschendorf in [3]. The scalarizations not only are of theoretical interest, but can be used to actually compute the values of set-valued risk measures. We do not focus on this point here, but we shall mention that there are strong relationships to recent developments in vector optimization theory; see [19] and earlier work of Löhne.

Let us mention that Cascos and Molchanov in [4] give a definition of risk measures with values in an abstract convex cone. Despite the fact that there are advantages and drawbacks of such an abstract approach (see Remark 2.5 below), they give a list of operations [4, section 4] that can also be used to construct new set-valued risk measures within the framework of the present paper. Moreover, they link the concepts of set-valued risk measures and depthtrimmed (or central) regions from multivariate statistics. However, a dual representation for general convex risk measures (and corresponding central regions) cannot be found in [4].

Finally, let us note that our report [12] also contains a duality result which could be transferred into the results of this paper by means of Lemma 4.1 below. Therein, the method
of proof is "purely set-valued," whereas the proof in section 6 below makes use of the scalar Fenchel-Moreau theorem from convex analysis. However, the mathematical essence of both proofs is the same. See [11] for a discussion of this question.

The rest of the paper is organized as follows. Section 2 contains a mathematical model of the situation, including definitions for set-valued (convex) measures of risk and acceptance sets. The link between set-valued risk measures and their acceptance sets is given in terms of bijection theorems in section 3. Section 4 contains the definition of set-valued upper expectations and the coupling condition result. In section 5 we will give scalar representations of set-valued convex risk measures, which have meaning in financial terms and are needed for the proof of the main dual representation result. Dual representations of convex and coherent risk measures can be found in section 6, including "penalty function" representation formulas and a subsection about $\sigma\left(L_{d}^{\infty}, L_{d}^{1}\right)$-lower semicontinuity ("Fatou property"). Every section contains-as far as possible - an interpretation of the concepts and results in financial terms. Examples are given throughout the paper, among them general versions of set-valued value at risk, average value at risk, and a set-valued variant of the entropic risk measure.

## 2. The mathematical model and examples.

2.1. Vector-valued random variables. Everywhere in this paper, $(\Omega, \mathcal{F}, P)$ is a probability space. By $L_{d}^{p}=L_{d}^{p}(\Omega, \mathcal{F}, P), 1 \leq p \leq \infty$, we denote the linear space of all $P$ measurable functions $X: \Omega \rightarrow \mathbb{R}^{d}$ such that $\int_{\Omega}|X(\omega)|^{p} d P<\infty$ for $1 \leq p<\infty$ and ess.sup ${ }_{\omega \in \Omega}|X(\omega)|<\infty$ for $p=\infty$. In all cases, $|\cdot|$ denotes an arbitrary but fixed norm on $\mathbb{R}^{d}$, and the usual identification of functions differing only on sets of $P$-measure zero is assumed; hence $L_{d}^{p}$ is a Banach space reflexive for $1<p<\infty$. In the latter case, the topological dual of $L_{d}^{p}$ can be identified with $L_{d}^{q}$ with $\frac{1}{p}+\frac{1}{q}=1$, where the bilinear form conveying the dual pairing between $L_{d}^{p}$ and $L_{d}^{q}$ is given by $(X, Y) \rightarrow E\left[X^{T} Y\right]$. Unless otherwise stated we will consider $L_{d}^{\infty}$ provided with the $\sigma\left(L_{d}^{\infty}, L_{d}^{1}\right)$-topology and the same dual pairing with $L_{d}^{1}$ as for the other $L_{d}^{p}$,s. We write $E[X]$ for the (componentwise) mathematical expectation of $X \in L_{d}^{p}$ under $P$.

An element $X \in L_{d}^{p}$ has components $X_{1}, \ldots, X_{d}$ in $L^{p}=L_{1}^{p}$. The symbol $\mathbb{I}$ denotes the random variable in $L^{p}$ which has $P$-almost surely the value 1 .

Let $M \subseteq \mathbb{R}^{d}$ be a linear subspace with dimension $m \geq 1$. Particular cases are $m=1$ and $m=d$. The introduction of $M$ is motivated by the idea that an investor or regulator accepts risk compensations or security deposits only in a certain subset of the $d$ markets or currencies, for example only in the first $m$, in which case $M=\mathbb{R}^{m} \times\{0\}^{d-m}$ (see [16]). We assume $M \cap \mathbb{R}_{+}^{d} \neq\{0\}$, which means that there is at least one position of reference instruments with nonnegative components which are accepted as risk compensation or deposit.

The frictions between the markets are modeled by the solvency cone $K \subseteq \mathbb{R}^{d}$, which is a closed convex cone with $\mathbb{R}_{+}^{d} \subseteq K$. The cone $K$ includes all positions of reference instruments which can be exchanged, by paying transaction costs, into reference positions with nonnegative entries only. If $K$ contains a line, then there is an exchange without transaction costs possible between at least two of the $d$ instruments. If $K=\mathbb{R}_{+}^{d}$, then there is no exchange possible. Typically, the solvency cone can be constructed from a bid-ask matrix, as is done in [24], which is a generalization of the geometric model for currency markets with proportional transaction costs in [17]. Note that in [17], [16], and [24], $K$ is a polyhedral cone by construction, but
none of our (duality) results below depends on this property.
Example 2.1. Consider a model with $d=3$ currency markets, say cash in euro, US dollar, and Japanese yen, respectively. Let the exchange rates (including transaction costs) be given by the following $3 \times 3$ bid-ask matrix:

$$
\Pi=\left(\begin{array}{ccc}
1 & 0.77 & 0.008 \\
1.4 & 1 & 0.01 \\
141.5 & 104.8 & 1
\end{array}\right)
$$

where the entry $\pi_{i j}$ means that one has to pay $\pi_{i j}$ units of currency $i$ in order to get one unit of currency $j$, where $i, j \in\{1,2,3\}$. The solvency cone $K$ for this model is the convex cone spanned by the six vectors

$$
\left(\begin{array}{c}
0.77 \\
-1 \\
0
\end{array}\right), \quad\left(\begin{array}{c}
0.008 \\
0 \\
-1
\end{array}\right), \quad\left(\begin{array}{c}
-1 \\
1.4 \\
0
\end{array}\right), \quad\left(\begin{array}{c}
0 \\
0.01 \\
-1
\end{array}\right), \quad\left(\begin{array}{c}
-1 \\
0 \\
141.5
\end{array}\right), \quad\left(\begin{array}{c}
0 \\
-1 \\
104.8
\end{array}\right) .
$$

The first vector, for example, has the meaning that one can clear a debt of 1 US dollar by paying 0.77 euros. Note that $\mathbb{R}^{3}$ with the order relation generated by $K$ is not a vector lattice since it is not possible to find three linearly independent vectors which generate $K$ (compare [21, Example 1.5]).

The part of the cone $K$ that is relevant for $M$ is $K_{M}=K \cap M$, also a closed convex cone. By int $K_{M}$ we denote the interior of $K_{M}$ in $M$, i.e., the interior of $K_{M}$ considered as a subset of the finite dimensional linear space $M$. Throughout the paper, it is assumed that int $K_{M} \neq \emptyset$. This is equivalent to the existence of a base of $M$ consisting only of ( $m$ ) elements of $K$.

For $1 \leq p \leq \infty$, the set

$$
\begin{equation*}
L_{d}^{p}(K)=\left\{X \in L_{d}^{p}: X \in K P \text {-a.s. }\right\} \tag{2.1}
\end{equation*}
$$

is a closed convex cone in $L_{d}^{p}$ generating a reflexive transitive relation for $\mathbb{R}^{d}$-valued random variables. We shall write $\left(L_{d}^{p}\right)_{+}$for $L_{d}^{p}\left(\mathbb{R}_{+}^{d}\right)$.
2.2. Definition of risk measures. By $\mathbb{F}_{M}=\left\{D \subseteq M: D=\operatorname{cl}\left(D+K_{M}\right)\right\}$ and $\mathbb{G}_{M}=$ $\left\{D \subseteq M: D=\mathrm{cl}\right.$ co $\left.\left(D+K_{M}\right)\right\}$ we denote the collections of the upper closed and upper closed convex subsets, respectively, of $M$. The families $\mathbb{F}_{M}, \mathbb{G}_{M}$ of subsets of $M$ are both closed under multiplication with positive reals and addition $\oplus$ defined by $D_{1} \oplus D_{2}=\mathrm{cl}\left(D_{1}+D_{2}\right)$ for $D_{1}, D_{2} \in \mathbb{F}_{M}$ or $\in \mathbb{G}_{M}$. The + sign denotes usual Minkowski addition with $\emptyset+D=D+\emptyset=\emptyset$ for all $D \subseteq M$. The multiplication is extended by $t \emptyset=\emptyset$ for $t>0$ and $0 D=K_{M}$ for all $D \in \mathbb{F}_{M}$; in particular, $0 \emptyset=K_{M}$.

A function $R: L_{d}^{p} \rightarrow \mathbb{F}_{M}$ is said to be closed (convex, sublinear) iff

$$
\operatorname{graph} R=\left\{(X, x) \in L_{d}^{p} \times M: x \in R(X)\right\}
$$

is closed (convex, a convex cone). If a function $R: L_{d}^{p} \rightarrow \mathbb{F}_{M}$ is convex, then $R(X)$ is convex for all $X \in L_{d}^{p}$. If $R$ is closed convex, then $R(X) \in \mathbb{G}_{M}$ for all $X \in L_{d}^{p}$. The effective domain of a function $R: L_{d}^{p} \rightarrow \mathbb{F}_{M}$ is the set

$$
\operatorname{dom} R=\left\{X \in L_{d}^{p}: R(X) \neq \emptyset\right\} .
$$

A function $R: L_{d}^{p} \rightarrow \mathbb{F}_{M}$ is said to be proper iff $\operatorname{dom} R \neq \emptyset$ and $R(X) \neq M$ for all $X \in L_{d}^{p}$.
Remark 2.1. The definition of convexity of an $\mathbb{F}_{M}$-valued function $R$ via its graph is in line with the usual terminology for set-valued maps; see [1, section 1.3], [10, section 2.4]. Moreover, for such functions the graph and the epigraph defined as epi $R=\left\{X \times u \in L_{d}^{p} \times M: u \in R(X)\right.$ $\left.+K_{M}\right\}$ do coincide; see also [10, section 2.4]. On the other hand, one could suggest defining the epigraph of a function $R: L_{d}^{p} \rightarrow \mathbb{F}_{M}$ by

$$
\text { epi } R=\left\{(X, D) \in L_{d}^{p} \times \mathbb{F}_{M}: R(X) \supseteq D\right\}
$$

Since $L_{d}^{p} \times \mathbb{F}_{M}$ is not a linear space one has to introduce an algebraic structure and a convexity concept on this set. This can be done in the framework of conlinear spaces (see [11]). However, it turns out that convexity in such a sense is equivalent to graph convexity as used in the present paper, and also to

$$
\forall X^{1}, X^{2} \in L_{d}^{p}, \forall t \in(0,1): R\left(t X^{1}+(1-t) X^{2}\right) \supseteq t R\left(X^{1}\right) \oplus(1-t) R\left(X^{1}\right) .
$$

Moreover, by means of the graph, closedness of $\mathbb{F}_{M}$-valued functions can be introduced easily. See also Remark 2.5 below. For more information about lower semicontinuity of set-valued functions, compare [14].

Definition 2.1. $A$ set-valued measure of risk is a function $R: L_{d}^{p} \rightarrow \mathbb{F}_{M}$ which is
(R0) normalized-i.e., $K_{M} \subseteq R(0)$ and $R(0) \cap-\operatorname{int} K_{M}=\emptyset$;
(R1) translative in $M$-i.e.,

$$
\begin{equation*}
\forall X \in L_{d}^{p}, \forall u \in M: R(X+u \mathbb{I})=R(X)-u ; \tag{2.2}
\end{equation*}
$$

(R2) $L_{d}^{p}(K)$-monotone-i.e., $X^{2}-X^{1} \in L_{d}^{p}(K)$ implies $R\left(X^{2}\right) \supseteq R\left(X^{1}\right)$.
If $R$ satisfies (R0), (R1), and (R2) and is convex (sublinear), then it is called a (set-valued) convex (coherent) measure of risk.

Interpretation. The value $R(X)$ includes all possible vectors of accepted reference instruments compensating for the risk of the multivariate position $X$. This set is closed, and if $u \in M$ compensates for the risk of $X$, then, of course, each element of $u+K_{M}$ compensates all the more. Condition (R0) is a generalization of the normalization condition $\varrho(0)=0$ (compare [9, p. 154]). The first condition means that "doing nothing" is an acceptable position in the sense that every deterministic position in $M$ which can be exchanged into a position with nonnegative entries only compensates for its risk. On the other hand, doing nothing should not allow for consuming capital: $u \in-\operatorname{int} K_{M} \cap R(0)$ would mean that the investor is still good with taking $-u$ of her capital, exchanging it into an element of int $R_{+}^{d}$, consuming it, and having zero payoff in the future. A stronger condition is
(R0S) $R(0)=K_{M}$,
which is often satisfied.
(R1) is basically the translation condition formulated in [16] generalized to a linear subspace $M$ instead of $\mathbb{R}^{m} \times\{0\}^{d-m}$ : If the reference position $u$ is added to the random payoff, its risk decreases by $u$. (R2) is monotonicity with respect to $\supseteq$ : It should be intuitively clear that for a less risky position there cannot be less possibility to compensate for its risk.

Remark 2.2. The convexity of a risk measure avoids the strange effect of risk cancellation by dividing a position. Indeed, if $R$ is, for example, positively homogeneous but not convex,
then there are positions $X, X^{1}, X^{2} \in L_{d}^{p}$ and elements $u^{1} \in R\left(X^{1}\right), u^{2} \in R\left(X^{2}\right)$ such that $X=X^{1}+X^{2}$ and $u^{1}$ compensates for the risk of $X^{1}, u^{2}$ for the risk of $X^{2}$, but the risk of $X$ cannot be compensated for by $u^{1}+u^{2}$. As in the scalar case, this possibility is one drawback of set-valued variants of the value at risk; see Example 3.3 below.

Remark 2.3. The axioms in Definition 2.1, the addition $\oplus$, and the order relation $\supseteq$ in $\mathbb{F}_{M}$ fit together: A position $X \in L_{d}^{p}$ with $\emptyset=R(X)$ is the worst possibility; its risk cannot be compensated. Thus, considering $\supseteq$ as a replacement for $\leq$ in $\mathbb{R} \cup\{ \pm \infty\}$, the element $\emptyset \in \mathbb{F}_{M}$ turns out to correspond to $+\infty$. Moreover, if one would have $u \in R(X)+R(Y)$ and $R(X)=\emptyset, v \in R(Y)$, then one would expect $u-v \in R(X)$, which leads to contradictions.

Remark 2.4. In the scalar case, the value $\varrho(X)$ of a risk measure for a univariate position $X$ can be seen as the infimal amount of the (single) reference instrument (cash) compensating for the risk of $X$ (rather than the set of all possible amounts). It is possible to define a counterpart for this infimal amount in the set-valued case. This would be the infimal set of $R(X)$ with respect to the cone $K_{M}$ in the sense of Tanino [25]. It is also possible to formulate the whole (duality) theory of this paper in terms of infimal sets. We abdicate this possibility here for the sake of technical clarity. The interested reader is referred to [19].

Remark 2.5. In [4], Cascos and Molchanov defined risk measures with values in an ordered topological monoid (a topological semigroup with neutral element), which is a complete lattice and equipped with a continuous multiplication with positive reals. They claim (see their Example 2.4) that the collection $\mathbb{G}_{\mathbb{R}^{d}}$ of upper closed convex subsets of $\mathbb{R}^{d}$ (denoted as $\mathbb{G}_{K}$ in [4]) form such an entity. This is true up to the fact that they do not provide a topology on $\mathbb{G}_{K}$, making it a topological monoid. Moreover, it seems desirable to have a multiplication with $0 \in \mathbb{R}$ as well. We think that it is not necessary to introduce such a topology when working with set-valued risk measures as defined above, not even for duality results as given below, and it is also possible to generalize the above definition to an abstract level (with ordered convex cones as image space), but this should be done in a way consistent with all possible applications and in terms which have a financial meaning. Moreover, in [4, Definition 2.3, R1] the translation condition is required for all elements $u \in \mathbb{R}^{d}$ rather than only for those of $M$.

We turn to properties of acceptance sets for set-valued risk measures.
Definition 2.2. We call a set $A \subseteq L_{d}^{p}$ directionally closed in $M$ iff $X \in L_{d}^{p},\left\{u^{k}\right\}_{k \in \mathbb{N}} \subset M$, $\lim _{k \rightarrow \infty} u^{k}=0$, and $X+u^{k} \mathbb{I} \in A$ for all $k \in \mathbb{N}$ implies $X \in A$.

This property of being directionally closed will turn out to be the weakest property ensuring the closedness of the values of a risk measure generated by an acceptance set. It is weaker than algebraic closedness, used in the case $m=d=1$ in [15]. For example, the set $A=\left\{X \in \mathbb{R}^{3}: X_{3}>0\right\}$ is directionally closed in $M=\operatorname{span}\left((1,0,0)^{T},(0,1,0)^{T}\right)$ but not algebraically closed.

Definition 2.3. An acceptance set is a subset $A \subseteq L_{d}^{p}$ satisfying the following:
(A0) $u \in K_{M}$ implies $u \mathbb{I} \in A$, and $u \in-\operatorname{int} K_{M}$ implies $u \mathbb{I} \notin A$;
(A1) $A$ is directionally closed in $M$ with $A+u \mathbb{I} \subseteq A$ whenever $u \in K_{M}$;
(A2) $A+L_{d}^{p}(K) \subseteq A$.
If $A$ satisfies (A0), (A1), (A2) and is convex (a convex cone), then it is called a convex (coherent) acceptance set.

A stronger condition than (A0) which will correspond to (R0S) is
(A0S) $u \mathbb{I} \in A$ iff $u \in K_{M}$.
Note that (A2) and the definitions of $L_{d}^{p}(K)$ (see (2.1)) and $K_{M}$ already imply the second part of (A1), i.e., $A+u \mathbb{I} \subseteq A$ whenever $u \in K_{M}$, but we will see in the next section that this property plays a particular role for the one-to-one correspondence between risk measures and acceptance sets. Moreover, by (A0), $0 \in A$. This and (A2) imply $L_{d}^{p}(K) \subseteq A$.
3. Primal representation of risk measures. In this section we shall establish bijection results for set-valued risk measures and acceptance sets. Let $R: L_{d}^{p} \rightarrow \mathbb{F}_{M}$ be a function. By means of

$$
\begin{equation*}
A_{R}=\left\{X \in L_{d}^{p}: 0 \in R(X)\right\}=\left\{X \in L_{d}^{p}: K_{M} \subseteq R(X)\right\} \tag{3.1}
\end{equation*}
$$

we assign to $R$ its zero sublevel set. Let $A \subseteq L_{d}^{p}$ be a set. By means of

$$
\begin{equation*}
R_{A}(X)=\{u \in M: X+u \mathbb{I} \in A\} \tag{3.2}
\end{equation*}
$$

we assign to $A$ a function $R_{A}$ mapping $L_{d}^{p}$ into the power set of $M$.
Interpretation. It should be clear that a multivariate position $X$ is acceptable if "depositing nothing" already compensates for its risk. Of course, if this is the case, every nonnegative accepted reference position (an element of $K_{M}$ ) does the job all the more. On the other hand, $R_{A}(X)$ contains all accepted reference positions, which make $X$ acceptable when added to it.

It turns out that (3.1) and (3.2) yield one-to-one correspondences between acceptance sets and set-valued risk measures.

Proposition 3.1. (i) Let $R: L_{d}^{p} \rightarrow \mathbb{F}_{M}$ be translative in $M$; i.e., it satisfies (R1). Then $A_{R}$ satisfies (A1) and we have $R=R_{A_{R}}$. (ii) Let $A \subseteq L_{d}^{p}$ be a set satisfying (A1). Then $R_{A}$ maps into $\mathbb{F}_{M}$, is translative, and $A=A_{R_{A}}$.

Proof. (i) First, we show (A1). Take $X \in L_{d}^{p},\left\{u^{k}\right\}_{k \in \mathbb{N}} \subset M$ with $\lim _{k \rightarrow \infty} u^{k}=0$, and $X+u^{k} \mathbb{I} \in A$. Then (3.1) and (R1) imply $0 \in R\left(X+u^{k} \mathbb{I}\right)=R(X)-u^{k}$; i.e., $u^{k} \in R(X)$ for all $k \in \mathbb{N}$. Since $R$ maps into $\mathbb{F}_{M}, R(X)$ is closed; hence $0 \in R(X)$, implying $X \in A_{R}$ by (3.1). For the second part of (A1) take $X \in A_{R}$ and $u \in K_{M}$. Since $K_{M}$ is a convex cone we get $K_{M} \subseteq K_{M}-u$. Since $X \in A_{R}$ we have $K_{M} \subseteq R(X)$; hence by (R1),

$$
K_{M} \subseteq K_{M}-u \subseteq R(X)-u=R(X+u \mathbb{I}),
$$

which gives $X+u \mathbb{I} \in A_{R}$.
Direct calculations using (3.2), (3.1), and (R1) yield

$$
\begin{aligned}
R_{A_{R}}(X) & =\left\{u \in M: X+u \mathbb{I} \in A_{R}\right\} \\
& =\{u \in M: 0 \in R(X+u \mathbb{I})\}=R(X) .
\end{aligned}
$$

(ii) First, we show that $R_{A}$ maps into $\mathbb{F}_{M}$. For this it is enough to show that $R_{A}(X)+$ $K_{M} \subseteq R_{A}(X)$ and $R_{A}(X)$ is closed for each $X \in L_{d}^{p}$. Take $u \in R_{A}(X)$ and $v \in K_{M}$. Then $X+u \mathbb{I} \in A$. (A1) implies $X+(u+v) \mathbb{I} \in A$; hence $u+v \in R_{A}(X)$ by (3.2). Take a sequence $\left\{u^{k}\right\}_{k \in \mathbb{N}} \subset R_{A}(X)$ with $\lim _{k \rightarrow \infty} u^{k}=u$. Then by (3.2), $X+u^{k} \mathbb{I}=(X+u \mathbb{I})+\left(u^{k}-u\right) \mathbb{I} \in A$ for all $k \in \mathbb{N}$. Since $A$ is directionally closed in $M$, this implies $X+u \mathbb{I} \in A$, which gives $u \in R_{A}(X)$.

We turn to (R1). Take $X \in L_{d}^{p}, u \in M$. Then by (3.2),

$$
\begin{aligned}
R_{A}(X+u \mathbb{I}) & =\{v \in M: X+(u+v) \mathbb{I} \in A\} \\
& =\{u+v \in M: X+(u+v) \mathbb{I} \in A\}-u=R_{A}(X)-u .
\end{aligned}
$$

From (3.2), (3.1) we get

$$
A_{R_{A}}=\left\{X \in L_{d}^{p}: 0 \in R_{A}(X)\right\}=\left\{X \in L_{d}^{p}: 0 \in\{u \in M: X+u \mathbb{I} \in A\}\right\},
$$

which proves the desired equality.
Proposition 3.2. (i) Let $R: L_{d}^{p} \rightarrow \mathbb{F}_{M}$ be a measure of risk. Then $A_{R}$ is an acceptance set. If $R$ is convex, so is $A_{R}$. If $R$ is coherent, then $A_{R}$ is a coherent acceptance set. If $R$ satisfies (R0S), then $A_{R}$ satisfies (A0S).
(ii) Let $A \subseteq L_{d}^{p}$ be an acceptance set. Then $R_{A}$ is a measure of risk. If $A$ is convex, so is $R_{A}$. If $A$ is a coherent acceptance set, then $R_{A}$ is a coherent measure of risk. If $A$ satisfies (A0S), then $R_{A}$ satisfies (R0S).

Proof. (i) First, we show (A0): Take $u \in K_{M}$. Then $u \in R(0)$ by (R0), and $0 \in R(u \mathbb{I})$ by (R1). The construction (3.1) implies $u \mathbb{I} \in A_{R}$. Now, take $u \in-\operatorname{int} K_{M}$. Then by (R0), $u \notin R(0)$; hence by (R1), $0 \notin R(u \mathbb{I})$ and $u \mathbb{I} \notin A_{R}$.
(A1) follows from Proposition 3.1.
In order to check (A2) take $X^{1} \in A_{R}$ and $X^{2} \in L_{d}^{p}(K)$. Then $\left(X^{1}+X^{2}\right)-X^{1} \in L_{d}^{p}(K)$, and by (R2), $0 \in R\left(X^{1}\right) \subseteq R\left(X^{1}+X^{2}\right)$, and hence $X^{1}+X^{2} \in A_{R}$ as desired.

It is left to the reader to check that $A_{R}$ is convex if $R$ is convex and that $A_{R}$ is a cone if $R$ is positively homogeneous.
(ii) From Proposition 3.1 we already know that $R_{A}$ maps into $\mathbb{F}_{M}$ and that (R1) holds true.

Next, we shall show (R0). By (A0) and (3.2) we get $u \in R_{A}(0)$ whenever $u \in K_{M}$ and $u \notin R_{A}(0)$ whenever $u \in-\operatorname{int} K_{M}$.

Finally, for a proof of (R2) take $X^{1}, X^{2} \in L_{d}^{p}$ such that $X^{2}-X^{1} \in L_{d}^{p}(K)$. Then

$$
\begin{aligned}
R_{A}\left(X^{1}\right) & =\left\{u \in M: X^{1}+u \mathbb{I} \in A\right\} \\
& =\left\{u \in M: X^{2}+u \mathbb{I} \in A+\left(X^{2}-X^{1}\right)\right\} \\
& \subseteq\left\{u \in M: X^{2}+u \mathbb{I} \in A\right\}=R_{A}\left(X^{2}\right),
\end{aligned}
$$

where the inclusion holds true since $A+L_{d}^{p}(K) \subseteq A$ by (A2).
Again, it is a matter of exercise to show that $R_{A}$ is convex if $A$ is convex and that $R_{A}$ is positively homogeneous if $A$ is a cone.

Finally, the correspondence between (R0S) and (A0S) can easily be verified.
The proof of the above propositions makes it apparent which property of a (set-valued) translative function corresponds to what property of its zero sublevel set. In particular, the closedness of the values of the function corresponds to directional closedness of its zero sublevel set.

Remark 3.1. Translativity of $R$ is equivalent to

$$
\begin{equation*}
\operatorname{graph} R=\left\{(X, u) \in L_{d}^{p} \times M: X+u \mathbb{I} \in A_{R}\right\} \tag{3.3}
\end{equation*}
$$

Indeed, if $R$ is translative in $M$, then $(X, u) \in \operatorname{graph} R$ iff $u \in R(X)$ iff $0 \in R(X+u \mathbb{I})$ by translativity iff $X+u \mathbb{I} \in A_{R}$ by (3.1). On the other hand, if (3.3) is true, then $w \in R(X+u \mathbb{I})$ for $u, w \in M$ iff $(X+u \mathbb{I}, w) \in \operatorname{graph} R$ iff $X+(u+w) \mathbb{I} \in A_{R}$ by $(3.3)$ iff $(X, u+w) \in \operatorname{graph} R$ again (3.3) iff $u+w \in$ graph $R$, which is translativity in $M$.

Using (3.3), we may observe that $R$ is closed (by definition, iff graph $R$ is closed) iff $A_{R}$ is closed. This is one more example for a bijection property, and closedness can be added, e.g., in Proposition 3.2.

Example 3.1 (set-valued worst case risk measure). The smaller the acceptance set, the higher the degree of risk aversion. Thus, the risk measure with the highest risk aversion is the one with the smallest acceptance set. The smallest set satisfying properties (A0)-(A2) is the cone $L_{d}^{p}(K)$. The corresponding coherent risk measure is

$$
W C^{M, S}(X)=\left\{u \in M: X+u \mathbb{I} \in L_{d}^{p}(K)\right\} .
$$

In the scalar case $\left(d=1, M=\mathbb{R}, K=K_{M}=\mathbb{R}_{+}\right)$this risk measure coincides with the negative of the essential infimum of $X$, which is called the worst case risk measure. Note that there are other possibilities for generalizing the scalar counterpart. For example,

$$
W C^{M, W}(X)=\{u \in M: P(X+u \mathbb{I} \in-\operatorname{int} K)=0\}
$$

defines a less risk averse risk measure, which coincides in the scalar case with $W C^{M, S}$ but is not convex in general in higher dimensions.

Example 3.2 (negative componentwise expectation). The set

$$
A=\left\{X \in L_{d}^{1}: E[X] \in K\right\}
$$

is a coherent acceptance set satisfying (A0S). Indeed, it is obviously a convex cone, and we have $u \mathbb{I} \in A$ if $u \in K_{M}$. Moreover, $u \in M \backslash K_{M}$ implies $u \notin K$ by definition of $K_{M}$; hence $u \mathbb{I} \notin A$.

The corresponding closed coherent risk measure that could be considered as a set-valued counterpart of the negative expected value is defined by

$$
\operatorname{NCE}(X)=R_{A}(X)=\{u \in M: E[X+u \mathbb{I}] \in K\}=(E[-X]+K) \cap M
$$

on $L_{d}^{1}$. We shall call it negative componentwise expectation. Observe that if $M=\mathbb{R}^{d}$, then $E[-X] \in N C E(X)$ for all $X \in L_{d}^{1}$.

Example 3.3 (set-valued value at risk). We shall define two extensions of the scalar value at risk, which will be called weak and strong value at risk. Let $0 \leq \lambda \leq 1$. The functions $X \mapsto V @ R_{\lambda}^{M, W}, V @ R_{\lambda}^{M, S}: L_{d}^{p} \rightarrow \mathbb{F}_{M}$ defined by

$$
V @ R_{\lambda}^{M, W}(X)=\{u \in M: P(X+u \mathbb{I} \in-\operatorname{int} K) \leq \lambda\}
$$

and

$$
V @ R_{\lambda}^{M, S}(X)=\{u \in M: P(X+u \mathbb{I} \notin K) \leq \lambda\},
$$

respectively, are set-valued measures of risk on $L_{d}^{1}$. If $d=1, M=\mathbb{R}$, and $K=K_{M}=\mathbb{R}_{+}$, then $V @ R_{\lambda}^{M, W}(X)=V @ R_{\lambda}^{M, S}(X)=V @ R_{\lambda}(X)+\mathbb{R}_{+}$, where $V @ R_{\lambda}(X)=\inf \{t \in \mathbb{R}$ :
$P(X+t<0) \leq \lambda\}$ denotes the usual scalar value at risk as defined, for example, in [9, section 4.4]. Therefore, $V @ R_{\lambda}^{M, W}, V @ R_{\lambda}^{M, S}$ cannot be convex in general, but they do satisfy (R0), (R1), and (R2). The strong variant even satisfies (R0S) for $0 \leq \lambda<1$. The respective acceptance sets are obviously

$$
A_{V @ R_{\lambda}^{M, W}}=\left\{X \in L_{d}^{p}: P(X \in-\operatorname{int} K) \leq \lambda\right\}
$$

and

$$
A_{V @ R_{\lambda}^{M, S}}=\left\{X \in L_{d}^{p}: P(X \notin K) \leq \lambda\right\} .
$$

Note that $V @ R_{\lambda}^{M, W}(X) \supseteq V @ R_{\lambda}^{M, S}(X)$ for all $X \in L_{d}^{1}$; i.e., using the strong variant means being more risk averse. Moreover, $V @ R_{0}^{M, S}(X)=W C^{M, S}(X)$ and $V @ R_{0}^{M, W}(X)=$ $W C^{M, W}(X)$.

An even more general definition unifying the above two constructions and producing more examples can be found in [13] along with continuity properties of set-valued risk measures.

Remark 3.2. In [5], two extensions of the scalar value at risk are defined. The "multivariate lower-orthant Value-at-Risk" $\underline{V a}_{\alpha}$ is given by

$$
\underline{V_{a}} \underline{R}_{\alpha}(X)=\operatorname{bd}\left\{u \in \mathbb{R}^{d}: P(X \leq u) \geq \alpha\right\}
$$

where bd $D$ is the topological boundary of $D \subseteq \mathbb{R}^{d}$ with respect to the usual separated, locally convex topology. The transformation $\lambda=1-\alpha$ yields

$$
\underline{V a}_{\alpha}(X)=\operatorname{bd} V @ R_{\lambda}^{\mathbb{R}^{d}, S}(-X),
$$

where $K=K_{M}=\mathbb{R}_{+}^{d}$. The "multivariate upper-orthant Value-at-Risk" $\overline{V a R}_{\alpha}$ of [5] is

$$
\overline{\operatorname{VaR}}_{\alpha}(X)=\operatorname{bd}\left\{u \in \mathbb{R}^{d}: P(X>u) \leq 1-\alpha\right\} .
$$

Interpreting $X>u$ as $X-u \in \operatorname{int} \mathbb{R}_{+}^{d}$ and setting $\lambda=1-\alpha$, we obtain

$$
\overline{V a R}_{\alpha}(X)=\operatorname{bd} V @ R_{\lambda}^{\mathbb{R}^{d}, W}(-X) .
$$

In [20], the (set-valued) upper-orthant value at risk for bivariate componentwise nonnegative random variables is defined, which corresponds, when extended to arbitrary ones, to $\overline{V a R}_{\alpha}$ of [5], and hence to $V @ R_{\lambda}^{\mathbb{R}^{d}, W}$ for $d=2$.
4. Set-valued upper expectations. This section is devoted to the introduction of certain set-valued functions which will replace the expectations in dual representation theorems for convex risk measures. They will involve two types of dual variable, one being vector probability measures absolutely continuous with respect to the physical measure $P$, the other one reflecting the order relations in $\mathbb{R}^{d}$ and $M$ generated by $K$ and $K_{M}$, respectively.

We shall fix some notation. By $K^{+}$and $K_{M}^{+}$we denote the positive dual cones of the cones $K$ in $\mathbb{R}^{d}$ and $K_{M}$ in $M$, respectively. Thus,

$$
K_{M}^{+}=\left\{v \in M: \forall u \in K_{M}: v^{T} u \geq 0\right\} \subseteq M .
$$

Note that $K_{M}^{+}=\left(K^{+}+M^{\perp}\right) \cap M$ with $M^{\perp}=\left\{v \in \mathbb{R}^{d}: \forall u \in M: v^{T} u=0\right\}$ since $K^{+}+M^{\perp}$ is the dual cone of $K_{M}$ in $\mathbb{R}^{d}$. The reader should be aware that both $K_{M}^{+}$and $K^{+}+M^{\perp}$ are dual cones of $K_{M}$, the first in $M$, the second in $\mathbb{R}^{d}$. We will also use $G(v)=\left\{x \in \mathbb{R}^{d}: 0 \leq v^{T} x\right\}$ as a subset of $\mathbb{R}^{d}$ not necessarily of $M$.

The positive dual cone of $L_{d}^{p}(K)$ is $L_{d}^{q}\left(K^{+}\right)$with $\frac{1}{p}+\frac{1}{q}=1(q=\infty$ for $p=1, q=1$ for $p=\infty)$. Let us denote by $\mathcal{M}_{1, d}^{P}=\mathcal{M}_{1, d}^{P}(\Omega, \mathcal{F})$ the set of all vector probability measures with components absolutely continuous with respect to $P$; i.e., $Q_{i}: \mathcal{F} \rightarrow[0,1]$ is a probability measure on $(\Omega, \mathcal{F})$ such that $\frac{d Q_{i}}{d P} \in L^{1}$ for $i=1, \ldots, d$.

Take $Y \in L_{d}^{q}$ and $v \in M$. Define a function $F_{(Y, v)}^{M}: L_{d}^{p} \rightarrow 2^{M}$ by

$$
\begin{equation*}
F_{(Y, v)}^{M}[X]=\left\{u \in M: E\left[X^{T} Y\right] \leq v^{T} u\right\} \tag{4.1}
\end{equation*}
$$

Remark 4.1. Let $Y=Y_{\mathbb{I}}=(\mathbb{I}, \ldots, \mathbb{I})^{T}$ with $v=e=(1, \ldots, 1)^{T}$. If $M=\mathbb{R}^{d}$, then

$$
F_{\left(Y_{\mathbb{I}}, e\right)}^{M}[X]=E[X]+\left\{u \in \mathbb{R}^{d}: e^{T} u \geq 0\right\}
$$

In particular, if $d=1, Y=\mathbb{I}$, then $F_{(\mathbb{I}, 1)}^{M}[X]=E[X]+\mathbb{R}_{+}$. This shows that the $F_{(Y, v)}^{M}$ 's in a certain sense extend the concept of the mathematical expectation to a set-valued function.

Elementary properties of the functions $X \mapsto F_{(Y, v)}^{M}[X]$ are collected in the following proposition.

Proposition 4.1. The following hold:
(i) $F_{(Y, 0)}^{M}[X]=\emptyset$ if $E\left[X^{T} Y\right]>0 ; F_{(Y, 0)}^{M}[X]=M$ if $E\left[X^{T} Y\right] \leq 0$.
(ii) $F_{(Y, v)}^{M}: L_{d}^{p} \rightarrow \mathbb{G}_{M}$ is proper iff $v \in K_{M}^{+} \backslash\{0\}$.
(iii) If $\widehat{u} \in M$ is such that $v^{T} \widehat{u}=1$, then

$$
F_{(Y, v)}^{M}[X]=E\left[X^{T} Y\right] \widehat{u}+F_{(Y, v)}^{M}[0]=E\left[X^{T} Y\right] \widehat{u}+G(v) \cap M
$$

(iv) Let $Y \in L_{d}^{q}, v \in M \backslash\{0\}$. For $X^{1}, X^{2} \in L_{d}^{p}$, $s>0$ we have $F_{(Y, v)}^{M}\left[X^{1}+X^{2}\right]=$ $F_{(Y, v)}^{M}\left[X^{1}\right]+F_{(Y, v)}^{M}\left[X^{2}\right]$ and $F_{(Y, v)}^{M}[s X]=s F_{(Y, v)}^{M}[X]$; moreover, $F_{(Y, v)}^{M}[0]=G(v) \cap M$.

Proof. (i) This part holds by definition of $F_{(Y, 0)}^{M}$. (ii) "if-part": We have to show cl co $\left(F_{(Y, v)}^{M}[X]+K_{M}\right)=F_{(Y, v)}^{M}[X]$ for $X \in L_{d}^{p}$. Since $F_{(Y, v)}^{M}[X]$ is a closed convex half space (hence $F_{(Y, v)}^{M}$ is proper) and one inclusion is trivial, it suffices to show $F_{(Y, v)}^{M}[X]+K_{M} \subseteq F_{(Y, v)}^{M}[X]$. But this is immediate from (4.1). "only-if-part": Assume that $F_{(Y, v)}^{M}$ maps into $\mathbb{G}_{M}$, is proper, and $v \notin K_{M}^{+} \backslash\{0\}$. If $v=0$, then $F_{(Y, v)}^{M}$ is not proper, and hence this case is not possible. If $v \neq 0$, there is $u^{0} \in K_{M}$ such that $v^{T} u^{0}<0$. Taking $u \in F_{(Y, v)}^{M}[X]$, we have $u+t u^{0} \in F_{(Y, v)}^{M}[X]$ for all $t>0$, by assumption. But this means $E\left[X^{T} Y\right] \leq v^{T}\left(u+t u^{0}\right)$ for all $t>0$, which is not possible. (iii) Follows immediately from $F_{(Y, v)}^{M}[X]=\left\{u \in M: 0 \leq v^{T}\left(u-E\left[X^{T} Y\right] \widehat{u}\right)\right\}$. (iv) For additivity, observe that if $v \neq 0, X^{1} \in L_{d}^{p}$, and $Y \in L_{d}^{q}$, then there is $u^{1} \in M$ such that $E\left[Y^{T} X^{1}\right]+v^{T}\left(u^{1}\right)=0$.

Proposition 4.2. Let $Y \in L_{d}^{q}\left(K^{+}\right)$and $v \in\left(E[Y]+M^{\perp}\right) \cap K_{M}^{+} \backslash\{0\}$. Then, the function $R: L_{d}^{p} \rightarrow 2^{M}$ defined by $R(X)=F_{(Y, v)}^{M}[-X]$ is a coherent measure of risk with acceptance set $A_{R}=\left\{X \in L_{d}^{p}: 0 \leq E\left[X^{T} Y\right]\right\}$.

Proof. First, note that $Y \in L_{d}^{q}\left(K^{+}\right)$implies $E[Y] \in K^{+}$. Proposition 4.1(ii) shows that $R$ indeed maps into $\mathbb{G}_{M}$. Moreover, $K_{M} \subseteq F_{(Y, v)}^{M}[0]=\left\{u \in M: 0 \leq v^{T} u\right\}$, and -int $K_{M} \cap$ $F_{(Y, v)}^{M}[0]=\emptyset$. Indeed, otherwise there is $u \in \operatorname{int} K_{M}$ such that $0 \leq-v^{T} u$; hence $v^{T} u=0$ since $v \in K_{M}^{+}$. Since $u \in \operatorname{int} K_{M}$ there is a circled neighborhood $N$ of $0 \in M$ such that $u+N \subseteq K_{M}$. Since $v \neq 0$ there is $w \in N$ such that $v^{T} w>0$. This implies $v^{T}(u+w)>0$ and $v^{T}(u-w)<0$. The last inequality is a contradiction since $u-w \in K_{M}$. Thus, (R0) is satisfied.

To check (R1) take $u \in M$. Then

$$
\begin{aligned}
F_{(Y, v)}^{M}[-X-u \mathbb{I}] & =\left\{z \in M: E\left[(X+u \mathbb{I})^{T} Y\right]+v^{T} z \geq 0\right\} \\
& =\left\{z \in M: E\left[X^{T} Y\right]+E[Y]^{T} u+v^{T} z \geq 0\right\} \\
& =\left\{z+u \in M: E\left[X^{T} Y\right]+(E[Y]-v)^{T} u+v^{T}(z+u) \geq 0\right\}-u \\
& =F_{(Y, v)}^{M}[-X]-u
\end{aligned}
$$

The last equation in this chain follows since, by assumption, $E[Y]-v \in M^{\perp}$.
To check (R2) take $X^{1}, X^{2} \in L_{d}^{p}$ such that $X^{2}-X^{1} \in L_{d}^{p}(K)$ and $u \in F_{(Y, v)}^{M}\left[-X^{1}\right]$. Then

$$
E\left[\left(-X^{2}\right)^{T} Y\right]+E\left[\left(X^{2}-X^{1}\right)^{T} Y\right] \leq v^{T} u
$$

and since $Y \in L_{d}^{q}\left(K^{+}\right)$we have $E\left[\left(X^{2}-X^{1}\right)^{T} Y\right] \geq 0$. This proves $u \in F_{(Y, v)}^{M}\left[-X^{2}\right]$.
The sublinearity follows from Proposition 4.1(iv).
Remark 4.2. Let $Y, v$ be as in Proposition 4.2 such that $E[Y]=e=(1, \ldots, 1)^{T}$; i.e., $Y$ is the density vector of some vector probability measure $Q$. The risk measure $X \mapsto F_{(Y, v)}^{M}[-X]$ can be seen as a set-valued generalization of the financial version of the correlation risk measure defined in Remark 2.2 of [23], for example. The corresponding maximal correlation risk measure is

$$
\widehat{R}(X)=\bigcap_{\tilde{X} \sim X} F_{(Y, v)}^{M}[-\tilde{X}],
$$

where the intersection runs through all $\widetilde{X}$ with the same joint distribution as $X$. This risk measure is of course law invariant in the sense that it gives the same value for all $X$ with the same joint distribution.

Via a change of variables we shall give another description of the functions $F_{(Y, v)}^{M}$. The new form will make the interpretation more apparent. Given an element $w \in \mathbb{R}^{d}$, we shall denote by $\operatorname{diag}(w)$ the diagonal matrix in $\mathbb{R}^{d \times d}$ with the components of $w$ in its main diagonal and zeros elsewhere.

Lemma 4.1. (i) Let $Y \in L_{d}^{q}\left(K^{+}\right), v \in\left(E[Y]+M^{\perp}\right) \cap K_{M}^{+} \backslash\{0\}$. Then there are $Q \in \mathcal{M}_{1, d}^{P}$, $w \in K^{+} \backslash M^{\perp}$ such that $\operatorname{diag}(w) \frac{d Q}{d P} \in L_{d}^{q}\left(K^{+}\right)$and $F_{(Y, v)}^{M}=\widetilde{F}_{(Q, w)}^{M}$ with

$$
\widetilde{F}_{(Q, w)}^{M}[X]=\left\{u \in M: w^{T} E^{Q}[X] \leq w^{T} u\right\}=\left(E^{Q}[X]+G(w)\right) \cap M .
$$

(ii) Conversely, if $Q \in \mathcal{M}_{1, d}^{P}, w \in K^{+} \backslash M^{\perp}$ such that $\operatorname{diag}(w) \frac{d Q}{d P} \in L_{d}^{q}\left(K^{+}\right)$, then there are $Y \in L_{d}^{q}\left(K^{+}\right), v \in\left(E[Y]+M^{\perp}\right) \cap K_{M}^{+} \backslash\{0\}$ such that $\widetilde{F}_{(Q, w)}^{M}=F_{(Y, v)}^{M}$.

Proof. (i) Set $w=E[Y] \in K^{+}$. Since $v \in\left(E[Y]+M^{\perp}\right) \cap K_{M}^{+} \backslash\{0\}$ we have $w \notin$ $M^{\perp}$. Choose $Z_{i}=\frac{1}{w_{i}} Y_{i}$ if $w_{i}>0$ and arbitrary in $\left(L_{d}^{q}\right)_{+}$such that $E\left[Z_{i}\right]=1$ if $w_{i}=0$, $i \in\{1, \ldots, d\}$. Define $Q$ via $\frac{d Q}{d P}=Z$. Then $Y=\operatorname{diag}(w) \frac{d Q}{d P} \in L_{d}^{q}\left(K^{+}\right)$and $E\left[X^{T} Y\right]=$ $E\left[X^{T} \operatorname{diag}(w) \frac{d Q}{d P}\right]=w^{T} E^{Q}[X]$. Since by assumption $v \in w+M^{\perp}$, and hence $v^{T} u=w^{T} u$ for all $u \in M$, we obtain

$$
\begin{equation*}
F_{(Y, v)}^{M}[X]=\left\{u \in M: E\left[X^{T} Y\right] \leq v^{T} u\right\}=\left\{u \in M: w^{T} E^{Q}[X] \leq w^{T} u\right\}=\widetilde{F}_{(Q, w)}^{M}[X] . \tag{4.2}
\end{equation*}
$$

(ii) Define $Y=\operatorname{diag}(w) \frac{d Q}{d P} \in L_{d}^{q}\left(K^{+}\right)$. Then $E[Y]=w$ and $w^{T} E^{Q}[X]=$ $E\left[X^{T} \operatorname{diag}(w) \frac{d Q}{d P}\right]=E\left[X^{T} Y\right]$. We claim that $w \in K_{M}^{+}+M^{\perp}$. Indeed, on the one hand, from $K_{M}^{+}=\left(K^{+}+M^{\perp}\right) \cap M$ we get $K_{M}^{+}+M^{\perp}=\left(K^{+}+M^{\perp}\right) \cap M+M^{\perp}$. On the other hand, $w=w_{M}+w_{M^{\perp}}$ with $w_{M} \in M, w_{M^{\perp}} \in M^{\perp}$, and hence $w_{M}=w-w_{M^{\perp}} \in K^{+}+M^{\perp}$, which in turn implies $w \in\left(K^{+}+M^{\perp}\right) \cap M+M^{\perp}$, and the claim is proven. Hence, there is $v \in K_{M}^{+}$such that $w \in v+M^{\perp}$. Since $w \notin M^{\perp}$ we have $v \neq 0$. Finally, (4.2) also holds in this case, which completes the proof of the lemma.

Interpretation. Of course, the above lemma is trivial in the case $m=d=1$. In higher dimensions, it provides a link between the risk attitude of the investor in the individual markets, on the one hand, and the preference of the investor among the reference instruments, on the other hand. This is done by the coupling conditions diag $(w) \frac{d Q}{d P} \in L_{d}^{q}\left(K^{+}\right)$(in terms of probability measures) and $v \in\left(E[Y]+M^{\perp}\right) \cap K_{M}^{+} \backslash\{0\}$ (in terms of density functions).

The investor can model his risk attitude in market $i$ by considering certain probability measures $Q_{i}$ as alternatives to the physical measure $P$. He can also have a preference between the reference instruments (i.e., the markets) themselves: For example, if he prefers to work in market $i$ rather than market $j$, then $w_{i}>w_{j}$. Moreover, this preference is quantified by the proportion $\frac{w_{i}}{w_{j}}$. For this reason, the vectors $w$ could be considered as internal weights or exchange rates of the investor.

Finally, the coupling condition tells him that he cannot choose his risk attitude in the individual markets, on the one hand, and his internal exchange rates between the markets, on the other hand, independently: They restrict each other. This should also explain why it makes sense to work with pairs $(Q, w)$ (or $(Y, v)$ ) of dual variables. Mathematically, the additional variables $w$ and $v$ reflect the order (generated by $K$ and $K_{M}$ ) in the image spaces, and this order should enter a (duality) theory somehow.

Remark 4.3. The variable $w$ reflects the investor's weighting of all $d$ reference instruments (markets), while the $v$ 's reflect only her weighting of the references instruments accepted for risk compensation. However, since $v-w \in M^{\perp}$ by construction $(v$ is the projection of $w$ on $M)$, for $v, w$ corresponding to each other according to Lemma 4.1 above they yield the same weighted sum $v^{T} u=w^{T} u$ for all $u \in M$, the candidates for risk compensation. Starting with a pair $(Y, v)$ as in Lemma 4.1(i), the investor's weighting of all $d$ markets is already contained in $Y$, as $E[Y]$ is going to be $w$, and $v$ is determined by this $w=E[Y]$ as its projection onto $M$. Note also that $w \in K^{+}$and $v \in K_{M}^{+}$means that for "good" reference instruments $x \in K$ and $u \in K_{M}$ the weighted sums $w^{T} x$ and $v^{T} u$ are nonnegative.

Remark 4.4. Let $Q: \mathcal{F} \rightarrow[0,1]^{d}$ be a vector probability measure and $w \in K^{+} \backslash M^{\perp}$. Proposition 4.2 and Lemma 4.1 taken together show that the function $X \mapsto \widetilde{F}_{(Q, w)}^{M}[-X]=$ $\left(E^{Q}[-X]+G(w)\right) \cap M$ is a coherent risk measure into $\mathbb{G}_{M}$, provided that diag $(w) \frac{d Q}{d P} \in$ $L_{d}^{q}\left(K^{+}\right)$.

Remark 4.5. If $w \in M^{\perp}$ would be admissible, then, since $u \in \widetilde{F}_{(Q, w)}^{M}[-X]=\left(E^{Q}[-X]+\right.$ $G(w)) \cap M$ iff $w^{T}\left(u+E^{Q}[X]\right) \geq 0$ and $u \in M$, we would have $\widetilde{F}_{(Q, w)}^{M}[-X]=M$ if $w^{T} E^{Q}[X] \geq$ 0 and $\widetilde{F}_{(Q, w)}^{M}[-X]=\emptyset$ otherwise. Thus, the result would be an improper risk measure, either being $\equiv \emptyset$ or having $M$ (which corresponds to $-\infty$ ) among its values.

Remark 4.6. Since

$$
u \in F_{(Y, v)}^{M}[X] \quad \Leftrightarrow \quad u \in\left(E^{Q}[X]+G(w)\right) \cap M
$$

with $(Y, v),(Q, w)$ related as in Lemma 4.1, we can see $F_{(Y, v)}^{M}[X]$ as the half space in $M$ which is "not below" $E^{Q}[X]$, where "below" is meant in direction $-w$, the normal of the hyperplane $\left\{u \in \mathbb{R}^{d}: 0=w^{T} u\right\}$. A total analogue for the real-valued expectation would be the boundary of $F_{(Y, v)}^{M}[X]$. However, the functions $X \mapsto \operatorname{bd} F_{(Y, v)}^{M}[X]=\left\{u \in M: E\left[X^{T} Y\right]=v^{T} u\right\}$ do not map into $\mathbb{G}_{M}$. In order to have the same image space for set-valued risk measures and expectations it is more appropriate to use the above constructions.

Remark 4.7. A version of Lemma 4.1 can be used to transform the end point of a consistent price process (see [24, Definition 1.5]) used as a dual variable for no-arbitrage results for markets with proportional transaction costs into a pair $(Q, w) \in \mathcal{M}_{1, d}^{P} \times K_{0}^{+} \backslash\{0\}$ with $\operatorname{diag}(w) \frac{d Q}{d P} \in L_{d}^{q}\left(K_{T}^{+}\right)$. In Schachermayer's model, $M=\mathbb{R}^{d}$ and $K_{T}$ is the random closed convex cone generated by a random bid-ask matrix at terminal time $T$, whereas $K_{0}$ is the (deterministic) closed convex cone modeling the bid-ask spread at initial time $t=0$. The cone $K_{0}$ corresponds to $K$ in this paper.

In order to interpret the set of initial endowments which superreplicate an $\mathbb{R}^{d}$-valued contingent claim as a set-valued coherent risk measure, the theory of the present paper needs to be extended to the case where random variables in $L_{d}^{p}$ are compared by means of the cone $\left\{X \in L_{d}^{p}: X \in K_{T} P\right.$-a.s. $\}$ generated by a measurable multifunction $\omega \mapsto K_{T}(\omega)$ with values in the set of closed convex cones in $\mathbb{R}^{d}$ including $\mathbb{R}_{+}^{d}$ and not being all of $\mathbb{R}^{d}$.
5. Scalar representation. Take $v \in K_{M}^{+}$, and define a function $\varphi_{v}: L_{d}^{p} \rightarrow \mathbb{R} \cup\{ \pm \infty\}$ by

$$
\begin{equation*}
\varphi_{v}(X)=\inf _{u \in R(X)} v^{T} u \tag{5.1}
\end{equation*}
$$

We set, as usual, $\varphi_{v}(X)=+\infty$ if $R(X)=\emptyset$. As a function of $v$ this is nothing other than the negative support function of $R(X)$ at $-v \in \mathbb{R}^{d}$ (in the sense of convex analysis; see, e.g., [26, p. 28]). Since we shall consider $\varphi_{v}$ as a function of $X$, not of $v$, we use the above notation.

Since for $v \in K_{M}^{+}$

$$
\varphi_{v}(X)=\inf _{u \in R(X)} v^{T} u=\inf _{u \in \operatorname{cl} \operatorname{co}\left(R(X)+K_{M}\right)} v^{T} u
$$

we may assume in this section that $R$ maps into $\mathbb{G}_{M}$.

Together with $\varphi_{v}$ we shall consider its closure (or lower semicontinuous envelope [26] or $\Gamma$-regularization [6], but different from the closure in [22]), which is defined by

$$
\left(\operatorname{cl} \varphi_{v}\right)\left(X^{0}\right)=\liminf _{X \rightarrow X^{0}} \varphi_{v}(X)=\sup _{U \in \mathcal{U}} \inf _{X \in U} \varphi_{v}\left(X^{0}+X\right)
$$

for $X^{0} \in L_{d}^{p}$, where $\mathcal{U}$ is a neighborhood base of $0 \in L_{d}^{p}$ for the appropriate topology consisting of convex balanced sets (norm topology for $p \in[1, \infty), \sigma\left(L_{d}^{\infty}, L_{d}^{1}\right)$-topology for $\left.p=\infty\right)$. This is mainly motivated by the fact that closedness of $R$ does not imply closedness of $\varphi_{v}$ in general. A simple counterexample in two dimensions is $R: \mathbb{R} \rightarrow \mathbb{G}_{\mathbb{R}^{2}}$ with $K=K_{M}=\mathbb{R}_{+}^{2}$ defined by $R(X)=\left\{\left(\frac{1}{X}, 0\right)^{T}\right\}+\mathbb{R}_{+}^{2}$ for $X>0$ and $R(X)=\emptyset$ for $X \leq 0$. This function is closed and convex, but $\varphi_{v}$ for $v=(0,1)^{T}$ is convex but not closed.

Apart from closedness, important properties of $R$ can be expressed as properties of the family $\left\{\varphi_{v}: v \in K_{M}^{+} \backslash\{0\}\right\}$. We shall give a selection in the following lemma.

Lemma 5.1. Let $R: L_{d}^{p} \rightarrow \mathbb{G}_{M}$ be a function. Then
(i) if $R$ is convex (positively homogeneous), then $\varphi_{v}$ is convex (positively homogeneous) for each $v \in K_{M}^{+}$;
(ii) if $R$ satisfies (R2), then $\varphi_{v}$ is monotone with respect to $L_{d}^{p}(K)$ for each $v \in K_{M}^{+}$;
(iii) $\operatorname{dom} R=\operatorname{dom} \varphi_{v}$ for each $v \in K_{M}^{+}$;
(iv) if $R$ is convex and closed, then

$$
\begin{equation*}
\forall X \in L_{d}^{p}: R(X)=\bigcap_{v \in K_{M}^{+} \backslash\{0\}}\left\{u \in M:\left(\operatorname{cl} \varphi_{v}\right)(X) \leq v^{T} u\right\} . \tag{5.2}
\end{equation*}
$$

Proof. (i), (ii), and (iii) are well known or easy to see.
(iv) " $\subseteq$ ": Pick $v \in K_{M}^{+} \backslash\{0\}$ and $u^{0} \in R\left(X^{0}\right)$. Since cl $\varphi_{v} \leq \varphi_{v}$ we get

$$
\left(\operatorname{cl} \varphi_{v}\right)\left(X^{0}\right) \leq \varphi_{v}\left(X^{0}\right)=\inf _{u \in R\left(X^{0}\right)} v^{T} u \leq v^{T} u^{0},
$$

as desired.
" $\supseteq$ ": Assume $u^{0} \in M \backslash R\left(X^{0}\right)$. Then there is $U_{0} \in \mathcal{U}$ such that

$$
u^{0} \notin \mathrm{cl} \bigcup_{X \in U_{0}} R\left(X^{0}+X\right)
$$

Since $R$ is convex, the right-hand side is a closed convex set. A separation argument yields $v \in K_{M}^{+} \backslash\{0\}$ such that

$$
v^{T} u^{0}<\inf \left\{v^{T} u: u \in \mathrm{cl} \bigcup_{X \in U_{0}} R\left(X^{0}+X\right)\right\}=\inf _{X \in U_{0}} \varphi_{v}\left(X^{0}+X\right)
$$

Since $\left(\operatorname{cl} \varphi_{v}\right)\left(X^{0}\right)=\sup _{U \in \mathcal{U}} \inf _{X \in U} \varphi_{v}\left(X^{0}+X\right)$ we get $u^{0} \notin\left\{u \in M:\left(\operatorname{cl} \varphi_{v}\right)\left(X^{0}\right) \leq v^{T} u\right\}$, and the proof is now complete.

Interpretation. Since the elements $v \in K_{M}^{+} \backslash\{0\}$ can be seen as internal weights of the investor/regulator for the reference instruments accepted for risk compensation, the value $\varphi_{v}(X)$ is the minimal internal price of the investor/regulator for risk compensation/security deposit for position $X$. If the risk measure $R$ is fixed (which already has to be done by a choice of the investor/regulator), then the investor/regulator has another degree of freedom: Among the elements of $K_{M}^{+} \backslash\{0\}$ she has to choose her internal weight/price system. The investor's minimal risk compensation or the regulator's minimal deposit requirement should be an element $u \in R(X)$ such that $v^{T} u=\varphi_{v}(X)$. Note that this corresponds to a weakly minimal (efficient) point of the set $R(X)$ with respect to the order generated by $K_{M}$ obtained by solving the scalarized problem $\min \left\{v^{T} u: u \in R(X)\right\}$.

The acceptance set of a risk measure can also be characterized using scalarizations. The result reads as follows.

Proposition 5.1. Let $R: L_{d}^{p} \rightarrow \mathbb{G}_{M}$ be a closed convex risk measure with acceptance set $A_{R}$. Then (i) $\varphi_{v}: L_{d}^{p} \rightarrow \mathbb{R} \cup\{ \pm \infty\}$ is a convex function satisfying

$$
\begin{equation*}
\forall u \in M, \forall X \in L_{d}^{p}: \varphi_{v}(X+u \mathbb{I})=\varphi_{v}(X)-v^{T} u, \tag{5.3}
\end{equation*}
$$

and (ii) we have

$$
A_{R}=\bigcap_{v \in K_{M}^{+} \backslash\{0\}} A_{(R, v)}
$$

where $A_{(R, v)}=\left\{X \in L_{d}^{p}:\left(\operatorname{cl} \varphi_{v}\right)(X) \leq 0\right\}$.
Proof. (i) In view of Lemma 5.1 we have to verify only (5.3), which is easily done. Part (ii) follows from (5.2).

Example 5.1 (one-dimensional risk measures for random vectors). In [3], a convex risk measure is defined to be a convex $L_{d}^{\infty}(K)$-decreasing function $\varrho: L_{d}^{\infty} \rightarrow \mathbb{R}$ satisfying

$$
\varrho\left(X+s e^{i}\right)=\varrho(X)-s
$$

for all $i=1, \ldots, d$ and for all $s \in R, X \in L_{d}^{\infty}$. This requirement does not seem to fit into the framework of Definition 2.1 for $M=R^{d}$, and for $M=\mathbb{R}^{1} \times\{0\}^{d-1}$ the above translativity condition is certainly stronger than (R1) in this case. The underlying assumption in [3] is that there is one (and only one) "currency" that can be used to compensate for the risk of a multivariate position $X$.

But another interpretation is possible: Such a risk measures $\varrho$ can be seen as scalarization of a set-valued convex risk measure $R: L_{d}^{\infty} \rightarrow \mathbb{G}_{M}$ with $M=\mathbb{R}^{d}, K_{M}=K$, and $v=$ $(1, \ldots, 1)^{T}$ : If $v$ is an admissible scalarization (which is not always the case - for example, if $d=2, K=\left\{x \in \mathbb{R}^{2}: x_{2} \geq-0.9 x_{1}, x_{2} \geq 0\right\}$; see condition $v \in\left(E[Y]+M^{\perp}\right) \cap K_{M}^{+} \backslash\{0\}=$ $\{E[Y]\} \cap K^{+}$in Lemma 4.1), then $\varphi_{v}$ is a convex risk measure in the sense of Definition 3.1 in [3].

Example 5.2 (scalarization of the upper expectation measure). We look for scalarizations of the coherent risk measure given in Proposition 4.2. Let $Y^{0} \in L_{d}^{q}\left(K^{+}\right)$and $v^{0} \in\left(E\left[Y^{0}\right]+M^{\perp}\right)$ $\cap K_{M}^{+} \backslash\{0\}$. Taking another $v \in K_{M}^{+} \backslash\{0\}$, we get

$$
\varphi_{v}(X)=\inf _{u \in F_{\left(Y^{0}, v^{0}\right)}^{\left.M^{[ }-X\right]}} v^{T} u=\inf \left\{v^{T} u: u \in M, E\left[-X^{T} Y^{0}\right] \leq\left(v^{0}\right)^{T} u\right\} .
$$

If there is $u \in M$ such that $\left(v^{0}\right)^{T} u>0$ and $v^{T} u<0$, then, taking $t u$ for $t>0$ and letting $t \rightarrow \infty$, we obtain $\varphi_{v}(X)=-\infty$. Hence

$$
\varphi_{v}(X)=\left\{\begin{array}{ccc}
E\left[-X^{T} Y^{0}\right] & : & G(v) \cap M=G\left(v^{0}\right) \cap M \\
-\infty & : & \text { otherwise }
\end{array}\right.
$$

This means that $\varphi_{v}$ is a linear function if $v$ and $v^{0}$ generate the same half space in $M$. Any other case does not make sense; there is basically one proper scalarization.

Example 5.3 (scalarization of the negative componentwise expectation). We consider (see Example 3.2) the coherent risk measure on $L_{d}^{1}$,

$$
\operatorname{NCE}(X)=\{u \in M: E[X+u \mathbb{I}] \in K\}=(E[-X]+K) \cap M .
$$

Taking $v \in K_{M}^{+} \backslash\{0\} \subseteq K^{+} \cap M \subseteq K^{+}$, we immediately get

$$
\varphi_{v}(X)=\inf _{u \in N C E(X)} v^{T} u=v^{T} E[-X] .
$$

Example 5.4 (scalarization of the worst case risk measure). The strong version of the worst case risk measure

$$
W C^{M, S}(X)=\left\{u \in M: X+u \mathbb{I} \in L_{d}^{p}(K)\right\}
$$

as defined in Example 3.1, is coherent with scalarization

$$
\varphi_{v}(X)=\inf \left\{v^{T} u: u \in M, X+u \mathbb{I} \in L_{d}^{p}(K)\right\} .
$$

If $v \in K_{M}^{+} \subseteq K^{+}$and since $K=\bigcap_{v \in K^{+}}\left\{x \in \mathbb{R}^{d}: v^{T} x \geq 0\right\}$, we have

$$
\forall u \in W C^{M, S}(X): v^{T} u \geq v^{T}(-X) \quad P \text {-a.s. }
$$

which gives

$$
\varphi_{v}(X)=\left\{\begin{array}{cl}
-\operatorname{ess} . \inf v^{T} X & : \quad R(X) \neq \emptyset \\
+\infty & : \quad \text { otherwise }
\end{array}\right.
$$

## 6. Dual representation.

6.1. Main result. In this section, we shall state and prove dual representation formulas for convex risk measures in terms of set-valued expectations. The following theorem is the main result of the paper. Recall that $\mathbb{G}_{M}=\left\{D \subseteq M: D=\operatorname{cl}\right.$ co $\left.\left(D+K_{M}\right)\right\}$ is the collection of upper closed convex subsets of $M$, and $\mathcal{M}_{1, d}^{P}$ is the set of all vector probability measures with components absolutely continuous with respect to $P$ and $G(w)=\left\{x \in \mathbb{R}^{d}: w^{T} x \geq 0\right\}$.

Theorem 6.1. A function $R: L_{d}^{p} \rightarrow \mathbb{G}_{M}$ is a proper, convex, and closed ( $\sigma\left(L_{d}^{\infty}, L_{d}^{1}\right)$-closed if $p=\infty$ ) risk measure iff there is a function $-\alpha_{R}: \mathcal{M}_{1, d}^{P} \times K^{+} \backslash M^{\perp} \rightarrow \mathbb{G}_{M}$ satisfying
(i) $-\alpha_{R}(Q, w)=\left(-\alpha_{R}(Q, w) \oplus G(w)\right) \cap M$ for $(Q, w) \in \mathcal{M}_{1, d}^{P} \times K^{+} \backslash M^{\perp}$,
(ii) $K_{M} \subseteq \alpha_{R}(Q, w)$ for all $(Q, w) \in \mathcal{W}$,
(iii) $\left(\bigcap_{(Q, w) \in \mathcal{W}}-\alpha_{R}(Q, w)\right) \cap-\operatorname{int} K_{M}=\emptyset$,
where

$$
\mathcal{W}=\left\{(Q, w) \in \mathcal{M}_{1, d}^{P} \times K^{+} \backslash M^{\perp}: \operatorname{diag}(w) \frac{d Q}{d P} \in L_{d}^{q}\left(K^{+}\right)\right\}
$$

such that

$$
\begin{equation*}
\forall X \in L_{d}^{p}: R(X)=\bigcap\left[-\alpha_{R}(Q, w)+\left(E^{Q}[-X]+G(w)\right) \cap M\right] . \tag{6.1}
\end{equation*}
$$

$$
(Q, w) \in \mathcal{W}
$$

In particular, (6.1) is satisfied with $-\alpha_{R}$ replaced by $-\alpha_{\min }$ with

$$
\begin{equation*}
-\alpha_{\min }(Q, w)=\mathrm{cl} \bigcup_{X^{\prime} \in A_{R}}\left(E^{Q}\left[X^{\prime}\right]+G(w)\right) \cap M . \tag{6.2}
\end{equation*}
$$

Moreover, if $-\alpha_{R}: \mathcal{M}_{1, d}^{P} \times K^{+} \backslash M^{\perp} \rightarrow \mathbb{G}_{M}$ satisfies (6.1), then $-\alpha_{R}(Q, w) \subseteq-\alpha_{\min }(Q, w)$ for all $(Q, w) \in \mathcal{W}$.

The function $R$ is a proper, closed ( $\sigma\left(L_{d}^{\infty}, L_{d}^{1}\right)$-closed if $p=\infty$ ), coherent measure of risk iff there is a nonempty set $\mathcal{W}_{R} \subseteq \mathcal{W}$ such that

$$
\begin{equation*}
\forall X \in L_{d}^{p}: R(X)=\bigcap_{(Q, w) \in \mathcal{W}_{R}}\left(E^{Q}[-X]+G(w)\right) \cap M \tag{6.3}
\end{equation*}
$$

In particular, (6.3) is satisfied with $\mathcal{W}_{R}$ replaced by $\mathcal{W}_{\max }$ with

$$
\begin{equation*}
\mathcal{W}_{\max }=\left\{(Q, w) \in \mathcal{M}_{1, d}^{P} \times K^{+} \backslash M^{\perp}: \operatorname{diag}(w) \frac{d Q}{d P} \in A_{R}^{+}\right\} . \tag{6.4}
\end{equation*}
$$

Moreover, if $\mathcal{W}_{R}$ satisfies (6.3), then $\mathcal{W}_{R} \subseteq \mathcal{W}_{\text {max }}$.
The proof is postponed to the end of this section. Note that dual representation theorems for coherent set-valued risk measures can be found in [16], [4], [18]. The results in these references are not formulated in terms of set-valued upper expectations but in terms of scalarizations. Dual representation theorems for convex, but not necessarily coherent, risk measures are not available in the literature so far. It should be mentioned that duality theories for set-valued convex functions in terms of Fenchel conjugates are a very recent achievement. The interested reader may consult [11] for related results and references.

Remark 6.1. Jouini, Meddeb, and Touzi [16] proved a dual representation result for (strongly) closed coherent risk measures on $L_{d}^{\infty}$. There is no principal difficulty in formulating a corresponding result parallel to the above theorem with $(Q, w) \in b a_{d} \times \mathbb{R}^{d}$; compare also [12]. At the end of this section one may find a condition characterizing the $\sigma\left(L_{d}^{\infty}, L_{d}^{1}\right)$-closedness, namely the so-called Fatou property for risk measures on $L_{d}^{\infty}$.

Interpretation. In complete analogy to the scalar case, formula (6.1) can be understood as a robust representation of the risk measure $R$. On the one hand, it involves a set of vector probability measures $Q$ that are absolutely continuous with respect to the reference measure $P$, and which represent alternative scenarios compared to $P$. These scenarios-and this is a new feature in the set-valued case - can be different in different markets. On the other hand, (6.1) involves the vectors $w \in K^{+} \backslash M^{\perp}$, which represent the investor's possible weighting of the $d$ markets. The "penalty function" $-\alpha_{R}$ (see [9, p. 161]) comprises the information on how serious the investor (with risk measure $R$ ) takes the pairs ( $Q, w$ ): The set $-\alpha_{R}(Q, w)$ influences the intersection in (6.1) more, the "fewer" elements it contains. Since $-\alpha_{R}(Q, w)=-\alpha_{R}(Q, w) \oplus G(w) \cap M\left(-\alpha_{R}(Q, w)\right.$ is a half space in $\left.M\right)$, it contains a smaller
portion of $M$ if it is "away from the origin" in $M$, in the direction of some element $u \in K_{M}$ with $w^{T} u=1$, for example. Finally, (6.1) can be seen as a worst case representation of $R$ : An element $u \in R(X)$ must be in $-\alpha_{R}(Q, w)+\left(E^{Q}[-X]+G(w)\right) \cap M$ for all alternative scenarios $Q$ and all corresponding weights $w$. The intersection is a supremum with respect to the order relation $\supseteq$ in $\mathbb{G}_{M}$; see Definition 2.1.

Remark 6.2. A risk measure is a subjective choice of the investor/regulator. If she wants to construct it in terms of scenarios alternative to $P$, (6.1) offers the following possibility: For each market, choose a collection of alternative probability measures $Q_{i}, i=1, \ldots, d$, absolutely continuous with respect to $P$. To each vector probability measure $Q$ having the chosen components assign a point $u_{Q} \in \mathbb{R}^{d}$. Take an internal weight system $w \in K^{+} \backslash M^{\perp}$ such that the coupling condition $\operatorname{diag}(w) \frac{d Q}{d P} \in L_{d}^{q}\left(K^{+}\right)$is satisfied, and let $-\alpha_{R}(Q, w)$ be the half space of $M$ which is obtained by intersecting the closed half space in $\mathbb{R}^{d}$ with normal $-w$ and $u_{Q}$ as boundary point with $M$. Formula (6.1) will produce a closed convex risk measure.

If the vector measure $Q$ seems to be a very likely alternative, $u_{Q}$ should be chosen in int $K$; if it seems very unlikely, in -int $K$. If it seems to be totally irrelevant, $-\alpha_{R}(Q, w)=M$ is the appropriate choice. In the coherent case, only the two choices $u_{Q}=0 \in M$ (hence $\left.-\alpha_{R}(Q, w)=G(w) \cap M\right)$ and $-\alpha_{R}(Q, w)=M$ are possible.

Remark 6.3. The dual representation formulas (6.3), (6.4) for closed coherent risk measures can be transferred into the version of Jouini, Meddeb, and Touzi [16] as follows. If $(Q, w) \in$ $\mathcal{W}_{\text {max }}$, then $\operatorname{diag}(w) \frac{d Q}{d P}$ is the density function of a vector measure $\widetilde{Q} \in c a_{d}$ that is absolutely continuous with respect to $P$, and the set of all such measures form a convex cone $\mathcal{Q} \subseteq c a_{d}$ since $(Q, w) \in \mathcal{W}_{\max }$ iff $(Q, t w) \in \mathcal{W}_{\max }$ for all $t>0$. We know, for $u \in M$,

$$
u \in E^{Q}[-X]+G(w) \Leftrightarrow w^{T} E^{Q}[X+u \mathbb{I}] \geq 0
$$

hence, by (6.3) with $\mathcal{W}_{R}$ replaced by $\mathcal{W}_{\text {max }}, u \in R(X)$ iff

$$
0 \leq \inf _{(Q, w) \in \mathcal{W}_{\max }} w^{T} E^{Q}[X+u \mathbb{I}]=\inf _{\widetilde{Q} \in \mathcal{Q}} \sum_{i=1}^{d} E^{\widetilde{Q}_{i}}\left[X_{i}+u_{i} \mathbb{I}\right]
$$

where $E^{\widetilde{Q}_{i}}[Z]=\int_{\Omega} Z d \widetilde{Q}_{i}$ for $Z \in L^{1}$, which gives an analogue of the dual representation formula of [16, Theorem 4.1] for $\sigma\left(L_{d}^{\infty}, L_{d}^{1}\right)$-closed coherent risk measures on $L_{d}^{\infty}$, and even for a general $M$. Note that the dual variables $\widetilde{Q}$ are not necessarily vector probability measures, which makes the analogy to the scalar case less striking.

Remark 6.4. It is well known that the minimal penalty function $\alpha$ occurring in the dual representation formula of a convex risk measure is its Legendre-Fenchel conjugate; compare [9, Remark 4.17, p. 164], for example. One might expect that the same is true in the set-valued case. Indeed, it will turn out that the right-hand side of (6.2) corresponds to the set-valued Legendre-Fenchel conjugate of $R$, as introduced in [11].

Before we turn to the proof we shall give some examples.
Example 6.1. The first example is the worst case risk measure. Since $A_{W C^{M, S}}=L_{d}^{p}(K)$, Theorem 6.1 yields

$$
W C^{M, S}(X)=\bigcap_{(Q, w) \in \mathcal{W}}\left(E^{Q}[-X]+G(w)\right) \cap M
$$

i.e., the set of dual variables entering the intersection is the largest possible one, which corresponds to the fact that $W C^{M, S}$ is the most risk averse risk measure.

Example 6.2. We investigate negative componentwise expectation (see Example 3.2). The polar cone of its acceptance set can be written as (see below)

$$
\begin{equation*}
A_{N C E}^{+}=\left\{Z \in L_{d}^{\infty}\left(K^{+}\right): \exists y \in \mathbb{R}^{d}: Z=y \mathbb{I}\right\} . \tag{6.5}
\end{equation*}
$$

Therefore, the dual representation of $N C E$ is

$$
\operatorname{NCE}(X)=\bigcap_{(Q, w) \in \mathcal{W}_{N C E}}\left[\left(E^{Q}[-X]+G(w)\right) \cap M\right]
$$

with

$$
\mathcal{W}_{N C E}=\left\{(Q, w) \in \mathcal{W}: \operatorname{diag}(w) \frac{d Q}{d P} \in A_{N C E}^{+}\right\}=\{P\}^{d} \times\left(K^{+} \backslash M^{\perp}\right)
$$

We sketch the proof for (6.5). Define a continuous linear operator $T: \mathbb{R}^{d} \rightarrow L_{d}^{\infty}\left(L_{d}^{\infty}\right.$ is considered with the $\sigma\left(L_{d}^{\infty}, L_{d}^{1}\right)$-topology) by $T y=y \mathbb{I}, y \in \mathbb{R}^{d}$. The adjoint operator $T^{*}$ maps $L_{d}^{1}$ into $\mathbb{R}^{d}$ by $T^{*} X=E[X]$, and we have $A_{N C E}=T^{*-1}(K)=\left\{X \in L_{d}^{1}: T^{*} X \in K\right\}=$ $\left\{X \in L_{d}^{1}: \forall y \in K^{+}: y^{T}\left(T^{*} X\right) \geq 0\right\}$. The latter set coincides with

$$
\begin{aligned}
T\left(K^{+}\right)^{+} & =\left\{X \in L_{d}^{1}: \forall Y \in T\left(K^{+}\right): E\left[X^{T} Y\right] \geq 0\right\} \\
& =\left\{X \in L_{d}^{1}: \forall y \in K^{+}: y^{T}\left(T^{*} X\right) \geq 0\right\}
\end{aligned}
$$

hence $A_{N C E}=T^{*-1}(K)=T\left(K^{+}\right)^{+}$and

$$
A_{N C E}^{+}=T\left(K^{+}\right)^{++}=\operatorname{cl} T\left(K^{+}\right)=\operatorname{cl}\left\{Y \in L_{d}^{\infty}: \exists y \in K^{+}: Y=y \mathbb{I}\right\}
$$

The closure may be dropped since $T\left(K^{+}\right)$is already closed with respect to the $\sigma\left(L_{d}^{\infty}, L_{d}^{1}\right)$ topology. This can be proven using Lemma A. 64 of [9].

Example 6.3. We give an extension of average value at risk to the set-valued case. Let $1 \leq p \leq \infty$ and $\lambda \in(0,1]^{d}$. We use the dual way of definition which gives via Theorem 4.47 of [9] the scalar special case. Define

$$
\begin{equation*}
\mathcal{W}_{\lambda}=\left\{(Q, w) \in \mathcal{W}: \forall i=1, \ldots, d: \frac{d Q_{i}}{d P} \leq \lambda_{i}\right\} \tag{6.6}
\end{equation*}
$$

Since $\mathcal{W}_{\lambda} \subseteq \mathcal{W}$,

$$
A V @ R_{\lambda}(X)=\bigcap_{(Q, w) \in \mathcal{W}_{\lambda}}\left(E^{Q}[-X]+G(w)\right) \cap M
$$

defines a coherent measure of risk on $L_{d}^{p}$ according to Theorem 6.1. A moment's thought will end with $\mathcal{W}_{N C E} \subseteq \mathcal{W}_{\lambda}$, and hence

$$
\forall \lambda \in(0,1]^{d}, \forall X \in L_{d}^{p}: N C E(X) \supseteq A V @ R_{\lambda}(X)
$$

Since $N C E$ satisfies (R0S) we obtain

$$
K_{M} \subseteq A V @ R_{\lambda}(0) \subseteq N C E(0)=K_{M},
$$

and hence $A V @ R_{\lambda}$ satisfies ( R 0 S ), too. The acceptance set of $A V @ R_{\lambda}$ is

$$
A_{\lambda}=\left\{X \in L_{d}^{p}: \forall(Q, w) \in \mathcal{W}_{\lambda}: E^{Q}[X] \in G(w)\right\}
$$

Moreover, since $\mathcal{W}_{\lambda} \supseteq \mathcal{W}_{\mu}$, for all $X \in L_{d}^{p}$ it holds that $A V @ R_{\lambda}(X) \subseteq A V @ R_{\mu}(X)$ whenever $\mu-\lambda \in \mathbb{R}_{+}^{d}$. Finally, the limit $\lambda \rightarrow 0$ goes along the lines of the scalar case, as follows.

Claim. The following holds:

$$
\forall X \in L_{d}^{p}: \bigcap_{\lambda \in(0,1]^{d}} A V @ R_{\lambda}(X)=W C^{M, S}(X) .
$$

Proof of the claim. Obviously, $\bigcap_{\lambda \in(0,1]^{d}} A V @ R_{\lambda}(X) \supseteq N E I^{M, S}(X)$. We shall show the converse. Take $u \in \bigcap_{\lambda \in(0,1]^{d}} A V @ R_{\lambda}(X)$; i.e., $u \in\left[\left(E^{Q}[-X]+G(w)\right) \cap M\right]$ for all $(Q, w) \in \mathcal{W}_{\lambda}, \lambda \in(0,1]^{d}$. We have to show $u \in\left[\left(E^{Q}[-X]+G(w)\right)\right]$ for all $(Q, w) \in \mathcal{W}$. Fix an arbitrary $(Q, w) \in \mathcal{W}$. Since each component $Z_{i}$ of $\frac{d Q}{d P}$ is a nonnegative random variable, there exist, for each $i \in\{1, \ldots, d\}$, a sequence of nonnegative simple functions $\left\{S_{i}^{k}\right\}_{k \in \mathbb{N}}$ with $S_{i}^{k} \rightarrow Z_{i}$ P-almost surely, $E\left[S_{i}^{k}\right]>0$, $\operatorname{diag}(w) S^{k} \in L_{d}^{q}\left(K^{+}\right)$, and $S_{i}^{k} \leq S_{i}^{k+1} \leq Z_{i}$. We have $\left|X_{i} S_{i}^{k}\right|=\left|X_{i}\right| S_{i}^{k} \leq\left|X_{i}\right| Z_{i} \in L^{1}$ since $X \in L_{d}^{p}, Z \in L_{d}^{q}$. From the dominated convergence theorem we get $E\left[X_{i} S_{i}^{k}\right] \rightarrow E\left[X_{i} Z_{i}\right]$. Of course $E\left[S_{i}^{k}\right] \rightarrow E\left[Z_{i}\right]$. Defining $\bar{S}_{i}^{k}:=\frac{S_{i}^{k}}{E\left[S_{i}^{k}\right]}$ and $\bar{w}_{i}^{k}:=w_{i} E\left[S_{i}^{k}\right]$, the $\bar{S}_{i}^{k}$ are densities of measures $\bar{Q}_{i}{ }^{k} \in \mathcal{M}_{1}^{P}$, and we get $\left(\bar{Q}^{k}, \bar{w}^{k}\right) \in$ $\bigcup_{\lambda \in(0,1]^{d}} \mathcal{W}_{\lambda}$. Now, we have $u \in\left[\left(E^{\bar{Q}^{k}}[-X]+G\left(\bar{w}^{k}\right)\right)\right]$, i.e., $\left.\bar{w}^{k T}\left(u+E^{\bar{Q}^{k}}\right)[X]\right) \leq 0$ for all $k \in \mathbb{N}$. Taking the limit $k \rightarrow \infty$, we get $w^{T}\left(u+E^{Q}[X]\right) \leq 0$, i.e., $u \in\left[\left(E^{Q}[-X]+G(w)\right)\right]$.

Remark 6.5. The average value at risk defined above is related to other coherent risk measures defined in the literature. In [16], Jouini, Meddeb, and Touzi introduced a set-valued version of the worst conditional expectation. Their definition, translated into our framework, is

$$
W C E_{\alpha}^{J}=\{u \in M: E[X+u \mathbb{I} \mid B] \in J \forall B \in \mathcal{F} \text { with } P(B)>\alpha\},
$$

where $J \supseteq K$ is a closed convex cone. The corresponding dual description is

$$
W C E_{\alpha}^{J}(X)=\bigcap_{(Q, w) \in \mathcal{W}_{W C E_{\alpha}}}\left(E^{Q}[-X]+G(w)\right) \cap M
$$

with

$$
\mathcal{W}_{W C E_{\alpha}}=\left\{(Q, w) \in \mathcal{W}: \exists B \in \mathcal{F} \text { with } P(B)>\alpha: \forall i=1, \ldots, d: \frac{d Q_{i}}{d P}=\frac{\mathbb{I}_{B}}{P(B)}\right\}
$$

The lower vector valued tail conditional expectation, defined in Bentahar [2], is

$$
T C E_{\alpha}(X)=\{u \in M: E[X+u \mathbb{I} \mid X \in A] \in J \forall A \in \mathcal{Q} \text { with } P(X \in A) \geq \alpha\},
$$

where $\mathcal{Q}=\left\{A \in \mathcal{B}\left(\mathbb{R}^{d}\right): A-K=A\right\}$ with dual description

$$
T C E_{\alpha}(X)=\bigcap_{(Q, w) \in \mathcal{W}_{T C E_{\alpha}}}\left(E^{Q}[-X]+G(w)\right) \cap M
$$

with

$$
\mathcal{W}_{T C E_{\alpha}}=\left\{(Q, w) \in \mathcal{W}: \exists A \in \mathcal{Q} \text { with } P(X \in A) \geq \alpha: \forall i=1, \ldots, d: \frac{d Q_{i}}{d P}=\frac{\mathbb{I}_{\{X \in A\}}}{P(X \in A)}\right\}
$$

The relation between these risk measures is $A V @ R_{\lambda}(X) \subseteq W C E_{\alpha}^{J}(X)$ for all $J \supseteq K$ if $\lambda=(\alpha, \ldots, \alpha)$ and $W C E_{\alpha}^{K}(X) \subseteq T C E_{\alpha}(X)$.

Example 6.4 (entropic risk measure). The entropic risk measure is an important example of a convex but noncoherent risk measure. We shall define a set-valued extension. Take $Q \in \mathcal{M}_{1, d}^{P}$ and define the vector entropy of $Q$ with respect to $P$ by

$$
H(Q \mid P)=\left(\begin{array}{c}
E^{Q_{1}}\left[\log \frac{d Q_{1}}{d P}\right] \\
\vdots \\
E^{Q_{d}}\left[\log \frac{d Q_{d}}{d P}\right]
\end{array}\right)=\left(\begin{array}{c}
H\left(Q_{1} \mid P\right) \\
\vdots \\
H\left(Q_{d} \mid P\right)
\end{array}\right)
$$

Take $\beta \in \operatorname{int} \mathbb{R}_{+}^{d}$ and define $H_{\beta}^{M}(Q, w)=[-\operatorname{diag}(\beta) H(Q \mid P)+G(w)] \cap M$ as a function on $\mathcal{M}_{1, d}^{P} \times K^{+} \backslash M^{\perp}$ into $\mathbb{G}_{M}$. Finally, set

$$
R_{\beta}(X)=\bigcap_{(Q, w) \in \mathcal{W}}\left[H_{\beta}^{M}(Q, w)+\left(E^{Q}[-X]+G(w)\right) \cap M\right] .
$$

We obtain from Theorem 6.1 that $R_{\beta}$ is a (dually defined) closed convex risk measure. If the investor chooses this risk measure, she has an entropic kind but possibly different risk attitude on each market, whereas the overall risk attitude depends on the coupling condition defining elements of $\mathcal{W}$.
6.2. Legendre-Fenchel conjugates of set-valued risk measures. Since dual representation results for convex risk measures are basically an application of the Fenchel-Moreau theorem of convex analysis, we have to introduce Legendre-Fenchel conjugates for set-valued functions. This can be done as follows. Let $R: L_{d}^{p} \rightarrow \mathbb{F}_{M}$ be a function. The Legendre-Fenchel conjugate and the biconjugate of $R$ are defined to be

$$
\begin{aligned}
& -R^{*}(Y, v)=\mathrm{cl} \bigcup_{X \in L_{d}^{p}}\left(R(X)+F_{(Y, v)}^{M}[-X]\right), \quad Y \in L_{d}^{q}, v \in \mathbb{R}^{d}, \\
& R^{* *}(X)=\bigcap_{Y \in L_{d}^{q}, v \in K_{M}^{+} \backslash\{0\}}\left(-R^{*}(Y, v)+F_{(Y, v)}^{M}[X]\right), \quad X \in L_{d}^{p} .
\end{aligned}
$$

Note that in these definitions the $F_{(Y, v)}^{M}$ 's act as substitutes for real-valued linear functionals.

The minus sign in $-R^{*}$ does not (in this paper) belong to an algebraic operation, but indicates that in the scalar case $\left(d=1, M=\mathbb{R}, K=K_{M}=\mathbb{R}_{+}\right)$we get the negative of the usual Fenchel conjugate in the sense of convex analysis by taking the infimum of the values of the corresponding set-valued conjugate. Again, the reason for keeping the minus is that we want to have the same image space for $R$ and its set-valued conjugate. The following simple result shows that the definitions above make sense.

Proposition 6.1. (i) $-R^{*}: L_{d}^{q} \times K_{M}^{+} \rightarrow \mathbb{G}_{M}$. Moreover, $-R^{*}(Y, v)$ is of the form $u+F_{(Y, v)}^{M}[0]$ for some $u \in M$ or an element of $\{M, \emptyset\}$. (ii) $R^{* *}: L_{d}^{p} \rightarrow \mathbb{G}_{M}$.

Proof. (i) We have to show $\mathrm{cl}\left(-R^{*}(Y, v)+K_{M}\right)=-R^{*}(Y, v)$, which is a consequence of the definitions of $-R^{*}$ and $F_{(Y, v)}^{M}$. For the second part, observe that $F_{(Y, v)}^{M}[-X]$ is a shifted closed half space in $M$ with normal vector $v$ (see, for example, Proposition 4.1(iii)), and that for $X \in L_{d}^{p}$ the set $R(X)+F_{[Y, v]}^{M}[-X]$ is a union of such half spaces and hence $-R^{*}(Y, v)$ is by definition the closure of the union of such half spaces or empty.
(ii) We have to show $R^{* *}(X)=\operatorname{cl}$ co $\left(R^{* *}(X)+K_{M}\right)$. This follows from the definitions of $R^{* *}$ and $F_{(Y, v)}^{M}[-X]$.
$R,-R^{*}$, and $R^{* *}$ can be characterized via scalarization. As one might expect, there are relationships to the conjugates of the scalarizations $\varphi_{v}$.

Lemma 6.1. Let $R: L_{d}^{p} \rightarrow \mathbb{F}_{M}$. Then
(i) $-R^{*}(Y, v)=\left\{u \in M:-\left(\varphi_{v}\right)^{*}(Y) \leq v^{T} u\right\}$ for each $Y \in L_{d}^{q}, v \in K_{M}^{+} \backslash\{0\}$;
(ii) $R^{* *}$ is closed and convex, and for all $X \in L_{d}^{p}$

$$
R^{* *}(X)=\bigcap_{v \in K_{M}^{+} \backslash\{0\}}\left\{u \in M:\left(\varphi_{v}\right)^{* *}(X) \leq v^{T} u\right\} .
$$

Proof. (i) First, we compute

$$
\begin{aligned}
\inf _{u \in-R^{*}(Y, v)} v^{T} u & =\inf _{X \in L_{d}^{p}} \inf _{u \in R(X)+F_{(Y, v)}^{M}[-X]} v^{T} u \\
& =\inf _{X \in L_{d}^{p}}\left(\inf _{u^{1} \in R(X)} v^{T} u^{1}+\inf _{u^{2} \in F_{(Y, v)}^{M}[-X]} v^{T} u^{2}\right) \\
& =\inf _{X \in L_{d}^{p}}\left(\varphi_{v}(X)+E\left[-X^{T} Y\right]\right)=-\left(\varphi_{v}\right)^{*}(Y) .
\end{aligned}
$$

The first equation is just the definition of $-R^{*}(Y, v)$, and the support function of a set equals the support function of its closure; the second equation holds since the support function of the sum of two sets equals the sum of the support functions; the third equation is the definition of $\varphi_{v}$ and $F_{(Y, v)}[-X]$; the fourth is the definition of the Fenchel conjugate of the extended real-valued function $\varphi_{v}$ on $L_{d}^{p}$ (see, for example, [26, p. 75]). Immediately, we get that $u \in-R^{*}(Y, v)$ implies $v^{T} u \geq-\left(\varphi_{v}\right)^{*}(Y)$, i.e., " $\subseteq$ " in (i).

To show " $\supseteq$ " in (i), first observe that $-R^{*}(Y, v)=\emptyset$ iff $R(X)=\emptyset$ for all $X \in L_{d}^{p}$. In this case, $\varphi_{v} \equiv+\infty$; hence $-\left(\varphi_{v}\right)^{*} \equiv+\infty$, which means there is no $u \in M$ such that $-\left(\varphi_{v}\right)^{*}(Y) \leq v^{T} u$, so the right-hand side in (i) is void, too. If $-R^{*}(Y, v) \neq \emptyset$, we have $-R^{*}(Y, v)=M$ or $-R^{*}(Y, v)=u+F_{(Y, v)}^{M}[0]$ for some $u \in M$ from Proposition 4.1(iii). In the first case, $-\left(\varphi_{v}\right)^{*}(Y)=-\infty$ (see formula above); hence the right-hand side in (i) is also
$M$. In the second case, if $u^{0} \in M \backslash-R^{*}(Y, v)$, then $v^{T} u^{0}<\inf _{u \in-R^{*}(Y, v)} v^{T} u=-\left(\varphi_{v}\right)^{*}(Y)$. This proves " $\supseteq$ " in (i) for this case.
(ii) By definition of $R^{* *}$ and (i) we have

$$
\begin{aligned}
R^{* *}(X) & =\bigcap_{v \in K_{M}^{+} \backslash\{0\}} \bigcap_{Y \in L_{d}^{q}}\left\{u \in M: v^{T} u \geq E\left[X^{T} Y\right]-\varphi_{v}^{*}(Y)\right\} \\
& =\bigcap_{v \in K_{M}^{+} \backslash\{0\}}\left\{u \in M: v^{T} u \geq \sup _{Y \in L_{d}^{q}}\left(E\left[X^{T} Y\right]-\varphi_{v}^{*}(Y)\right)\right\} \\
& =\bigcap_{v \in K_{M}^{+} \backslash\{0\}}\left\{u \in M: v^{T} u \geq \varphi_{v}^{* *}(X)\right\}
\end{aligned}
$$

This completes the proof of the lemma.
Using Lemma 6.1, one immediately obtains relationships between the Legendre-Fenchel conjugates of $R$ and those of $\varphi_{v}$. The conjugate of the scalarization turns out to be the support function at $-v$ of the values of $-R^{*}$.

Corollary 6.1. Let $R: L_{d}^{p} \rightarrow \mathbb{F}_{M}$ and $v \in K_{M}^{+} \backslash\{0\}$. Then, for $Y \in L_{d}^{q}$,

$$
\left(\varphi_{v}\right)^{*}(Y)=\sup _{u \in-R^{*}(Y, v)}-v^{T} u
$$

Proof. This claim is immediate from the first part of the proof of (i) of Lemma 6.1.
What is special for the Fenchel conjugate of a risk measure? In the scalar case, the conjugate basically coincides with the support function of the acceptance set of the risk measure. The same is true for set-valued risk measures, provided that the right definition for a set-valued support function is used.

Proposition 6.2. Let $R: L_{d}^{p} \rightarrow \mathbb{G}_{M}$ be a convex risk measure with acceptance set $A_{R}$ and $v \in K_{M}^{+} \backslash\{0\}$. Then

$$
-R^{*}(Y, v)=\left\{\begin{array}{clc}
-S_{A_{R}}(Y, v) & : & Y \in-L_{d}^{q}\left(K^{+}\right), v \in E[-Y]+M^{\perp}  \tag{6.7}\\
M & : & \text { else }
\end{array}\right.
$$

where

$$
-S_{A_{R}}(Y, v)=-\left(I_{A_{R}}\right)^{*}(Y, v)=\mathrm{cl} \bigcup_{X \in A_{R}} F_{(Y, v)}^{M}[-X]
$$

If $R$ is additionally positively homogeneous, then

$$
-R^{*}(Y, v)=\left\{\begin{array}{ccc}
G(v) \cap M & : & Y \in-A_{R}^{+}, v \in E[-Y]+M^{\perp}  \tag{6.8}\\
M & : & \text { else. }
\end{array}\right.
$$

Proof. Obviously,

$$
-R^{*}(Y, v) \supseteq \operatorname{cl} \bigcup_{X \in A_{R}}\left(R(X)+F_{(Y, v)}^{M}[-X]\right) \supseteq \mathrm{cl} \bigcup_{X \in A_{R}} F_{(Y, v)}^{M}[-X]
$$

and hence $-R^{*}(Y, v) \supseteq-S_{A_{R}}(Y, v)$ for all $Y \in L_{d}^{q}, v \in \mathbb{R}^{d}$. If $Y \notin-L_{d}^{q}\left(K^{+}\right)$, then there is $X^{0} \in L_{d}^{p}(K)$ such that $E\left[\left(X^{0}\right)^{T} Y\right]>0$. Observing $L_{d}^{p}(K) \subseteq A_{R}$ and using the definition of $F_{(Y, v)}^{M}\left[-t X^{0}\right]$ we obtain

$$
-S_{A_{R}}(Y, v) \supseteq \mathrm{cl} \bigcup_{X \in L_{d}^{p}(K)} F_{(Y, v)}^{M}[-X] \supseteq \bigcup_{t>0} F_{(Y, v)}^{M}\left[-t X^{0}\right]=M .
$$

Hence $-R^{*}(Y, v) \supseteq-S_{A_{R}}(Y, v)=M$ whenever $Y \notin-L_{d}^{q}\left(K^{+}\right)$. Now, assume $E[Y]+v \notin M^{\perp}$ and take $w \in M$. Then

$$
\begin{aligned}
F_{(Y, v)}^{M}[-X-w \mathbb{I}] & =\left\{u \in M: E\left[-X^{T} Y\right] \leq v^{T} u+E[Y]^{T} w\right\} \\
& =\left\{u-w \in M: E\left[-X^{T} Y\right] \leq v^{T}(u-w)+(E[Y]+v)^{T} w\right\}+w \\
& =\left\{u \in M: E\left[-X^{T} Y\right] \leq v^{T} u+(E[Y]+v)^{T} w\right\}+w
\end{aligned}
$$

Since $E[Y]+v \notin M^{\perp}$ for each $u \in M$ we can find $w \in M$ such that $E\left[-X^{T} Y\right] \leq v^{T} u+$ $(E[Y]+v)^{T} w$. Therefore,

$$
\bigcup_{w \in M}\left(F_{(Y, v)}^{M}[-X-w \mathbb{I}]-w\right)=M
$$

and a fortiori

$$
\begin{aligned}
-R^{*}(Y, v) & =\mathrm{cl} \bigcup_{X \in L_{d}^{p}, w \in M}\left(R(X+w \mathbb{I})+F_{(Y, v)}^{M}[-X-w \mathbb{I}]\right) \\
& =\mathrm{cl} \bigcup_{X \in L_{d}^{p}, w \in M}\left(R(X)-w+F_{(Y, v)}^{M}[-X-w \mathbb{I}]\right)=M .
\end{aligned}
$$

It remains to show that $-R^{*}(Y, v) \subseteq-S_{A_{R}}(Y, v)$ for $Y \in-L_{d}^{q}\left(K^{+}\right), E[Y]+v \in M^{\perp}$. Indeed, taking $u \in R(X)$ (hence $X+u \mathbb{I} \in A_{R}$ ), we obtain with Proposition 4.2

$$
-S_{A_{R}}(Y, v) \supseteq F_{(Y, v)}^{M}[-X-u \mathbb{I}]=F_{(Y, v)}^{M}[-X]+u .
$$

Therefore, $R(X)+F_{(Y, v)}^{M}[-X] \subseteq-S_{A_{R}}(Y, v)$ for all $X \in L_{d}^{p}$, and hence $-R^{*}(Y, v) \subseteq$ $-S_{A_{R}}(Y, v)$.

Now, let $R$ be additionally positively homogeneous. If $Y \notin-A_{R}^{+}$, then there is $X^{0} \in A_{R}$ such that $E\left[Y^{T} X^{0}\right]>0$, which yields $\bigcup_{X \in A_{R}} F_{(Y, v)}^{M}[-X] \supseteq \bigcup_{t>0} F_{(Y, v)}^{M}\left[-t X^{0}\right]=M$, and a fortiori $-R^{*}(Y, v)=M$ by (6.7). If $Y \in-A_{R}^{+}$, then $E\left[Y^{T} X\right] \leq 0$ for all $X \in A_{R}$; hence $F_{(Y, v)}^{M}[-X] \subseteq F_{(Y, v)}^{M}[0]$ for all $X \in A_{R}$, and therefore, since $0 \in A_{R}$,

$$
F_{(Y, v)}^{M}[0] \subseteq \bigcup_{X \in A_{R}} F_{(Y, v)}^{M}[-X] \subseteq F_{(Y, v)}^{M}[0] .
$$

This completes the proof of the proposition.
6.3. Proof of Theorem 6.1. The proof will be given with the help of a couple of auxiliary lemmas.

Lemma 6.2. Let $R: L_{d}^{p} \rightarrow \mathbb{G}_{M}$ be proper closed convex. Then, for each $\left(X^{0}, u^{0}\right) \notin \operatorname{graph} R$ there is $v \in K_{M}^{+} \backslash\{0\}$ such that $\varphi_{v}: L_{d}^{p} \rightarrow \mathbb{R} \cup\{+\infty\}$ has a continuous affine minorant, and moreover, $\left(\operatorname{cl} \varphi_{v}\right)\left(X^{0}\right)>v^{T} u^{0}$.

Proof. The point $\left(X^{0}, u^{0}\right)$ can be strongly separated from the closed convex set graph $R$ by means of $(Y, v) \in L_{d}^{q} \times M$ and $r \in \mathbb{R}$ such that

$$
\begin{equation*}
E\left[Y^{T} X^{0}\right]+v^{T} u^{0}<r<\inf _{(X, u) \in \operatorname{graph} R}\left(E\left[Y^{T} X\right]+v^{T} u\right) \tag{6.9}
\end{equation*}
$$

Then $v \in K_{M}^{+}$, since otherwise the right-hand side of (6.9) would be $-\infty$.
Take $X \in \operatorname{dom} R$. Then

$$
\begin{aligned}
\varphi_{v}(X) & =\inf _{u \in R(X)} v^{T} u=\inf _{u \in R(X)}\left(E\left[Y^{T} X\right]+v^{T} u\right)-E\left[Y^{T} X\right] \\
& \geq \inf _{\left(X^{\prime}, u^{\prime}\right) \in \operatorname{graph} R}\left(E\left[Y^{T} X^{\prime}\right]+v^{T} u^{\prime}\right)-E\left[Y^{T} X\right] \\
& >-E\left[Y^{T} X\right]+r>E\left[Y^{T}\left(X^{0}-X\right)\right]+v^{T} u^{0} .
\end{aligned}
$$

The function $X \mapsto E\left[Y^{T}\left(X^{0}-X\right)\right]+r$ is a continuous affine minorant of $\varphi_{v}$ since $\varphi_{v}(X)>$ $E\left[Y^{T}\left(X^{0}-X\right)\right]+r$ is all the more true for $X \notin \operatorname{dom} R$.

If $v \neq 0$ (which is automatic if $X^{0} \in \operatorname{dom} R$ ), we obtain a continuous affine minorant of $\mathrm{cl} \varphi_{v}$ which also satisfies $\left(\operatorname{cl} \varphi_{v}\right)\left(X^{0}\right)>v^{T} u^{0}$.

If $v=0$, (6.9) yields

$$
\begin{equation*}
E\left[Y^{T} X^{0}\right]<r<\inf _{X \in \operatorname{dom} R} E\left[Y^{T} X\right] \tag{6.10}
\end{equation*}
$$

Since $R$ is proper there are $X^{1} \in \operatorname{dom} R$ and $u^{1} \in M \backslash R\left(X^{1}\right)$, and the point ( $X^{1}, u^{1}$ ) can be separated from graph $R$ by means of $(Z, w) \in L_{d}^{q} \times M$ and $s \in \mathbb{R}$ such that

$$
\begin{equation*}
E\left[Z^{T} X^{1}\right]+w^{T} u^{1}<s<\inf _{(X, u) \in \operatorname{graph} R}\left(E\left[Z^{T} X\right]+w^{T} u\right) \tag{6.11}
\end{equation*}
$$

Again, $w \in K_{M}^{+}$. Moreover, $w \neq 0$, since otherwise (choose $X=X^{1}$ on the right-hand side) (6.11) leads to a contradiction.

Define $v_{t}=t w$ and $Y_{t}=t Z+(1-t) Y$ for $t \in(0,1)$. Then $v_{t} \in K_{M}^{+} \backslash\{0\}$. Set $\alpha_{1}=$ $E\left[Y^{T} X^{0}\right]$ and $\alpha_{2}=E\left[Z^{T} X^{0}\right]+w^{T} u^{0}$. Then $\alpha_{1}<r$ because of (6.10). Hence, there is $\bar{t}>0$ such that $\alpha_{1}+\bar{t}\left(\alpha_{2}-\alpha_{1}\right)<r+\bar{t}(s-r)$. This gives

$$
\begin{equation*}
E\left[Y_{\bar{t}}^{T} X^{0}\right]+v_{\bar{t}}^{T} u^{0}<\bar{t} s+(1-\bar{t}) r<\inf _{(X, u) \in \operatorname{graph} R}\left(E\left[Y_{\bar{t}}^{T} X\right]+v_{\bar{t}}^{T} u\right) \tag{6.12}
\end{equation*}
$$

The same argument as in the case $v \neq 0$ produces a continuous affine minorant of $\varphi_{v_{\bar{t}}}$ and $\left(\mathrm{cl} \varphi_{v_{\bar{t}}}\right)\left(X^{0}\right)>v_{\bar{t}}^{T} u^{0}$.

Lemma 6.3. If $R: L_{d}^{p} \rightarrow \mathbb{G}_{M}$ is proper convex closed, then $R=R^{* *}$.

Proof. Taking into account Lemmas 5.1(iv) and 6.1(ii), we have to show that

$$
\begin{equation*}
\bigcap_{v \in K_{M}^{+} \backslash\{0\}}\left\{u \in M:\left(\operatorname{cl} \varphi_{v}\right)(X) \leq v^{T} u\right\}=\bigcap_{v \in K_{M}^{+} \backslash\{0\}}\left\{u \in M:\left(\varphi_{v}\right)^{* *}(X) \leq v^{T} u\right\} \tag{6.13}
\end{equation*}
$$

holds true for all $X \in L_{d}^{p}$. We have

$$
\bigcap_{v \in K_{M}^{+} \backslash\{0\}}\left\{u \in M:\left(\operatorname{cl} \varphi_{v}\right)(X) \leq v^{T} u\right\}=\bigcap_{\substack{v \in K_{M}^{+} \backslash\{0\}, \\ \text { cl } \varphi_{v} \text { proper }}}\left\{u \in M:\left(\operatorname{cl} \varphi_{v}\right)(X) \leq v^{T} u\right\}
$$

since " $\subseteq$ " is obvious and " $\supseteq$ " follows since if $u^{0} \in M$ does not belong to the left-hand side, then by Lemma 6.2 there is $v \in K_{M}^{+} \backslash\{0\}$ such that $\operatorname{cl} \varphi_{v}$ is proper and $v^{T} u^{0}<\left(\operatorname{cl} \varphi_{v}\right)(X)$; i.e., $u^{0}$ does not belong to the right-hand side.

Moreover, obviously

$$
\bigcap_{v \in K_{M}^{+} \backslash\{0\}}\left\{u \in M: v^{T} u \geq\left(\varphi_{v}\right)^{* *}(X)\right\}=\bigcap_{\substack{v \in K^{+} \backslash\{0\},\left(\varphi_{v}\right)^{* *} \neq-\infty}}\left\{u \in M: v^{T} u \geq\left(\varphi_{v}\right)^{* *}(X)\right\}
$$

Since $\left(\varphi_{v}\right)^{* *} \equiv-\infty$ iff $\operatorname{cl} \varphi_{v}$ is not proper, and $\left(\varphi_{v}\right)^{* *}=\operatorname{cl} \varphi_{v}$ if $\operatorname{cl} \varphi_{v}$ is proper, the result follows.

The dual representation theorem for set-valued convex risk measures can now be proven. Proof of Theorem 6.1. From Lemma 6.3 we know $R=R^{* *}$; hence for all $X \in L_{d}^{p}$

$$
R(X)=R^{* *}(X)=\bigcap_{Y \in L_{d}^{q}, v \in K_{M}^{+} \backslash\{0\}}\left(-R^{*}(Y, v)+F_{(Y, v)}^{M}[X]\right) .
$$

Replacing $Y$ by $-Y$ and observing $F_{(Y, v)}^{M}[X]=F_{(-Y, v)}^{M}[-X]$, we obtain from Proposition 6.2

$$
R^{* *}(X)=\bigcap_{Y \in L_{d}^{q}\left(K^{+}\right), v \in\left(E[Y]+M^{\perp}\right) \cap K_{M}^{+} \backslash\{0\}}\left(\mathrm{cl} \bigcup_{X^{\prime} \in A_{R}} F_{(Y, v)}^{M}\left[X^{\prime}\right]+F_{(Y, v)}^{M}[-X]\right)
$$

for all $X \in L_{d}^{p}$. By Lemma 4.1 we can replace $Y \in L_{d}^{q}\left(K^{+}\right)$and $v \in\left(E[Y]+M^{\perp}\right) \cap K_{M}^{+} \backslash\{0\}$ by a vector probability measure $Q$ and some $w \in K^{+} \backslash M^{\perp}$ such that $\operatorname{diag}(w) \frac{d Q}{d P} \in L_{d}^{q}\left(K^{+}\right)$, and $F_{(Y, v)}^{M}[X]=\left(E^{Q}[X]+G(w)\right) \cap M$ holds for all $X \in L_{d}^{p}$. Substituting the corresponding $(Q, w)$ for $(Y, v)$ into the above formula, we obtain the dual representation (6.1) with

$$
-\alpha_{R}(Q, w)=-\alpha_{\min }(Q, w)=\mathrm{cl} \bigcup_{X^{\prime} \in A_{R}}\left(E^{Q}\left[X^{\prime}\right]+G(w)\right) \cap M .
$$

The reader may check that in this case $-\alpha_{R}(Q, w)=\left(-\alpha_{R}(Q, w) \oplus G(w)\right) \cap M$ holds true. Moreover, $R$ satisfies (R0) iff $K_{M} \subseteq \alpha_{R}(Q, w)$ for all $(Q, w) \in \mathcal{W}$ and $\left(\bigcap_{(Q, w) \in \mathcal{W}}-\alpha_{R}(Q, w)\right) \cap$ $-\operatorname{int} K_{M}=\emptyset$.

Conversely, if (6.1) holds true for some $-\alpha_{R}$ having the latter property, then for all $X \in L_{d}^{p}$, for all $(Q, w) \in \mathcal{W}$,

$$
R(X) \subseteq-\alpha_{R}(Q, w)+\left(E^{Q}[-X]+G(w)\right) \cap M
$$

Adding $\left(E^{Q}[X]+G(w)\right) \cap M$ on both sides, we get

$$
R(X)+\left(E^{Q}[X]+G(w)\right) \cap M \subseteq-\alpha_{R}(Q, w)
$$

for all $X \in L_{d}^{p}$, for all $(Q, w) \in \mathcal{W}$. Replacing $(Q, w) \in \mathcal{W}$ by $(-Y, v)$ with $Y \in-L_{d}^{q}\left(K^{+}\right)$, $v \in\left(E[Y]+M^{\perp}\right) \cap K_{M}^{+} \backslash\{0\}$, using the definition of $-R^{*}$ as well as (6.7), and re-substituting $(Q, w)$ for $(Y, v)$ in $-R^{*}$, we obtain (6.2).

The coherent case can be dealt with using the corresponding part of Proposition 6.2.
6.4. The Fatou property for set-valued risk measures. One main question concerning risk measures on $L^{\infty}$ is under what circumstances a dual representation is possible based on a dual pairing of $L_{d}^{\infty}$ and $L_{d}^{1}$, i.e., when $\sigma\left(L_{d}^{\infty}, L_{d}^{1}\right)$-closedness of the graph is present. An answer can be given that is parallel to the scalar case: A so-called Fatou property for set-valued functions can be formulated and proven to be equivalent to $\sigma\left(L_{d}^{\infty}, L_{d}^{1}\right)$-closedness. The result reads as follows.

Theorem 6.2. The following are equivalent for a convex risk measure $R: L_{d}^{\infty} \rightarrow \mathbb{G}_{M}$ :
(i) $R$ has a dual representation (6.1);
(ii) $R$ has the Fatou property; i.e., if $\left\{X^{k}\right\}_{k \in \mathbb{N}} \subseteq L_{d}^{\infty}$ is a bounded sequence with $X^{k} \rightarrow X$ $P$-almost surely, then

$$
R(X) \supseteq \liminf _{k \rightarrow \infty} R\left(X^{k}\right)=\left\{u \in M: \forall k \in \mathbb{N}: \exists u^{k} \in R\left(X^{k}\right): \lim _{k \rightarrow \infty} u^{k}=u\right\} ;
$$

(iii) graph $R$ is closed in the product topology on $L_{d}^{\infty} \times M$ generated by $\sigma\left(L_{d}^{\infty}, L_{d}^{1}\right)$ and the usual separated locally convex topology on $M$;
(iv) $A_{R}$ is $\sigma\left(L_{d}^{\infty}, L_{d}^{1}\right)$-closed.

Proof. (iii) implies (i): Since $R$ is proper convex and $\sigma\left(L_{d}^{\infty}, L_{d}^{1}\right)$-closed, we can apply Theorem 6.1 for $L_{d}^{\infty}$ with $\sigma\left(L_{d}^{\infty}, L_{d}^{1}\right)$ as topology, getting the desired representation.
(i) implies (ii): Denote $T_{(Q, w)}=\mathrm{cl} \bigcup_{X \in A_{R}} \tilde{F}_{(Q, w)}^{M}[X]=\mathrm{cl} \bigcup_{X \in A_{R}}\left(E^{Q}[X]+G(w)\right) \cap$ M. Take a bounded sequence $\left\{X^{k}\right\}_{k \in \mathbb{N}} \subseteq L_{d}^{\infty}$ with $X^{k} \rightarrow X P$-almost surely and $u \in$ $\liminf _{k \rightarrow \infty} R\left(X^{k}\right)$. Then there is a sequence $u^{k} \in R\left(X^{k}\right)$ such that $u^{k} \rightarrow u$ in $M$. Because of (i) we have

$$
\forall k \in \mathbb{N}, \forall(Q, w) \in \mathcal{W}: u^{k} \in \tilde{F}_{(Q, w)}^{M}\left[-X^{k}\right]+T_{(Q, w)}
$$

We have to show $u \in R(X)$, i.e., $u \in \tilde{F}_{(Q, w)}^{M}[-X]+T_{(Q, w)}$ for all $(Q, w) \in \mathcal{W}$. If $T_{(Q, w)}=M$, then there is nothing to prove. If $T_{(Q, w)} \neq M$, then there is $\widehat{u} \in M$ such that $T_{(Q, w)}=$ $\widehat{u}+G(w) \cap M$; hence

$$
\forall k \in \mathbb{N}, \forall(Q, w) \in \mathcal{W}: u^{k} \in \tilde{F}_{(Q, w)}^{M}\left[-X^{k}\right]+\widehat{u}+G(w) \cap M \subseteq \tilde{F}_{(Q, w)}^{M}\left[-X^{k}\right]+\widehat{u}
$$

This means $u^{k}-\widehat{u} \in M$ for all $k \in \mathbb{N}$ and

$$
\forall k \in \mathbb{N}, \forall(Q, w) \in \mathcal{W}: w^{T}\left(u^{k}-\widehat{u}+E^{Q}\left[X^{k}\right]\right) \geq 0
$$

Since $X^{k}$ is bounded and converges $P$-almost surely to $X$, we have $E^{Q}\left[X^{k}\right] \rightarrow E^{Q}[X]$ for each $Q \in \mathcal{M}_{1, d}^{P}$ by Lebesgue's dominated convergence theorem; hence taking the limit $k \rightarrow \infty$ gives the desired result.
(ii) implies (iv): According to Lemma A. 64 in [9] it suffices to show that the sets

$$
A_{R}^{r}=A_{R} \cap\left\{X \in L_{d}^{\infty}:\|X\|_{L_{d}^{\infty}} \leq r\right\}, \quad r>0,
$$

are closed in $L_{d}^{1}$. Take a sequence $\left\{X^{k}\right\}_{k \in \mathbb{N}} \subseteq A_{R}^{r}$ with $X^{k} \rightarrow X$ in $L_{d}^{1}$. Then there is a subsequence, again denoted by $\left\{X^{k}\right\}_{k \in \mathbb{N}}$, that converges $P$-almost surely to $X$. The Fatou property ensures that $X \in A_{R}$. The $P$-almost surely convergence and boundedness of the $X^{k}$, s ensure $X \in A_{R}^{r}$. Hence $A_{R}^{r}$ is closed in $L_{d}^{1}$ and $A_{R}$ is $\sigma\left(L_{d}^{\infty}, L_{d}^{1}\right)$-closed.
(iv) implies (iii): Compare Remark 3.1.

## REFERENCES

[1] J.-P. Aubin and H. Frankowska, Set-Valued Analysis, Birkhäuser-Verlag, Basel, Switzerland, 1990.
[2] I. Bentahar, Tail conditional expectation for vector-valued risks, Sonderforschungsbereich Discussion Paper 2006-029, Humboldt Universität Berlin, Berlin, 2006.
[3] C. Burgert and L. Rüschendorf, Consistent risk measures for portfolio vectors, Insurance Math. Econom., 38 (2006), pp. 289-297.
[4] I. Cascos and I. Molchanov, Multivariate risks and depth-trimmed regions, Finance Stoch., 11 (2007), pp. 373-397.
[5] P. Embrechts and G. Puccetti, Bounds for functions of multivariate risks, J. Multivariate Anal., 97 (2006), pp. 526-547.
[6] I. Ekeland and R. Temam, Convex Analysis and Variational Problems, Stud. Math. Appl. Vol. 1, North-Holland, Amsterdam, Elsevier, New York, 1976.
[7] European Central Bank, Measures to further expand the collateral framework and enhance the provision of liquidity, press release, October 15, 2008, available online at http://www.ecb.int/press/pr/ date $/ 2008 / \mathrm{html} / \mathrm{pr081015} . \mathrm{en} . \mathrm{html}$.
[8] European Central Bank, Further technical specifications for the temporary expansion of the collateral framework, press release November 12, 2008, available online at http://www.ecb.int/press/pr/ date/2008/html/pr081112_ 1.en.html.
[9] H. Föllmer and A. Schied, Stochastic Finance. An Introduction in Discrete Time, 2nd ed., de Gruyter, Berlin, Germany, 2004.
[10] A. Göpfert, H. Riahi, C. Tammer, and C. Zălinescu, Variational Methods in Partially Ordered Spaces, CMS Books in Math. 17, Springer-Verlag, New York, 2003.
[11] A. Hamel, A duality theory for set-valued functions I: Fenchel conjugation theory, Set-Valued and Variational Anal., 17 (2009), pp. 153-182.
[12] A. Hamel, F. Heyde, and M. Höhne, Set-valued measures of risk, Report on Optimization and Stochastics, Report 15, Martin-Luther-Universität Halle-Wittenberg, Halle, Germany, 2007; available online at http://webdoc.sub.gwdg.de/ebook/serien/e/reports_Halle-Wittenberg_math/07-15report.pdf.
[13] A. Hamel and B. Rudloff, Continuity and finite-valuedness of set-valued risk measures, in Festschrift in Celebration of Prof. Dr. Wilfried Grecksch's 60th Birthday, C. Tammer and F. Heyde, eds., ShakerVerlag, Aachen, Germany, 2008, pp. 49-64.
[14] F. Heyde and A. Löhne, Solution concepts in vector optimization. A fresh look at an old story, Report on Optimization and Stochastics, Report 19, Martin-Luther-Universität HalleWittenberg, Halle, Germany, 2008; available online at http://www.mathematik.uni-halle.de/reports/ sources/2008/08-19report.pdf.
[15] S. Jaschke and U. Küchler, Coherent risk measures and good-deal bounds, Finance Stoch., 6 (2001), pp. 181-200.
[16] E. Jouini, M. Meddeb, and N. Touzi, Vector-valued coherent risk measures, Finance Stoch., 8 (2004), pp. 531-552.
[17] Y. M. Kabanov, Hedging and liquidation under transaction costs in currency markets, Finance Stoch., 3 (1999), pp. 237-248.
[18] A. V. Kulikov, Multidimensional convex and coherent risk measures, Theory Probab. Appl., 52 (2008), pp. 614-635.
[19] A. Löhne and C. Tammer, A new approach to duality in vector optimization, Optimization, 56 (2007), pp. 221-239.
[20] G. Nappo and F. Spizzichino, Kendall distributions and level sets in bivariate exchangeable survival models, Inform. Sci., 179 (2009), pp. 2878-2890.
[21] A. L. Peressini, Ordered Topological Vector Spaces, Harper's Series in Modern Mathematics, Harper \& Row, New York, 1967.
[22] R. T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, NJ, 1970.
[23] L. Rüschendorf, Law invariant convex risk measures for portfolio vectors, Statistics and Decisions, 24 (2006), pp. 97-108.
[24] W. Schachermayer, The fundamental theorem of asset pricing under proportional transaction costs in finite discrete time, Math. Finance, 14 (2004), pp. 19-48.
[25] T. Tanino, On the supremum of a set in a multidimensional space, J. Math. Anal. Appl., 130 (1988), pp. 386-397.
[26] C. ZĂLinescu, Convex Analysis in General Vector Spaces, World Scientific, Singapore, 2002.

# Continuous-Time Markowitz's Model with Transaction Costs* 

Min Dai ${ }^{\dagger}$, Zuo Quan $\mathrm{Xu}^{\ddagger}$, and Xun Yu Zhou ${ }^{\S}$


#### Abstract

A continuous-time Markowitz's mean-variance portfolio selection problem is studied in a market with one stock, one bond, and proportional transaction costs. This is a singular stochastic control problem, inherently with a finite time horizon. Via a series of transformations, the problem is turned into a so-called double obstacle problem, a well-studied problem in physics and PDE literature, featuring two time-varying free boundaries. The two boundaries, which define the buy, sell, and no-trade regions, are proved to be smooth in time. This in turn characterizes the optimal strategy, via a Skorokhod problem, as one that tries to keep a certain adjusted bond-stock position within the no-trade region. Several features of the optimal strategy are revealed that are remarkably different from its no-transaction-cost counterpart. It is shown that there exists a critical length in time, which is dependent on the stock excess return as well as the transaction fees but independent of the investment target and the stock volatility, so that an expected terminal return may not be achievable if the planning horizon is shorter than that critical length (while in the absence of transaction costs any expected return can be reached in an arbitrary period of time). It is further demonstrated that anyone following the optimal strategy should not buy the stock beyond the point when the time to maturity is shorter than the aforementioned critical length. Moreover, the investor would be less likely to buy the stock and more likely to sell the stock when the maturity date is getting closer. These features, while consistent with the widely accepted investment wisdom, suggest that the planning horizon is an integral part of the investment opportunities.


Key words. continuous-time, mean-variance, transaction costs, singular stochastic control, planning horizon, Lagrange multiplier, double-obstacle problem, Skorokhod problem

AMS subject classifications. 93E20, 47J20, 49L20
DOI. 10.1137/080742889

1. Introduction. Markowitz's (single-period) mean-variance (MV) portfolio selection model [27] marked the start of modern quantitative finance theory. Extensions to the dynamicespecially continuous-time - setting in the asset allocation literature have, however, been dominated by the expected utility maximization (EUM) models, which are a considerable departure

[^4]from the MV model. While the utility approach was theoretically justified by von Neumann and Morgenstern [33], in practice "few if any investors know their utility functions; nor do the functions which financial engineers and financial economists find analytically convenient necessarily represent a particular investor's attitude towards risk and return" [28]. Meanwhile, there are technical and conceptual difficulties in studying a dynamic MV model. In particular, an optimal trading strategy generated initially may no longer be optimal halfway through. This so-called time inconsistency means that dynamic programming-which is the main tool for solving dynamic optimization problems - is not directly applicable. Furthermore, one could argue that it would be hard for an investor to follow a time-inconsistent strategy. Kydland and Prescott [19] instead argue that time-inconsistent solutions are economically meaningful if the investor can commit at the initial time to follow a strategy (called a precommitted strategy).

Time-inconsistent control problems have recently attracted some interest; see Björk and Murgoci [4] and Ekeland and Lazrak [13]. We note also that there are problems other than dynamic MV analysis that are inherently time inconsistent. For example, a dynamic behavioral portfolio selection problem is time inconsistent due to the distortions in probabilities [16].

Basak and Chabakauri [2] specifically address the time inconsistency problem in MV analysis by proposing the construction of a trading strategy that is locally optimal in an MV sense and time consistent, although it is not globally optimal in the sense of Problem 2.1 (to be formulated in section 2). Basak and Chabakauri [2] further show that their strategy solves a global optimization problem with a state-dependent CARA utility function.

In this paper, we solve the global Problem 2.1 and do so when trading is subject to transaction costs. The solution obtained is precommitted, instead of time consistent. We solve the problem by reformulating it in a way that makes it amenable to dynamic programming.

Richardson [30] is probably the earliest paper that studies a faithful extension of the MV model to the continuous-time setting (albeit in the context of a single stock with a constant risk-free rate), followed by Bajeux-Besnainou and Portait [1]. Li and Ng [20], in a discrete-time setting, developed an embedding technique for changing the originally time-inconsistent MV problem into a stochastic linear-quadratic (LQ) control problem. This technique was extended by Zhou and Li [35], along with a stochastic LQ control approach, to the continuous-time case. ${ }^{1}$ Further extensions and improvements are carried out by, among many others, Lim and Zhou [23], Lim [22], Bielecki et al. [3], and Xia [34].

All the existing works on continuous-time MV models have assumed that there is no transaction cost, leading to results that are analytically elegant, and sometimes truly surprising (for example, it is shown in Li and Zhou [21] that any efficient strategy realizes its goal-no matter how high it is-with a probability of at least $80 \%$ ). However elegant they may be, certain investment behaviors derived from the results simply contradict the conventional wisdom, which in turn hints that the models may not have been properly formulated. For instance, the results dictate that an optimal strategy must trade all the time; moreover, there must be risky exposures at any time (see [5]). These are certainly not consistent with the common investment advice. Indeed, the assumption that there is no transaction cost is flawed, which misleadingly allows an investor to continuously trade without any penalty.

[^5]Portfolio selection subject to transaction costs has been studied extensively, albeit in the realm of utility maximization. Mathematically such a problem is a singular stochastic control problem. Two different types of models must be distinguished: one in an infinite planning horizon and the other in a finite horizon. See Magill and Constantinides [26], Davis and Norman [10], and Shreve and Soner [31] for the former, and Davis, Panas and Zariphopoulou [11], Cvitanic and Karatzas [6], and Gennotte and Jung [15] for the latter. Technically, the latter is substantially more difficult than the former, since in the finite-horizon case there is an additional time variable in the related Hamilton-Jacobi-Bellman (HJB) equation or variational inequality (VI). This is why the research on finite-horizon problems had predominantly addressed qualitative and numerical solutions until Liu and Loewenstein [25] devised an analytical approach based on an approximation of the finite horizon by a sequence of Erlang distributed random horizons. Dai and Yi [8] subsequently employed a different analytical approach-a PDE approach - to study the same problem.

This paper aims to analytically solve the MV model with transaction costs. Note that such a problem inherently occurs in a finite time horizon, because the very nature of the Markowitz problem is about striking a balance between the risk and return of the wealth at a finite, terminal time. Compared with its EUM counterpart, this model has a feasibility issue that must be addressed before an optimal solution is sought. Precisely speaking, the MV model is to minimize the variance of the terminal wealth subject to the constraint that an investment target - certain expected net terminal wealth-is achieved. The feasibility is about whether such a target is achievable by at least one admissible investment strategy. For a Black-Scholes market without transaction costs, it has been shown [23] that any target can be reached in an arbitrary length of time (so long as the risk involved is not a concern, that is). For a more complicated model with random investment opportunities and a no-bankruptcy constraint, the feasibility is painstakingly investigated in Bielecki et al. [3]. In this paper we show that the length of the planning horizon is a determinant of this issue. In fact, there exists a critical length of time, which is dependent only on the stock excess return and the transaction fees, so that a sufficiently high target is not achievable if the planning horizon is shorter than that critical length. This certainly makes good sense intuitively.

To obtain an optimal strategy, technically we follow the idea of Dai and Yi [8] of eventually turning the associated VI into a double-obstacle problem, a problem that has been well studied in physics and PDE theory. That said, there are indeed intriguing subtleties when actually carrying it out. In particular, this paper is the first to prove (to the best of the authors' knowledge) that the two free boundaries that define the buy, sell, and no-trade regions are smooth. This smoothness is critical in deriving the optimal strategy via a Skorokhod problem. ${ }^{2}$ The optimal strategy is rather simple in implementation; it is to keep a certain adjusted bond-stock position within the no-trade region. Several features of the optimal strategy are revealed that are remarkably different from its no-transaction-cost counterpart. Among these, it is notable that one should no longer buy stock beyond the point when the time to maturity is shorter than the aforementioned critical length associated with the feasibility. Moreover,

[^6]one is less likely to buy the stock and more likely to sell the stock when the maturity date is getting closer. These results are consistent with the widely accepted financial advice and suggest that the planning horizon should be regarded as a part of the investment opportunity set when it comes to continuous-time portfolio selection.

The remainder of the paper is organized as follows. The model under consideration is formulated in section 2, and the feasibility issue is addressed in section 3 . The optimal strategy is derived in sections 4-6 via several steps, including Lagrange relaxation, transformation of the HJB equation to a double-obstacle problem, and the Skorokhod problem. Finally, the paper concludes with remarks in section 7. Some technical proofs are relegated to Appendices A and B .
2. Problem formulation. We consider a continuous-time market where there are only two investment instruments: a bond and a stock with price dynamics given, respectively, by

$$
\begin{aligned}
& \mathrm{d} R(t)=r R(t) \mathrm{d} t \\
& \mathrm{~d} S(t)=\alpha S(t) \mathrm{d} t+\sigma S(t) \mathrm{d} B(t) .
\end{aligned}
$$

Here $r>0, \alpha>r$, and $\sigma>0$ are constants, and the process $\{B(t)\}_{t \in[0, T]}$ is a standard one-dimensional Brownian motion on a filtered probability space $\left(\Omega, \mathcal{F},\left\{\mathcal{F}_{t}\right\}_{t \in[0, T]}, \mathbf{P}\right)$ with $B(0)=0$ almost surely. We assume that the filtration $\left\{\mathcal{F}_{t}\right\}_{t \in[0, T]}$ is generated by the Brownian motion and is right continuous, and that each $\mathcal{F}_{t}$ contains all the $\mathbf{P}$-null sets of $\mathcal{F}$. We denote by $L_{\mathcal{F}}^{2}$ the set of square integrable $\left\{\mathcal{F}_{t}\right\}_{t \in[0, T]}$-adapted processes,

$$
L_{\mathcal{F}}^{2} \xlongequal{\text { def }}\left\{X \left\lvert\, \begin{array}{l}
\text { The process } X=\{X(t)\}_{t \in[0, T]} \text { is an }\left\{\mathcal{F}_{t}\right\}_{t \in[0, T]^{-}} \\
\text {adapted process such that } \int_{0}^{T} \mathbf{E}\left[X^{2}(t)\right] \mathrm{d} t<\infty
\end{array}\right.\right\},
$$

and by $L_{\mathcal{F}_{T}}^{2}$ the set of square integrable $\mathcal{F}_{T}$-measurable random variables,

$$
L_{\mathcal{F}_{T}}^{2} \xlongequal{\text { def }}\left\{X \mid X \text { is an } \mathcal{F}_{T} \text {-measurable random variable such that } \mathbf{E}\left[X^{2}\right]<\infty\right\} .
$$

There is a self-financing investor with a finite investment horizon $[0, T]$ who invests $X(t)$ dollars in the bond and $Y(t)$ dollars in the stock at time $t$. Any stock transaction incurs a proportional transaction fee, with $\lambda \in[0,+\infty)$ and $\mu \in[0,1)$ being the proportions paid when buying and selling the stock, respectively. Throughout this paper, we assume that $\lambda+\mu>0$, which means transaction costs must be involved. The bond-stock value process, starting from $(x, y)$ at $t=0$, evolves according to the equations

$$
\begin{align*}
& X^{x, M, N}(t)=x+r \int_{0}^{t} X^{x, M, N}(s) \mathrm{d} s-(1+\lambda) M(t)+(1-\mu) N(t),  \tag{2.1}\\
& Y^{y, M, N}(t)=y+\alpha \int_{0}^{t} Y^{y, M, N}(s) \mathrm{d} s+\sigma \int_{0}^{t} Y^{y, M, N}(s) \mathrm{d} B(s)+M(t)-N(t), \tag{2.2}
\end{align*}
$$

where $M(t)$ and $N(t)$ denote, respectively, the cumulative stock purchase and sell up to time $t$. Sometimes we simply use $X, Y$ or $X^{M, N}, Y^{M, N}$ instead of $X^{x, M, N}, Y^{y, M, N}$ if there is no ambiguity.

The admissible strategy set $\mathcal{A}$ of the investor is defined as follows:

$$
\mathcal{A} \xlongequal{\text { def }}\left\{\begin{array}{l|l}
(M, N) & \begin{array}{l}
\text { The processes } M=\{M(t)\}_{t \in[0, T]} \text { and } N=\{N(t)\}_{t \in[0, T]} \text { are } \\
\left\{\mathcal{F}_{t}\right\}_{t \in[0, T]-\text {-adapted, RCLL, nonnegative, and nondecreas- }} \\
\text { ing, and the processes } X^{x, M, N} \text { and } Y^{y, M, N} \text { are both in } L_{\mathcal{F}}^{2} \\
\text { for any }(x, y) \in \mathbb{R}^{2}
\end{array}
\end{array}\right\} .
$$

$(M, N)$ is called an admissible strategy if $(M, N) \in \mathcal{A}$. Correspondingly, $\left(X^{x, M, N}, Y^{y, M, N}\right)$ is called an admissible (bond-stock) process if $(x, y) \in \mathbb{R}^{2}$ and $(M, N) \in \mathcal{A}$.

For an admissible process ( $X^{x, M, N}, Y^{y, M, N}$ ), we define the investor's net wealth process by

$$
W^{X, Y}(t) \xlongequal{\text { def }} X(t)+(1-\mu) Y(t)^{+}-(1+\lambda) Y(t)^{-}, \quad t \in[0, T] .
$$

Namely, $W^{X, Y}(t)$ is the net worth of the investor's portfolio at $t$ after the transaction cost is deducted. The investor's attainable net wealth set at the maturity time $T$ is defined as

$$
\mathcal{W}_{0}^{x, y} \xlongequal{\text { def }}\left\{\begin{array}{l|l}
W^{X, Y}(T) & \begin{array}{l}
W^{X, Y}(T) \text { is the net wealth at } T \\
\text { of an admissible process }(X, Y) \text { with } \\
X(0-)=x, Y(0-)=y .
\end{array}
\end{array}\right\} .
$$

In the spirit of the original Markowitz's MV portfolio theory, an efficient strategy is a trading strategy for which there does not exist another strategy that has higher mean and no higher variance, and/or has less variance and no lower mean at the terminal time $T$. In other words, an efficient strategy is one that is Pareto optimal. Clearly, there could be many efficient strategies, and the terminal means and variances corresponding to all the efficient strategies form an efficient frontier. The positioning on the efficient frontier of a specific investor is dictated by his/her risk preference.

It is now well known that the efficient frontier can be obtained by solving the following variance minimizing problem.

Problem 2.1.

$$
\begin{aligned}
& \text { Minimize } \operatorname{Var}(W) \\
& \text { subject to } \mathbf{E}[W]=z, \quad W \in \mathcal{W}_{0}^{x, y}
\end{aligned}
$$

Here $z$ is a parameter satisfying

$$
z>e^{r T} x+(1-\mu) e^{r T} y^{+}-(1+\lambda) e^{r T} y^{-}
$$

which means that the target expected terminal wealth is higher than that of the simple "allbond" strategy (i.e., initially liquidating the stock investment and putting all the money in the bond account). The optimal solutions to the above problem with varying values of $z$ will trace out the efficient frontier we are looking for. For this reason, although Problem 2.1 is indeed an auxiliary mathematical problem introduced to help solve the original MV problem, it is sometimes (as in this paper) itself called the MV problem.

It is immediate to see that Problem 2.1 is equivalent to the following problem.
Problem 2.2.

$$
\begin{array}{ll}
\text { Minimize } & \mathbf{E}\left[W^{2}\right] \\
\text { subject to } & \mathbf{E}[W]=z, \quad W \in \mathcal{W}_{0}^{x, y}
\end{array}
$$

3. Feasibility. In contrast to the EUM problem, the MV model, Problem 2.2, has an inherent constraint $\mathbf{E}[W]=z$. Is there always an admissible strategy to meet this constraint no matter how aggressive the target $z$ is? This is the so-called feasibility issue. The issue is important and unique to the MV problem, and will be addressed fully in this section. To begin with, we introduce two lemmas.

Lemma 3.1. If $W_{1} \in \mathcal{W}_{0}^{x, y}, W_{2} \in L_{\mathcal{F}_{T}}^{2}$, and $W_{2} \leqslant W_{1}$, then $W_{2} \in \mathcal{W}_{0}^{x, y}$.
Proof. By the definition of $\mathcal{W}_{0}^{x, y}$, there exists $(M, N) \in \mathcal{A}$ such that $X^{M, N}(0-)=x$, $Y^{M, N}(0-)=y$, and $W^{X^{M, N}, Y^{M, N}}(T)=W_{1}$. We define

$$
\bar{M}(t)=\left\{\begin{array}{ll}
M(t) & \text { if } t<T, \\
M(T)+\frac{W_{1}-W_{2}}{\lambda+\mu} & \text { if } t=T,
\end{array} \quad \bar{N}(t)= \begin{cases}N(t) & \text { if } t<T, \\
N(T)+\frac{W_{1}-W_{2}}{\lambda+\mu} & \text { if } t=T\end{cases}\right.
$$

Then $(\bar{M}, \bar{N}) \in \mathcal{A}$ and

$$
X^{\bar{M}, \bar{N}}(t)=\left\{\begin{array}{ll}
X^{M, N}(t) & \text { if } t<T, \\
X^{M, N}(T)-W_{1}+W_{2} & \text { if } t=T,
\end{array} \quad Y^{\bar{M}, \bar{N}}(t)=Y^{M, N}(t), \quad t \in[0, T] .\right.
$$

Therefore $W_{2}=W^{X^{\bar{M}, \bar{N}}, Y^{\bar{M}, \bar{N}}}(T) \in \mathcal{W}_{0}^{x, y}$.
The above proof is very intuitive. If a higher terminal wealth is achievable by an admissible strategy, then so is a lower one, by simply "wasting money," i.e., buying and selling the same amount of the stock at $T$, thanks to the presence of the transaction costs. This is not necessarily true when there is no transaction cost.

Lemma 3.2. For any $(x, y) \in \mathbb{R}^{2}$, we have the following:
(1) The set $\mathcal{W}_{0}^{x, y}$ is convex.
(2) If $\left(x_{i}, y_{i}\right) \in \mathbb{R}^{2}, W_{i} \in \mathcal{W}_{0}^{x_{i}, y_{i}}, i=1,2$, then $W_{1}+W_{2} \in \mathcal{W}_{0}^{x_{1}+x_{2}, y_{1}+y_{2}}$.
(3) If $x_{1} \leqslant x_{2}$ and $y_{1} \leqslant y_{2}$, then $\mathcal{W}_{0}^{x_{1}, y_{1}} \subseteq \mathcal{W}_{0}^{x_{2}, y_{2}}$.
(4) $\mathcal{W}_{0}^{x-(1+\lambda) \rho, y+\rho} \subseteq \mathcal{W}_{0}^{x, y}$ and $\mathcal{W}_{0}^{x+(1-\mu) \rho, y-\rho} \subseteq \mathcal{W}_{0}^{x, y}$ for any $\rho>0$.
(5) $\mathcal{W}_{0}^{\rho x, \rho y}=\rho \mathcal{W}_{0}^{x, y}$ for any $\rho>0$.
(6) If $x+(1-\mu) y^{+}-(1+\lambda) y^{-} \geqslant 0$, then $0 \in \mathcal{W}_{0}^{x, y}$.

Proof. (1) For any $W_{1}, W_{2} \in \mathcal{W}_{0}^{x, y}$, assume that $W_{i}=W^{X_{i}, Y_{i}}(T)$, where $\left(X_{i}, Y_{i}\right)=$ $\left(X^{x, M_{i}, N_{i}}, Y^{y, M_{i}, N_{i}}\right),\left(M_{i}, N_{i}\right) \in \mathcal{A}, i=1,2$. For any $k \in(0,1)$, let $M=k M_{1}+(1-k) M_{2}$, $N=k N_{1}+(1-k) N_{2}$. Then $(M, N) \in \mathcal{A}$, and

$$
\begin{aligned}
X^{x, M, N} & =k X_{1}+(1-k) X_{2}, \\
Y^{y, M, N} & =k Y_{1}+(1-k) Y_{2} .
\end{aligned}
$$

Thus

$$
\begin{aligned}
W^{X, Y}(T)= & X(T)+(1-\mu) Y(T)^{+}-(1+\lambda) Y(T)^{-}=X(T)+(1-\mu) Y(T)-(\mu+\lambda) Y(T)^{-} \\
\geqslant & k X_{1}(T)+(1-k) X_{2}(T)+(1-\mu)\left(k Y_{1}(T)+(1-k) Y_{2}(T)\right) \\
& -(\mu+\lambda)\left(k Y_{1}(T)^{-}+(1-k) Y_{2}(T)^{-}\right) \\
= & k\left(X_{1}(T)+(1-\mu) Y_{1}(T)^{+}-(1+\lambda) Y_{1}(T)^{-}\right) \\
& +(1-k)\left(X_{2}(T)+(1-\mu) Y_{2}(T)^{+}-(1+\lambda) Y_{2}(T)^{-}\right) \\
= & k W_{1}+(1-k) W_{2} .
\end{aligned}
$$

Since $W^{X, Y}(T) \in \mathcal{W}_{0}^{x, y}, k W_{1}+(1-k) W_{2} \in L_{\mathcal{F}_{T}}^{2}$, it follows from Lemma 3.1 that $k W_{1}+$ $(1-k) W_{2} \in \mathcal{W}_{0}^{x, y}$.
(2) This can be proved by the same argument as above.
(3) If $x_{1} \leqslant x_{2}$ and $y_{1} \leqslant y_{2}$, then for any $(M, N) \in \mathcal{A}$, we have $X^{x_{1}, M, N} \leqslant X^{x_{2, M, N}}$, $Y^{y_{1}, M, N} \leqslant Y^{y_{2}, M, N}$. So $W^{X^{x_{1}, M, N}, Y^{y_{1}, M, N}}(T) \leqslant W^{X^{x_{2}, M, N}, Y^{y_{2}, M, N}}(T) \in \mathcal{W}_{0}^{x_{2}, y_{2}}$. Therefore by Lemma 3.1, we have $W^{X^{x_{1}, M, N}, Y^{y_{1}, M, N}}(T) \in \mathcal{W}_{0}^{x_{2}, y_{2}}$. This shows that $\mathcal{W}_{0}^{x_{1}, y_{1}} \subseteq \mathcal{W}_{0}^{x_{2}, y_{2}}$.
(4) For any $\rho>0,(M, N) \in \mathcal{A}$, we define

$$
\bar{M}(s)=M(s)+\rho \quad \forall s \in[0, T] .
$$

Then $\left(X^{x-(1+\lambda) \rho, M, N}, Y^{y+\rho, M, N}\right)=\left(X^{x, \bar{M}, N}, Y^{y, \bar{M}, N}\right)$. So

$$
W^{X^{x-(1+\lambda) \rho, M, N}, Y^{y+\rho, M, N}}=W^{X^{x, \bar{M}, N}, Y^{y, \bar{M}, N}} \in \mathcal{W}_{0}^{x, y} .
$$

Hence $\mathcal{W}_{0}^{x-(1+\lambda) \rho, y+\rho} \subseteq \mathcal{W}_{0}^{x, y}$. Similarly, we can prove that $\mathcal{W}_{0}^{x+(1-\mu) \rho, y-\rho} \subseteq \mathcal{W}_{0}^{x, y}$.
 have $\mathcal{W}_{0}^{\rho x, \rho y}=\rho \mathcal{W}_{0}^{x, y}$.
(6) If $y \geqslant 0$ and $x+(1-\mu) y \geqslant 0$, then obviously $0 \in \mathcal{W}_{0}^{x+(1-\mu) y, 0}$ by Lemma 3.1. Noting (4) proved above, we conclude that $0 \in \mathcal{W}_{0}^{x, y}$. Similarly we can prove the case of $y<0$ and $x+(1+\lambda) y \geqslant 0$.

Denote

$$
\begin{equation*}
\hat{z} \xlongequal{\text { def }} \sup \left\{\mathbf{E}[W] \mid W \in \mathcal{W}_{0}^{x, y}\right\} . \tag{3.1}
\end{equation*}
$$

In view of Lemma 3.1, Problem 2.2 is feasible when

$$
\begin{equation*}
z \in \mathcal{D} \xlongequal{\text { def }}\left(e^{r T} x+(1-\mu) e^{r T} y^{+}-(1+\lambda) e^{r T} y^{-}, \hat{z}\right) \tag{3.2}
\end{equation*}
$$

It is clear that Problem 2.2 is not feasible when $z>\hat{z}$. Thus, it remains to investigate whether Problem 2.2 admits a feasible solution when $z=\hat{z}$.

It is well known that in the absence of transaction costs (i.e., $\lambda=\mu=0$ ), we have $\hat{z}=+\infty$ and thus Problem 2.2 is always feasible for any $z>e^{r T} x+(1-\mu) e^{r T} y^{+}-(1+\lambda) e^{r T} y^{-}$(see [23]). In other words, no matter how small the investor's initial wealth is, the investor can always arrive at an arbitrarily large expected return in a split second by taking a huge leverage on the stock. The following theorem indicates that things become very different when the transaction costs get involved.

Theorem 3.3. Assume that $T \leqslant T^{*} \xlongequal{\text { def }} \frac{1}{\alpha-r} \ln \left(\frac{1+\lambda}{1-\mu}\right)$. Then

$$
\hat{z}= \begin{cases}e^{r T} x+(1-\mu) e^{\alpha T} y & \text { if } y>0 \\ e^{r T} x+(1+\lambda) e^{r T} y & \text { if } y \leqslant 0\end{cases}
$$

Moreover, if $y>0$ and $z=\hat{z}$, then Problem 2.1 admits a unique feasible (thus optimal) solution and the optimal strategy is $(M, N) \equiv(0,0)$. If $y \leqslant 0$, then $\mathcal{D}=\emptyset$.

Proof. For any $(M, N) \in \mathcal{A}$, due to (2.1), (2.2), and Itô's formula, we have

$$
\begin{aligned}
X(T)+(1-\mu) Y(T)= & e^{r T} x+(1-\mu) e^{\alpha T} y+\int_{0}^{T} e^{r(T-t)}(\mathrm{d} X(t)-r X(t) \mathrm{d} t) \\
& +\int_{0}^{T}(1-\mu) e^{\alpha(T-t)}(\mathrm{d} Y(t)-\alpha Y(t) \mathrm{d} t) \\
= & e^{r T} x+(1-\mu) e^{\alpha T} y+\int_{0}^{T}\left((1-\mu) e^{\alpha(T-t)}-(1+\lambda) e^{r(T-t)}\right) \mathrm{d} M^{c}(t) \\
& +\int_{0}^{T}(1-\mu)\left(e^{r(T-t)}-e^{\alpha(T-t)}\right) \mathrm{d} N^{c}(t)+\int_{0}^{T}(1-\mu) e^{\alpha(T-t)} \sigma Y(t) \mathrm{d} B(t) \\
& +\sum_{0 \leqslant t \leqslant T}\left((1-\mu) e^{\alpha(T-t)}-(1+\lambda) e^{r(T-t)}\right)(M(t)-M(t-)) \\
& +\sum_{0 \leqslant t \leqslant T}(1-\mu)\left(e^{r(T-t)}-e^{\alpha(T-t)}\right)(N(t)-N(t-)) \\
\leqslant & e^{r T} x+(1-\mu) e^{\alpha T} y+\int_{0}^{T}(1-\mu) e^{\alpha(T-t)} \sigma Y(t) \mathrm{d} B(t),
\end{aligned}
$$

where $M^{c}$ and $N^{c}$ are, respectively, the contiguous parts of $M$ and $N$ and we have noted $T \leqslant \frac{1}{\alpha-r} \ln \left(\frac{1+\lambda}{1-\mu}\right)$. Thus

$$
\mathbf{E}[X(T)+(1-\mu) Y(T)] \leqslant e^{r T} x+(1-\mu) e^{\alpha T} y
$$

It follows that

$$
\begin{aligned}
\mathbf{E}\left[W^{X, Y}\right] & =\mathbf{E}\left[X(T)+(1-\mu) Y(T)^{+}-(1+\lambda) Y(T)^{-}\right] \leqslant \mathbf{E}[X(T)+(1-\mu) Y(T)] \\
& \leqslant e^{r T} x+(1-\mu) e^{\alpha T} y .
\end{aligned}
$$

When $y \geqslant 0$, we have $\mathbf{E}\left[W^{X, Y}\right]=e^{r T} x+(1-\mu) e^{\alpha T} y$ if and only if $(M, N) \equiv(0,0)$. Thus, if $y \geqslant 0$, then $\hat{z}=e^{r T} x+(1-\mu) e^{\alpha T} y$, and $(0,0)$ is the unique feasible strategy when $z=\hat{z}$.

Now we turn to the case of $y<0$. Define $\tau=\inf \{t \in[0, T): Y(t)>0\} \wedge T$. Then $\tau$ is a stopping time (cf. [18, Theorem 2.33]). On the set $\{Y(\tau) \geqslant 0\}$, by the same argument as above, we have

$$
W_{T}^{X, Y} \leqslant e^{r(T-\tau)} X(\tau)+(1-\mu) e^{\alpha(T-\tau)} Y(\tau) \leqslant e^{r(T-\tau)} X(\tau)+(1+\lambda) e^{r(T-\tau)} Y(\tau)
$$

On the set $\{Y(\tau)<0\}$, we have $\tau=T$,

$$
W_{T}^{X, Y}=X(T)+(1+\lambda) Y(T)=e^{r(T-\tau)} X(\tau)+(1+\lambda) e^{r(T-\tau)} Y(\tau)
$$

On the other hand, noting that $Y(t) \leqslant 0, t \in[0, \tau)$, we have

$$
\begin{aligned}
& e^{r(T-\tau)} X(\tau)+(1+\lambda) e^{r(T-\tau)} Y(\tau) \\
= & e^{r T} x+(1+\lambda) e^{r T} y-\int_{0}^{\tau}(\lambda+\mu) e^{r(T-t)} \mathrm{d} N^{c}(t)+\int_{0}^{\tau}(\alpha-r) e^{r(T-t)} Y(t) \mathrm{d} t \\
& +\int_{0}^{\tau} e^{r(T-t)}(1-\mu) \sigma Y(t) \mathrm{d} B(t)-\sum_{0 \leqslant t \leqslant \tau}(\lambda+\mu) e^{r(T-t)}(N(t)-N(t-)) \\
\leqslant & e^{r T} x+(1+\lambda) e^{r T} y+\int_{0}^{\tau} e^{r(T-t)}(1-\mu) \sigma Y(t) \mathrm{d} B(t) .
\end{aligned}
$$

It follows that

$$
\mathbf{E}\left[e^{r(T-\tau)} X(\tau)+(1+\lambda) e^{r(T-\tau)} Y(\tau)\right] \leqslant e^{r T} x+(1+\lambda) e^{r T} y .
$$

Therefore,

$$
\mathbf{E}\left[W_{T}^{X, Y}\right] \leqslant \mathbf{E}\left[e^{r(T-\tau)} X(\tau)+(1+\lambda) e^{r(T-\tau)} Y(\tau)\right] \leqslant e^{r T} x+(1+\lambda) e^{r T} y
$$

This indicates that $\mathbf{E}\left[W_{T}^{X, Y}\right]=e^{r T} x+(1+\lambda) e^{r T} y$ if and only if the investor puts all of his wealth in the bond at time 0 . Thus $\hat{z}=e^{r T} x+(1+\lambda) e^{r T} y$ if $y<0$.

This result demonstrates the importance of the length of the investment planning horizon, $T$, by examining the situation when $T$ is not long enough. In this "short-horizon" case, if the investor starts with a short position in stock, then the only sensible strategy is the all-bond one, since any other strategy will just be worse off in both mean and variance. On the other hand, if one starts with a long stock position, then the highest expected terminal net wealth (without considering the variance) is achieved by the "stay-put" strategy, one that does not switch at all between bond and stock from the very beginning. Therefore, any efficient strategy is between the two extreme strategies, those of "all-bond" and "stay-put," according to an individual investor's risk preference.

More significantly, Theorem 3.3 specifies explicitly this critical length of horizon, $T^{*}=$ $\frac{1}{\alpha-r} \ln \left(\frac{1+\lambda}{1-\mu}\right)$. It is intriguing that $T^{*}$ depends only on the excess return $\alpha-r$ and the transaction fees $\lambda, \mu$, not on the individual target $z$ or the stock volatility $\sigma$. Later we will show that, indeed, $\hat{z}=+\infty$ when $T>T^{*}$ in Corollary 6.3. Therefore $T^{*}$ is a critical value in time that distinguishes between "global feasibility" and "limited feasibility" of the underlying MV portfolio selection problem. It signifies the opportunity that a longer time horizon would provide for achieving a higher potential gain. In this sense, the length of the planning horizon should really be included in the set of the investment opportunities, as opposed to the hitherto widely accepted notion that the investment opportunity set consists of only the probabilistic characteristics of the returns. Moreover, it follows from the expression of $T^{*}$ that the lower the transaction cost is and/or the higher the excess return of the stock is, the shorter is the time required to attain the global feasibility. These observations, of course, all make perfect sense economically.

In the remaining part of this paper, we consider only the case when $\mathcal{D} \neq \emptyset$ and $z \in \mathcal{D}$.

## 4. Unconstrained problem and double-obstacle problem.

4.1. Lagrangian relaxation and HJB equation. By virtue of Lemma 3.2, Problem 2.2 is a convex constrained optimization problem. We shall utilize the well-known Lagrange multiplier method to remove the constraint.

Let us introduce the following unconstrained problem.
Problem 4.1 (unconstrained problem).

$$
\begin{array}{ll}
\text { Minimize } & \mathbf{E}\left[W^{2}\right]-2 \ell(\mathbf{E}[W]-z) \\
\text { subject to } & W \in \mathcal{W}_{0}^{x, y}
\end{array}
$$

or, equivalently, the following problem.
Problem 4.2.

$$
\begin{array}{ll}
\text { Minimize } & \mathbf{E}\left[(W-\ell)^{2}\right] \\
\text { subject to } & W \in \mathcal{W}_{0}^{x, y}
\end{array}
$$

Define the value function of Problem 2.2 as follows:

$$
V_{1}(x, y ; z) \xlongequal{\text { def }} \inf _{\substack{W \in \mathcal{W}_{0}^{x, y} \\ \mathbf{E}[W]=z}} \mathbf{E}\left[W^{2}\right], \quad z \in \mathcal{D} .
$$

The following result, showing the connection between Problems 2.2 and 4.2 , can be proved by a standard convex analysis argument.

Proposition 4.1. Problems 2.2 and 4.2 have the following relations.
(1) If $W_{z}^{*}$ solves Problem 2.2 with parameter $z \in \mathcal{D}$, then there exists $\ell \in \mathbb{R}$ such that $W_{z}^{*}$ also solves Problem 4.2 with parameter $\ell$.
(2) Conversely, if $W_{\ell}$ solves Problem 4.2 with parameter $\ell \in \mathbb{R}$, then it must also solve Problem 2.2 with parameter $z=\mathbf{E}\left[W_{\ell}\right]$.
It is easy to see that $\mathcal{W}_{0}^{x, y}-\ell=\mathcal{W}_{0}^{x-\ell e^{-r T}, y}$. As a consequence, we consider the following problem instead of Problem 4.2.

Problem 4.3.

$$
\begin{array}{ll}
\text { Minimize } & \mathbf{E}\left[W^{2}\right] \\
\text { subject to } & W \in \mathcal{W}_{0}^{x-\ell e^{-r T}, y}
\end{array}
$$

To solve the above problem, we use dynamic programming. In doing so we need to parameterize the initial time. Consider the dynamics (2.1)-(2.2), where the initial time 0 is revised to some $s \in[0, T)$, and define $\mathcal{W}_{s}^{x, y}$ as the counterpart of $\mathcal{W}_{0}^{x, y}$, where the initial time is $s$ and initial bond-stock position is $(x, y)$. We then define the value function of Problem 4.3 as

$$
\begin{equation*}
V(t, x, y) \xlongequal{\text { def }} \inf _{W \in \mathcal{W}_{t}^{x, y}} \mathbf{E}\left[W^{2}\right], \quad(t, x, y) \in[0, T) \times \mathbb{R}^{2} \tag{4.1}
\end{equation*}
$$

The following proposition establishes a link between Problems 4.3 and 2.2.

Proposition 4.2. If $z \in \mathcal{D}$, then

$$
\sup _{\ell \in \mathbb{R}}\left(V\left(0, x-\ell e^{-r T}, y\right)-(\ell-z)^{2}\right)=V_{1}(x, y ; z)-z^{2}
$$

Proof. Note that

$$
\begin{aligned}
& \sup _{\ell \in \mathbb{R}}\left(V\left(0, x-\ell e^{-r T}, y\right)-(\ell-z)^{2}\right)=\sup _{\ell \in \mathbb{R}} \inf _{W \in \mathcal{W}_{0}^{x-\ell e^{-r T}, y}} \mathbf{E}\left[W^{2}-(\ell-z)^{2}\right] \\
= & \sup _{\ell \in \mathbb{R}} \inf _{W \in \mathcal{W}_{0}^{x, y}} \mathbf{E}\left[(W-\ell)^{2}-(\ell-z)^{2}\right] \leqslant \sup _{\ell \in \mathbb{R}} \inf _{\substack{ \\
\mathbf{E} \mathcal{W}_{0}^{x, y} \\
\mathbf{E}[W]=z}} \mathbf{E}\left[(W-\ell)^{2}-(\ell-z)^{2}\right] \\
= & \sup _{\ell \in \mathbb{R}} \inf _{\substack{W \in \mathcal{W}_{0}^{x, y} \\
\mathbf{E}[W]=z}}\left(\mathbf{E}\left[W^{2}\right]-z^{2}\right)=\inf _{\substack{W \in \mathcal{W}_{0}^{x, y} \\
\mathbf{E}[W]=z}}\left(\mathbf{E}\left[W^{2}\right]-z^{2}\right)=V_{1}(x, y ; z)-z^{2} .
\end{aligned}
$$

Therefore

$$
\sup _{\ell \in \mathbb{R}}\left(V\left(0, x-\ell e^{-r T}, y\right)-(\ell-z)^{2}\right) \leqslant V_{1}(x, y ; z)-z^{2}
$$

Since $V_{1}$ is convex and $z$ is an interior point of $\mathcal{D}$, by convex analysis, there exists $\ell^{*} \in \mathbb{R}$ such that

$$
V_{1}(x, y ; z)-2 \ell^{*} z \leqslant V_{1}(x, y ; \tilde{z})-2 \ell^{*} \tilde{z} \quad \forall \tilde{z} \in \mathcal{D}
$$

For any $W \in \mathcal{W}_{0}^{x, y}$, by the definition of $V_{1}$, we have

$$
\begin{aligned}
\mathbf{E}\left[\left(W-\ell^{*}\right)^{2}-\left(\ell^{*}-z\right)^{2}\right]=\mathbf{E}\left[W^{2}\right]-2 \ell^{*}(\mathbf{E}[W]-z)-z^{2} & \geqslant V_{1}(x, y ; \mathbf{E}[W])-2 \ell^{*}(\mathbf{E}[W]-z)-z^{2} \\
& \geqslant V_{1}(x, y ; z)-z^{2}
\end{aligned}
$$

It follows that

$$
\begin{aligned}
& \sup _{\ell \in \mathbb{R}}\left(V\left(0, x-\ell e^{-r T}, y\right)-(\ell-z)^{2}\right)=\sup _{\ell \in \mathbb{R}} \inf _{W \in \mathcal{W}_{0}^{x, y}} \mathbf{E}\left[(W-\ell)^{2}-(\ell-z)^{2}\right] \\
\geqslant & \inf _{W \in \mathcal{W}_{0}^{x, y}} \mathbf{E}\left[\left(W-\ell^{*}\right)^{2}-\left(\ell^{*}-z\right)^{2}\right] \geqslant V_{1}(x, y ; z)-z^{2},
\end{aligned}
$$

which yields the desired result.
Therefore, we need only study the value function $V(t, x, y)$, which we set out to do now.
Lemma 4.3. The value function $V$ defined in (4.1) has the following properties.
(1) For any $t \in[0, T), V(t, \cdot, \cdot)$ is convex and continuous in $\mathbb{R}^{2}$.
(2) For any $t \in[0, T), V(t, x, y)$ is nonincreasing in $x$ and $y$.
(3) For any $\rho>0, t \in[0, T)$, we have $V(t, x+(1-\mu) \rho, y-\rho) \geqslant V(t, x, y), V(t, x-(1+$ $\lambda) \rho, y+\rho) \geqslant V(t, x, y)$.
(4) For any $\rho>0, t \in[0, T)$, we have $V(t, \rho x, \rho y)=\rho^{2} V(t, x, y)$.
(5) If $x+(1-\mu) y^{+}-(1+\lambda) y^{-} \geqslant 0$, then $V(t, x, y)=0$.

Proof. All the results can be easily proved in terms of the definition of $V$ and Lemma 3.2.

Due to part (5) of the above lemma, we need only consider Problem 4.3 in the insolvency region

$$
\mathscr{S} \xlongequal{\text { def }}\left\{(x, y) \in \mathbb{R}^{2} \mid x+(1-\mu) y^{+}-(1+\lambda) y^{-}<0\right\} .
$$

It is well known that the value function $V$ is a viscosity solution to the following HJB equation or VI with terminal condition:

$$
\begin{cases}\min \left\{\varphi_{t}+\mathcal{L}_{0} \varphi,(1-\mu) \varphi_{x}-\varphi_{y}, \varphi_{y}-(1+\lambda) \varphi_{x}\right\}=0 & \forall(t, x, y) \in[0, T) \times \mathscr{S},  \tag{4.2}\\ \varphi(T, x, y)=\left(x+(1-\mu) y^{+}-(1+\lambda) y^{-}\right)^{2} & \forall(x, y) \in \mathscr{S}\end{cases}
$$

where

$$
\mathcal{L}_{0} \varphi \xlongequal{\text { def }} \frac{1}{2} \sigma^{2} y^{2} \varphi_{y y}+\alpha y \varphi_{y}+r x \varphi_{x} .
$$

The idea of the subsequent analysis is to construct a particular solution to the HJB equation and then employ the verification theorem to obtain an optimal strategy. The construction of the solution is built upon a series of transformations of (4.2) until we reach an equation related to the so-called double-obstacle problem in physics, which has been well studied in the PDE literature.

We will show in Proposition 4.4 below that the constructed solution $\varphi$ satisfies $\varphi_{y}-$ $(1+\lambda) \varphi_{x}=0$ when $y<0$. Hence, we need only focus on $y>0$. A substantial technical difficulty arises with the HJB equation (4.2) in that the spatial variable $(x, y)$ is two-dimensional. However, the homogeneity of Lemma 4.3 (4) motivates us to make the transformation

$$
\varphi(t, x, y)=y^{2} \bar{V}\left(t, \frac{x}{y}\right) \text { for } y>0
$$

so as to reduce the dimension by one. Accordingly, (4.2) is transformed into
$\begin{cases}\min \left\{\bar{V}_{t}+\mathcal{L}_{1} \bar{V},(x+1-\mu) \bar{V}_{x}-2 \bar{V},-(x+1+\lambda) \bar{V}_{x}+2 \bar{V}\right\}=0 & \forall(t, x) \in[0, T) \times \mathscr{X}, \\ \bar{V}(T, x)=(x+1-\mu)^{2} & \forall x \in \mathscr{X},\end{cases}$
where $\mathscr{X} \xlongequal{\text { def }}(-\infty,-(1-\mu))$ and

$$
\mathcal{L}_{1} \bar{V}=\frac{1}{2} \sigma^{2} x^{2} \bar{V}_{x x}-\left(\alpha-r+\sigma^{2}\right) x \bar{V}_{x}+\left(2 \alpha+\sigma^{2}\right) \bar{V} .
$$

Further, let

$$
w(t, x) \xlongequal{\text { def }} \frac{1}{2} \ln \bar{V}(t, x), \quad(t, x) \in[0, T) \times \mathscr{X} .
$$

It is not hard to show that $w(t, x)$ is governed by

$$
\begin{cases}\min \left\{w_{t}+\mathcal{L}_{2} w, \frac{1}{w_{x}}-(x+1-\mu),(x+1+\lambda)-\frac{1}{w_{x}}\right\}=0 & \forall(t, x) \in[0, T) \times \mathscr{X},  \tag{4.3}\\ w(t, x)=\ln (-x-(1-\mu)) & \forall x \in \mathscr{X},\end{cases}
$$

where

$$
\mathcal{L}_{2} w \xlongequal{\text { def }} \frac{1}{2} \sigma^{2} x^{2}\left(w_{x x}+2 w_{x}^{2}\right)-\left(\alpha-r+\sigma^{2}\right) x w_{x}+\alpha+\frac{1}{2} \sigma^{2} .
$$

4.2. A related double-obstacle problem. Equation (4.3) is a VI with gradient constraints, which is hard to study. As in Dai and Yi [8], we will relate it to a double-obstacle problem that is tractable. We refer interested readers to Friedman [14] for obstacle problems.

Let

$$
\begin{equation*}
v(t, x) \xlongequal{\text { def }} \frac{1}{w_{x}(t, x)}, \quad(t, x) \in[0, T) \times \mathscr{X} . \tag{4.4}
\end{equation*}
$$

Notice that

$$
\frac{\partial}{\partial x} \mathcal{L}_{2} w=-\frac{1}{v^{2}}\left[\frac{1}{2} \sigma^{2} x^{2} v_{x x}-(\alpha-r) x v_{x}+\left(\alpha-r+\sigma^{2}\right) v+\sigma^{2}\left(\frac{2 x^{2} v_{x}-x^{2} v_{x}^{2}}{v}-2 x\right)\right] .
$$

This inspires us to consider the following double-obstacle problem:

$$
\left\{\begin{array}{l}
\max \left\{\min \left\{-v_{t}-\mathcal{L} v, v-(x+1-\mu)\right\}, v-(x+1+\lambda)\right\}=0 \quad \forall(t, x) \in[0, T) \times \mathscr{X},  \tag{4.5}\\
v(T, x)=x+1-\mu \quad \forall x \in \mathscr{X},
\end{array}\right.
$$

where

$$
\begin{equation*}
\mathcal{L} v \xlongequal{\text { def }} \frac{1}{2} \sigma^{2} x^{2} v_{x x}-(\alpha-r) x v_{x}+\left(\alpha-r+\sigma^{2}\right) v+\sigma^{2}\left(\frac{2 x^{2} v_{x}-x^{2} v_{x}^{2}}{v}-2 x\right) . \tag{4.6}
\end{equation*}
$$

It should be emphasized that at this stage we have yet to know if (4.5) is mathematically equivalent to (4.3) via the transformation (4.4). However, the following propositions show that (4.5) is solvable, and the solution to (4.3) can be constructed through the solutions of (4.5).

Proposition 4.4. Equation (4.5) has a solution $v \in W_{p}^{1,2}([0, T) \times(-N,-(1-\mu)))$ for any $N>1-\mu, p \in(1, \infty)$. Moreover,

$$
\begin{gather*}
v_{t} \leqslant 0  \tag{4.7}\\
0 \leqslant v_{x} \leqslant 1 \tag{4.8}
\end{gather*}
$$

and there exist two decreasing functions $x_{s}^{*}(\cdot) \in C^{\infty}[0, T)$ and $x_{b}^{*}(\cdot) \in C^{\infty}\left[0, T_{0}\right)$ such that

$$
\begin{equation*}
\{(t, x) \in[0, T) \times \mathscr{X}: v(t, x)=x+1-\mu\}=\left\{(t, x) \in[0, T) \times \mathscr{X}: x \geqslant x_{s}^{*}(t)\right\} \tag{4.9}
\end{equation*}
$$

and

$$
\begin{equation*}
\{(t, x) \in[0, T) \times \mathscr{X}: v(t, x)=x+1+\lambda\}=\left\{(t, x) \in\left[0, T_{0}\right) \times \mathscr{X}: x \leqslant x_{b}^{*}(t)\right\} \tag{4.10}
\end{equation*}
$$ where

$$
\begin{equation*}
T_{0}=\max \left\{T-\frac{1}{\alpha-r} \ln \left(\frac{1+\lambda}{1-\mu}\right), 0\right\} . \tag{4.11}
\end{equation*}
$$

Further, we have
(4.12) $\lim _{t \uparrow T} x_{s}^{*}(t)=(1-\mu) x_{M}, \quad \lim _{T \uparrow \infty} x_{s}^{*}(0)=x_{s, \infty}^{*}, \quad \lim _{t \uparrow T_{0}} x_{b}^{*}(t)=-\infty, \quad \lim _{T \uparrow \infty} x_{b}^{*}(0)=x_{b, \infty}^{*}$,
where $x_{M}=-\frac{\alpha-r+\sigma^{2}}{\alpha-r}$ and $x_{s, \infty}^{*}$ and $x_{b, \infty}^{*}$ are defined in (A.1) and (A.2).
Proposition 4.5. Define

$$
\begin{equation*}
w(t, x) \stackrel{\text { def }}{=} \mathfrak{A}(t)+\ln \left(-x_{s}^{*}(t)-(1-\mu)\right)+\int_{x_{s}^{*}(t)}^{x} \frac{1}{v(t, y)} \mathrm{d} y \tag{4.13}
\end{equation*}
$$

where

$$
\begin{equation*}
\mathfrak{A}(t) \xlongequal{\text { def }} \int_{t}^{T} \frac{r x_{s}^{* 2}(\tau)+(\alpha+r)(1-\mu) x_{s}^{*}(\tau)+\left(\alpha+\frac{1}{2} \sigma^{2}\right)(1-\mu)^{2}}{\left(x_{s}^{*}(\tau)+1-\mu\right)^{2}} \mathrm{~d} \tau \tag{4.14}
\end{equation*}
$$

Then $w \in C^{1,2}([0, T) \times \mathscr{X})$ is a solution to (4.3). Moreover, for any $(t, x, y) \in[0, T) \times \mathscr{S}$, we define

$$
\varphi(t, x, y) \xlongequal{\text { def }} \begin{cases}y^{2} e^{2 w\left(t, \frac{x}{y}\right)} & \text { if } y>0  \tag{4.15}\\ e^{2 \mathfrak{B}(t)}(x+(1+\lambda) y)^{2} & \text { if } y \leqslant 0\end{cases}
$$

where $w(t, x)$ is given in (4.13) and

$$
\mathfrak{B}(t) \xlongequal{\text { def }} \int_{t}^{T} \frac{r x_{b}^{* 2}(\tau)+(\alpha+r)(1+\lambda) x_{b}^{*}(\tau)+\left(\alpha+\frac{1}{2} \sigma^{2}\right)(1+\lambda)^{2}}{\left(x_{b}^{*}(\tau)+1+\lambda\right)^{2}} d \tau
$$

Then $\varphi \in C^{1,2,2}([0, T) \times \mathscr{S} \backslash\{y=0\})$ is a solution to the HJB equation (4.2).
Due to their considerable technicality, the proofs of the preceding two propositions are placed in Appendix A. Most of the above results are similar to those obtained by Dai and Yi [8] where they considered the expected utility portfolio selection with transaction costs. Nonetheless, there is one breakthrough made by the present paper: both $x_{s}^{*}(\cdot)$ and $x_{b}^{*}(\cdot)$ are proved to be $C^{\infty}$, whereas Dai and Yi [8] obtained the smoothness of only $x_{s}^{*}(\cdot)$. In fact, Dai and Yi [8] essentially considered a double-obstacle problem for $w_{x}$. In contrast, the present paper takes the double-obstacle problem for $1 / w_{x}$ into consideration. This seemingly innocent modification in fact simplifies the proof greatly. More importantly, it allows us to provide a unified framework to obtain the smoothness of $x_{s}^{*}(\cdot)$ and $x_{b}^{*}(\cdot)$. Later we will see that the smoothness of $x_{s}^{*}(\cdot)$ and $x_{b}^{*}(\cdot)$ plays a critical role in the proof of the existence of an optimal strategy.

Note that (4.8) is important in the study of the double-obstacle problem (4.5). In essence, the result is based on parts (1) and (2) of Lemma 4.3.

In the subsequent section, we plan to show, through the verification theorem and a Skorokhod problem, that $\varphi(t, x, y)$ is nothing but the value function. At the same time, an optimal strategy will be constructed.
5. Skorokhod problem and optimal strategy. Due to (4.9)-(4.10) and Proposition 4.5, we define

$$
\begin{align*}
\mathcal{S R} & =\left\{(t, x, y) \in[0, T) \times \mathscr{S} \mid y>0, x \geqslant x_{s}^{*}(t) y\right\} \\
\mathcal{B R} & =\left\{(t, x, y) \in[0, T) \times \mathscr{S} \mid y>0, x \leqslant x_{b}^{*}(t) y \text { or } y \leqslant 0\right\}  \tag{5.1}\\
\mathcal{N} \mathcal{T} & =\left\{(t, x, y) \in[0, T) \times \mathscr{S} \mid y>0, x_{b}^{*}(t) y<x<x_{s}^{*}(t) y\right\}
\end{align*}
$$

which stand for the sell region, buy region, and no-trade region, respectively. Here we set $x_{b}^{*}(t)=-\infty$ when $t \in\left[T_{0}, T\right]$ in view of the fact that $\lim _{t \uparrow T_{0}} x_{b}^{*}(t)=-\infty$; hence $\mathcal{B} \mathcal{R}=\emptyset$ when $t \in\left[T_{0}, T\right]$. Notice that these regions do not depend on the target $z$.
5.1. Skorokhod problem and verification theorem. In order to find the optimal solution to the MV problem, we need to study the so-called Skorokhod problem.

Problem 5.1 (Skorokhod problem). Given $(0, X(0), Y(0)) \in \overline{\mathcal{N T}}$, find an admissible strategy $(M, N)$ such that the corresponding bond-stock value process $(X, Y)$ is continuous in $[0, T]$, and $(t, X(t), Y(t)) \in \overline{\mathcal{N T}}$ for any $t \in[0, T]$.

In other words, a solution to the Skorokhod problem is an investment strategy with which the trading takes place only on the boundary of the no-trade region. It turns out that the solution can be constructed via solving the following more specific problem.

Problem 5.2. Given $(0, X(0), Y(0)) \in \overline{\mathcal{N T}}$, find a process of bounded total variation $k$ and a continuous process $(X, Y)$ such that for any $t \in[0, T]$,

$$
\begin{array}{r}
(t, X(t), Y(t)) \in \overline{\mathcal{N} \mathcal{T}}, \\
\mathrm{d} X(t)=r X(t) \mathrm{d} t+\gamma_{1}(X(t), Y(t)) \mathrm{d}|k|(t), \\
\mathrm{d} Y(t)=\alpha Y(t) \mathrm{d} t+\sigma Y(t) \mathrm{d} B(t)+\gamma_{2}(X(t), Y(t)) \mathrm{d}|k|(t), \\
|k|(t)=\int_{0}^{t} \mathbf{1}_{\{(s, X(s), Y(s)) \in \partial \mathcal{N} \mathcal{T}\}} \mathrm{d}|k|(s),
\end{array}
$$

where $|k|(t)$ stands for the total variation of $k$ on $[0, t]$,

$$
\left(\gamma_{1}(x, y), \gamma_{2}(x, y)\right) \xlongequal{\text { def }} \begin{cases}\frac{1}{\sqrt{(1+\lambda)^{2}+1}}(-(1+\lambda), 1) & \text { if }(t, x, y) \in \partial_{1} \mathcal{N} \mathcal{T} \\ \frac{1}{\sqrt{(1-\mu)^{2}+1}}(1-\mu,-1) & \text { if }(t, x, y) \in \partial_{2} \mathcal{N} \mathcal{T}\end{cases}
$$

and

$$
\begin{aligned}
& \partial_{1} \mathcal{N} \mathcal{T} \xlongequal{\text { def }}\left\{(t, x, y) \in[0, T) \times \mathscr{S} \mid y>0, x=x_{b}^{*}(t) y\right\}, \\
& \partial_{2} \mathcal{N} \mathcal{T} \xlongequal{\text { def }}\left\{(t, x, y) \in[0, T) \times \mathscr{S} \mid y>0, x=x_{s}^{*}(t) y\right\} .
\end{aligned}
$$

Letting a triplet ( $k, X, Y$ ) solve Problem 5.2, define

$$
\begin{align*}
& M(t) \xlongequal{\text { def }} \frac{1}{\sqrt{(1+\lambda)^{2}+1}} \int_{0}^{t} \mathbf{1}_{\left\{(s, X(s), Y(s)) \in \partial_{1} \mathcal{N} \mathcal{T}\right\}} \mathrm{d}|k|(s),  \tag{5.2}\\
& N(t) \xlongequal{\text { def }} \frac{1}{\sqrt{(1-\mu)^{2}+1}} \int_{0}^{t} \mathbf{1}_{\left\{(s, Y(s), Y(s)) \in \partial_{2} \mathcal{N} \mathcal{T}\right\}} \mathrm{d}|k|(s) . \tag{5.3}
\end{align*}
$$

Then $(M, N)$ is a solution to the Skorokhod problem. Moreover, since $(M, N, X, Y)$ satisfies (2.1) and (2.2), we can prove that the corresponding terminal net wealth $W^{X, Y}(T)$ is the optimal solution to Problem 4.3. To prove that we need the verification theorem.

Note that one can naturally extend the definition of the Skorokhod problem to the time horizon $[s, T]$ for any $s \in[0, T)$.

Theorem 5.1 (verification theorem). Let $\varphi$ be defined in (4.15) and $V$ be the value function defined in (4.1). If the Skorokhod problem admits a solution in $[s, T]$, where $s \in[0, T)$, then

$$
V(t, x, y)=\varphi(t, x, y) \quad \forall(t, x, y) \in[s, T] \times \mathbb{R}^{2} .
$$

Proof. The proof is rather standard in the singular control literature. We give only a sketch and refer interested readers to Karatzas and Shreve [17]. Similarly to Karatzas and Shreve [17], we can show that the function $\varphi \leqslant V$ in $\overline{\mathcal{N T}}$. Moreover, if $\tau$ is a stopping time valued in $[s, T]$, where $s \in[0, T)$, and $(X, Y)$ is a solution to Problem 5.1 in $[s, \tau]$, then

$$
\varphi(s, X(s), Y(s))=\mathbf{E}[\varphi(\tau, X(\tau), Y(\tau))] .
$$

Particularly, if $\tau=T$, then $\varphi=V$ in $\overline{\mathcal{N T}} \cap[s, T] \times \mathbb{R}^{2}$, and

$$
\begin{equation*}
V(s, X(s), Y(s))=\mathbf{E}\left[\left(W^{X, Y}(T)\right)^{2}\right] . \tag{5.4}
\end{equation*}
$$

The verification theorem in $\mathcal{N} \mathcal{T}$ follows. By Lemma 4.3, we know that $V(t, \cdot, \cdot)$ is convex in $\mathbb{R}^{2}$. So we can define its subdifferential as

$$
\partial V(t, x, y) \xlongequal{\text { def }}\left\{\left(\delta_{x}, \delta_{y}\right) \mid V(t, \bar{x}, \bar{y}) \geqslant V(t, x, y)+\delta_{x} \cdot(\bar{x}-x)+\delta_{y} \cdot(\bar{y}-y) \forall(\bar{x}, \bar{y}) \in \mathbb{R}^{2}\right\}
$$

Then, we are able to utilize the convex analysis as in Shreve and Soner [31] to obtain the verification theorem in $\mathcal{B R}$ and $\mathcal{S R}$.
5.2. Solution to Skorokhod problem. Note that, in the Skorokhod problem, Problem 5.1, the reflection boundary depends on time $t$. This is very different from the standard Skorokhod problem in the literature; see, e.g., Lions and Sznitman [24]. To remove the dependence of the reflection boundary on time, we introduce a new state variable $Z(t)$ and instead consider an equivalent problem.

Problem 5.3. Given $(0, X(0), Y(0)) \in \overline{\mathcal{N T}}$, find a process of bounded total variation $k$ and a continuous process $(Z, X, Y)$ such that, for any $t \in[0, T]$,

$$
\begin{array}{r}
(Z(t), X(t), Y(t)) \in \overline{\mathcal{N T} \mathcal{T}}, \\
\mathrm{d} X(t)=r X(t) \mathrm{d} t+\gamma_{1}(X(t), Y(t)) \mathrm{d}|k|(t), \\
\mathrm{d} Y(t)=\alpha Y(t) \mathrm{d} t+\sigma Y(t) \mathrm{d} B(t)+\gamma_{2}(X(t), Y(t)) \mathrm{d}|k|(t), \\
\mathrm{d} Z(t)=\mathrm{d} t+\gamma_{3}(Z(t), X(t), Y(t)) \mathrm{d}|k|(t), \\
|k|(t)=\int_{0}^{t} \mathbf{1}_{\{(Z(s), X(s), Y(s)) \in \partial \mathcal{N} \mathcal{T}\}} \mathrm{d}|k|(s),
\end{array}
$$

where $Z(0)=0, \gamma_{3} \equiv 0$.
Clearly $Z(t) \equiv t$ because $\gamma_{3}(z, x, y) \equiv 0$. Therefore if $(k, Z, X, Y)$ solves Problem 5.3, then ( $k, X, Y$ ) solves Problem 5.2. It is worthwhile pointing out that the reflection boundary of Problem 5.3 becomes time independent.

Now let us consider Problem 5.3.
Theorem 5.2. There exists a unique solution to Problem 5.3 in $[0, T]$.
Proof. Both Lions and Sznitman [24] and Dupus and Ishii [12] have established existence and uniqueness for the Skorokhod problem on a domain with sufficiently smooth boundary. Since the $C^{\infty}$ smoothness of $x_{s}^{*}(\cdot)$ and $x_{b}^{*}(\cdot)$ is in place, the proof is similar to that of Lemma 9.3 of Shreve and Soner [31]. It is worthwhile pointing out that Shreve and Soner [31] were not concerned with the smoothness of $x_{s}^{*}$ and $x_{b}^{*}$ because they took into consideration a stationary problem which leads to time-independent policies (free boundaries).

Thanks to Theorems 5.1 and 5.2 , we have the following corollary.
Corollary 5.3. $V(t, x, y)=\varphi(t, x, y)$ for all $(t, x, y) \in[0, T] \times \mathbb{R}^{2}$.

## 6. Main results.

Theorem 6.1. For any initial position $\left(x_{0}, y_{0}\right) \in \mathscr{S}$, define

$$
(X(0), Y(0)) \xlongequal{\text { def }} \begin{cases}\left(\frac{\left(x_{0}+(1-\mu) y_{0}\right) x_{s}^{*}(0)}{x_{s}^{*}(0)+1-\mu}, \frac{x_{0}+(1-\mu) y_{0}}{x_{s}^{*}(0)+1-\mu}\right) & \text { if }\left(0, x_{0}, y_{0}\right) \in \mathcal{S R} \\ \left(x_{0}, y_{0}\right) & \text { if }\left(0, x_{0}, y_{0}\right) \in \overline{\mathcal{N} \mathcal{T}} \\ \left(\frac{\left(x_{0}+(1+\lambda) y_{0}\right) x_{b}^{*}(0)}{x_{b}^{*}(0)+1+\lambda}, \frac{x_{0}+(1+\lambda) y_{0}}{x_{b}^{*}(0)+1+\lambda}\right) & \text { if }\left(0, x_{0}, y_{0}\right) \in \mathcal{B R}\end{cases}
$$

Let $(k, Z, X, Y)$ be the solution to Problem 5.3 as stipulated in Theorem 5.2. Then $(X, Y)$ is the unique solution to the Skorokhod problem, Problem 5.1, and

$$
V\left(0, x_{0}, y_{0}\right)=\mathbf{E}\left[\left(W^{X, Y}(T)\right)^{2}\right]
$$

Moreover, the optimal strategy ( $M, N$ ) is defined by (5.2) and (5.3).
Proof. Noting that $(0, X(0), Y(0)) \in \overline{\mathcal{N T}}$, by (5.4),

$$
V(0, X(0), Y(0))=\mathbf{E}\left[\left(W^{X, Y}(T)\right)^{2}\right] .
$$

Since $V=\varphi$, it is not hard to check that

$$
V\left(0, x_{0}, y_{0}\right)=V(0, X(0), Y(0)) .
$$

The proof is complete.
As a final task before reaching the main result, we prove the existence of the Lagrange multiplier.

Proposition 6.2. For any $(x, y, z) \in \mathbb{R}^{2} \times \mathcal{D}$, there exists a unique $\ell^{*} \in \mathbb{R}$ such that

$$
V\left(0, x-\ell^{*} e^{-r T}, y\right)-\left(\ell^{*}-z\right)^{2}=\sup _{\ell \in \mathbb{R}}\left(V\left(0, x-\ell e^{-r T}, y\right)-(\ell-z)^{2}\right) .
$$

Moreover, $\ell^{*}$ is determined by

$$
\begin{equation*}
e^{-r T} V_{x}\left(0, x-\ell^{*} e^{-r T}, y\right)+2 \ell^{*}=2 z \tag{6.1}
\end{equation*}
$$

The proof is found in Appendix B.
Corollary 6.3. If $T>\frac{1}{\alpha-r} \ln \left(\frac{1+\lambda}{1-\mu}\right)$, then

$$
\hat{z}=+\infty, \quad \mathcal{D}=\left(e^{r T} x+(1-\mu) e^{r T} y^{+}-(1+\lambda) e^{r T} y^{-},+\infty\right) .
$$

Proof. Suppose $\hat{z}<+\infty$. Then by the definition of $\hat{z}$, for any $W \in \mathcal{W}_{0}^{x, y}$, we have $\mathbf{E}[W] \leqslant \hat{z}$. So for any $z>\hat{z}, \ell \geqslant 0$,

$$
\begin{aligned}
V\left(0, x-\ell e^{-r T}, y\right)-(\ell-z)^{2} & =\inf _{W \in \mathcal{W}_{0}^{x-\ell e^{-r T}, y}} \mathbf{E}\left[W^{2}-(\ell-z)^{2}\right]=\inf _{W \in \mathcal{W}_{0}^{x, y}} \mathbf{E}\left[(W-\ell)^{2}-(\ell-z)^{2}\right] \\
& =\inf _{W \in \mathcal{W}_{0}^{x, y}}\left(\mathbf{E}\left[W^{2}\right]+2 \ell(z-\mathbf{E}[W])-z^{2}\right) \\
& \geqslant 2 \ell(z-\hat{z})-z^{2} .
\end{aligned}
$$

Consequently,

$$
\sup _{\ell \in \mathbb{R}}\left(V\left(0, x-\ell e^{-r T}, y\right)-(\ell-z)^{2}\right)=+\infty .
$$

However, the proof of Proposition 6.2 (Appendix B) shows that the above supremum is finite under the condition $T>\frac{1}{\alpha-r} \ln \left(\frac{1+\lambda}{1-\mu}\right)$; see (B.1). The proof is complete.

Now we arrive at the complete solution to the MV problem, Problem 2.1.
Theorem 6.4. Problem 2.1 admits an optimal solution if and only if $z \in \widetilde{\mathcal{D}}$, where

$$
\widetilde{\mathcal{D}} \xlongequal{\text { def }} \begin{cases}\left(e^{r T} x+(1-\mu) e^{r T} y^{+}-(1+\lambda) y^{-},+\infty\right) & \text { if } T>\frac{1}{\alpha-r} \ln \left(\frac{1+\lambda}{1-\mu}\right), \\ \left(e^{r T} x+(1-\mu) e^{r T} y, e^{r T} x+(1-\mu) e^{\alpha T} y\right] & \text { if } T \leqslant \frac{1}{\alpha-r} \ln \left(\frac{1+\lambda}{1-\mu}\right), y>0, \\ \emptyset & \text { if } T \leqslant \frac{1}{\alpha-r} \ln \left(\frac{1+\lambda}{1-\mu}\right), y \leqslant 0 .\end{cases}
$$

Moreover, if $z$ is on the boundary of $\widetilde{\mathcal{D}}$, i.e., $T \leqslant \frac{1}{\alpha-r} \ln \left(\frac{1+\lambda}{1-\mu}\right)$, $y>0$, and $z=e^{r T} x+$ $(1-\mu) e^{\alpha T} y$, then the optimal strategy is $(M, N) \equiv(0,0)$; otherwise, the value function and the optimal solution are given by Theorem 6.1, in which the initial position $\left(x_{0}, y_{0}\right)=(x-$ $\left.\ell^{*} e^{-r T}, y\right)$, where $\ell^{*}$ is determined by (6.1).

Proof. If $z \notin \widetilde{\mathcal{D}}$, then there is no feasible solution by Theorem 3.3; thus Problem 2.1 admits no optimal solution. If $z=e^{r T} x+(1-\mu) e^{\alpha T} y$ while $T \leqslant \frac{1}{\alpha-r} \ln \left(\frac{1+\lambda}{1-\mu}\right)$ and $y>0$, then Theorem 3.3 again shows that the optimal strategy is $(M, N) \equiv(0,0)$.

In all the other cases, it follows from Proposition 6.2 that there exists a unique Lagrange multiplier $\ell^{*}$ such that

$$
V\left(0, x-\ell^{*} e^{-r T}, y\right)-\left(\ell^{*}-z\right)^{2}=\sup _{\ell \in \mathbb{R}}\left(V\left(0, x-\ell e^{-r T}, y\right)-(\ell-z)^{2}\right) .
$$

Appealing to Proposition 4.2, we have

$$
V\left(0, x-\ell^{*} e^{-r T}, y\right)-\left(\ell^{*}-z\right)^{2}=V_{1}(x, y ; z)-z^{2} .
$$

Theorem 6.1 then dictates that there exists an admissible strategy $\left(M^{*}, N^{*}\right) \in \mathcal{A}$ such that

$$
V\left(0, x-\ell^{*} e^{-r T}, y\right)=\mathbf{E}\left[\left(W^{X^{x-\ell^{*}} e^{-r T}, M^{*}, N^{*}, Y^{y, M^{*}, N^{*}}}(T)\right)^{2}\right]
$$

Noting that for any $(M, N) \in \mathcal{A}$, we have

$$
\begin{aligned}
X^{x-\ell^{*} e^{-r T}, M, N}(T) & =X^{x, M, N}(T)-\ell^{*} \\
Y^{y, M, N}(T) & =Y^{y, M, N}(T) \\
W^{X^{x-\ell^{*} e^{-r T}, M, N}, Y^{y, M, N}}(T) & =W^{X^{x, M, N}, Y^{y, M, N}}(T)-\ell^{*}
\end{aligned}
$$

Thus

$$
V\left(0, x-\ell^{*} e^{-r T}, y\right)=\mathbf{E}\left[\left(W^{X^{x, M^{*}, N^{*}}, Y^{y, M^{*}, N^{*}}}(T)-\ell^{*}\right)^{2}\right] .
$$

By the definition of $V$, for any $(M, N) \in \mathcal{A}$,

$$
\begin{aligned}
V\left(0, x-\ell^{*} e^{-r T}, y\right) & \leqslant \mathbf{E}\left[\left(W^{X^{x-\ell^{*}} e^{-r T}, M, N, Y^{y, M, N}}(T)\right)^{2}\right] \\
& =\mathbf{E}\left[\left(W^{X^{x, M, N}, Y^{y, M, N}}(T)-\ell^{*}\right)^{2}\right] .
\end{aligned}
$$

Therefore $W^{*}$ is optimal to Problem 4.2 with parameter $\ell^{*}$, where $W^{*}$ is defined by

$$
W^{*}=W^{X^{x, M^{*}, N^{*}}, Y^{y, M^{*}, N^{*}}}(T) .
$$

Owing to Proposition 4.1, $W^{*}$ is optimal to Problem 2.2 with parameter $\mathbf{E}\left[W^{*}\right]$. By the uniqueness of $\ell^{*}$, we have $\mathbf{E}\left[W^{*}\right]=z$. Thus $W^{*}$ is the optimal solution to both Problems 2.2 and 2.1, and $\left(M^{*}, N^{*}\right)$ is the optimal strategy.

The preceding theorem fully describes the behavior of an optimal MV investor under transaction costs. If the planning horizon is not long enough (the precise critical length depends only on the stock excess return and the transaction fees), then what could be achieved at the terminal time (in terms of the expected wealth) is rather limited. Otherwise, any terminal target is achievable by an investment strategy, while an optimal (efficient) strategy is to minimize the corresponding risk (represented by the variance). The optimal strategy is characterized by three regions (those of sell, buy, and no-trade) defined by (5.1). The implementation of the strategy is very simple: a transaction takes place only when the "adjusted" bond-stock process, $\left(X(t)-\ell^{*} e^{-r(T-t)}, Y(t)\right)$, reaches the boundary of the no-trade zone so that the process stays within the zone (if initially the process is outside of the no-trade zone, then a transaction is carried out as in Theorem 6.1 to move it instantaneously into the no-trade zone).

The optimal strategy presented here is markedly different from its no-transaction counterpart (see, e.g., [35]). With transaction costs, an investor tries not to trade unless absolutely necessary, so as to keep the "adjusted" bond-stock ratio, $\frac{x(t)-\ell^{*} e^{-r(T-t)}}{y(t)}$, between the two barriers, $x_{b}^{*}(t)$ and $x_{s}^{*}(t)$, at any given time $t$. When there is no transaction cost, however, the two barriers coincide; thus the optimal strategy is to keep the above ratio exactly at the barrier. ${ }^{3}$ This, in turn, requires the optimal strategy to trade all the time. Clearly, the strategy presented here is more consistent with the actual investors' behaviors.

Let us examine more closely the trade zone consisting of the sell and buy regions, defined in (5.1). By and large, when the adjusted bond-stock ratio, $\frac{x(t)-\ell^{*} e^{-r(T-t)}}{y(t)}$, starts to be greater than a critical barrier (namely, $x_{s}^{*}(t)$, which is time-varying), one then needs to reduce the stock holdings. When the ratio starts to be smaller than another barrier $\left(x_{b}^{*}(t)\right.$, again timevarying), one then must accumulate the stock. It is interesting to see that $y \leqslant 0$ always triggers buying; in other words, shorting the stock is never favored, and any short position must be covered immediately. The essential reason behind this is the standing assumption that $\alpha>r$, so there is no good reason to short the stock.

[^7]Another not so obvious yet extremely intriguing behavior of the optimal strategy is that when the time to maturity is short enough (precisely, when the remaining time is less than $T^{*}=\frac{1}{\alpha-r} \ln \left(\frac{1+\lambda}{1-\mu}\right)$ ), one should not buy stock any longer (except to cover a possible short position). This is seen from the fact that $\lim _{t \uparrow T_{0}} x_{b}^{*}(t)=-\infty$ as stated in Proposition 4.4, along with the definition of the buy region. Moreover, since both barriers are decreasing in time, the buy region gets smaller and the sell region gets bigger as time passes. This suggests that the investor would be less likely to buy the stock and more likely to sell the stock when the maturity date is getting closer. These phenomena, again, are in line with what prevails in practice.

We end this section with a numerical example. Consider a market with the following parameters:

$$
(\alpha, r, \sigma, \lambda, \mu, T)=(0.15,0.05,0.2,0.02,0.02,2)
$$

In this case, $T>\frac{1}{\alpha-r} \ln \left(\frac{1+\lambda}{1-\mu}\right)$. In terms of a penalty method developed by Dai and Zhong [9], we numerically solve (4.5) and then construct the two free boundaries by Proposition 4.4.

Consider an investor having the initial position $(x, y)=(-1,1)$ with an expected return $z=1.1$ at time $T$. Based on (6.1) we can calculate that $\ell^{*}=4.5069$; thus the adjusted initial position is $\left(x-\ell^{*} e^{-r T}, y\right)=(-5.078,1)$. The optimal strategy is the following. At time 0 , applying Theorem 6.1, the investor carries out a transaction so as to move his adjusted position to the boundary of $\mathcal{N} \mathcal{T}$. This is realized by buying 4.3395 units (in terms of the dollar amount) of the stock, with the new adjusted position to be ( $-9.5043,5.3395$ ). After the initial time, the investor trades only on the boundaries of the $\mathcal{N} \mathcal{T}$ region just to keep his adjusted position within the $\mathcal{N} \mathcal{T}$ region.

Next consider another investor with an initial position $(x, y)=(1,0)$ and expected return $z=1.2$. In this case $\ell^{*}=2.3690$ and $\left(x-\ell^{*} e^{-r T}, y\right)=(-1.1436,0)$. So initially the investor buys 1.5047 worth of the stock, moving the adjusted position to $(-2.6784,1.5047)$, which is on the boundary of $\mathcal{N} \mathcal{T}$. Afterwards, the trading strategy is simply to keep the adjusted position within $\mathcal{N} \mathcal{T}$. The two boundaries are depicted in Figure 1, where the horizontal axis is time $t$ and the vertical axis is the ratio between $x-\ell^{*} e^{-r(T-t)}$ and $y$. A sample path corresponding to the optimal strategy is also illustrated.

Finally we compare the MV efficiency with and without transaction costs, by plotting the respective efficient frontiers. We use the same model parameters as above and take the initial position $(x, y)=(1,0)$. Figure 2 depicts the efficient frontiers with different transaction costs. ${ }^{4}$ It is worth pointing out that the frontiers in the presence of transaction costs are still straight lines due to the availability of a risk-free asset. The figure clearly shows that the MV efficiency declines as the transaction costs increase. Indeed, one could easily derive numerically the rate of the efficiency decline with respect to the transaction costs.

Now that the frontiers are straight lines, we plot in Figure 3 the slopes of the lines (known as the prices of risk) against different times (the expiration date $T=2$ is fixed). When transaction costs are incurred, we have shown that there is a threshold time value after which one never buys stock, and hence the corresponding price of risk is 0 . The threshold values $T-\frac{1}{\alpha-r} \log ((1+\lambda) /(1-\mu))=1.6$ or 1.8 for $\lambda=\mu=0.02$ or 0.01 , respectively. These are verified by Figure 3.

[^8]

Figure 1. A sample path corresponding to optimal strategy.


Figure 2. Efficient frontiers with different transaction costs.
7. Concluding remarks. This paper investigates a continuous-time Markowitz's MV portfolio selection model with proportional transaction costs. In the terminology of stochastic control theory, this is a singular control problem. We use the Lagrangian multiplier and the PDE theory to approach the problem. The problem has been completely solved in the following sense. First, the feasibility of the model has been fully characterized by certain relationships among the parameters. Second, the value function is given via a PDE, which is analytically proven to be uniquely solvable and numerically tractable, whereas the Lagrange multiplier is determined by an algebraic equation. Third, the optimal strategy is expressed in terms of the free boundaries of the PDE. Economically, the results in the paper have revealed three critical differences arising from the presence of transaction costs. First, the expected return on the portfolio may not be achievable if the time to maturity is not long enough, while without


Figure 3. Prices of risk in time.
transaction costs, any expected return can be achieved in an arbitrarily short time. Second, instead of trading all the time so as to keep a constant adjusted ratio between the stock and bond, there exist time-dependent upper and lower boundaries so that a transaction is carried out only when the ratio is on the boundaries. Third, there is a critical time which depends only on the stock excess return and the transaction fees, such that beyond that time it is optimal not to buy stock at all. Finally, although shorting is allowed in our model, it is never favored by an optimal strategy. Our results are closer to real investment practice where people tend not to invest more in risky assets toward the end of the investment horizon.

Methodologically, this paper employs the PDE approach of Dai and Yi [8] developed for EUM (CRRA utility). In both MV (this paper) and CRRA [25, 8] cases it is shown that if the investor is holding the stock, then he should stop buying more shares if the time left is too short. This is an interesting and important feature of the finite-horizon problem with transaction costs. The intuition behind this is that the investor should not purchase any additional shares if the remaining investment period is not long enough to offset at least the transaction costs. However, there are significant differences between the MV and EUM models. ${ }^{5}$ First of all, the issue of feasibility is unique to the MV model, which itself is interesting both mathematically and economically. Second, the present paper has shown that, whenever the investor is shorting the stock, the MV problem (with transaction costs) has no feasible solution if the remaining time is short enough; otherwise one should immediately buy shares. There are no corresponding results in the EUM setting. ${ }^{6}$ Last but not least, the smoothness of the switching boundaries is proved for the first time in this paper, which is instrumental in rigorously deriving the optimal trading strategies.

[^9]Appendix A. Proofs of Propositions 4.4 and 4.5. Proposition 4.5 is straightforward once Proposition 4.4 is proved. So we prove Proposition 4.4 only. Note that (4.5) could have a singularity if $v=0$. To remove the possible singularity, let us begin with the stationary counterpart of the problem. As in Theorem 6.1 of Dai and Yi [8], we are able to show that the semiexplicit stationary solution is available through a Riccati equation

$$
\left\{\begin{array}{l}
\mathcal{L} v_{\infty}=0 \quad \text { if } x \in\left(x_{b, \infty}^{*}, x_{s, \infty}^{*}\right), \\
v_{\infty}\left(x_{b, \infty}^{*}\right)=x_{b, \infty}^{*}+1+\lambda, \quad v_{\infty}^{\prime}\left(x_{b, \infty}^{*}\right)=1 \\
v_{\infty}\left(x_{s, \infty}^{*}\right)=x_{s, \infty}^{*}+1-\mu, \quad v_{\infty}^{\prime}\left(x_{s, \infty}^{*}\right)=1
\end{array}\right.
$$

with

$$
\begin{equation*}
x_{s, \infty}^{*} \stackrel{\text { def }}{=}-\frac{a}{a+k^{*}}(1-\mu), \tag{A.1}
\end{equation*}
$$

$$
\begin{equation*}
x_{b, \infty}^{*} \xlongequal{\text { def }}-\frac{a}{a+\frac{k^{*}}{k^{*}-1}}(1+\lambda), \tag{A.2}
\end{equation*}
$$

where $\mathcal{L}$ is defined in (4.6),

$$
a \xlongequal{\text { def }}-\frac{2\left(\alpha-r+\sigma^{2}\right)}{\sigma^{2}} \in(-\infty,-2)
$$

$k^{*} \in(1,2)$ is the solution to

$$
\begin{gathered}
F(k)=\frac{1+\lambda}{1-\mu}, \\
F(k) \xlongequal{\text { def }} \begin{cases}\frac{a+\frac{k}{k-1}}{a+k}\left(\frac{\left(c_{1}+\frac{k-1}{k} a\right)\left(c_{2}+\frac{1}{k} a\right)}{\left(c_{2}+\frac{k-1}{k} a\right)\left(c_{1}+\frac{1}{k} a\right)}\right)^{\frac{1}{2\left(c_{2}-c_{1}\right)}} & \text { if } \Delta_{k}<0, \\
\frac{a+\frac{k}{k-1}}{a+k} \exp \left(\frac{1}{2}\left(\frac{1}{\frac{1}{k} a+\frac{1-a}{4}}-\frac{1}{\frac{k-1}{k} a+\frac{1-a}{4}}\right)\right) & \text { if } \Delta_{k}=0, \\
\frac{a+\frac{k}{k-1}}{a+k} \exp \left(\frac{1}{\sqrt{2 \Delta_{k}}}\left(\arctan \frac{k(a-1)-4 a}{2 k \sqrt{2 \Delta_{k}}}-\arctan \frac{4 a-k(3 a+1)}{2 k \sqrt{2 \Delta_{k}}}\right)\right) & \text { if } \Delta_{k}>0,\end{cases}
\end{gathered}
$$

$c_{1}, c_{2}$ are the two roots of $2 c^{2}+(a-1) c+\frac{k-1}{k^{2}} a^{2}=0$, and

$$
\Delta_{k} \xlongequal{\text { def }} \frac{k-1}{k^{2}} a^{2}-\frac{1}{8}(a-1)^{2} .
$$

Let $v(t, x)$ be a solution to (4.5) restricted to the region $[0, T) \times\left(-\infty, x_{s, \infty}^{*}\right)$ with a boundary condition $v\left(t, x_{s, \infty}^{*}\right)=x_{s, \infty}^{*}+1-\mu$. Apparently, $v_{\infty}$ is a supersolution to (4.5) in the region; i.e., $v(t, x) \leq v_{\infty}(x)$ for all $x<-(1-\mu), t \in[0, T)$. It is easy to show that $v_{\infty}(x)$ is increasing in $x$. We then deduce

$$
\begin{equation*}
v(t, x) \leq v_{\infty}\left(x_{s, \infty}^{*}\right)=x_{s, \infty}^{*}+1-\mu \xlongequal{\text { def }}-C_{0}<0 \quad \text { for } x<x_{s, \infty}^{*} . \tag{A.3}
\end{equation*}
$$

In what follows, we will confine (4.5) to the restricted region $[0, T) \times\left(-\infty, x_{s, \infty}^{*}\right)$ in which, due to (A.3), the equation has no singularity. It is worthwhile pointing out that $v(t, x)$ can be trivially extended to the original region by letting $v(t, x)=x+1-\mu$ for $x \geq x_{s, \infty}^{*}$.

In terms of a penalized approximation (see, for example, Friedman [14]), it is not hard to show that $v(t, x) \in W_{p}^{1,2}\left([0, T) \times\left(-N, x_{s, \infty}^{*}\right)\right)$ for any $-N<x_{s, \infty}^{*}, p>1$. By the maximum principle, (4.7) and (4.8) follow. Then we have

$$
\frac{\partial}{\partial x}[v-(x+1-\mu)]=\frac{\partial}{\partial x}[v-(x+1+\lambda)]=v_{x}-1 \leqslant 0
$$

which implies the existence of $x_{s}(\cdot)$ (or $x_{b}(\cdot)$ ) as a single-value function. The monotonicity of $x_{s}(\cdot)$ and $x_{b}(\cdot)$ is due to

$$
\frac{\partial}{\partial t}[v-(x+1-\mu)]=\frac{\partial}{\partial t}[v-(x+1+\lambda)]=v_{t} \leqslant 0 .
$$

The proof of (4.12) is the same as those of Theorems 4.5 and 4.7 of Dai and Yi [8].
It remains to show the smoothness of $x_{b}^{*}(\cdot)$ and $x_{s}^{*}(\cdot)$. To begin with, let us make a transformation and introduce two lemmas.

Let $z=\log (-x), u(t, z)=v(t, x)$. Then

$$
\left\{\begin{array}{l}
\max \left\{\min \left\{-u_{t}-\mathcal{L}_{1} u, u-\left(-e^{z}+1-\mu\right)\right\}, u-\left(-e^{z}+1+\lambda\right)\right\}=0, \\
u(T, z)=-e^{z}+1-\mu, \\
u\left(t, \log \left(-x_{s, \infty}^{*}\right)\right)=x_{s, \infty}^{*}+1-\mu,
\end{array} \quad(t, z) \in[0, T) \times \mathscr{Z},\right.
$$

where $\mathscr{Z}=\left(\log \left(-x_{s, \infty}^{*}\right),+\infty\right)$, and

$$
\mathcal{L}_{1} u=\frac{\sigma^{2}}{2} u_{z z}-\left(\alpha-r+\frac{\sigma^{2}}{2}\right) u_{z}+\left(\alpha-r+\sigma^{2}\right) u-\sigma^{2}\left[\frac{u_{z}^{2}}{u}+2 e^{z} \frac{u_{z}}{u}-2 e^{z}\right] .
$$

Lemma A.1. For any $(t, z) \in[0, T) \times \mathscr{Z}$, we have
(1) $u \leqslant-C_{0}$;
(2) $u_{z}=x v_{x} \geqslant x=-e^{z}$, i.e., $u_{z}+e^{z} \geqslant 0$;
(3) $u-u_{z} \geqslant 0$. Moreover, there is a constant $C_{1}>0$ such that $u-u_{z} \geqslant C_{1}$.

Proof. Parts (1) and (2) are immediate from (A.3) and (4.8). Let us prove part (3). Denote $w=u_{z}$. Thus,

$$
\frac{\partial}{\partial z}\left(-u_{t}-\mathcal{L}_{1} u\right)=-w_{t}-\mathcal{L}_{2} w+2 e^{z} \sigma^{2}\left(\frac{u_{z}}{u}-1\right)
$$

where
$\mathcal{L}_{2} w=\frac{\sigma^{2}}{2} w_{z z}-\left(\alpha-r+\frac{\sigma^{2}}{2}\right) w_{z}+\left(\alpha-r+\sigma^{2}\right) w-\sigma^{2}\left(\frac{2\left(u_{z}+e^{z}\right)}{u} w_{z}-\frac{u_{z}\left(u_{z}+2 e^{z}\right)}{u^{2}} w\right)$.
We define

$$
\begin{aligned}
\mathrm{SR} & =\left\{(t, z) \in[0, T) \times \mathscr{Z} \mid u=-e^{z}+1-\mu\right\}, \\
\mathrm{BR} & =\left\{(t, z) \in[0, T) \times \mathscr{Z} \mid u=-e^{z}+1+\lambda\right\}, \\
\mathrm{NT} & =\left\{(t, z) \in[0, T) \times \mathscr{Z} \mid-e^{z}+1-\mu<u<-e^{z}+1+\lambda\right\} .
\end{aligned}
$$

Notice that we can rewrite

$$
-u_{t}-\mathcal{L}_{1} u=-u_{t}-\mathcal{L}_{2} u+2 e^{z} \sigma^{2}\left(\frac{u_{z}}{u}-1\right)
$$

Denote $H=u-u_{z}$. Then

$$
-H_{t}-\mathcal{L}_{2} H=0 \quad \text { in NT. }
$$

Clearly $H=1-\mu$ in SR and at $\mathrm{t}=\mathrm{T}$, and $H=1+\lambda$ in BR. Hence, applying the maximum principle yields $H \geq 0$. Moreover, it is not hard to verify that the coefficients in $\mathcal{L}_{2} H$ are bounded. We then infer that there is a constant $C_{1}>0$, such that $H \geqslant C_{1}$.

Lemma A.2. There is a constant $C_{2}>0$ such that $u_{t} \geqslant-C_{2}$.
Proof. Let $z_{s}(t)=\log \left(-x_{s}^{*}(t)\right)$ be the corresponding selling boundary. For $z>z_{s}(t)$

$$
\begin{aligned}
\left.u_{t}\right|_{t=T} & =-\mathcal{L}_{1}\left(-e^{z}+1-\mu\right)=-(\alpha-r)(1-\mu)-\frac{(1-\mu)^{2}}{-e^{z}+1-\mu} \\
& \geqslant-(\alpha-r)(1-\mu) .
\end{aligned}
$$

Applying the maximum principle gives the desired result.
We will now prove that both $z_{s}(\cdot)$ and $z_{b}(\cdot)$ are $C^{\infty}$, where $z_{s}(t)=\log \left(-x_{s}^{*}(t)\right)$ and $z_{b}(t)=$ $\log \left(-x_{b}^{*}(t)\right)$. Thanks to the bootstrap technique, we need only show that they are Lipschitzcontinuous. Hence, it suffices to prove the cone property; namely, for any $\left(t, z_{0}\right) \in[0, T) \times \mathscr{Z}$, there exists a constant $C>0$ such that

$$
\begin{aligned}
& (T-t) u_{t}+\left.C \frac{\partial}{\partial z}\left(u-\left(-e^{z}+1-\mu\right)\right)\right|_{\left(t, z_{0}\right)} \geqslant 0 \\
& (T-t) u_{t}+\left.C \frac{\partial}{\partial z}\left(u-\left(-e^{z}+1+\lambda\right)\right)\right|_{\left(t, z_{0}\right)} \geqslant 0
\end{aligned}
$$

which is equivalent to

$$
\begin{equation*}
(T-t) u_{t}+\left.C\left(u_{z}+e^{z}\right)\right|_{\left(t, z_{0}\right)} \geqslant 0 \tag{A.4}
\end{equation*}
$$

Now let us prove (A.4). We can focus on only the NT region. Note that

$$
\frac{\partial}{\partial t}\left(-u_{t}-\mathcal{L}_{1} u\right)=\left(-\frac{\partial}{\partial t}-\mathcal{L}_{2}\right) u_{t}
$$

It follows that

$$
\left(-\frac{\partial}{\partial t}-\mathcal{L}_{2}\right)\left[(T-t) u_{t}\right]=u_{t} \quad \text { in NT }
$$

On the other hand, it is not hard to check that

$$
\begin{aligned}
\left(-\frac{\partial}{\partial t}-\mathcal{L}_{2}\right)\left(u_{z}+e^{z}\right) & =\sigma^{2} e^{z} \frac{u-u_{z}}{u^{2}}\left(u+u_{z}+2 e^{z}\right) \\
& \geqslant \sigma^{2} e^{z} \frac{u-u_{z}}{u^{2}}\left(u+u_{z}-2 u_{z}\right) \\
& =\sigma^{2} e^{z} \frac{\left(u-u_{z}\right)^{2}}{u^{2}} \geqslant \frac{C_{1}^{2} \sigma^{2} e^{z}}{u^{2}} \quad \text { in NT, }
\end{aligned}
$$

where $u-u_{z} \geqslant C_{1}$ is used in the last inequality. Thus,

$$
\begin{gathered}
\left(-\frac{\partial}{\partial t}-\mathcal{L}_{2}\right)\left[(T-t) u_{t}+C\left(u_{z}+e^{z}\right)\right] \\
\geqslant u_{t}+C \frac{C_{1}^{2} \sigma^{2} e^{z}}{u^{2}} \geqslant-C_{2}+C \frac{C_{1}^{2} \sigma^{2} e^{z}}{u^{2}} \quad \text { in NT. }
\end{gathered}
$$

Since NT is unbounded, we can follow Soner and Shreve [32] to introduce an auxiliary function $\psi\left(t, z ; z_{0}\right)=e^{a(T-t)}\left(z-z_{0}\right)^{2}$ with a constant $a>0$. We can choose $a$ big enough so that

$$
\left(-\frac{\partial}{\partial t}-\mathcal{L}_{2}\right) \psi\left(t, z ; z_{0}\right) \geqslant C_{3}\left(z-z_{0}\right)^{2}-C_{4}
$$

where $C_{3}$ and $C_{4}$ are positive constants independent of $(t, z)$. It follows that

$$
\begin{aligned}
& \left(-\frac{\partial}{\partial t}-\mathcal{L}_{2}\right)\left[(T-t) u_{t}+C\left(u_{z}+e^{z}\right)+\psi\left(t, z ; z_{0}\right)\right] \\
\geqslant & -C_{2}+C \frac{C_{1}^{2} \sigma^{2} e^{z}}{u^{2}}+C_{3}\left(z-z_{0}\right)^{2}-C_{4}
\end{aligned}
$$

Then we can choose $r>0$ such that

$$
C_{3} r^{2}-C_{2}-C_{4} \geqslant 0
$$

and choose $C>0$ big enough such that

$$
C \frac{C_{1}^{2} \sigma^{2} e^{z}}{u^{2}}-C_{2}-C_{4} \geqslant 0 \quad \text { for }\left|z-z_{0}\right| \leqslant r
$$

It then follows that

$$
\left(-\frac{\partial}{\partial t}-\mathcal{L}_{2}\right)\left[(T-t) u_{t}+C\left(u_{z}+e^{z}\right)+\psi\left(t, z ; z_{0}\right)\right] \geqslant 0 \quad \text { in NT. }
$$

Applying the maximum principle and penalty approximation, we conclude that

$$
(T-t) u_{t}+C\left(u_{z}+e^{z}\right)+\psi\left(t, z ; z_{0}\right) \geqslant 0, \quad(t, z) \in[0, T) \times \mathscr{Z} .
$$

Letting $z=z_{0}$, we get the desired result.

## Appendix B. Proof of Proposition 6.2.

## Proof. From

$$
\begin{aligned}
V\left(0, x-\ell e^{-r T}, y\right)-(\ell-z)^{2} & =\inf _{W \in \mathcal{W}_{0}^{x-\ell e^{-r T}, y}} \mathbf{E}\left[W^{2}-(\ell-z)^{2}\right]=\inf _{W \in \mathcal{W}_{0}^{x, y}} \mathbf{E}\left[(W-\ell)^{2}-(\ell-z)^{2}\right] \\
& =\inf _{W \in \mathcal{W}_{0}^{x, y}}\left(\mathbf{E}\left[W^{2}\right]-2 \ell(\mathbf{E}[W]-z)\right)
\end{aligned}
$$

it follows that $V\left(0, x-\ell e^{-r T}, y\right)-(\ell-z)^{2}$ is concave in $\ell$. So its maximum attains at point $\ell^{*}$ which satisfies

$$
\left.\frac{\partial}{\partial \ell}\left(V\left(0, x-\ell e^{-r T}, y\right)-(\ell-z)^{2}\right)\right|_{\ell=\ell^{*}}=0
$$

i.e.,

$$
e^{-r T} V_{x}\left(0, x-\ell^{*} e^{-r T}, y\right)+2 \ell^{*}=2 z
$$

Define

$$
f(\ell) \xlongequal{\text { def }} e^{-r T} V_{x}\left(0, x-\ell e^{-r T}, y\right)+2 \ell
$$

Then by the convexity of $V\left(0, x-\ell e^{-r T}, y\right)-(\ell-z)^{2}$ in $\ell$, we have that $f$ is increasing. Since $V_{x} \leqslant 0$, we have $f(z) \leqslant 2 z$. By the monotonicity of $f$, the existence of $\ell^{*}$ depends on $\lim _{\ell \rightarrow+\infty} f(\ell)$.

- First we consider the case when $T_{0}>0$. In this case $x_{b}^{*}(0) \in(-\infty, 0)$. If $y \leqslant 0$, then

$$
\left(0, x-\ell e^{-r T}, y\right) \in \mathcal{B R} \quad \forall \ell \geqslant z .
$$

If $y>0$, then

$$
\left(0, x-\ell e^{-r T}, y\right) \in \mathcal{B} \mathcal{R} \quad \forall \ell \geqslant e^{r T}\left(x-x_{b}^{*}(0) y\right) .
$$

Therefore,

$$
\begin{aligned}
\lim _{\ell \rightarrow+\infty} f(\ell) & =\lim _{\ell \rightarrow+\infty}\left(e^{-r T} V_{x}\left(0, x-\ell e^{-r T}, y\right)+2 \ell\right) \\
& =\lim _{\ell \rightarrow+\infty}\left(2 e^{-r T} e^{2 \mathfrak{B}(0)}\left(x-\ell e^{-r T}+(1+\lambda) y\right)+2 \ell\right) \\
& =\lim _{\ell \rightarrow+\infty} 2\left(1-e^{-2 r T+2 \mathfrak{B}(0)}\right) \ell+2 e^{-r T}(x+(1+\lambda) y) \\
& =+\infty,
\end{aligned}
$$

where we have used the fact that $\mathfrak{B}(0)<r T$ when $T_{0}>0$. Therefore, for any

$$
z \in\left(e^{r T} x+(1-\mu) e^{r T} y^{+}-(1+\lambda) e^{r T} y^{-},+\infty\right)
$$

there exists $\ell^{*}$ such that

$$
\begin{align*}
e^{-r T} V_{x}\left(0, x-\ell^{*} e^{-r T}, y\right)+2 \ell^{*} & =2 z \\
V\left(0, x-\ell^{*} e^{-r T}, y\right)-\left(\ell^{*}-z\right)^{2} & =\sup _{\ell \in \mathbb{R}}\left(V\left(0, x-\ell e^{-r T}, y\right)-(\ell-z)^{2}\right) . \tag{B.1}
\end{align*}
$$

Now we prove the uniqueness. For $T_{0}>0$, we have

$$
\mathfrak{A}(0)<r T, \quad \mathfrak{B}(0)<r T .
$$

If $\left(0, x-\ell e^{-r T}, y\right) \in \mathcal{S R}$, then

$$
f^{\prime}(\ell)=-e^{-2 r T} V_{x x}\left(0, x-\ell e^{-r T}, y\right)+2=-2 e^{-2 r T+2 \mathfrak{A}(0)}+2>0 .
$$

Similarly, if $\left(0, x-\ell e^{-r T}, y\right) \in \mathcal{B R}$, then

$$
f^{\prime}(\ell)=-e^{-2 r T} V_{x x}\left(0, x-\ell e^{-r T}, y\right)+2=-2 e^{-2 r T+2 \mathfrak{B}(0)}+2>0 .
$$

By the maximum principle, we have

$$
f^{\prime}(\ell)>0 \quad \text { for }\left(0, x-\ell e^{-r T}, y\right) \in \mathcal{N T} .
$$

This implies the uniqueness of $\ell^{*}$.

- Now, we move to the case when $T_{0}=0$. According to Theorem 3.3, we have

$$
\mathcal{D}= \begin{cases}\left(e^{r T} x+(1-\mu) e^{r T} y, e^{r T} x+(1-\mu) e^{\alpha T} y\right) & \text { if } y>0 \\ \emptyset & \text { if } y \leqslant 0\end{cases}
$$

We need only consider the case of $y>0$. Note that in this case,

$$
\left(0, x-\ell e^{-r T}, y\right) \in \mathcal{N} \mathcal{T} \quad \forall \ell \geqslant e^{r T}\left(x-x_{s}^{*}(0) y\right)
$$

By the homogeneity property, we have

$$
V_{x}(t, \rho x, \rho y)=\rho V_{x}(t, x, y) \quad \forall(t, x, y, \rho) \in[0, T) \times \mathbb{R}^{2} \times \mathbb{R}_{+}
$$

Thus we can make the following transformation in $\mathcal{N} \mathcal{T}$ :

$$
z=-\frac{y}{x} \in\left(0, \frac{-1}{x_{s}^{*}(0)}\right), \quad \bar{v}(t, z)=-\frac{1}{x} V_{x}(t, x, y)
$$

Then

$$
\left\{\begin{array}{l}
\bar{v}_{t}+\frac{1}{2} \sigma^{2} z^{2} \bar{v}_{z z}+(\alpha-r) z \bar{v}_{z}+2 r \bar{v}=0, \quad(t, z) \in[0, T) \times\left(0, \frac{-1}{x_{s}^{*}(0)}\right) \\
\bar{v}(T, z)=2(-1+(1-\mu) z)
\end{array}\right.
$$

Therefore

$$
\bar{v}(t, 0)=-2 e^{2 r(T-t)}
$$

Let $\tilde{v}(t, z)=\bar{v}_{z}(t, z)$, which satisfies

$$
\left\{\begin{array}{l}
\tilde{v}_{t}+\frac{1}{2} \sigma^{2} z^{2} \tilde{v}_{z z}+\left(\alpha-r+\sigma^{2}\right) z \tilde{v}_{z}+(\alpha+r) \tilde{v}=0, \quad(t, z) \in[0, T) \times\left(0, \frac{-1}{x_{s}^{*}(0)}\right) \\
\tilde{v}(T, z)=2(1-\mu)
\end{array}\right.
$$

Therefore

$$
\tilde{v}(t, 0)=2(1-\mu) e^{(\alpha+r)(T-t)}
$$

It follows that

$$
\begin{aligned}
V_{x}(0, x, y) & =-x \bar{v}\left(0,-\frac{y}{x}\right)=-x\left(\bar{v}(0,0)-\frac{y}{x} \bar{v}_{z}(0,0)+O\left(\frac{y^{2}}{x^{2}}\right)\right) \\
& =2 x e^{2 r T}+2 y(1-\mu) e^{(\alpha+r) T}+O\left(\frac{y^{2}}{|x|}\right)
\end{aligned}
$$

Thus

$$
\begin{aligned}
\lim _{\ell \rightarrow+\infty} f(\ell) & =\lim _{\ell \rightarrow+\infty}\left(e^{-r T} V_{x}\left(0, x-\ell e^{-r T}, y\right)+2 \ell\right) \\
& =\lim _{\ell \rightarrow+\infty}\left(e^{-r T}\left(2\left(x-\ell e^{-r T}\right) e^{2 r T}+2 y(1-\mu) e^{(\alpha+r) T}\right)+O\left(\frac{y^{2}}{\left|x-\ell e^{-r T}\right|}\right)+2 \ell\right) \\
& =2\left(e^{r T} x+(1-\mu) e^{\alpha T} y\right)
\end{aligned}
$$

The monotonicity of $f$ ensures the existence of $\ell^{*}$. The proof for the uniqueness is similar as above.
The proof is complete.

Acknowledgment. The authors thank the referees and the associate editor for their constructive comments which have led to a much improved version of the paper.

## REFERENCES

[1] I. Bajeux-Besnainou and R. Portait, Dynamic asset allocation in a mean-variance framework, Management Sci., 44 (1998), pp. S79-S95.
[2] S. Basak and G. Chabakauri, Dynamic Mean-Variance Asset Allocation, working paper, London Business School, London, 2008
[3] T. R. Bielecki, H. Jin, S. R. Pliska, and X. Y. Zhou, Continuous-time mean-variance portfolio selection with bankruptcy prohibition, Math. Finance, 15 (2005), pp. 213-244.
[4] T. Björk and A. Murgoci, A General Theory of Markovian Time Inconsistent Stochastic Control Problems, working paper, Stockholm School of Economics, 2008.
[5] C. H. Chiu and X. Zhou, The premium of dynamic trading, Quant. Finance, to appear.
[6] J. Cvitanic and I. Karatzas, Hedging and portfolio optimization under transaction costs: A martingale approach, Math. Finance, 6 (1996), pp. 133-165.
[7] M. Dai, L. Jiang, P. Li, and F. Yi, Finite-horizon optimal investment and consumption with transaction costs, SIAM J. Control Optim., 48 (2009), pp. 1134-1154.
[8] M. Dai and F. H. Yi, Finite-horizon optimal investment with transaction costs: A parabolic double obstacle problem, J. Differential Equations, 246 (2009), pp. 1445-1469.
[9] M. Dai and Y. F. Zhong, Penalty methods for continuous-time portfolio selection with proportional transaction costs, J. Comput. Finance, to appear.
[10] M. H. A. Davis and A. R. Norman, Portfolio selection with transaction costs, Math. Oper. Res., 15 (1990), pp. 676-713.
[11] M. H. A. Davis, V. G. Panas, and T. Zariphopoulou, European option pricing with transaction costs, SIAM J. Control Optim., 31 (1993), pp. 470-493.
[12] P. Dupus and H. Ishir, Stochastic differential equations with oblique reflections on nonsmooth domains, Ann. Probab., 21 (1993), pp. 554-580.
[13] I. Ekeland and A. Lazrak, Being Serious about Non-Commitment: Subgame Perfect Equilibrium in Continuous Time, working paper, University of British Columbia, Vancouver, 2007.
[14] A. Friedman, Variational Principles and Free-Boundary Problems, Krieger Publishing, Malabar, FL, 1988.
[15] G. Gennotte and A. Jung, Investment strategies under transaction costs: The finite horizon case, Management Sci., 40 (1994), pp. 385-404.
[16] H. Jin and X. Y. Zhou, Behavioral portfolio selection in continuous time, Math. Finance, 18 (2008), pp. 385-426.
[17] I. Karatzas and S. E. Shreve, Methods of Mathematical Finance, Springer, New York, 1998.
[18] F. C. Klebaner, Introduction to Stochastic Calculus with Applications, World Scientific, Beijing, 2004.
[19] F. E. Kydland and E. C. Prescott, Rules rather than discretion: The inconsistency of optimal plans, J. Political Economy, 85 (1977), pp. 473-492.
[20] D. Li and W. L. NG, Optimal dynamic portfolio selection: Multi-period mean-variance formulation, Math. Finance, 10 (2000), pp. 387-406.
[21] X. Li and X. Y. Zhou, Continuous-time mean-variance efficiency: The $80 \%$ rule, Ann. Appl. Probab., 16 (2006), pp. 1751-1763.
[22] A. E. B. Lim, Quadratic hedging and mean-variance portfolio selection with random parameters in an incomplete market., Math. Oper. Res., 29 (2004), pp. 132-161.
[23] A. E. B. Lim and X. Y. Zhou, Mean-variance portfolio selection with random parameters in a complete market, Math. Oper. Res., 27 (2002), pp. 101-120.
[24] P.-L. Lions and A. S. Sznitman, Stochastic differential equations with reflecting boundary conditions, Comm. Pure Appl. Math., 37 (1984), pp. 511-537.
[25] H. Liu and M. Loewenstein, Optimal portfolio selection with transaction costs and finite horizons, Rev. Financial Stud., 15 (2002), pp. 805-835.
[26] M. J. P. Magill and G. M. Constantinides, Portfolio selection with transaction costs, J. Econom. Theory, 13 (1976), pp. 245-263.
[27] H. M. Markowitz, Portfolio selection, J. Finance, 7 (1952), pp. 77-91.
[28] H. M. Markowitz and X. Y. Zhou, private communication, 2004.
[29] M. H. Protter and H. F. Weinberger, Maximum Principles in Differential Equations, Prentice-Hall, Englewood Cliffs, NJ, 1967.
[30] H. R. Richardson, A minimum variance result in continuous trading portfolio optimization, Management Sci., 35 (1989), pp. 1045-1055.
[31] S. E. Shreve and H. M. Soner, Optimal investment and consumption with transaction costs, Ann. Appl. Probab., 4 (1994), pp. 609-692.
[32] H. M. Soner and S. E. Shreve, A free boundary problem related to singular stochastic control: The parabolic case, Comm. Partial Differential Equations, 16 (1991), pp. 373-424.
[33] J. von Neumann and O. Morgenstern, Theory of Games and Economic Behavior, Princeton University Press, Princeton, NJ, 1947.
[34] J. M. XiA, Mean-variance portfolio choice: Quadratic partial hedging, Math. Finance, 15 (2005), pp. 533538.
[35] X. Y. Zhou and D. Li, Continuous-time mean-variance portfolio selection: A stochastic LQ framework, Appl. Math. Optim., 42 (2000), pp. 19-33.

# Short-Maturity Asymptotics for a Fast Mean-Reverting Heston Stochastic Volatility Model* 

Jin Feng ${ }^{\dagger}$, Martin Forde ${ }^{\ddagger}$, and Jean-Pierre Fouque ${ }^{\S}$


#### Abstract

In this paper, we study the Heston stochastic volatility model in a regime where the maturity is small but large compared to the mean-reversion time of the stochastic volatility factor. We derive a large deviation principle and compute the rate function by a precise study of the moment generating function and its asymptotic. We then obtain asymptotic prices for out-of-the-money call and put options and their corresponding implied volatilities.


Key words. stochastic volatility, Heston model, multiscale asymptotics, large deviation principle, implied volatility smile/skew

AMS subject classifications. 60F10, 91 B 70
DOI. 10.1137/090745465

1. Introduction. Large deviations theory provides a natural framework for approximating the exponentially small probabilities associated with the behavior of a diffusion process over a small time interval. In the context of financial mathematics, large deviations theory arises in the computation of small-maturity, out-of-the-money (OTM) call or put option prices or the probability of reaching a default level in a small time period. The theory of large deviations has recently been applied to local and stochastic volatility models [3, 4, 5, 6, 14, 23] and has given very interesting results on the behavior of implied volatilities near maturity. (An implied volatility is the volatility parameter needed in the Black-Scholes formula in order to match a call option price. It is common practice to quote prices in volatility through this transformation.) In the context of stochastic volatility models, the rate function involved in the large deviation estimates is given in terms of a distance function, which in general cannot be calculated in closed form. For particular models, such as the SABR model [13, 15], approximations obtained by expansion techniques have been proposed (see also [9, 12, 18]).

Multifactor stochastic volatility models have been studied during the last 10 years by many authors (see, for instance, $[8,10,12,19,20]$ ). They are quite efficient in capturing the main features of implied volatilities known as smiles and skews. They are usually not simple to calibrate, in particular with respect to the stability of parameter estimation. In the presence of separated time scales, an asymptotic theory has been proposed in [10, 11]. It has

[^10]the advantage of capturing the main effects of stochastic volatility through a small number of group parameters arising in the asymptotic. The fast time scale expansion is related to the ergodic property of the corresponding fast mean-reverting stochastic volatility factor.

In this paper, we study the Heston stochastic volatility model in the regime in which the maturity is small but large compared to the mean-reversion time of the stochastic volatility factor. This is a realistic situation where, for instance, the maturity is one month and the volatility mean-reversion time is of the order of a few days. We derive a large deviation principle and compute the rate function by a precise study of the moment generating function and its asymptotic.
1.1. The Heston model. We consider the risk-neutral Heston stochastic volatility model for the price $S_{t}$ and its square volatility $Y_{t}$ :

$$
\begin{align*}
& d S_{t}=r S_{t} d t+S_{t} \sqrt{Y_{t}} d W_{t}^{1}  \tag{1.1}\\
& d Y_{t}=\kappa\left(\theta-Y_{t}\right) d t+\nu \sqrt{Y_{t}} d W_{t}^{2}
\end{align*}
$$

where $W^{1}, W^{2}$ are two standard Brownian motions with covariation $d\left\langle W^{1}, W^{2}\right\rangle_{t}=\rho d t$, where $\rho$ is constant such that $|\rho|<1$. The short rate $r$ is constant, and throughout we assume that $2 \kappa \theta>\nu^{2}, \nu, \kappa, \theta, Y_{0}>0$, so that the square-root (or CIR) process $\left(Y_{t}\right)$ stays positive at all times (see, for instance, [17]). In this paper, we are mainly interested in small-time asymptotics for $S_{t}$ when the stochastic volatility factor $Y_{t}$ runs on a fast scale.
1.2. Fast mean-reverting stochastic volatility scaling. By fast mean-reverting stochastic volatility we mean that the rate of mean reversion $\kappa$ is large. In order to ensure that volatility is not "dying" or "exploding" we also impose that the volatility-of-volatility parameter $\nu$ be large of the order of the square root of $\kappa$. In order to achieve this scaling, we introduce a small parameter $0<\epsilon \ll 1$, and we replace $(\kappa, \nu)$ by $\left(\kappa / \epsilon^{2}, \nu / \epsilon\right)$ in (1.1)-(1.2) so that the model becomes

$$
\begin{align*}
d S_{t} & =r S_{t} d t+S_{t} \sqrt{Y_{t}} d W_{t}^{1}  \tag{1.3}\\
d Y_{t} & =\frac{\kappa}{\epsilon^{2}}\left(\theta-Y_{t}\right) d t+\frac{\nu}{\epsilon} \sqrt{Y_{t}} d W_{t}^{2} \tag{1.4}
\end{align*}
$$

The small quantity $\epsilon^{2}$ represents the intrinsic time scale of the volatility process $\left(Y_{t}\right)$, or, in other words, its decorrelation time (we refer the reader to [10] for more details). Observe that the condition $2\left(\frac{\kappa}{\epsilon^{2}}\right) \theta>\left(\frac{\nu}{\epsilon}\right)^{2}$ is equivalent to $2 \kappa \theta>\nu^{2}$ and therefore independent of $\epsilon$. Derivatives and implied volatilities have been studied extensively in [10] for a range of maturities and for general stochastic volatility processes by means of singular perturbation techniques around the Black-Scholes model. In this regime, as $\epsilon \rightarrow 0$ and fixed maturity, a call option can be approximated by its Black-Scholes price at a constant effective volatility plus a small correction of order $\epsilon$ and proportional to $\rho$, involving the Delta and Gamma of the leading order Black-Scholes price. In terms of the implied volatility surface, it turns out that the skew is asymptotically affine in $\log$-moneyness-to-maturity-ratio $\log (K / S) / T$, which leads to a particularly simple calibration procedure.
1.3. Short-maturity scaling. Since here we are interested in short maturities, but long compared with the volatility time scale $\epsilon^{2}$, we rescale time by the change of variable $t \mapsto \epsilon t$, so that typical maturities will be of the order of $\epsilon$. This regime will be of practical use when, for instance, the maturity is a couple of weeks and the time scale for the volatility to mean revert is of the order of a couple of days.

Performing the change of variable $t \mapsto \epsilon t$ in (1.3)-(1.4) gives rise to the rescaled process denoted by $\left(S_{\epsilon, t}, Y_{\epsilon, t}\right)$ and defined by

$$
\begin{align*}
d S_{\epsilon, t} & =\epsilon r S_{\epsilon, t} d t+S_{\epsilon, t} \sqrt{\epsilon Y_{\epsilon, t}} d W_{t}^{1}  \tag{1.5}\\
d Y_{\epsilon, t} & =\frac{\kappa}{\epsilon}\left(\theta-Y_{\epsilon, t}\right) d t+\frac{\nu}{\sqrt{\epsilon}} \sqrt{Y_{\epsilon, t}} d W_{t}^{2} \tag{1.6}
\end{align*}
$$

where we have used that $\left(W_{\epsilon t}^{1}, W_{\epsilon t}^{2}\right)=\left(\sqrt{\epsilon} W_{t}^{1}, \sqrt{\epsilon} W_{t}^{2}\right)$ in distribution, therefore preserving the constant correlation $\rho$.

We will use the discounted price $\tilde{S}_{\epsilon, t}=e^{-r \epsilon t} S_{\epsilon, t}$ which satisfies

$$
\begin{equation*}
d \tilde{S}_{\epsilon, t}=\tilde{S}_{\epsilon, t} \sqrt{\epsilon Y_{\epsilon, t}} d W_{t}^{1} \tag{1.7}
\end{equation*}
$$

It will also be useful to consider the $\log$-price $X_{\epsilon, t}=\log S_{\epsilon, t}$ which satisfies

$$
\begin{equation*}
d X_{\epsilon, t}=r \epsilon d t-\frac{1}{2} \epsilon Y_{\epsilon, t} d t+\sqrt{\epsilon Y_{\epsilon, t}} d W_{t}^{1} . \tag{1.8}
\end{equation*}
$$

In both cases $Y_{\epsilon, t}$ satisfies (1.6), and we note that

$$
\begin{equation*}
\tilde{S}_{\epsilon, t}=x \exp \left(-\frac{\epsilon}{2} \int_{0}^{t} Y_{\epsilon, s} d s+\sqrt{\epsilon} \int_{0}^{t} \sqrt{Y_{\epsilon, s}} d W_{s}^{1}\right) \tag{1.9}
\end{equation*}
$$

1.4. Main results. In section 2 we derive the following result, which describes the asymptotic behavior of $X_{\epsilon, t}$ as $\epsilon \rightarrow 0$ for fixed $t>0$.

Theorem 1.1. Assume $X_{\epsilon, 0}=x_{0}$. For each $t>0,\left\{X_{\epsilon, t}: \epsilon>0\right\}$ satisfies the large deviation principle with good rate function

$$
I\left(q ; x_{0}, t\right)=\Lambda^{*}\left(q-x_{0} ; 0, t\right)
$$

where $\Lambda^{*}$ is the Legendre transform of $\Lambda$

$$
\Lambda^{*}(q ; x, t) \equiv \sup _{p \in R}\{q p-\Lambda(p ; x, t)\}
$$

and $\Lambda(p ; x, t): R \times R \times R_{+} \mapsto R \cup\{+\infty\}$ is given explicitly by

$$
\begin{align*}
\Lambda(p ; x, t)= & x p+\frac{\kappa \theta t}{\nu^{2}}\left((\kappa-\nu \rho p)-\sqrt{(\kappa-\rho \nu p)^{2}-\nu^{2} p^{2}}\right)  \tag{1.10}\\
& \quad \text { for }-\frac{\kappa}{\nu(1-\rho)} \leq p \leq \frac{\kappa}{\nu(1+\rho)}, \\
=+\infty \quad & \text { otherwise. }
\end{align*}
$$

The function $\Lambda(p ; x, t)$, and the rate function $\Lambda^{*}(q)$ given below, are plotted in Figure 1 in section 2.3 in the three cases $\rho>0, \rho=0$, and $\rho<0$.

Lemma 1.2. The rate function $\Lambda^{*}$ is given explicitly by

$$
\Lambda^{*}(q ; 0, t)=q p(q ; t)-\Lambda(p(q ; t) ; 0, t),
$$

where $p(q ; t)$ is defined by

$$
\begin{align*}
p(q ; t)= & \frac{\kappa}{\nu\left(1-\rho^{2}\right)}\left(-\rho+\frac{q \nu+\kappa \theta t \rho}{\sqrt{(q \nu+\kappa \theta t \rho)^{2}+\left(1-\rho^{2}\right) \kappa^{2} \theta^{2} t^{2}}}\right)  \tag{1.11}\\
& \in \operatorname{int}(\operatorname{Dom}(\Lambda))=\left(-\frac{\kappa}{\nu(1-\rho)}, \frac{\kappa}{\nu(1+\rho)}\right)
\end{align*}
$$

$\Lambda^{*}(q ; 0, t)$ is finite for all $q \in R$; it is strictly increasing for $q>0$ and strictly decreasing for $q<0$; and $\Lambda^{*}(0 ; 0, t)=0$.
$\Lambda^{*}(q ; 0, t)$ is continuous in $(q, t) \in R \times R_{+}$.

## Remarks.

1. $\Lambda^{*}(q ; x, t)=\Lambda^{*}(q-x ; 0, t)$ since the only $x$ dependence in $\Lambda$ is the linear term $x p$.
2. Note also the scaling property $\Lambda(p ; x, t)=t \Lambda\left(p ; \frac{x}{t}, 1\right)$. In the following we choose to keep the $t$-dependence.
3. In this asymptotic regime, the limiting quantities $\Lambda$ and $\Lambda^{*}$ do not depend on the starting level of volatility $y$, and they depend on the $\kappa$ (mean-reversion rate) and $\nu$ (volatility-of-volatility) parameters only through their ratio $\nu / \kappa$.
4. The previous remark will also apply to asymptotic option prices and implied volatilities described below. In this regime, therefore, the relevant features of the Heston model are captured by just three parameters: the ergodic mean $\theta$, the correlation $\rho$, and the ratio $\nu / \kappa$. They control, respectively, the implied volatility skew's level, slope, and convexity.
The proof of Theorem 1.1 is the object of section 2 and is concluded after Lemma 2.4. The proof of Lemma 1.2 is given at the end of section 2 .

A practical application of this result is the following rare event estimate for pricing OTM options of small maturity, derived in section 2.2.

Corollary 1.3. Suppose that $\log$-moneyness is positive, $\log \left(\frac{K}{S_{0}}\right)>0$, and $t>0$ fixed. Then

$$
\lim _{\epsilon \rightarrow 0^{+}} \epsilon \log E\left[e^{-r \epsilon t}\left(S_{\epsilon, t}-K\right)^{+} \mid S_{\epsilon, 0}=S_{0}, Y_{\epsilon, 0}=y_{0}\right]=-\Lambda^{*}\left(\log \left(\frac{K}{S_{0}}\right) ; 0, t\right)
$$

independently of the initial square-volatility level $y_{0}$. Note that the maturity of the option is $T=\epsilon t$, which goes to zero in the limit. The discounting factor $e^{-r \epsilon t}$ plays no role in this asymptotic result.

Moreover, the asymptotic implied volatility can be computed. Let $\sigma_{\epsilon}(t, x)$ denote the Black-Scholes implied volatility for the European call option with strike price $K$, OTM so that $x=\log \left(K / S_{0}\right)>0$, with short maturity $T=\epsilon t$ for $t>0$ fixed, and computed under the dynamics given by (1.3), (1.4). In section 2.3 we prove the following result.

Corollary 1.4.

$$
\lim _{\epsilon \rightarrow 0^{+}} \sigma_{\epsilon}^{2}(t, x)=\frac{x^{2}}{2 \Lambda^{*}(x ; 0, t) t}, \quad x=\log \left(\frac{K}{S_{0}}\right)>0
$$

Similarly, by considering OTM put options, one obtains the same formula for $x<0$. The at-the-money (ATM) volatility is obtained by taking the limit $x \rightarrow 0$ (a precise statement is given in Lemma 2.6).

In fact, the results in Corollaries 1.3 and 1.4 hold for any fast mean-reverting stochastic volatility model (other than Heston's) which satisfies a large deviation principle, as in Theorem 1.1, provided the asymptotic rate function satisfies the following: $\Lambda^{*}(q ; 0, t)$ is finite for all $q \in R$; it is strictly increasing for $q>0$ and strictly decreasing for $q<0$; and $\Lambda^{*}(0 ; 0, t)=0$. This last remark is easily justified by going through the proofs of these results given in sections 2.2 and 2.3.
2. Moment generating function and its asymptotic. Much of our analysis relies on an explicit calculation of a moment generating function and evaluating its limit. First we define the quantity

$$
\begin{align*}
\Lambda_{\epsilon}(p) & =\Lambda_{\epsilon}(p ; x, y, t)=\epsilon \log E\left[e^{\left.\left.\frac{p}{\epsilon} X_{\epsilon, t} \right\rvert\, X_{\epsilon, 0}=x, Y_{\epsilon, 0}=y\right]}\right. \\
& =\epsilon \log E\left[\left.S_{\epsilon, t}^{\frac{p}{\epsilon}} \right\rvert\, S_{\epsilon, 0}=e^{x}, Y_{\epsilon, 0}=y\right] \\
& =\epsilon r p t+\epsilon \log E\left[\tilde{S}_{\epsilon, t}^{\frac{p}{\epsilon}} \tilde{S}_{\epsilon, 0}=e^{x}, Y_{\epsilon, 0}=y\right], \tag{2.1}
\end{align*}
$$

where $S_{\epsilon, t}, Y_{\epsilon, t}, X_{\epsilon, t}$, and $\tilde{S}_{\epsilon, t}$ are defined in section 1.3. Using (1.9) and introducing a Brownian motion $W^{3}$ independent of $W^{2}$, the moments of $\tilde{S}_{\epsilon, t}$ can be formally rewritten as follows:

$$
\begin{aligned}
& E\left[\left.\tilde{S}_{\epsilon, t}^{\frac{p}{\epsilon}} \right\rvert\, \tilde{S}_{\epsilon, 0}=e^{x}, Y_{\epsilon, 0}=y\right] \\
& =e^{\frac{x p}{\epsilon}} E\left[\left.e^{-\frac{p}{2} \int_{0}^{t} Y_{\epsilon, s} d s+\frac{p}{\sqrt{\epsilon}} \int_{0}^{t} \sqrt{Y_{\epsilon, s}} d W_{s}^{1}} \right\rvert\, Y_{\epsilon, 0}=y\right] \\
& =e^{\frac{x p}{\epsilon}} E\left[\left.e^{-\frac{p}{2} \int_{0}^{t} Y_{\epsilon, s} d s+\frac{p \rho}{\sqrt{\epsilon}} \int_{0}^{t} \sqrt{Y_{\epsilon, s}} d W_{s}^{2}+\frac{p \sqrt{1-\rho^{2}}}{\sqrt{\epsilon}} \int_{0}^{t} \sqrt{Y_{\epsilon, s}} d W_{s}^{3}} \right\rvert\, Y_{\epsilon, 0}=y\right] \\
& =e^{\frac{x p}{\epsilon}} E\left[\left.e^{-\frac{p}{2} \int_{0}^{t} Y_{\epsilon, s} d s+\frac{p \rho}{\sqrt{\epsilon}} \int_{0}^{t} \sqrt{Y_{\epsilon, s}} d W_{s}^{2}+\frac{p^{2}\left(1-\rho^{2}\right)}{2 \epsilon} \int_{0}^{t} Y_{\epsilon, s} d s} \right\rvert\, Y_{\epsilon, 0}=y\right] \\
& =e^{\frac{x p}{\epsilon}} E\left[\left.e^{\frac{p \rho}{\sqrt{\epsilon}} \int_{0}^{t} \sqrt{Y_{\epsilon, s}} d W_{s}^{2}-\frac{p^{2} \rho^{2}}{2 \epsilon} \int_{0}^{t} Y_{\epsilon, s} d s} e^{\frac{p(p-\epsilon)}{2 \epsilon}} \int_{0}^{t} Y_{\epsilon, s} d s \right\rvert\, Y_{\epsilon, 0}=y\right] \text {, }
\end{aligned}
$$

where we integrated with respect to the independent Brownian motion $W^{3}$ and redistributed the bounded variation terms. Using the Girsanov transform, one obtains that

$$
\begin{equation*}
E\left[\left.\tilde{S}_{\epsilon, t}^{\frac{p}{\epsilon}} \right\rvert\, \tilde{S}_{\epsilon, 0}=e^{x}, Y_{\epsilon, 0}=y\right]=e^{\frac{x p}{\epsilon}} E^{Q}\left[\left.e^{\frac{p(p-\epsilon)}{2 \epsilon}} \int_{0}^{t} Z_{\epsilon, s} d s \right\rvert\, Z_{\epsilon, 0}=y\right], \tag{2.2}
\end{equation*}
$$

where, under the measure $Q$, the process $Z_{\epsilon, t}$ satisfies the equation

$$
\begin{equation*}
d Z_{\epsilon, t}=\frac{1}{\epsilon}\left(\kappa \theta-(\kappa-\nu \rho p) Z_{\epsilon, t}\right) d t+\frac{\nu}{\sqrt{\epsilon}} \sqrt{Z_{\epsilon, t}} d W_{t}^{Q} \tag{2.3}
\end{equation*}
$$

driven by a Brownian motion $W^{Q}$. The result (2.2)-(2.3) is given in Lemma 2.3 in Andersen and Piterbarg [2] (with a proof in B. 1 of their supplementary material). Note that the proof
of (2.2) in Andersen and Piterbarg [2] allows the possibility of " $+\infty=+\infty$." Although the statement of their Lemma 2.3 is limited to the case of $p(p-\epsilon)>0$, the proof is not limited to that case, allowing $p \in R$.
2.1. Explicit evaluation of $\Lambda_{\epsilon}$. The following two inequalities play important roles:

$$
\begin{align*}
(\rho \nu p-\kappa)^{2} & \geq p(p-\epsilon) \nu^{2},  \tag{2.4}\\
\rho \nu p & <\kappa . \tag{2.5}
\end{align*}
$$

When (2.4) and (2.5) are both satisfied, then by results concerning exponential functionals of CIR processes (see, e.g., Corollary 3 of Albanese and Lawi [1] or Theorem 3.1 of Hurd and Kuznetsov [16]) we have

$$
E^{Q}\left[\left.e^{\frac{p(p-\epsilon)}{2 \epsilon} \int_{0}^{t} Z_{\epsilon, s} d s} \right\rvert\, Z_{\epsilon, 0}=y\right]=e^{m(t)-n(t) y}
$$

where

$$
\begin{aligned}
m(t) & =m_{\epsilon}(t)=\frac{\kappa \theta t}{\nu^{2}}(b-\bar{b})+\frac{2 \kappa \theta}{\nu^{2}} \log \left(\frac{\bar{b} e^{\bar{b} t / 2}}{\bar{b} \cosh \left(\frac{\bar{b} t}{2}\right)+b \sinh \left(\frac{\bar{b} t}{2}\right)}\right), \\
n(t) & =n_{\epsilon}(t)=\frac{-p(p-\epsilon)}{\epsilon}\left(\frac{\sinh \left(\frac{\bar{b} t}{2}\right)}{\bar{b} \cosh \left(\frac{\bar{b} t}{2}\right)+b \sinh (\overline{\bar{b} t})}\right), \\
\bar{b} & =\frac{1}{\epsilon} \sqrt{(\kappa-\nu \rho p)^{2}-\nu^{2} p(p-\epsilon)}, \\
b & =\frac{\kappa-\nu \rho p}{\epsilon},
\end{aligned}
$$

and consequently, $\Lambda_{\epsilon}$ defined in (2.1) is given explicitly by

$$
\begin{equation*}
\Lambda_{\epsilon}(p ; x, y, t)=\epsilon r p t+x p+\epsilon\left(m_{\epsilon}(t)-n_{\epsilon}(t) y\right) . \tag{2.6}
\end{equation*}
$$

Note that when the limit exists as $\epsilon \rightarrow 0^{+}$, the only contribution from $\epsilon\left(m_{\epsilon}(t)-n_{\epsilon}(t) y\right)$ comes from the first term of $m(t)$, which leads to formula (1.10) for $\Lambda(p ; x, t)$.

Next, we show that if (2.4) or (2.5) is violated, then $\Lambda_{\epsilon}=+\infty$. First, we sort out (2.4)(2.5) more explicitly. The inequality (2.4) is equivalent to

$$
c_{1, \epsilon} \leq p \leq c_{2, \epsilon}
$$

where

$$
\begin{aligned}
& c_{1, \epsilon}=\frac{(\epsilon \nu-2 \kappa \rho)-\sqrt{(\epsilon \nu-2 \kappa \rho)^{2}+4 \kappa^{2}\left(1-\rho^{2}\right)}}{2 \nu\left(1-\rho^{2}\right)} \leq 0, \\
& c_{2, \epsilon}=\frac{(\epsilon \nu-2 \kappa \rho)+\sqrt{(\epsilon \nu-2 \kappa \rho)^{2}+4 \kappa^{2}\left(1-\rho^{2}\right)}}{2 \nu\left(1-\rho^{2}\right)} \geq 0 .
\end{aligned}
$$

We denote the case $\epsilon=0$ as follows:

$$
c_{1}=-\frac{\kappa}{\nu(1-\rho)}, \quad c_{2}=\frac{\kappa}{\nu(1+\rho)} .
$$

Then in the limit $\epsilon \rightarrow 0^{+}$, (2.4) becomes

$$
\begin{equation*}
(\rho \nu p-\kappa)^{2} \geq p^{2} \nu^{2} \Leftrightarrow c_{1} \leq p \leq c_{2} \tag{2.7}
\end{equation*}
$$

Lemma 2.1. For $\epsilon$ small enough,

$$
\begin{equation*}
c_{1}<c_{1, \epsilon}<0<c_{2}<c_{2, \epsilon} . \tag{2.8}
\end{equation*}
$$

Moreover, (2.5) always holds if (2.4) is satisfied for $\epsilon$ small enough. In fact,

1. if $0<\rho<1$, then $c_{2}<c_{2, \epsilon}<\frac{\kappa}{\rho \nu}$;
2. if $-1<\rho<0$, then $\frac{\kappa}{\rho \nu}<c_{1}<c_{1, \epsilon}$;
3. if $\rho=0$, then (2.5) always holds.

Proof. Equation (2.8) follows from the definition by direct verification.
Assume that $\rho>0$. Then $(1+\rho)^{-1}<\rho^{-1}$, and therefore $c_{2}=\frac{\kappa}{\nu(1+\rho)}<\frac{\kappa}{\rho \nu}$, since $c_{2}<c_{2, \epsilon}$ and $\lim _{\epsilon} c_{2, \epsilon}=c_{2}, c_{2}<c_{2, \epsilon}<\frac{\kappa}{\rho \nu}$ when $\epsilon$ is small enough.

The other case follows by a similar computation (note that if $-1<\rho<0$, then $\rho^{-1}<$ $-(1-\rho)^{-1}$, implying $\left.\frac{\kappa}{\rho \nu}<c_{1}\right)$.

We have the following result.
Lemma 2.2. $\Lambda_{\epsilon}(p)$ is lower semicontinuous and convex in $p$. For $\epsilon>0$ small enough, (2.6) holds when $c_{1, \epsilon} \leq p \leq c_{2, \epsilon}$.

Proof. The lower semicontinuity and convexity of $\Lambda_{\epsilon}(p)$ follow from its definition as a logarithmic transform of moment generating function for $X_{\epsilon, t}$.

The other conclusion follows from Lemma 2.1.
Using the convexity of $\Lambda_{\epsilon}$, we conclude next that $\Lambda_{\epsilon}(p)=+\infty$ whenever $p \notin\left[c_{1, \epsilon}, c_{2, \epsilon}\right]$ (again, when $\epsilon>0$ is small enough). This is implied by the behavior of $\partial_{p} \Lambda_{\epsilon}$ for $p \in\left(c_{1, \epsilon}, c_{2, \epsilon}\right)$. From (2.6), we get

$$
\begin{aligned}
\partial_{p} \Lambda_{\epsilon}= & \epsilon r t+x+\epsilon \partial_{p} m-\epsilon y \partial_{p} n \\
=\epsilon r t+ & x+\epsilon \frac{2 \kappa \theta}{\nu^{2}}\left(\frac{\partial \bar{b}}{\bar{b}}+\frac{t}{2} \partial b-\frac{(\partial \bar{b})\left(1+\frac{t}{2} b\right) \cosh \left(\frac{\bar{b} t}{2}\right)+(\partial b)\left(1+\frac{t}{2} \bar{b}\right) \sinh \left(\frac{\bar{b} t}{2}\right)}{\bar{b} \cosh \left(\frac{\bar{b} t}{2}\right)+b \sinh \left(\frac{\bar{b} t}{2}\right)}\right) \\
& +y\left(\frac{(2 p-\epsilon) \sinh \left(\frac{\bar{b}}{2} t\right)+\left(p^{2}-p \epsilon\right)(\partial \bar{b}) \frac{t}{2} \cosh \left(\frac{\bar{b}}{2} t\right)}{\bar{b} \cosh \left(\frac{\bar{b} t}{2}\right)+b \sinh \left(\frac{\bar{b} t}{2}\right)}\right. \\
& \left.\quad-\left(p^{2}-p \epsilon\right) \frac{\left(1+\frac{t}{2} \bar{b}\right)(\partial b) \sinh ^{2}\left(\frac{\bar{b}}{2} t\right)+(\partial \bar{b})\left(1+\frac{t}{2} b\right) \sinh \left(\frac{\bar{b}}{2} t\right) \cosh \left(\frac{\bar{b}}{2} t\right)}{\left(\bar{b} \cosh \left(\frac{\bar{b} t}{2}\right)+b \sinh \left(\frac{\bar{b} t}{2}\right)\right)^{2}}\right) .
\end{aligned}
$$

When $p \rightarrow c_{1, \epsilon}^{+}$, or $c_{2, \epsilon}^{-}$, then $\bar{b} \rightarrow 0^{+}$and $\partial_{p} \bar{b} \rightarrow \pm \infty$. By Lemma 2.1, in both cases we have $\lim _{p} b>0$. Therefore, by the above expression for $\partial_{p} \Lambda_{\epsilon}$, we conclude that, for $\epsilon>0$ small enough,

$$
\begin{equation*}
\lim _{p \rightarrow c_{1, \epsilon}^{+}} \partial_{p} \Lambda_{\epsilon}(p)=-\infty, \quad \lim _{p \rightarrow c_{2, \epsilon}^{-}} \partial_{p} \Lambda_{\epsilon}(p)=\infty . \tag{2.9}
\end{equation*}
$$

Since $\Lambda_{\epsilon}$ is convex and lower semicontinuous, the limits (2.9) imply the following result.

Lemma 2.3. For $\epsilon>0$ small enough, $\Lambda_{\epsilon}$ is given by (2.6) when $c_{1, \epsilon} \leq p \leq c_{2, \epsilon}$, and $+\infty$ otherwise.

Proof. The proof follows from Lemma 3.1 given in the appendix.
The following result follows easily.
Lemma 2.4. The function $\Lambda$ given in (1.10) is lower semicontinuous and essentially smooth in p. Moreover, $\Lambda_{\epsilon}(\cdot ; x, y, t) \Gamma$-converges to $\Lambda(\cdot ; x, t)$ (see Definition 3.2). In particular, for each $x \in R, y>0, t>0$, we have the following:

1. For every $p \in R$, there exists $\left\{p_{\epsilon}\right\}$ with $p_{\epsilon} \rightarrow p$ such that

$$
\lim _{\epsilon \rightarrow 0^{+}} \Lambda_{\epsilon}\left(p_{\epsilon} ; x, y, t\right)=\Lambda(p ; x, t) .
$$

2. For every $p \in R$ and every $p_{\epsilon} \rightarrow p$,

$$
\liminf _{\epsilon \rightarrow 0^{+}} \Lambda_{\epsilon}\left(p_{\epsilon} ; x, y, t\right) \geq \Lambda(p ; x, t) .
$$

By Lemma 3.3, Theorem 1.1 follows.
We now give the proof of Lemma 1.2, where we derive an explicit formula for $\Lambda^{*}$, the Legendre transform of $\Lambda$ defined by (1.10).

Proof of Lemma 1.2. By the essential smoothness property of $\Lambda(p)=\Lambda(p ; 0, t)$ in $p$, the equation

$$
\frac{\partial}{\partial p}(q p-\Lambda(p))=0
$$

has a solution $p \in \operatorname{int}(\operatorname{Dom}(\Lambda))=\left(-\frac{\kappa}{\nu(1-\rho)}, \frac{\kappa}{\nu(1+\rho)}\right)$, which equivalently solves

$$
\begin{equation*}
q=\partial_{p} \Lambda=\frac{\kappa \theta t}{\nu}\left(-\rho+\frac{(\kappa-\rho \nu p) \rho+\nu p}{\sqrt{(\kappa-\rho \nu p)^{2}-\nu^{2} p^{2}}}\right) \tag{2.10}
\end{equation*}
$$

If $\nu q=-\rho \kappa \theta t$, it follows from (2.10) that $p=-\frac{\kappa \rho}{\nu\left(1-\rho^{2}\right)}$ and therefore that (1.11) is satisfied.

If $\nu q \neq-\rho \kappa \theta t$, then (2.10) is a quadratic equation in $p$ with the sign condition

$$
\frac{\kappa \rho+\nu\left(1-\rho^{2}\right) p}{\nu q+\rho \kappa \theta t}>0 .
$$

One can easily verify that $p$ given by (1.11) is the only root satisfying the sign condition. Consequently, the expression of $\Lambda^{*}(q ; 0, t)$ follows.

It follows by direct verification that $\Lambda^{*}(0 ; 0, t)=0$, and that $\Lambda^{*}(q ; 0, t)$ is continuous, finitely defined for all $q \in R$.

Also by direct calculation, $\partial_{p} \Lambda(0 ; 0, t)=0$ and

$$
\partial_{p p} \Lambda(p ; 0, t)=\frac{\kappa^{3} \theta t}{\left((\kappa-\rho \nu p)^{2}-\nu^{2} p^{2}\right)^{3 / 2}}>0 .
$$

Therefore, for $p \in \operatorname{int}(\operatorname{Dom}(\Lambda)), \partial_{p} \Lambda(p ; 0, t)$ is negative when $p<0$ and is positive when $p>0$. By a convex analysis result, $q=\partial_{p} \Lambda(p ; 0, t), p \in \operatorname{int}(\operatorname{Dom}(\Lambda))$ if and only if $p=\partial_{q} \Lambda^{*}(q ; 0, t)$. Consequently, $\Lambda^{*}(q ; 0, t)$ is strictly increasing when $q>0$ and strictly decreasing for $q<0$, and it achieves its minimum (zero) when $q=0$.
2.2. Pricing. We now prove Corollary 1.3.

Recall that $S_{\epsilon, t}=e^{X_{\epsilon, t}}$ and $S_{\epsilon, 0}=S_{0}$. For $\delta>0$ we have

$$
\begin{align*}
E\left[\left(S_{\epsilon, t}-K\right)^{+}\right] & \geq E\left[\mathbf{1}_{\left\{S_{\epsilon, t}-K>\delta\right\}}\left(S_{\epsilon, t}-K\right)^{+}\right]  \tag{2.11}\\
& \geq \delta P\left(S_{\epsilon, t}>K+\delta\right) .
\end{align*}
$$

By Theorem 1.1, it follows that

$$
\begin{aligned}
\liminf _{\epsilon \rightarrow 0^{+}} \epsilon \log E\left[\left(S_{\epsilon, t}-K\right)^{+}\right] & \geq \liminf _{\epsilon \rightarrow 0^{+}} \epsilon \log P\left(X_{\epsilon, t}>\log (K+\delta)\right) \\
& \geq-\inf _{q>\log (K+\delta)} \Lambda^{*}\left(q-\log S_{0} ; 0, t\right)=-\Lambda^{*}\left(\log \left(\frac{K+\delta}{S_{0}}\right) ; 0, t\right) .
\end{aligned}
$$

The last equality follows from the fact that $\log \left(\frac{K}{S_{0}}\right)>0$ and that $\Lambda^{*}(q ; 0, t)$ is nondecreasing for $q$ in the region $q \geq 0$ (see Lemma 1.2). Taking $\delta \rightarrow 0^{+}$, by continuity of $\Lambda^{*}$, we obtain the desired lower bound.

To show the upper bound, we note that for $p, q>1$ such that $p^{-1}+q^{-1}=1$,

$$
E\left[\left(S_{\epsilon, t}-K\right)^{+}\right] \leq E^{1 / p}\left[\left|\left(S_{\epsilon, t}-K\right)^{+}\right|^{p}\right] E^{1 / q}\left[\mathbf{1}_{\left\{S_{\epsilon, t}-K \geq 0\right\}}\right] .
$$

Therefore

$$
\begin{aligned}
\epsilon \log E\left[\left(S_{\epsilon, t}-K\right)^{+}\right] & \leq \frac{\epsilon}{p} \log E\left[\left(S_{\epsilon, t}\right)^{p}\right]+\epsilon\left(1-\frac{1}{p}\right) \log P\left(S_{\epsilon, t} \geq K\right) \\
& \leq \frac{1}{p} \Lambda_{\epsilon}(\epsilon p)+\left(1-\frac{1}{p}\right) \epsilon \log P\left(S_{\epsilon, t} \geq K\right) .
\end{aligned}
$$

Taking $\lim _{p \rightarrow+\infty} \lim \sup _{\epsilon \rightarrow 0^{+}}$on both sides and noting that $\lim _{\epsilon \rightarrow 0^{+}} \Lambda_{\epsilon}(\epsilon p)=0$, we deduce (by Theorem 1.1) the desired upper bound

$$
\limsup _{\epsilon \rightarrow 0^{+}} \epsilon \log E\left[\left(S_{\epsilon, t}-K\right)^{+}\right] \leq-\Lambda^{*}\left(\log \left(\frac{K}{S_{0}}\right) ; 0, t\right) .
$$

2.3. Implied volatility. We prove Corollary 1.4 , which gives the asymptotic behavior of the implied volatility $\sigma_{\epsilon}(t, x)$. Throughout, we denote the log-moneyness by $x=\log \left(K / S_{0}\right)>0$, and for simplicity $\sigma_{\epsilon}(t, x)=\sigma_{\epsilon}, t$ and $x$ being fixed in the following analysis.

First, we show that

$$
\begin{equation*}
\lim _{\epsilon \rightarrow 0^{+}} \sigma_{\epsilon} \sqrt{\epsilon t}=0 \tag{2.12}
\end{equation*}
$$

By Lemma $1.2, \Lambda^{*}(x ; 0, t)>0$. Let $0<\delta<\Lambda^{*}(x ; 0, t)$. By the definition of $\sigma_{\epsilon}$ and Corollary 1.3, for $\epsilon>0$ small enough

$$
\begin{aligned}
e^{-\left(\Lambda^{*}(x ; 0, t)-\delta\right) / \epsilon} & \geq E\left[\left(S_{\epsilon, t}-K\right)^{+}\right] \\
& =e^{r \epsilon t} S_{0} \Phi\left(\frac{-x+r \epsilon t+\frac{1}{2} \sigma_{\epsilon}^{2} \epsilon t}{\sigma_{\epsilon} \sqrt{\epsilon t}}\right)-K \Phi\left(\frac{-x+r \epsilon t-\frac{1}{2} \sigma_{\epsilon}^{2} \epsilon t}{\sigma_{\epsilon} \sqrt{\epsilon t}}\right),
\end{aligned}
$$

where we have used the Black-Scholes formula and denoted by $\Phi$ the $\mathcal{N}(0,1)$ cumulative distribution function. Since $E\left[\left(S_{\epsilon, t}-K\right)^{+}\right] \geq 0$, one deduces that the right-hand side must converge to zero as $\epsilon \rightarrow 0^{+}$. Let $l \geq 0$ be the limit of $\sigma_{\epsilon} \sqrt{\epsilon t}$ along a converging subsequence; then $l$ must satisfy

$$
S_{0} \Phi\left(-\frac{x}{l}+\frac{l}{2}\right)-K \Phi\left(-\frac{x}{l}-\frac{l}{2}\right)=0,
$$

with $x=\log \left(K / S_{0}\right)>0$. One can easily check that $l=0$ is the only solution, and therefore (2.12) holds.

The following estimate on $\Phi$ using its derivative denoted by $\phi$ is classical and will be useful (we refer the reader to [21, section 14.8], for instance).

Lemma 2.5. For $x>0$,

$$
\begin{equation*}
\left(x+\frac{1}{x}\right)^{-1} \phi(x) \leq 1-\Phi(x) \leq \frac{1}{x} \phi(x) . \tag{2.13}
\end{equation*}
$$

Next, we establish the lower bound for the limit in Corollary 1.4. We will use the classical notation

$$
d_{1}=\frac{\log \left(\frac{S_{0}}{K}\right)+r \epsilon t+\frac{\sigma_{\epsilon}^{2}}{2} \epsilon t}{\sigma_{\epsilon} \sqrt{\epsilon t}} .
$$

Let $\delta>0$; by the definition of $\sigma_{\epsilon}(t)$ and Corollary 1.3, for $\epsilon>0$ small enough, we have

$$
\begin{aligned}
e^{-\left(\Lambda^{*}(x ; 0, t)+\delta\right) / \epsilon} & \leq E\left[\left(S_{\epsilon, t}-K\right)^{+}\right] \\
& \leq e^{r \epsilon t} S_{0} \Phi\left(d_{1}\right)=e^{r \epsilon t} S_{0}\left(1-\Phi\left(-d_{1}\right)\right) \\
& \leq e^{r \epsilon t} S_{0}\left(\frac{1}{-d_{1}}\right) \phi\left(-d_{1}\right)
\end{aligned}
$$

where the last line follows from (2.13). By (2.12) and $S_{0}<K$, we know that $\lim _{\epsilon \rightarrow 0^{+}} d_{1}=-\infty$. Taking $(\epsilon \log )$ on both sides, one sees that the leading order term on the right-hand side is given by

$$
-\epsilon \frac{\left(\log \left(\frac{S_{0}}{K}\right)\right)^{2}}{2\left(\sigma_{\epsilon} \sqrt{\epsilon t}\right)^{2}}=-\frac{x^{2}}{2 \sigma_{\epsilon}^{2} t} .
$$

Therefore any limit point of $\sigma_{\epsilon}$ along a converging subsequence $\left(\epsilon_{n}\right)$ will satisfy

$$
\begin{equation*}
-\left(\Lambda^{*}(x, ; 0, t)+\delta\right) \leq-\frac{x^{2}}{2 \lim _{\epsilon_{n} \rightarrow 0^{+}} \sigma_{\epsilon_{n}}^{2} t} \tag{2.14}
\end{equation*}
$$

for all $\delta>0$ and consequently the desired lower bound.
Next, we justify the upper bound. To avoid confusion, we denote by $P$ the measure under which $S_{\epsilon}$ is defined in section 1.3 and by $P_{B S}=P_{B S}\left(\sigma_{\epsilon}\right)$ the measure under which $S_{\epsilon}$ follows the Black-Scholes model with constant volatility $\sigma_{\epsilon}=\sigma_{\epsilon}(t, x)$ :

$$
d S_{\epsilon, s}=S_{\epsilon, s}\left(r d s+\sigma_{\epsilon} d W_{s}\right),
$$

where $W$ is a Brownian motion under $P_{B S}$ (note that here $t$ is fixed and the maturity of the call option is $\epsilon t$ ). Then, using the classical notation

$$
d_{2}=\frac{\log \left(\frac{S_{0}}{K+\delta}\right)+r \epsilon t-\frac{\sigma_{\epsilon}^{2}}{2} \epsilon t}{\sigma_{\epsilon} \sqrt{\epsilon t}}
$$

one obtains

$$
\begin{aligned}
e^{-\left(\Lambda^{*}(x ; 0, t)-\delta\right) / \epsilon} & \geq E^{P}\left[\left(S_{\epsilon, t}-K\right)^{+}\right]=E^{P_{B S}}\left[\left(S_{\epsilon, t}-K\right)^{+}\right] \\
& \geq \delta P_{B S}\left(S_{\epsilon, t}>K+\delta\right) \\
& =\delta\left(1-\Phi\left(-d_{2}\right)\right) \\
& \geq \delta\left(\frac{-d_{2}}{1+d_{2}^{2}}\right) \phi\left(-d_{2}\right)
\end{aligned}
$$

where the second inequality follows by (2.11). Arguing as above, in the case of the lower bound, we know that $\lim _{\epsilon \rightarrow 0^{+}} d_{2}=-\infty$. Taking $(\epsilon \log )$ on both sides, the leading order term on the right-hand side is given by

$$
-\frac{\left(\log \left(\frac{S_{0}}{K+\delta}\right)\right)^{2}}{2 \sigma_{\epsilon}^{2} t}
$$

and therefore, along any converging subsequence,

$$
-\left(\Lambda^{*}(x ; 0, t)-\delta\right) \geq-\frac{\left(\log \left(\frac{K+\delta}{S_{0}}\right)\right)^{2}}{2 \lim _{\epsilon_{n} \rightarrow 0^{+}} \sigma_{\epsilon_{n}}^{2} t}
$$

Sending $\delta \rightarrow 0^{+}$gives the desired upper bound, which concludes the proof of Corollary 1.4.
To summarize, we proved that in this regime (fast mean-reverting volatility and short maturity) the asymptotic implied volatility of an OTM call option $(x>0)$ is given by

$$
\sigma(t, x)^{2}=\frac{x^{2}}{2 \Lambda^{*}(x ; 0, t) t},
$$

where $\Lambda^{*}$ is given in Lemma 1.2. The same formula for $x<0$ is derived similarly by considering OTM put options. Using the explicit formula for $\Lambda^{*}(x ; 0, t)$, one can derive the ATM limit

$$
\lim _{x \rightarrow 0} \sigma(t, x)^{2}=\theta
$$

by checking that near zero $p(q ; t)=\frac{q}{\theta t}+O\left(q^{2}\right)$ and $\Lambda(p ; 0, t)=\frac{\theta t}{2} p^{2}+O\left(p^{3}\right)$, and consequently

$$
\Lambda^{*}(q ; 0, t)=q\left(\frac{q}{\theta t}\right)-\frac{\theta t}{2}\left(\frac{q}{\theta t}\right)^{2}+O\left(q^{3}\right)=\frac{q^{2}}{2 \theta t}+O\left(q^{3}\right)
$$

In fact, we can also derive the limit as $\epsilon \rightarrow 0^{+}$of the ATM volatility $\sigma_{\epsilon}(t, 0)$.

Lemma 2.6. The asymptotic ATM volatility is given by

$$
\lim _{\epsilon \rightarrow 0^{+}} \sigma_{\epsilon}(t, 0)^{2}=\lim _{x \rightarrow 0} \sigma(t, x)^{2}=\theta
$$

Proof. This is not a large deviation result but rather an averaging result of the type studied in [10]. Since it involves convergence in distribution, it is more convenient to work with put options whose payoffs are continuous and bounded. The ATM volatility is defined by the unique positive number $\sigma_{\epsilon}(0, t)$ satisfying

$$
E\left[\left(S_{0}-S_{\epsilon, t}\right)^{+}\right]=S_{0} \Phi\left(-d_{2}\right)-e^{r \epsilon t} S_{0} \Phi\left(-d_{1}\right),
$$

where here

$$
\begin{equation*}
d_{1,2}=\frac{\left(r \pm \frac{1}{2} \sigma_{\epsilon}(0)^{2}\right) \sqrt{\epsilon t}}{\sigma_{\epsilon}(0)}, \tag{2.15}
\end{equation*}
$$

and we have denoted $\sigma_{\epsilon}(0, t)=\sigma_{\epsilon}(0)$ since $t$ is fixed. Using (1.5) and dividing on both sides by $\sqrt{\epsilon} S_{0}$, one gets

$$
\begin{gather*}
E\left[\left(-\sqrt{\epsilon} \int_{0}^{t} r \frac{S_{\epsilon, s}}{S_{0}} d s-\int_{0}^{t} \frac{S_{\epsilon, s}}{S_{0}} \sqrt{Y_{\epsilon, s}} d W_{s}^{1}\right)^{+}\right]  \tag{2.16}\\
=\frac{1}{\sqrt{\epsilon}}\left(\Phi\left(-d_{2}\right)-e^{r \epsilon t} \Phi\left(-d_{1}\right)\right)
\end{gather*}
$$

One easily obtains the convergence in probability to zero of the following integrals:

$$
\sqrt{\epsilon} \int_{0}^{t} r \frac{S_{\epsilon, s}}{S_{0}} d s \quad \text { and } \quad \int_{0}^{t}\left(\frac{S_{\epsilon, s}}{S_{0}}-1\right) \sqrt{Y_{\epsilon, s}} d W_{s}^{1} .
$$

The convergence of the quadratic variation of the martingale term, $\int_{0}^{t} Y_{\epsilon, s} d s \rightarrow \bar{\sigma}^{2} t$, implies the convergence in distribution

$$
\left(-\sqrt{\epsilon} \int_{0}^{t} r \frac{S_{\epsilon, s}}{S_{0}} d s-\int_{0}^{t} \frac{S_{\epsilon, s}}{S_{0}} \sqrt{Y_{\epsilon, s}} d W_{s}^{1}\right) \rightarrow \int_{0}^{t} \bar{\sigma} d W_{s}^{1}=\bar{\sigma} W_{t}^{1},
$$

where $\bar{\sigma}^{2}$ is the ergodic average of the square volatility $Y_{\epsilon, \text {, }}$, that is,

$$
\bar{\sigma}^{2}=\int_{0}^{+\infty} y \Gamma(d y)
$$

where $\Gamma$ is the invariant distribution of the ergodic process $Y$ defined by (1.2). A complete proof of this result involves introducing a solution $\psi$ of the Poisson equation

$$
\mathcal{L} \psi(y)=y-\bar{\sigma}^{2},
$$

where $\mathcal{L}$ is the infinitesimal generator of the process $Y$, and using Itô's formula to show that

$$
\int_{0}^{t}\left(Y_{\epsilon, s}-\bar{\sigma}^{2}\right) d s=\int_{0}^{t} \mathcal{L} \psi\left(Y_{\epsilon, s}\right) d s=\epsilon\left(\psi\left(Y_{\epsilon, t}\right)-\psi\left(Y_{0}\right)\right)-\sqrt{\epsilon} \int_{0}^{t} \sigma \psi^{\prime}\left(Y_{\epsilon, s}\right) Y_{\epsilon, s} d W_{s}^{2}
$$



Implied Volatility in the small-epsilon limit


Figure 1. Here we have plotted $\Lambda, \Lambda^{*}$, and the implied volatility in the small- limit as a function of the $\log$-moneyness $x=\log \left(K / S_{0}\right)$. The parameters are $t=1$, ergodic mean $\theta=.04$, convexity $\nu / \kappa=1.74$ ( $\kappa=1.15, \nu=.2$ ), and skew $\rho=-.4$ (dashed blue), $\rho=0$ (solid black), $\rho=+.4$ (dotted red).
converges to zero (we refer the reader to [10] for details).
In this case, the invariant distribution is a Gamma with mean $\theta$, and consequently $\bar{\sigma}^{2}=\theta$. Therefore, the left-hand side of (2.16) converges to $E\left[\left(\bar{\sigma} W_{t}^{1}\right)^{+}\right]=\bar{\sigma} \sqrt{t} / \sqrt{2 \pi}=\sqrt{\theta t} / \sqrt{2 \pi}$. By direct inspection of the right-hand side of (2.16) and the relation (2.15) between $d_{1,2}$ and $\sigma_{\epsilon}(0)$, one deduces that $\sigma_{\epsilon}(0)$ must converge to $\theta$ as $\epsilon \rightarrow 0^{+}$.

In Figure 1 we show plots of the functions $\Lambda$ and $\Lambda^{*}$ and of the implied volatility smile/skew obtained in the limit $\epsilon \rightarrow 0^{+}$.

## 3. Appendix.

### 3.1. A property of convex functions in $R$.

Lemma 3.1. Suppose $\Lambda: R \mapsto \bar{R}$ is convex and for some $c \in R$

$$
\liminf _{x \rightarrow c^{-}} \Lambda(x)>-\infty \quad \text { and } \quad \lim _{x \rightarrow c^{-}} \partial \Lambda(x)=+\infty
$$

then $\Lambda(y)=+\infty$ for all $y>c$. Similarly, if for some $c \in R$

$$
\limsup _{x \rightarrow c^{+}} \Lambda(x)>-\infty \quad \text { and } \quad \lim _{x \rightarrow c^{+}} \partial \Lambda(x)=-\infty
$$

then $\Lambda(y)=+\infty$ for all $y<c$.
Proof. Let $y>c>x$, and denote $\delta=y-c>0$. Then

$$
\Lambda(y) \geq \Lambda(x)+\partial \Lambda(x)(y-x)
$$

Taking $x \rightarrow c^{-}$gives

$$
\Lambda(y) \geq \liminf _{x \rightarrow c^{-}} \Lambda(x)+\limsup _{x \rightarrow c^{-}} \partial \Lambda(x) \delta=+\infty
$$

3.2. Gärtner-Ellis theorem via $\Gamma$-convergence. We generalize the Gärtner-Ellis theorem (see, e.g., Theorem 2.3.6 in Dembo and Zeitouni [7]) for Euclidean space valued random variables.

Definition 3.2. Let sequence $\Lambda_{n}, \Lambda: R^{d} \mapsto \bar{R}$. We say that $\Lambda_{n} \Gamma$-converges to $\Lambda$ (denoted $\Lambda_{n} \xrightarrow{\Gamma} \Lambda$ ) if, for all $p \in R^{d}$,

1. (limsup inequality) there exists a sequence of $\left\{p_{n}\right\}$ converging to $p$ such that

$$
\Lambda(p) \geq \limsup _{n \rightarrow \infty} \Lambda_{n}\left(p_{n}\right)
$$

2. (liminf inequality) for every sequence $\left\{p_{n}\right\}$ converging to $p$, we have

$$
\Lambda(p) \leq \liminf _{n \rightarrow \infty} \Lambda_{n}\left(p_{n}\right)
$$

Let $\left\{X_{n}: n=1,2, \ldots\right\}$ be a sequence of $R^{d}$-valued random variables, and denote

$$
\Lambda_{n}(p)=\frac{1}{n} \log E\left[e^{n p X_{n}}\right], \quad p \in R^{d}
$$

Lemma 3.3. Suppose that the limsup property in $\Gamma$-convergence holds for $\Lambda_{n}$ to $a \Lambda$ : $R^{d} \mapsto \bar{R}$. Then the large deviation upper bound holds for all compact $F \subset R^{d}$ :

$$
\begin{equation*}
\limsup _{n \rightarrow \infty} \frac{1}{n} \log P\left(X_{n} \in F\right) \leq-\inf _{x \in F} \Lambda^{*}(x) \tag{3.1}
\end{equation*}
$$

Furthermore, if $0 \in$ interior $(D(\Lambda))$, then $\left\{X_{n}\right\}$ is exponentially tight and (3.1) holds for all closed $F \subset R^{d}$.

In addition to the above, suppose that the liminf property in $\Gamma$-convergence holds for $\Lambda_{n}$ to a $\Lambda: R^{d} \mapsto \bar{R}$, and assume $\Lambda(0)=0$. Then the following upper bound holds:

$$
\begin{equation*}
\liminf _{n \rightarrow \infty} \frac{1}{n} \log P\left(X_{n} \in G\right) \geq-\inf _{x \in G \cap \mathcal{F}} \Lambda^{*}(x), \quad G \text { open in } R^{d}, \tag{3.2}
\end{equation*}
$$

where $\mathcal{F}$ is the set of exposed points of $\Lambda^{*}$ with exposing hyperplane in interior $(D(\Lambda))$, where $D(\Lambda)=\{x: \Lambda(x)<\infty\}$.

If $\Lambda$ is lower semicontinuous and essentially smooth (see, e.g., Definition 2.3.5 in Dembo and Zeitouni [7]), then $\inf _{x \in G \cap \mathcal{F}} \Lambda^{*}(x)=\inf _{x \in G} \Lambda^{*}(x)$ for all $G$ open, and $\Lambda^{*}$ is a good rate function.

Proof. The upper bound (3.1) for compact set $F$ has been shown in Theorem 1.2 of Zabell [22] for more general case. Under the condition $0 \in \operatorname{interior}(D(\Lambda))$, the exponential tightness follows (see, e.g., the proof on pages 48-49 of [7]).

We prove the lower bound (3.2) by highlighting the new ingredients needed to modify [7]. For each $y \in \mathcal{F}$ and $\eta \in \operatorname{interior}(D(\Lambda))$ the exposing hyperplane for $y$, by the limsup inequality of $\Lambda_{n}$ to $\Lambda$, there exists $\eta_{n} \rightarrow \eta$ such that $\Lambda_{n}\left(\eta_{n}\right)<\infty$. We define a new probability measure

$$
\frac{d \tilde{\mu}_{n}}{d P X_{n}^{-1}}(z)=e^{n\left(\eta_{n} \cdot z-\Lambda_{n}\left(\eta_{n}\right)\right)}
$$

Using the liminf inequality of $\Lambda_{n}$ to $\Lambda$, then, as in [7],

$$
\lim _{\delta \rightarrow 0} \liminf _{n \rightarrow \infty} \frac{1}{n} \log P\left(X_{n} \in B(y, \delta)\right) \geq-\Lambda^{*}(y)+\lim _{\delta \rightarrow 0+} \liminf _{n \rightarrow \infty} \frac{1}{n} \log \tilde{\mu}_{n}(B(y, \delta))
$$

Now let $\tilde{\Lambda}(\cdot)=\Lambda(\cdot+\eta)-\Lambda(\eta)$. Then $\tilde{\Lambda}(0)=0$ and $0 \in \operatorname{interior}(D(\tilde{\Lambda}))$. Let

$$
\tilde{\Lambda}_{n}(p)=\frac{1}{n} \int_{R^{d}} e^{n p x} \tilde{\mu}_{n}(d x) .
$$

Then $\tilde{\Lambda}_{n} \xrightarrow{\Gamma} \tilde{\Lambda}$. The rest of the proof follows verbatim that in [7], concluding that

$$
\liminf _{n \rightarrow \infty} \frac{1}{n} \log \tilde{\mu}_{n}(B(y, \delta))=0
$$

Hence (3.2) follows.

## REFERENCES

[1] C. Albanese and S. Lawi, Laplace transforms for integrals of Markov processes, Markov Process. Related Fields, 11 (2005), pp. 677-724.
[2] L. Andersen and V. V. Piterbarg, Moment explosions in stochastic volatility models, Finance Stoch., 11 (2007), pp. 29-50.
[3] M. Avellaneda, D. Boyer-Olson, J. Busca, and P. Friz, Reconstructing the smile, Risk Magazine, October (2002).
[4] M. Avellaneda, D. Boyer-Olson, J. Busca, and P. Friz, Application of large deviation methods to the pricing of index options in finance, C. R. Math. Acad. Sci. Paris, 336 (2003), pp. 263-266.
[5] H. Berestycki, J. Busca, and I. Florent, Asymptotics and calibration of local volatility models, Quant. Finance, 2 (2002), pp. 61-69.
[6] H. Berestycki, J. Busca, and I. Florent, Computing the implied volatility in stochastic volatility models, Comm. Pure Appl. Math., 57 (2004), pp. 1352-1373.
[7] A. Dembo and O. Zeitouni, Large Deviation Techniques and Applications, 2nd ed., Springer, New York, 1998.
[8] D. Duffie, J. Pan, and K. Singleton, Transform analysis and asset pricing for affine jump diffusions, Econometrica, 68 (2000), pp. 1343-1376.
[9] M. Forde, Small-time asymptotics for implied volatility under the Heston model, Int. J. Theor. Appl. Finance, 12 (2009), pp. 861-876.
[10] J.-P. Fouque, G. Papanicolaou, and R. K. Sircar, Derivatives in Financial Markets with Stochastic Volatility, Cambridge University Press, Cambridge, UK, 2000.
[11] J.-P. Fouque, G. Papanicolaou, R. Sircar, and K. Solna, Multiscale stochastic volatility asymptotics, Multiscale Model. Simul., 2 (2003), pp. 22-42.
[12] J. Gatheral, The Volatility Surface: A Practitioner's Guide, Wiley, New York, 2006.
[13] P. S. Hagan, D. Kumar, A. S. Lesniewski, and D. E. Woodward, Managing smile risk, Willmott Magazine, September (2002), pp. 84-108.
[14] P. Henry-Labordère, A general asymptotic implied volatility for stochastic volatility models, in Proceedings "Petit déjeuner de la Finance," 2005; also available online from http://ssrn.com/abstract= 698601.
[15] P. Henry-Labordère, Combining the $S A B R$ and LVM models, Risk Magazine, October (2007).
[16] T. R. Hurd and A. Kuznetsov, Explicit formulas for Laplace transforms of stochastic integrals, Markov Process. Related Fields, 14 (2008), pp. 277-290.
[17] D. Lamberton and B. Lapeyre, Introduction to Stochastic Calculus Applied to Finance, Chapman \& Hall, London, 1996.
[18] A. Lewis, Option Valuation under Stochastic Volatility, Finance Press, Orange County, CA, 2000; also available online from http://www.optioncity.net/publications.htm.
[19] B. LeBaron, Stochastic volatility as a simple generator of apparent financial power laws and long memory, Quant. Finance, 1 (2001), pp. 621-631.
[20] J. Perello, J. Masoliver, and J.-Ph. Bouchaud, Multiple Time Scales in Volatility and Leverage Correlations: An Stochastic Volatility Model, http://ssrn.com/abstract=381780 (2003).
[21] D. Williams, Probability with Martingales, Cambridge University Press, Cambridge, UK, 1991.
[22] S. L. Zabell, Mosco convergence and large deviations, in Probability in Banach Spaces, 8 (Brunswick, ME, 1991), Progr. Probab. 30, R. M. Dudley, M. G. Hahn, and J. Kuelbs, eds., Birkhäuser Boston, Boston, MA, 1992, pp. 245-252.
[23] J. Zhu and M. Avellaneda, A risk-neutral stochastic volatility model, Int. J. Theor. Appl. Finance, 1 (1997), pp. 289-310.

# A Fourier Transform Method for Spread Option Pricing* 

T. R. Hurd ${ }^{\dagger}$ and Zhuowei Zhou $^{\dagger}$


#### Abstract

Spread options are a fundamental class of derivative contracts written on multiple assets and are widely traded in a range of financial markets. There is a long history of approximation methods for computing such products, but as yet there is no preferred approach that is accurate, efficient, and flexible enough to apply in general asset models. The present paper introduces a new formula for general spread option pricing based on Fourier analysis of the payoff function. Our detailed investigation, including a flexible and general error analysis, proves the effectiveness of a fast Fourier transform implementation of this formula for the computation of spread option prices. It is found to be easy to implement, stable, efficient, and applicable in a wide variety of asset pricing models.


Key words. spread options, multivariate spread options, jump diffusions, fast Fourier transform, gamma function

AMS subject classifications. 33B15, 65T50, 91B28
DOI. 10.1137/090750421

1. Introduction. When $S_{j t}, j=1,2, t \geq 0$, are two asset price processes, the basic spread option with maturity $T$ and strike $K \geq 0$ is the contract that pays $\left(S_{1 T}-S_{2 T}-K\right)^{+}$ at time $T$. If we assume the existence of a risk-neutral pricing measure, the risk-neutral expectation formula for the time 0 price of this option, assuming a constant interest rate $r$, is

$$
\begin{equation*}
\operatorname{Spr}\left(S_{0} ; T, K\right)=e^{-r T} \mathbb{E}_{S_{0}}\left[\left(S_{1 T}-S_{2 T}-K\right)^{+}\right] \tag{1.1}
\end{equation*}
$$

The literature on applications of spread options is extensive and is reviewed by Carmona and Durrleman [2], who explore further applications of spread options beyond the case of equities modeled by geometric Brownian motion (GBM), in particular to energy trading. For example, the difference between the price of crude oil and a refined fuel such as natural gas is called a "crack spread." "Spark spreads" refer to differences between the price of electricity and the price of fuel: options on spark spreads are widely used by power plant operators to optimize their revenue streams. Energy pricing requires models with mean reversion and jumps very different from GBM, and pricing spread options in such situations can be challenging.

Closed formulas for (1.1) are known only for a limited set of asset models. In the Bachelier stock model, $S_{t}=\left(S_{1 t}, S_{2 t}\right)$ is an arithmetic Brownian motion, and in this case (1.1) has a Black-Scholes-type formula for any $T, K$. In the special case $K=0$ when $S_{t}$ is a GBM, (1.1) is given by the Margrabe formula [14].

[^11]In the basic case where $S_{t}$ is a GBM and $K>0$, no explicit pricing formula is known. Instead, there is a long history of approximation methods for this problem. Numerical integration methods, typically Monte Carlo based, are often employed. When possible, however, the fastest option pricing engines by numerical integration are usually those based on the fast Fourier transform (FFT) methods introduced by Carr and Madan [4]. Their first interest was in single asset option pricing for geometric Lévy process models like the variance gamma (VG) model, but their basic framework has since been adapted to a variety of option payoffs and a host of asset return models where the characteristic function is known. In this work, when the payoff function is not square integrable, it is important to account for singularities in the Fourier transform variables.

Dempster and Hong [5] introduced a numerical integration method for spread options based on two-dimensional FFTs that was shown to be efficient when the asset price processes are GBMs or to have stochastic volatility. Three more recent papers study the use of multidimensional convolution FFT methods to price a wide range of multiasset options, including basket and spread options. These newer methods also compute by discretized Fourier transforms over truncated domains, but unlike earlier work using the FFT, they apparently do not rely on knowing the analytic Fourier transform of the payoff function or integrability of the payoff function. Lord et al. [11] provide error analysis that explains their observation that errors decay as a negative power of the size $N$ of the grid used in computing the FFT, provided the truncation is taken large enough. Leentvaar and Oosterlee [9] propose a parallel partitioning approach to tackle the so-called curse of dimensionality when the number of underlying assets becomes large. Jackson, Jaimungal, and Surkov [6] proposed a general FFT pricing framework for multiasset options, including variations with Bermudan early exercise features. These three papers all find that the FFT applied to the payoff function can perform well even if the payoff function is not square integrable and observe that errors can be made to decay as a negative power of $N$.

As an alternative to numerical integration methods, another stream uses analytical methods applicable to log-normal models that involve linear approximations of the nonlinear exercise boundary. Such methods are often very fast, but their accuracy is usually not easy to determine. Kirk [7] presented an analytical approximation that performs well in practice. Carmona and Durrleman [3] and later Li, Deng, and Zhou [10] demonstrate a number of lower and upper bounds for the spread option price that combine to produce accurate analytical approximation formulas in log-normal asset models. These results extend to approximate values for the Greeks.

The main purpose of the present paper is to give a numerical integration method for computing spread options in two or higher dimensions using the FFT. Unlike the above multiasset FFT methods, it is based on square integrable integral formulas for the payoff function, and like those methods it is applicable to a variety of spread option payoffs in any model for which the characteristic function of the joint return process is given analytically. Since our method involves only smooth square integrable integrands, the error estimates we present are quite straightforward and standard. In fact, we demonstrate that the asymptotic decay of errors is exponential, rather than polynomial, in the size $N$ of the Fourier grid. For option payoffs that can be made square integrable, our method has the flexibility to handle a wide range of desirable asset return models, all with a very competitive computational expense.

The results we describe stem from the following new formula, ${ }^{1}$ which gives a square integrable Fourier representation of the basic spread option payoff function $P\left(x_{1}, x_{2}\right)=$ $\left(e^{x_{1}}-e^{x_{2}}-1\right)^{+}$.

Theorem 1.1. For any real numbers $\epsilon=\left(\epsilon_{1}, \epsilon_{2}\right)$ with $\epsilon_{2}>0$ and $\epsilon_{1}+\epsilon_{2}<-1$ and $x=$ $\left(x_{1}, x_{2}\right),{ }^{2}$

$$
\begin{equation*}
P(x)=(2 \pi)^{-2} \iint_{\mathbb{R}^{2}+i \epsilon} e^{i u x^{\prime}} \hat{P}(u) d^{2} u, \quad \hat{P}(u)=\frac{\Gamma\left(i\left(u_{1}+u_{2}\right)-1\right) \Gamma\left(-i u_{2}\right)}{\Gamma\left(i u_{1}+1\right)} . \tag{1.2}
\end{equation*}
$$

Here $\Gamma(z)$ is the complex gamma function defined for $\Re e(z)>0$ by the integral $\Gamma(z)=$ $\int_{0}^{\infty} e^{-t} t^{z-1} d t$.

Using this theorem, whose proof is given in the appendix, we will find that we can follow the logic of Carr and Madan to derive numerical algorithms for efficient computation of a variety of spread options and their Greeks. The basic strategy to compute (1.1) is to combine (1.2) with an explicit formula for the characteristic function of the bivariate random variable $X_{t}=\left(\log S_{1 t}, \log S_{2 t}\right)$. For the remainder of this paper, we make a simplifying assumption.

Assumption 1. For any $t>0$, the increment $X_{t}-X_{0}$ is independent of $X_{0}$.
This implies that the characteristic function of $X_{T}$ factorizes

$$
\begin{equation*}
\mathbb{E}_{X_{0}}\left[e^{i u X_{T}^{\prime}}\right]=e^{i u X_{0}^{\prime}} \Phi(u ; T), \quad \Phi(u ; T):=\mathbb{E}_{X_{0}}\left[e^{i u\left(X_{T}-X_{0}\right)^{\prime}}\right], \tag{1.3}
\end{equation*}
$$

where $\Phi(u ; T)$ is independent of $X_{0}$. Although the above assumption rules out mean-reverting processes that often arise in energy applications, it holds for typical stock models: moreover, the method we propose can be generalized to a variety of mean-reverting processes. Using Theorem 1.1 and (1.3), the spread option formula can be written as an explicit two-dimensional Fourier transform in the variable $X_{0}$ :

$$
\begin{align*}
\operatorname{Spr}\left(X_{0} ; T\right) & =e^{-r T} \mathbb{E}_{X_{0}}\left[\left(e^{X_{1 T}}-e^{X_{2 T}}-1\right)^{+}\right] \\
& =e^{-r T} \mathbb{E}_{X_{0}}\left[(2 \pi)^{-2} \iint_{\mathbb{R}^{2}+i \epsilon} e^{i u X_{T}^{\prime}} \hat{P}(u) d^{2} u\right] \\
& =(2 \pi)^{-2} e^{-r T} \iint_{\mathbb{R}^{2}+i \epsilon} \mathbb{E}_{X_{0}}\left[e^{i u X_{T}^{\prime}}\right] \hat{P}(u) d^{2} u \\
& =(2 \pi)^{-2} e^{-r T} \iint_{\mathbb{R}^{2}+i \epsilon} e^{i u X_{0}^{\prime}} \Phi(u ; T) \hat{P}(u) d^{2} u . \tag{1.4}
\end{align*}
$$

The Greeks are handled in exactly the same way. For example, the Delta $\Delta^{1}:=\partial \operatorname{Spr} / \partial S_{10}$ is obtained as a function of $S_{0}$ by replacing $\Phi$ in (1.4) by $\partial \Phi / \partial S_{10}$.

Double Fourier integrals like this can be approximated numerically by a two-dimensional FFT. Such approximations involve both a truncation and discretization of the integral, and the two properties that determine their accuracy are the decay of the integrand of (1.4) in $u$-space

[^12]and the decay of the function Spr in $x$-space. The remaining issue of computing the gamma function is not really difficult. Fast and accurate computation of the complex gamma function in, for example, MATLAB, is based on the Lanczos approximation popularized by [15]. ${ }^{3}$

In this paper, we demonstrate how our method performs for computing spread options in three different two-asset stock models, namely GBM, a three factor stochastic volatility (SV) model, and the VG model. Section 2 provides the essential definitions of the three types of asset return models, including explicit formulas for their bivariate characteristic functions. Section 3 discusses how the two-dimensional FFT can be implemented for our problem. Section 4 provides error analysis that shows how the accuracy and speed will depend on the implementation choices made. Section 5 describes briefly how the method extends to the computation of spread option Greeks. Section 6 gives the detailed results of the performance of the method in the three asset return models. In this section, the accuracy of each model is compared to benchmark values computed by an independent method for a reference set of option prices. We also demonstrate that the computation of the spread option Greeks in such models is equally feasible. Section 7 extends all the above results to several kinds of basket options on two or more assets. Although the formulation is simple, the resulting FFTs become, in practice, much slower to compute in higher dimensions, due to the so-called curse of dimensionality: in such cases, one can implement the parallel partitioning approach of [9].

## 2. Three kinds of stock models.

2.1. The case of GBM. In the two-asset Black-Scholes model, the vector $S_{t}=\left(S_{1 t}, S_{2 t}\right)$ has components

$$
S_{j t}=S_{j 0} \exp \left[\left(r-\sigma_{j}^{2} / 2\right) t+\sigma_{j} W_{t}^{j}\right], \quad j=1,2,
$$

where $\sigma_{1}, \sigma_{2}>0$ and $W^{1}, W^{2}$ are risk-neutral Brownian motions with constant correlation $\rho,|\rho|<1$. The joint characteristic function of $X_{T}=\left(\log S_{1 T}, \log S_{2 T}\right)$ as a function of $u=$ $\left(u_{1}, u_{2}\right)$ is of the form $e^{i u X_{0}^{\prime}} \Phi(u ; T)$ with

$$
\begin{equation*}
\Phi(u ; T)=\exp \left[i u\left(r T e-\sigma^{2} T / 2\right)^{\prime}-u \Sigma u^{\prime} T / 2\right], \tag{2.1}
\end{equation*}
$$

where $e=(1,1), \Sigma=\left[\sigma_{1}^{2}, \sigma_{1} \sigma_{2} \rho ; \sigma_{1} \sigma_{2} \rho, \sigma_{2}^{2}\right]$, and $\sigma^{2}=\operatorname{diag} \Sigma$. We remind the reader that we use implied matrix multiplication and that $u^{\prime}$ denotes the (unconjugated) matrix transpose. Substituting this expression into (1.4) yields the spread option formula

$$
\begin{equation*}
\operatorname{Spr}\left(X_{0} ; T\right)=(2 \pi)^{-2} e^{-r T} \iint_{\mathbb{R}^{2}+i \epsilon} e^{i u X_{0}^{\prime}} \exp \left[i u\left(r T e-\sigma^{2} T / 2\right)^{\prime}-u \Sigma u^{\prime} T / 2\right] \hat{P}(u) d^{2} u . \tag{2.2}
\end{equation*}
$$

As we discuss in section 3, we recommend that this be computed numerically using the FFT.
2.2. Three factor SV model. The spread option problem in a three factor stochastic volatility model was given as an example by Dempster and Hong [5]. Their asset model is

[^13]defined by SDEs for $X_{t}=\left(\log S_{1 t}, \log S_{2 t}\right)$ and the squared volatility $v_{t}$ :
\[

$$
\begin{aligned}
d X_{1} & =\left[\left(r-\delta_{1}-\sigma_{1}^{2} / 2\right) d t+\sigma_{1} \sqrt{v} d W^{1}\right], \\
d X_{2} & =\left[\left(r-\delta_{2}-\sigma_{2}^{2} / 2\right) d t+\sigma_{2} \sqrt{v} d W^{2}\right], \\
d v & =\kappa(\mu-v) d t+\sigma_{v} \sqrt{v} d W^{v},
\end{aligned}
$$
\]

where the three Brownian motions have correlations:

$$
\begin{aligned}
& E\left[d W^{1} d W^{2}\right]=\rho d t, \\
& E\left[d W^{1} d W^{v}\right]=\rho_{1} d t \\
& E\left[d W^{2} d W^{v}\right]=\rho_{2} d t
\end{aligned}
$$

As discussed in that paper, the asset return vector has the joint characteristic function $e^{i u X_{0}^{\prime}} \Phi\left(u ; T, v_{0}\right)$, where

$$
\begin{aligned}
\Phi\left(u ; T, v_{0}\right)= & {\left[\left(\frac{2 \zeta\left(1-e^{-\theta T}\right)}{2 \theta-(\theta-\gamma)\left(1-e^{-\theta T}\right)}\right) v_{0}\right.} \\
& \left.+i u(r e-\delta)^{\prime} T-\frac{\kappa \mu}{\sigma_{v}^{2}}\left[2 \log \left(\frac{2 \theta-(\theta-\gamma)\left(1-e^{-\theta T}\right)}{2 \theta}\right)+(\theta-\gamma) T\right]\right]
\end{aligned}
$$

and

$$
\begin{aligned}
\zeta & :=-\frac{1}{2}\left[\left(\sigma_{1}^{2} u_{1}^{2}+\sigma_{2}^{2} u_{2}^{2}+2 \rho \sigma_{1} \sigma_{2} u_{1} u_{2}\right)+i\left(\sigma_{1}^{2} u_{1}+\sigma_{2}^{2} u_{2}\right)\right], \\
\gamma & :=\kappa-i\left(\rho_{1} \sigma_{1} u_{1}+\rho_{2} \sigma_{2} u_{2}\right) \sigma_{\nu}, \\
\theta & :=\sqrt{\gamma^{2}-2 \sigma_{v}^{2} \zeta} .
\end{aligned}
$$

2.3. Exponential Lévy models. Many stock price models are of the form $S_{t}=e^{X_{t}}$, where $X_{t}$ is a Lévy process for which the characteristic function is explicitly known. We illustrate with the example of the VG process introduced by [13] the three parameter process $Y_{t}$ with Lévy characteristic triple $(0,0, \nu)$, where the Lévy measure is $\nu(x)=\lambda\left[e^{-a_{+} x} \mathbf{1}_{x>0}+\right.$ $\left.e^{a_{-} x} \mathbf{1}_{x<0}\right] /|x|$ for positive constants $\lambda, a_{ \pm}$. The characteristic function of $Y_{t}$ is

$$
\begin{equation*}
\Phi_{Y_{t}}(u)=\left[1+i\left(\frac{1}{a_{-}}-\frac{1}{a_{+}}\right) u+\frac{u^{2}}{a_{-} a_{+}}\right]^{-\lambda t} . \tag{2.3}
\end{equation*}
$$

To demonstrate the effects of correlation, we take a bivariate VG model driven by three independent VG processes $Y_{1}, Y_{2}, Y$ with common parameters $a_{ \pm}$and $\lambda_{1}=\lambda_{2}=(1-\alpha) \lambda$, $\lambda^{Y}=\alpha \lambda$. The bivariate $\log$ return process $X_{t}=\log S_{t}$ is a mixture:

$$
\begin{equation*}
X_{1 t}=X_{10}+Y_{1 t}+Y_{t}, \quad X_{2 t}=X_{20}+Y_{2 t}+Y_{t} . \tag{2.4}
\end{equation*}
$$

Here $\alpha \in[0,1]$ leads to dependence between the two return processes but leaves their marginal laws unchanged. An easy calculation leads to the bivariate characteristic function $e^{i u X_{0}^{\prime}} \Phi(u ; T)$
with

$$
\begin{align*}
& \Phi(u ; T)=\left[1+i\left(\frac{1}{a_{-}}-\frac{1}{a_{+}}\right)\left(u_{1}+u_{2}\right)+\frac{\left(u_{1}+u_{2}\right)^{2}}{a_{-} a_{+}}\right]^{-\alpha \lambda t}  \tag{2.5}\\
& \times\left[1+i\left(\frac{1}{a_{-}}-\frac{1}{a_{+}}\right) u_{1}+\frac{u_{1}^{2}}{a_{-} a_{+}}\right]^{-(1-\alpha) \lambda t}\left[1+i\left(\frac{1}{a_{-}}-\frac{1}{a_{+}}\right) u_{2}+\frac{u_{2}^{2}}{a_{-} a_{+}}\right]^{-(1-\alpha) \lambda t}
\end{align*}
$$

3. Numerical integration by FFT. To compute (1.4) in these models we approximate the double integral by a double sum over the lattice

$$
\Gamma=\left\{u(k)=\left(u_{1}\left(k_{1}\right), u_{2}\left(k_{2}\right)\right) \mid k=\left(k_{1}, k_{2}\right) \in\{0, \ldots, N-1\}^{2}\right\}, \quad u_{i}\left(k_{i}\right)=-\bar{u}+k_{i} \eta
$$

for appropriate choices of $N, \eta, \bar{u}:=N \eta / 2$. For the FFT it is convenient to take $N$ to be a power of 2 and lattice spacing $\eta$ such that truncation of the $u$-integrals to $[-\bar{u}, \bar{u}]$ and discretization leads to an acceptable error. Finally, we choose initial values $X_{0}=\log S_{0}$ to lie on the reciprocal lattice with spacing $\eta^{*}=2 \pi / N \eta=\pi / \bar{u}$ and width $2 \bar{x}, \bar{x}=N \eta^{*} / 2$ :

$$
\Gamma^{*}=\left\{x(\ell)=\left(x_{1}\left(\ell_{1}\right), x_{2}\left(\ell_{2}\right)\right) \mid \ell=\left(\ell_{1}, \ell_{2}\right) \in\{0, \ldots, N-1\}^{2}\right\}, x_{i}\left(\ell_{i}\right)=-\bar{x}+\ell_{i} \eta^{*}
$$

For any $S_{0}=e^{X_{0}}$ with $X_{0}=x(\ell) \in \Gamma^{*}$ we then have the approximation

$$
\begin{equation*}
\operatorname{Spr}\left(X_{0} ; T\right) \sim \frac{\eta^{2} e^{-r T}}{(2 \pi)^{2}} \sum_{k_{1}, k_{2}=0}^{N-1} e^{i(u(k)+i \epsilon) x(\ell)^{\prime}} \Phi(u(k)+i \epsilon ; T) \hat{P}(u(k)+i \epsilon) \tag{3.1}
\end{equation*}
$$

Now, as usual for the discrete FFT, as long as $N$ is even,

$$
i u(k) x(\ell)^{\prime}=i \pi\left(k_{1}+k_{2}+\ell_{1}+\ell_{2}\right)+2 \pi i k \ell^{\prime} / N \quad(\bmod 2 \pi i)
$$

This leads to the double inverse discrete Fourier transform (i.e., the MATLAB function ifft2)

$$
\begin{align*}
\operatorname{Spr}\left(X_{0} ; T\right) & \sim(-1)^{\ell_{1}+\ell_{2}} e^{-r T}\left(\frac{\eta N}{2 \pi}\right)^{2} e^{-\epsilon x(\ell)^{\prime}}\left[\frac{1}{N^{2}} \sum_{k_{1}, k_{2}=0}^{N-1} e^{2 \pi i k \ell^{\prime} / N} H(k)\right] \\
& =(-1)^{\ell_{1}+\ell_{2}} e^{-r T}\left(\frac{\eta N}{2 \pi}\right)^{2} e^{-\epsilon x(\ell)^{\prime}}[\operatorname{ifft} 2(H)](\ell) \tag{3.2}
\end{align*}
$$

where

$$
H(k)=(-1)^{k_{1}+k_{2}} \Phi(u(k)+i \epsilon ; T) \hat{P}(u(k)+i \epsilon)
$$

4. Error discussion. The selection of suitable values for $\epsilon, N$, and $\eta$ when implementing the above FFT approximation of (1.4) is a somewhat subtle issue whose details depend on the asset model in question. We now give a general discussion of the pure truncation error and pure discretization error in (3.1): a more complete analysis of the combined errors using methods described in [8] will lead to the same broad conclusions.

The pure truncation error, defined by taking $\eta \rightarrow 0, N \rightarrow \infty$ while keeping $\bar{u}=N \eta / 2$ fixed, can be made smaller than $\delta_{1} \ll 1$ if the integrand of (1.4) is small and decaying outside
the square $\left[-\bar{u}+i \epsilon_{1}, \bar{u}+i \epsilon_{1}\right] \times\left[-\bar{u}+i \epsilon_{2}, \bar{u}+i \epsilon_{2}\right]$. Corollary A.1, proved in the appendix, gives a uniform $O\left(|u|^{-2}\right)$ upper bound on $\hat{P}$, while $\Phi(u)$ can generally be seen directly to have some $u$-decay. Thus the truncation error will be less than $\delta_{1}$ if one picks $\bar{u}$ large enough so that $|\Phi|<O\left(\delta_{1}\right)$ and has decay outside the square.

The pure discretization error, defined by taking $\bar{u} \rightarrow \infty, N \rightarrow \infty$ while keeping $\bar{x}=\pi / \eta$ fixed, can be made smaller than $\delta_{2} \ll 1$ if $e^{\epsilon X_{0}^{\prime}} \operatorname{Spr}\left(X_{0}\right)$, taken as a function of $X_{0} \in \mathbb{R}^{2}$, has rapid decay in $X_{0}$. This is related to the smoothness of $\Phi(u)$ and the choice of $\epsilon$. The first two models are not very sensitive to $\epsilon$, but in the VG model the following conditions are needed to ensure that singularities in $u$-space are avoided:

$$
-a_{+}<\epsilon_{1}, \epsilon_{2}, \epsilon_{1}+\epsilon_{2}<a_{-}
$$

By applying the Poisson summation formula to $e^{\epsilon X_{0}^{\prime}} \operatorname{Spr}\left(X_{0}\right)$, one can write the discretization error as

$$
\begin{equation*}
\operatorname{Spr}^{(\bar{x})}\left(X_{0}\right)-\operatorname{Spr}\left(X_{0}\right)=\sum_{\ell \in \mathbb{Z}^{2} \backslash\{(0,0)\}} e^{2 \bar{x} \epsilon \ell^{\prime}} \operatorname{Spr}\left(X_{0}+2 \bar{x} \ell\right) . \tag{4.1}
\end{equation*}
$$

One can verify using brute force bounds that the terms on the right-hand side of (4.1) are all small and decay in all lattice directions, provided $\bar{x}$ is sufficiently large. Thus the discretization error will be less than $\delta_{2}$ for all $X_{0} \in[-c \bar{x}, c \bar{x}]^{2}$ with $0<c \ll 1$ if one picks $\bar{x}$ large enough so that $\left|e^{\epsilon X_{0}^{\prime}} \operatorname{Spr}\left(X_{0}\right)\right|<O\left(\delta_{2}\right)$ and has decay outside the square $[-\bar{x}, \bar{x}]^{2}$.

In summary, one expects that the combined truncation and discretization error will be close to $\delta_{1}+\delta_{2}$ if $\bar{u}=N \eta / 2$ and $\eta=\pi / \bar{x}$ are each chosen as above. We shall see in section 6 that the observed errors are consistent with the above analysis that predicts an asymptotic exponential decay with the size $N$ of the Fourier lattice for the models we address.
5. Greeks. The FFT method can also be applied to the Greeks, enabling us to tackle hedging and other interesting problems. It is particularly efficient for the GBM model, where differentiation under the integral sign is always permissible. For instance, the FFT formula for vega (the sensitivity to $\sigma$ ) takes the form

$$
\begin{aligned}
\frac{\partial \operatorname{Spr}\left(S_{0} ; T\right)}{\partial \sigma_{1}} & =(-1)^{\ell_{1}+\ell_{2}} e^{-r T}\left(\frac{\eta N}{2 \pi}\right)^{2} e^{-\epsilon x(\ell)^{\prime}}\left[\operatorname{ifft2}\left(\frac{\partial H}{\partial \sigma_{1}}\right)\right](\ell), \\
\frac{\partial H(k)}{\partial \sigma_{1}} & =\left[-(u(k)+i \epsilon)\left(i \frac{\partial{\sigma^{2}}^{\prime}}{\partial \sigma_{1}}+\frac{\partial \Sigma}{\partial \sigma_{1}}(u(k)+i \epsilon)^{\prime}\right) \frac{T}{2}\right] H(k),
\end{aligned}
$$

where $\frac{\partial \sigma^{2}}{\partial \sigma_{1}}=\left[2 \sigma_{1}, 0\right]$ and $\frac{\partial \Sigma}{\partial \sigma_{1}}=\left[2 \sigma_{1}, \rho \sigma_{2} ; \rho \sigma_{2}, 0\right]$. Other Greeks including those of higher orders can be computed in a similar fashion. This method needs to be used with care for the SV and VG models, since it is possible that differentiation leads to an integrand that decays slowly.
6. Numerical results. Our numerical experiments were coded and implemented in MATLAB version 7.6.0 on an Intel 2.80 GHz machine running under Linux with 1 GB physical memory. If they were coded in $\mathrm{C}++$ with similar algorithms, we should expect to see faster performance.


Figure 1. This graph shows the objective function Err for the numerical computation of the GBM spread option versus the benchmark. Errors are plotted against the grid size for different choices of $\bar{u}$. The parameter values are those of the GBM model used by [5]: $r=0.1, T=1.0, \rho=0.5, \delta_{1}=0.05, \sigma_{1}=0.2, \delta_{2}=0.05$, $\sigma_{2}=0.1$.

The strength of the FFT method is demonstrated by comparison with accurate benchmark prices computed by an independent (usually extremely slow) method. Based on a representative selection of initial $\log$-asset value pairs $\log S_{10}^{i}=\frac{i \pi}{10}, \log S_{20}^{j}=-\frac{\pi}{5}+\frac{j \pi}{10}, i, j \in 1,2,3, \ldots, 6$, the objective function we measure is defined as

$$
\begin{equation*}
\operatorname{Err}=\frac{1}{36} \sum_{i, j=1}^{6}\left|\log \left(M^{i j}\right)-\log \left(B^{i j}\right)\right|, \tag{6.1}
\end{equation*}
$$

where $M^{i j}$ and $B^{i j}$ are the corresponding FFT computed prices and benchmark prices. These choices cover a wide range of moneyness, from deep out-of-the-money to deep in-the-money. Since these combinations all lie on lattices $\Gamma^{*}$ corresponding to $N=2^{n}$ and $\bar{u} / 10=2^{m}$ for integers $n, m$, all 36 prices $M^{i j}$ can be computed simultaneously with a single FFT.

Figure 1 shows how the FFT method performs in the two-dimensional GBM model for different choices of $N$ and $\bar{u}$. Since the two factors are bivariate normal, benchmark prices can be calculated to high accuracy by one-dimensional integrations. In Figure 1 we can clearly see the effects of both truncation errors and discretization errors. For a fixed $\bar{u}$, the objective function decreases when $N$ increases. The $\bar{u}=20$ curve flattens out near $10^{-5}$ due to its truncation error of that magnitude. In turn, we can quantify its discretization errors with respect to $N$ by subtracting the truncation error from the total error. The flattening of the curves with $\bar{u}=40,80$, and 160 near $10^{-14}$ should be attributed to MATLAB round-off errors: because of the rapid decrease of the characteristic function $\Phi$, their truncation error is negligible. For a fixed $N$, increasing $\bar{u}$ brings two effects: reducing truncation error and enlarging discretization error. These effects are well demonstrated in Figure 1.


Figure 2. This graph shows the objective function Err for the numerical computation of the SV spread option versus the benchmark computed using the FFT method itself with parameters $N=2^{12}$ and $\bar{u}=80$. The parameter values are those of the $S V$ model used by [5]: $r=0.1, T=1.0, \rho=0.5, \delta_{1}=0.05, \sigma_{1}=1.0$, $\rho_{1}=-0.5, \delta_{2}=0.05, \sigma_{2}=0.5, \rho_{2}=0.25, v_{0}=0.04, \kappa=1.0, \mu=0.04, \sigma_{v}=0.05$.

For the SV model, no analytical or numerical method we know is consistently accurate enough to serve as an independent benchmark. Instead, we computed benchmark prices using the FFT method itself with $N=2^{12}$ and $\bar{u}=80$. The resulting objective function, shown in Figure 2, exhibits behavior similar to Figure 1 and is consistent with accuracies at the level of roundoff. We also verified that the benchmark prices are consistent to a level of $4 \times 10^{-4}$ with those resulting from an intensive Monte Carlo computation using 1,000,000 simulations, each consisting of 2000 time steps. The computational cost to further reduce the Monte Carlo simulation error becomes prohibitive.

Because the VG process has an explicit probability density function in terms of a Bessel function [12], rather accurate benchmark spread option values for the VG model can be computed by a three-dimensional integration. ${ }^{4}$ We used a Gaussian quadrature algorithm set with a high tolerance of $10^{-9}$ to compute the integrals for these benchmarks. The resulting objective function for various values of $\bar{u}, N$ is shown in Figure 3. The truncation error for $\bar{u}=20$ is about $2 \times 10^{-5}$. The other three curves flatten out near $5 \times 10^{-8}$, a level we identify as the accuracy of the benchmark. A comparable graph (not shown), using benchmark prices computed with the FFT method with $N=2^{12}$ and $\bar{u}=80$, showed behavior similar to Figures 1 and 2 and is consistent with the FFT method being capable of producing accuracies at the level of roundoff.

The strength of the FFT method is further illustrated by the computation of individual prices and relative errors shown in Tables 1, 2, and 3. One can observe that an FFT with $N=256$ is capable of producing very high accuracy in all three models. It is interesting to note that FFT prices in almost all cases were biased low compared to the benchmark. Exceptions

[^14]

Figure 3. This graph shows the objective function Err for the numerical computation of the VG spread option versus the benchmark values computed with a three-dimensional integration. Errors are plotted against the grid size for five different choices of $\bar{u}$. The parameters are $r=0.1, T=1.0, \rho=0.5, a_{+}=20.4499$, $a_{-}=24.4499, \alpha=0.4, \lambda=10$.

## Table 1

Benchmark prices for the two-factor GBM model of [5] and relative errors for the FFT method with different choices of $N$. The parameter values are the same as Figure 1 except we fix $S_{10}=100, S_{20}=96$, $\bar{u}=40$. The interpolation is based on a matrix of prices with discretization of $N=256$ and a polynomial with a degree of 8 .

| Strike $K$ | Benchmark | 64 | 128 | 256 | 512 | Interpolation |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0.4 | 8.312461 | -3.8 | $-4.5 \mathrm{E}-4$ | $-1.9 \mathrm{E}-8$ | $-1.7 \mathrm{E}-14$ | $1.9 \mathrm{E}-8$ |
| 0.8 | 8.114994 | $-3.8 \mathrm{E}-1$ | $-4.6 \mathrm{E}-4$ | $-2.0 \mathrm{E}-8$ | $-7 \mathrm{E}-15$ | $2.0 \mathrm{E}-8$ |
| 1.2 | 7.920820 | $-7.3 \mathrm{E}-2$ | $-4.6 \mathrm{E}-4$ | $-2.0 \mathrm{E}-8$ | $-2.8 \mathrm{E}-14$ | $2.0 \mathrm{E}-8$ |
| 1.6 | 7.729932 | $-7.2 \mathrm{E}-2$ | $-4.7 \mathrm{E}-4$ | $-2.0 \mathrm{E}-8$ | $-4.8 \mathrm{E}-14$ | $2.0 \mathrm{E}-8$ |
| 2.0 | 7.542324 | $-7.3 \mathrm{E}-2$ | $-4.8 \mathrm{E}-4$ | $-2.1 \mathrm{E}-8$ | $-4.9 \mathrm{E}-14$ | $2.1 \mathrm{E}-8$ |
| 2.4 | 7.357984 | $-7.5 \mathrm{E}-2$ | $-4.9 \mathrm{E}-4$ | $-2.1 \mathrm{E}-8$ | $-7.3 \mathrm{E}-14$ | $2.1 \mathrm{E}-8$ |
| 2.8 | 7.176902 | $-7.6 \mathrm{E}-2$ | $-5.0 \mathrm{E}-4$ | $-2.2 \mathrm{E}-8$ | $-6.8 \mathrm{E}-14$ | $2.2 \mathrm{E}-8$ |
| 3.2 | 6.999065 | $-7.8 \mathrm{E}-2$ | $-5.1 \mathrm{E}-4$ | $-2.2 \mathrm{E}-8$ | $-9.7 \mathrm{E}-14$ | $2.2 \mathrm{E}-8$ |
| 3.6 | 6.824458 | $-8.0 \mathrm{E}-2$ | $-5.3 \mathrm{E}-4$ | $-2.3 \mathrm{E}-8$ | $-8.2 \mathrm{E}-14$ | $2.3 \mathrm{E}-8$ |
| 4.0 | 6.653065 | $-8.1 \mathrm{E}-2$ | $-5.4 \mathrm{E}-4$ | $-2.3 \mathrm{E}-8$ | $-9.0 \mathrm{E}-14$ | $2.3 \mathrm{E}-8$ |

to this observation seem only to appear at a level of the accuracy of the benchmark itself.
The FFT computes in a single iteration an $N \times N$ panel of prices spread corresponding to initial values $S_{10}=e^{x_{10}+\ell_{1} \eta^{*}}, S_{20}=e^{x_{20}+\ell_{2} \eta^{*}}, K=1,\left(\ell_{1}, \ell_{2}\right) \in\{0, \ldots, N-1\}^{2}$. If the desired selection of $\left\{S_{10}, S_{20}, K\right\}$ fits into this panel of prices, or its scaling, a single FFT suffices. If not, then one has to match $\left(x_{10}, x_{20}\right)$ with each combination, and run several FFTs, with a consequent increase in computation time. However, we have found that an interpolation technique is very accurate for practical purposes. For instance, prices for multiple strikes with the same $S_{10}$ and $S_{20}$ are approximated by a polynomial fit along the diagonal

## Table 2

Benchmark prices for the three factor SV model of [5] and relative errors for the FFT method with different choices of $N$. The parameter values are the same as Figure 2 except we fix $S_{10}=100, S_{20}=96, \bar{u}=40$. The interpolation is based on a matrix of prices with discretization of $N=256$ and a polynomial with a degree of 8 .

| Strike $K$ | Benchmark | 64 | 128 | 256 | 512 | Interpolation |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 2.0 | 7.548502 | $-7.3 \mathrm{E}-2$ | $-4.8 \mathrm{E}-4$ | $-2.1 \mathrm{E}-8$ | $1.6 \mathrm{E}-11$ | $-2.1 \mathrm{E}-8$ |
| 2.2 | 7.453536 | $-7.4 \mathrm{E}-2$ | $-4.9 \mathrm{E}-4$ | $-2.1 \mathrm{E}-8$ | $1.2 \mathrm{E}-11$ | $-2.1 \mathrm{E}-8$ |
| 2.4 | 7.359381 | $-7.5 \mathrm{E}-2$ | $-4.8 \mathrm{E}-4$ | $-2.1 \mathrm{E}-8$ | $8.6 \mathrm{E}-12$ | $-2.1 \mathrm{E}-8$ |
| 2.6 | 7.266037 | $-7.5 \mathrm{E}-2$ | $-5.0 \mathrm{E}-4$ | $-2.1 \mathrm{E}-8$ | $4.6 \mathrm{E}-12$ | $-2.1 \mathrm{E}-8$ |
| 2.8 | 7.173501 | $-7.6 \mathrm{E}-2$ | $-5.0 \mathrm{E}-4$ | $-2.2 \mathrm{E}-8$ | $6.1 \mathrm{E}-13$ | $-2.2 \mathrm{E}-8$ |
| 3.0 | 7.081775 | $-7.7 \mathrm{E}-2$ | $-5.1 \mathrm{E}-4$ | $-2.2 \mathrm{E}-8$ | $-3.5 \mathrm{E}-12$ | $-2.2 \mathrm{E}-8$ |
| 3.2 | 6.990857 | $-7.8 \mathrm{E}-2$ | $-5.2 \mathrm{E}-4$ | $-2.2 \mathrm{E}-8$ | $-7.7 \mathrm{E}-12$ | $-2.2 \mathrm{E}-8$ |
| 3.4 | 6.900745 | $-7.9 \mathrm{E}-2$ | $-5.2 \mathrm{E}-4$ | $-2.2 \mathrm{E}-8$ | $-1.2 \mathrm{E}-11$ | $-2.2 \mathrm{E}-8$ |
| 3.6 | 6.811440 | $-8.0 \mathrm{E}-2$ | $-5.3 \mathrm{E}-4$ | $-2.3 \mathrm{E}-8$ | $-1.7 \mathrm{E}-11$ | $-2.3 \mathrm{E}-8$ |
| 3.8 | 6.722939 | $-8.1 \mathrm{E}-2$ | $-5.3 \mathrm{E}-4$ | $-2.3 \mathrm{E}-8$ | $-2.0 \mathrm{E}-11$ | $-2.3 \mathrm{E}-8$ |
| 4.0 | 6.635242 | $-8.1 \mathrm{E}-2$ | $-5.4 \mathrm{E}-4$ | $-2.3 \mathrm{E}-8$ | $-2.4 \mathrm{E}-11$ | $-2.3 \mathrm{E}-8$ |

Table 3
Benchmark prices for the VG model and relative errors for the FFT method with different choices of $N$. The parameter values are the same as Figure 3 except we fix $S_{10}=100, S_{20}=96, \bar{u}=40$. The interpolation is based on a matrix of prices with discretization of $N=256$ and a polynomial with a degree of 8 .

| Strike $K$ | Benchmark | 64 | 128 | 256 | 512 | Interpolation |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 2.0 | 9.727458 | $-5.9 \mathrm{E}-2$ | $-3.9 \mathrm{E}-4$ | $1.5 \mathrm{E}-8$ | $3.2 \mathrm{E}-8$ | $1.5 \mathrm{E}-8$ |
| 2.2 | 9.630005 | $-5.9 \mathrm{E}-2$ | $-3.9 \mathrm{E}-4$ | $1.7 \mathrm{E}-8$ | $3.4 \mathrm{E}-8$ | $1.7 \mathrm{E}-8$ |
| 2.4 | 9.533199 | $-6.0 \mathrm{E}-2$ | $-3.9 \mathrm{E}-4$ | $1.8 \mathrm{E}-8$ | $3.5 \mathrm{E}-8$ | $1.8 \mathrm{E}-8$ |
| 2.6 | 9.437040 | $-6.0 \mathrm{E}-2$ | $-4.0 \mathrm{E}-4$ | $2.0 \mathrm{E}-8$ | $3.7 \mathrm{E}-8$ | $2.0 \mathrm{E}-8$ |
| 2.8 | 9.341527 | $-6.0 \mathrm{E}-2$ | $-4.0 \mathrm{E}-4$ | $2.5 \mathrm{E}-8$ | $4.3 \mathrm{E}-8$ | $2.5 \mathrm{E}-8$ |
| 3.0 | 9.246662 | $-6.1 \mathrm{E}-2$ | $-4.0 \mathrm{E}-4$ | $2.5 \mathrm{E}-8$ | $4.3 \mathrm{E}-8$ | $2.5 \mathrm{E}-8$ |
| 3.2 | 9.152445 | $-6.1 \mathrm{E}-2$ | $-4.1 \mathrm{E}-4$ | $2.3 \mathrm{E}-8$ | $4.1 \mathrm{E}-8$ | $2.3 \mathrm{E}-8$ |
| 3.4 | 9.058875 | $-6.2 \mathrm{E}-2$ | $-4.1 \mathrm{E}-4$ | $3.0 \mathrm{E}-8$ | $4.8 \mathrm{E}-8$ | $3.0 \mathrm{E}-8$ |
| 3.6 | 8.965954 | $-6.2 \mathrm{E}-2$ | $-4.1 \mathrm{E}-4$ | $3.0 \mathrm{E}-8$ | $4.8 \mathrm{E}-8$ | $3.0 \mathrm{E}-8$ |
| 3.8 | 8.873681 | $-6.3 \mathrm{E}-2$ | $-4.2 \mathrm{E}-4$ | $2.8 \mathrm{E}-8$ | $4.6 \mathrm{E}-8$ | $2.8 \mathrm{E}-8$ |
| 4.0 | 8.782057 | $-6.4 \mathrm{E}-2$ | $-4.2 \mathrm{E}-4$ | $2.9 \mathrm{E}-8$ | $4.7 \mathrm{E}-8$ | $2.9 \mathrm{E}-8$ |

of the price panel: $\operatorname{Spr}\left(S_{0} ; K_{1}\right)=K_{1} \cdot \operatorname{spread}(1,1), \operatorname{Spr}\left(S_{0} ; K_{1} e^{-\eta^{*}}\right)=K_{1} e^{-\eta^{*}} \cdot \operatorname{spread}(2,2)$, $\operatorname{Spr}\left(S_{0} ; K_{1} e^{-2 \eta^{*}}\right)=K_{1} e^{-2 \eta^{*}} \cdot \operatorname{spread}(3,3), \ldots$ The results of this technique are recorded in Tables 2 and 3 in the column "Interpolation." We can see this technique generates very accurate results and moreover, saves computational resources.

Finally, we computed first order Greeks using the method described at the beginning of section 3 and compared them with finite differences. As seen in Table 4, the two methods come up with very consistent results. The Greeks of our at-the-money spread option exhibit some resemblance to those of the at-the-money European put/call option. The delta of $S_{1}$ is close to the delta of the call option, which is about 0.5 . And the delta of $S_{2}$ is close to the delta of the put option, which is also about 0.5 . The time premium of the spread option is positive. The option price is much more sensitive to $S_{1}$ volatility than to $S_{2}$ volatility. It is an important feature that the option price is negatively correlated with the underlying correlation: Intuitively speaking, if the two underlyings are strongly correlated, their comovements

Table 4
The Greeks for the GBM model compared between the FFT method and the finite difference method. The FFT method uses $N=2^{10}$ and $\bar{u}=40$. The finite difference uses a two-point central formula, in which the displacement is $\pm 1 \%$. Other parameters are the same as Table 1 except that we fix the strike $K=4.0$ to make the option at-the-money.

|  | Delta(S1) | Delta(S2) | Theta | $\operatorname{Vega}\left(\sigma_{1}\right)$ | $\operatorname{Vega}\left(\sigma_{2}\right)$ | $\partial$ Spr $/ \partial \rho$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| FD | 0.512648 | -0.447127 | 3.023823 | 33.114315 | -0.798959 | -4.193749 |
| FFT | 0.512705 | -0.447079 | 3.023777 | 33.114834 | -0.798972 | -4.193728 |

Table 5
Computing time of FFT for a panel of prices.

| Discretization | GBM | SV | VG |
| :---: | :---: | :---: | :---: |
| 64 | 0.091647 | 0.083326 | 0.109537 |
| 128 | 0.099994 | 0.120412 | 0.139276 |
| 256 | 0.126687 | 0.234024 | 0.220364 |
| 512 | 0.240938 | 0.711395 | 0.621074 |
| 1024 | 0.609860 | 2.628901 | 2.208770 |
| 2048 | 2.261325 | 10.243228 | 8.695122 |

diminish the probability that $S_{1 T}$ develops a wide spread over $S_{2 T}$. This result is consistent with observations made by [10].

Since the FFT method naturally generates a panel of prices and interpolation can be implemented accurately with negligible additional computational cost, it is appropriate to measure the efficiency of the method by timing the computation of a panel of prices. Such computing times are shown in Table 5. For the FFT method, the main computational cost comes from the calculation of the matrix $H$ in (3.2) and the subsequent FFT of $H$. We see that the GBM model is noticeably faster than the SV and VG models: This is due to a recursive method used to calculate the $H$ matrix entries of the GBM model, which is not applicable for the SV and VG models. The number of calculations for $H$ is of order $N^{2}$, which for large $N$ exceeds the $N \log N$ of the FFT of $H$, and thus the advantage of this efficient algorithm for the GBM model is magnified as $N$ increases. However, our FFT method is still very fast for the SV and VG models and is able to generate a large panel of prices within a couple of seconds.
7. High-dimensional basket options. The ideas of section 2 turn out to extend naturally to two particular classes of basket options on $M \geq 2$ assets.

Proposition 7.1. Let $M \geq 2$.

1. For any real numbers $\epsilon=\left(\epsilon_{1}, \ldots, \epsilon_{M}\right)$ with $\epsilon_{m}>0$ for $2 \leq m \leq M$ and $\epsilon_{1}<$ $-1-\sum_{m=2}^{M} \epsilon_{m}$,

$$
\begin{equation*}
\left(e^{x_{1}}-\sum_{m=2}^{M} e^{x_{m}}-1\right)^{+}=(2 \pi)^{-M} \int_{\mathbb{R}^{M}+i \epsilon} e^{i u x^{\prime}} \hat{P}^{M}(u) d^{M} u, \tag{7.1}
\end{equation*}
$$

where for $u=\left(u_{1}, \ldots, u_{M}\right) \in \mathbb{C}^{M}$

$$
\begin{equation*}
\hat{P}^{M}(u)=\frac{\Gamma\left(i\left(u_{1}+\sum_{m=2}^{M} u_{m}\right)-1\right) \prod_{m=2}^{M} \Gamma\left(-i u_{m}\right)}{\Gamma\left(i u_{1}+1\right)} . \tag{7.2}
\end{equation*}
$$

2. For any real numbers $\epsilon=\left(\epsilon_{1}, \ldots, \epsilon_{M}\right)$ with $\epsilon_{m}>0$ for all $m \leq M$,

$$
\begin{equation*}
\left(1-\sum_{m=1}^{M} e^{x_{m}}\right)^{+}=(2 \pi)^{-M} \int_{\mathbb{R}^{M}+i \epsilon} e^{i u x^{\prime}} \hat{Q}^{M}(u) d^{M} u \tag{7.3}
\end{equation*}
$$

where for $u=\left(u_{1}, \ldots, u_{M}\right) \in \mathbb{C}^{M}$

$$
\begin{equation*}
\hat{Q}^{M}(u)=\frac{\prod_{m=1}^{M} \Gamma\left(-i u_{m}\right)}{\Gamma\left(-i \sum_{m=1}^{M} u_{m}+2\right)} \tag{7.4}
\end{equation*}
$$

Remark. Clearly, these two results can be applied directly to obtain an $M$-dimensional FFT method to price $M$-asset basket options that pay off either $\left(S_{1 T}-S_{2 T}-\cdots-S_{M T}-1\right)^{+}$ or $\left(1-S_{1 T}-S_{2 T}-\cdots-S_{M T}\right)^{+}$. However, it is important to also note that by a generalized "put-call parity" one can also price options that pay off either $\left(1+S_{2 T}+\cdots+S_{M T}-S_{1 T}\right)^{+}$ or $\left(S_{1 T}+S_{2 T}+\cdots+S_{M T}-1\right)^{+}$.

Proof. The proof of both parts of the above proposition is based on a simple lemma proved in the appendix.

Lemma 7.2. Let $z \in \mathbb{R}$ and $u=\left(u_{1}, \ldots, u_{M}\right)^{\prime} \in \mathbb{C}^{M}$ with $\Im m\left(u_{m}\right)>0$ for all $m \leq M$. Then

$$
\begin{equation*}
\int_{\mathbb{R}^{M}} e^{z} \delta\left(e^{z}-\sum_{m=1}^{M} e^{x_{m}}\right) e^{-i u x^{\prime}} d^{M} x=\frac{\prod_{m=1}^{M} \Gamma\left(-i u_{m}\right)}{\Gamma\left(-i \sum_{m=1}^{M} u_{m}\right)} e^{-i\left(\sum_{m=1}^{M} u_{m}\right) z} \tag{7.5}
\end{equation*}
$$

To prove (7.2), we need to compute, for $u \in \mathbb{C}^{M}$,

$$
\hat{P}^{M}(u)=\int_{\mathbb{R}^{M}}\left(e^{x_{1}}-\sum_{m=2}^{M} e^{x_{m}}-1\right)^{+} e^{-i \tilde{u} \tilde{x}} d^{M} x
$$

We introduce the factor $1=\int_{\mathbb{R}} \delta\left(e^{z}-\sum_{m=2}^{M} e^{x_{m}}\right) e^{z} d z$ and interchange the $z$ integral with the $x$ integrals. Then using Lemma 7.2 one finds

$$
\begin{aligned}
\hat{P}^{M}(u) & =\int_{\mathbb{R}^{2}}\left(e^{x_{1}}-e^{z}-1\right)^{+}\left[\int_{\mathbb{R}^{M-1}} e^{z} \delta\left(e^{z}-\sum_{m=2}^{M} e^{x_{m}}\right) e^{-i u x^{\prime}} d x_{2} \ldots d x_{M}\right] d x_{1} d z \\
& =\frac{\prod_{m=2}^{M} \Gamma\left(-i u_{m}\right)}{\Gamma\left(-i \sum_{m=2}^{M} u_{m}\right)} \int_{\mathbb{R}^{2}} e^{-i u_{1} x_{1}} e^{-i\left(\sum_{m=2}^{M} u_{m}\right) z}\left(e^{x_{1}}-e^{z}-1\right)^{+} d x_{1} d z .
\end{aligned}
$$

We can then apply Theorem 1.1 and obtain the result.
The proof of (7.4) is similar to the proof of (7.2), where the two-dimensional problem can be deduced first and extended to higher dimensions with the application of Lemma 7.2.
8. Conclusion. This paper presents a new approach to the valuation of spread options, an important class of financial contracts. The method is based on a newly discovered explicit formula for the Fourier transform of the spread option payoff in terms of the gamma function.

In the final section we extended this formula to spread options in all dimensions and a certain class of basket options.

This mathematical result leads to simple and transparent algorithms for pricing spread options and other basket options in all dimensions. We have shown that the powerful tool of the FFT provides an accurate and efficient implementation of the pricing formula in low dimensions. For implementation of higher-dimensional problems, the curse of dimensionality sets in, and such cases should proceed using parallel partitioning methods as introduced in [9]. The difficulties and pitfalls of the FFT, of which there are admittedly several, are by now well understood, and thus the reliability and stability properties of our method are clear. We present a detailed discussion of errors and show which criteria determine the optimal choice of implementation parameters.

Many important processes in finance, particularly affine models and Lévy jump models, have well-known explicit characteristic functions and can be included in the method with little difficulty. Thus the method can easily be applied to important problems arising in energy and commodity markets.

Finally, the Greeks can be systematically evaluated for such models, with similar performance and little extra work.

While our method provides a basic analytic framework for spread options, much as has been done for one-dimensional options, it is certainly possible to add refinements that will improve convergence rates. Such techniques might include, for example, analytic computation of residues combined with contour deformation.

## Appendix. Proofs of Theorem 1.1 and Lemma 7.2.

Proof of Theorem 1.1. Suppose $\epsilon_{2}>0, \epsilon_{1}+\epsilon_{2}<-1$. One can then verify either directly or from the argument that follows that $e^{\epsilon \cdot x} P(x), \epsilon=\left(\epsilon_{1}, \epsilon_{2}\right)$, is in $\mathbb{L}^{2}\left(\mathbb{R}^{2}\right)$. Therefore, application of the Fourier inversion theorem to $e^{\epsilon \cdot x} P(x), \epsilon=\left(\epsilon_{1}, \epsilon_{2}\right)$, implies that

$$
\begin{equation*}
P(x)=(2 \pi)^{-2} \iint_{\mathbb{R}^{2}+i \epsilon} e^{i u \cdot x} g(u) d^{2} u \tag{A.1}
\end{equation*}
$$

where

$$
g(u)=\iint_{\mathbb{R}^{2}} e^{-i u \cdot x} P(x) d^{2} x
$$

By restricting to the domain $\left\{x: x_{1}>0, e^{x_{2}}<e^{x_{1}}-1\right\}$ we have

$$
\begin{aligned}
g(u) & =\int_{0}^{\infty} e^{-i u_{1} x_{1}}\left[\int_{-\infty}^{\log \left(e^{x_{1}}-1\right)} e^{-i u_{2} x_{2}}\left[\left(e^{x_{1}}-1\right)-e^{x_{2}}\right] d x_{2}\right] d x_{1} \\
& =\int_{0}^{\infty} e^{-i u_{1} x_{1}}\left(e^{x_{1}}-1\right)^{1-i u_{2}}\left[\frac{1}{-i u_{2}}-\frac{1}{1-i u_{2}}\right] d x_{1} .
\end{aligned}
$$

The change of variables $z=e^{-x_{1}}$ then leads to

$$
g(u)=\frac{1}{\left(1-i u_{2}\right)\left(-i u_{2}\right)} \int_{0}^{1} z^{i u_{1}}\left(\frac{1-z}{z}\right)^{1-i u_{2}} \frac{d z}{z} .
$$

The beta function

$$
B(a, b)=\frac{\Gamma(a) \Gamma(b)}{\Gamma(a+b)}
$$

is defined for any complex $a, b$ with $\Re e(a), \Re e(b)>0$ by

$$
B(a, b)=\int_{0}^{1} z^{a-1}(1-z)^{b-1} d z
$$

From this and the property $\Gamma(z)=(z-1) \Gamma(z-1)$ we have the formulas

$$
\begin{equation*}
g(u)=\frac{\Gamma\left(i\left(u_{1}+u_{2}\right)-1\right) \Gamma\left(-i u_{2}+2\right)}{\left(1-i u_{2}\right)\left(-i u_{2}\right) \Gamma\left(i u_{1}+1\right)}=\frac{\Gamma\left(i\left(u_{1}+u_{2}\right)-1\right) \Gamma\left(-i u_{2}\right)}{\Gamma\left(i u_{1}+1\right)} \tag{A.2}
\end{equation*}
$$

The above derivation also leads to the following bound on $\hat{P}$.
Corollary A.1. Fix $\epsilon_{2}=\epsilon, \epsilon_{1}=-1-2 \epsilon$ for some $\epsilon>0$. Then

$$
\begin{equation*}
\left|\hat{P}\left(u_{1}, u_{2}\right)\right| \leq \frac{\Gamma(\epsilon) \Gamma(2+\epsilon)}{\Gamma(2+2 \epsilon)} \cdot \frac{1}{Q\left(|u|^{2} / 5\right)^{1 / 2}} \tag{A.3}
\end{equation*}
$$

where $Q(z)=\left(z+\epsilon^{2}\right)\left(z+(1+\epsilon)^{2}\right)$.
Proof. First note that for $z_{1}, z_{2} \in \mathbb{C},\left|B\left(z_{1}, z_{2}\right)\right| \leq B\left(\Re e\left(z_{1}\right), \Re e\left(z_{2}\right)\right)$. Then (A.2) and a symmetric formula with $u_{2} \leftrightarrow-1-u_{1}-u_{2}$ lead to the upper bound

$$
\left|\hat{P}\left(u_{1}-i(\epsilon+1), u_{2}+i \epsilon\right)\right| \leq B(\epsilon, 2+\epsilon) \min \left(\frac{1}{Q\left(\left|u_{2}\right|\right)}, \frac{1}{Q\left(\left|u_{1}+u_{2}\right|\right)}\right)
$$

But since $Q$ is monotonic and $|u| \leq \sqrt{5} \max \left(\left|u_{2}\right|,\left|u_{1}+u_{2}\right|\right)$ for all $u \in \mathbb{R}^{2}$, the required result follows.

Proof of Lemma 7.2. We make the change of variables $p=e^{z}$ and $q_{m}=e^{x_{m}}$ and prove by induction that

$$
\begin{equation*}
\int_{\mathbb{R}^{M}} p \delta\left(p-\sum_{m=1}^{M} q_{m}\right) \prod_{m=1}^{M} q_{m}^{-i u_{m}-1} d^{M} q=\frac{\prod_{m=1}^{M} \Gamma\left(-i u_{m}\right)}{\Gamma\left(-i \sum_{m=1}^{M} u_{m}\right)} p^{-i\left(\sum_{m=1}^{M} u_{m}\right)} \tag{A.4}
\end{equation*}
$$

The above equation trivially holds when $M=1$. If it holds for $M=N$, then for $M=N+1$ one finds

$$
\begin{align*}
& L H S=\int_{\mathbb{R}^{N+1}} p \delta\left(p-q_{N+1}-\sum_{m=1}^{N} q_{m}\right) q_{N+1}^{-i u_{N+1}-1} \prod_{m=1}^{N} q_{m}^{-i u_{m}-1} d^{N+1} q \\
& \quad=\frac{\prod_{m=1}^{N} \Gamma\left(-i u_{m}\right)}{\Gamma\left(-i \sum_{m=1}^{N} u_{m}\right)} \int_{0}^{p} \frac{p}{p-q_{N+1}}\left(p-q_{N+1}\right)^{-i\left(\sum_{m=1}^{N} u_{m}\right)} q_{N+1}^{-i u_{N+1}-1} d q_{N+1} \tag{A.5}
\end{align*}
$$

The proof is complete when one notices that the $q_{N+1}$ integral is simply $p^{-i\left(\sum_{m=1}^{N+1} u_{m}\right)}$ multiplied by a beta function with parameters $-i\left(\sum_{m=1}^{N} u_{m}\right)$ and $-i u_{N+1}$.

## REFERENCES

[1] A. Antonov and M. Arneguy, Analytical Formulas for Pricing CMS Products in the LIBOR Market Model with the Stochastic Volatility, SSRN eLibrary, 2009.
[2] R. Carmona and V. Durrleman, Pricing and hedging spread options, SIAM Rev., 45 (2003), pp. 627-685.
[3] R. Carmona and V. Durrleman, Pricing and Hedging Spread Options in a Log-Normal Model, Tech. report, Princeton University, Princeton, NJ, 2003. Available online from http://orfe.princeton.edu/ ~rcarmona/download/fe/spread.pdf.
[4] P. Carr and D. Madan, Option valuation using the fast Fourier transform, J. Comput. Finance, 2 (1999), pp. 61-73.
[5] M. A. H. Dempster and S. S. G. Hong, Spread option valuation and the fast Fourier transform, in Mathematical Finance - Bachelier Congress 2000, Springer, Berlin, 2002, pp. 203-220.
[6] K. R. Jackson, S. Jaimungal, and V. Surkov, Fourier space time-stepping for option pricing with Lévy models, J. Comput. Finance, 12(2) (2008), pp. 1-29.
[7] E. Kirk, Correlations in the energy markets, in Managing Energy Price Risk, Risk Publications and Enron, London, 1995, pp. 71-78.
[8] R. Lee, Option pricing by transform methods: Extensions, unification, and error control, J. Comput. Finance, 7 (2004), pp. 51-86.
[9] C. C. W. Leentvaar and C. W. Oosterlee, Multi-asset option pricing using a parallel Fourier-based technique, J. Comput. Finance, 12(1) (2008), pp. 1-26.
[10] M. Li, S. Deng, and J. Zhou, Closed-form approximations for spread option prices and Greeks, J. Derivatives, 15 (2008), pp. 58-80.
[11] R. Lord, F. Fang, F. Bervoets, and C. W. Oosterlee, A fast and accurate FFT-based method for pricing early-exercise options under Lévy processes, SIAM J. Sci. Comput., 30 (2008), pp. 1678-1705.
[12] D. Madan, P. Carr, and E. Chang, The variance gamma model and option pricing, Eur. Finance Rev., 3 (1998), pp. 79-105.
[13] D. Madan and E. Seneta, The VG model for share market returns, J. Business, 63 (1990), pp. 511-524.
[14] W. Margrabe, The value of an option to exchange one asset for another, J. Finance, 33 (1978), pp. 177-186.
[15] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, Numerical Recipes: The Art of Scientific Computing, 3rd ed., Cambridge University Press, Cambridge, UK, 2007.

# Hedging of Claims with Physical Delivery under Convex Transaction Costs* 

Teemu Pennanen ${ }^{\dagger}$ and Irina Penner ${ }^{\ddagger}$


#### Abstract

We study superhedging of contingent claims with physical delivery in a discrete-time market model with convex transaction costs. Our model extends Kabanov's currency market model by allowing for nonlinear illiquidity effects. We show that an appropriate generalization of Schachermayer's robust no-arbitrage condition implies that the set of claims hedgeable with zero cost is closed in probability. Combined with classical techniques of convex analysis, the closedness yields a dual characterization of premium processes that are sufficient to superhedge a given claim process. We also extend the fundamental theorem of asset pricing for general conical models.


Key words. superhedging, physical delivery, illiquidity, transaction costs, convex duality
AMS subject classifications. 91B25, 52A07, 46A20
DOI. 10.1137/090754182

1. Introduction. This paper studies superhedging of contingent claims with physical delivery in markets with temporary illiquidity effects. Our market model is a generalization of the currency market model of Kabanov [12]. In Kabanov's model, price dynamics and transaction costs are modeled implicitly by solvency cones, i.e., sets of portfolios which can be transformed into the zero portfolio by self-financing transactions at a given time and state. An essential difference between Kabanov's model and more traditional models of mathematical finance (including, e.g., the transaction cost models of Jouini and Kallal [11], Cvitanić and Karatzas [6], and Kaval and Molchanov [18]) is that Kabanov's model focuses on contingent claims with physical delivery, i.e., claims whose payouts are given in terms of portfolios of assets instead of a single reference asset like cash. Accordingly, notions of arbitrage as well as the corresponding dual variables are defined in terms of vector-valued processes; see, e.g., [15, 31].

Astic and Touzi [2] extended Kabanov's model by allowing general convex solvency regions in the case of finite probability spaces. Nonconical solvency regions allow for the modeling of temporary illiquidity effects where marginal trading costs may depend on the magnitude of a trade as, e.g., in Çetin, Jarrow, and Protter [3], Çetin and Rogers [4], Çetin, Soner, and Touzi [5], Rogers and Singh [27], or Pennanen [19, 20]. These models cover nonlinear illiquidity effects, but they assume that agents have no market power in the sense that their trades do not affect the costs of subsequent trades; see [27] for further motivation of this assumption.

[^15]Temporal illiquidity effects act essentially as nonlinear transaction costs. Moreover, most modern stock exchanges are organized so that the costs are convex with respect to transacted amounts; see [19].

This paper studies general convex solvency regions in general probability spaces in finite discrete time. Using some classical techniques of convex analysis we extend the approaches developed in [15] and [31] for the models with proportional transaction costs. We give a sufficient condition for the closedness of the set of contingent claims with physical delivery that can be hedged with zero investment under convex transaction costs. Classical separation arguments then yield dual characterizations of superhedging conditions much as in Pennanen [21] in the case of claims with cash delivery. In the conical case, our sufficient condition coincides with the robust no-arbitrage condition and the dual variables become consistent price systems in the sense of Schachermayer [31]. We also give a version of the "fundamental theorem of asset pricing" for general conical models. Even in the conical case our results improve the existing ones since we do not assume polyhedrality of the solvency cones.

The rest of this paper is organized as follows. Section 2 describes the market model, and in section 3 the robust no scalable arbitrage property is introduced. The relevance of this property becomes clear in Theorem 3.3, which provides a basis for the main results in sections 4 and 5 . Section 4 combines Theorem 3.3 with some classical techniques of convex analysis to derive dual characterizations of superhedging conditions. Section 5 generalizes the fundamental theorem of asset pricing to the general conical case. The proof of Theorem 3.3 is contained in section 6 .
2. The market model. We consider a financial market in which $d$ securities can be traded over finite discrete time $t=0, \ldots, T$. The information evolves according to a filtration $\left(\mathcal{F}_{t}\right)_{t=0}^{T}$ on a probability space $(\Omega, \mathcal{F}, P)$.

For each $t$ and $\omega$ we denote by $C_{t}(\omega) \subset \mathbb{R}^{d}$ the set of portfolios that are freely available in the market. We assume that for each $t$ the set-valued mapping $C_{t}: \Omega \rightrightarrows \mathbb{R}^{d}$ is $\mathcal{F}_{t}$-measurable in the sense that

$$
C_{t}^{-1}(U):=\left\{\omega \in \Omega \mid C_{t}(\omega) \cap U \neq \emptyset\right\} \in \mathcal{F}_{t}
$$

for every open set $U \subset \mathbb{R}^{d}$.
Definition 2.1. A market model is an $\left(\mathcal{F}_{t}\right)_{t=0}^{T}$-adapted sequence $C=\left(C_{t}\right)_{t=0}^{T}$ of closedvalued mappings $C_{t}: \Omega \rightrightarrows \mathbb{R}^{d}$ with $\mathbb{R}_{-}^{d} \subset C_{t}(\omega)$ for every $t$ and $\omega$. A market model $C$ is convex, conical, polyhedral, etc., if $C_{t}(\omega)$ has the corresponding property for every $t$ and $\omega$.

Traditionally, portfolios in financial market models have been defined in terms of a reference asset such as cash or some other numéraire; see Example 2.3. This is natural when studying financial contracts with cash payments only. Treating all assets symmetrically as in Definition 2.1 was initiated by Kabanov [12].

Example 2.2 (currency markets with proportional transaction costs). If $\left(s_{t}\right)_{t=0}^{T}$ is an adapted price process with values in $\mathbb{R}_{+}^{d}$ and $\left(\Lambda_{t}\right)_{t=0}^{T}$ an adapted $\mathbb{R}_{+}^{d \times d}$-valued process of transaction cost coefficients, the solvency regions (in physical units) were defined in Kabanov [12] as

$$
\hat{K}_{t}:=\left\{x \in \mathbb{R}^{d} \mid \exists a \in \mathbb{R}_{+}^{d \times d}: s_{t}^{i} x^{i}+\sum_{j=1}^{d}\left(a^{j i}-\left(1+\lambda_{t}^{i j}\right) a^{i j}\right) \geq 0,1 \leq i \leq d\right\}
$$

A portfolio $x$ belongs to the solvency region $\hat{K}_{t}$ if and only if, after some possible transfers $\left(a^{j i}\right)_{1 \leq i, j \leq d}$, it has only nonnegative components. Thus the solvency region describes the set of all portfolios with "positive" values.

One can also define solvency regions directly in terms of bid-ask spreads as in Schachermayer [31]. If $\left(\Pi_{t}\right)_{t=0}^{T}$ is an adapted sequence of bid-ask matrices, then

$$
\hat{K}_{t}=\left\{x \in \mathbb{R}^{d} \mid \exists a \in \mathbb{R}_{+}^{d \times d}: x^{i}+\sum_{j=1}^{d}\left(a^{j i}-\pi_{t}^{i j} a^{i j}\right) \geq 0,1 \leq i \leq d\right\}
$$

For each $\omega$ and $t$, the set $\hat{K}_{t}(\omega)$ is a polyhedral cone and

$$
C_{t}(\omega):=-\hat{K}_{t}(\omega)
$$

defines a conical market model in the sense of Definition 2.1.
Example 2.3 (illiquid markets with cash). A convex cost process is a sequence $S=\left(S_{t}\right)_{t=0}^{T}$ of extended real-valued functions on $\mathbb{R}^{d} \times \Omega$ such that for all $t$ the function $S_{t}$ is $\mathcal{B}\left(\mathbb{R}^{d}\right) \otimes \mathcal{F}_{t^{-}}$ measurable and for each $\omega$ the function $S_{t}(\cdot, \omega)$ is lower semicontinuous and convex and vanishes at 0 ; see $[20,21]$. The quantity $S_{t}(x, \omega)$ denotes the cost (in cash) of buying a portfolio $x$ at time $t$ and scenario $\omega$. If $S$ is a convex cost process, then

$$
C_{t}(\omega)=\left\{x \in \mathbb{R}^{d} \mid S_{t}(x, \omega) \leq 0\right\}, \quad t=0, \ldots, T
$$

defines a convex market model provided $S_{t}(\cdot, \omega)$ are componentwise nondecreasing for every $t$ and $\omega$. The main difference between the model studied in [20,21] and the one studied here is that those papers studied contingent claims with cash delivery in markets with portfolio constraints possibly on all traded assets, including cash. The present paper studies contingent claims with physical delivery but allows for free transfer of all positions through time.

Models with convex cost processes include, in particular, classical frictionless markets with a cash account (where $S_{t}\left(\left(x^{0}, x^{1}\right), \omega\right)=x^{0}+s_{t}(\omega) \cdot x^{1}$ for an adapted $\mathbb{R}^{d}$-valued price process $s$ ) as well as models with bid-ask spreads or proportional transaction costs as, e.g., in Jouini and Kallal [11]. Convex cost processes also allow for modeling of illiquidity effects as, e.g., in Çetin and Rogers [4], where $d=2$ and $S_{t}\left(\left(x^{0}, x^{1}\right), \omega\right)=x^{0}+s_{t}(\omega) \varphi\left(x^{1}\right)$ for a strictly positive adapted process $\left(s_{t}\right)_{t=0}^{T}$ and an increasing convex function $\varphi: \mathbb{R} \rightarrow(-\infty, \infty]$ that describes the costs of illiquidity.

A convex cost process can be identified with the liquidation function $P_{t}$ in Astic and Touzi [2] through $S_{t}(x, \omega)=-P_{t}(-x, \omega)$. Convex cost processes are also related to the supply curve introduced in Çetin, Jarrow, and Protter [3]. A supply curve $s_{t}(x, \omega)$ gives a price per unit when buying $x$ units of the risky asset so that the total cost is $S_{t}(x, \omega)=s_{t}(x, \omega) x$. Instead of convexity of $S$, [3] assumed that the supply curve is smooth in $x$; see also Example 2.2 in [2].

Example 2.4 (currency markets with illiquidity costs). In order to model nonproportional illiquidity effects in a currency market model as in Example 2.2, one can replace a bid-ask matrix $\left(\Pi_{t}\right)_{t=0}^{T}$ by a matrix of convex cost processes $S^{i j}=\left(S_{t}^{i j}\right)_{t=0}^{T}(1 \leq i, j \leq d)$ on $\mathbb{R}_{+}$. Here $S^{i j}(x, \omega)$ denotes the number of units of asset $i$ for which one can buy $x$ units of asset $j$. If $\left(S_{t}^{i j}\right)_{t=0}^{T}, i, j=1, \ldots, d$, are convex cost processes on $\mathbb{R}_{+}$in the sense of Example 2.3, then

$$
C_{t}(\omega)=\left\{x \in \mathbb{R}^{d} \mid \exists a \in \mathbb{R}_{+}^{d \times d}: x^{i} \leq \sum_{j=1}^{d}\left(a^{j i}-S_{t}^{i j}\left(a^{i j}, \omega\right)\right), 1 \leq i \leq d\right\}
$$

for $t=0, \ldots, T$ defines a convex market model. Markets with proportional transaction costs correspond to $S^{i j}(x, \omega)=\pi^{i j}(\omega) x$.
3. No-arbitrage criteria. An $\mathbb{R}^{d}$-valued adapted process $x=\left(x_{t}\right)_{t=0}^{T}$ is a self-financing portfolio process in a market model $C=\left(C_{t}\right)_{t=0}^{T}$ if

$$
\Delta x_{t}:=x_{t}-x_{t-1} \in C_{t} \quad P \text {-a.s. }
$$

for every $t=0, \ldots, T$; i.e., the increments $\Delta x_{t}$ are freely available in the market. Here and in what follows, we always define $x_{-1}=0$ and denote by $L^{0}\left(C_{t}, \mathcal{F}_{t}\right)$ the set of all $\mathcal{F}_{t}$-measurable selectors of a set-valued mapping $C_{t}: \Omega \rightrightarrows \mathbb{R}^{d}$, i.e., the set of all $\mathcal{F}_{t}$-measurable random vectors $x$ such that $x \in C_{t}$ almost surely. In particular, $L^{0}\left(\mathbb{R}^{d}, \mathcal{F}_{t}\right)$ denotes the set of all $\mathcal{F}_{t}$-measurable random vectors $x: \Omega \rightarrow \mathbb{R}^{d}$.

We say that a market model $C$ has the no-arbitrage property if

$$
\begin{equation*}
A_{T}(C) \cap L^{0}\left(\mathbb{R}_{+}^{d}, \mathcal{F}\right)=\{0\} \tag{3.1}
\end{equation*}
$$

where $A_{T}(C)$ denotes the set in $L^{0}$ formed by final values $x_{T}$ of all self-financing portfolio processes $x=\left(x_{t}\right)_{t=0}^{T}$. Since $x_{T}=\sum_{t=0}^{T}\left(x_{t}-x_{t-1}\right)$, we have the expression

$$
A_{T}(C)=L^{0}\left(C_{0}, \mathcal{F}_{0}\right)+\cdots+L^{0}\left(C_{T}, \mathcal{F}_{T}\right)
$$

Condition (3.1) was introduced in Kabanov and Stricker [16] for conical models under the name "weak no-arbitrage property" in a formally different way. But it is equivalent to (3.1) if $C_{t}$ contains $\mathbb{R}_{-}^{d}$ for all $t$ as noted in Lemma 3.5 of Kabanov [13].

In classical market models the no-arbitrage condition implies the closedness of the set of contingent claims with cash delivery that can be superhedged at zero cost, a result which is of vital importance in deriving dual characterizations of superhedging and the absence of arbitrage. However, as shown in Schachermayer [31], in a market with proportional transaction costs as in Example 2.2, the no-arbitrage property (3.1) does not, in general, imply the closedness of the set $A_{T}(C)$. Schachermayer [31] also showed, in the case of the conical model of Example 2.2, that $A_{T}(C)$ is closed in probability if $C$ satisfies the robust no-arbitrage condition, which can be defined in the general conical case as follows.

Given a market model $C$ let $C_{t}^{0}(\omega)$ be the largest linear subspace contained in $C_{t}(\omega)$. Using the terminology of Kabanov, Rásonyi, and Stricker [15] we say that $C$ is dominated by another market model $\tilde{C}$ if

$$
C_{t} \subset \tilde{C}_{t} \quad \text { and } \quad C_{t} \backslash C_{t}^{0} \subset \operatorname{ri} \tilde{C}_{t} \quad \forall t=0, \ldots, T
$$

where ri $\tilde{C}_{t}$ denotes the relative interior of a convex set $\tilde{C}_{t}$; see [ 24 , Chapter 6$]$ for the definition. Since in our setting $\mathbb{R}_{-}^{d} \subseteq C_{t} \subseteq \tilde{C}_{t}$, the relative interior of $\tilde{C}_{t}$ coincides with the interior.

Definition 3.1. A conical market model has the robust no-arbitrage property if it is dominated by another conical market model that has the no-arbitrage property.

When moving to general convex market models it is not immediately clear how the condition of robust no-arbitrage should be extended in order to have the closedness of $A_{T}(C)$. Indeed, in general convex models even the traditional notion of arbitrage has two natural
extensions, one being the possibility of making something out of nothing and the other one being the possibility of making arbitrarily much out of nothing; see [20]. These correspond to the notions of the tangent cone and the recession cone from convex analysis. It turns out to be the latter one which is more relevant for closedness of $A_{T}(C)$. This was noted in [21, Theorem 18] for the case of claims with cash delivery under convex constraints and is, in fact, already suggested by classical closedness criteria in convex analysis; see [24, Chapter 9].

Given a convex market model $C$, let

$$
C_{t}^{\infty}(\omega)=\left\{x \in \mathbb{R}^{d} \mid C_{t}(\omega)+\alpha x \subset C_{t}(\omega) \forall \alpha>0\right\} .
$$

This is a closed convex cone known as the recession cone of $C_{t}(\omega)$; see [24, Chapter 8]. The recession cone describes the asymptotic behavior of a convex set infinitely far from the origin. Since $C_{t}(\omega)$ is a closed convex set containing $\mathbb{R}_{-}^{d}$ we have, by [24, Theorem 8.1, Theorem 8.2, Corollary 8.3.2, Theorem 8.3], that $C^{\infty}$ is a closed convex cone containing $\mathbb{R}_{-}^{d}$ and

$$
C_{t}^{\infty}(\omega)=\bigcap_{\alpha>0} \alpha C_{t}(\omega)
$$

and

$$
\begin{equation*}
C_{t}^{\infty}(\omega)=\left\{x \in \mathbb{R}^{d} \mid \exists x^{n} \in C_{t}(\omega), \alpha^{n} \searrow 0, \text { with } \alpha^{n} x^{n} \rightarrow x\right\} . \tag{3.2}
\end{equation*}
$$

By [26, Exercise 14.21], the set-valued mappings $\omega \mapsto C_{t}^{\infty}(\omega)$ are $\mathcal{F}_{t}$-measurable, so they define a convex conical market model in the sense of Definition 2.1.

Definition 3.2. A convex market model has the robust no scalable arbitrage property if $C^{\infty}$ has the robust no-arbitrage property.

The term "scalable arbitrage" refers to arbitrage opportunities that may be scaled by arbitrarily large positive numbers to yield arbitrarily "large" arbitrage opportunities; see [20]. Such scalable arbitrage opportunities can be related to the market model $C^{\infty}$ much as in [20, Proposition 17].

We are now ready to state our main result, the proof of which can be found in the last section.

Theorem 3.3. If $C$ is a convex market model with the robust no scalable arbitrage property, then $A_{T}(C)$ is closed in probability.

Remark 3.4. If $C$ is conical, we have $C_{t}^{\infty}(\omega)=C_{t}(\omega)$ and Theorem 3.3 coincides with [15, Lemma 2], which extends [31, Theorem 2.1] to general conical models.

Remark 3.5. For general convex models, $C_{t}^{\infty}(\omega) \subseteq C_{t}(\omega)$ and the condition in Theorem 3.3 may be satisfied even if $C$ fails the no-arbitrage condition. Consider, for example, a deterministic model where $\Omega$ is a singleton and $C_{t}(\omega)=\mathbb{R}_{-}^{d}+\mathbb{B}$ for every $t$. Here $\mathbb{B}$ denotes the unit ball of $\mathbb{R}^{d}$. We get

$$
A_{T}(C)=L^{0}\left(C_{0}, \mathcal{F}_{0}\right)+\cdots+L^{0}\left(C_{T}, \mathcal{F}_{T}\right)=\mathbb{R}_{-}^{d}+(T+1) \mathbb{B},
$$

so $C$ does not have the no-arbitrage property. On the other hand, $C_{t}^{\infty}(\omega)=\mathbb{R}_{-}^{d}$ is dominated by $\tilde{C}_{t}(\omega)=\left\{x \in \mathbb{R}^{d} \mid \sum_{i=1}^{d} x^{i} \leq 0\right\}$, which does have the no-arbitrage property. Indeed, $A_{T}(\tilde{C})=\left\{x \in \mathbb{R}^{d} \mid \sum_{i=1}^{d} x^{i} \leq 0\right\}$, so $A_{T}(\tilde{C}) \cap L^{0}\left(\mathbb{R}_{+}^{d}, \mathcal{F}\right)=\{0\}$.

The following illustrates the robust no scalable arbitrage condition in a model with a cash account.

Example 3.6 (the model of Çetin and Rogers [4]). As observed in Example 2.3, the model of Çetin and Rogers [4] corresponds to the two-dimensional model

$$
\begin{equation*}
C_{t}(\omega)=\left\{\left(x^{0}, x^{1}\right) \in \mathbb{R}^{2} \mid x^{0}+s_{t}(\omega) \varphi\left(x^{1}\right) \leq 0\right\}, \quad t=0, \ldots, T, \tag{3.3}
\end{equation*}
$$

for a strictly positive adapted marginal price process $\left(s_{t}\right)_{t=0}^{T}$ and an increasing convex function $\varphi: \mathbb{R} \rightarrow(-\infty, \infty]$ with $\varphi(0)=0$. In this case, the recession cones become

$$
\begin{aligned}
C_{t}^{\infty}(\omega) & =\left\{\left(x^{0}, x^{1}\right) \mid \alpha x^{0}+s_{t}(\omega) \varphi\left(\alpha x^{1}\right) \leq 0 \forall \alpha>0\right\} \\
& =\left\{\left(x^{0}, x^{1}\right) \mid x^{0}+s_{t}(\omega) \varphi^{\infty}\left(x^{1}\right) \leq 0\right\},
\end{aligned}
$$

where $\varphi^{\infty}(x)=\sup _{\alpha>0} \alpha^{-1} \varphi(\alpha x)$ is the recession function of $\varphi$; see [24, Corollary 8.5.2]. By [24, Theorem 13.3], we have the expression

$$
\varphi^{\infty}(x)=\sup _{y \in I} x y
$$

where $I$ is the range of the gradient mapping $x \mapsto \varphi^{\prime}(x)$; see [24, page 227].
If we assume, as in [4], that $\inf _{x} \varphi^{\prime}(x)=0, \sup _{x} \varphi^{\prime}(x)=\infty$, we have

$$
\varphi^{\infty}(x)= \begin{cases}0 & \text { if } x \leq 0 \\ +\infty & \text { if } x>0\end{cases}
$$

and thus $C_{t}^{\infty}(\omega)=\mathbb{R}_{-}^{2}$ almost surely for all $t$. In this case the robust no scalable arbitrage condition holds with an arbitrary strictly positive marginal price process $\left(s_{t}\right)_{t=0}^{T}$ whether it satisfies the classical no-arbitrage condition or not. In [4] the above growth properties of $\varphi$ were used to prove the existence of optimal strategies in utility maximization problems, and it was noted that the usual no-arbitrage condition on the price process $\left(s_{t}\right)_{t=0}^{T}$ played no role in the proof. The nonexistence of arbitrage opportunities (not just the scalable ones) is related to a stronger no-arbitrage condition as formulated, e.g., in [2] or [20].

More precise characterizations of the robust no scalable arbitrage condition for models with a cash account will be given at the end of section 5 .
4. Superhedging. A contingent claim with physical delivery is a security that, at some future time, gives its owner a random portfolio of securities (instead of a single security like in the case of cash delivery). A contingent claim process with physical delivery $c=\left(c_{t}\right)_{t=0}^{T}$ is a security that, at each time $t=0, \ldots, T$, gives its owner an $\mathcal{F}_{t}$-measurable random portfolio $c_{t} \in \mathbb{R}^{d}$. The linear space of $\mathbb{R}^{d}$-valued adapted processes will be denoted by $\mathcal{A}$, i.e.,

$$
\mathcal{A}=L^{0}\left(\mathbb{R}^{d}, \mathcal{F}_{0}\right) \times \cdots \times L^{0}\left(\mathbb{R}^{d}, \mathcal{F}_{T}\right)
$$

We call the elements of $\mathcal{A}$ portfolio processes and denote by $\mathcal{A}_{+}$the set of all nonnegative portfolio processes:

$$
\mathcal{A}_{+}=L^{0}\left(\mathbb{R}_{+}^{d}, \mathcal{F}_{0}\right) \times \cdots \times L^{0}\left(\mathbb{R}_{+}^{d}, \mathcal{F}_{T}\right)
$$

Given a market model $C$, we will say that a process $p \in \mathcal{A}$ is a superhedging premium process for a claim process $c \in \mathcal{A}$ if there is a portfolio process $x \in \mathcal{A}$ such that $x_{T}=0$ and

$$
x_{t}-x_{t-1}+c_{t}-p_{t} \in C_{t}
$$

almost surely for every $t=0, \ldots, T$. The requirement that $x_{T}=0$ means that everything is liquidated at the terminal date. One could relax this condition to $x_{T} \geq 0$, but since $\mathbb{R}_{-}^{d} \subset C_{T}$ it would amount to the same thing.

We use claim and premium processes (rather than claims and premiums (prices) with payouts only at the end and the beginning) in the present paper mainly for mathematical convenience. However, when moving to market models with portfolio constraints it is essential to distinguish between payments at different points in time, and then claim processes become the natural object of study; see [20,21]. Claim and premium processes are common in various insurance applications where payments are made, e.g., annually.

The superhedging condition can be written as

$$
c-p \in A(C)
$$

where

$$
A(C)=\left\{c \in \mathcal{A} \mid \exists x \in \mathcal{A}: x_{t}-x_{t-1}+c_{t} \in C_{t}, x_{T}=0\right\}
$$

is the set of all claim processes with physical delivery that can be superhedged without any investment. It is easily checked that $A(C)$ is a convex subset (a convex cone) of $\mathcal{A}$ if $C$ is a convex (conical) market model.

Lemma 4.1. Let $C$ be a convex market model.

1. The sets $A(C)$ and $A_{T}(C)$ are related through

$$
\begin{aligned}
A_{T}(C) & =\left\{c_{T} \mid\left(0, \ldots, 0, c_{T}\right) \in A(C)\right\} \\
A(C) & =\left\{\left(c_{0}, \ldots, c_{T}\right) \mid \sum_{t=0}^{T} c_{t} \in A_{T}(C)\right\} .
\end{aligned}
$$

2. The set $A(C)$ is closed in probability if and only if $A_{T}(C)$ is closed in probability.
3. $C$ has the no-arbitrage property if and only if $A(C) \cap \mathcal{A}_{+}=\{0\}$.

Proof. It suffices to prove the first part since the other two follow from it. The first equation is immediate. As to the second, we have $c \in A(C)$ if and only if there is an $x \in \mathcal{A}$ such that $x_{T}=0$ and $x_{t}-x_{t-1}+c_{t} \in C_{t}$. Defining $\tilde{x}_{-1}=0$ and $\tilde{x}_{t}=\tilde{x}_{t-1}+x_{t}-x_{t-1}+c_{t}$, we get that $\tilde{x} \in \mathcal{A}$ is self-financing and $\tilde{x}_{T}=\sum_{t=0}^{T} c_{t}$. This just means that $\sum_{t=0}^{T} c_{t} \in A_{T}(C)$.

Example 4.2. The classical case of a single premium payment at the beginning and a single claim payment at maturity corresponds to $p=\left(p_{0}, 0, \ldots, 0\right)$ and $c=\left(0, \ldots, 0, c_{T}\right)$. In this case, Lemma 4.1 says that $p$ is a superhedging premium for $c$ if and only if

$$
\begin{equation*}
c_{T}-p_{0} \in A_{T}(C) . \tag{4.1}
\end{equation*}
$$

Dual characterizations of the set of all initial endowments satisfying condition (4.1) have been given in the conical case as in Example 2.2 in [16, 7, 14, 22, 31]. In conical models the superhedging endowments are characterized in terms of the same dual elements that
characterize the no-arbitrage condition. When moving to nonconical models, a larger class of dual variables is needed in order to capture the structure of the sets; see [21] for the case of claims with cash delivery.

By Lemma 4.1, the set $A(C)$ is closed if and only if $A_{T}(C)$ is closed. Combining Theorem 3.3 with classical techniques of convex analysis, we can derive dual characterizations of superhedging premium processes in terms of martingales much as in [21] in the case of claims with cash delivery. Consider the Banach space

$$
\mathcal{A}^{1}:=L^{1}\left(\Omega, \mathcal{F}_{0}, P ; \mathbb{R}^{d}\right) \times \cdots \times L^{1}\left(\Omega, \mathcal{F}_{T}, P ; \mathbb{R}^{d}\right)
$$

and its dual

$$
\mathcal{A}^{\infty}:=L^{\infty}\left(\Omega, \mathcal{F}_{0}, P ; \mathbb{R}^{d}\right) \times \cdots \times L^{\infty}\left(\Omega, \mathcal{F}_{T}, P ; \mathbb{R}^{d}\right)
$$

Note that both $\mathcal{A}^{1}$ and $\mathcal{A}^{\infty}$ are linear subspaces of $\mathcal{A}$. One can then use the classical bipolar theorem to characterize the superhedging condition in terms of the support function $\sigma_{A^{1}(C)}$ : $\mathcal{A}^{\infty} \rightarrow \overline{\mathbb{R}}$ of the set $A^{1}(C):=A(C) \cap \mathcal{A}^{1}$ of integrable claim processes. The support function is given by

$$
\sigma_{A^{1}(C)}(y)=\sup _{c \in A^{1}(C)} E \sum_{t=0}^{T} c_{t} \cdot y_{t}, \quad y \in \mathcal{A}^{\infty} .
$$

The lemma below expresses $\sigma_{A^{1}(C)}$ in terms of the support functions of the random sets $C_{t}(\omega)$ :

$$
\sigma_{C_{t}(\omega)}(z)=\sup _{x \in C_{t}(\omega)} x \cdot z, \quad z \in \mathbb{R}^{d}
$$

By [26, Example 14.51] the function $\sigma_{C_{t}}: \Omega \times \mathbb{R}^{d} \rightarrow \overline{\mathbb{R}}$ is an $\mathcal{F}_{t}$-measurable convex normal integrand (see [26, Definition 14.27]). This implies, in particular, that $\sigma_{C_{t}(\omega)}\left(y_{t}\right)$ is an $\mathcal{F}_{t}$-measurable function whenever $y_{t}$ is $\mathcal{F}_{t}$-measurable. The following corresponds to [20, Lemma 28].

Lemma 4.3. Let $C$ be a convex market model, and let $y \in \mathcal{A}^{\infty}$. Then

$$
\sigma_{A^{1}(C)}(y)= \begin{cases}E \sum_{t=0}^{T} \sigma_{C_{t}}\left(y_{t}\right) & \text { if } y \text { is a nonnegative martingale, } \\ +\infty & \text { otherwise. }\end{cases}
$$

Proof. In the following, we define the expectation of an arbitrary random variable $\varphi$ by setting it equal to $-\infty$ if the negative part of $\varphi$ is not integrable (the remaining cases being defined unambiguously as real numbers or as $+\infty$ ). This allows us, in particular, to apply the results of [26, section 14.F] on extended real-valued integral functionals.

On one hand,

$$
\begin{aligned}
\sigma_{A^{1}(C)}(y) & =\sup \left\{E \sum_{t=0}^{T} c_{t} \cdot y_{t} \mid x \in \mathcal{A}, c \in \mathcal{A}^{1}: x_{t}-x_{t-1}+c_{t} \in C_{t}\right\} \\
& \leq \sup \left\{E \sum_{t=0}^{T} c_{t} \cdot y_{t} \mid x, c \in \mathcal{A}: x_{t}-x_{t-1}+c_{t} \in C_{t}\right\}
\end{aligned}
$$

$$
\begin{aligned}
& =\sup \left\{E \sum_{t=0}^{T}\left(w_{t}-x_{t}+x_{t-1}\right) \cdot y_{t} \mid x, w \in \mathcal{A}: w_{t} \in C_{t}\right\} \\
& \leq \sup \left\{E \sum_{t=0}^{T}\left[\sigma_{C_{t}}\left(y_{t}\right)+\left(x_{t-1}-x_{t}\right) \cdot y_{t}\right] \mid x \in \mathcal{A}\right\} \\
& =E \sum_{t=0}^{T} \sigma_{C_{t}}\left(y_{t}\right)+\sup _{x \in \mathcal{A}} E \sum_{t=0}^{T}\left(x_{t-1}-x_{t}\right) \cdot y_{t} \\
& =E \sum_{t=0}^{T} \sigma_{C_{t}}\left(y_{t}\right)+\sup _{x \in \mathcal{A}} E \sum_{t=0}^{T-1} x_{t} \cdot\left(y_{t+1}-y_{t}\right)
\end{aligned}
$$

where the last term vanishes if $y$ is a martingale and equals $+\infty$ otherwise. On the other hand,

$$
\begin{aligned}
\sigma_{A^{1}(C)}(y)= & \sup \left\{E \sum_{t=0}^{T} c_{t} \cdot y_{t} \mid x \in \mathcal{A}, c \in \mathcal{A}^{1}: x_{t}-x_{t-1}+c_{t} \in C_{t}\right\} \\
\geq & \sup \left\{E \sum_{t=0}^{T} c_{t} \cdot y_{t} \mid x, c \in \mathcal{A}^{1}: x_{t}-x_{t-1}+c_{t} \in C_{t}\right\} \\
= & \sup \left\{E \sum_{t=0}^{T}\left(w_{t}-x_{t}+x_{t-1}\right) \cdot y_{t} \mid x, w \in \mathcal{A}^{1}: w_{t} \in C_{t}\right\} \\
= & \sup \left\{E \sum_{t=0}^{T} w_{t} \cdot y_{t} \mid w \in \mathcal{A}^{1}: w_{t} \in C_{t}\right\}+\sup _{x \in \mathcal{A}^{1}} E \sum_{t=0}^{T}\left(x_{t-1}-x_{t}\right) \cdot y_{t} \\
= & \sum_{t=0}^{T} \sup \left\{E w_{t} \cdot y_{t} \mid w_{t} \in L^{1}\left(\Omega, \mathcal{F}_{t}, P ; \mathbb{R}^{d}\right): w_{t} \in C_{t}\right\} \\
& +\sup _{x \in \mathcal{A}^{1}} E \sum_{t=0}^{T-1} x_{t} \cdot\left(y_{t+1}-y_{t}\right)
\end{aligned}
$$

Here again the last term vanishes if $y$ is a martingale and equals $+\infty$ otherwise. Applying [26, Theorem 14.60] to the function

$$
f(w, \omega)= \begin{cases}-w \cdot y_{t}(\omega) & \text { if } w \in C_{t}(\omega) \\ +\infty & \text { otherwise }\end{cases}
$$

(which is a normal integrand by [26, Example 14.32]), we get

$$
\sup \left\{E w_{t} \cdot y_{t} \mid w_{t} \in L^{1}\left(\Omega, \mathcal{F}_{t}, P ; \mathbb{R}^{d}\right): w_{t} \in C_{t}\right\}=E \sigma_{C_{t}}\left(y_{t}\right)
$$

Thus,

$$
\sigma_{A^{1}(C)}(y)= \begin{cases}E \sum_{t=0}^{T} \sigma_{C_{t}}\left(y_{t}\right) & \text { if } y \text { is a martingale } \\ +\infty & \text { otherwise }\end{cases}
$$

Moreover, since $\mathbb{R}_{-}^{d} \subseteq C_{t}$ for all $t$, we have $\sigma_{C_{t}}\left(y_{t}\right)=\infty$ on the set $\left\{y_{t} \notin \mathbb{R}_{+}^{d}\right\}$. This completes the proof.

Theorem 4.4. Assume that the convex market model C has the robust no scalable arbitrage property, and let $c, p \in \mathcal{A}$ be such that $c-p \in \mathcal{A}^{1}$. Then the following are equivalent:
(i) $p$ is a superhedging premium process for $c$.
(ii) $E \sum_{t=0}^{T}\left(c_{t}-p_{t}\right) \cdot y_{t} \leq 1$ for every bounded nonnegative martingale $y=\left(y_{t}\right)_{t=0}^{T}$ such that $E \sum_{t=0}^{T} \sigma_{C_{t}}\left(y_{t}\right) \leq 1$.
(iii) $E \sum_{t=0}^{T}\left(c_{t}-p_{t}\right) \cdot y_{t} \leq E \sum_{t=0}^{T} \sigma_{C_{t}}\left(y_{t}\right)$ for every bounded nonnegative martingale $y=$
$\left(y_{t}\right)_{t=0}^{T}$.

Proof. By definition, $p$ is a superhedging premium for $c$ if and only if $c-p \in A(C)$. Since $c-p \in \mathcal{A}^{1}$, this can be written as $c-p \in A^{1}(C)$, where $A^{1}(C)=A(C) \cap \mathcal{A}^{1}$ is a closed subset of $\mathcal{A}^{1}$, by Theorem 3.3 and Lemma 4.1. Since $A^{1}(C)$ is also a convex set containing the origin, the bipolar theorem (see, e.g., [1, Theorem 5.91]) implies that $c-p \in A^{1}(C)$ if and only if

$$
E \sum_{t=0}^{T}\left(c_{t}-p_{t}\right) \cdot y_{t} \leq 1 \quad \forall y \in A^{1}(C)^{\circ},
$$

where

$$
A^{1}(C)^{\circ}=\left\{y \in \mathcal{A}^{\infty} \mid \sigma_{A^{1}(C)}(y) \leq 1\right\}
$$

Thus the equivalence of (i) and (ii) and the proof of (i) $\Rightarrow$ (iii) follow from Lemma 4.3. And obviously (iii) implies (ii).

As noted in Remark 3.5, arbitrage opportunities may very well exist under the conditions of Theorem 3.3. They do not interfere with the hedging description in Theorem 4.4.

If $C$ is conical, we have

$$
\sigma_{C_{t}(\omega)}(y)= \begin{cases}0 & \text { if } y \in C_{t}^{*}(\omega)  \tag{4.2}\\ +\infty & \text { otherwise }\end{cases}
$$

where the closed convex cone

$$
C_{t}^{*}(\omega):=\left\{y \in \mathbb{R}^{d} \mid x \cdot y \leq 0 \forall x \in C_{t}(\omega)\right\}
$$

is known as the polar of $C_{t}(\omega)$; see [24]. We then have $E \sum_{t=0}^{T} \sigma_{C_{t}}\left(y_{t}\right)<\infty$ for a $y \in \mathcal{A}^{\infty}$ if and only if $y_{t} \in C_{t}^{*}$ almost surely for all $t$. Thus, in the conical case, superhedging premiums can be characterized in terms of the following dual elements introduced in Kabanov [12].

Definition 4.5. An adapted $\mathbb{R}^{d}$-valued nonzero process $y=\left(y_{t}\right)_{t=0}^{T}$ is called a consistent (resp., strictly consistent) price system for a conical model $C$ if $y$ is a martingale such that $y_{t} \in C_{t}^{*}$ (resp., y has strictly positive components and $y_{t} \in \operatorname{ri} C_{t}^{*}$ ) almost surely for all $t=$ $0, \ldots, T$.

Note that the condition $y_{t} \in$ ri $C_{t}^{*}$ in general does not imply the strict positivity of $y_{t}$.
Applying Theorem 4.4 in the conical case and making use of (4.1) and (4.2), we obtain the following.

Corollary 4.6. Assume that $C$ is a conical market model with the robust no-arbitrage property. Assume further that $\mathcal{F}_{0}$ is a trivial $\sigma$-algebra. Let $c_{T} \in L^{1}\left(\Omega, \mathcal{F}_{T}, P ; \mathbb{R}^{d}\right)$ and $p_{0} \in \mathbb{R}^{d}$. Then the following are equivalent:
(i) $p=\left(p_{0}, 0, \ldots, 0\right)$ is a superhedging premium for $c=\left(0, \ldots, 0, c_{T}\right)$.
(ii) $E\left(c_{T} \cdot y_{T}\right) \leq p_{0} \cdot y_{0}$ for every bounded consistent price system $y=\left(y_{t}\right)_{t=0}^{T}$.

We illustrate Theorem 4.4 and Corollary 4.6 with the following example.
Example 4.7 (the model of Çetin and Rogers [4]). Consider again the model of Çetin and Rogers [4] described in Example 3.6. Given $z \in \mathbb{R}^{2}$, we have

$$
\begin{align*}
\sigma_{C_{t}(\omega)}(z) & =\sup _{x \in \mathbb{R}^{2}}\left\{x^{0} z^{0}+x^{1} z^{1} \mid x^{0}+s_{t}(\omega) \varphi\left(x^{1}\right) \leq 0\right\} \\
& = \begin{cases}\sup _{x^{1} \in \mathbb{R}}\left\{-z^{0} s_{t}(\omega) \varphi\left(x^{1}\right)+x^{1} z^{1}\right\} & \text { if } z^{0} \geq 0, \\
+\infty & \text { otherwise }\end{cases} \\
& = \begin{cases}z^{0} s_{t}(\omega) \varphi^{*}\left(\frac{z^{1}}{z^{0} s_{t}(\omega)}\right) & \text { if } z^{0}>0, \\
0 & \text { if } z^{0}=z^{1}=0, \\
+\infty & \text { otherwise },\end{cases} \tag{4.3}
\end{align*}
$$

where $\varphi^{*}(z)=\sup _{x}\{x z-\varphi(x)\}$ is the convex conjugate of $\varphi$.
In the frictionless case where $\varphi(x)=x$, we have $\varphi^{*}(z)=0$ if $z=1$ and $\varphi^{*}(z)=+\infty$ otherwise so that

$$
\sigma_{C_{t}(\omega)}(z)= \begin{cases}0 & \text { if } z^{0} \geq 0 \text { and } z^{1}=s_{t}(\omega) z_{0} \\ +\infty & \text { otherwise }\end{cases}
$$

It is easily checked that since $s$ is strictly positive, the robust no scalable arbitrage condition is equivalent to the classical no-arbitrage condition on $s$. The consistent price systems are the nonnegative martingales $y=\left(y^{0}, y^{1}\right)$ such that $y^{1}=s y^{0}$. A nonzero consistent price system defines a density of an absolutely continuous martingale measure for the price process $s$ via $d_{T}:=y_{T}^{0} / y_{0}^{0}$. Vice versa, every density process $d=\left(d_{t}\right)_{t=0}^{T}$ of an absolutely continuous martingale measure for $s$ defines a consistent price system $y$ by

$$
y_{t}:=\left(d_{t}, d_{t} s_{t}\right), \quad t=0, \ldots, T .
$$

The hedging condition (ii) of Corollary 4.6 can then be written as

$$
\begin{equation*}
E_{P^{*}}\left[c_{T}^{0}+c_{T}^{1} s_{T}\right] \leq p_{0}^{0}+p_{0}^{1} s_{0} \quad \forall P^{*} \in \tilde{\mathcal{P}} \tag{4.4}
\end{equation*}
$$

where $\tilde{\mathcal{P}}$ is the set of absolutely continuous martingale measures. In particular, if $c_{T}^{1}=0$, $p_{0}^{1}=0$, and if $s$ satisfies the no-arbitrage condition, we have that $p_{0}^{0}$ units of cash at time 0 are sufficient to superhedge $c_{T}^{0}$ units of cash at time $T$ if and only if

$$
\sup _{P^{*} \in \tilde{\mathcal{P}}} E_{P^{*}}\left[c_{T}^{0}\right] \leq p_{0}^{0}
$$

in accordance with classical results for claims with cash delivery; see, e.g., [9, Remark 1.35].
Now assume that there are nontrivial illiquidity costs in the sense that $\varphi(x) \geq x$ for every $x \in \mathbb{R}$ and $\varphi(x)>x$ for some $x \in \mathbb{R}$. Assume also that the model satisfies the robust no scalable arbitrage condition. This happens, e.g., when

$$
\varphi(x)=\frac{e^{\alpha x}-1}{\alpha}
$$

as in Çetin and Rogers [4, section 6]. Increasing the illiquidity costs shrinks the set of claims that can be superhedged with a given premium so that (4.4) is no longer a sufficient condition for superhedging. In other words, the superhedging premium under nontrivial illiquidity costs will in general be higher than in the corresponding frictionless model.

To illustrate this, consider the claim $c=\left(0, \ldots, 0, c_{T}\right)$ with $c_{T}:=\left(b s_{T},-b\right)$ for $b \in \mathbb{R}$. Condition (4.4) is clearly satisfied with $p_{0}=0$, so $c_{T}$ can be hedged without any costs in the frictionless model provided it has the no-arbitrage property. The existence of an $x \in$ $\mathbb{R}$ such that $\varphi(x)>x$, on the other hand, implies that the recession function $\varphi^{\infty}(x)=$ $\sup _{\alpha \geq 0} \alpha^{-1} \varphi(\alpha x)$ dominates the identity nontrivially. By [24, Theorem 13.3], this means that there is an $a>0, a \neq 1$, such that $\varphi^{*}(a)<\infty$. Consider the martingale $v$ defined by $v_{t}:=\left(d_{t} / a, d_{t} s_{t}\right)$, where $\left(d_{t}\right)$ is the density process of an absolutely continuous martingale measure for the price process $s$. Using (4.3), we get

$$
E \sigma_{C_{t}}\left(v_{t}\right)=\frac{\varphi^{*}(a)}{a} E\left[d_{t} s_{t}\right]=\frac{\varphi^{*}(a)}{a} E\left[d_{0} s_{0}\right]
$$

Thus, for some $b \in \mathbb{R}$,

$$
E\left[v_{T} c_{T}\right]=b\left(\frac{1}{a}-1\right) E\left[d_{T} s_{T}\right]>\sum_{t=0}^{T} E \sigma_{C_{t}}\left(v_{t}\right)
$$

so that, by Theorem 4.4, the zero premium is not sufficient to hedge $c_{T}$ under the illiquidity costs given by $\varphi$.

Characterizations similar to Corollary 4.6 were given in $[16,7,14,22,31]$ under less restrictive integrability conditions on $c_{T}$. In our case the integrability condition in Theorem 4.4 can be relaxed in the following way. If the process $c-p$ is not $P$-integrable, we can always find a probability measure $\tilde{P} \approx P$ with bounded density such that $c_{t}-p_{t} \in L^{1}\left(\Omega, \mathcal{F}, \tilde{P}, \mathbb{R}^{d}\right)$ for all $t$, e.g.,

$$
\frac{d \tilde{P}}{d P}:=\frac{a}{1+\sum_{t=0}^{T}\left|c_{t}-p_{t}\right|},
$$

where $a$ is a normalizing constant. Then Theorem 4.4 holds with $\tilde{P}$ instead of $P$, and we obtain the following corollary.

Corollary 4.8. Assume that $C$ is a conical market model with the robust no-arbitrage property, and let $c, p \in \mathcal{A}$. Further, let $\tilde{P} \approx P$ be a probability measure with bounded density process $z=\left(z_{t}\right)_{t=0}^{T}$ such that $c_{t}-p_{t} \in L^{1}\left(\Omega, \mathcal{F}_{t}, \tilde{P}, \mathbb{R}^{d}\right)$ for all $t$. Then the following are equivalent:
(i) $p$ is a superhedging premium process for $c$.
(ii) $E \sum_{t=0}^{T}\left(c_{t}-p_{t}\right) \cdot y_{t} \leq 1$ for every bounded nonnegative $P$-martingale $\left(y_{t}\right)_{t=0}^{T}$ such that $E \sum_{t=0}^{T} \sigma_{C_{t}}\left(y_{t}\right) \leq 1$ and such that $y_{t} / z_{t}:=\left(y_{t}^{1} / z_{t}, \ldots, y_{t}^{d} / z_{t}\right)$ is almost surely bounded for all $t$.
Proof. Theorem 4.4 applied with $\tilde{P}$ instead of $P$ yields the equivalence of the following:
(i) $p$ is a superhedging premium process for $c$.
(ii) $E \sum_{t=0}^{T}\left(c_{t}-p_{t}\right) \cdot z_{t} \tilde{y}_{t} \leq 1$ for every nonnegative bounded $\tilde{P}$-martingale $\tilde{y}=\left(\tilde{y}_{t}\right)_{t=0}^{T}$ such that $E \sum_{t=0}^{T} z_{t} \sigma_{C_{t}}\left(\tilde{y}_{t}\right) \leq 1$.

Note further that $z \sigma_{C_{t}}(y)=\sigma_{C_{t}}(z y)$ for all $y \in \mathbb{R}^{d}, z \in \mathbb{R}^{+}$and that $\tilde{y}$ is a bounded $\tilde{P}$ martingale if and only if $z \tilde{y}$ is a bounded $P$-martingale. Thus (ii) holds for all nonnegative bounded $P$-martingales $y$ such that $\frac{1}{z} y$ is almost surely bounded.
5. Fundamental theorem of asset pricing. The fundamental theorem of asset pricing describes the absence of arbitrage by the existence of certain pricing functionals. In the classical frictionless model with a cash account these pricing functionals can be identified with densities of equivalent martingale measures. Under proportional transaction costs the robust no-arbitrage condition is equivalent to the existence of strictly consistent price systems. This result was proved in $[16,14,22,31,15]$ for conical polyhedral market models, i.e., where each $C_{t}(\omega)$ is a cone spanned by a finite number of vectors. Nonpolyhedral conical models are considered in [23] and in [30] under the assumption of efficient friction; i.e., the cones $C_{t}(\omega)$ are assumed to be pointed. In [25, section 3.2] Rokhlin shows, without assuming the polyhedrality of the cones or efficient friction, that the existence of a strictly consistent price system is equivalent to the following condition:

$$
\begin{equation*}
\text { If } \sum_{t=0}^{T} x_{t}=0 \text {, where } x_{t} \in L^{0}\left(C_{t}, \mathcal{F}_{t}\right), \text { then } x_{t} \in L^{0}\left(C_{t}^{0}, \mathcal{F}_{t}\right), \quad t=0, \ldots, T \tag{5.1}
\end{equation*}
$$

However, to show that condition (5.1) is equivalent to the robust no-arbitrage condition one needs [15, Lemma 5, Corollary 1], where the polyhedrality of the cones is used; see [29, Theorem 2].

In this section we derive a version of the fundamental theorem of asset pricing for general conical models. This is achieved through the following lemma, which allows the application of strict separation arguments much as in [31]. It is similar to Proposition A. 5 in [31] but does not rely on the polyhedrality of $C$.

Lemma 5.1. If $K$ is a conical market model that has the robust no-arbitrage property, then it is dominated by another conical market model $\hat{K}$ which still has the robust no-arbitrage property.

Proof. If $K$ has the robust no-arbitrage property, there is an arbitrage-free model $\tilde{K}$ such that $K_{t} \backslash K_{t}^{0} \subset \operatorname{int} \tilde{K}_{t}$, or equivalently, $\tilde{K}_{t}^{*} \backslash\{0\} \subset$ ri $K_{t}^{*}$. Let $e \in \mathbb{R}^{d}$ be the vector with all components equal to one. Since $-e \in \operatorname{int} K_{t}(\omega)$, we have

$$
K_{t}^{*}(\omega)=\bigcup_{\alpha \geq 0} \alpha G_{t}(\omega) \quad \text { and } \quad \tilde{K}_{t}^{*}(\omega)=\bigcup_{\alpha \geq 0} \alpha \tilde{G}_{t}(\omega),
$$

where $G_{t}(\omega)=\left\{v \in \mathbb{R}^{d} \mid v \in K_{t}^{*}(\omega), e \cdot v=1\right\}$ and $\tilde{G}_{t}(\omega)=\left\{v \in \mathbb{R}^{d} \mid v \in \tilde{K}_{t}^{*}(\omega), e \cdot v=1\right\}$. Since $\tilde{K}_{t}^{*} \backslash\{0\} \subset$ ri $K_{t}^{*}$ we have

$$
\begin{equation*}
\tilde{G}_{t}(\omega) \subset \operatorname{ri} G_{t}(\omega) \tag{5.2}
\end{equation*}
$$

Let $\hat{G}_{t}(\omega):=\frac{1}{2} G_{t}(\omega)+\frac{1}{2} \tilde{G}_{t}(\omega)$, and let $\hat{K}_{t}(\omega)$ be the polar of the cone

$$
\hat{K}_{t}^{*}(\omega)=\bigcup_{\alpha \geq 0} \alpha \hat{G}_{t}(\omega)
$$

It suffices to show that $\hat{K}$ is a market model that dominates $K$ and is dominated by $\tilde{K}$. By construction, $\hat{K}_{t}(\omega)$ is closed and convex and, by Proposition 14.11 and Exercise 14.12 of [26], $\hat{K}_{t}$ is $\mathcal{F}_{t}$-measurable, so $\hat{K}$ is indeed a market model. The dominance relations are equivalent to the conditions $\hat{G}_{t}(\omega) \subset \operatorname{ri} G_{t}(\omega)$ and $\tilde{G}_{t}(\omega) \subset \operatorname{ri} \hat{G}_{t}(\omega)$. By (5.2) and [24, Theorem 6.1],

$$
\hat{G}_{t}(\omega) \subset \frac{1}{2} G_{t}(\omega)+\frac{1}{2} \operatorname{ri} G_{t}(\omega)=\operatorname{ri} G_{t}(\omega)
$$

To verify the second condition, we first note that (5.2) implies

$$
\begin{aligned}
\tilde{G}_{t}(\omega) & =\frac{1}{2} \tilde{G}_{t}(\omega)+\frac{1}{2} \tilde{G}_{t}(\omega) \\
& \subset \frac{1}{2} \mathrm{ri} G_{t}(\omega)+\frac{1}{2} \tilde{G}_{t}(\omega),
\end{aligned}
$$

where the last set is relatively open. We then get

$$
\begin{aligned}
\tilde{G}_{t}(\omega) & \subset \operatorname{ri}\left[\frac{1}{2} \operatorname{ri} G_{t}(\omega)+\frac{1}{2} \tilde{G}_{t}(\omega)\right] \\
& =\frac{1}{2} \operatorname{ri} G_{t}(\omega)+\frac{1}{2} \operatorname{ri} \tilde{G}_{t}(\omega) \\
& =\operatorname{ri} \hat{G}_{t}(\omega),
\end{aligned}
$$

where the equalities hold by [24, Corollary 6.6.2].
Equipped with Theorem 3.3 and Lemma 5.1 it is easy to extend the proof of [31, Theorem 1.7] to get the following.

Theorem 5.2. A conical market model $K$ has the robust no-arbitrage property if and only if there exists a strictly consistent price system $y$ for $K$. Moreover, the price system $y$ can be chosen bounded.

Proof. Lemma 5.1 implies the existence of another conical market model $\hat{K}$ such that $\hat{K}_{t}^{*} \backslash\{0\} \subset \operatorname{ri}\left(K_{t}\right)^{*}$. By Theorem 3.3 and Lemma 4.1, the set $A(\hat{K})$ is closed with respect to convergence in measure. Thus $A^{1}(\hat{K})=A(\hat{K}) \cap \mathcal{A}^{1}$ is a convex cone in $\mathcal{A}^{1}$ which is closed in the norm topology. Moreover, the no-arbitrage property implies $A^{1}(\hat{K}) \cap \mathcal{A}_{+}^{1}=\{0\}$, where $\mathcal{A}_{+}^{1}$ denotes the $\mathbb{R}_{+}^{d}$-valued processes in $\mathcal{A}^{1}$.

Since the Banach space $\mathcal{A}^{1}$ is weakly compactly generated as a finite product of $L^{1}$-spaces, the Kreps-Yan theorem holds true on $\mathcal{A}^{1}$; see [28, Theorem 1 and the discussion after it]. Thus there exists a $y \in \mathcal{A}^{\infty}$ such that

$$
\begin{equation*}
E \sum_{t=0}^{T} y_{t} \cdot c_{t} \leq 0 \quad \forall c \in A^{1}(\hat{K}) \tag{5.3}
\end{equation*}
$$

and

$$
\begin{equation*}
E \sum_{t=0}^{T} y_{t} \cdot c_{t}>0 \quad \forall c \in \mathcal{A}_{+}^{1} \backslash\{0\} \tag{5.4}
\end{equation*}
$$

Condition (5.3) can be written as $\sigma_{A^{1}(\hat{K})}(y) \leq 0$, which, by Lemma 4.3, means that $y$ is a nonnegative martingale with $\sigma_{\hat{K}_{t}(\omega)}\left(y_{t}(\omega)\right) \leq 0$ almost surely for every $t$. Since $\hat{K}_{t}(\omega)$ is a cone, we have $y_{t}(\omega) \in \hat{K}_{t}(\omega)^{*}$. Condition (5.4) means that $y$ is componentwise strictly positive, so $y_{t} \in \hat{K}_{t}^{*} \backslash\{0\} \subset$ ri $K_{t}^{*}$ and $y$ is a strictly consistent price system for $K$.

To prove the converse let $y$ be a strictly consistent price system for $K$, and define a conical model

$$
\tilde{K}_{t}:=\left\{x \in \mathbb{R}^{d} \mid y_{t} \cdot x \leq 0\right\}, \quad t=0, \ldots, T
$$

Since $y \in \operatorname{ri} K$ the model $\tilde{K}$ dominates $K$. It suffices to show that $\tilde{K}$ satisfies the no-arbitrage condition. To this end let $c \in A(\tilde{K}) \cap \mathcal{A}_{+}$, and let $x \in \mathcal{A}$ be a self-financing strategy that hedges $c$. Then we have

$$
y_{t} \cdot\left(x_{t}-x_{t-1}+c_{t}\right) \leq 0 \quad \forall t=0, \ldots, T
$$

and hence

$$
0 \leq \sum_{t=0}^{T} y_{t} \cdot c_{t} \leq-\sum_{t=0}^{T} y_{t} \cdot\left(x_{t}-x_{t-1}\right)=\sum_{t=0}^{T} x_{t-1} \cdot\left(y_{t}-y_{t-1}\right) .
$$

Since $y$ is a martingale and $x$ is adapted, the process

$$
M_{s}:=\sum_{t=0}^{s} x_{t-1} \cdot\left(y_{t}-y_{t-1}\right), \quad s=0, \ldots, T
$$

is a local martingale by Theorem 1 in Jacod and Shiryaev [10]. Moreover, $M$ is a martingale by [10, Theorem 2] since $M_{T} \geq 0$ almost surely. Since $M_{0}=0$ we obtain

$$
E\left[\sum_{t=0}^{T} y_{t} \cdot c_{t}\right]=0
$$

and hence $\sum_{t=0}^{T} y_{t} \cdot c_{t}=0$ almost surely. Since $y$ has strictly positive components and $c \in \mathcal{A}_{+}$ this implies $c_{t}=0 P$-almost surely for all $t$. Thus the no-arbitrage condition holds for $\tilde{K}$.

Using Theorem 5.2 we can restate Corollary 4.6 in terms of strictly consistent price systems.

Corollary 5.3. Assume that $C$ is a conical market model with the robust no-arbitrage property. Assume further that $\mathcal{F}_{0}$ is a trivial $\sigma$-algebra. Let $c_{T} \in L^{1}\left(\Omega, \mathcal{F}_{T}, P, \mathbb{R}^{d}\right)$ and $p_{0} \in \mathbb{R}^{d}$. Then the following are equivalent:
(i) $p=\left(p_{0}, 0, \ldots, 0\right)$ is a superhedging premium for $c=\left(0, \ldots, 0, c_{T}\right)$.
(ii) $E\left(c_{T} \cdot y_{T}\right) \leq p_{0} \cdot y_{0}$ for every bounded strictly consistent price system $y=\left(y_{t}\right)_{t=0}^{T}$.

Proof. By Corollary 4.6, it suffices to show that (ii) implies (i). Theorem 5.2 says that there exists a strictly consistent price system $y^{*}$ for $C$. Then for $\varepsilon \in(0,1]$ and for any consistent price system $y$ the process $y^{\varepsilon}:=\varepsilon y^{*}+(1-\varepsilon) y$ defines a strictly consistent price system and for $\varepsilon$ small enough $E\left(c_{T} \cdot y_{T}^{\varepsilon}\right)>p_{0} \cdot y_{0}^{\varepsilon}$ if $E\left(c_{T} \cdot y_{T}\right)>p_{0} \cdot y_{0}$.

Example 5.4. Consider the frictionless model $C$ with a cash account defined by

$$
C_{t}(\omega):=\left\{\left(x^{0}, x^{1}\right) \in \mathbb{R}^{d+1} \mid x^{0}+x^{1} \cdot s_{t}(\omega) \leq 0\right\}, \quad t=0, \ldots, T,
$$

where $\left(s_{t}\right)_{t=0}^{T}$ is a nonnegative $d$-dimensional adapted price process. If $s$ is strictly positive, the robust no scalable arbitrage property coincides with the classical no-arbitrage property. Theorem 5.2 then implies the existence of a strictly consistent price system $y$ for $C$. Much as in Example 4.7, we can identify every strictly consistent price system with a density of an equivalent martingale measure for the price process $s$, and we can write the superhedging condition (ii) of Corollary 5.3 as

$$
E_{P^{*}}\left[c_{T}^{0}+c_{T}^{1} \cdot s_{T}\right] \leq p_{0}^{0}+p_{0}^{1} \cdot s_{0} \quad \forall P^{*} \in \mathcal{P}
$$

where $\mathcal{P}$ is the set of equivalent martingale measures. In particular, if $c_{T}^{1}=0$ and $p^{1}=0$, the condition can be written as

$$
\sup _{P^{*} \in \mathcal{P}} E_{P^{*}}\left[c_{T}^{0}\right] \leq p_{0}^{0}
$$

in accordance with the classical result for claims with cash delivery.
For general convex models, Theorem 5.2 can be stated in the following form.
Theorem 5.5. A convex market model $C$ has the robust no scalable arbitrage property if and only if there exists a strictly positive martingale $y$ such that $y_{t} \in \operatorname{ridom} \sigma_{C_{t}}$ almost surely for all $t$.

Proof. Applying Theorem 5.2 to $C^{\infty}$, we get that $C^{\infty}$ has the robust no-arbitrage property if and only if there exists a strictly consistent price system for $C^{\infty}$. The claim follows by noting that $\operatorname{ri}\left(C_{t}^{\infty}\right)^{*}=\operatorname{ridom} \sigma_{C_{t}}$ by [24, Theorem 6.3], the bipolar theorem, and [24, Corollary 14.2.1].

Corollary 5.6. Let $S$ be a convex cost process on $\mathbb{R}^{d-1} \times \Omega$ in the sense of Example 2.3, such that $S_{t}(\cdot, \omega)$ is nondecreasing on $\mathbb{R}^{d-1}$ for all $t$ and $\omega$, and let $S^{*}$ denote the convex conjugate of $S$. The model

$$
C_{t}(\omega)=\left\{\left(x^{0}, x^{1}\right) \in \mathbb{R}^{d} \mid x^{0}+S_{t}\left(x^{1}, \omega\right) \leq 0\right\}, \quad t=0, \ldots, T,
$$

satisfies the robust no scalable arbitrage condition if and only if there exists a strictly positive adapted process $s$ with $s_{t} \in$ ridom $S_{t}^{*}$ almost surely for all $t$ such that one of the following equivalent conditions holds:
(i) the frictionless model

$$
\tilde{C}_{t}(\omega):=\left\{\left(x^{0}, x^{1}\right) \in \mathbb{R}^{d} \mid x^{0}+s_{t}(\omega) \cdot x^{1} \leq 0\right\}, \quad t=0, \ldots, T,
$$

is arbitrage-free;
(ii) there exists a probability measure $P^{*}$ equivalent to $P$ such that $s$ is a martingale un$\operatorname{der} P^{*}$.

## Proof. We get

$$
\begin{aligned}
\sigma_{C_{t}(\omega)}(z) & =\sup _{x^{0}, x^{1}}\left\{z^{0} x^{0}+z^{1} \cdot x^{1} \mid x^{0}+S_{t}\left(x^{1}, \omega\right) \leq 0\right\} \\
& = \begin{cases}\sup _{x^{1}}\left\{z^{1} \cdot x^{1}-z^{0} S_{t}\left(x^{1}, \omega\right)\right\} & \text { if } z^{0} \geq 0, \\
+\infty & \text { otherwise }\end{cases} \\
& = \begin{cases}z^{0} S_{t}^{*}\left(z^{1} / z^{0}, \omega\right) & \text { if } z^{0}>0, \\
0 & \text { if } z^{0}=0 \text { and } z^{1}=0, \\
+\infty & \text { otherwise },\end{cases}
\end{aligned}
$$

and thus ridom $\sigma_{C_{t}(\omega)}=\left\{\left(z^{0}, z^{1}\right) \mid z^{0}>0, z^{1} / z^{0} \in \operatorname{ridom} S_{t}^{*}(\omega)\right\}$. By Theorem 5.5, the model $C$ has the robust no scalable arbitrage property if and only if there exists a strictly positive martingale $\left(y^{0}, y^{1}\right)$ such that $y_{t}^{1} / y_{t}^{0} \in \operatorname{ridom} S_{t}^{*}$ almost surely for all $t$. The latter condition means that there is a strictly positive adapted process $s$ with $s_{t} \in$ ridom $S_{t}^{*}$ such that $y^{1}=y^{0} s$. In particular, $s$ is a martingale under the equivalent martingale measure defined by $d P^{*} / d P:=y_{T}^{0} / y_{0}^{0}$. Equivalence of (i) and (ii) follows from Theorem 5.2 as in Example 5.4.

Remark 5.7. By Theorem 3.3 and Lemma 4.1, the no-arbitrage criterion in the above corollary guarantees that the set $A(C) \subset \mathcal{A}$ is closed in probability. This implies, in particular, the closedness of the set

$$
\begin{aligned}
A^{0}(C) & :=\left\{c^{0} \in \mathcal{M} \mid\left(c^{0}, 0\right) \in A(C)\right\} \\
& =\left\{c^{0} \in \mathcal{M} \mid \exists x \in \mathcal{A}: \Delta x_{t}^{0}+c_{t}^{0}+S_{t}\left(\Delta x_{t}^{1}\right) \leq 0 \forall t, x_{T}=0\right\}
\end{aligned}
$$

where $\mathcal{M}$ denotes the space of adapted real-valued processes. The set $A^{0}(C)$ consists of the claim processes with cash delivery that can be superhedged with zero cost. An alternative condition for the closedness of $A^{0}(C)$ was given in [21, section 6]. The condition from [21] does not require the cost process $S$ to be nondecreasing, and it applies under portfolio constraints of the form $\left(x_{t}^{0}, x_{t}^{1}\right) \in\{0\} \times D_{t}(\omega)$, where $D_{t}$ is an $\mathcal{F}_{t}$-measurable set-valued mapping with values in closed convex subsets of $\mathbb{R}^{d-1}$.

The equivalence of conditions (i) and (ii) in Corollary 5.6 is reminiscent of [20, Theorem 5.4] and, in the case of proportional transaction costs, also of Jouini and Kallal [11, Theorem 3.2.i]. If $S^{\infty}$ is finite-valued, then, by [20, Theorem 5.4], the condition

$$
\operatorname{cl} A^{0}\left(C^{\infty}\right) \cap \mathcal{M}_{+}=\{0\}
$$

where cl denotes the closure with respect to convergence in measure and $\mathcal{M}_{+}$denotes the nonnegative elements of $\mathcal{M}$, implies the existence of an adapted process $s$ and a measure $P^{*}$ equivalent to $P$ such that $s$ is a martingale under $P^{*}$ and $s_{t} \in \operatorname{cldom} S_{t}^{*}$ almost surely for all $t$.
6. Proof of Theorem 3.3. The proof of Theorem 3.3 requires some preparation. First we will use projection techniques similar to those in [31] to extract self-financing portfolio processes ending with 0 in the model $C^{\infty}$. For a convex market model $C$ we consider the set

$$
\mathcal{N}(C):=\left\{x \in \mathcal{A} \mid \Delta x_{t} \in C_{t} P \text {-a.s. } \forall t=0, \ldots, T, x_{T}=0\right\} .
$$

The next lemma is a version of Lemma 5 in [15]; see also Lemma 2.6 in [31].
Lemma 6.1. If $C$ is a convex market model that has the robust no scalable arbitrage property, then $\mathcal{N}\left(C^{\infty}\right)=\mathcal{N}\left(C^{0}\right)$.

Proof. We have $\mathcal{N}\left(C^{\infty}\right) \supset \mathcal{N}\left(C^{0}\right)$ simply because $C_{t}^{\infty} \supset C_{t}^{0}$ almost surely for every $t$, so it suffices to prove the reverse. Let $x \in \mathcal{N}\left(C^{\infty}\right)$, and assume that $\Delta x_{t} \in C_{t}^{\infty} \backslash C_{t}^{0}$ on some set with positive probability for some $t=0, \ldots, T$. This contradicts the robust no scalable arbitrage property. Indeed, if $\tilde{C}$ is a market model such that $C_{t}^{\infty} \backslash C_{t}^{0} \subset \operatorname{int} \tilde{C}_{t}$, we have $\Delta x_{t} \in \operatorname{int} \tilde{C}_{t}$ on a nontrivial set, and then for any $e \in \mathbb{R}_{+}^{d} \backslash\{0\}$

$$
\varepsilon_{t}(\omega)=\sup \left\{\varepsilon \in \mathbb{R} \mid \Delta x_{t}(\omega)+\varepsilon e \in \tilde{C}_{t}(\omega)\right\}
$$

defines an $\mathcal{F}_{t}$-measurable $\mathbb{R}_{+}^{d}$-valued random variable (see, e.g., [26, Theorem 14.37]) which does not vanish almost surely. By Lemma 4.1, this would mean that $\tilde{C}$ violates the no-arbitrage condition.

For each $t \in\{0, \ldots, T\}$ we denote by $\mathcal{N}_{t}$ the set of all $\mathcal{F}_{t}$-measurable random vectors that may be extended to a portfolio in $\mathcal{N}\left(C^{0}\right)$, i.e.,

$$
\mathcal{N}_{t}:=\left\{x_{t} \in L^{0}\left(\mathbb{R}^{d}, \mathcal{F}_{t}\right) \mid \exists x_{t+1}, \ldots, x_{T} \text { s.t. }\left(0, \ldots, 0, x_{t}, \ldots, x_{T}\right) \in \mathcal{N}\left(C^{0}\right)\right\} .
$$

Lemma 6.2. Let $C$ be a convex market model. Then for each $t \in\{0, \ldots, T\}$ there is an $\mathcal{F}_{t}$-measurable map $N_{t}: \Omega \rightrightarrows \mathbb{R}^{d}$ whose values are linear subspaces of $\mathbb{R}^{d}$ and $\mathcal{N}_{t}=L^{0}\left(N_{t}, \mathcal{F}_{t}\right)$.

Proof. We consider first the larger set

$$
\begin{array}{r}
\mathcal{M}_{t}=\left\{x_{t} \in L^{0}\left(\mathbb{R}^{d}, \mathcal{F}_{t}\right) \mid \exists x_{t+1}, \ldots, x_{T} \text { s.t. } \Delta x_{s} \in L^{0}\left(C_{s}^{0}, \mathcal{F}_{s}\right)\right. \\
\left.\forall s=t+1, \ldots, T \text { and } x_{T}=0\right\}
\end{array}
$$

and show that there is an $\mathcal{F}_{t}$-measurable map $M_{t}: \Omega \rightrightarrows \mathbb{R}^{d}$ whose values are linear subspaces of $\mathbb{R}^{d}$ and $\mathcal{M}_{t}=L^{0}\left(M_{t}, \mathcal{F}_{t}\right)$. Then the $\mathcal{F}_{t}$-measurable map $N_{t}: \Omega \rightrightarrows \mathbb{R}^{d}$ with $N_{t}(\omega):=$ $M_{t}(\omega) \cap C_{t}^{0}(\omega)$ has the desired properties.

In order to obtain the maps $M_{t}$ we argue by induction on $T$. For $T=0$ we have $\mathcal{M}_{T}=$ $M_{T}=\{0\}$. Now assume that the claim holds for any $T$-step model, and consider the ( $T+1$ )step model $\left(C_{t}\right)_{t=0}^{T}$. By the induction hypothesis there exist $\mathcal{F}_{s}$-measurable maps $M_{s}$ whose values are linear subspaces of $\mathbb{R}^{d}$ such that $\mathcal{M}_{s}=L^{0}\left(M_{s}, \mathcal{F}_{s}\right)$ for each $s \in\{1, \ldots, T\}$. Note that $x_{0} \in \mathcal{M}_{0}$ if and only if $x_{0} \in M_{1}+C_{1}^{0}$ almost surely. Indeed, $x_{0} \in \mathcal{M}_{0}$ if and only if there exists an $x_{1} \in \mathcal{M}_{1}$ such that $x_{1}-x_{0} \in C_{1}^{0}$ almost surely. The mapping $L(\omega):=$ $M_{1}(\omega)+C_{1}^{0}(\omega)$ is $\mathcal{F}_{1}$-measurable (see [26, Proposition 14.11(c)]), and its values are linear subspaces of $\mathbb{R}^{d}$ and, in particular, closed. By the theorem on page 135 of [32], there exists the largest closed $\mathcal{F}_{0}$-measurable set-valued map $M_{0}$ such that $M_{0} \subseteq L$ almost surely and $L^{0}\left(M_{0}, \mathcal{F}_{0}\right)=L^{0}\left(L, \mathcal{F}_{0}\right)=\mathcal{M}_{0}$. Clearly $M_{0}(\omega)$ is a linear subspace of $\mathbb{R}^{d}$ for each $\omega$.

For each $\omega$ we denote by $N_{t}^{\perp}(\omega)$ the orthogonal complement of $N_{t}(\omega)$ in $\mathbb{R}^{d}$. Then $N_{t}^{\perp}$ : $\Omega \rightrightarrows \mathbb{R}^{d}$ is an $\mathcal{F}_{t}$-measurable (see [26, Exercise 14.12(e)]) map whose values are linear subspaces of $\mathbb{R}^{d}$.

Lemma 6.3. Let $C$ be a convex market model, and let $c \in A(C)$. Then there exists a process $x^{\perp} \in \mathcal{A}$ such that $x_{t}^{\perp} \in N_{t}^{\perp}, \Delta x_{t}^{\perp}+c_{t} \in C_{t}$ for all $t=0, \ldots, T$ and $x_{T}^{\perp}=0$.

Proof. We will prove the existence of the process $x^{\perp}$ by induction on $T$. For $T=0$ we have $x_{T}^{\perp}:=x_{T}=0$. Assume that the claim holds for any $T$-step model, and consider the $(T+1)$-step model $\left(C_{t}\right)_{t=0}^{T}$. Let $c \in A(C)$, and let $x \in \mathcal{A}$ be such that $\Delta x_{t}+c_{t} \in C_{t}$ for all $t=0, \ldots, T$ and $x_{T}=0$. We denote by $x_{0}^{0}$ the $\mathcal{F}_{0}$-measurable projection of $x_{0}$ on $N_{0}$ and by $\left(x_{1}^{0}, \ldots, x_{T-1}^{0}, 0\right)$ an extension of $x_{0}^{0}$ to a self-financing portfolio process in the model $C^{0}$ (the existence of such an extension is given by Lemma 6.2). Then

$$
y_{t}:=x_{t}-x_{t}^{0}, \quad t=0, \ldots, T-1, \quad y_{T}:=0
$$

defines an adapted process $y$ with $y_{0}=x_{0}-x_{0}^{0} \in N_{0}^{\perp}$ almost surely. Moreover, $y$ hedges $c$. Indeed, since $\Delta x_{t}^{0} \in C_{t}^{0}$ we have

$$
\Delta y_{t}+c_{t}=\Delta x_{t}+c_{t}-\Delta x_{t}^{0} \in C_{t}-C_{t}^{0}=C_{t} \quad P \text {-a.s. }
$$

for all $t=0, \ldots, T$.
The process $\left(y_{1}, \ldots, y_{T-1}, 0\right)$ hedges the claim $\left(c_{1}-y_{0}, c_{2}, \ldots, c_{T}\right)$ in the $T$-step model $\left(C_{t}\right)_{t=1}^{T}$. Thus by the induction hypothesis there is a process $y^{\perp}=\left(y_{1}^{\perp}, \ldots, y_{T-1}^{\perp}, 0\right)$ such that $y_{t}^{\perp} \in N_{t}^{\perp}$ almost surely and $y^{\perp}$ hedges $\left(c_{1}-y_{0}, c_{2}, \ldots, c_{T}\right)$. Then the process $x^{\perp}:=$ $\left(y_{0}, y_{1}^{\perp}, \ldots, y_{T-1}^{\perp}, 0\right)$ has the required properties.

The following lemma was used first in Kabanov and Stricker [17]; it is also documented, e.g., in $[31,15,9,8]$. We refer to these works for the proof.

Lemma 6.4. Let $\left(x^{n}\right)_{n=1}^{\infty}$ be a sequence of random vectors in $L^{0}\left(\mathbb{R}^{d}, \mathcal{F}\right)$ such that $\lim _{\inf }^{n}\left|x^{n}\right|$ $<\infty P$-almost surely. Then there exists an $\mathcal{F}$-measurable increasing $\mathbb{N}$-valued random sequence $\left(\tau^{n}\right)_{n=1}^{\infty}$ such that $\left(x^{\tau^{n}}\right)_{n=1}^{\infty}$ converges almost surely to some $x \in L^{0}\left(\mathbb{R}^{d}, \mathcal{F}\right)$.

Proof of Theorem 3.3. By Lemma 4.1, the closedness of $A_{T}(C)$ is equivalent to the closedness of $A(C)$. By Lemma 6.3, we may assume that $x_{t} \in N_{t}^{\perp}$ for all $t=0, \ldots, T-1$ in the definition of $A(C)$.

In order to prove the closedness of $A(C)$ we will prove the following more precise statement: Let $\left(c^{n}\right)_{n=1}^{\infty}$ be a sequence of claim processes in $A(C)$ such that $c_{t}^{n} \rightarrow c_{t}$ for all $t$ in measure, and let $\left(x^{n}\right)_{n=1}^{\infty}$ be any sequence in $\mathcal{A}$ such that $\Delta x_{t}^{n}+c_{t}^{n} \in C_{t}, x_{t}^{n} \in N_{t}^{\perp}$ for all $t$ and $x_{T}^{n}=0$. Then there exists an $\mathcal{F}$-measurable random subsequence $\left(\tau^{n}\right)_{n=1}^{\infty}$ of $\mathbb{N}$ and an $x \in \mathcal{A}$ such that $x_{t}^{\tau^{n}} \rightarrow x_{t}$ almost surely and $\Delta x_{t}+c_{t} \in C_{t}, x_{t} \in N_{t}^{\perp}$ for all $t$ and $x_{T}=0$. In particular, $c \in A(C)$.

The proof will follow by induction on $T$. For $T=0$ the statement is obvious since the set $L^{0}\left(C_{0}, \mathcal{F}_{0}\right)$ is closed in $L^{0}\left(\mathbb{R}^{d}, \mathcal{F}_{0}\right)$ and any hedging strategy is identically 0 . Now assume that the statement holds for any $T$-step model, and consider a $(T+1)$-step model $C=\left(C_{t}\right)_{t=0}^{T}$. Let $c^{n} \in A(C), n=0,1, \ldots$, be such that $c_{t}^{n} \rightarrow c_{t} \in \mathcal{A}$ for all $t$ in measure. By passing to a subsequence if necessary, we may assume that $c_{t}^{n} \rightarrow c_{t}$ for all $t$ almost surely. Let $x^{n} \in \mathcal{A}$, $n=0,1, \ldots$, be such that $\Delta x_{t}^{n}+c_{t}^{n} \in C_{t}, x_{t}^{n} \in N_{t}^{\perp}$ for all $t$ and $x_{T}^{n}=0$.

Case 1: the sequence $\left(x_{0}^{n}\right)_{n=1}^{\infty}$ is almost surely bounded. In this case, we can apply Lemma 6.4 to find an $\mathcal{F}_{0}$-measurable random sequence $\left(\sigma^{n}\right)_{n=1}^{\infty}$ such that $x_{0}^{\sigma^{n}}$ converges to some $x_{0} \in L^{0}\left(\mathbb{R}^{d}, \mathcal{F}_{0}\right)$ almost surely. Then $x_{0} \in L^{0}\left(N_{0}^{\perp}, \mathcal{F}_{0}\right)$ since each $N_{0}^{\perp}(\omega)$ is a closed subspace of $\mathbb{R}^{d}$. Moreover, the sequence of claim processes

$$
\tilde{c}^{n}:=\left(c_{1}^{\sigma^{n}}-x_{0}^{\sigma^{n}}, c_{2}^{\sigma^{n}}, \ldots, c_{T}^{\sigma^{n}}\right), \quad n \in \mathbb{N}
$$

belongs to $A\left(\left(C_{t}\right)_{t=1}^{T}\right)$ with the hedging sequence $\tilde{x}^{n}:=\left(x_{1}^{\sigma^{n}}, \ldots, x_{T-1}^{\sigma^{n}}, 0\right)$. Since $\tilde{c}^{n} \rightarrow$ $\left(c_{1}-x_{0}, c_{2}, \ldots, c_{T}\right)$ almost surely, we apply the induction hypothesis to the $T$-step model $\left(C_{t}\right)_{t=1}^{T}$ and obtain an $\mathcal{F}$-measurable random subsequence $\left(\tau^{n}\right)_{n=0}^{\infty}$ of $\left(\sigma^{n}\right)_{n=0}^{\infty}$ and an $\left(\mathcal{F}_{t}\right)_{t=1^{-}}^{T}$ adapted process $\tilde{x}=\left(x_{1}, \ldots, x_{T}\right)$ such that $\tilde{x}_{t}^{\tau^{n}} \rightarrow \tilde{x}_{t}$ for all $t$ almost surely, $\tilde{x}$ hedges $\tilde{c}$, and $x_{t} \in N_{t}^{\perp}$ for $t=1, \ldots, T$. This proves the claim, since $x_{t}^{\tau^{n}} \rightarrow x_{t}$ almost surely for all $t=0, \ldots, T$ and the process $x:=\left(x_{0}, \ldots, x_{T}\right)$ has the desired properties. Indeed, $x$ is adapted, $x_{t} \in N_{t}^{\perp}$ for all $t, x_{T}=0$, and we have

$$
\Delta x_{t}+c_{t} \in C_{t} \quad P \text {-a.s. } \quad \forall t=0, \ldots, T
$$

since $C_{t}(\omega)$ is a closed subset of $\mathbb{R}^{d}$ for each $\omega$.
Case 2: the sequence $\left(x_{0}^{n}\right)_{n=1}^{\infty}$ is not almost surely bounded. We will show that this leads to a contradiction with the robust no scalable arbitrage property. Indeed, assume that the set

$$
A:=\left\{\omega \in \Omega|\liminf | x_{0}^{n}(\omega) \mid=\infty\right\}
$$

has positive probability, and consider the sequences

$$
\hat{x}^{n}:=\alpha^{n} x^{n} \quad \text { and } \quad \hat{c}^{n}:=\alpha^{n} c^{n}, \quad n \in \mathbb{N}
$$

where $\alpha^{n}=\frac{\chi_{A}}{\max \left\{\left|x_{0}^{n}\right|, 1\right\}}$. The processes $\left(\hat{x}^{n}\right)_{n=1}^{\infty}$ and $\left(\hat{c}^{n}\right)_{n=1}^{\infty}$ are adapted, $\hat{x}_{t}^{n} \in L^{0}\left(N_{t}^{\perp}, \mathcal{F}_{t}\right)$ for all $t, \hat{x}_{T}^{n}=0, \hat{c}_{t}^{n} \rightarrow 0$ almost surely for each $t$, and the sequence $\left(\hat{x}_{0}^{n}\right)_{n=0}^{\infty}$ is almost surely bounded. Moreover, we have

$$
\begin{equation*}
\hat{x}_{t}^{n}-\hat{x}_{t-1}^{n}+\hat{c}_{t}^{n} \in \alpha^{n} C_{t} \quad P \text {-a.s. } \tag{6.1}
\end{equation*}
$$

where $\alpha^{n} C_{t} \subset C_{t}$ since $C_{t}$ is convex, $0 \in C_{t}$, and $\left|\alpha^{n}\right| \leq 1$. Thus we have the same situation as in Case 1, and using the same reasoning we obtain an $\mathcal{F}$-measurable random sequence $\left(\tau^{n}\right)_{n=0}^{\infty}$ and an adapted process $x$ such that $\hat{x}^{\tau^{n}} \rightarrow x$ almost surely, $x_{t} \in N_{t}^{\perp}$ for all $t$, and $x_{T}=0$. Moreover, since $\alpha^{n} \rightarrow 0$ almost surely, (6.1) and (3.2) imply

$$
\Delta x_{t} \in C_{t}^{\infty} \quad P \text {-a.s. } \quad \forall t=0, \ldots, T
$$

Thus $x \in \mathcal{N}\left(C^{0}\right)$ by Lemma 6.1, and then $x_{0} \in N_{0}$ almost surely by Lemma 6.2. Since also $x_{0} \in N_{0}^{\perp}$, we have $x_{0}=0$ almost surely. But we also have $\left|\hat{x}_{0}^{n}\right| \rightarrow 1$ and hence $\left|x_{0}\right|=1$ on the nontrivial set $A$, which is a contradiction. Thus, Case 2 cannot occur under the robust no scalable arbitrage condition.

Acknowledgment. We thank the anonymous referees for useful comments that helped us to improve this paper.

## REFERENCES

[1] C. D. Aliprantis and K. C. Border, Infinite-Dimensional Analysis. A Hitchhiker's Guide, 2nd ed., Springer-Verlag, Berlin, 1999.
[2] F. Astic and N. Touzi, No arbitrage conditions and liquidity, J. Math. Econom., 43 (2007), pp. 692-708.
[3] U. Çetin, R. A. Jarrow, and P. Protter, Liquidity risk and arbitrage pricing theory, Finance Stoch., 8 (2004), pp. 311-341.
[4] U. Çetin and L. C. G. Rogers, Modelling liquidity effects in discrete time, Math. Finance, 17 (2007), pp. 15-29.
[5] U. Çetin, M. H. Soner, and N. Touzi, Option hedging for small investors under liquidity costs, Finance Stoch., to appear.
[6] J. Cvitanić and I. Karatzas, Hedging and portfolio optimization under transaction costs: A martingale approach, Math. Finance, 6 (1996), pp. 133-165.
[7] F. Delbaen, Yu. M. Kabanov, and E. Valkeila, Hedging under transaction costs in currency markets: A discrete-time model, Math. Finance, 12 (2002), pp. 45-61.
[8] F. Delbaen and W. Schachermayer, The Mathematics of Arbitrage, Springer Finance, SpringerVerlag, Berlin, Heidelberg, 2006.
[9] H. Föllmer and A. Schied, Stochastic Finance. An Introduction in Discrete Time, de Gruyter Stud. Math. 27, Walter de Gruyter, Berlin, 2004.
[10] J. Jacod and A. N. Shiryaev, Local martingales and the fundamental asset pricing theorems in the discrete-time case, Finance Stoch., 2 (1998), pp. 259-273.
[11] E. Jouini and H. Kallal, Martingales and arbitrage in securities markets with transaction costs, J. Econom. Theory, 66 (1995), pp. 178-197.
[12] Yu. M. Kabanov, Hedging and liquidation under transaction costs in currency markets, Finance Stoch., 3 (1999), pp. 237-248.
[13] Yu. M. Kabanov, Arbitrage theory, in Option Pricing, Interest Rates and Risk Management, Handb. Math. Finance, Cambridge University Press, Cambridge, UK, 2001, pp. 3-42.
[14] Yu. M. Kabanov, M. Rásonyi, and Ch. Stricker, No-arbitrage criteria for financial markets with efficient friction, Finance Stoch., 6 (2002), pp. 371-382.
[15] Yu. M. Kabanov, M. Rásonyi, and Ch. Stricker, On the closedness of sums of convex cones in $L^{0}$ and the robust no-arbitrage property, Finance Stoch., 7 (2003), pp. 403-411.
[16] Yu. M. Kabanov and Ch. Stricker, The Harrison-Pliska arbitrage pricing theorem under transaction costs, J. Math. Econom., 35 (2001), pp. 185-196.
[17] Yu. M. Kabanov and Ch. Stricker, A teachers' note on no-arbitrage criteria, in Séminaire de Probabilités, XXXV, Lecture Notes in Math. 1755, Springer-Verlag, Berlin, 2001, pp. 149-152.
[18] K. Kaval and I. Molchanov, Link-save trading, J. Math. Econom., 42 (2006), pp. 710-728.
[19] T. Pennanen, Nonlinear Price Processes, preprint, 2006.
[20] T. Pennanen, Arbitrage and deflators in illiquid markets, Finance Stoch., to appear.
[21] T. Pennanen, Superhedging in illiquid markets, Math. Finance, to appear.
[22] I. Penner, Arbitragefreiheit in Finanzmärkten mit Transaktionskosten, master's thesis, HumboldtUniversität zu Berlin, Berlin, Germany, 2001.
[23] M. RÁsonyi, New methods in the arbitrage theory of financial markets with transaction costs, in Séminaire de Probabilités XLI, Springer-Verlag, Berlin, 2008, pp. 455-462.
[24] R. T. Rockafellar, Convex Analysis, Princeton Math. Ser. 28, Princeton University Press, Princeton, NJ, 1970.
[25] R. T. Rockafellar, Coherent approaches to risk in optimization under uncertainty, in Tutorials in Operations Research, INFORMS, Hanover, MD, 2007, pp. 38-61.
[26] R. T. Rockafellar and R. J.-B. Wets, Variational Analysis, Grundlehren Math. Wiss. 317, SpringerVerlag, Berlin, 1998.
[27] L. C. G. Rogers and S. Singh, The Costs of Illiquidity and Its Effects on Hedging, preprint, 2007.
[28] D. B. Rokhlin, The Kreps-Yan theorem for $L^{\infty}$, Int. J. Math. Math. Sci., no. 17 (2005), pp. 2749-2756.
[29] D. B. Rokhlin, A constructive no-arbitrage criterion with transaction costs in the case of finite discrete time, Teor. Veroyatn. Primen., 52 (2007), pp. 41-59 (in Russian); Theory Probab. Appl., 52 (2008), pp. 93-107 (in English).
[30] D. B. Rokhlin, Martingale selection problem and asset pricing in finite discrete time, Electron. Comm. Probab., 12 (2007), pp. 1-8.
[31] W. Schachermayer, The fundamental theorem of asset pricing under proportional transaction costs in finite discrete time, Math. Finance, 14 (2004), pp. 19-48.
[32] A. Truffert, Conditional expectation of integrands and random sets, Ann. Oper. Res., 30 (1991), pp. 117-156.

# Weak Kyle-Back Equilibrium Models for Max and ArgMax* 

A. Kohatsu-Higa ${ }^{\dagger}$ and S. Ortiz-Latorre ${ }^{\ddagger}$

Abstract. The goal of this article is to introduce a new approach to model equilibrium in financial markets with an insider. We prove the existence and uniqueness in law of equilibrium for these markets. Our setting is weaker than that of Back, and it can be interpreted as a first theoretical step towards developing statistical test procedures. Additionally, it allows various forms of insider information to be considered under the same framework and compared. As major examples, we consider the cases of the maximum of the demand and the time at which this maximum is taken, which have not previously been treated in the literature of equilibrium in financial markets with inside information. Simulations indicate that the expected wealth for the maximum is greater than the expected wealth for its argument.

Key words. large insider trading, equilibrium theory, semimartingale decomposition
AMS subject classifications. 49J40, 60G48, 93E20
DOI. 10.1137/080739768

1. Introduction. In recent years, the study of mathematical models for financial markets with asymmetry of information has been gaining increasing attention from mathematical finance researchers. In a seminal paper and from the market microstructure point of view, Kyle [16] introduced a model in which an insider, who knows the value of the stock at some future time, optimizes his/her wealth while the market maker makes prices rational, that is, a rational expectations equilibrium model. The main features of Kyle's model are that it gives finite utilities and that it is a model of price formation. That is, the insider controls the price process through his/her demand of stock shares. Kyle's model has been extended by Back [2], Lasserre [17], Cho [5], and Campi and Çetin [4], among others.

We consider a continuous time market composed of one risk-free asset and one risky asset. We assume, without loss of generality, that the risk-free rate is zero. Trading in the risky asset is continuous in time and quantity. Furthermore, the market is order-driven; that is, prices are determined by the demand on the risky asset.

There is to be a public release of information at time $t=1$. This information reveals the value of the risky asset, which we denote by $\xi$. As the market is order-driven, this entails that $\xi$ will be the price at which the asset will be traded just after the release of information and, therefore, the final profit obtained through trading on this asset will depend on $\xi$.

There are three representative agents in the market: the market maker, the insider, and

[^16]the noise trader. The role of the market maker is to organize the market. That is, according to the asset's aggregate demand, the market maker sets the price of the asset and clears the market. The insider is assumed to know at the beginning of the trading period some strong information, say $\lambda=L(Y)$, not necessarily equal to $\xi$, which depends exclusively on the total demand $Y$. This agent uses this information in order to maximize his/her expected profit. The noise trader represents all the other participants in the market. The noise trader's orders are a consequence of liquidity or hedging issues and are assumed to be independent of $\lambda$ but not necessarily of $\xi$. Thanks to the demand of the noise trader, denoted by $Z$, the market maker cannot observe the demand of the insider.

Our formulation is weak in the sense that the vector $(\xi, \lambda, Z)$ is not given beforehand, in contrast with the previous literature on this subject. The initial data in our formulation is $(\mu, L)$, where $\mu$ is the law of $\xi$ and the other ingredients of an equilibrium are part of the problem. The mathematical motivation for using a weak setup is due to the fact that in a strong formulation the relationship between $\lambda, \xi$, and $Z$ cannot be simply stated in general. This relationship is not unique if one wants only to give as initial data the law of the final price. Furthermore, in general, $\xi$ is not independent of $\lambda$ or $Z$. However, it is assumed that $\xi$ is made public at the end of the trading period. Hence, $\xi$ is incorporated in the functional to be optimized in the equilibrium.

From the economic modeling point of view, the situation can be explained as follows. Suppose a financial controller exists, say a member of an exchange commission, which would like to test after the time interval has been totally observed (say $[0,1]$ ) the large trader/insider behavior in a sector of the market. By a sector we mean a collection of homogeneous companies sharing a similar activity for which one can assign a law for $\xi$, its value at $t=1$. The financial controller observes the data for different companies in the sector, and after some renormalization we can regard the data as different realizations or sample points in his/her universe.

With the data, a law $\mu$ for $\xi$ can be inferred and a functional $L$ of the total demand is fixed for testing. The first step for the controller is to know if it is possible that there exists insider trading in the stocks of this sector using the information $L(Y)$ and being in equilibrium. Our paper addresses this question. The next step would be to design a statistical test according to the probabilistic properties of the equilibrium, but we do not pursue this goal in this paper.

Now, we briefly discuss the concept of weak equilibrium used in this paper.
The difference between the classical notion of equilibrium used in Back [2, section 1] and the one proposed here is that in Back the information is exogenously given while here is also part of the definition of weak equilibrium.

In particular, condition (vii) in Definition 2.7 states that if we fix the noise trader process and the information, the strategy used by the insider maximizes his/her expected final wealth within a suitable admissible space. This condition can be also interpreted as a local equilibrium condition because the noise trader process and the information are fixed.

This interpretation is linked to the notion of partial (or local) equilibrium. If the insider finds himself/herself at such a partial equilibrium point, there is no particular reason to move from such a point.

From the point of view of a financial controller, the procedure is carried out after all the data is available. That is, the final price has already been announced and the controller wants
to test the existence of some insiders in the market. Once the type of information is selected, one can statistically check whether the strategy used by the insider(s) is locally optimal.

It is important to point out that this optimal strategy has the same functional form as the compensator of the Brownian motion $W$ with respect to its natural filtration enlarged with the random variable $L(W)$.

Besides the weak equilibrium feature, there are various delicate mathematical points where our results and techniques differ from previously mentioned research. Briefly summarizing, we mention the following:
(1) Due to the generality of the functional $L$, we use variational calculus (or dynamic principle) and do not obtain an HJB equation formulation. In particular, optimal strategies do not depend only on the insider's additional information and the value process, the admissible strategies do not form a linear space, and the expected profit depends on $\xi$, which is not measurable with respect to the insider's filtration. These features introduce some difficulties in obtaining the optimality results.
(2) One of the conditions of admissibility requires that the optimal strategy has to be adapted to the filtration generated by the noise trader process and the insider information. This result, which was easy to obtain in previous articles (in fact, this was just a property of Brownian bridges), becomes extremely difficult in the generality presented here. In fact, we consider as examples the case where $L(Y)=\max _{t} Y_{t}$ corresponds to the maximum of the demand and to the argument of this maximum $L(Y)=\arg \max _{t} Y_{t}$. This leads to the study of the existence and uniqueness of the solutions of stochastic differential equations (SDEs) with path dependent coefficients which degenerate at random times.
Finally, we compare the expected wealth obtained by the large insider/trader in the two main examples considered. The simulations indicate that knowing $\tau$, the time at which the maximum of the total demand is achieved, gives less expected wealth than knowing $M$, the maximum of the total demand.

As a final remark, we want to state that one of the main goals of this article is to raise/ contribute to the discussion on the issue of the equilibrium concept for the large insider/ trader for general information as explained in this article. We do not pretend that this is the unique way to solve the problem. We hope that other researchers will also present alternative proposals and comments on this model.

The paper is organized as follows. In section 2 we give some basic definitions and introduce our weak formulation of equilibrium. Section 3 contains the discussion of the optimization problem for the insider, and an optimality equation is deduced. In section 4 we relate the properties of the solutions of the optimality equation with the rationality of prices. Section 5 is devoted to stating the main results on the existence and uniqueness in law of a weak equilibrium. In section 6 we deal with two basic examples, previously treated in the literature: Back's example and an example on binary information. Section 7 aims to introduce two new examples in the literature of equilibrium for asymmetric markets. The first is in the case that the additional information held by the insider is the maximum of the total demand, and the second is in the case of the time at which this maximum is attained. We state and prove the existence and uniqueness in law of a weak equilibrium in both cases. Finally, we numerically compare the expected wealth obtained by the insider in these two examples. Section 8 is
dedicated to conclusions. Finally, section 9 contains an appendix devoted to proving some technical results.

Throughout this article $C$ will denote a constant that may change from line to line. $\mathcal{L}(\mathcal{X})$ denotes the law of the random element $\mathcal{X}$.
2. Weak formulation of equilibrium. In this section we introduce the concept of weak equilibrium. First we define the class of pricing rules and admissible strategies.

Definition 2.1. We say that a function $F:[0,1] \times \mathbb{R} \rightarrow \mathbb{R}$ satisfies an exponential growth condition if there exist positive constants $A, B$ such that $|F(t, y)| \leq A e^{B|y|}$ for all $(t, y) \in$ $[0,1] \times \mathbb{R}$.

Definition 2.2. A pricing rule is a function $H \in \mathcal{C}^{1,3}((0,1) \times \mathbb{R})$ such that $H_{y}(t, y)>0$ for all $t \in[0,1]$ and $H, H_{t}, H_{y}, H_{y y}$ satisfy an exponential growth condition. We denote by $\mathcal{H}$ the set of functions $H$ satisfying these properties.

In the previous definition we require the pricing rules to satisfy some regularity and growth conditions for technical reasons. From a modeling point of view, the important assumption is the requirement that $H_{y}(t, y)>0$ for all $t \in[0,1]$. This implies that the insider can invert the price process to obtain the total demand and, hence, the noise trader demand; see Remark 2.8(c).

Definition 2.3. Given a process $Z$ and a random variable $M$, we define $\Theta_{\text {sup }}(M, Z)$ as the class of $\mathbb{F}^{I}=\mathbb{F}^{Z} \vee \sigma(M)$-adapted càglàd processes in $[0,1)$ which satisfy

$$
\begin{equation*}
\int_{0}^{1}\left|\theta_{s}\right| d s \in L^{1}(\Omega) \tag{2.1}
\end{equation*}
$$

$$
\begin{equation*}
\sup _{0 \leq t \leq 1}\left|\int_{0}^{t} H\left(s, \int_{0}^{s} \theta_{u} d u+Z_{s}\right) \theta_{s} d s\right| \in L^{1}(\Omega) \tag{2.2}
\end{equation*}
$$

$$
\begin{equation*}
\sup _{0 \leq t \leq 1}\left|\int_{0}^{t} H_{y}\left(s, \int_{0}^{s} \theta_{u} d u+Z_{s}\right) \theta_{s} d s\right| \in L^{1+\varepsilon}(\Omega) \quad \text { for some } \varepsilon>0, \tag{2.3}
\end{equation*}
$$

and

$$
\begin{equation*}
\exp \left(C \sup _{0 \leq t \leq 1}\left|\int_{0}^{t} \theta_{s} d s\right|\right) \in L^{1}(\Omega) \quad \forall C>0 \tag{2.4}
\end{equation*}
$$

for all $H \in \mathcal{H}$.
Remark 2.4. We could replace the technical conditions (2.1), (2.2), (2.3), and (2.4) in the definition of $\Theta_{\sup }(M, Z)$ by the stronger ones

$$
\begin{equation*}
\int_{0}^{1}\left|\theta_{s}\right|^{1+\varepsilon} d s \in L^{1}(\Omega) \quad \text { for some } \varepsilon>0 \tag{2.5}
\end{equation*}
$$

and

$$
\begin{equation*}
\exp \left(C \sup _{0 \leq t \leq 1}\left|\int_{0}^{t} \theta_{s} d s\right|\right) \in L^{1}(\Omega) \quad \forall C>0 \tag{2.6}
\end{equation*}
$$

The advantage is that conditions (2.5) and (2.6) define a linear space. On the other hand, condition (2.5) is difficult to verify in specific examples.

Definition 2.5. Given a process $Z$ and a random variable $M$, independent of $Z$, we say that a process $X$ is an ( $M, Z$ )-strategy process if there exists $\theta \in \Theta_{\sup }(M, Z)$ such that $X_{t}=\int_{0}^{t} \theta_{s} d s, t \in[0,1]$.

Definition 2.6. Given a final price $\xi \in L^{2}(\Omega)$, a stochastic process $Z$, a random variable $M$, independent of $Z$, a price semimartingale process $P \equiv\left(P_{t}\right)_{t \in[0,1]}$ with respect to $\mathbb{F}^{Z} \vee \sigma(M)$, and an $(M, Z)$-strategy process $X=\left(X_{t}\right)_{t \in[0,1]}$, we denote by $V=V(X, P, \xi)$ the agent final wealth defined by

$$
V(X, P, \xi)=V_{0}+\int_{0}^{1} X_{s} d P_{s}+\left(\xi-P_{1}\right) X_{1}
$$

whenever the above stochastic integral is well defined. Here $V_{0}$ is a constant.
Definition 2.7 (weak equilibrium). Let $L: C[0,1] \rightarrow \mathbb{R}^{k}$ be a measurable functional on the canonical Wiener space and $\mu$ be a probability measure on $\mathbb{R}$ with $\int_{\mathbb{R}} x^{2} \mu(d x)<\infty$. We say that there exists an $(L, \mu)$-weak equilibrium if there exists some probability space $(\Omega, \mathcal{F}, P)$ where there exist three processes $Y^{*}, \theta^{*}$, and $Z^{*}$, a random variable $\xi^{*}$, a random vector $\lambda^{*}$, and a function $H^{*} \in \mathcal{H}$ such that the following hold:
(i) $Y_{t}^{*}=X_{t}^{*}+Z_{t}^{*}$, where $X_{t}^{*}=\int_{0}^{t} \theta_{s}^{*} d s$ for $t \in[0,1]$.
(ii) $\lambda^{*}=L\left(Y^{*}\right)$ is independent of the process $Z^{*}$.
(iii) $Z^{*}$ is a Brownian motion.
(iv) $\xi^{*}$ has the law $\mu$.
(v) $\theta^{*} \in \Theta_{\text {sup }}\left(\lambda^{*}, Z^{*}\right)$.
(vi) Prices are rational. That is, $P_{t}^{*} \triangleq H^{*}\left(t, Y_{t}^{*}\right)=\mathbb{E}\left[\xi^{*} \mid \mathcal{F}_{t}^{Y^{*}}\right]$ for $t \in[0,1]$.
(vii) For all $\theta \in \Theta_{\sup }\left(\lambda^{*}, Z^{*}\right)$, one has

$$
\mathbb{E}\left[V\left(X, P, \xi^{*}\right)\right] \leq \mathbb{E}\left[V\left(X^{*}, P^{*}, \xi^{*}\right)\right]
$$

where $X .=\int_{0}^{\cdot} \theta_{s} d s, Y^{\theta}=X+Z^{*}$, and $P .=H^{*}\left(\cdot, Y^{\theta}\right)$.
Now we give a series of remarks related to this definition.
Remark 2.8. (a) It is clear that $Z^{*}$ is an $\mathbb{F}^{Z^{*}} \vee \sigma\left(\lambda^{*}\right)$-Brownian motion, as $Z^{*}$ is adapted to this filtration and is independent of $\lambda^{*}$.
(b) The price of the asset will be equal to $\xi^{*}$, just after the release of information at time $t=1$. This price has to have the prespecified law $\mu$. Furthermore, the relationship between $\xi^{*}$ and $\lambda^{*}$ is specified through the rationality of prices (property (vi) above), and in general $\xi^{*}$ is not independent of $Z^{*}$.
(c) The natural definition of the insider filtration is $\mathbb{F}^{*}=\mathbb{F}^{Z^{*}} \vee \sigma\left(\lambda^{*}\right)$. This is due to the monotonicity of the pricing rule and the fact that the insider observes the prices; one has that at time $t$ the insider can infer $Z_{t}^{*}$. Note that in general $\mathbb{F}^{Y^{*}}$ is not necessarily included in $\mathbb{F}^{Z^{*}}$.
(d) The set of $\left(\lambda^{*}, Z^{*}\right)$-strategies is usually nonempty. Furthermore, in the optimization problem (vii) above, one may think that it is more natural to restrict the opportunity set of strategies to the ones satisfying $L\left(Y^{\theta}\right)=\lambda^{*}$. This means that the insider would realize the giving of strong information on the total demand. In the next section, we will see that the optimum, with or without this restriction, is the same, given condition (ii) in Definition 2.7.
(e) From now on we will always assume that $\mu$ is a probability measure on $\mathbb{R}$ satisfying $\int_{\mathbb{R}} x^{2} \mu(d x)<\infty$ without any further mention.

Note that in the above (partial) equilibrium setup the insider optimizes his/her expected profit given the information $\lambda^{*}$ and $Z^{*}$. In this aspect, the above equilibrium is a partial one. In other words, if the agent uses the strategy $\theta^{*}$, there is no (local) reason to change strategy. It can also be considered as a stable point where the insider can actually realize all the conditions for a stable market. The above setup and the subsequent proofs to follow are not constructive.
3. Optimization problem for the insider. In this section we give necessary conditions for a process to solve the optimization problem stated in property (vii) of Defintion 2.7. Given a Wiener process $Z$ and a fixed random variable $M$, which is independent of $Z$, we define $\mathbb{F}^{I}=\mathbb{F}^{Z} \vee \sigma(M)$. As pointed out in the introduction, we use the classical approach of variational calculus.

From now on, we denote by a superindex $\theta$ on $Y_{t}$ the dependence of the total demand on the strategy of the insider. Then, $Y_{t}^{\theta}=\int_{0}^{t} \theta_{s} d s+Z_{t}$. Before studying the optimization problem we state the following property for the portfolio process $\theta$.

Lemma 3.1. If $\theta \in \Theta_{\text {sup }}(M, Z)$, then the price process $P_{t}^{\theta}=H\left(t, Y_{t}^{\theta}\right)$ is an $\mathbb{F}^{I}$-semimartingale and its decomposition is given by

$$
P_{t}^{\theta}=P_{0}^{\theta}+\int_{0}^{t}\left\{H_{t}\left(s, Y_{s}^{\theta}\right)+\frac{1}{2} H_{y y}\left(s, Y_{s}^{\theta}\right)+H_{y}\left(s, Y_{s}^{\theta}\right) \theta_{s}\right\} d s+\int_{0}^{t} H_{y}\left(s, Y_{s}^{\theta}\right) d Z_{s} .
$$

The proof is a straightforward application of Itô's formula, as $H \in C^{1,2}((0,1) \times \mathbb{R})$ and $Y^{\theta}$ is a semimartingale in the filtration $\mathbb{F}^{I}=\mathbb{F}^{Z} \vee \sigma(M)$. The next lemma is obtained using the integration by parts formula.

Lemma 3.2. Let $\theta$ be any $\mathbb{F}^{I}$-adapted process such that $\int_{0}^{1}\left|\theta_{s}\right| d s<\infty$ a.s. Then, we have that

$$
V \triangleq V\left(X, P^{\theta}, \xi\right)=V_{0}+\int_{0}^{1}\left(\xi-H\left(t, Y_{t}^{\theta}\right)\right) \theta_{t} d t
$$

Without loss of generality we assume from now on that $V_{0}=0$. The optimization problem we consider in this section is $\max _{\theta \in \Theta_{\sup }(M, Z)} J(\theta)$, where

$$
\begin{equation*}
J(\theta) \triangleq \mathbb{E}\left[\int_{0}^{1}\left(\xi-H\left(t, Y_{t}^{\theta}\right)\right) \theta_{t} d t\right], \quad \theta \in \Theta_{\sup }(M, Z) \tag{3.1}
\end{equation*}
$$

We also denote by $\theta^{*} \triangleq \arg \max _{\theta \in \Theta_{\sup }(M, Z)} J(\theta)$ when this process exists. The difficulties in solving this problem are due to the nonlinearity of the functional $J$ and the fact that $\Theta_{\text {sup }}(M, Z)$ is not a linear space.

Remark 3.3. Note that, for $\theta \in \Theta_{\text {sup }}(M, Z)$, we have that

$$
|J(\theta)| \leq \mathbb{E}\left[|\xi|\left|\int_{0}^{1} \theta_{t} d t\right|\right]+\mathbb{E}\left[\left|\int_{0}^{1} H\left(t, Y_{t}^{\theta}\right) \theta_{t} d t\right|\right]<\infty
$$

due to $\xi \in L^{2}(\Omega)$ and the fact that $\theta$ satisfies conditions (2.4) and (2.2). Furthermore, if $\theta$ is $\mathbb{F}^{I}$-adapted and satisfies the integrability conditions that define $\Theta_{\text {sup }}(M, Z)$, but it is not necessarily càglàd, then we also have that $|J(\theta)|<\infty$.

The first step in our strategy to solve the problem is to study the properties of $J(\theta)$ in the following linear subset of $\Theta_{\text {sup }}(M, Z)$ :

$$
\Theta_{b}(M, Z)=\left\{\theta \in \Theta_{\sup }(M, Z): \text { there exists } K>0 \text { such that } \forall \omega,\left|\theta_{s}(\omega)\right| \leq K\right\} .
$$

Lemma 3.4. If $v, \theta \in \Theta_{b}(M, Z)$, then

$$
\begin{align*}
D_{v} J(\theta) & \left.\triangleq \frac{d}{d \varepsilon} J(\theta+\varepsilon v)\right|_{\varepsilon=0} \\
& =\mathbb{E}\left[\int_{0}^{1} v_{t}\left(\xi-H\left(t, Y_{t}^{\theta}\right)\right) d t\right]-\mathbb{E}\left[\int_{0}^{1}\left(\int_{0}^{t} v_{s} d s\right) H_{y}\left(t, Y_{t}^{\theta}\right) \theta_{t} d t\right] . \tag{3.2}
\end{align*}
$$

Furthermore, the operator $D . J(\theta): v \rightarrow D_{v} J(\theta)$ is linear.
Proof. First note that for any $\theta, v \in \Theta_{b}(M, Z)$ and $\varepsilon>0$ one has that $\theta+\varepsilon v \in \Theta_{b}(M, Z)$, and differentiating under the integral sign (see Lemma 9.3) and applying Fubini's theorem one obtains

$$
\begin{aligned}
& \left.\frac{d}{d \varepsilon} J(\theta+\varepsilon v)\right|_{\varepsilon=0}=\mathbb{E}\left[\int_{0}^{1} v_{t}\left\{\left(\xi-H\left(t, Y_{t}^{\theta}\right)\right)-\int_{t}^{1} H_{y}\left(s, Y_{s}^{\theta}\right) \theta_{s} d s\right\} d t\right] \\
& =\mathbb{E}\left[\int_{0}^{1} v_{t}\left(\xi-H\left(t, Y_{t}^{\theta}\right)\right) d t\right]-\mathbb{E}\left[\int_{0}^{1}\left(\int_{0}^{t} v_{s} d s\right) H_{y}\left(t, Y_{t}^{\theta}\right) \theta_{t} d t\right]
\end{aligned}
$$

Remark 3.5. If $\theta \in \Theta_{\text {sup }}(M, Z)$ is such that

$$
\begin{equation*}
\mathbb{E}\left[\left(\xi-H\left(t, Y_{t}^{\theta}\right)\right)-\int_{t}^{1} H_{y}\left(s, Y_{s}^{\theta}\right) \theta_{s} d s \mid \mathcal{F}_{t}^{I}\right]=0, \tag{3.3}
\end{equation*}
$$

then for any $v \in \Theta_{b}(M, Z)$ we have

$$
\mathbb{E}\left[\int_{0}^{1} v_{t}\left(\xi-H\left(t, Y_{t}^{\theta}\right)\right) d t\right]-\mathbb{E}\left[\int_{0}^{1}\left(\int_{0}^{t} v_{s} d s\right) H_{y}\left(t, Y_{t}^{\theta}\right) \theta_{t} d t\right]=0 .
$$

From now on, we refer to (3.3) as the optimality equation.
The next step is to prove the concavity of $J(\theta)$ in $\Theta_{b}(M, Z)$. This is done in the following proposition, which makes use of a general result on convex analysis; see Proposition 9.2.

Proposition 3.6. If $H \in \mathcal{H}$ satisfies

$$
\begin{equation*}
H_{t y}(t, y)+\frac{1}{2} H_{y y y}(t, y) \leq 0, \tag{3.4}
\end{equation*}
$$

then $J(\theta)$ is concave in $\Theta_{b}(M, Z)$.
Proof. We will show that for every $\theta, \eta \in \Theta_{b}(M, Z)$ we have $D_{\eta-\theta} J(\theta) \geq J(\eta)-J(\theta)$, which thanks to Proposition 9.2 is equivalent to $J(\theta)$ being concave. Given $\theta, \eta \in \Theta_{b}(M, Z)$, $\alpha \in[0,1]$, define $\delta \triangleq \eta-\theta$ and $\Psi^{\alpha} \triangleq \theta+\alpha \delta$. For $\alpha \in[0,1]$, define $\varphi(\alpha) \triangleq D_{\eta-\theta} J\left(\Psi^{\alpha}\right)=$ $\frac{d}{d \alpha} J\left(\Psi^{\alpha}\right)$. We will show that $\varphi^{\prime}(\alpha)=\frac{d^{2}}{d \alpha^{2}} J\left(\Psi^{\alpha}\right) \leq 0$. First, by Lemma 9.3 we have that

$$
\varphi^{\prime}(\alpha)=-\mathbb{E}\left[\int_{0}^{1}\left(\int_{0}^{t} \delta_{s} d s\right)^{2} H_{y y}\left(t, Y_{t}^{\alpha}\right) \Psi_{t}^{\alpha} d t\right]-\mathbb{E}\left[2\left(\int_{0}^{t} \delta_{s} d s\right) H_{y}\left(t, Y_{t}^{\alpha}\right) \delta_{t} d t\right] .
$$

Then we apply integration by parts in the second expectation to obtain that

$$
\begin{aligned}
& \mathbb{E}\left[\int_{0}^{1} 2\left(\int_{0}^{t} \delta_{s} d s\right) H_{y}\left(t, Y_{t}^{\Psi^{\alpha}}\right) \delta_{t} d t\right]=\mathbb{E}\left[\left(\int_{0}^{1} \delta_{s} d s\right)^{2} H_{y}\left(1, Y_{1}^{\Psi^{\alpha}}\right)\right] \\
& -\mathbb{E}\left[\int_{0}^{1}\left(\int_{0}^{t} \delta_{s} d s\right)^{2}\left\{H_{t y}\left(t, Y_{t}^{\Psi^{\alpha}}\right)+\frac{1}{2} H_{y y y}\left(t, Y_{t}^{\Psi^{\alpha}}\right)+H_{y y}\left(t, Y_{t}^{\Psi^{\alpha}}\right) \Psi_{t}^{\alpha}\right\} d t\right]
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\varphi^{\prime}(\alpha) & =-\mathbb{E}\left[\left(\int_{0}^{1} \delta_{s} d s\right)^{2} H_{y}\left(1, Y_{1}^{\Psi^{\alpha}}\right)\right] \\
& +\mathbb{E}\left[\int_{0}^{1}\left(\int_{0}^{t} \delta_{s} d s\right)^{2}\left\{H_{t y}\left(t, Y_{t}^{\Psi^{\alpha}}\right)+\frac{1}{2} H_{y y y}\left(t, Y_{t}^{\Psi^{\alpha}}\right)\right\} d t\right] .
\end{aligned}
$$

Using the fact that $H_{y}>0$ and (3.4) we conclude that $\varphi^{\prime}(\alpha) \leq 0$, and therefore $\varphi$ is a decreasing function. On the other hand, an application of the mean value theorem gives that $J(\eta)-J(\theta)=D_{\eta-\theta} J\left(\Psi^{\alpha^{*}}\right)=\varphi\left(\alpha^{*}\right)$ for some $\alpha^{*} \in[0,1]$. Therefore $J(\eta)-J(\theta) \leq \varphi(0)=$ $D_{\eta-\theta} J(\theta)$.

The following theorem gives a sufficient condition to find the optimal process in $\Theta_{\text {sup }}(M, Z)$.
Theorem 3.7. Let $H \in \mathcal{H}$ satisfy (3.4). If $\theta^{*} \in \Theta_{\text {sup }}(M, Z)$ is such that

$$
\mathbb{E}\left[\left(\xi-H\left(t, Y_{t}^{\theta^{*}}\right)\right)-\int_{t}^{1} H_{y}\left(s, Y_{s}^{\theta^{*}}\right) \theta_{s}^{*} d s \mid \mathcal{F}_{t}^{I}\right]=0
$$

then $J(\theta) \leq J\left(\theta^{*}\right), \theta \in \Theta_{\text {sup }}(M, Z)$.
Proof. According to Propositions 9.5 and 9.6 , there exists a sequence $\left\{\theta^{*, n}\right\}_{n \in \mathbb{N}} \subseteq \Theta_{b}(M, Z)$ such that $\lim _{n \rightarrow \infty} J\left(\theta^{*, n}\right)=J\left(\theta^{*}\right)$ and

$$
\lim _{n \rightarrow \infty} D_{\theta-\theta^{*, n}} J\left(\theta^{*, n}\right)=0 \quad \forall \theta \in \Theta_{b}(M, Z) .
$$

Using Proposition 3.6, we obtain that $J$ is concave in $\Theta_{b}(M, Z)$, and therefore we have that $J(\theta) \leq J\left(\theta^{*, n}\right)+D_{\theta-\theta^{*, n}} J\left(\theta^{*, n}\right), \theta \in \Theta_{b}(M, Z)$. Therefore, taking limits one gets that $J(\theta) \leq J\left(\theta^{*}\right), \theta \in \Theta_{b}(M, Z)$. Using Proposition 9.5 again, we have that for all $\theta \in \Theta_{\text {sup }}(M, Z)$ there exists a sequence $\left\{\theta^{n}\right\}_{n \in \mathbb{N}} \subseteq \Theta_{b}(M, Z)$ such that $\lim _{n \rightarrow \infty} J\left(\theta^{n}\right)=J(\theta) \leq J\left(\theta^{*}\right)$.
4. Properties of the solutions to the optimality equation. The following proposition is important for finding a strategy $\theta^{*}$ satisfying the optimality equation and yielding a rational price. It tell us that given an insider's strategy satisfying the optimality equation, the price process associated with this strategy is rational if and only if the market maker sees the associated total demand as a Brownian motion. In other words, the associated price process is rational if and only if the market maker sees the total demand as if only the noise trader were buying or selling stocks. Moreover, this suggests the connection between the optimal insider's demand and the compensator of a Brownian motion with respect to an enlarged filtration; see Remark 5.2.

Proposition 4.1. Assume there exists a process $\theta^{*} \in \Theta_{\text {sup }}(M, Z)$ satisfying the optimality equation (3.3). Then $H\left(\cdot, Y^{\theta^{*}}\right)$ is an $\mathbb{F}^{Y^{\theta^{*}}}$-martingale if and only if $Y^{\theta^{*}}$ is an $\mathbb{F}^{Y^{\theta^{*}}}$-Brownian motion.

Proof. Assume that $\theta^{*}$ and $Y_{t}^{\theta^{*}}=\int_{0}^{t} \theta_{s}^{*} d s+Z_{t}$ satisfy the optimality equation. Note that this equation is equivalent to

$$
\begin{equation*}
H\left(t, Y_{t}^{\theta^{*}}\right)-\int_{0}^{t} H_{y}\left(s, Y_{s}^{\theta^{*}}\right) \theta_{s}^{*} d s=\mathbb{E}\left[\xi \mid \mathcal{F}_{t}^{I}\right]-M_{t} \tag{4.1}
\end{equation*}
$$

where $M_{t}=\mathbb{E}\left[\int_{0}^{1} H_{y}\left(s, Y_{s}^{\theta^{*}}\right) \theta_{s}^{*} d s \mid \mathcal{F}_{t}^{I}\right]$. Making $t=0$ in (4.1), we obtain $H(0,0)+M_{0}=$ $\mathbb{E}\left[\xi \mid \mathcal{F}_{0}^{I}\right]$. Applying Itô's formula to $H\left(t, Y_{t}^{\theta^{*}}\right)$ in (4.1), we get

$$
\begin{align*}
& \int_{0}^{t}\left\{H_{t}\left(s, Y_{s}^{\theta^{*}}\right)+\frac{1}{2} H_{y y}\left(s, Y_{s}^{\theta^{*}}\right)\right\} d s \\
& =-\int_{0}^{t} H_{y}\left(s, Y_{s}^{\theta^{*}}\right) d Z_{s}+\mathbb{E}\left[\xi \mid \mathcal{F}_{t}^{I}\right]-\mathbb{E}\left[\xi \mid \mathcal{F}_{0}^{I}\right]-\left(M_{t}-M_{0}\right) \tag{4.2}
\end{align*}
$$

for all $t \in[0,1]$. The right-hand side of (4.2) is a continuous $\mathbb{F}^{I}$-local martingale with initial value 0 , and the left-hand side is a finite variation process with continuous paths. Therefore, both processes must be identically zero. Therefore, we have that

$$
\begin{equation*}
H_{t}\left(t, Y_{t}^{\theta^{*}}\right)+\frac{1}{2} H_{y y}\left(t, Y_{t}^{\theta^{*}}\right)=0, \quad t \in[0,1] . \tag{4.3}
\end{equation*}
$$

Combining the above equation with Itô's formula, we have

$$
\begin{equation*}
H\left(t, Y_{t}^{\theta^{*}}\right)=H(0,0)+\int_{0}^{t} H_{y}\left(s, Y_{s}^{\theta^{*}}\right) d Y_{s}^{\theta^{*}} . \tag{4.4}
\end{equation*}
$$

If $Y^{\theta^{*}}$ is an $\mathbb{F}^{r^{\theta^{*}}}$-Brownian motion, then the stochastic integral $\int_{0}^{t} H_{y}\left(s, Y_{s}^{\theta^{*}}\right) d Y_{s}^{\theta^{*}}$ is a martingale due to Lemma 9.1. Therefore $H\left(t, Y_{t}^{\theta^{*}}\right)$ is an $\mathbb{F}^{Y^{\theta^{*}}}$-martingale. Conversely, note that as $H_{y}>0$, we can write $Y_{t}^{\theta^{*}}=\int_{0}^{t} \frac{d H\left(s, Y_{\theta^{*}}^{\theta^{*}}\right)}{H_{y}\left(s, Y_{\theta^{\theta^{*}}}\right)}$. Hence, if we assume that $H\left(t, Y_{t}^{\theta^{*}}\right)$ is an $\mathbb{F}^{Y^{\theta^{*}}}$ martingale, then $Y^{\theta^{*}}$ is an $\mathbb{F}^{Y^{\theta^{*}}}$-local martingale. As $Y^{\theta^{*}}$ has the same quadratic variation as $Z$, we obtain that $Y^{\theta^{*}}$ is actually a Brownian motion with respect to its own filtration.

Corollary 4.2. If there exists a process $\theta^{*} \in \Theta_{\text {sup }}(M, Z)$ satisfying (3.3) and $H\left(\cdot, Y_{.}^{\theta^{*}}\right)$ is an $\mathbb{F}^{Y^{\theta^{*}}}$-martingale, then $H$ and $\xi$ must satisfy

$$
\begin{equation*}
H_{t}(t, y)+\frac{1}{2} H_{y y}(t, y)=0 \tag{4.5}
\end{equation*}
$$

and

$$
\begin{equation*}
H\left(1, Y_{1}^{\theta^{*}}\right)=\mathbb{E}\left[\xi \mid \mathcal{F}_{1}^{I}\right] . \tag{4.6}
\end{equation*}
$$

Proof. Equation (4.3) and the fact that $Y^{\theta^{*}}$ is a Brownian motion in its own filtration lead to (4.5). Making $t=1$ in the optimality equation (3.3), one obtains (4.6).
5. Existence and uniqueness in law of weak equilibrium. We start this section with a result giving sufficient conditions to obtain an $(L, \mu)$-weak equilibrium. The first condition in Theorem 5.1 essentially says that the law $\mu$ of the asset value $\xi$ must be a smooth transformation of a standard normal random variable. Actually, in the examples of the following sections we do not specify the law $\mu$ but the pricing rule $H$, which gives this smooth transformation. We will consider the pricing rules studied previously in the literature, which are $H(t, y)=y$ and $H(t, y)=e^{y+(1-t) / 2}$; see [2]. Notice that the exponential pricing rule has much more economical interpretation, as it implies that prices are lognormally distributed. The second condition says that there exists a Brownian motion $W$ such that $W$ is a semimartingale with respect to enlarged filtration $\mathbb{F}^{W} \vee \sigma(L(W))$. According to Proposition 4.1, if the insider wants to obtain a rational price process, then the total demand $Y$ must be a Brownian motion with respect to its natural filtration. Therefore, it is natural to impose that the additional information of the insider, given by $L(Y)$, be such that the total demand remains a Brownian motion with respect to the enlarged filtration $\mathbb{F}^{Y} \vee \sigma(L(Y))$ and then use the compensator as the insider's strategy. The technical condition in order to carry out this argument is $\mathbb{F}^{Y} \vee \sigma(L(Y))=\mathbb{F}^{I}$ plus some integrability conditions, which is the third condition in the theorem.

From an economical point of view, it seems reasonable to expect that the insider cannot possess "too much additional information" for an equilibrium to hold. In Theorem 5.1, this is reflected in the semimartingale property of $W$. In fact, if $L(W)$ gives too much information to the insider, then $W$ will not be a semimartingale with respect to the enlarged filtration and, therefore, prices will not be rational. Although there exists a general criterion to ensure that a given functional satisfies this semimartingale condition, known as Jacod's criterion (see, for instance, Theorem 10 in Chapter VI of [19]), this criterion does not apply to our main examples $L(W)=\max _{0 \leq t \leq 1} W_{t}$ or $L(W)=\arg \max _{0 \leq t \leq 1} W_{t}$. In other models of insider trading, where the rationality of prices is not taken into account, this condition is not sufficient to provide realistic models with finite expected wealth for the insider optimization problem; see [12]. Usually the information held by the insider has to be perturbed by some noise; see [12] and [6].

Theorem 5.1 (existence). Given a measurable functional $L: \mathcal{C}[0,1] \rightarrow \mathbb{R}^{k}$ and $\mu$ a probability measure on $\mathbb{R}$ satisfying $\int_{\mathbb{R}} x^{2} \mu(d x)<\infty$, we assume the following:
(1) There exists $H \in \mathcal{H}$ such that it satisfies (4.5) and

$$
\mu(A)=\frac{1}{\sqrt{2 \pi}} \int_{H(1, \cdot)^{-1}(A)} e^{-x^{2} / 2} d x \quad \forall A \in \mathcal{B}(\mathbb{R})
$$

(2) There exists a probability space $(\Omega, \mathcal{F}, P)$ supporting a Brownian motion $W$ which is a semimartingale in the filtration $\mathbb{F}^{W} \vee \sigma(\lambda), \lambda \triangleq L(W)$, with semimartingale decomposition $W_{t}=\int_{0}^{t} \alpha(s, \lambda) d s+W_{t}^{\lambda}$, where $W^{\lambda}$ is an $\mathbb{F}^{W} \vee \sigma(\lambda)$-Brownian motion.
(3) $\alpha \in \Theta_{\sup }\left(\lambda, W^{\lambda}\right)$.

Then

$$
\left(Y^{*}, \theta^{*}, Z^{*}, H^{*}, \xi^{*}, \lambda^{*}\right)=\left(W, \alpha(\cdot, \lambda), W^{\lambda}, H, H\left(1, W_{1}\right), \lambda\right)
$$

is an $(L, \mu)$-weak equilibrium.
Proof. Verification of properties (i), (iii), (iv), and (v) in the definition of weak equilibrium (Definition 2.7) is straightforward. Property (ii) follows from the fact that $W_{t}^{\lambda}$ is an
$\mathbb{F}^{W} \vee \sigma(\lambda)$-Brownian motion and, hence, is independent of $\mathcal{F}_{0}^{W} \vee \sigma(\lambda)=\sigma(\lambda)$. From hypothesis (1), together with (4.4), we have that $H\left(\cdot, W\right.$.) is an $\mathbb{F}^{W}$-martingale. As $H\left(t, W_{t}\right)=$ $\mathbb{E}\left[H\left(1, W_{1}\right) \mid \mathcal{F}_{t}^{W}\right]$, property (vi) follows. To check property (vii), we apply Itô's formula to $H(\cdot, W$.) in the left-hand side of the optimality equation (3.3) with $\theta=\alpha$. Due to hypothesis (1) we obtain that it is equal to

$$
\mathbb{E}\left[\int_{t}^{1} H_{y}\left(s, W_{s}\right) d W_{s}^{\lambda} \mid \mathcal{F}_{t}^{W^{\lambda}} \vee \sigma(\lambda)\right]
$$

On the other hand, hypothesis (3) implies that $\alpha(\cdot, \lambda)$ is $\mathbb{F}^{W^{\lambda}} \vee \sigma(\lambda)$-adapted, which entails that $W$ is $\mathbb{F}^{W^{\lambda}} \vee \sigma(\lambda)$-adapted, and one can conclude that $\mathbb{F}^{W} \vee \sigma(\lambda)=\mathbb{F}^{W^{\lambda}} \vee \sigma(\lambda)$. Hence, due to Lemma 9.1, the above conditional expectation equals zero, and the conclusion follows from Theorem 3.7.

Remark 5.2. Of the three hypotheses in the previous theorem, hypothesis (3) is difficult to verify in general. Besides the integrability conditions in the definition of $\Theta_{\sup }\left(\lambda, W^{\lambda}\right), \alpha(\cdot, \lambda)$ must be $\mathbb{F}^{W^{\lambda}} \vee \sigma(\lambda)$-adapted. This property will follow if $\mathbb{F}^{W} \vee \sigma(\lambda)=\mathbb{F}^{W \lambda} \vee \sigma(\lambda)$. This problem seems to be difficult to solve in general.

We deal with this problem in each of the examples to follow in the next sections. The general strategy is to show existence and uniqueness for SDEs of the form

$$
X_{t}=\int_{0}^{t} \alpha\left(s, G, X_{[0, s]}\right) d s+V_{t}
$$

where $V$ is a Brownian motion, $\alpha$ is a (degenerate) functional, and $G$ is a random variable independent of $V$. Therefore, $X$ would be $\mathbb{F}^{V} \vee \sigma(G)$-adapted.

The following theorem gives a uniqueness result for the $(L, \mu)$-weak equilibrium found in the previous theorem. Condition (6) in the following theorem deserves a comment. This assumption roughly says that two weak equilibriums have the same law whenever they are obtained through a semimartingale decomposition of a Brownian motion with respect to an enlarged filtration. In other words, if in condition (2) of Theorem 5.1 we use two different Brownian motions possibly defined in two different probability spaces, the two different weak equilibriums obtained have the same law. From the economic point of view, this assumption states that if the market maker knew the insider's additional information, then he/she would have exactly the same information flow as the insider.

Theorem 5.3 (uniqueness in law). Assume the same hypotheses of Theorem 5.1, and denote by $\left(Y^{*}, \theta^{*}, Z^{*}, H^{*}, \xi^{*}, \lambda^{*}\right)$ the $(L, \mu)$-weak equilibrium. Suppose that there exists another probability space supporting processes $(Y, \theta, Z)$ such that
(1) $Y_{t}=\int_{0}^{t} \theta_{s} d s+Z_{t}$;
(2) $\lambda \triangleq L(Y)$ is independent of $Z$;
(3) $Z$ is a Brownian motion in its own filtration;
(4) $\theta \in \Theta_{\sup }(\lambda, Z)$;
(5) $H^{*}\left(t, Y_{t}\right)=\mathbb{E}\left[H^{*}\left(1, Y_{1}\right) \mid \mathcal{F}_{t}^{Y}\right]$ for $t \in[0,1]$.
(6) $\mathbb{F}^{Z} \vee \sigma(\lambda)=\mathbb{F}^{Y} \vee \sigma(\lambda)$.

Then, we have that $\mathcal{L}\left(Y^{*}, X^{*}, Z^{*}, \xi^{*}, \lambda^{*}\right)=\mathcal{L}(Y, X, Z, \xi, \lambda)$, where $\xi \triangleq H^{*}\left(1, Y_{1}\right)$, and therefore $\mathbb{E}[V(X, P, \xi)]=\mathbb{E}^{*}\left[V\left(X^{*}, P^{*}, \xi^{*}\right)\right]$.

Proof. Applying Itô's formula in the filtration $\mathbb{F}^{I}=\mathbb{F}^{Z} \vee \sigma(\lambda)$, we have that

$$
\xi-H^{*}\left(t, Y_{t}\right)-\int_{t}^{1} H_{y}^{*}\left(s, Y_{s}\right) \theta_{s} d s=\int_{t}^{1} H_{y}^{*}\left(s, Y_{s}\right) d Z_{s}
$$

where in the last equality we have used that $H^{*}$ satisfies (4.5). After taking conditional expectation, this yields

$$
\mathbb{E}\left[\xi-H^{*}\left(t, Y_{t}\right)-\int_{t}^{1} H_{y}^{*}\left(s, Y_{s}\right) \theta_{s} d s \mid \mathcal{F}_{t}^{I}\right]=0
$$

Then, by Theorem 3.7 we have that $J(\eta) \leq J(\theta)$ for all $\eta \in \Theta_{\text {sup }}(L(Y), Z)$. By hypothesis (5) and Proposition 4.1 one gets that $Y$ is a Brownian motion in its own filtration. Therefore, $\mathcal{L}(Y, \lambda, \xi)=\mathcal{L}\left(Y^{*}, \lambda^{*}, \xi^{*}\right)$. As the process $\theta^{*}$ is adapted to $\mathbb{F}^{Y^{*}} \vee \sigma\left(\lambda^{*}\right)$, it can be written as $\theta_{t}^{*}=\Lambda\left(t, Y_{[0, t]}^{*}, L\left(Y_{[0,1]}^{*}\right)\right), P \times \lambda$-a.s. Then, defining $\hat{\theta}_{t} \triangleq \Lambda\left(t, Y_{[0, t]}, L\left(Y_{[0,1]}\right)\right)$ and using that $\mathcal{L}\left(Y^{*}\right)=\mathcal{L}(Y)$, we have that

$$
\mathbb{E}\left[Y_{t}-Y_{s}-\int_{s}^{t} \hat{\theta}_{u} d u \mid \mathcal{F}_{s}^{Y} \vee \sigma(\lambda)\right]=0
$$

Thus, $Y_{t}=\int_{0}^{t} \hat{\theta}_{s} d s+M_{t}=\int_{0}^{t} \theta_{s} d s+Z_{t}$, where $M$ is an $\mathbb{F}^{Y} \vee \sigma(\lambda)$-martingale. Given assumption (6), the uniqueness of the semimartingale decomposition of $Y$ with respect to $\mathbb{F}^{Y} \vee \sigma(\lambda)$ proves that $\widehat{\theta}=\theta, P \times \lambda$-a.s.

The following result is helpful when proving that $\alpha \in \Theta_{\text {sup }}(\lambda, Z)$.
Proposition 5.4. Let $Y$ be a Brownian motion, and let $\lambda=L(Y)$. Assume that $Y$ has a semimartingale decomposition with respect to $\mathbb{F}^{Y} \vee \sigma(\lambda)$ given by $Y_{t}=\int_{0}^{t} \alpha_{s} d s+Z_{t}$, where $Z$ is an $\mathbb{F}^{Y} \vee \sigma(\lambda)$-Brownian motion. Then, $\exp \left(C \sup _{0 \leq t \leq 1}\left|\int_{0}^{t} \alpha_{s} d s\right|\right) \in L^{p}(\Omega), p \geq 1$, for all $C>0$, and $\sup _{0 \leq t \leq 1}\left|\int_{0}^{t} F\left(s, Y_{s}\right) \alpha_{s} d s\right| \in L^{p}(\Omega), p \geq 1$, where $F$ is any function satisfying an exponential growth condition.

Proof. To prove the first statement, notice that

$$
\left|\exp \left(C \sup _{0 \leq t \leq 1}\left|\int_{0}^{t} \alpha_{s} d s\right|\right)\right|^{p} \leq \exp \left(p C \sup _{0 \leq t \leq 1}\left|Y_{t}\right|\right) \exp \left(p C \sup _{0 \leq t \leq 1}\left|Z_{t}\right|\right) .
$$

By the Cauchy-Schwarz inequality, taking into account that $Y$ and $Z$ are Brownian motions, we obtain that

$$
\mathbb{E}\left[\left|\exp \left(C \sup _{0 \leq t \leq 1}\left|\int_{0}^{t} \alpha_{s} d s\right|\right)\right|^{p}\right] \leq\left(\mathbb{E}\left[\exp \left(2 p C \sup _{0 \leq t \leq 1}\left|Y_{t}\right|\right)\right]\right)^{2}<\infty
$$

To prove the second statement, note that

$$
\left|\int_{0}^{t} F\left(s, Y_{s}\right) \alpha_{s} d s\right|^{p} \leq C(p)\left(\left|\int_{0}^{t} F\left(s, Y_{s}\right) d Y_{s}\right|^{p}+\left|\int_{0}^{t} F\left(s, Y_{s}\right) d Z_{s}\right|^{p}\right) .
$$

Define $M_{t}^{1} \triangleq \int_{0}^{t} F\left(s, Y_{s}\right) d Y_{s}$ and $M_{t}^{2} \triangleq \int_{0}^{t} F\left(s, Y_{s}\right) d Z_{s}$. Here, $M^{1}$ is an $\mathbb{F}^{Y}$-local martingale. By the BDG inequality (see Theorem 73 on page 222 of [19]), taking into account that
$F$ satisfies an exponential growth condition and that $Y$ is a Brownian motion, we obtain that

$$
\begin{aligned}
\mathbb{E}\left[\sup _{0 \leq t \leq 1}\left|M_{t}^{1}\right|^{p}\right] & \leq C_{p} \mathbb{E}\left[\left(\int_{0}^{1} F\left(s, Y_{s}\right)^{2} d s\right)^{p / 2}\right] \\
& \leq C_{p} \mathbb{E}\left[\left(\int_{0}^{1} A^{2} \exp \left(2 B\left|Y_{s}\right|\right) d s\right)^{p / 2}\right] \\
& \leq C_{p} A^{p} \mathbb{E}\left[\exp \left(p B \sup _{0 \leq t \leq 1}\left|Y_{t}\right|\right)\right]<\infty .
\end{aligned}
$$

Thus $M^{1}$ is an $\mathbb{F}^{Y}$-martingale and $\sup _{0 \leq t \leq 1}\left|M_{t}^{1}\right| \in L^{p}(\Omega), p \geq 1$. We can repeat the same argument for $M^{2}$, taking into account that $M^{2}$ is an $\mathbb{F}^{Y} \vee \sigma(\lambda)$-local martingale.
6. Back's example and an example of binary information. In this section we comment on two known examples where the general result in Theorem 5.1 applies. Throughout this section we will consider a Brownian motion $W$ defined on a complete probability space $(\Omega, \mathcal{F}, P)$. From now on, we denote by $\phi(x, t)$ the density of a centered Gaussian random variable with variance $t$ and by $\Phi(x, t)$ its distribution function, and $\bar{\Phi}(x, t)=1-\Phi(x, t)$.

In all the examples to follow in the next sections, we assume that $\mu$ is a probability measure on $\mathbb{R}$ with $\int_{\mathbb{R}} x^{2} \mu(d x)<\infty$ and that there exists $H \in \mathcal{H}$ satisfying (4.5) and

$$
\mu(A)=\int_{H(1, \cdot)^{-1}(A)} \phi(x, 1) d x \quad \forall A \in \mathcal{B}(\mathbb{R}) .
$$

Theorem 6.1. Let $W$ be a Brownian motion. Then $W$ is a semimartingale with respect to the filtration $\mathbb{F}^{W} \vee \sigma\left(W_{1}\right)$ with decomposition

$$
\begin{equation*}
W_{t}=\int_{0}^{t} \alpha\left(u, W_{1}\right) d u+W_{t}^{W_{1}} \quad \forall t \in[0,1], \tag{6.1}
\end{equation*}
$$

where $W^{W_{1}}$ is an $\mathbb{F}^{W} \vee \sigma\left(W_{1}\right)$-Brownian motion,

$$
\begin{equation*}
\alpha\left(t, W_{1}\right)=\frac{W_{1}-W_{t}}{1-t}, \tag{6.2}
\end{equation*}
$$

for all $t \in[0,1)$.
The previous result is well known, and its proof can be found, for instance, in [11, Théorème 1]. In [11, Corollaire 1.1], the connection between the Brownian bridge $\left\{W_{t}-\right.$ $\left.t W_{1}\right\}_{0 \leq t<1}$ and the Brownian motion $\left\{W_{t}^{W_{1}}\right\}_{0 \leq t<1}$ is also discussed, showing that these two processes have the same natural filtration and that $W_{1}$ is independent of $\left\{W_{t}^{W_{1}}\right\}_{0 \leq t<1}$. This idea is later used in order to consider (6.1) as a linear equation, where the unknown function is $W .(\omega)$ and $W_{t}^{W_{1}}(\omega)$ and $W_{1}(\omega)$ are given. The following result is slightly more general than Corollaire 1.1 in [11], in the sense that if we assume that we are given a Brownian motion $B$ and a random variable $G$, independent of $B$ and not necessarily Gaussian, we can construct a process $X$ with terminal value $G$. In the particular case that $\mathcal{L}(G)=\mathcal{N}(0,1)$, the process $X$ is a Brownian bridge with $X_{1}=G$.

Theorem 6.2. Let $B$ be a Brownian motion and $G$ be a random variable independent of $B$, both defined in the same probability space $(\Omega, \mathcal{F}, P)$. Then there exists a unique strong solution $X$ adapted to the filtration $\mathbb{F}^{B} \vee \sigma(G)$ of the following SDE:

$$
\begin{equation*}
X_{t}=\int_{0}^{t} \frac{G-X_{s}}{1-s} d s+B_{t}, \quad t \in[0,1] . \tag{6.3}
\end{equation*}
$$

Furthermore, if we assume that the law of $G$ is $N(0,1)$, then $X_{t}$ is a Brownian motion with respect to its own filtration.

Proof. As $G$ is independent of $B$, one has that $B$ is an $\mathbb{F}^{B} \vee \sigma(G)$-Brownian motion. Using $(1-t)^{-1}$ as an integrating factor, we obtain

$$
d\left(\frac{X_{t}}{1-t}\right)=\frac{G}{(1-t)^{-2}} d t+\frac{d B_{t}}{1-t} .
$$

Therefore, one has that $X_{t}=t G+(1-t) \int_{0}^{t} \frac{d B_{s}}{1-s}, 0 \leq t<1$. In Lemma 6.9 on page 358 of [13], it is proved that the process

$$
\begin{aligned}
& \bar{B}_{t}=(1-t) \int_{0}^{t} \frac{d B_{s}}{1-s}, \quad 0 \leq t<1, \\
& \bar{B}_{1}=0
\end{aligned}
$$

is a continuous, centered Gaussian process with covariance function $s \wedge t-s t$. Hence we have proved the existence and uniqueness for the solutions of (6.3). If we assume that $G \sim N(0,1)$, we have that $t G$ is a continuous, centered Gaussian process with covariance function st. As the sum of two independent Gaussian processes is still a Gaussian process and $B_{t}$ and $G$ are independent, we obtain that $X$ is a continuous, centered Gaussian process with covariance function $s \wedge t$ and thus a standard Brownian motion.

The following property is important to determine the finiteness of optimal utilities. For $p>0, \mathbb{E}\left[\int_{0}^{1}|\alpha(t, G)|^{p} d t\right]<\infty$ if and only if $p<2$, where $\alpha(t, x)=\frac{x-X_{t}}{1-t}$ for all $t \in[0,1]$.

Let's state the weak equilibrium result for this case.
Theorem 6.3. Let $L(Y)=Y_{1}$. Then

$$
\left(Y^{*}, \theta^{*}, Z^{*}, H^{*}, \xi^{*}, \lambda^{*}\right)=(X, \alpha(\cdot, G), B, H, H(1, G), L(X))
$$

is an $(L, \mu)$-weak equilibrium.
In this particular case the above weak equilibrium is in fact a strong-type equilibrium. For this, see Theorem 1 in [2] or Proposition 2 in [5].

Theorem 6.4. Assume that we are given a Brownian motion $Z$ and strong information $\xi$. Assume that $H^{*} \in \mathcal{H}$ satisfies (4.5) and $\xi \sim H^{*}(1, N(0,1))$. Set $\theta_{t}^{*} \triangleq \alpha\left(t,\left(H^{*}\right)^{-1}(1, \xi)\right)$. Then $\left(H^{*}, \theta^{*}\right)$ is an equilibrium. That is, the following hold:

- $H^{*}\left(t, Y^{*}\right)$ is a rational price, that is, $H^{*}\left(t, Y^{*}\right)=\mathbb{E}\left[\xi \mid \mathcal{F}_{t}^{Y^{*}}\right]$.
- For all $\theta \in \Theta_{\text {sup }}(\xi, Z)$, one has

$$
\mathbb{E}[V(X, P, \xi)] \leq \mathbb{E}\left[V\left(X^{*}, P^{*}, \xi\right)\right]
$$

where $X^{(*)}=\int_{0}^{r} \theta_{s}^{(*)} d s, Y^{(*)}=X^{(*)}+Z$, and $P^{(*)}=H^{*}\left(\cdot, Y_{.}^{(*)}\right)$.

Now we consider the case in which the insider knows that the total demand at time 1 is greater than or equal to a fixed constant $a$. The next two results are quoted from [12, Example 4.6].

Theorem 6.5. Let $W$ be a Brownian motion. Then $W$ is a semimartingale with respect to the filtration $\mathbb{F}^{W} \vee \sigma\left(\mathbf{1}_{[a, \infty)}\left(W_{1}\right)\right)$ with decomposition

$$
W_{t}=\int_{0}^{t} \alpha\left(u, \mathbf{1}_{[a, \infty)}\left(W_{1}\right)\right) d u+W_{t}^{a} \quad \forall t \in[0,1],
$$

where $W^{a}$ is an $\mathbb{F}^{W} \vee \sigma\left(\mathbf{1}_{[a, \infty)}\left(W_{1}\right)\right)$-Brownian motion,

$$
\alpha\left(t, \mathbf{1}_{[a, \infty)}\left(W_{1}\right)\right)=\frac{\phi\left(W_{t}-a, 1-t\right)}{\Phi\left(W_{t}-a, 1-t\right)} \mathbf{1}_{[a, \infty)}\left(W_{1}\right)+\frac{\phi\left(W_{t}-a, 1-t\right)}{\Phi\left(a-W_{t}, 1-t\right)} \mathbf{1}_{[a, \infty)^{c}}\left(W_{1}\right),
$$

for all $t \in[0,1]$.
Lemma 6.6. We have that $\mathbb{E}\left[\int_{0}^{1}\left|\alpha\left(t, \mathbf{1}_{[a, \infty)}\left(W_{1}\right)\right)\right|^{2} d t\right]<\infty$.
Theorem 6.7. Let $B$ be a Brownian motion and $G$ be a Bernoulli random variable independent of $B$, both defined on the same probability space $(\Omega, \mathcal{F}, P)$. Then there exists a unique strong solution $X$ adapted to the filtration $\mathbb{F}^{B} \vee \sigma(G)$ of the following SDE:

$$
\begin{equation*}
X_{t}=\int_{0}^{t}\left(\frac{\phi\left(X_{t}-a, 1-t\right)}{\Phi\left(X_{t}-a, 1-t\right)} \mathbf{1}_{\{1\}}(G)+\frac{\phi\left(X_{t}-a, 1-t\right)}{\Phi\left(a-X_{t}, 1-t\right)} \mathbf{1}_{\{0\}}(G)\right) d s+B_{t} \tag{6.4}
\end{equation*}
$$

for $0 \leq t<1$.
Proof. First we will prove that $\Psi_{a}^{1}(x, t) \triangleq \phi(x-a, 1-t) / \Phi(x-a, 1-t)$ is Lipschitz in the $x$ variable for $t \in[0,1)$ fixed. Note that we can take $a=0$, without loss of generality. Furthermore, $\Psi_{0}^{1}(x, t)=\Psi_{0}^{1}(x / \sqrt{1-t}, 0) / \sqrt{1-t}$.

We have that

$$
\begin{aligned}
\partial_{x} \Psi_{0}^{1}(x, t) & =\frac{-\frac{x}{1-t} \phi(x, 1-t) \Phi(x, 1-t)-(\phi(x, 1-t))^{2}}{(\Phi(x, 1-t))^{2}} \\
& =-\frac{x}{1-t} \Psi_{0}^{1}(x, t)-\left(\Psi_{0}^{1}(x, t)\right)^{2} \\
& =-\frac{1}{1-t}\left\{\frac{x}{\sqrt{1-t}} \Psi_{0}^{1}\left(\frac{x}{\sqrt{1-t}}, 0\right)+\left(\Psi_{0}^{1}\left(\frac{x}{\sqrt{1-t}}, 0\right)\right)^{2}\right\} .
\end{aligned}
$$

Fix $t^{*}<1$; then

$$
\sup _{t \in\left[0, t^{*}\right], x \in \mathbb{R}}\left|\partial_{x} \Psi_{0}^{1}(x, t)\right| \leq \frac{1}{1-t^{*}} \sup _{y \in \mathbb{R}}\left|y \Psi_{0}^{1}(y, 0)+\left(\Psi_{0}^{1}(y, 0)\right)^{2}\right| .
$$

Applying l'Hôpital's rule, it can be shown that

$$
\lim _{y \rightarrow-\infty} y \Psi_{0}^{1}(y, 0)+\left(\Psi_{0}^{1}(y, 0)\right)^{2}=1, \quad \lim _{y \rightarrow \infty} y \Psi_{0}^{1}(y, 0)+\left(\Psi_{0}^{1}(y, 0)\right)^{2}=0
$$

which entails that $\sup _{t \in\left[0, t^{*}\right], x \in \mathbb{R}}\left|\partial_{x} \Psi_{0}^{1}(x, t)\right|<\infty$. Therefore, $\Psi_{a}^{1}(x, t)$ is Lipschitz in the $x$ variable uniformly in $t \in\left[0, t^{*}\right], t^{*}<1$. To study the growth of $\Psi_{a}^{1}(x, t)$ we take $a=$ 0 . Then, $\Psi_{0}^{1}(x, t) \leq \frac{1}{\sqrt{1-t^{*}}} \sup _{y \in \mathbb{R}} \Psi_{0}^{1}(y, 0)$ for $t \in\left[0, t^{*}\right], t^{*}<1$. It can be shown that
$\lim _{y \rightarrow-\infty} \Psi_{0}^{1}(y, 0) / y=-1$ and $\lim _{y \rightarrow \infty} \Psi_{0}^{1}(y, 0)=0$, which implies that $\sup _{y \in \mathbb{R}} \Psi_{0}^{1}(y, 0)<\infty$. Hence, $\Psi_{a}^{1}(x, t)$ satisfies a linear growth condition for $t \in\left[0, t^{*}\right], t^{*}<1$. Using the classical results on SDEs, we have that there exists a unique strong solution to the following equation:

$$
Y_{t}^{1}=\int_{0}^{t} \Psi_{a}^{1}\left(Y_{s}^{1}, s\right) d s+B_{t}, \quad 0 \leq t<1 .
$$

We can use a similar reasoning for $\Psi_{a}^{2}(x, t) \triangleq \phi(x-a, 1-t) / \Phi(a-x, 1-t)$ and get the same conclusions. Finally, the $\mathbb{F}^{B} \vee \sigma(G)$-adapted process $X_{t} \triangleq Y_{t}^{1} 1_{\{1\}}(G)+Y_{t}^{2} 1_{\{0\}}(G)$ solves our problem.

Theorem 6.8. Let $L(Y)=\mathbf{1}_{[a, \infty)}\left(Y_{1}\right)$. Then

$$
\left(Y^{*}, \theta^{*}, Z^{*}, H^{*}, \xi^{*}, \lambda^{*}\right)=\left(W, \alpha\left(\cdot, \mathbf{1}_{[a, \infty)}\left(W_{1}\right)\right), W^{a}, H, H\left(1, W_{1}\right), L(W)\right)
$$

is an $(L, \mu)$-weak equilibrium.
Proof. We apply Theorem 5.1. The first hypothesis of the theorem is assumed. The second hypothesis follows from Theorem 6.5. Finally the fact that $\alpha \in \Theta_{\text {sup }}\left(\mathbf{1}_{[a, \infty)}\left(W_{1}\right), W^{a}\right)$ follows from Lemma 6.6, Proposition 5.4, and Theorem 6.7 (see Remark 5.2).
7. The maximum and its argument. In this section we deal with two examples that are more complicated but by far more interesting. In particular, the second example is new in the literature of insider trading with initial strong information. Throughout this section we will consider a Brownian motion $W$ defined on a complete probability space $(\Omega, \mathcal{F}, P)$. We consider the maximum process in the interval $[s, t], M_{s, t}, 0 \leq s<t \leq 1$, defined by $M_{s, t} \triangleq \max _{s \leq u \leq t} W_{u}$. To simplify notation we use $M_{t} \triangleq M_{0, t}, \tau_{t} \triangleq \arg \max _{0 \leq s \leq t} W_{s}, M \triangleq M_{1}$, $\tau \triangleq \tau_{1}$, and $\gamma_{s, t} \triangleq M_{s, t}-W_{s}$. The density and distribution functions of $\gamma_{s, t}$ are given by $p_{2}(x, t-s) \triangleq 2 \phi(x, t-s) \mathbf{1}_{(0, \infty)}(x)$ and $\Pi_{2}(x, t-s) \triangleq \int_{0}^{x} p_{2}(z, t-s) d z$. Similarly, the density of the random vector ( $\gamma_{s, t}, W_{t}-W_{s}$ ) is given by

$$
p_{1}(x, y, t-s) \triangleq \frac{2(2 x-y)}{\sqrt{2 \pi(t-s)^{3}}} \exp \left\{\frac{-(2 x-y)^{2}}{2(t-s)}\right\} \mathbf{1}_{(0, \infty) \times(-\infty, x)}(x, y)
$$

Let us recall a theorem by Lévy that links the maximum process $M_{t}$ with the Brownian local time $L_{t}^{x}(W)$.

Theorem 7.1. The pairs of processes $\left\{\left(M_{t}-W_{t}, M_{t}\right) ; 0 \leq t<\infty\right\}$ and $\left\{\left(\left|W_{t}\right|, 2 L_{t}^{0}(W)\right)\right.$; $0 \leq t<\infty\}$ have the same laws under $P$.

For more details, see [13, Chapter 3, Theorem 6.17]. Furthermore, it is easy to show that, for a fixed $t, M_{t}-W_{t} \sim \gamma_{0, t}$. Finally, we set

$$
\varphi(x, t) \triangleq \frac{p_{2}(x, t)}{\Pi_{2}(x, t)}=\frac{e^{-\frac{x^{2}}{2 t}}}{\int_{0}^{x} e^{-\frac{y^{2}}{2 t}} d y} \mathbf{1}_{(0, \infty)}(x)
$$

7.1. $L(\boldsymbol{Y})=\max _{t \in[0,1]} \boldsymbol{Y}_{t}$. In this subsection we consider the case in which the insider knows the maximum of the total demand. A more general version of the following result is proved in Jeulin [10] (see Proposition 3.24, page 49). See also Mansuy and Yor [18] for an updated reference on enlargement of filtrations theory.

Theorem 7.2. Let $W$ be a Brownian motion. Then $W$ is a semimartingale with respect to the filtration $\mathbb{F}^{W} \vee \sigma(M)$ with decomposition

$$
W_{t}=\int_{0}^{t} \alpha(u, M) d u+W_{t}^{M} \quad \forall t \in[0,1],
$$

where $W^{M}$ is an $\mathbb{F}^{W} \vee \sigma(M)$-Brownian motion,

$$
\alpha(t, M)=\frac{M-W_{t}}{1-t} \mathbf{1}_{\left\{M_{t}<M\right\}}-\varphi\left(M-W_{t}, 1-t\right) \mathbf{1}_{\left\{M_{t}=M\right\}} .
$$

Note that $\mathbf{1}_{\left\{M_{t}<M\right\}}=\mathbf{1}_{[0, \tau)}(t)$.
Lemma 7.3. We have that $\mathbb{E}\left[\int_{0}^{1}|\alpha(t, M)| d t\right]<\infty$ and $\mathbb{E}\left[\int_{0}^{1}|\alpha(t, M)|^{2} d t\right]=\infty$.
Proof. To deduce that $\mathbb{E}\left[\int_{0}^{1}|\alpha(t, M)| d t\right]<\infty$, notice that $\mathbb{E}\left[\int_{0}^{1} \alpha(t, M) d t\right]=\mathbb{E}\left[W_{1}\right]-$ $\mathbb{E}\left[W_{1}^{M}\right]=0$, which implies

$$
\mathbb{E}\left[\int_{0}^{1} \mathbf{1}_{\left\{M_{t}<M\right\}} \frac{M-W_{t}}{1-t} d t\right]=\mathbb{E}\left[\int_{0}^{1} \mathbf{1}_{\left\{M_{t}=M\right\}} \varphi\left(M-W_{t}, 1-t\right) d t\right] .
$$

As the integrands in the above expectations are positive, the problem is reduced to show that $\mathbb{E}\left[\int_{0}^{1} \mathbf{1}_{\left\{M_{t}<M\right\}} \frac{M-W_{t}}{1-t} d t\right]<\infty$. Let us compute this expectation:

$$
\begin{aligned}
\mathbb{E}\left[\int_{0}^{1} \mathbf{1}_{\left\{M>M_{t}\right\}} \frac{M-W_{t}}{1-t} d t\right] & =\mathbb{E}\left[\int_{0}^{1} \mathbf{1}_{\left\{M_{t, 1}>M_{t}\right\}} \frac{M_{t, 1}-W_{t}}{1-t} d t\right] \\
& =\mathbb{E}\left[\int_{0}^{1} \mathbf{1}_{\left\{\gamma_{t, 1}>M_{t}-W_{t}\right\}} \frac{\gamma_{t, 1}}{1-t} d t\right]
\end{aligned}
$$

Conditioning with respect to $\mathcal{F}_{t}^{W}$ and using Lemma 9.7, this expectation is equal to

$$
\int_{0}^{1} \int_{0}^{\infty} \int_{y}^{\infty} \frac{x}{1-t} p_{2}(x, 1-t) p_{2}(y, t) d x d y d t=\sqrt{\frac{2}{\pi}}<\infty
$$

To show the divergence of the second moment, notice that

$$
\begin{aligned}
& \mathbb{E}\left[\int_{0}^{1} \alpha(t, M)^{2} d t\right] \\
& =\mathbb{E}\left[\int_{0}^{1} \mathbf{1}_{\left\{M>M_{t}\right\}}\left(\frac{M-W_{t}}{1-t}\right)^{2} d t\right]+\mathbb{E}\left[\int_{0}^{1} \mathbf{1}_{\left\{M=M_{t}\right\}}\left(\varphi\left(M-W_{t}, 1-t\right)\right)^{2} d t\right]
\end{aligned}
$$

Therefore, it suffices to show the divergence of one of the above expectations. The second expectation above is equal to

$$
\begin{aligned}
& \mathbb{E}\left[\int_{0}^{1} \mathbf{1}_{\left\{M_{t}>M_{t, 1}\right\}}\left(\varphi\left(M_{t}-W_{t}, 1-t\right)\right)^{2} d t\right] \\
& =\mathbb{E}\left[\int_{0}^{1} \int_{0}^{M_{t}-W_{t}}\left(\varphi\left(M_{t}-W_{t}, 1-t\right)\right)^{2} p_{2}(x, 1-t) d x d t\right] \\
& =\int_{0}^{1} \int_{0}^{\infty} \int_{0}^{y}(\varphi(y, 1-t))^{2} p_{2}(x, 1-t) p_{2}(y, t) d x d y d t \\
& =\int_{0}^{1} \int_{0}^{\infty} \frac{\left(p_{2}(y, 1-t)\right)^{2}}{\Pi_{2}(y, 1-t)} p_{2}(y, t) d y d t
\end{aligned}
$$

But this integral is infinite, because

$$
\lim _{y \rightarrow 0^{+}} y \frac{\left(p_{2}(y, 1-t)\right)^{2}}{\Pi_{2}(y, 1-t)} p_{2}(y, t)=p_{2}(0,1-t) p_{2}(0, t)=\frac{2}{\pi \sqrt{t(1-t)}} \neq 0
$$

and this implies that $\int_{0}^{\infty} \frac{\left(p_{2}(y, 1-t)\right)^{2}}{\Pi_{2}(y, 1-t)} p_{2}(y, t) d y=\infty$ for all $t \in[\varepsilon, 1-\varepsilon]$, which is a set of positive Lebesgue measure provided $\varepsilon<1 / 2$.

In order to verify that $\alpha(\cdot, M)$ is $\mathbb{F}^{W^{M}} \vee \sigma(M)$-adapted we prove that $W$ is $\mathbb{F}^{W^{M}} \vee \sigma(M)$ adapted, which follows from the following result.

Theorem 7.4. Let $B$ be a Brownian motion and $G$ be a positive random variable independent of $B$, both defined in the same probability space $(\Omega, \mathcal{F}, P)$. Then there exists a unique strong solution $X$ adapted to the filtration $\mathbb{F}^{B} \vee \sigma(G)$ of the following $S D E$ :

$$
\begin{equation*}
X_{t}=\int_{0}^{t}\left(\frac{G-X_{s}}{1-s} \mathbf{1}_{\left\{M_{s}^{X}<G\right\}}-\varphi\left(G-X_{s}, 1-s\right) \mathbf{1}_{\left\{M_{s}^{X}=G\right\}}\right) d s+B_{t} \tag{7.1}
\end{equation*}
$$

where $M_{t}^{X} \triangleq \max _{0 \leq s \leq t} X_{s}$.
Proof. Our approach to the solution of (7.1) is to write $X_{t}=X_{t}^{1} \mathbf{1}_{[0, \rho)}(t)+X_{t}^{2} \mathbf{1}_{[\rho, 1)}(t)$, where $\rho \triangleq \inf \left\{t: X_{t}^{1}=G\right\}$, and $X_{t}^{1}$ and $X_{t}^{2}$ are the solutions to the following SDEs:

$$
\mathcal{E}_{1}: \quad X_{t}^{1}=\int_{0}^{t} \frac{G-X_{s}^{1}}{1-s} d s+B_{t}, \quad 0 \leq t<\rho
$$

and

$$
\mathcal{E}_{2}: \quad X_{t}^{2}=G-\int_{\rho}^{t} \varphi\left(G-X_{s}^{2}, 1-s\right) d s+B_{t}-B_{\rho}, \quad \rho \leq t<1
$$

we denote them by $\mathcal{E}_{1}$ and $\mathcal{E}_{2}$, respectively. The next step is to show the existence and uniqueness of the solutions to $\mathcal{E}_{1}$ and $\mathcal{E}_{2}$. Note that $\rho \leq 1$ is an $\mathbb{F}^{B} \vee \sigma(G)$-stopping time.

- Existence and uniqueness for the solution of $\mathcal{E}_{1}$ : This follows as in the case of the Brownian bridge (see Lemma 6.9 on page 358 of [13]).
- Existence and uniqueness for the solution of $\mathcal{E}_{2}$ : Note that the drift has a singularity at $t=\rho$. That is, $\lim _{x \rightarrow 0^{+}} \varphi(x, t)=\infty, t>0$. Instead of proving the existence and uniqueness for $\mathcal{E}_{2}$, we will prove them for the following equivalent SDE :

$$
\mathcal{E}_{2}^{\prime}: \quad R_{t}=\int_{0}^{t} \varphi\left(R_{s}, 1-\rho-s\right) d s+N_{t}, \quad 0 \leq t<1-\rho
$$

The $\operatorname{SDE} \mathcal{E}_{2}^{\prime}$ is obtained from $\mathcal{E}_{2}$ through the change of variables $R_{t}=G-X_{t+\rho}^{2}$ and $N_{t}=-\left(B_{t+\rho}-B_{\rho}\right)$. The existence is proved in Proposition 7.5. To prove the uniqueness, we may consider $\Delta_{t} \triangleq R_{t}^{1}-R_{t}^{2}$ the difference of two positive solutions $R^{1}$ and $R^{2}$ of $\mathcal{E}_{2}^{\prime}$. Then, applying Itô's formula to $\Delta_{t \wedge(1-\rho)}^{2}$, we obtain that $P$-a.s.

$$
\begin{aligned}
& \Delta_{t \wedge(1-\rho)}^{2} \\
& =2 \int_{0}^{t \wedge(1-\rho)}\left(R_{s}^{1}-R_{s}^{2}\right)\left(\varphi\left(R_{s}^{1}, 1-\rho-s\right)-\varphi\left(R_{s}^{2}, 1-\rho-s\right)\right) d s \leq 0
\end{aligned}
$$

as $\varphi(x, t) \geq \varphi(y, t)$ if $x \leq y$ for all $t \in(0,1)$.

Proposition 7.5. There exists a positive, continuous, strong solution with respect to $\mathbb{F}^{N} \vee$ $\sigma(\rho)$ to

$$
\begin{equation*}
R_{t}=\int_{0}^{t} \varphi\left(R_{s}, 1-\rho-s\right) d s+N_{t}, \quad 0 \leq t<1-\rho, \tag{7.2}
\end{equation*}
$$

where $N$ is a Wiener process and $\rho \in(0,1)$ is a random variable independent of $\mathbb{F}^{N}$.
Proof. First note that

$$
\begin{equation*}
x \varphi(x, t) \leq 1 \quad \forall t>0, \quad x \in \mathbb{R} . \tag{7.3}
\end{equation*}
$$

We define $\varphi^{n}(x, t) \triangleq \exp \left\{-\frac{1}{n}\left(\frac{1}{x}+\frac{1}{t}\right)\right\} \varphi(x, t)$, which satisfies (7.3) with $\varphi^{n}$ instead of $\varphi$. This sequence of functions is monotone increasing in $n$ and bounded and converges to $\varphi(x, t)$ for each $x \in \mathbb{R}, t>0$ such that $x^{-1}+t^{-1}>0$. Furthermore,

$$
\begin{aligned}
\partial_{x} \varphi^{n}(x, t) & =\exp \left\{-\frac{1}{n}\left(\frac{1}{x}+\frac{1}{t}\right)\right\}\left(\frac{1}{n} x^{-2} \varphi(x, t)+\partial_{x} \varphi(x, t)\right) \\
& =\left(\frac{1}{n x^{2}}-\frac{x}{t}-\varphi(x, t)\right) \varphi^{n}(x, t) .
\end{aligned}
$$

Using inequality (7.3), one obtains that $\sup _{x \in[0, \infty), t \in[0,1]}\left|\partial_{x} \varphi^{n}(x, t)\right|<\infty$, which implies that $\varphi^{n}(x, t)$ is a Lipschitz function. Therefore, for a fixed $n \in \mathbb{N}$, we have the existence and uniqueness of the solutions for the following SDE:

$$
R_{t}^{n}=\int_{0}^{t} \varphi^{n}\left(R_{s}^{n}, 1-\rho-s\right) d s+N_{t}, \quad 0 \leq t<1-\rho .
$$

By a comparison theorem, we have that $P\left(R_{t}^{n+1} \geq R_{t}^{n} ; 0 \leq t<1-\rho\right)=1$, which shows that $R_{t} \triangleq \lim _{n \rightarrow \infty} R_{t}^{n}, 0 \leq t<1-\rho$, exists a.s. in $(-\infty, \infty]$ and is a measurable process, as it is a limit of measurable processes. Now, we show that for $t \in[0,1-\rho), R_{t}<\infty, P$-a.s. and $R$ satisfies equation $\mathcal{E}_{2}^{\prime}$. In order to prove the first property, we show the uniform integrability in $n \in \mathbb{N}$ of $R_{t}^{n}, 0 \leq t<1-\rho$. Applying Itô's formula, we obtain

$$
\left(R_{t}^{n}\right)^{2}=t+2 \int_{0}^{t} R_{s}^{n} \varphi^{n}\left(R_{s}^{n}, 1-\rho-s\right) d s+2 \int_{0}^{t} R_{s}^{n} d N_{s}, \quad 0 \leq t<1-\rho .
$$

Next, we bound the expectation of the second term above. We obtain

$$
\begin{aligned}
& \mathbb{E}\left[\int_{0}^{t \wedge(1-\rho)} R_{s}^{n} \varphi^{n}\left(R_{s}^{n}, 1-\rho-s\right) d s\right] \leq \mathbb{E}\left[\int_{0}^{t \wedge(1-\rho)}\left|R_{s}^{n}\right| \varphi\left(R_{s}^{n}, 1-\rho-s\right) d s\right] \\
& =\mathbb{E}\left[\int_{0}^{t \wedge(1-\rho)} \mathbf{1}_{\left\{R_{s}^{n}>0\right\}} R_{s}^{n} \varphi\left(R_{s}^{n}, 1-\rho-s\right) d s\right] \leq \mathbb{E}\left[\int_{0}^{t \wedge(1-\rho)} \mathbf{1}_{\left\{R_{s}^{n}>0\right\}} d s\right] \leq t .
\end{aligned}
$$

For the third term, one has $\mathbb{E}\left[\int_{0}^{t \wedge(1-\rho)} R_{u}^{n} d N_{u}\right]=0$. Thus, $\sup _{n \in \mathbb{N}} \mathbb{E}\left[\left[\left(R_{t \wedge(1-\rho)}^{n}\right)^{2}\right]\right] \leq 3 t$. This implies the uniform integrability of $R_{t \wedge(1-\rho)}^{n}$, and therefore $R_{t \wedge(1-\rho)} \in L^{1}(\Omega)$. Next, we show that $R_{t}$ satisfies $\mathcal{E}_{2}^{\prime}$. First note that

$$
R_{t \wedge(1-\rho)}=\lim _{n \rightarrow \infty} R_{t \wedge(1-\rho)}^{n}=\lim _{n \rightarrow \infty} \int_{0}^{t \wedge(1-\rho)} \varphi^{n}\left(R_{s}^{n}, 1-\rho-s\right) d s+N_{t \wedge(1-\rho)} .
$$

To conclude the proof we show that

$$
\lim _{n \rightarrow \infty} \int_{0}^{t \wedge(1-\rho)} \varphi^{n}\left(R_{s}^{n}, 1-\rho-s\right) d s=\int_{0}^{t \wedge(1-\rho)} \varphi\left(R_{s}, 1-\rho-s\right) d s
$$

$0 \leq t<1$, with probability 1 . This will also give the continuity for the paths of $R$. Fix $\varepsilon>0$, and define

$$
\begin{aligned}
& \rho_{0}^{\varepsilon} \triangleq \inf \left\{t \in(0,1-\rho): N_{t}=\varepsilon\right\} \\
& \rho_{l}^{\varepsilon} \triangleq \inf \left\{t \in\left(\rho_{l-1}^{\varepsilon}, 1-\rho\right): N_{t}-N_{\rho_{l-1}^{\varepsilon}}=-R_{\rho_{l-1}^{\varepsilon}}^{1} / 2\right\}, \quad l \geq 1 .
\end{aligned}
$$

By construction, the sequence $\left\{\rho_{l}^{\varepsilon}\right\}_{l \in \mathbb{N}}$ is nondecreasing, and therefore we can define $\sigma^{\varepsilon} \triangleq$ $\lim _{l \rightarrow \infty} \rho_{l}^{\varepsilon}$. For fixed $\omega \in \Omega$, we apply the dominated convergence theorem in each interval $\left[\rho_{l-1}^{\varepsilon}, \rho_{l}^{\varepsilon}\right], l \geq 1$. One has that

$$
R_{t}^{1}=R_{\rho_{l-1}^{\varepsilon}}^{1}+\int_{\rho_{l-1}^{\varepsilon}}^{t} \varphi^{1}\left(R_{s}^{1}, 1-\rho-s\right) d s+N_{t}-N_{\rho_{l-1}^{\varepsilon}}>\frac{R_{\rho_{l-1}^{\varepsilon}}^{1}}{2} \geq \frac{\varepsilon}{2^{l}}
$$

for $t \in\left[\rho_{l-1}^{\varepsilon}, \rho_{l}^{\varepsilon}\right]$ and $l \geq 1$, due to the positivity of the integral. Then, using inequality (7.3), we have for $s \in\left[\rho_{l-1}^{\varepsilon}, \rho_{l}^{\varepsilon}\right]$ that

$$
\varphi^{n}\left(R_{s}^{n}, 1-\rho-s\right) \leq \varphi\left(R_{s}^{n}, 1-\rho-s\right) \leq \varphi\left(R_{s}^{1}, 1-\rho-s\right) \leq \frac{1}{R_{s}^{1}} \leq \frac{2^{l}}{\varepsilon}
$$

Hence, by the dominated convergence theorem

$$
\lim _{n \rightarrow \infty} \int_{\rho_{l-1}^{\varepsilon}}^{\rho_{l}^{\varepsilon}} \varphi^{n}\left(R_{s}^{n}, 1-\rho-s\right) d s=\int_{\rho_{l-1}^{\varepsilon}}^{\rho_{l}^{\varepsilon}} \varphi\left(R_{s}, 1-\rho-s\right) d s
$$

This implies that

$$
R_{t}=R_{\rho_{0}^{\varepsilon}}+\int_{\rho_{0}^{\varepsilon}}^{t} \varphi\left(R_{s}, 1-\rho-s\right) d s+N_{t}-N_{\rho_{0}^{\varepsilon}}, \quad \rho_{0}^{\varepsilon} \leq t<\sigma^{\varepsilon} .
$$

We now prove that $\sigma^{\varepsilon}=1-\rho$. If $\omega \in \Omega$ is such that there exists $l$ for which $\rho_{l}^{\varepsilon}=1-\rho$, we have finished. By contradiction, assume that the sequence $\left\{\rho_{l}^{\varepsilon}\right\}_{l \in \mathbb{N}}$ is strictly increasing. First, by the definition of $\left\{\rho_{l}^{\varepsilon}\right\}_{l \in \mathbb{N}}$ and the fact that the sequence is strictly increasing, one has that $N_{\rho_{l}^{\varepsilon}}-N_{\rho_{l-1}^{\varepsilon}}=-R_{\rho_{l-1}^{\varepsilon}}^{1} / 2$. Taking limits we obtain that $R_{\sigma^{\varepsilon}}^{1}=0$, due to the continuity of Brownian paths. Then $R_{t}^{1}+\int_{t}^{\sigma^{\varepsilon}} \varphi^{1}\left(R_{s}^{1}, 1-\rho-s\right) d s=N_{t}-N_{\sigma^{\varepsilon}}$, but this contradicts the law of the iterated logarithm when $t$ tends to $\sigma^{\varepsilon}$, because the left-hand side is positive a.s. for $t \in\left[\rho_{0}^{\varepsilon}, \sigma^{\varepsilon}\right)$. Hence we can conclude that the set of $\omega \in \Omega$ for which there does not exist a finite $l$ such that $\rho_{l}^{\varepsilon}=1-\rho$ is a null set. Now, notice that $\rho_{0}^{\varepsilon} \downarrow 0$ when $\varepsilon \downarrow 0$. Hence, $N_{\rho_{0}^{\varepsilon}} \longrightarrow{ }_{\varepsilon \downarrow 0} 0$ and by monotone convergence $\lim _{\varepsilon \downarrow 0} \int_{\rho_{0}^{\varepsilon}}^{t} \varphi\left(R_{s}, 1-\rho-s\right) d s=\int_{0}^{t} \varphi\left(R_{s}, 1-\rho-s\right) d s, 0 \leq t<1-\rho$. Therefore, $R_{t}=\lim _{\varepsilon \downarrow 0} R_{\rho_{0}^{\varepsilon}}+\int_{0}^{t} \varphi\left(R_{s}, 1-\rho-s\right) d s+N_{t}, 0 \leq t<1-\rho$. As $R_{0}=\lim _{n \rightarrow \infty} R_{0}^{n}=0$, making $t=0$ in the above equation we obtain $\lim _{\varepsilon \downarrow 0} R_{\rho_{0}^{\varepsilon}}=0$. Furthermore, as $\left|R_{t}\right|<\infty$,
$P$-a.s. we obtain that $\int_{0}^{t} \varphi\left(R_{s}, 1-\rho-s\right) d s<\infty, P$-a.s. for $t<\sigma$. Hence we have shown that $R$ satisfies (7.2). Note that, in particular, we have also proved that $R_{t}>0$.

Theorem 7.6. Let $L(Y)=\max _{0 \leq t \leq 1} Y_{t}$. Then

$$
\left(Y^{*}, \theta^{*}, Z^{*}, H^{*}, \xi^{*}, \lambda^{*}\right)=\left(W, \alpha(\cdot, M), W^{M}, H, H\left(1, W_{1}\right), L(W)\right)
$$

satisfies all the requirements to be an $(L, \mu)$-weak equilibrium except the càglàd property in condition (v) of Definition 2.7.

Proof. Properties (i) through (iv) in the definition of weak equilibrium (Definition 2.7) follow directly. Property (v) with the exception of the càglàd property follows from Lemma 7.3, Proposition 5.4, and Theorem 7.4 (see Remark 5.2). From the assumptions on $H$ and $\mu$ and (4.4), we have that $H(\cdot, W$.$) is an \mathbb{F}^{W}$-martingale. As $H\left(t, W_{t}\right)=\mathbb{E}\left[H\left(1, W_{1}\right) \mid \mathcal{F}_{t}^{W}\right]$, property (vi) follows. Let us check property (vii). To simplify the notation we set $\alpha_{t} \triangleq \alpha(t, M)$, $0 \leq t \leq 1$. Note that $\alpha_{t} \geq 0$ if $t \leq \tau$ and $\alpha_{t} \leq 0$ if $t>\tau$. From this property, the following inequality easily follows:

$$
\begin{equation*}
\int_{0}^{1}\left|\alpha_{t}\right| d t \leq \int_{0}^{\tau} \alpha_{t} d t-\int_{\tau}^{1} \alpha_{t} d t \leq 3 \sup _{0 \leq t \leq 1}\left|\int_{0}^{t} \alpha_{s} d s\right| \tag{7.4}
\end{equation*}
$$

which combined with Proposition 5.4 gives that

$$
\begin{equation*}
\int_{0}^{1}\left|\alpha_{t}\right| d t \in L^{p}(\Omega), \quad p \geq 1 \tag{7.5}
\end{equation*}
$$

For $\varepsilon \in(0,1)$, define $\tau^{\varepsilon,+}=(\tau+\varepsilon) \wedge 1$. Then the process $\alpha^{\varepsilon}=\left\{\alpha_{t}^{\varepsilon} \triangleq \alpha_{t} \mathbf{1}_{\left(\tau, \tau^{\varepsilon},+\mathrm{j}\right.}(t)\right.$, $t \in[0,1]\}$ converges $P \times \lambda$-a.e. to $\alpha$ as $\varepsilon \downarrow 0$ and satisfies $\left|\alpha^{\varepsilon}\right| \leq|\alpha|$. Now we will prove that $\alpha^{\varepsilon} \in \Theta_{\text {sup }}\left(M, W^{M}\right)$ for all $\varepsilon \in(0,1)$. First, the càglàd property of $\alpha^{\varepsilon}$ follows from the fact that this approximation avoids the essential discontinuity of $\alpha_{t}$ in $t=\tau$. The integrability property (2.1) is trivial. Property (2.4) follows from (7.4). The proofs of properties (2.2) and (2.3) are similar. We will prove property (2.2). We have that $\sup _{0 \leq t \leq 1}\left|\int_{0}^{t} H\left(s, Y_{s}^{\alpha^{\varepsilon}}\right) \alpha_{s}^{\varepsilon} d s\right| \leq$ $\sup _{0 \leq t \leq 1}\left|H\left(t, Y_{t}^{\alpha^{\varepsilon}}\right)\right| \int_{0}^{1}\left|\alpha_{t}\right| d t$, which belongs to $L^{1}(\Omega)$ by the Cauchy-Schwarz inequality, property (7.5), and Lemma 9.1. According to Proposition 9.5, $\lim _{n \rightarrow \infty} J\left(\alpha^{\varepsilon, n}\right)=J\left(\alpha^{\varepsilon}\right)$ for all $\varepsilon \in(0,1)$, where $\alpha^{\varepsilon, n}$ is defined according to Definition 9.4 with $\theta=\alpha^{\varepsilon}$. As the functional $J$ is concave in $\Theta_{b}\left(M, W^{M}\right)$, we obtain that $J(\eta) \leq J\left(\alpha^{\varepsilon, n}\right)+D_{\eta-\alpha^{\varepsilon, n}} J\left(\alpha^{\varepsilon, n}\right)$ for $\eta \in \Theta_{b}\left(M, W^{M}\right)$.

- $\lim _{\varepsilon \downarrow 0} J\left(\alpha^{\varepsilon}\right)=J(\alpha)$ : This is analogous to the proof of Proposition 9.5. Note that using property (7.4), we have that

$$
\left(\int_{0}^{1} \alpha_{t}-\alpha_{t}^{\varepsilon} d t\right)^{2} \leq C \sup _{0 \leq t \leq 1}\left(\int_{0}^{1} \alpha_{t}\right)^{2}
$$

and

$$
\begin{aligned}
& \left|\int_{0}^{1} H\left(t, Y_{t}^{\alpha}\right) \alpha_{t}-H\left(t, Y_{t}^{\alpha^{\varepsilon}}\right) \alpha_{t}^{\varepsilon} d t\right| \\
& \leq\left|\int_{0}^{1} H\left(t, Y_{t}^{\alpha}\right)\left(\alpha_{t}-\alpha_{t}^{\varepsilon}\right) d t\right|+\left|\int_{0}^{1}\left(H\left(t, Y_{t}^{\alpha}\right)-H\left(t, Y_{t}^{\alpha^{\varepsilon}}\right)\right) \alpha_{t}^{\varepsilon} d t\right|
\end{aligned}
$$

$$
\begin{aligned}
& \leq C \sup _{0 \leq t \leq 1}\left|H\left(t, Y_{t}^{\alpha}\right)\right| \int_{0}^{1}\left|\alpha_{t}\right| d t \\
& +C \sup _{0 \leq t \leq 1} \int_{0}^{1}\left|H_{y}\left(t, Y_{t}^{\alpha^{\varepsilon}+r\left(\alpha-\alpha^{\varepsilon}\right)}\right)\right| d r\left(\int_{0}^{1}\left|\alpha_{t}\right| d t\right)^{2}
\end{aligned}
$$

This gives sufficient integrability properties to apply the dominated convergence theorem. Note that as in the proof of Lemma 9.1,

$$
\begin{equation*}
\sup _{0 \leq t \leq 1}\left|H_{y}\left(t, Y_{t}^{\alpha^{\varepsilon}+r\left(\alpha-\alpha^{\varepsilon}\right)}\right)\right| \leq C \exp \left\{9 B \sup _{0 \leq t \leq 1}\left|\int_{0}^{t} \alpha_{s} d s\right|\right\} \exp \left\{B \sup _{0 \leq t \leq 1}\left|Z_{t}\right|\right\} \tag{7.6}
\end{equation*}
$$

- $\lim _{\varepsilon \downarrow 0} \lim _{n \rightarrow \infty} D_{\eta-\alpha^{\varepsilon, n}} J\left(\alpha^{\varepsilon, n}\right)=0$ : Repeating the proof of Proposition 9.6, we obtain $\mid \overline{D_{\eta-\alpha^{\varepsilon, n}} J\left(\alpha^{\varepsilon, n}\right) \mid \leq B_{1}^{\varepsilon, n}+B_{2}^{\varepsilon, n}, \text { where }}$

$$
B_{1}^{\varepsilon, n} \triangleq\left|\mathbb{E}\left[\int_{0}^{1}\left(\eta_{t}-\alpha_{t}^{\varepsilon, n}\right)\left(H\left(t, Y_{t}^{\alpha}\right)-H\left(t, Y_{t}^{\alpha^{\varepsilon, n}}\right)\right) d t\right]\right|
$$

and

$$
B_{2}^{\varepsilon, n} \triangleq\left|\mathbb{E}\left[\int_{0}^{1}\left(\int_{0}^{t}\left(\eta_{s}-\alpha_{s}^{\varepsilon, n}\right) d s\right)\left(H_{y}\left(t, Y_{t}^{\alpha}\right) \alpha_{t}-H_{y}\left(t, Y_{t}^{\alpha^{\varepsilon, n}}\right) \alpha_{t}^{\varepsilon, n}\right) d t\right]\right|
$$

Let us show that $\lim _{\varepsilon \downarrow 0} \lim _{n \rightarrow \infty} B_{1}^{\varepsilon, n}=0$. This follows by dominated convergence, once we have shown that

$$
\sup _{\varepsilon, n} \int_{0}^{1}\left(\eta_{t}-\alpha_{t}^{\varepsilon, n}\right)\left(H\left(t, Y_{t}^{\alpha}\right)-H\left(t, Y_{t}^{\alpha^{\varepsilon, n}}\right)\right) d t \in L^{1}(\Omega),
$$

because $\lim _{n \rightarrow \infty} \alpha^{\varepsilon, n}=\alpha^{\varepsilon}, P \times \lambda$-a.s. and $\lim _{\varepsilon \downarrow 0} \alpha^{\varepsilon}=\alpha, P \times \lambda$-a.s. Using inequalities (7.4) and (7.6) we obtain

$$
\begin{aligned}
& \left|\int_{0}^{1}\left(\eta_{t}-\alpha_{t}^{\varepsilon, n}\right)\left(H\left(t, Y_{t}^{\alpha}\right)-H\left(t, Y_{t}^{\alpha^{\varepsilon, n}}\right)\right) d t\right| \\
& \leq \int_{0}^{1}\left|\eta_{t}-\alpha_{t}^{\varepsilon, n}\right| \int_{0}^{1} H_{y}\left(t, Y_{t}^{\alpha^{\varepsilon, n}+r\left(\alpha-\alpha^{\varepsilon, n}\right)}\right) d r\left|Y_{t}^{\alpha}-Y_{t}^{\alpha^{\varepsilon, n}}\right| d t \\
& \leq\left\{C+\int_{0}^{1}\left|\alpha_{t}\right| d t\right\} \sup _{0 \leq t \leq 1} \int_{0}^{1} H_{y}\left(t, Y_{t}^{\alpha^{\varepsilon, n}+r\left(\alpha-\alpha^{\varepsilon, n}\right)}\right) d r \int_{0}^{1}\left|\alpha_{t}-\alpha_{t}^{\varepsilon, n}\right| d t
\end{aligned}
$$

which is in $L^{1}(\Omega)$, because as in Lemma 9.1, $\sup _{0 \leq t \leq 1}\left|\int_{0}^{t} \alpha_{s} d s\right|$ and $\sup _{0 \leq t \leq 1}\left|Z_{t}\right|$ have exponential moments. The proof of $\lim _{\varepsilon \downarrow 0} \lim _{n \rightarrow \infty} \bar{B}_{2}^{\varepsilon, n}=0$ can be obtained similarly. Therefore, we have proved that $J(\eta) \leq J(\alpha)$ for all $\eta \in \Theta_{b}\left(M, W^{M}\right)$. The final result follows from the application of Proposition 9.5, using an argument as in the end of the proof of Theorem 3.7.
7.2. $L(\boldsymbol{Y})=\arg \max _{t \in[0,1]} \boldsymbol{Y}_{t}$. In this section we consider the case in which the insider knows the time at which the total demand achieves its maximum. The first part of this subsection is devoted to obtaining the compensator of $W$ with respect to the filtration $\mathbb{F}^{W} \vee \sigma(\tau)$, which we will denote by $\mathbb{F}^{\tau}=\left\{\mathcal{F}_{t}^{\tau}, 0 \leq t \leq 1\right\}$. This will be done dividing the problem into two parts: before the random time $\tau$ and after it. But first, we give the conditional law of $\tau$ given $\mathcal{F}_{t}^{W}$.

Proposition 7.7. The conditional law of $\tau$, given $\mathcal{F}_{t}^{W}$, is

$$
P\left(\tau>u \mid \mathcal{F}_{t}^{W}\right)=\left\{\begin{array}{cl}
1-\mathbf{1}_{\left\{M_{u} \geq \gamma_{u, t}+W_{u}\right\}} \bar{p}_{2}\left(M_{u}-W_{t}, 1-t\right) & \text { if } u<t \\
\int_{u}^{1} r\left(M_{t}-W_{t}, v-t, 1-v\right) d v & \text { if } u \geq t
\end{array}\right.
$$

where $r(x, s, t)$ is given by $r(x, s, t) \triangleq \frac{2}{\sqrt{2 \pi t}} \phi(x, s) \mathbf{1}_{(0, \infty)}(x)=\frac{1}{\sqrt{2 \pi t}} p_{2}(x, s)$. Moreover, $P\left(\tau>u \mid \mathcal{F}_{t}^{W}\right)$ is continuous in $u, P$-a.s.

Proof. If $u<t$, then

$$
\begin{aligned}
& P\left(\tau>u \mid \mathcal{F}_{t}^{W}\right) \\
& =P\left(M_{u}<M_{u, 1} \mid \mathcal{F}_{t}^{W}\right)=P\left(M_{u}<M_{u, t} \vee M_{t, 1} \mid \mathcal{F}_{t}^{W}\right) \\
& =P\left(M_{u}<M_{u, t}, M_{u, t}>M_{t, 1} \mid \mathcal{F}_{t}^{W}\right)+P\left(M_{u}<M_{t, 1}, M_{u, t} \leq M_{t, 1} \mid \mathcal{F}_{t}^{W}\right) \\
& =\mathbf{1}_{\left\{M_{u}<M_{u, t}\right\}} P\left(M_{u, t}>\gamma_{t, 1}+W_{t} \mid \mathcal{F}_{t}^{W}\right)+P\left(\gamma_{t, 1}>\left(M_{u} \vee M_{u, t}\right)-W_{t} \mid \mathcal{F}_{t}^{W}\right) \\
& =\mathbf{1}_{\left\{M_{u}<M_{u, t}\right\}} \int_{0}^{M_{u, t}-W_{t}} p_{2}(z, 1-t) d z+\int_{\left(M_{u} \vee M_{u, t}\right)-W_{t}}^{\infty} p_{2}(z, 1-t) d z \\
& =\mathbf{1}_{\left\{M_{u}<M_{u, t}\right\}} \int_{0}^{M_{u, t}-W_{t}} p_{2}(z, 1-t) d z+1-\int_{0}^{\left(M_{u} \vee M_{u, t}\right)-W_{t}} p_{2}(z, 1-t) d z \\
& =1-\mathbf{1}_{\left\{M_{u} \geq \gamma_{u, t}+W_{u}\right\}} \int_{0}^{M_{u}-W_{t}} p_{2}(z, 1-t) d z .
\end{aligned}
$$

If $u>t$, the calculations are more involved; the idea is to break the maximum processes into pieces that are independent of $\mathcal{F}_{t}^{W}$ and pieces that are $\mathcal{F}_{t}^{W}$-measurable:

$$
\begin{aligned}
& P\left(\tau>u \mid \mathcal{F}_{t}^{W}\right) \\
& =P\left(M_{u}<M_{u, 1} \mid \mathcal{F}_{t}^{W}\right)=P\left(M_{t} \vee M_{t, u}<M_{u, 1} \mid \mathcal{F}_{t}^{W}\right) \\
& =P\left(M_{t}<M_{u, 1}, M_{t} \geq M_{t, u} \mid \mathcal{F}_{t}^{W}\right)+P\left(M_{t, u}<M_{u, 1}, M_{t}<M_{t, u} \mid \mathcal{F}_{t}^{W}\right) \\
& =P\left(M_{t}-W_{t}<\gamma_{u, 1}+W_{u}-W_{t}, M_{t}-W_{t} \geq \gamma_{t, u} \mid \mathcal{F}_{t}^{W}\right) \\
& +P\left(\alpha_{t, u}<\gamma_{u, 1}+W_{u}-W_{t}, M_{t}-W_{t}<\gamma_{t, u} \mid \mathcal{F}_{t}^{W}\right) .
\end{aligned}
$$

Hence,

$$
\begin{aligned}
& P\left(M_{t}-W_{t}<\gamma_{u, 1}+W_{u}-W_{t}, M_{t}-W_{t} \geq \gamma_{t, u} \mid \mathcal{F}_{t}^{W}\right) \\
& =\int_{0}^{M_{t}-W_{t}} \int_{-\infty}^{x} \int_{M_{t}-W_{t}-y}^{\infty} p_{1}(x, y, u-t) p_{2}(z, 1-u) d z d y d x
\end{aligned}
$$

$$
\begin{aligned}
& =\int_{0}^{M_{t}-W_{t}} \int_{-\infty}^{x} 2 p_{1}(x, y, u-t) \Phi\left(M_{t}-W_{t}-y, 1-u\right) d y d x \\
& =\int_{-\infty}^{M_{t}-W_{t}} \int_{0 \vee y}^{M_{t}-W_{t}} 2 p_{1}(x, y, u-t) \bar{\Phi}\left(M_{t}-W_{t}-y, 1-u\right) d x d y \\
& =\int_{-\infty}^{M_{t}-W_{t}} 2\left(\phi(|y|, u-t)-\phi\left(2\left(M_{t}-W_{t}\right)-y, u-t\right)\right) \bar{\Phi}\left(M_{t}-W_{t}-y, 1-u\right) d y \\
& =\int_{0}^{\infty} 2\left(\phi\left(\left|M_{t}-W_{t}-z\right|, u-t\right)-\phi\left(M_{t}-W_{t}+z, u-t\right)\right) \bar{\Phi}(z, 1-u) d z
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
& P\left(\gamma_{t, u}<\gamma_{u, 1}+W_{u}-W_{t}, M_{t}-W_{t}<\gamma_{t, u} \mid \mathcal{F}_{t}^{W}\right) \\
& =\int_{M_{t}-W_{t}}^{\infty} \int_{-\infty}^{x} \int_{x-y}^{\infty} p_{1}(x, y, u-t) p_{2}(z, 1-u) d z d y d x \\
& =\int_{M_{t}-W_{t}}^{\infty} \int_{-\infty}^{x} 2 p_{1}(x, y, u-t) \bar{\Phi}(x-y, 1-u) d y d x \\
& =\int_{M_{t}-W_{t}}^{\infty} \int_{0}^{\infty} 2 p_{1}(x, x-z, u-t) \bar{\Phi}(z, 1-u) d z d x \\
& =\int_{0}^{\infty} 4 \phi\left(M_{t}-W_{t}+z, u-t\right) \bar{\Phi}(z, 1-u) d z .
\end{aligned}
$$

Summing up and taking into account that $\phi(|z|, u-t)=\phi(z, u-t)$, we obtain

$$
\begin{aligned}
& P\left(\tau>u \mid \mathcal{F}_{t}^{W}\right) \\
& =\int_{0}^{\infty} 2\left\{\phi\left(M_{t}-W_{t}-z, u-t\right)+\phi\left(M_{t}-W_{t}+z, u-t\right)\right\} \bar{\Phi}(z, 1-u) d z
\end{aligned}
$$

Differentiating under the integral sign, we obtain that there exists a density function $r$ such that $P\left(\tau>u \mid \mathcal{F}_{t}^{W}\right)=\int_{u}^{1} r\left(M_{t}-W_{t}, v-t, 1-v\right) d v$. Furthermore, this density is smooth in all its variables due to the regularity of $\phi$ and $\bar{\Phi}$. For the explicit computation of this density we refer the reader to [14]. To conclude the proof we need only show that $P\left(\tau>u \mid \mathcal{F}_{t}^{W}\right)$, as a function of $u$, is continuous in $u=t$. We have that

$$
\lim _{u \rightarrow t} P\left(\tau>u \mid \mathcal{F}_{t}^{W}\right)=\lim _{u \rightarrow t} P\left(M_{u}<M_{u, 1} \mid \mathcal{F}_{t}^{W}\right)=P\left(M_{t} \leq M_{t, 1} \mid \mathcal{F}_{t}^{W}\right)
$$

where we have used the dominated convergence theorem for conditional expectations and the continuity in $t$ of the paths of $M_{t}$ and $M_{t, 1}, P$-a.s.

Proposition 7.8. If $0 \leq s \leq t \leq 1$, we have that

$$
\mathbb{E}\left[\left.\mathbf{1}_{\{\tau>t\}}\left(W_{t}-\int_{0}^{t} \frac{M_{u}-W_{u}}{\tau-u} d u\right) \right\rvert\, \mathcal{F}_{s}^{\tau}\right]=\mathbf{1}_{\{\tau>t\}}\left(W_{s}-\int_{0}^{s} \frac{M_{u}-W_{u}}{\tau-u} d u\right) .
$$

Proof. Let $A \in \mathcal{F}_{s}^{W}$, and let $f$ be a bounded Borel measurable function; then taking into account that $\tau$ has a conditional density given $\mathcal{F}_{t}^{W}$, in the set $\{\tau>t\}$ we have that

$$
\begin{aligned}
& \mathbb{E}\left[\mathbf{1}_{A} f(\tau) \mathbf{1}_{\{\tau>t\}}\left(W_{t}-W_{s}\right)\right]=\mathbb{E}\left[\mathbf{1}_{A} \mathbb{E}\left[f(\tau) \mathbf{1}_{\{\tau>t\}} \mid \mathcal{F}_{t}^{W}\right]\left(W_{t}-W_{s}\right)\right] \\
& =\mathbb{E}\left[\mathbf{1}_{A} \int_{t}^{1} f(u) r\left(M_{t}-W_{t}, u-t, 1-u\right) d u\left(W_{t}-W_{s}\right)\right] .
\end{aligned}
$$

Applying Theorem 7.1 and Tanaka's formula, we obtain that the last expectation is equal to

$$
\begin{aligned}
& \mathbb{E}\left[\mathbf{1}_{A} \int_{t}^{1} f(u) r\left(\left|W_{t}\right|, u-t, 1-u\right) d u\left(2 \int_{s}^{t} d L_{v}^{0}(W)-\int_{s}^{t} d\left|W_{v}\right|\right)\right] \\
& =\mathbb{E}\left[\mathbf{1}_{A} \int_{t}^{1} f(u) r\left(\left|W_{t}\right|, u-t, 1-u\right) d u\left(-\int_{s}^{t} \operatorname{sgn}\left(W_{v}\right) d W_{v}\right)\right] .
\end{aligned}
$$

Notice that $r\left(\left|W_{t}\right|, u-t, 1-u\right)=\frac{2}{\sqrt{2 \pi(1-u)}} \phi\left(W_{t}, u-t\right)$. Using Itô's formula, we can write

$$
r\left(\left|W_{t}\right|, u-t, 1-u\right)=\frac{2}{\sqrt{2 \pi(1-u)}} \phi(0, u)+\int_{0}^{t} \frac{2}{\sqrt{2 \pi(1-u)}} \partial_{x} \phi\left(W_{v}, u-v\right) d W_{v} .
$$

Then, the former expectation is equal to

$$
\begin{aligned}
& \mathbb{E}\left[\mathbf{1}_{A} \int_{t}^{1} f(u) \int_{s}^{t}-\operatorname{sgn}\left(W_{v}\right) \frac{2}{\sqrt{2 \pi(1-u)}} \partial_{x} \phi\left(W_{v}, u-v\right) d v d u\right] \\
& =\mathbb{E}\left[\mathbf{1}_{A} \int_{t}^{1} f(u) \int_{s}^{t} \frac{\left|W_{v}\right|}{u-v} r\left(\left|W_{v}\right|, u-v, 1-u\right) d v d u\right] \\
& =\mathbb{E}\left[\mathbf{1}_{A} \int_{t}^{1} f(u) \int_{s}^{t} \frac{M_{v}-W_{v}}{u-v} r\left(M_{v}-W_{v}, u-v, 1-u\right) d v d u\right] \\
& =\mathbb{E}\left[\mathbf{1}_{A} \int_{s}^{t} \mathbb{E}\left[\left.\mathbf{1}_{\{\tau>t\}} f(\tau) \frac{M_{v}-W_{v}}{\tau-v} \right\rvert\, \mathcal{F}_{v}\right] d v\right]=\mathbb{E}\left[\mathbf{1}_{A} f(\tau) \mathbf{1}_{\{\tau>t\}} \int_{s}^{t} \frac{M_{v}-W_{v}}{\tau-v} d v\right] .
\end{aligned}
$$

As the $\sigma$-algebra $\mathcal{F}_{s}^{\tau}$ is generated by elements of the form $\mathbf{1}_{A} f(\tau)$, where $A \in \mathcal{F}_{s}^{W}$ and $f$ is a bounded Borel function, we obtain the result using elementary properties of the conditional expectation. Note also that $\left(M_{t}-W_{t}, M_{t}\right)$ and $\left(\left|W_{t}\right|, 2 L_{t}(0)\right)$ are not the same processes. We can interchange them because we are dealing with expectations, and therefore they depend only on the law of the processes, which are equal by Theorem 7.1.

Now, we are going to prove an analogous result for the case after the time $\tau$. In the proof we will use the decomposition of $W$ with respect to $\mathbb{F}^{W} \vee \sigma(M)$ (see Theorem 7.2).

Proposition 7.9. If $0 \leq s \leq t \leq 1$, we have that

$$
\begin{aligned}
& \mathbb{E}\left[\mathbf{1}_{\{\tau \leq s\}}\left(W_{t}+\int_{0}^{t} \varphi\left(M-W_{u}, 1-u\right) d u\right) \mid \mathcal{F}_{s}^{\tau}\right] \\
& =\mathbf{1}_{\{\tau \leq s\}}\left(W_{s}+\int_{0}^{s} \varphi\left(M-W_{u}, 1-u\right) d u\right) .
\end{aligned}
$$

Proof. Let $A \in \mathcal{F}_{s}^{W}$ and $f(\tau)=\mathbf{1}_{\{\tau \leq r\}}$, where $0 \leq r \leq 1$. We have that

$$
\begin{aligned}
& \mathbb{E}\left[\mathbf{1}_{A} f(\tau) \mathbf{1}_{\{\tau \leq s\}}\left(W_{t}-W_{s}\right)\right]=\mathbb{E}\left[\mathbf{1}_{A} \mathbf{1}_{\{\tau \leq r\}} \mathbf{1}_{\{\tau \leq s\}}\left(W_{t}-W_{s}\right)\right] \\
& =\mathbb{E}\left[\mathbf{1}_{A} \mathbf{1}_{\{\tau \leq r \wedge s\}} \mathbf{1}_{\{\tau \leq s\}}\left(W_{t}-W_{s}\right)\right]=\mathbb{E}\left[\mathbf{1}_{A} \mathbf{1}_{\left\{M_{s \wedge r}=M\right\}} \mathbf{1}_{\{\tau \leq s\}}\left(W_{t}-W_{s}\right)\right] \\
& =-\mathbb{E}\left[\mathbf{1}_{A} f(\tau) \mathbf{1}_{\{\tau \leq s\}} \int_{s}^{t} \varphi\left(M-W_{u}, 1-u\right) d u\right] .
\end{aligned}
$$

Notice that $\mathbf{1}_{\{\tau \leq r \wedge s\}}=\mathbf{1}_{\left\{M_{s \wedge r}=M\right\}}$ is $\mathcal{F}_{s}^{W} \vee \sigma(M)$-measurable and that $\varphi\left(M-W_{u}, 1-u\right)$ is $\mathcal{F}_{u}^{\tau}$-measurable because $M=M_{\tau}$. The elements of the form $\mathbf{1}_{A} f(\tau)$, where $A \in \mathcal{F}_{s}^{W}$ and $f(\tau)=\mathbf{1}_{\{\tau \leq r\}}, 0 \leq r \leq 1$, generate the $\sigma$-algebra $\mathcal{F}_{s}^{\tau}$. Therefore as in the proof of the previous proposition we obtain the result using elementary properties of conditional expectations.

The next lemma gives us an integrability result for the drift term in the $\mathbb{F}^{\boldsymbol{\tau}}$-decomposition of $W$.

Lemma 7.10. We have that $\mathbb{E}\left[\int_{0}^{1}|\alpha(t, \tau)| d t\right]<\infty$ and $\mathbb{E}\left[\int_{0}^{1}|\alpha(t, \tau)|^{2} d t\right]=\infty$.
Proof. As in Lemma 7.3, we have that

$$
\mathbb{E}\left[\int_{0}^{1} \mathbf{1}_{[0, \tau)} \frac{M_{t}-W_{t}}{\tau-t} d t\right]=\mathbb{E}\left[\int_{0}^{1} \mathbf{1}_{[\tau, 1]}(t) \varphi\left(M-W_{t}, 1-t\right) d t\right],
$$

where the integrands are positive. The second part of the statement follows as in Lemma 7.3.

The main result of this section is the following theorem, which gives the semimartingale decomposition of $W$ in the filtration $\mathbb{F}^{\tau}$.

Theorem 7.11. $W$ is an $\mathbb{F}^{\tau}$-semimartingale with the following decomposition:

$$
\begin{equation*}
W_{t}=\int_{0}^{t} \alpha(u, \tau) d u+W_{t}^{\tau} \tag{7.7}
\end{equation*}
$$

where $\alpha(u, \tau)=\frac{M_{u}-W_{u}}{\tau-u} \mathbf{1}_{[0, \tau)}(u)-\varphi\left(M-W_{u}, 1-u\right) \mathbf{1}_{[\tau, 1]}(u)$ and $W^{\tau}$ is an $\mathbb{F}^{\tau}$-Brownian motion.

Proof. If we define $W_{t}^{\tau} \triangleq W_{t}-\int_{0}^{t} \alpha(u, \tau) d u$, we have a process in $L^{1}(\Omega)$, because $\mathbb{E}\left[\left|\int_{0}^{t} \alpha(u, \tau) d u\right|\right] \leq \mathbb{E}\left[\int_{0}^{t}|\alpha(u, \tau)| d u\right]<\infty$ by Lemma 7.10. Furthermore, the quadratic variation of $W^{\tau}$ is $t$, because $W$ is an $\mathcal{F}_{t}^{W}$-Brownian motion and $\int_{0}^{\sim} \alpha(u, \tau) d u$ is a process of finite variation. Hence, by Levy's characterization of the Brownian motion, we need only prove that $W^{\tau}$ is an $\mathbb{F}^{\tau}$-local martingale. To show this, using Propositions 7.8 and 7.9 , we obtain the conclusion as in the proof of Theorem 2 in [15].

Theorem 7.12. The Brownian motion $W$ in the decomposition (7.7) is $\mathbb{F}^{W^{\tau}} \vee \sigma(\tau)$-adapted. Proof. First we will show that the following SDE has a unique strong solution:

$$
\mathcal{E}_{1}: \quad X_{t}=\int_{0}^{t} \frac{M_{s}^{X}-X_{s}}{\rho-s} d s+B_{t}, \quad 0 \leq t<\rho
$$

where $B$ is a Brownian motion with respect its own filtration and $\rho$ is a random variable independent of $B$ and taking values in $[0,1]$. We will prove first that $\Psi\left(t, X_{[0, t]}\right)=\frac{M_{t}^{X}-X_{t}}{\tau-t}$ is functional Lipschitz. We have that

$$
\left|\Psi\left(t, X_{[0, t]}\right)-\Psi\left(t, Y_{[0, t]}\right)\right| \leq \frac{1}{\tau-t}\left\{\left|X_{t}-Y_{t}\right|+\left|M_{t}^{X}-M_{t}^{Y}\right|\right\} .
$$

Obviously, $\left|X_{t}-Y_{t}\right| \leq M_{t}^{|X-Y|}$. On the other hand $M_{t}^{X} \leq M_{t}^{X-Y}+M_{t}^{Y}$, which gives that $M_{t}^{X}-M_{t}^{Y} \leq M_{t}^{|X-Y|}$. We also have that $M_{t}^{Y} \leq M_{t}^{Y-X}+M_{t}^{X}$, which yields $M_{t}^{Y}-M_{t}^{X} \leq$ $-M_{t}^{X-Y} \leq-M_{t}^{|X-Y|}$. Hence, $\left|\Psi\left(t, X_{[0, t]}\right)-\Psi\left(t, Y_{[0, t]}\right)\right| \leq \frac{2}{\tau-t} M_{t}^{|X-Y|}$. By Theorem 7 in Chapter V of Protter [19], we obtain the existence and uniqueness of the solutions for $\mathcal{E}_{1}$.

Now, if we take $B_{t}=W_{t}^{\tau}$ and $\rho=\tau$, we obtain that $X$ must coincide with $W_{t}$ for $t<\tau$. This means that $W_{t}$ is $\mathcal{F}_{t}^{W^{\tau}} \vee \sigma(\tau)$-adapted. Furthermore, as $\lim _{t \rightarrow \tau} W_{t}=W_{\tau}$, one also has that $W_{\tau}$ is $\mathcal{F}_{t}^{W^{\tau}} \vee \sigma(\tau)$-adapted.

The next step is to show the existence and uniqueness for the solutions of

$$
\mathcal{E}_{2}: \quad X_{t}=G-\int_{\rho}^{t} \varphi\left(G-X_{s}, 1-s\right) d s+B_{t}-B_{\rho}, \quad \rho \leq t<1
$$

where $B$ is a Brownian motion with respect its own filtration, $\rho$ is a random variable independent of $B$ and taking values in $[0,1]$, and $G$ is an $\mathcal{F}_{\rho}^{B}$-adapted random variable. The existence is proved in Proposition 7.5 and the uniqueness in Theorem 7.4. If we take $B_{t}=W_{t}^{\tau}, \rho=\tau$, and $G=M$, we obtain that $X$ must coincide with $W_{t}$ for $t \geq \tau$. This means that $W_{t}$ is $\mathcal{F}_{t}^{W^{\tau}} \vee \sigma(\tau)$-adapted.

Theorem 7.13. Let $L(Y)=\arg \max _{0 \leq t \leq 1} Y_{t}$. Then

$$
\left(Y^{*}, \theta^{*}, Z^{*}, H^{*}, \xi^{*}, \lambda^{*}\right)=\left(W, \alpha(\cdot, \tau), W^{\tau}, H, H\left(1, W_{1}\right), L(W)\right)
$$

satisfies all the requirements to be an $(L, \mu)$-weak equilibrium except the càglàd property in condition (v) of Definition 2.7.

Proof. The proof of this result is exactly the same as the one for Theorem 7.6, except for the càglàd approximations used. In this case, for $\varepsilon \in(0,1)$, define $\tau^{\varepsilon,+}=(\tau+\varepsilon) \wedge 1$ and $\tau^{\varepsilon,-}=(\tau-\varepsilon) \wedge 0$. Then the process $\alpha^{\varepsilon}=\left\{\alpha_{t}^{\varepsilon} \triangleq \alpha_{t} \mathbf{1}_{\left(\tau^{\varepsilon,-,} \tau^{\varepsilon,+]^{c}}\right.}(t), t \in[0,1]\right\}$ converges $P \times \lambda$-a.e. to $\alpha$ as $\varepsilon \downarrow 0$ and satisfies $\left|\alpha^{\varepsilon}\right| \leq|\alpha|$.
7.3. Comparing the expected wealth of $M$ and $\tau$. In this subsection we show using numerical calculations that the information about the time at which the total demand achieves its maximum gives less expected profit than the information about the maximum, that is, $J(\alpha(\cdot . \tau)) \leq J(\alpha(\cdot . M))$.

We can write

$$
\begin{aligned}
J(\alpha(\cdot, M)) & =\mathbb{E}\left[\int_{0}^{1}\left(\mathbb{E}\left[\xi \mid \mathcal{F}_{1}^{W}\right]-H\left(t, W_{t}\right)\right) \alpha(t, M) d t\right] \\
& =\mathbb{E}\left[\int_{0}^{1}\left(H\left(1, W_{1}\right)-H\left(t, W_{t}\right)\right) \alpha(t, M) d t\right]
\end{aligned}
$$

Note that

$$
\mathbb{E}\left[\int_{0}^{1} H\left(t, W_{t}\right) \alpha(t, M) d t\right]=\mathbb{E}\left[\int_{0}^{1} H\left(t, W_{t}\right) d W_{t}\right]-\mathbb{E}\left[\int_{0}^{1} H\left(t, W_{t}\right) d W_{t}^{M}\right]=0
$$

because the integrability properties of $H$ yield that $\int_{0}^{t} H\left(s, W_{s}\right) d W_{s}$ and $\int_{0}^{t} H\left(s, W_{s}\right) d W_{s}^{M}$ are an $\mathbb{F}^{W}$-martingale and an $\mathbb{F}^{W} \vee \sigma(M)$-martingale, respectively. As the same arguments
work for $\alpha(\cdot, \tau)$, we obtain that for $\lambda \in\{M, \tau\}$

$$
J(\alpha(\cdot, \lambda))=\mathbb{E}\left[H\left(1, W_{1}\right) \int_{0}^{1} \alpha(t, \lambda) d t\right] .
$$

Note also that, after $\tau$, the compensators of $M$ and $\tau$ coincide. Hence, the problem is reduced to verify whether

$$
A(M) \triangleq \mathbb{E}\left[H\left(1, W_{1}\right) \int_{0}^{\tau} \frac{M-W_{t}}{1-t} d t\right] \geq \mathbb{E}\left[H\left(1, W_{1}\right) \int_{0}^{\tau} \frac{M_{t}-W_{t}}{\tau-t} d t\right] \triangleq A(\tau)
$$

7.3.1. Computation of $\boldsymbol{A}(\boldsymbol{M})$ and $\boldsymbol{A}(\boldsymbol{\tau})$. An exact computation of $A(M)$ and $A(\tau)$ is difficult. This is due to the fact that we need to compute integrals with respect to the joint density of $\left(W_{1}, \tau\right)$ conditioned to $\mathcal{F}_{t}^{W}$, which is unknown. Although we have computed this density explicitly, it turns out that it is useless because of its complicated expression. Therefore, we perform a Monte Carlo simulation.

First, we have considered a uniform partition $\pi_{m}=\left\{t_{i}=i / m\right\}_{i=0, \ldots, m}$ of the interval $[0,1]$. We sample the paths of a Brownian motion in this partition and approximate the integral inside the expectation by its upper Riemann sum in $\pi_{m}$. We have used $H\left(1, W_{1}\right)$ as a control variate to reduce the variance of our estimators. Recall that the variance of $H\left(1, W_{1}\right)$ can be computed analytically and the covariances $\operatorname{Cov}\left(H\left(1, W_{1}\right), H\left(1, W_{1}\right) \int_{0}^{\tau} \frac{M-W_{t}}{1-t} d t\right)$ and $\operatorname{Cov}\left(H\left(1, W_{1}\right), H\left(1, W_{1}\right) \int_{0}^{\tau} \frac{M_{t}-W_{t}}{\tau-t} d t\right)$ have been estimated doing a pilot simulation with number of simulations $n=1000$. The main simulations, including the control variate, have length $n=10^{5}$. We have repeated the simulations for different partitions. We quote here the results with $m=10000$ and compute a $99 \%$ confidence interval $[L, U]$ for each simulation. The results are shown in Table 1. Here $\beta$ denotes the value of the control variate. The pricing rules that we use in our experiments are $H(t, y)=y$ and $H(t, Y)=e^{y+(1-t) / 2}$, which are solutions to the heat equation (4.5). We denote them by the letters L and E , respectively. These examples of the pricing rule are the examples considered in Back [2] and yield that the price process follows a Brownian motion and a geometric Brownian motion, respectively. In the first case, note that price or demand information is the same.

Table 1
Monte Carlo estimation of $A(M)$ and $A(\tau)$.

|  | $\widehat{A}$ | $L$ | $U$ | $\widehat{\sigma}_{n}$ | $\beta$ | $H$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\widehat{A}(M)$ | 0.684 | 0.675 | 0.693 | 1.074 | 1.062 | L |
| $\widehat{A}(\tau)$ | 0.189 | 0.186 | 0.192 | 0.376 | 0.616 | L |
| $\widehat{A}(M)$ | 2.656 | 2.625 | 2.687 | 3.775 | 1.955 | E |
| $\widehat{A}(\tau)$ | 1.386 | 1.379 | 1.393 | 0.829 | 0.718 | E |

From these simulations one is inclined to postulate that $J(\alpha(\cdot \cdot \tau)) \leq J(\alpha(\cdot . M))$.
Remark 7.14. It is worth pointing out that these examples can also be considered in the Karatzas-Pikovsky setting; see [12]. In this setting, one studies the portfolio optimization problem of an agent with additional information with respect to the small investor. This model
assumes that the price dynamics is given exogenously and that the insider cannot influence the price process (for more information on this type of formulation, see, e.g., [12], [1], [9], [8], [7], [3], and [6], among others). In fact, in this framework the finiteness of logarithmic utilities for the insider is determined by the quantity $\mathbb{E}\left[\int_{0}^{1}|\alpha(s)|^{2} d s\right]$. By Lemmas 7.10 and 7.3 , we have that in both cases the logarithmic utility for the insider is infinite. However, as the compensator is the same after $\tau$, we decided to compute the previous expectation until $\tau$. The result is

$$
\mathbb{E}\left[\int_{0}^{\tau}|\alpha(s, M)|^{2} d s\right]<\mathbb{E}\left[\int_{0}^{\tau}|\alpha(s, \tau)|^{2} d s\right]=\infty
$$

which is the reverse conclusion of the one shown in Table 1. Nevertheless, there are two major differences between our approach and the Karatzas-Pikovsky one. First, in the KaratzasPikovsky approach the insider is risk averse, while in our approach he/she is risk neutral. Moreover, in the Karatzas-Pikovsky approach the insider has no influence in the price dynamics, while in our model the price process is driven by the insider's demand. Therefore, it would be interesting to extend our model to risk averse insiders to try to examine this issue further.
8. Conclusions. In this paper we construct a model which allows the existence of a rational expectations equilibrium, in a sense weaker than that of the Kyle-Back setting, with an insider possessing information different from the value of the asset at the end of the trading interval. We provide sufficient conditions for the existence and uniqueness in law of a weak equilibrium. Our model allows us to compare the expected wealth obtained by insiders with different kinds of information. We study in some detail the examples of the maximum and the time at which the maximum of the demand is achieved, finding that the first provides more expected final wealth than the second. In order to deal with these examples we prove a new initial enlargement formula for the argument of the maximum of a Brownian motion. Moreover, we prove the existence and uniqueness of a strong solution for an SDE with a drift degenerating at a random time.

## 9. Appendix.

Lemma 9.1. Let $F$ be a function satisfying an exponential growth condition and $\theta$ be a process satisfying (2.4); then $\sup _{0 \leq t \leq 1}\left|F\left(t, Y_{t}^{\theta}\right)\right|$ belongs to $L^{p}(\Omega)$, for any $p \geq 0$, where $Y_{t}^{\theta}=\int_{0}^{t} \theta_{s} d s+Z_{t}$ and $Z$ is a Brownian motion.

Proof. Thanks to the exponential growth condition on $F$, one has that

$$
\begin{aligned}
\sup _{0 \leq t \leq 1}\left|F\left(t, Y_{t}^{\theta}\right)\right|^{p} & \leq A \sup _{0 \leq t \leq 1} \exp \left\{p B\left|\int_{0}^{t} \theta_{s} d s+Z_{t}\right|\right\} \\
& \leq A \exp \left\{p B \sup _{0 \leq t \leq 1}\left|\int_{0}^{t} \theta_{s} d s\right|\right\} \exp \left\{p B \sup _{0 \leq t \leq 1}\left|Z_{t}\right|\right\}
\end{aligned}
$$

The result follows from (2.4) and the fact that the law of $\sup _{0 \leq t \leq 1}\left|Z_{t}\right|$ has finite exponential moments.

Proposition 9.2. Let $\Theta$ be a convex real linear space and $J$ be a functional defined on $\Theta$. Assume that for any $\theta \in \Theta$ there exists the Gâteaux derivative of $J$. That is, for all $v \in \Theta$
the following limit exists:

$$
D_{v} J(\theta) \triangleq \lim _{\varepsilon \rightarrow 0} \frac{J(\theta+\varepsilon v)-J(\theta)}{\varepsilon}
$$

and the application $v \mapsto D_{v} J(\theta)$ is linear for every $\theta \in \Theta$. Then, the following statements are equivalent:
(1) $J$ is concave;
(2) $J\left(\theta^{2}\right) \leq J\left(\theta^{1}\right)+D_{\theta^{2}-\theta^{1}} J\left(\theta^{1}\right)$ for all $\theta^{1}, \theta^{2} \in \Theta$.

Proof. If $J$ is concave, then $J\left(\alpha \theta^{2}+(1-\alpha) \theta^{1}\right) \geq \alpha J\left(\theta^{2}\right)+(1-\alpha) J\left(\theta^{1}\right)$ for all $\theta^{1}, \theta^{2} \in \Theta$, $\alpha \in[0,1]$. This implies

$$
\frac{J\left(\theta^{1}+\alpha\left(\theta^{2}-\theta^{1}\right)\right)-J\left(\theta^{1}\right)}{\alpha} \geq J\left(\theta^{2}\right)-J\left(\theta^{1}\right) \quad \forall \theta^{1}, \theta^{2} \in \Theta, \quad \alpha \in[0,1]
$$

Taking the limit when $\alpha$ tends to zero, by the assumption on the existence of the Gâteaux derivative, we obtain $D_{\theta^{2}-\theta^{1}} J\left(\theta^{1}\right) \geq J\left(\theta^{2}\right)-J\left(\theta^{1}\right)$ for all $\theta^{1}, \theta^{2} \in \Theta$. Conversely, assume that statement (2) is satisfied. Set $\theta=\alpha \theta^{2}+(1-\alpha) \theta^{1}$; then we have that

$$
J\left(\theta^{1}\right) \leq J(\theta)+D_{\theta^{1}-\theta} J(\theta), \quad J\left(\theta^{2}\right) \leq J(\theta)+D_{\theta^{2}-\theta} J(\theta)
$$

Multiplying the first inequality by $\alpha$ and the second one by ( $1-\alpha$ ) and adding them one obtains

$$
\alpha J\left(\theta^{1}\right)+(1-\alpha) J\left(\theta^{2}\right) \leq J(\theta)+D_{\alpha\left(\theta^{1}-\theta\right)+(1-\alpha)\left(\theta^{2}-\theta\right)} J(\theta)
$$

As $\alpha\left(\theta^{1}-\theta\right)+(1-\alpha)\left(\theta^{1}-\theta\right)=0$, the result follows.
Lemma 9.3. Assume that $H \in \mathcal{H}$ and $\theta, v \in \Theta_{b}(M, Z)$. Then, for all $\varepsilon \in(-1,1)$, we have that for $i=1,2$

$$
\begin{aligned}
& \frac{d^{i}}{d \varepsilon^{i}}\left(\mathbb{E}\left[\int_{0}^{1}\left(\xi-H\left(t, Y_{t}^{\theta+\varepsilon v}\right)\right)\left(\theta_{t}+\varepsilon v_{t}\right) d t\right]\right) \\
& =\mathbb{E}\left[\int_{0}^{1} \frac{d^{i}}{d \varepsilon^{i}}\left(\left(\xi-H\left(t, Y_{t}^{\theta+\varepsilon v}\right)\right)\left(\theta_{t}+\varepsilon v_{t}\right)\right) d t\right] .
\end{aligned}
$$

Proof. We do the proof for $i=2$. First, we estimate

$$
\begin{aligned}
& \left|\frac{d^{2}}{d \varepsilon^{2}}\left\{\left(\xi-H\left(t, Y_{t}^{\theta+\varepsilon v}\right)\right)\left(\theta_{t}+\varepsilon v_{t}\right)\right\}\right| \\
& \leq\left|H_{y y}\left(t, Y_{t}^{\theta+\varepsilon v}\right)\left(\int_{0}^{t} v_{s} d s\right)^{2}\left(\theta_{t}+\varepsilon v_{t}\right)\right|+\left|2 H_{y}\left(t, Y_{t}^{\theta+\varepsilon v}\right)\left(\int_{0}^{t} v_{s} d s\right) v_{t}\right| \\
& \leq C\left\{\left|H_{y y}\left(t, Y_{t}^{\theta+\varepsilon v}\right)\right|+\left|H_{y}\left(t, Y_{t}^{\theta+\varepsilon v}\right)\right|\right\}
\end{aligned}
$$

where $C$ is a constant which is independent of $\varepsilon$. These quantities are bounded in $L^{p}(\Omega)$ as Lemma 9.1 shows. Hence, the result follows by the dominated convergence theorem.

Definition 9.4. Let $\theta \in \Theta_{\sup }(M, Z)$. Define $\theta_{t}^{n} \triangleq \theta_{t} \mathbf{1}_{\left\{\sup _{s \leq t}\left|\theta_{s}\right| \leq n\right\}}$ for all $n \in \mathbb{N}$. Clearly, the sequence $\left\{\theta^{n}\right\}_{n \in \mathbb{N}} \subseteq \Theta_{b}(M, Z)$. We also have that $\left\{\theta^{n}\right\}_{n \in \mathbb{N}}$ converges $P \times \lambda$-a.s. to $\theta$. Furthermore, $\left\{\theta^{n}\right\}_{n \in \mathbb{N}}$ converges to $\theta$ in $L^{1}(P \times \lambda)$ by dominated convergence, because $\left|\theta^{n}\right| \leq|\theta|$.

Proposition 9.5. Assume that $\theta \in \Theta_{\text {sup }}(M, Z)$. Then, $\lim _{n \rightarrow \infty} J\left(\theta^{n}\right)=J(\theta)$, where $J$ is the functional defined in (3.1).

Proof. We can define the sequence of $\mathbb{F}^{I}$-stopping times $\tau^{n}=\inf \left\{t \leq 1: \sup _{s \leq t}\left|\theta_{s}\right|>n\right\}$. In the set $\left\{\tau^{n}>t\right\}$, one has that for all $s \leq t,\left|\theta_{s}\right| \leq n$ and $\theta_{s}^{n}=\theta_{s}$. On the other hand, in the set $\left\{\tau^{n} \leq t\right\}$, one has that $\sup _{s \leq t}\left|\theta_{s}\right|>n$ and $\theta_{t}^{n}=0$. Moreover, $\tau^{n} \uparrow 1, P$-a.s. when $n$ tends to infinity. We have that

$$
\begin{aligned}
\left|J(\theta)-J\left(\theta^{n}\right)\right| & \leq\left|\mathbb{E}\left[\xi \int_{0}^{1}\left(\theta_{t}-\theta_{t}^{n}\right) d t\right]\right|+\left|\mathbb{E}\left[\int_{0}^{1}\left(H\left(t, Y_{t}^{\theta}\right) \theta_{t}-H\left(t, Y_{t}^{\theta^{n}}\right) \theta_{t}^{n}\right) d t\right]\right| \\
& \triangleq A_{1}^{n}+A_{2}^{n}
\end{aligned}
$$

Applying the Cauchy-Schwarz inequality, we obtain

$$
A_{1}^{n} \leq \mathbb{E}\left[|\xi|\left|\int_{0}^{1}\left(\theta_{t}-\theta_{t}^{n}\right) d t\right|\right] \leq \mathbb{E}\left[|\xi|^{2}\right]^{1 / 2} \mathbb{E}\left[\left(\int_{0}^{1}\left(\theta_{t}-\theta_{t}^{n}\right) d t\right)^{2}\right]^{1 / 2}
$$

The first expectation is finite, because $\xi$ has moments of second order. For the second expectation, notice that if we fix $\omega \in \Omega$, by dominated convergence we have that $\lim _{n \rightarrow \infty}\left(\int_{0}^{1}\left(\theta_{t}-\right.\right.$ $\left.\left.\theta_{t}^{n}\right) d t\right)^{2}=0, P$-a.s., because $\theta \in L^{1}(\Omega \times[0,1])$. Furthermore,

$$
\left(\int_{0}^{1}\left(\theta_{t}-\theta_{t}^{n}\right) d t\right)^{2}=\left(\int_{\tau^{n}}^{1} \theta_{t} d t\right)^{2} \leq C \sup _{0 \leq t \leq 1}\left(\int_{0}^{t} \theta_{s} d s\right)^{2}
$$

which is in $L^{1}(\Omega)$, by hypothesis (2.4). Therefore, also by dominated convergence one has that $\lim _{n \rightarrow \infty} \mathbb{E}\left[\left(\int_{0}^{1}\left(\theta_{t}-\theta_{t}^{n}\right) d t\right)^{2}\right]=0$. For the term $A_{2}^{n}$, we have that

$$
\int_{0}^{1}\left(H\left(t, Y_{t}^{\theta}\right) \theta_{t}-H\left(t, Y_{t}^{\theta^{n}}\right) \theta_{t}^{n}\right) d t=\int_{\tau^{n}}^{1} H\left(t, Y_{t}^{\theta}\right) \theta_{t} d t, \quad P \text {-a.s. }
$$

Notice that $\int_{0}^{1}\left|H\left(t, Y_{t}^{\theta}\right) \theta_{t}\right| d t \leq \sup _{0 \leq t \leq 1}\left|H\left(t, Y_{t}^{\theta}\right)\right| \int_{0}^{1}\left|\theta_{t}\right| d t<\infty, P$-a.s., because $\theta \in L^{1}(\Omega \times$ $[0,1])$. Hence, $\lim _{n \rightarrow \infty} \int_{\tau^{n}}^{1} H\left(t, Y_{t}^{\theta}\right) \theta_{t} d t=0, P$-a.s. by dominated convergence. Furthermore, $\left|\int_{\tau^{n}}^{1} H\left(t, Y_{t}^{\theta}\right) \theta_{t} d t\right| \leq C \sup _{0 \leq t \leq 1}\left|\int_{0}^{t} H\left(s, Y_{s}^{\theta}\right) \theta_{s} d s\right|$, which is in $L^{1}(\Omega)$, by hypothesis (2.2). Thus, by the dominated convergence theorem we obtain that $\lim _{n \rightarrow \infty} A_{2}^{n}=0$.

Proposition 9.6. Assume that $\theta \in \Theta_{\text {sup }}(M, Z)$ satisfies the optimality equation (3.3). Then, $\lim _{n \rightarrow \infty} D_{\eta-\theta^{n}} J\left(\theta^{n}\right)=0$, for all $\eta \in \Theta_{b}(M, Z)$, where $D_{\eta-\theta^{n}} J\left(\theta^{n}\right)$ is given by (3.2).

Proof. As $\theta \in \Theta_{\text {sup }}(M, Z)$ satisfies (3.3), Remark 3.5 yields

$$
\mathbb{E}\left[\int_{0}^{1}\left(\eta_{t}-\theta_{t}^{n}\right)\left\{\left(\xi-H\left(t, Y_{t}^{\theta}\right)\right)-\int_{t}^{1} H_{y}\left(s, Y_{s}^{\theta}\right) \theta_{s} d s\right\} d t\right]=0 .
$$

Therefore,

$$
\begin{aligned}
& \left|D_{\eta-\theta^{n}} J\left(\theta^{n}\right)\right| \leq\left|\mathbb{E}\left[\int_{0}^{1}\left(\eta_{t}-\theta_{t}^{n}\right)\left(H\left(t, Y_{t}^{\theta}\right)-H\left(t, Y_{t}^{\theta^{n}}\right)\right) d t\right]\right| \\
& +\left|\mathbb{E}\left[\int_{0}^{1}\left(\int_{0}^{t}\left(\eta_{s}-\theta_{s}^{n}\right) d s\right)\left(H_{y}\left(t, Y_{t}^{\theta}\right) \theta_{t}-H_{y}\left(t, Y_{t}^{\theta^{n}}\right) \theta_{t}^{n}\right) d t\right]\right| \\
& \triangleq B_{1}^{n}+B_{2}^{n} .
\end{aligned}
$$

For the term $B_{1}^{n}$, one has that

$$
\begin{aligned}
& \mathbb{E}\left[\int_{0}^{1}\left(\eta_{t}-\theta_{t}^{n}\right)\left(H\left(t, Y_{t}^{\theta}\right)-H\left(t, Y_{t}^{\theta^{n}}\right)\right) d t\right] \\
& =\mathbb{E}\left[\int_{0}^{1} \mathbf{1}_{\left\{t>\tau^{n}\right\}} \eta_{t}\left(H\left(t, Y_{t}^{\theta}\right)-H\left(t, Y_{\tau^{n}}^{\theta}+Z_{t}-Z_{\tau^{n}}\right)\right) d t\right] .
\end{aligned}
$$

When $n$ tends to infinity the integrand in the last equation tends to $0, P \otimes \lambda$-a.s. So we need only justify the application of the dominated convergence theorem. We have, $P \otimes \lambda$-a.s., that

$$
\begin{aligned}
& \left|\eta_{t}\left(H\left(t, Y_{t}^{\theta}\right)-H\left(t, Y_{\tau^{n}}^{\theta}+Z_{t}-Z_{\tau^{n}}\right)\right)\right| \\
& \leq C\left\{\sup _{0 \leq t \leq 1}\left|H\left(t, Y_{t}^{\theta}\right)\right|+\left|H\left(t, Y_{\tau^{n}}^{\theta}+Z_{t}-Z_{\tau^{n}}\right)\right|\right\}
\end{aligned}
$$

By Lemma 9.1, $\sup _{0 \leq t \leq 1}\left|H\left(t, Y_{t}^{\theta}\right)\right|$ is an integrable random variable. These quantities are bounded in $L^{p}(\Omega)$ as the proof of Lemma 9.1 shows. Hence by dominated convergence $\lim _{n \rightarrow \infty} B_{1}^{n}=0$. For the term $B_{2}^{n}$, one has

$$
\begin{aligned}
& \mathbb{E}\left[\int_{0}^{1}\left(\int_{0}^{t}\left(\eta_{s}-\theta_{s}^{n}\right) d s\right)\left(H_{y}\left(t, Y_{t}^{\theta}\right) \theta_{t}-H_{y}\left(t, Y_{t}^{\theta^{n}}\right) \theta_{t}^{n}\right) d t\right] \\
& =\mathbb{E}\left[\int_{\tau^{n}}^{1}\left(\int_{0}^{t}\left(\eta_{s}-\theta_{s}^{n}\right) d s\right) H_{y}\left(t, Y_{t}^{\theta}\right) \theta_{t} d t\right] \\
& =\mathbb{E}\left[\int_{\tau^{n}}^{1}\left(\int_{\tau^{n}}^{t} \eta_{s} d s\right) H_{y}\left(t, Y_{t}^{\theta}\right) \theta_{t} d t\right]+\mathbb{E}\left[\int_{\tau^{n}}^{1}\left(\int_{0}^{\tau^{n}}\left(\eta_{s}-\theta_{s}\right) d s\right) H_{y}\left(t, Y_{t}^{\theta}\right) \theta_{t} d t\right] \\
& \triangleq B_{2,1}^{n}+B_{2,2}^{n} .
\end{aligned}
$$

The term $B_{2,1}^{n}$ converges to zero due to the dominated convergence theorem as

$$
\begin{aligned}
& \left|\int_{\tau^{n}}^{1}\left(\int_{\tau^{n}}^{t} \eta_{s} d s\right) H_{y}\left(t, Y_{t}^{\theta}\right) \theta_{t} d t\right|=\left|\int_{\tau^{n}}^{1} \eta_{s}\left(\int_{s}^{1} H_{y}\left(t, Y_{t}^{\theta}\right) \theta_{t} d t\right) d s\right| \\
& \leq C \sup _{0 \leq t \leq 1}\left|\int_{0}^{t} H_{y}\left(s, Y_{s}^{\theta}\right) \theta_{s} d s\right| \int_{\tau^{n}}^{1}\left|\eta_{s}\right| d s \leq C \sup _{0 \leq t \leq 1}\left|\int_{0}^{t} H_{y}\left(s, Y_{s}^{\theta}\right) \theta_{s} d s\right| \in L^{1}(\Omega),
\end{aligned}
$$

thanks to condition (2.2) and the fact that $\int_{\tau^{n}}^{1}\left|\eta_{s}\right| d s$ converges to zero as $n$ tends to infinity. The term $B_{2,2}^{n}$ converges to zero due to the dominated convergence theorem as

$$
\begin{aligned}
\left|\int_{\tau^{n}}^{1}\left(\int_{0}^{\tau^{n}}\left(\eta_{s}-\theta_{s}\right) d s\right) H_{y}\left(t, Y_{t}^{\theta}\right) \theta_{t} d t\right| & \leq C \sup _{0 \leq t \leq 1}\left|\int_{0}^{t}\left(\eta_{s}-\theta_{s}\right) d s\right| \\
\times \sup _{0 \leq t \leq 1}\left|\int_{0}^{t} H_{y}\left(s, Y_{s}^{\theta}\right) \theta_{s} d s\right| & \in L^{1}(\Omega)
\end{aligned}
$$

as a result of conditions (2.3) and (2.4), the fact that $\eta \in \Theta_{b}(M, Z)$, and the fact that $\left|\int_{\tau^{n}}^{t} H_{y}\left(s, Y_{s}^{\theta}\right) \theta_{s} d s\right|$ converges to zero as $n$ tends to infinity.

Lemma 9.7. Let $a, b \in \mathbb{R}$, and let $s, t>0$; then

$$
\int_{0}^{\infty} \phi(x+a, s) \phi(x+b, t) d x=\phi(b-a, s+t)\left(1-\Phi\left(\frac{a t+b s}{s+t}, \frac{s t}{s+t}\right)\right) .
$$

Proof. In order to prove this statement, rewrite the product of the two density functions $\phi(x+a, s) \phi(x+b, t)$ as a single density function by completing squares in the exponent.

Acknowledgment. We would like to thank two anonymous referees for helpful comments.

## REFERENCES

[1] J. Amendiger, P. Imkeller, and M. Schweizer, Additional logarithmic utility of an insider, Stochastic Process. Appl., 75 (1998), pp. 263-268.
[2] K. Back, Insider trading in continuous time, Rev. Financial Stud., 5 (1992), pp. 387-409.
[3] F. Baudoin, Modeling anticipations in financial markets, in Paris-Princeton Lectures on Mathematical Finance 2002, Springer-Verlag, Berlin, 2003, pp. 43-94.
[4] L. Campi and U. Çetin, Insider trading in an equilibrium model with default: A passage from reducedform to structural modelling, Finance Stoch., 11 (2007), pp. 591-602.
[5] K. H. Сho, Continuous auctions and insider trading: Uniqueness and risk aversion, Finance Stoch., 7 (2003), pp. 47-71.
[6] J. M. Corcuera, P. Imkeller, A. Kohatsu-Higa, and D. Nualart, Additional utility of insiders with imperfect dynamical information, Finance Stoch., 8 (2004), pp. 437-450.
[7] A. Grorud and M. Pontier, Insider trading in a continuous time market model, IJTAF, 3 (1998), pp. 331-347.
[8] P. Imkeller, Malliavin's calculus in insider models: Additional utility and free lunches, Math. Finance, 13 (2003), pp. 153-169.
[9] P. Imkeller, M. Pontier, and F. Weisz, Free lunch and arbitrage possibilities in a financial market with an insider, Stochastic Process. Appl., 92 (2001), pp. 103-130.
[10] T. Jeulin, Semi-martingales et grossissement d'une filtration, Lecture Notes in Math. 833, SpringerVerlag, Berlin, 1980.
[11] T. Jeulin and M. Yor, Inégalité de Hardy, semimartingales, et faux-amis, in Séminaire de Probabilités, XIII (Univ. Strasbourg, Strasbourg, 1977/78), Lecture Notes in Math. 721, Springer-Verlag, Berlin, 1979, pp. 332-359.
[12] I. Karatzas and I. Pikovsky, Anticipative portfolio optimization, Adv. in Appl. Probab., 28 (1996), pp. 1095-1122.
[13] I. Karatzas and E. Shreve, Brownian Motion and Stochastic Calculus, 2nd ed., Grad. Texts in Math. 113, Springer-Verlag, Berlin, 1991.
[14] A. Kohatsu-Higa and S. Ortiz-Latorre, Insider Trading and Weak Equilibrium, Mathematics Preprint 402, IMUB, Barcelona, Spain, 2008.
[15] A. Kohatsu-Higa and M. Yamazato, Enlargement of filtrations with random times for processes with jumps, Stochastic Process. Appl., 118 (2008), pp. 1136-1158.
[16] A. Kyle, Continuous auctions and insider trading, Econometrica, 53 (1985), pp. 317-356.
[17] G. Lasserre, Asymmetric information and imperfect competition in a continuous time multivariate security model, Finance Stoch., 8 (2004), pp. 285-309.
[18] R. Mansuy and M. Yor, Random Times and Enlargement of Filtrations in a Brownian Setting, Lecture Notes in Math. 1873, Springer-Verlag, Berlin, 2006.
[19] P. E. Protter, Stochastic Integration and Differential Equations, Appl. Math. (N. Y.) 21, SpringerVerlag, Berlin, 2004.

# Common Forward Rate Volatility* 

Victor Goodman ${ }^{\dagger}$ and Kyounghee $\mathrm{Kim}^{\ddagger}$


#### Abstract

Statistical analyses of forward interest rate behavior provide evidence that these rates share a common volatility. We develop a risk-neutral term structure model based on this assumption. The main feature of this model is that each discounted bond price is both an explicit local martingale and a diffusion. The Markov property of discounted bonds is convenient for pricing interest rate derivatives. We give price formulas for caps and swaptions and compare caplet prices to the market-standard Black formula. The two formulas have nearly identical numerical values.


Key words. term structure, Black formula, LIBOR derivative
AMS subject classifications. Primary, 91B28, 60H30; Secondary, 60J65
DOI. 10.1137/090750676

1. Introduction. Empirical studies of forward interest rates were carried out in the early nineties by Litterman and Scheinkman [7], Wilson [16], and others. A summary account of several statistical analyses and their similar conclusions can be found in Rebonato [12]. These studies found that a major factor in explaining the movement of forward rate curves is a shared level of volatility. That is, a promising choice to model the evolution of forward interest rates, $f(t, T)$, with $T$ denoting the maturity parameter, has the form

$$
\begin{equation*}
d f(t, T)=\alpha(t, T) d t+\sigma f(t, T) d W \tag{1.1}
\end{equation*}
$$

The parameter $\sigma$ is the common forward rate volatility, and $W$ is a single Brownian motion.
The difficulties of setting up such models are noted in Heath, Jarrow, and Morton [5] and in Miltersen, Sandmann, and Sondermann [8]. We avoid these difficulties by introducing a hypothetical forward risk-neutral measure as the context for defining asset prices. One can easily develop a theory which identifies the structure of the market drift $\alpha(t, T)$ in terms of a market price of risk in order to characterize arbitrage-free models. The crucial steps avoid the use of a money market account. A forward risk-neutral approach allows one to relate drift and no-arbitrage in a fashion parallel to the well-known general result in [5]. However, we will not do that in this paper. Instead, we focus on developing and using forward risk-neutral models to price and hedge basic interest rate derivatives.

Each forward risk-neutral model has a benign, Markovian structure. Each discounted $T$-maturity bond process is explicitly expressed in terms of basic Brownian functionals, but, more importantly, each discounted bond, $Z(t, T)$, is a diffusion process whose SDE is

$$
\begin{equation*}
d Z=\sigma Z \log Z d W \tag{1.2}
\end{equation*}
$$

[^17]That is, all bond prices satisfy the same SDE and differ only because of their calibration prices. Consequently, the pricing of interest rate caps and swaptions is exceptionally easy since values of the same diffusion process, with differing initial conditions, enter into the cap and swaption calculations.

Previous work by Ritchken and Sanakarasubramanian [11] and Rogers [14] has developed Markov models where short rate processes are functions of Markov processes. Here it is the asset prices themselves which are Markov.

We return to the issue of statistical evidence for this type of model by examining some recent bond trading data. One implication of (1.1) is that a bond's stochastic volatility is proportional to the logarithm of the bond price. That is, the volatility of a $T$-maturity bond has the form $\sigma \log B(t, T)$, where $B(t, T)$ denotes the price of the bond.

If bond trading is consistent with this volatility choice, then one should be able to identify the shared volatility parameter by observing a common variance parameter for several time series of the type

$$
\frac{\Delta B}{B \log B},
$$

where different maturity prices are observed over a specific time period. In Appendix I, we graph trading results for U.S. Treasury bond prices during July and August of 2000. Figure 2 displays individual bond volatility estimates based on four-week trailing variances for the bond returns $\Delta B / B$. In contrast, Figure 3 displays trailing variances for the modified time series incorporating the log factor. The coalescing sigma plots in Figure 3 offer striking evidence in favor of this type of model.

The paper is organized as follows. Section 2 contains explicit formulas for forward interest rates, bond prices, and their SDEs within forward risk-neutral models. Section 3 has a general pricing formula. Section 4 has the cap price formula and compares caplet prices to marketstandard prices based on the Black formula. The two formulas give numerically identical results. Section 5 contains the swaption price formula.
2. Forward rates and bond prices. Two well-known choices for $\alpha(t, T)$ in (1.1) produce risk-neutral modeling equations. The first choice, as Health, Jarrow, and Morton discovered in [5], allows discounting with a money market account. The second choice, later used by Musiela and Rutkowski [9], allows discounting with a bond price, whose maturity is denoted by $T^{*}$. We refer to this choice as forward risk-neutral dynamics.

Generic forms for the two choices are

$$
\alpha=\beta \int_{t}^{T} \beta(t, u) d u \quad \text { or } \quad \alpha=-\beta \int_{T}^{T^{*}} \beta(t, u) d u \text {. }
$$

Here, $\beta(t, T)$ denotes the Brownian coefficient in an SDE for the forward rate $f(t, T)$. In our case, $\beta=\sigma f$.

It is known (see [5] and [8]) that the risk-neutral choice, involving money market discounting, leads to arbitrage. Therefore, we use the second choice, with a bond as numéraire, in our crucial modeling equation:

$$
\begin{equation*}
d f(t, T)=-\sigma^{2} f(t, T) \int_{T}^{T^{*}} f(t, u) d u d t+\sigma f(t, T) d W \tag{2.1}
\end{equation*}
$$

One obvious feature of $(2.1)$ is that any positive solution is likely to be finite due to a negative drift for the case of earlier maturities $\left(T \leq T^{*}\right)$. In fact, (2.1) has explicit positive solutions. We formulate our model in terms of these solutions.

Definition 2.1. The forward rate process $f(t, T)$, defined for a common volatility parameter $\sigma$ and a range of maturities $0 \leq T \leq T^{*}$ with $t \leq T$, is given by

$$
f(t, T)=\frac{f(0, T) M_{t}}{\left(1+\frac{\sigma^{2}}{2} F(T) \int_{0}^{t} M_{s} d s\right)^{2}}
$$

In this equation $M_{t}=\exp \left(\sigma W_{t}-\frac{\sigma^{2}}{2} t\right)$; the initial (given) forward rate curve, $f(0, T)$, and its integral, $F(T):=\int_{T}^{T^{*}} f(0, u) d u$, define the calibration parameters in the forward rate formula.

We assume for now that the initial forward rate curve is nonnegative, so that $F(T) \geq 0$. This assumption may be relaxed to allow some negative rates as long as the function $F(T)=$ $\log \left(B(0, T) / B\left(0, T^{*}\right)\right)$ is nonnegative.

Proposition 2.2. Each maturity integral of rates in Definition 2.1 is given by

$$
\begin{equation*}
\int_{T}^{T^{*}} f(t, u) d u=\frac{F(T) M_{t}}{1+\frac{\sigma^{2}}{2} F(T) \int_{0}^{t} M_{s} d s} \tag{2.2}
\end{equation*}
$$

Moreover, each T-maturity bond price has the explicit form

$$
B(t, T)=\exp \left(\frac{F(T) M_{t}}{1+\frac{\sigma^{2}}{2} F(T) \int_{0}^{t} M_{s} d s}-\frac{F(t) M_{t}}{1+\frac{\sigma^{2}}{2} F(t) \int_{0}^{t} M_{s} d s}\right)
$$

In particular,

$$
B\left(t, T^{*}\right)=\exp \left(-\frac{F(t) M_{t}}{1+\frac{\sigma^{2}}{2} F(t) \int_{0}^{t} M_{s} d s}\right) .
$$

Proof. Equation (2.2) can verified by differentiating with respect to $T$. The two bond equations follow from the identity

$$
B(t, T)=\exp \left(-\int_{t}^{T} f(t, u) d u\right)
$$

Theorem 2.3. Each T-maturity bond price process is a semimartingale, and each discounted price

$$
\frac{B(t, T)}{B\left(t, T^{*}\right)}
$$

is a positive local martingale if $T \leq T^{*}$. Moreover, each discounted bond price satisfies the SDE

$$
\begin{equation*}
d Z=\sigma Z \log Z d W \tag{2.3}
\end{equation*}
$$

Proof. Let $Y_{t}=\log \left(B(t, T) / B\left(t, T^{*}\right)\right)$. From (2.2) one sees that

$$
\begin{equation*}
Y_{t}=\frac{F(T) M_{t}}{1+\frac{\sigma^{2}}{2} F(T) \int_{0}^{t} M_{s} d s} \tag{2.4}
\end{equation*}
$$

It is easily checked that $Y$ satisfies

$$
\begin{equation*}
d Y=-\frac{\sigma^{2}}{2} Y^{2} d t+\sigma Y d W \tag{2.5}
\end{equation*}
$$

It follows that the process $Z=\exp (Y)$ is a local martingale and satisfies (2.3). If $T \leq T^{*}$, so that $F(T) \geq 0$, then $Y_{t}$ remains positive and so $Z$ is positive and is larger than one.
3. European derivative prices. Each discounted bond price in Theorem 2.3 is strictly a supermartingale (see Goodman and Kim [4]). That is, a bond's risk-neutral price is larger than its expected discounted future value under the forward risk-neutral measure. One may say that these prices form a market bubble as described in Cox and Hobson [2] and Heston, Loewenstein, and Willard [6]. Some care must be taken to define admissible trading strategies and fair prices of derivatives based on appropriate hedges.

We select a probability space $\left(\Omega, \mathcal{F}, \mathcal{F}_{t}, P\right)$ which supports a Brownian motion $W_{t}$, where $t \leq T^{*}$ and $\mathcal{F}=\mathcal{F}_{T^{*}}$. Proposition 2.2 determines bond price processes, and we follow the pricing framework in [2], where Theorem 3.3 states that the price at time $t$ of a European option with payoff $H(s) \geq 0$ at time $T$, based on an asset price $S_{T}$, is given by

$$
\begin{equation*}
E\left\{H\left(S_{T}\right) \mid \mathcal{F}_{t}\right\} . \tag{3.1}
\end{equation*}
$$

This general formula is valid if the process (3.1) can be expressed as a stochastic integral with respect to the discounted asset $S_{t}$. We consider interest rate derivatives which have a discounted payoff of the form

$$
\begin{equation*}
\mathcal{Z}=\frac{B\left(T, T_{0}\right)}{B\left(T, T^{*}\right)} H\left(\frac{B\left(T, T_{1}\right)}{B\left(T, T_{0}\right)}, \ldots, \frac{B\left(T, T_{k}\right)}{B\left(T, T_{0}\right)}\right), \tag{3.2}
\end{equation*}
$$

where $T^{*} \geq T_{i}>T, i=0,1, \ldots, k ; H\left(z_{1}, \ldots, z_{k}\right)$ is measurable and nonnegative.
The main result of this section identifies the process (3.1) for this type of payoff. Since both interest rate cap and swaption payoffs can be expressed as combinations of (3.2), we use this result to derive specific cap and swaption price formulas in sections 4 and 5.

Theorem 3.1. Let $E^{T *}$ denote expectation with respect to a $T^{*}$-forward measure where bond price processes are defined as in Proposition 2.2. If a payoff amount $\mathcal{Z}$ has the form (3.2), then

$$
\begin{equation*}
E\left\{\mathcal{Z} \mid \mathcal{F}_{t}\right\}=G\left(\frac{B\left(t, T^{*}\right)}{B\left(t, T_{0}\right)}, \frac{B\left(t, T_{1}\right)}{B\left(t, T_{0}\right)}, \ldots, \frac{B\left(t, T_{k}\right)}{B\left(t, T_{0}\right)}\right) . \tag{3.3}
\end{equation*}
$$

The function $G\left(z_{0}, z_{1}, \ldots, z_{k}\right)$ is given by

$$
\begin{equation*}
G=z_{0}^{-1} E^{T_{0}}\left\{H\left(\frac{B\left(\tau, T_{1}\right)}{B\left(\tau, T_{0}\right)}, \ldots, \frac{B\left(\tau, T_{k}\right)}{B\left(\tau, T_{0}\right)}\right) ; \frac{B\left(\tau, T^{*}\right)}{B\left(\tau, T_{0}\right)}>0\right\} . \tag{3.4}
\end{equation*}
$$

In this equation, each $z_{i}, i \geq 1$, is the initial discounted bond price $B\left(0, T_{i}\right) / B\left(0, T_{0}\right)$ in a $T_{0}$-forward model, $z_{0}=B\left(0, T^{*}\right) / B\left(0, T_{0}\right)$, and $\tau=T-t$.

Proof. We use the Markov property of each process $t \rightarrow B\left(t, T_{i}\right) / B\left(t, T^{*}\right)$ to verify (3.3). Each argument, $B\left(T, T_{i}\right) / B\left(T, T_{0}\right)$, in $\mathcal{Z}$ is the ratio of two Markov processes, and so $\mathcal{Z}$ is a
function of a (degenerate) Markov vector that is evaluated at time $T$. A basic conditioning identity reduces the answer to an unconditional expectation involving the vector Markov process at time $\tau=T-t$. The function $G$ is computed in Appendix II.

Bond prices appear in (3.4) in cases where bond maturities exceed the numéraire maturity $T_{0}$. We use the bond price formula in Proposition 2.2 to define a price process in each case where $T_{i}>T_{0}$. An initial discounted bond price $z_{i}$ is smaller than one. That is, the $\log$ ratio is negative and determines a negative initial value $F\left(T_{i}\right)=\log B\left(0, T_{i}\right) / B\left(0, T_{0}\right)$. The discounted asset value $Z_{i}(t)$ is a bounded martingale which may collapse to zero. The formula (3.4) restricts the expectation to the event $B\left(\tau, T^{*}\right) / B\left(\tau, T_{0}\right)>0$ so that all possible asset collapses are avoided within the expectation.

Example. Suppose we take $H \equiv 1$ in (3.2) and $t=0$. Theorem 3.1 reduces in this case to

$$
\begin{equation*}
E\left\{\frac{B\left(T, T_{0}\right)}{B\left(T, T^{*}\right)}\right\}=\frac{B\left(0, T_{0}\right)}{B\left(0, T^{*}\right)} P^{T_{0}}\left\{\frac{B\left(T, T^{*}\right)}{B\left(T, T_{0}\right)}>0\right\} \tag{3.5}
\end{equation*}
$$

a formula for the expected future discounted value of the $T_{0}$ bond. We wish to estimate this quantity for various bonds, yielding, say, $4 \%$, that might be involved in a 10 -year interest rate swap, assuming a common volatility of $\sigma=0.2$.

Let $y=\log B\left(0, T_{0}\right) / B\left(0, T^{*}\right)$, a positive quantity. For the choice of the longest bond $\left(T^{*}=10\right)$ and the midmaturity bond $\left(T_{0}=5\right)$ in the swap, $y=0.2$. Equation (3.4) requires the use of a longer-term discounted bond $B\left(t, T^{*}\right) / B\left(t, T_{0}\right)=\exp (Y)$, where

$$
Y=\frac{-y M_{t}}{1+\frac{\sigma^{2}}{2}(-y) \int_{0}^{t} M_{s} d s}
$$

$Y$ is negative for small values of $t$, as it should be, since a discounted long bond price is less than one. However, $Y$ explodes to $-\infty$ at the random time when the denominator hits zero. That is, the discounted long bond price collapses at this moment. Therefore,

$$
\begin{equation*}
P^{T_{0}}\left\{\frac{B\left(T, T^{*}\right)}{B\left(T, T_{0}\right)}>0\right\}=P\left\{\frac{\sigma^{2} y}{2} \int_{0}^{T} M_{s} d s<1\right\} \tag{3.6}
\end{equation*}
$$

This identity shows that the discounted $T_{0}$-bond price process is not a martingale since the probability (3.6) is a decreasing function of $T$. However, the probability is essentially equal to one.

The random quantity $\int_{0}^{T} M_{s} d s$ is a time integral of geometric Brownian motion. These random quantities first surfaced in financial models as payoff amounts in Asian options and have been studied by Geman and Yor [3], Rogers and Shi [15], and many others. Although much information concerning the probability distributions of time integrals has been found (see Yor [17], for example), the known results do not seem to be useful for deriving the risk-neutral density of discounted bonds in this work.

On the other hand, market data suggests that the effective time in such integrals is extremely short and this allows a simple estimate of their size. The time scaling $\sigma^{2} s \rightarrow t$ reduces the integral

$$
\sigma^{2} \int_{0}^{T} M_{s} d s
$$

to one with the same probability distribution, $\int_{0}^{\sigma^{2} T} M_{s}^{(1)} d s$, where $M^{(1)}$ denotes the exponential martingale with $\sigma=1$.

With $T=T_{0}=5$, and $T^{*}=10$, the effective time is $\tau=5 \sigma^{2}=0.2$, and we use the estimate

$$
\int_{0}^{\tau} \exp \left(W_{s}-s / 2\right) d s<\int_{0}^{\tau} \exp \left(\max _{s \leq \tau} W_{s}\right) d s
$$

Since $\max _{s \leq \tau} W_{s}$ has the same distribution as $\sqrt{\tau}|Z|$, where $Z$ denotes a standard normal random variable, we may replace the time integral by the stochastically larger random variable

$$
\begin{equation*}
\tau \exp (\sqrt{\tau}|Z|) \tag{3.7}
\end{equation*}
$$

to obtain the following lower bound on the probability in (3.6):

$$
P\{0.2 \exp (\sqrt{0.2}|Z|)<10)\} \approx P\{|Z|<8.75\} \approx 1-10^{-16} .
$$

Even if a 30 -year bond were used as the numéraire in this 10 -year swap example, each probability in (3.6) for the 10 -year bond would exceed 0.9962 .
4. Interest rate caps. A cap is a collection of call options on LIBOR rates over several consecutive LIBOR periods. Each single call option is a caplet, and a caplet's price before the LIBOR period reset date agrees with the price of a European put option on a specific bond (see [10]). We first derive a price formula for bond puts and then state the corresponding interest rate caplet formula.

If $T$ denotes the exercise date for a put on a $T^{\prime}$-maturity bond $\left(T^{\prime}>T\right)$, we write the payoff function in the scaled form

$$
\left(1-\kappa B\left(T, T^{\prime}\right)\right)^{+} .
$$

The payoff has been scaled up by the reciprocal, $\kappa>1$, of the put exercise price. This form of a payoff is more convenient for relating put and caplet prices. We denote the put price by

$$
P\left(t, T, T^{\prime}, \kappa\right)
$$

Our answer is expressed in terms of two probability functions of $x>0, y>0, d \geq x$, and $s>0$ :

$$
\begin{align*}
\mathcal{K}^{+}(s, x, y, d) & =\operatorname{Pr}\left\{\frac{M_{s}^{(1)}}{x^{-1}-\frac{1}{2} \int_{0}^{s} M_{u}^{(1)} d u} \geq y \quad \text { and } \quad \int_{0}^{s} M_{u}^{(1)} d u<\frac{2}{d}\right\},  \tag{4.1}\\
\mathcal{K}^{-}(s, x, y, d) & =\operatorname{Pr}\left\{\frac{M_{s}^{(1)}}{x^{-1}+\frac{1}{2} \int_{0}^{s} M_{u}^{(1)} d u} \geq y \quad \text { and } \quad \int_{0}^{s} M_{u}^{(1)} d u<\frac{2}{d}\right\},
\end{align*}
$$

where $M_{s}^{(1)}=\exp \left(W_{s}-\frac{s}{2}\right)$.
Proposition 4.1. A put on a $T^{\prime}$-maturity bond, with exercise date $T$ and exercise price $1 / \kappa$, having the scaled payoff $\left(1-\kappa B\left(T, T^{\prime}\right)\right)^{+}$, has a price given by

$$
\begin{equation*}
P=B(t, T) \mathcal{K}^{+}\left(\sigma^{2} \tau, x, y, d_{1}\right)-\kappa B\left(t, T^{\prime}\right) \mathcal{K}^{-}\left(\sigma^{2} \tau, x, y, d_{2}\right) . \tag{4.2}
\end{equation*}
$$

In this formula, $\tau=T-t, x=\log \left(B(t, T) / B\left(t, T^{\prime}\right)\right), y=\log \kappa, d_{1}=\log \left(B(t, T) / B\left(t, T^{*}\right)\right)$, and $d_{2}=\log \left(B\left(t, T^{\prime}\right) / B\left(t, T^{*}\right)\right)$.

Proof. See Appendix II.
Corollary 4.2. The put price in (4.2) is given approximately by

$$
P\left(t, T, T^{\prime}, \kappa\right) \doteq B(t, T) \Phi\left(b_{1}\right)-\kappa B\left(t, T^{\prime}\right) \Phi\left(b_{2}\right)
$$

where

$$
\begin{equation*}
b_{i}:=\frac{\log (x / y)}{\sigma \sqrt{\tau}}-\frac{\sigma \sqrt{\tau}}{2}(1 \mp x) . \tag{4.3}
\end{equation*}
$$

$\Phi$ denotes the standard normal distribution function.
Proof. The functions $\mathcal{K}^{ \pm}$involve the probability

$$
\operatorname{Pr}\left\{\int_{0}^{s} M_{u}^{(1)} d u<\frac{2}{d}\right\} .
$$

If $d=0.4$, the value of a 10 -year discount at the annual rate of $4 \%$, and if a volatility of $\sigma=.2$ is assumed, then an expiration time of two years in the put formula gives an approximate $z$-score of +14 for the probability in question. (See (3.7) for a simple estimate of the time integral.) This probability is essentially equal to one. Consider the approximations

$$
\begin{equation*}
\tilde{Y}=x \exp \left( \pm \frac{s x}{2}\right) M_{s}^{(1)} \tag{4.4}
\end{equation*}
$$

to the processes

$$
Y=\frac{M_{s}^{(1)}}{x^{-1} \mp \frac{1}{2} \int_{0}^{s} M_{u}^{(1)} d u}
$$

Notice that $d \tilde{Y}= \pm \frac{1}{2} \tilde{Y}_{0} \tilde{Y} d s+\tilde{Y} d W$ so that the distributions in (4.4) are approximately equal for small times to the processes $Y^{ \pm}$(see (2.5)). Then

$$
\begin{aligned}
\mathcal{K}^{ \pm} & \approx \operatorname{Pr}\left\{x \exp \left( \pm \frac{s x}{2}\right) M_{s}^{(1)} \geq y\right\} \\
& =\Phi\left(b_{i}\right),
\end{aligned}
$$

where

$$
b_{i}=\frac{\log (x / y)}{\sqrt{s}}-\frac{\sqrt{s}}{2}(1 \mp x) .
$$

4.1. Caplet price. A caplet payoff at time $T+\Delta$ with a cap $K$ (see [13]) is a call on the LIBOR rate

$$
L=(1 / B(T, T+\Delta)-1) \Delta^{-1}
$$

The payoff at the end of the LIBOR period is

$$
(L-K)^{+} \Delta
$$

Since the payoff is known at the reset time, $T$, the value (at this time) of the caplet is the product of the bond $B(T, T+\Delta)$ and the payoff. This gives us the well-known relation between caplet prices and put prices (see [9]):

$$
B(T, T+\Delta)(L-K)^{+} \Delta=(1-(1+K \Delta) B(T, T+\Delta))^{+} .
$$

The right-hand expression is a payoff for a put, whose parameter $\kappa$, appearing in Proposition 4.1, equals $1+K \Delta$. Two caplet price formulas are consequences of the result in Proposition 4.1.

Corollary 4.3. A caplet on the LIBOR rate over the time period $[T, T+\Delta]$ with cap $K$ has the price

$$
\begin{align*}
C(t, T, T+\Delta, K) & =B(t, T) \mathcal{K}^{+}\left(\sigma^{2} \tau, x, y, d_{1}\right)  \tag{4.5}\\
& -(1+\Delta K) B(t, T+\Delta) \mathcal{K}^{-}\left(\sigma^{2} \tau, x, y, d_{2}\right) .
\end{align*}
$$

Here, $\tau=T-t, x=\log (B(t, T) / B(t, T+\Delta)), y=\log (1+\Delta K), d_{1}=\log \left(B(t, T) / B\left(t, T^{*}\right)\right)$, and $d_{2}=\log \left(B(t, T+\Delta) / B\left(t, T^{*}\right)\right)$. An approximate caplet price is given by

$$
\begin{align*}
& C \doteq B(t, T) \Phi\left(b_{1}\right)-(1+\Delta K) B(t, T+\Delta) \Phi\left(b_{2}\right), \\
& b_{i}:=\frac{1}{\sigma \sqrt{\tau}} \log \left(\frac{\log \left(B(t, T) / B\left(t, T^{\prime}\right)\right)}{\log (1+\Delta K)}\right)-\frac{\sigma \sqrt{\tau}}{2}(1 \mp x) . \tag{4.6}
\end{align*}
$$

4.2. Comparison with the market-standard formula. The approximate caplet price formula is numerically very close to the Black formula. Our caplet price is

$$
B(t, T) \Phi\left(b_{1}\right)-B(t, T+\Delta) \Phi\left(b_{2}\right)-\Delta K B(t, T+\Delta) \Phi\left(b_{2}\right)
$$

whereas the market-standard formula (equation (2.9) of [13]) is

$$
(B(t, T)-B(t, T+\Delta)) \Phi\left(h_{1}\right)-\Delta K B(t, T+\Delta) \Phi\left(h_{2}\right) .
$$

In practice, the ratio of the two bonds is near one and also $\log (1+\Delta K) \doteq \Delta K$. One sees that the first term of $b_{i}$ in (4.6) is very nearly

$$
\frac{\log (f / K)}{\sigma \sqrt{\tau}} .
$$

In this expression, $f$ denotes forward LiBor, and the expression is the leading term of $h_{i}$. Both quantities $b_{1}$ and $b_{2}$ are numerically very close to $h_{2}$.

Despite differences between the two formulas, they agree numerically. Figure 1 displays numerical values for (4.6) compared with values for the standard caplet formula. The swap principal amount is $\$ 1,000$.
5. Swaptions. A swaption is an option whose payoff is the difference between the floating leg and the fixed leg of an interest rate swap [13]. The payoff of a European swaption is

$$
\begin{equation*}
\left(1-B\left(T, T_{n}\right)-K \sum_{i=1}^{n} B\left(T, T_{i}\right)\right)^{+} . \tag{5.1}
\end{equation*}
$$

$\left.\begin{array}{cccccccc}\mathrm{B}(\mathrm{t}, \mathrm{T}) & \mathrm{B}\left(\mathrm{t}, \mathrm{T}^{\prime}\right) & \begin{array}{l}\text { Time } \\ \text { to } \\ \text { exp }\end{array} & \begin{array}{c}\text { Cap } \\ \text { value } \\ \text { (K) }\end{array} & \begin{array}{l}\text { libor } \\ \text { yrly } \\ \text { vol }\end{array} & \begin{array}{c}\text { 3 Mo } \\ \text { Forwd } \\ \text { libor }\end{array} & \begin{array}{l}\text { Caplet } \\ \text { Eq. } \\ \text { Price }\end{array} & \begin{array}{l}\text { Market }\end{array} \\ & & & & & & \text { Standard } \\ \text { Price }\end{array}\right\}$

Figure 1.

In this equation, $K$ represents a given fixed coupon amount for periodic payments in a swap that is offered to the holder at the expiration time, $T$, and $T<T_{i}<T_{i+1}$.

We denote the swaption price at time $t \leq T$ by $S(t, K)$. The price formula contains $n+1$ probability functions $\mathcal{S}_{0}, \mathcal{S}_{1}, \ldots, \mathcal{S}_{n}$, and the formula has the form

$$
\begin{equation*}
S(t, K)=B(t, T) \mathcal{S}_{0}-B\left(t, T_{n}\right) \mathcal{S}_{n}-K \sum_{i=1}^{n} B\left(t, T_{i}\right) \mathcal{S}_{i} \tag{5.2}
\end{equation*}
$$

Each term $\mathcal{S}_{i}$ is a transition probability for the $(n+1)$-dimensional diffusion process

$$
\begin{equation*}
\mathbf{D}(t)=\left\{\exp \left(\frac{x_{i} M_{t}}{1+\frac{x_{i}}{2} \int_{0}^{t} M_{s} d s}\right)\right\}_{i=0, \ldots, n} \tag{5.3}
\end{equation*}
$$

where $M_{t}=\exp \left(W_{t}-t / 2\right)$ and all components are dependent. We include cases where some initial conditions $x_{i}$ are negative.

Each $\mathcal{S}_{i}$ is a fixed function of variables $s, \mathbf{x}$, and $K$, where the vector $\mathbf{x}$ is the initial value of the exponents in (5.3). We let

$$
\begin{equation*}
\mathcal{S}_{0}(s, \mathbf{x}, K)=\operatorname{Pr}\left\{K \sum_{i=1}^{n} \mathbf{D}_{i}(s)+\mathbf{D}_{n}(s) \leq 1, \mathbf{D}_{0}(s)>0\right\} \tag{5.4}
\end{equation*}
$$

Each component of $\mathbf{x}$ is negative. The function $\mathcal{S}_{n}$ is given by

$$
\begin{equation*}
\mathcal{S}_{n}(s, \mathbf{x}, K)=\operatorname{Pr}\left\{1+K+K \sum_{j=1}^{n-1} \mathbf{D}_{j}(s) \leq \mathbf{D}_{n}(s), \mathbf{D}_{0}(s)>0\right\} \tag{5.5}
\end{equation*}
$$

All components $x_{i}, i \geq 1$, are positive. Each of the other functions $\mathcal{S}_{i}, 0<i<n$, is given by (5.6):

$$
\begin{equation*}
\mathcal{S}_{i}=\operatorname{Pr}\left\{K \sum_{j \neq i}^{n} \mathbf{D}_{j}(s)+K+(K+1) \mathbf{D}_{n}(s) \leq \mathbf{D}_{i}(s), \mathbf{D}_{0}(s)>0\right\} \tag{5.6}
\end{equation*}
$$

Each component $x_{j}, 1 \leq j \leq i$, is positive, but both $x_{0}$ and $x_{j}$ are negative for $j>i$.
Proposition 5.1. The swaption price for the payoff in (5.1) is

$$
\begin{align*}
S(t, K) & =B(t, T) \mathcal{S}_{0}\left(\sigma^{2} \tau, \boldsymbol{x}, K\right)-B\left(t, T_{n}\right) \mathcal{S}_{n}\left(\sigma^{2} \tau, \boldsymbol{x}^{\prime}, K\right) \\
& -K \sum_{i=1}^{n} B\left(t, T_{i}\right) \mathcal{S}_{i}\left(\sigma^{2} \tau, \boldsymbol{x}^{(i)}, K\right) \tag{5.7}
\end{align*}
$$

In this formula $\tau=T-t$ and the components of $\boldsymbol{x}, \boldsymbol{x}^{\prime}$, and $\boldsymbol{x}^{(i)}$ are the following asset ratios:

$$
x_{i}=\log \frac{B\left(t, T_{i}\right)}{B(t, T)}, \quad x_{i}^{\prime}=\log \frac{B\left(t, T_{i}\right)}{B\left(t, T_{n}\right)}, \quad i<n, \quad \text { and } x_{n}^{\prime}=\log \frac{B(t, T)}{B\left(t, T_{n}\right)}
$$

For each $j \neq i$

$$
\begin{gathered}
x_{j}^{(i)}=\log \frac{B\left(t, T_{j}\right)}{B\left(t, T_{i}\right)}, \quad x_{i}^{(i)}=\log \frac{B(t, T)}{B\left(t, T_{i}\right)} . \quad \text { Also } x_{0}=\log \frac{B\left(t, T^{*}\right)}{B(t, T)} \\
x_{0}^{\prime}=\log \frac{B\left(t, T^{*}\right)}{B\left(t, T_{n}\right)}, \quad \text { and } x_{0}^{(i)}=\log \frac{B\left(t, T^{*}\right)}{B\left(t, T_{i}\right)} .
\end{gathered}
$$

Proof. See Appendix II.
Remark 5.2. Each term in the swaption formula is the product of an asset price and a transition probability for the vector-valued diffusion process $\mathbf{D}(t)$ to occupy a particular halfspace in $\mathbb{R}^{n}$ at time $t=\sigma^{2} \tau$. The diffusion is highly degenerate since the components are different values of the same process. Numerical techniques can be applied to obtain these transition probabilities.

Each $x_{i}$ is the logarithm of a discounted bond price, so that the swaption formula directly determines a hedge with positions in the bond assets involved in the swap.
6. Conclusion. This paper introduces a Markov model for bond prices with various bond prices serving as the numéraire. A Markovian framework, where discounted assets are expressed as Markov processes, is better suited for pricing interest rate derivatives than earlier models developed in [1], [11], and [14]. Previous work expresses short rate processes as functions of Markov processes. While these risk-neutral models are convenient for pricing calculations, often the asset prices do not explicitly appear in derivative price formulas.

In contrast, this paper studies forward rate processes which have a characteristic trait matching market data analysis; forward rates have a common volatility. In addition, asset prices determine the basic LIBOR derivative prices within these models. We have presented caplet and swaption formulas as explicit functions of the underlying bond prices.

This suggests questions for future work. First, hedging positions could be calculated for caps and swaptions and tested against market derivative price data to determine the viability of hedging strategies for commercial use. Second, PDE solutions to many pricing questions can be developed. Each discounted asset satisfies a simple SDE, (2.3), and functions of these prices form local martingales when the appropriate PDE is satisfied.

In addition, these models may be generalized to include stochastic choices for the common rate volatility. If $\sigma(t, \omega)$ is a Markov process, the pair (asset price, $\sigma(t, \omega)$ ) retains its Markov character, and it seems possible to generate other tractable pricing models.
7. Appendix I. Figure 2 uses Federal Reserve constant maturity data for zero coupon bonds. The graph was prepared from daily quotes of various bond prices. The 20-day trailing standard deviations of relative bond returns were computed and plotted in Figure 2. The plots show high volatility for long maturities compared to low volatility of short maturities.

July 2000 bond volatilities


Figure 2. 30 trading days, from July to mid-August 2000.
Figure 3 displays the volatility common to several maturities. The plot was obtained from the same data that produced Figure 2. In this case the relative bond returns were adjusted by dividing by the logarithm of the bond price, and then 20-day trailing standard deviations were computed and plotted for this adjusted data. Figure 3 shows that the plots coalesce.


Figure 3. 30 trading days, with adjusted relative returns.

## 8. Appendix II.

8.1. Proof of Theorem 3.1. As mentioned in section 3, the Markov property of each discounted bond reduces the conditional expectation computation

$$
E\left\{\mathcal{Z} \mid \mathcal{F}_{t}\right\}
$$

to the unconditional expected value

$$
\begin{equation*}
E^{T^{*}}\left\{\frac{B\left(\tau, T_{0}\right)}{B\left(\tau, T^{*}\right)} H\left(\frac{B\left(\tau, T_{1}\right)}{B\left(\tau, T_{0}\right)}, \ldots, \frac{B\left(\tau, T_{k}\right)}{B\left(\tau, T_{0}\right)}\right)\right\} . \tag{8.1}
\end{equation*}
$$

Prescribed initial values for the discounted bond prices within the expectation produce a deterministic function of $k+1$ variables. These are set equal to discounted asset prices at time $t$ to produce the $\mathcal{F}_{t}$-measurable function.

To evaluate this expectation and, in effect, to change probabilities in order to eliminate the first factor in the integrand, we use some computations appearing in the proof of Theorem 2.3.

Let $Y_{t}=\log B\left(t, T_{0}\right) / B\left(\left(t, T^{*}\right)\right.$, and notice from (2.5) that

$$
Y_{t}-Y_{0}=\int_{0}^{t} \sigma Y_{s} d W_{s}-\frac{1}{2} \int_{0}^{t} \sigma^{2} Y_{s}^{2} d s
$$

One sees that

$$
\exp \left(Y_{t}-Y_{0}\right)=\frac{B\left(t, T_{0}\right)}{B\left(t, T^{*}\right)} \frac{B\left(0, T^{*}\right)}{B\left(0, T_{0}\right)}
$$

is a stochastic exponential; a stopped version of it will allow a change of measure. For each $n$ let $\eta_{n}$ be the first time that $Y_{t}$ reaches level $n$.

The stopped process $\Lambda_{t}=\exp \left(Y_{t \wedge \eta_{n}}-Y_{0}\right)$ is a bounded stochastic exponential and is therefore a martingale. We apply the Girsanov theorem. If $Q$ denotes the probability measure

$$
d Q=\Lambda_{T_{0}} d P^{T^{*}}
$$

on the $\sigma$-algebra $\mathcal{F}_{T_{0}}$, the process

$$
\tilde{W}_{t}=W_{t}-\sigma \int_{0}^{t \wedge \eta_{n}} Y_{s} d s
$$

is a Brownian motion. Although each process $\tilde{W}_{t}$ depends on $n$, we use these Brownian motions only to obtain an integral identity where $n$ appears explicitly. Therefore, in some basic calculations, we will suppress the $n$-dependence and merely use the notation $\tilde{W}_{t}$.

First, consider the new dynamics for $Y_{t}$ up to the time $\eta_{n}$ :

$$
\begin{aligned}
d Y=- & \sigma^{2} \frac{1}{2} Y^{2} d t+\sigma Y\left\{d \tilde{W}_{t}+\sigma Y_{t} d t\right\} \\
= & \sigma^{2} \frac{1}{2} Y^{2} d t+\sigma Y d \tilde{W}_{t}
\end{aligned}
$$

Its explicit solution is

$$
\begin{equation*}
Y=\frac{Y_{0} \tilde{M}_{t}}{1-\sigma^{2} Y_{0} \frac{1}{2} \int_{0}^{t} \tilde{M}_{s} d s} \tag{8.2}
\end{equation*}
$$

Recall that $\eta_{n}$ is the first time for this process to hit level $n$. We next find explicit formulas for each ratio $B\left(t, T_{i}\right) / B\left(t, T_{0}\right)$ in terms of $\tilde{W}_{t}$. Let $Y_{t}^{i}$ denote the logarithm of this bond ratio. Then $Y^{i}=L-Y$, where $L=\log B\left(t, T_{i}\right) / B\left(t, T^{*}\right)$.

$$
\begin{aligned}
d Y^{i} & =d L-d Y=-\frac{1}{2} \sigma^{2}\left\{L^{2}+Y^{2}\right\} d t+\sigma\left\{L d W_{t}-Y d \tilde{W}_{t}\right\} \\
& =-\frac{1}{2} \sigma^{2}\left\{L^{2}+Y^{2}\right\} d t+\sigma\left\{L\left(d \tilde{W}_{t}+\sigma Y d t\right)-Y d \tilde{W}_{t}\right\} \\
& =-\frac{1}{2} \sigma^{2}\left\{L^{2}+Y^{2}-2 L Y\right\} d t+\sigma\left\{L d \tilde{W}_{t}-Y d \tilde{W}_{t}\right\}
\end{aligned}
$$

so that

$$
d Y^{i}=-\frac{1}{2} \sigma^{2}\left(Y^{i}\right)^{2} d t+\sigma Y^{i} d \tilde{W}_{t}
$$

This is exactly the dynamics for the $\log$ of the discounted $T_{i}$-maturity bond process with respect to a $T_{0}$-forward risk-neutral measure. In those cases where $T_{i}>T_{0}$, the initial value $Y_{0}^{i}$ is negative, and the explicit solution, which is given by the formula (2.2), is negative and there is a possibility of an explosion of $Y^{i}$ to $-\infty$. Nevertheless, the discounted bond price $\exp \left(Y^{i}\right)$ is a bounded martingale which collapses to zero at this random time.

Finally, $-Y_{t}$, explicitly given in (8.2), also has the correct dynamics with respect to an extended use of a $T_{0}$-forward risk-neutral measure. Equation (8.2) states that

$$
\begin{equation*}
-Y=\frac{-Y_{0} \tilde{M}_{t}}{1+\sigma^{2}\left(-Y_{0}\right) \frac{1}{2} \int_{0}^{t} \tilde{M}_{s} d s}, \tag{8.3}
\end{equation*}
$$

matching the formula (2.2) for $\log B\left(t, T^{*}\right) / B\left(t, T_{0}\right)$. Even though we have explicit formulas for all bond ratios in terms of the ( $n$-dependent) Brownian motion $\tilde{W}$, we prefer to express an identity for an approximate value of the expected value (8.1) as

$$
\begin{gathered}
E^{T^{*}}\left\{\frac{B\left(\tau, T_{0}\right)}{B\left(\tau, T^{*}\right)} H\left(\frac{B\left(\tau, T_{1}\right)}{B\left(\tau, T_{0}\right)}, \ldots, \frac{B\left(\tau, T_{k}\right)}{B\left(\tau, T_{0}\right)}\right) ; \eta_{n}>T_{0}\right\} \\
=\frac{B\left(0, T_{0}\right)}{B\left(0, T^{*}\right)} E^{T_{0}}\left\{H\left(\frac{B\left(\tau, T_{1}\right)}{B\left(\tau, T_{0}\right)}, \ldots, \frac{B\left(\tau, T_{k}\right)}{B\left(\tau, T_{0}\right)}\right) ; \frac{B\left(\tau, T^{*}\right)}{B\left(\tau, T_{0}\right)}>\exp (-n)\right\} .
\end{gathered}
$$

We are able to use a generic Brownian motion here since the SDEs have identified all joint distributions up to the time $\eta_{n}$, which itself is expressible in terms of one bond ratio within a $T_{0}$-forward model. Now, we take a limit as $n \rightarrow \infty$ to establish the desired result.

In the first expected value $\eta_{n} \rightarrow \infty$ almost surely, whereas the distributions in the second integral confine the expectation to the event where the discounted long bond remains positive. All the other bond ratios are positive as well; this fact follows from the explicit formula in (2.2) which shows that their logarithms decrease with increasing maturity.
8.2. Derivation of put prices. The payoff for a scaled bond put on a $T^{\prime}$-maturity bond, expiring at time $T<T^{\prime}$, is

$$
\left(1-\kappa B\left(T, T^{\prime}\right)\right)^{+},
$$

where $\kappa>1$ is the reciprocal of the exercise price. Let $\mathcal{I}$ denote the indicator function

$$
I_{\left\{1 \geq \kappa B\left(T, T^{\prime}\right)\right\}} .
$$

We derive discounted price processes for two derivatives whose payoffs are $\mathcal{I}$ and $B\left(T, T^{\prime}\right) \mathcal{I}$, paid at time $T$. The discounted payoff function $\mathcal{I}$ has the form

$$
\frac{B(T, T)}{B\left(T, T^{*}\right)} H\left(\frac{B\left(T, T^{\prime}\right)}{B(T, T)}\right),
$$

where

$$
H=I_{\left\{1 \geq \kappa B\left(T, T^{\prime}\right) / B(T, T)\right\}} .
$$

We apply Theorem 3.1 and identify the deterministic function $G\left(z_{0}, z_{1}\right)$ whose formula is given in (3.4):

$$
\begin{gathered}
G=z_{0}^{-1} E^{T}\left\{H\left(\frac{B\left(\tau, T^{\prime}\right)}{B(\tau, T)}\right) ; \frac{B\left(\tau, T^{*}\right)}{B(\tau, T)}>0\right\} \\
=\exp (d) \operatorname{Pr}\left\{-\log \kappa \geq \frac{-x M_{\tau}}{1+\frac{1}{2} \sigma^{2}(-x) \int_{0}^{\tau} M_{s} d s} \text { and } \frac{1}{2} \sigma^{2} d \int_{0}^{\tau} M_{s} d s<1\right\},
\end{gathered}
$$

where $x=\log B(0, T) / B\left(0, T^{\prime}\right)>0$ and $d=\log B(0, T) / B\left(0, T^{*}\right)>0$.
The final form of the payoff function is stated in terms of the process $M_{s}^{(1)}$ because the time scaling $s \rightarrow \sigma^{2} t$ produces the distributions for the general case.

The second discounted payoff function is $\left(B\left(T, T^{\prime}\right) / B\left(T, T^{*}\right)\right) \mathcal{I}$, which has the form

$$
\frac{B\left(T, T^{\prime}\right)}{B\left(T, T^{*}\right)} H\left(\frac{B(T, T)}{B\left(T, T^{\prime}\right)}\right)
$$

where

$$
H=I_{\left\{B(T, T) / B\left(T, T^{\prime}\right) \geq \kappa\right\}}
$$

The deterministic function $G\left(z_{0}, z_{1}\right)$ in Theorem 3.1 is

$$
\begin{gathered}
G=z_{0}^{-1} E^{T^{\prime}}\left\{H\left(\frac{B(\tau, T)}{B\left(\tau, T^{\prime}\right)}\right) ; \frac{B\left(\tau, T^{*}\right)}{B\left(\tau, T^{\prime}\right)>0}\right\} \\
=\exp (d) \operatorname{Pr}\left\{\frac{x M_{\tau}}{1+\frac{1}{2} \sigma^{2} x \int_{0}^{\tau} M_{s} d s} \geq \log \kappa \text { and } \frac{1}{2} \sigma^{2} d \int_{0}^{\tau} M_{s} d s<1\right\},
\end{gathered}
$$

where $x=\log B(0, T) / B\left(0, T^{\prime}\right)>0$ and $d=\log B\left(0, T^{\prime}\right) / B\left(0, T^{*}\right)>0$.
We show that the supermartingale $E\left\{\mathcal{Z}_{\mathcal{T}} \mid \mathcal{F}_{t}\right\}$ is a stochastic integral with respect to discounted asset prices and so satisfies the hypotheses of Theorem 3.3 in [2]. The Markov property allows us to express the discounted put price as

$$
\begin{equation*}
E^{T^{*}}\left\{\left(B(\tau, T) / B\left(\tau, T^{*}\right)-\kappa B\left(\tau, T^{\prime}\right) / B\left(\tau, T^{*}\right)\right)^{+}\right\} \tag{8.4}
\end{equation*}
$$

where initial prices $z_{0}$ and $z_{0}^{\prime}$ are set equal to the time- $t$ asset prices. It is sufficient to show that (8.4) is a smooth function of $z_{0}$ and $z_{0}^{\prime}$. Now $B(\tau, T) / B\left(\tau, T^{*}\right)=\exp \left(Y_{t}\right)$, where $Y_{t}$, given by $(2.4)$, is a function of $F=\log z_{0}$. We first consider the dependence of (8.4) on $\log z_{0}$ :

$$
\begin{gathered}
\frac{\partial}{\partial F}\left(\frac{B(\tau, T)}{B\left(\tau, T^{*}\right)}-\frac{\kappa B\left(\tau, T^{\prime}\right)}{B\left(\tau, T^{*}\right)}\right)^{+} \\
=\exp \left(Y_{t}\right) \frac{\partial Y}{\partial F} I_{\left\{B(\tau, T) / B\left(\tau, T^{\prime}\right) \geq \kappa\right\}} \quad \text { almost surely. }
\end{gathered}
$$

Since $x \rightarrow(x+A)^{+}$is a Lipschitz function, the expected value of the derivative is the limit of the put price difference quotient if the random variable

$$
\exp \left(Y_{t}\right) \frac{\partial Y}{\partial F}=\exp \left(Y_{t}\right) \frac{M_{t}}{\left(1+\frac{\sigma^{2}}{2} F \int_{0}^{t} M_{s} d s\right)^{2}}
$$

has finite expectation. The function $F \rightarrow Y_{t}$ is pointwise increasing for fixed $t$, and, by inspection, the function $F^{2} \partial Y_{t} / \partial F$ is also pointwise increasing. Therefore, if

$$
E\left\{\exp \left(Y_{t}\right) \frac{\partial Y}{\partial F} F^{2}\right\}=\infty
$$

for some $F_{0}$, then the expectation would be infinite for all larger $F$ values, say, to $F_{1}$. But, then $E\left\{\exp \left(Y_{t}\right) \partial Y / \partial F\right\} \equiv \infty$ over this range. By integrating over $F$ we would see that $E\left\{\exp \left(Y_{t}\right)\right\}=\infty$ at $F_{1}$, which contradicts the fact that $\exp \left(Y_{t}\right)$ is a supermartingale.

A similar argument shows that the second derivative of $\exp \left(Y_{t}\right)$ has a finite expected value, and so (8.4) has two continuous derivatives. The same type of argument applies to (8.4) as a function of the discounted $T^{\prime}$-bond.
8.3. Derivation of swaption prices. The payoff of a European swaption (see (5.1)) is

$$
\left(1-B\left(T, T_{n}\right)-K \sum_{i=1}^{n} B\left(T, T_{i}\right)\right)^{+}
$$

This may be expressed as a sum of derivative payoffs:

$$
\mathcal{I}-B\left(T, T_{n}\right) \mathcal{I}-K \sum_{i=1}^{n} B\left(T, T_{i}\right) \mathcal{I}
$$

where $\mathcal{I}$ is the indicator function of the event

$$
\begin{equation*}
B\left(T, T_{n}\right)+K \sum_{i=1}^{n} B\left(T, T_{i}\right) \leq 1 \tag{8.5}
\end{equation*}
$$

We derive a price process for each payoff amount in the sum using Theorem 3.1. The first payoff term, $\mathcal{I}$, is expressed in discounted form by introducing the factor $B(T, T) / B\left(T, T^{*}\right)$. The indicator event can be expressed as

$$
\frac{B\left(T, T_{n}\right)}{B(T, T)}+K \sum_{i=1}^{n} \frac{B\left(T, T_{i}\right)}{B(T, T)} \leq 1
$$

Therefore, the discounted payoff equals

$$
\frac{B(T, T)}{B\left(T, T^{*}\right)} H\left(\frac{B\left(T, T_{1}\right)}{B(T, T)}, \ldots, \frac{B\left(T, T_{n}\right)}{B(T, T)}\right)
$$

and the function $G\left(z_{0}, z_{1}, \ldots, z_{n}\right)$ in Theorem 3.1 is

$$
z_{0}^{-1} P^{T}\left\{\frac{B\left(\tau, T_{n}\right)}{B(\tau, T)}+K \sum_{i=1}^{n} \frac{B\left(\tau, T_{i}\right)}{B(\tau, T)} \leq 1 \text { and } \frac{B\left(\tau, T^{*}\right)}{B(\tau, T)}>0\right\}
$$

Equation (2.2) gives explicit formulas for each ratio, and the probability is

$$
\begin{gathered}
\exp \left(-x_{0}\right) P\left\{\exp \left(\frac{x_{n} M_{\tau}}{1+\frac{\sigma^{2}}{2} x_{n} \int_{0}^{\tau} M_{s} d s}\right)+K \sum_{i=1}^{n} \exp \left(\frac{x_{i} M_{\tau}}{1+\frac{\sigma^{2}}{2} x_{i} \int_{0}^{\tau} M_{s} d s}\right) \leq 1\right. \\
\text { and } \left.\frac{-\sigma^{2} x_{0}}{2} \int_{0}^{\tau} M_{s} d s<1\right\}
\end{gathered}
$$

where $x_{i}=\log B\left(0, T_{i}\right) / B(0, T), i \geq 1$, and $x_{0}=\log B\left(0, T^{*}\right) / B(0, T)$.
Each remaining payoff term in the swaption is a scaled payoff of the amount

$$
B\left(T, T_{i}\right) \mathcal{I}
$$

The discounted payoff has the factor $B\left(T, T_{i}\right) / B\left(T, T^{*}\right)$, and it is convenient to write the event in the indicator function as

$$
\frac{B\left(T, T_{n}\right)}{B\left(T, T_{i}\right)}+K \sum_{j=1}^{n} \frac{B\left(T, T_{j}\right)}{B\left(T, T_{i}\right)} \leq \frac{B(T, T)}{B\left(T, T_{i}\right)}
$$

The function $H$ in Theorem 3.1 is the indicator of this event, and the function $G$ is then

$$
\begin{gathered}
z_{0}^{-1} P^{T_{i}}\left\{\frac{B\left(T, T_{n}\right)}{B\left(T, T_{i}\right)}+K \sum_{j=1}^{n} \frac{B\left(T, T_{j}\right)}{B\left(T, T_{i}\right)} \leq \frac{B(T, T)}{B\left(T, T_{i}\right)}\right. \\
\text { and } \left.\frac{B\left(\tau, T^{*}\right)}{B\left(\tau, T_{i}\right)}>0\right\}
\end{gathered}
$$

Equation (2.2) gives the explicit formulas for this probability:

$$
\begin{gathered}
\exp \left(-x_{0}^{i}\right) \operatorname{Pr}\left\{\exp \left(\frac{x_{n}^{i} M_{\tau}}{1+\frac{\sigma^{2}}{2} x_{n}^{i} \int_{0}^{\tau} M_{s} d s}\right)+K \sum_{j=1}^{n} \exp \left(\frac{x_{j}^{i} M_{\tau}}{1+\frac{\sigma^{2}}{2} x_{j}^{i} \int_{0}^{\tau} M_{s} d s}\right)\right. \\
\left.\leq \exp \left(\frac{x_{0}^{i} M_{\tau}}{1+\frac{\sigma^{2}}{2} x_{0}^{i} \int_{0}^{\tau} M_{s} d s}\right) \text { and } \frac{-\sigma^{2} x_{0}^{i}}{2} \int_{0}^{\tau} M_{s} d s<1\right\} .
\end{gathered}
$$

Here, $x_{0}^{i}=\log \frac{B\left(0, T^{*}\right)}{B\left(0, T_{i}\right)}$ and $x_{j}^{i}=\log \frac{B\left(0, T_{j}\right)}{B\left(0, T_{i}\right)}$.
The swaption price can be expressed as a stochastic integral with respect to the (discounted) bonds involved in the swap. To see this one expresses the price as in (8.4), using the entire swaption payoff function, in discounted form, in place of the put payoff. Following the steps in the argument for the put price, one differentiates a single term in the payoff with respect to its initial value, using the Lipschitz property of $x \rightarrow(x+A)^{+}$to compute the derivative of the integrand as the product of $\exp \left(Y_{t}\right) \partial Y_{t} / \partial F$ and an indicator function. This product has a finite expected value, and its expectation, which depends continuously on $F$, is a formula for the derivative of the swaption price with respect to the initial condition $F$.

Acknowledgment. The authors wish to thank two anonymous referees for helpful comments and suggestions that enlarged the scope of the pricing method.

## REFERENCES

[1] A. Carverhill, When is the short rate Markovian?, Math. Finance, 4 (1994), pp. 305-312.
[2] A. M. G. Cox and D. G. Hobson, Local martingales, bubbles and option prices, Finance Stoch., 9 (2005), pp. 477-492.
[3] H. Geman and M. Yor, Bessel processes, Asian options, and perpetuities, Math. Finance, 2 (1993), pp. 349-375.
[4] V. Goodman and K. Kim, Exponential Martingales and Time Integrals of Brownian Motion, http:// arxiv.org/abs/math/0612034v2 (2006).
[5] D. Heath, R. Jarrow, and A. Morton, Bond pricing and the term structure of interest rates: A new methodology for contingent claim valuation, Econometrica, 60 (1992), pp. 77-105.
[6] S. L. Heston, M. Loewenstein, and G. A. Willard, Options and bubbles, Rev. Financ. Stud., 20 (2007), pp. 359-390.
[7] R. Litterman and J. Scheinkman, Common factors affecting bond returns, J. Fixed Income, 1 (1991), pp. 54-61.
[8] K. Miltersen, K. Sandmann, and D. Sondermann, Closed form solutions for term structure derivatives with log-normal interest rates, J. Finance, 52 (1997), pp. 409-430.
[9] M. Musiela and M. Rutkowski, Continuous-time term structure models: Forward measure approach, Finance Stoch., 1 (1997), pp. 261-291.
[10] M. Musiela and M. Rutkowski, Martingale Methods in Financial Modelling, Appl. Math. (N. Y.) 36, Springer-Verlag, New York, 1997.
[11] P. Ritchken and L. Sanakarasubramanian, Volatility structures of forward rates and the dynamics of the term structure, Math. Finance, 5 (1995), pp. 55-72.
[12] R. Rebonato, Interest-Rate Option Models, 2nd ed., John Wiley \& Sons, New York, 1998.
[13] R. Rebonato, Modern Pricing of Interest-Rate Derivatives, Princeton University Press, Princeton, NJ, 2002.
[14] L. C. G. Rogers, The potential approach to the term structure of interest rates and foreign exchange rates, Math. Finance, 7 (1997), pp. 157-176.
[15] L. C. G. Rogers and Z. Shi, The value of an Asian option, J. Appl. Probab., 32 (1995), pp. 1077-1088.
[16] T. Wilson, Debunking the myths, Risk, 7 (1994), pp. 67-73.
[17] M. Yor, Exponential Functionals of Brownian Motion and Related Processes, Springer-Verlag, New York, 2001.

# Convergence by Viscosity Methods in Multiscale Financial Models with Stochastic Volatility* 

Martino Bardi ${ }^{\dagger}$, Annalisa Cesaroni ${ }^{\dagger}$, and Luigi Manca ${ }^{\ddagger}$

Abstract. We study singular perturbations of a class of stochastic control problems under assumptions motivated by models of financial markets with stochastic volatilities evolving on a fast time scale. We prove the convergence of the value function to the solution of a limit (effective) Cauchy problem for a parabolic equation of HJB type. We use methods of the theory of viscosity solutions and of the homogenization of fully nonlinear PDEs. We test the result on some financial examples, such as Merton portfolio optimization problem.

Key words. singular perturbations, viscosity solutions, stochastic volatility, asymptotic approximation, portfolio optimization

AMS subject classifications. 35B25, 91B28, 93C70, 49L25
DOI. 10.1137/090748147

1. Introduction. In this paper we consider stochastic control systems with a small parameter $\varepsilon>0$ in the form

$$
\left\{\begin{array}{l}
d X_{t}=\tilde{\phi}\left(X_{t}, Y_{t}, u_{t}\right) d t+\sqrt{2} \tilde{\sigma}\left(X_{t}, Y_{t}, u_{t}\right) d W_{t},  \tag{1}\\
d Y_{t}=\frac{1}{\varepsilon} b\left(Y_{t}\right) d t+\sqrt{\frac{2}{\varepsilon}} \tau\left(Y_{t}\right) d W_{t},
\end{array}\right.
$$

where $X_{t} \in \mathbb{R}^{n}, Y_{t} \in \mathbb{R}^{m}, u_{t}$ is the control taking values in a given compact set $U, W_{t}$ is a multidimensional Brownian motion, and the components of drift and diffusion of the slow variables $X_{t}$ have the form

$$
\tilde{\phi}^{i}:=x^{i} \phi^{i}(x, y, u), \quad \tilde{\sigma}_{i j}:=x^{i} \sigma_{i}^{j}(x, y, u)
$$

with $\phi^{i}, \sigma_{i}^{j}$ bounded and Lipschitz continuous uniformly in $u$, so that $X_{t}^{i} \geq 0$ for $t>t_{o}$ if $X_{t_{o}}^{i} \geq 0$. On the fast process $Y_{t}$ we will assume that the matrix $\tau \tau^{T}$ is positive definite and a condition implying the ergodicity (see (3)). We also take payoff functionals of the form

$$
\mathbf{E}\left[e^{\lambda(t-T)} g\left(X_{T}, Y_{T}\right) \mid X_{t}=x, Y_{t}=y\right], \quad 0 \leq t \leq T, \quad \lambda \geq 0,
$$

[^18]with $g$ continuous and growing at most quadratically at infinity, and call $V^{\varepsilon}(t, x, y)$ the value function of this optimal control problem, i.e.,
$$
V^{\varepsilon}(t, x, y):=\sup _{u .} \mathbf{E}\left[e^{\lambda(t-T)} g\left(X_{T}, Y_{T}\right) \mid X_{t}=x, Y_{t}=y,(X ., Y .) \text { satisfy (1) with } u .\right] .
$$

We are interested in the limit $V$ as $\varepsilon \rightarrow 0$ of $V^{\varepsilon}$, in particular in understanding the PDE satisfied by $V$ and interpreting it as the HJB equation for an effective limit control problem. This is a singular perturbation problem for the system (1) and for the HJB equation associated with it. We treat it by methods of the theory of viscosity solutions to such equations.

Our motivations are the models of pricing and trading derivative securities in financial markets with stochastic volatility. The book by Fleming and Soner [20] is a general presentation of viscosity solution methods in stochastic control, and in Chapter 10 it gives an excellent introduction to the applications of this theory to the mathematical models of financial markets. In such markets with stochastic volatility the asset prices are affected by correlated economic factors, modeled as diffusion processes. This is motivated by empirical studies of stock price returns in which the estimated volatility exhibits random behavior. So, typically, volatility is assumed to be a function of an Itô process $Y_{t}$ driven by another Brownian motion, which is often negatively correlated with the one driving the stock prices (this is the empirically observed leverage effect; i.e., asset prices tend to go down as volatility goes up). This approach seems to have success in taking into account the so-called smile effect, due to the discrepancy between the predicted and market traded option prices, and in reproducing much more realistic returns distributions (i.e., with fatter and asymmetric tails).

An important extension of the stochastic volatility approach was introduced recently by Fouque, Papanicolaou, and Sircar in the book [23] (see, in particular, Chapter 3). The idea is trying to describe the bursty behavior of volatility: in empirical observations volatility often tends to fluctuate to a high level for a while, then to a low level for another small time period, then again at high level, and so on, for several times during the life of a derivative contract. These phenomena are also related to another feature of stochastic volatility, which is mean reversion. A mathematical framework which takes into account both bursting and mean reverting behavior of the volatility is that of multiple time scale systems and singular perturbations. In this setting volatility is modeled as a process which evolves on a faster time scale than the asset prices and which is ergodic, in the sense that it has a unique invariant distribution (the long-run distribution) and asymptotically decorellates (in the sense that it becomes independent of the initial distribution). We refer the reader to the book [23] and to the references therein for a detailed presentation of these models and for their empirical justification.

Several extensions and applications to a variety of financial problems appeared afterward; see $[31,24,25,22,43,30,41,29,37]$ and the references therein.

According to the previous discussion, stochastic control systems of the form (1) are appropriate for studying financial problems in this setting. Indeed, here the slow variables represent prices of assets or the wealth of the investor, whereas $Y_{t}$ is an ergodic process representing the volatility and evolving on a faster time scale for $\varepsilon$ small. The main example for $Y_{t}$ is the Ornstein-Uhlenbeck process. The asymptotic analysis of such systems as $\varepsilon \rightarrow 0$ then yields a simple pricing and hedging theory which provides a correction to classical Black-Scholes
formulas, taking into account the effect of uncertain and changing volatility.
Most of the papers we cited on fast mean reverting stochastic volatility use formal asymptotic expansions of the value function in powers of $\varepsilon$ and compute the first terms of the expansions by solving suitable auxiliary elliptic and parabolic PDEs. These methods are closely related to homogenization theory and can be found in earlier papers of Papanicolaou and coauthors and, e.g., in the book [9]. They are particularly fit to problems without control, such as the pricing of many options, so that the price function is smooth and satisfies a linear PDE. In these cases the accuracy of the expansion can often be proved.

There is a wide literature on singular perturbations of diffusion processes, with and without controls. For results based on probabilistic methods we refer the reader to the books [33, 32], the recent papers $[38,39,40,12]$, and the references therein. An approach based on PDEviscosity methods for the HJB equations was developed by Alvarez and one of the authors in $[1,2,3]$; see also [4] for problems with an arbitrary number of scales. It allows one to identify the appropriate limit PDE governed by the effective Hamiltonian and gives general convergence theorems of the value function of the singularly perturbed system to the solution of the effective PDE, under assumptions that include deterministic control (i.e., $\sigma \equiv 0$ and/or $\tau \equiv 0$ ) as well as differential games, deterministic and stochastic. However, this theory originating in periodic homogenization problems $[35,18]$ was developed so far for fast variables restricted to a compact set, mostly the $m$-dimensional torus. As we already observed, though, an a priori assumption of boundedness does not appear natural to model volatility in financial markets, according to the empirical data and in the discussion presented in [23] and the references therein.

The goal of this paper is extending the methods based on viscosity solutions of $[1,2,3]$ to singular perturbation problems of the form (1), including several models of mathematical finance. The main new difficulty is that the fast variables $Y_{t}$ are unbounded.

We first check that the value function $V^{\varepsilon}$ is the unique (viscosity) solution to a Cauchy problem for the HJB equation under very general assumptions on the data. In particular, the diffusion matrix of the slow variables $\sigma \sigma^{T}$ may degenerate and $V^{\varepsilon}$ may be merely continuous. The possible degeneration of the diffusion matrix $\sigma \sigma^{T}$ can also have interesting financial applications, e.g., to path-dependent options and to interest rate models in the Heath-JarrowMorton framework (see section 6.5 for more comments on this).

Next we assume that the fast subsystem

$$
\begin{equation*}
d Y_{t}=b\left(Y_{t}\right) d t+\sqrt{2} \tau\left(Y_{t}\right) d W_{t} \tag{2}
\end{equation*}
$$

has a Lyapunov-like function $w$ satisfying

$$
\begin{equation*}
-\mathcal{L} w(y) \geq k>0 \text { for }|y|>R_{0}, \quad \lim _{|y| \rightarrow+\infty} w(y)=+\infty \tag{3}
\end{equation*}
$$

where $\mathcal{L}$ is the infinitesimal generator of the process (2). We prove a Liouville property for sub- and supersolution of $\mathcal{L} v=0$, the existence of a unique invariant measure $\mu$ for (2) (by exploiting the theory of Hasminskii [28]), and some crucial properties of the effective Hamiltonian and terminal cost

$$
\bar{H}\left(x, D_{x} V, D_{x x}^{2} V\right):=\int_{\mathbb{R}^{m}} H\left(x, y, D_{x} V, D_{x x}^{2} V, 0\right) d \mu(y), \quad \bar{g}(x):=\int_{\mathbb{R}^{m}} g(x, y) d \mu(y)
$$

where $H$ is the Bellman Hamiltonian associated with the slow variables of (1) and its last entry is for the mixed derivatives $D_{x y}$. The condition (3) is easier to check and looks weaker than other known sufficient conditions for ergodicity [28, 36]. It appears also in a remark of [34], where the proof of the existence of $\mu$ is different from ours. Lions and Musiela [34] also state that (3) is indeed equivalent to the ergodicity of (2) and to the classical Lyapunov-type condition of Hasminskii [28].

Our main result is the convergence of $V^{\varepsilon}(t, x, y)$ to $V(t, x)$ as $\varepsilon \rightarrow 0$ uniformly on compact subsets of $[0, T) \times \mathbb{R}_{+}^{n} \times \mathbb{R}^{m}$, where $V$ is the unique (viscosity) solution to

$$
\begin{equation*}
-V_{t}+\bar{H}\left(x, D_{x} V, D_{x x}^{2} V\right)+\lambda V(x)=0 \quad \text { in }(0, T) \times \mathbb{R}_{+}^{n}, \tag{4}
\end{equation*}
$$

with final data $V(T, x)=\bar{g}(x)$ in $\overline{\mathbb{R}_{+}^{n}}$. Note that there is a boundary layer at the terminal time $T$ if the utility $g$ depends on $y$.

We test this convergence theorem on two examples of financial models chosen from [23]. The first is the problem of pricing $n$ assets with an $m$-dimensional vector of volatilities. The second is the Merton portfolio optimization problem with one riskless bond and $n$ risky assets. The control system driving wealth and volatility is

$$
\left\{\begin{array}{l}
d \mathcal{W}_{t}=\mathcal{W}_{t}\left(r+\sum_{i=1}^{n}\left(\alpha^{i}-r\right) u_{t}^{i}\right) d t+\sqrt{2} \mathcal{W}_{t} \sum_{i=1}^{n} u_{t}^{i} f_{i}\left(Y_{t}\right) \cdot d \bar{W}_{t}  \tag{5}\\
d Y_{t}=\frac{1}{\varepsilon} b\left(Y_{t}\right) d t+\sqrt{\frac{2}{\varepsilon}} \nu\left(Y_{t}\right) d \bar{Z}_{t}
\end{array}\right.
$$

with $\mathcal{W}_{t_{o}}=w>0$, where $\bar{W}_{t}, \bar{Z}_{t}$ are possibly correlated Brownian motions, and the value function is

$$
V^{\varepsilon}(t, w, y):=\sup _{u .} \mathbf{E}\left[g\left(\mathcal{W}_{T}, Y_{T}\right) \mid \mathcal{W}_{t}=w, Y_{t}=y\right] .
$$

Our convergence result for this problem appears to be new, to the best of our knowledge, although the formula for the limit is derived in [23] (by a different method and for $n=1$, $g$ independent of $y$; another term of an asymptotic expansion in powers of $\varepsilon$ is also computed in [23]). We also show that we can handle a periodic day effect, i.e., $f_{i}=f_{i}\left(\frac{t}{\varepsilon}, Y_{t}\right)$ periodic in the first entry, as in section 10.2 of [23], and the presence of a component of the volatility evolving on a very slow time scale (dependent or not on $\varepsilon$ ), as in [25, 37]. A similar result for the infinite horizon Merton problem of optimal consumption [19, 20] is under investigation.

Finally we observe that our methods work if an additional unknown disturbance $\tilde{u}_{t}$ affects the dynamics of $X_{t}$ and we maximize the payoff under the worst possible behavior of $\tilde{u}_{t}$. This situation is modeled as a 0 -sum differential game: its value function is characterized by a Hamilton-Jacobi-Isaacs PDE that can be analyzed in the framework of viscosity solutions $[21,3]$. In $[1,2,3]$ the disturbance $\tilde{u}_{t}$ and/or the controls $u_{t}$ may also affect the fast variables $Y_{t}$ (constrained to a compact set). Then there is no invariant measure and the definition of the effective Hamiltonian and terminal cost is less explicit, but the convergence theorem still holds.

Our conclusion is that the theory of viscosity solutions is the appropriate mathematical framework for fully nonlinear Bellman-Isaacs equations that provides general methods for treating singular perturbation problems (relaxed semilimits, perturbed test function method, comparison principles, etc.). These can be useful additional tools for the rigorous analysis of multiscale financial problems with stochastic volatility, in particular when some variables are
controlled, the value function is not smooth, or the complexity of the model prevents more explicit calculations.

The paper is organized as follows. Section 2 presents the standing assumptions and the HJB equation. Section 3 studies the initial value problem satisfied by $V^{\varepsilon}$. Section 4 is devoted to the ergodicity of a diffusion process in the whole spaces and the properties of the effective Hamiltonian and terminal cost. In section 5 we prove our main result, Theorem 5.1, on the convergence of $V^{\varepsilon}$ to the solution of the effective Cauchy problem. In section 6 we apply our results to a multidimensional option pricing model and to the Merton portfolio optimization problem and then illustrate some extensions. Section 7 is the conclusion.

## 2. The two-scale stochastic control problem.

2.1. The control system. We consider stochastic control problems that can be written in the form

$$
\begin{cases}d X_{t}^{i}=X_{t}^{i} \phi^{i}\left(X_{t}, Y_{t}, u_{t}\right) d t+\sqrt{2} X_{t}^{i} \sigma_{i}\left(X_{t}, Y_{t}, u_{t}\right) \cdot d W_{t}, & i=1, \ldots, n  \tag{6}\\ d Y_{t}^{k}=\frac{1}{\varepsilon} b^{k}\left(Y_{t}\right) d t+\sqrt{\frac{2}{\varepsilon}} \tau_{k}\left(Y_{t}\right) \cdot d W_{t}, & k=1, \ldots, m\end{cases}
$$

with $X_{t_{o}}^{i}=x^{i} \geq 0, Y_{t_{o}}^{k}=y^{k}$, where $\varepsilon>0, U$ is a given compact set, $\phi=\left(\phi^{1}, \ldots, \phi^{n}\right)$ : $\mathbb{R}^{n} \times \mathbb{R}^{m} \times U \rightarrow \mathbb{R}^{n}, \sigma^{i}: \mathbb{R}^{n} \times \mathbb{R}^{m} \times U \rightarrow \mathbb{R}^{r}$ are bounded continuous functions, Lipschitz continuous in $(x, y)$ uniformly with respect to $u \in U, b=\left(b^{1}, \ldots, b^{m}\right): \mathbb{R}^{m} \rightarrow \mathbb{R}^{m}, \tau_{k}: \mathbb{R}^{m} \rightarrow \mathbb{R}^{r}$ are locally Lipschitz continuous functions with linear growth, i.e.,

$$
\begin{equation*}
\text { for some } K_{c}>0 \quad|b(y)|,\left\|\tau_{k}(y)\right\| \leq K_{c}(1+|y|) \quad \forall y \in \mathbb{R}^{m}, \quad k=1, \ldots, m \tag{7}
\end{equation*}
$$

and $W_{t}$ is an $r$-dimensional standard Brownian motion. These assumptions will hold throughout the paper.

We will use the symbols $\mathbb{M}^{k, j}$ and $\mathbb{S}^{k}$ to denote, respectively, the set of $k \times j$ matrices and the set of $k \times k$ symmetric matrices, and we set

$$
\mathbb{R}_{+}^{n}:=\left\{x \in \mathbb{R}^{n}: x^{i}>0 \forall i=1, \ldots, n\right\} .
$$

To shorten the notation we call $\tilde{\phi}: \mathbb{R}^{n} \times \mathbb{R}^{m} \times U \rightarrow \mathbb{R}^{n}$ the drift of the slow variables $X_{t}$, $\tilde{\sigma} \in \mathbb{M}^{n, r}$ the matrix whose $i$ th row is $x^{i} \sigma_{i}$, and $\tau \in \mathbb{M}^{m, r}$ the matrix whose $k$ th row is $\tau_{k}$, i.e.,

$$
\tilde{\phi}^{i}:=x^{i} \phi^{i}, \quad \tilde{\sigma}_{i j}:=x^{i} \sigma_{i}^{j}, \quad \tau_{k j}:=\tau_{k}^{j}, \quad j=1, \ldots, r .
$$

Then the system (6) can be rewritten with vector notations in the form

$$
\begin{cases}d X_{t}=\tilde{\phi}\left(X_{t}, Y_{t}, u_{t}\right) d t+\sqrt{2} \tilde{\sigma}\left(X_{t}, Y_{t}, u_{t}\right) d W_{t}, & X_{t_{o}}=x \in \overline{\mathbb{R}}_{+}^{n}  \tag{8}\\ d Y_{t}=\frac{1}{\varepsilon} b\left(Y_{t}\right) d t+\sqrt{\frac{2}{\varepsilon}} \tau\left(Y_{t}\right) d W_{t}, & Y_{t_{o}}=y\end{cases}
$$

The set of admissible control functions is
$\mathcal{U}:=\{u$. progressively measurable processes taking values in $U\}$.

In the following we will assume the uniform nondegeneracy of the diffusion driving the fast variables $Y_{t}$, i.e.,
(9) $\exists e(y)>0$ such that $\xi \tau(y) \tau^{T}(y) \cdot \xi=|\xi \tau(y)|^{2} \geq e(y)|\xi|^{2} \quad$ for every $y, \xi \in \mathbb{R}^{m}$.

We will not make any nondegeneracy assumption on the matrix $\sigma$ and remark that, in any case, $\tilde{\sigma}$ degenerates near the boundary of $\mathbb{R}_{+}^{n}$.
2.2. The optimal control problem. We consider a payoff functional depending only on the position of the system at a fixed terminal time $T>0$ (Mayer problem). The utility function $g: \overline{\mathbb{R}}_{+}^{n} \times \mathbb{R}^{m} \rightarrow \mathbb{R}$ is continuous and satisfies

$$
\begin{equation*}
\exists K_{g}>0 \quad \text { such that } \sup _{y \in \mathbb{R}^{d}}|g(x, y)| \leq K_{g}\left(1+|x|^{2}\right) \quad \forall x \in \mathbb{R}_{+}^{n}, \tag{10}
\end{equation*}
$$

and the discount factor is

$$
\lambda \geq 0
$$

Therefore the value function of the optimal control problem is

$$
\begin{equation*}
V^{\varepsilon}(t, x, y):=\sup _{u \in \mathcal{U}} \mathbf{E}\left[e^{\lambda(t-T)} g\left(X_{T}, Y_{T}\right) \mid X_{t}=x, Y_{t}=y\right], \quad 0 \leq t \leq T, \tag{11}
\end{equation*}
$$

where $\mathbf{E}$ denotes the expectation. This choice of the payoff is sufficiently general for the application to finance models presented in this paper, but we could easily include in the payoff an integral term keeping track of some running costs or earnings.
2.3. The HJB equation. For a fixed control $u \in U$ the generator of the diffusion process is

$$
\operatorname{trace}\left(\tilde{\sigma} \tilde{\sigma}^{T} D_{x x}^{2}\right)+\frac{2}{\sqrt{\varepsilon}} \operatorname{trace}\left(\tilde{\sigma} \tau^{T}\left(D_{x y}^{2}\right)^{T}\right)+\tilde{\phi} \cdot D_{x}+\frac{1}{\varepsilon} \operatorname{trace}\left(\tau \tau^{T} D_{y y}^{2}\right)+\frac{1}{\varepsilon} b \cdot D_{y},
$$

where the last two terms give the generator of the fast process $Y_{t}$.
The HJB equation associated via dynamic programming with the value function of this control problem is

$$
\begin{equation*}
-V_{t}+H\left(x, y, D_{x} V, D_{x x}^{2} V, \frac{D_{x y}^{2} V}{\sqrt{\varepsilon}}\right)-\frac{1}{\varepsilon} \mathcal{L}\left(y, D_{y} V, D_{y y}^{2} V\right)+\lambda V=0 \tag{12}
\end{equation*}
$$

in $(0, T) \times \mathbb{R}_{+}^{n} \times \mathbb{R}^{m}$, where

$$
\begin{equation*}
H(x, y, p, X, Z):=\min _{u \in U}\left\{-\operatorname{trace}\left(\tilde{\sigma} \tilde{\sigma}^{T} X\right)-\tilde{\phi} \cdot p-2 \operatorname{trace}\left(\tilde{\sigma} \tau^{T} Z^{T}\right)\right\} \tag{13}
\end{equation*}
$$

with $\tilde{\sigma}$ and $\tilde{\phi}$ computed at $(x, y, u), \tau=\tau(y)$, and

$$
\begin{equation*}
\mathcal{L}(y, q, Y):=b(y) \cdot q+\operatorname{trace}\left(\tau(y) \tau^{T}(y) Y\right) . \tag{14}
\end{equation*}
$$

This is a fully nonlinear degenerate parabolic equation (strictly parabolic in the $y$ variables by the assumption (9)).

The HJB equation is complemented with the obvious terminal condition

$$
V(T, x, y)=g(x, y) .
$$

However, there is no natural boundary condition on the space boundary of the domain, i.e.,

$$
(0, T) \times \partial \mathbb{R}_{+}^{n} \times \mathbb{R}^{m}=\left\{(t, x, y): 0<t<T, x^{i}=0 \text { for some } i\right\} .
$$

We will prove in the next section that the initial boundary value problem is well posed without prescribing any boundary condition because the PDE "holds up to boundary"; namely, the value function is a viscosity solution in the set $(0, T) \times \overline{\mathbb{R}}_{+}^{n} \times \mathbb{R}^{m}$, and there is at most one such solution. The irrelevance of the space boundary $(0, T) \times \partial \mathbb{R}_{+}^{n} \times \mathbb{R}^{m}$ is essentially due to the fact that $\overline{\mathbb{R}}_{+}^{n} \times \mathbb{R}^{m}$ is an invariant set for the system (6) for all admissible control functions (almost surely); that is, the state variables cannot exit this closed domain.
2.4. The main assumption. Consider the diffusion process in $\mathbb{R}^{m}$ obtained putting $\varepsilon=1$ in (1),

$$
\begin{equation*}
d Y_{t}=b\left(Y_{t}\right) d t+\sqrt{2} \tau\left(Y_{t}\right) d W_{t} \tag{15}
\end{equation*}
$$

called the fast subsystem, and observe that its infinitesimal generator is $\mathcal{L} w:=\mathcal{L}\left(y, D_{y} w, D_{y y}^{2} w\right)$, with $\mathcal{L}$ defined by (14). We assume the following condition:

There exist $w \in \mathcal{C}\left(\mathbb{R}^{d}\right)$ and constants $k, R_{0}>0$ such that

$$
\begin{equation*}
-\mathcal{L} w \geq k \text { for }|y|>R_{0} \text { in viscosity sense, and } w(y) \rightarrow+\infty \text { as }|y| \rightarrow+\infty . \tag{16}
\end{equation*}
$$

It is reminiscent of other similar conditions about ergodicity of diffusion processes in the whole space; see, for example, $[28,9,34,12,36]$.

Remark 2.1. Condition (16) can be interpreted as a weak Lyapunov condition for the process (15) relative to the set $\left\{|y| \leq R_{0}\right\}$. Indeed, a Lyapunov function for the system (15) relative to a compact invariant set $K$ is a continuous, positive definite function $L$ such that $L(x)=0$ if and only if $x \in K$, the sublevel sets $\{y \mid L(y) \leq k\}$ are compact, and $-\mathcal{L} L(x)=l(x)$ in $\mathbb{R}^{m}$, where $l$ is a continuous function with $l=0$ on $K$ and $l>0$ outside. For more details see [28].

Example 2.1. The motivating model problem studied in [23] is the Ornstein-Uhlenbeck process with equation

$$
d Y_{t}=\left(m-Y_{t}\right) d t+\sqrt{2} \tau d W_{t}
$$

where the vector $m$ and matrix $\tau$ are constant. In this case it is immediate to check condition (16) by choosing $w(y)=|y|^{2}$ and $R_{0}$ sufficiently large.

Example 2.2. More generally, condition (16) is satisfied if

$$
\limsup _{|y| \rightarrow+\infty}\left[b(y) \cdot y+\operatorname{trace}\left(\tau \tau^{T}(y)\right)\right]<0 .
$$

Indeed, also in this case it is sufficient to choose $w(y)=|y|^{2}$. Pardoux and Veretennikov $[38,39,40]$ assume $\tau \tau^{T}$ bounded and $\lim _{|y| \rightarrow+\infty} b(y) \cdot y=-\infty$, and they call it the recurrence condition.
3. The Cauchy problem for the HJB equation. We characterize the value function $V^{\varepsilon}$ as the unique continuous viscosity solution with quadratic growth to the parabolic problem with terminal data in the form

$$
\begin{cases}-V_{t}+F\left(x, y, V, D_{x} V, \frac{D_{y} V}{\varepsilon}, D_{x x}^{2} V, \frac{D_{y y}^{2} V}{\varepsilon}, \frac{D_{x y}^{2} V}{\sqrt{\varepsilon}}\right)=0 & \text { in }(0, T) \times \mathbb{R}_{+}^{n} \times \mathbb{R}^{m},  \tag{17}\\ V(T, x, y)=g(x, y) & \text { in } \overline{\mathbb{R}}_{+}^{n} \times \mathbb{R}^{m}\end{cases}
$$

where the Hamiltonian $F: \mathbb{R}^{n} \times \mathbb{R}^{m} \times \mathbb{R} \times \mathbb{R}^{n} \times \mathbb{R}^{m} \times \mathbb{S}^{n} \times \mathbb{S}^{m} \times \mathbb{M}^{n, m} \rightarrow \mathbb{R}$ is defined as

$$
\begin{equation*}
F(x, y, s, p, q, X, Y, Z):=H(x, y, p, X, Z)-\mathcal{L}(y, q, Y)+\lambda s . \tag{18}
\end{equation*}
$$

This is a variant of a standard result (see [20] and the references therein) where we must take care of the lack of boundary condition on $\partial \mathbb{R}_{+}^{n}$ and the unboundedness of the solution.

Proposition 3.1. For any $\varepsilon>0$, the function $V^{\varepsilon}$ defined in (11) is the unique continuous viscosity solution to the Cauchy problem (17) with at most quadratic growth in $x$ and $y$. Moreover the functions $V^{\varepsilon}$ are locally equibounded.

Proof. The proof is divided into several steps.
Step 1 (bounds on $V^{\varepsilon}$ ). Observe that, using the definition of $V^{\varepsilon}$ and (10),

$$
\left|V^{\varepsilon}(t, x, y)\right| \leq K_{g} \mathbf{E}\left(1+\left|X_{T}(t, x, y)\right|^{2}\right) .
$$

So, using standard estimates on the second moment of the solution to (49) (see, for instance, [27, Thms. 1 and 4, Chap. 2] or [20, App. D]) and the boundedness of $\tilde{\phi}$ and $\tilde{\sigma}$ with respect to $y$, we get that there exist $C, c>0$ such that

$$
\begin{equation*}
\left|V^{\varepsilon}(t, x, y)\right| \leq C e^{c T}\left(1+|x|^{2}\right)=K_{V}\left(1+|x|^{2}\right), \quad t \in[0, T], \quad x \in \mathbb{R}_{+}^{n}, \quad y \in \mathbb{R}^{m} . \tag{19}
\end{equation*}
$$

This estimate in particular implies that the sequence $V^{\varepsilon}$ is locally equibounded.
Step 2 (the semicontinuous envelopes are sub- and supersolutions). We define the lower and upper semicontinuous envelopes of $V^{\varepsilon}$ as

$$
\begin{gathered}
V_{*}^{\varepsilon}(t, x, y)=\liminf _{\left(t^{\prime}, x^{\prime}, y^{\prime}\right) \rightarrow(t, x, y)} V^{\varepsilon}\left(t^{\prime}, x^{\prime}, y^{\prime}\right), \\
\left(V^{\varepsilon}\right)^{*}(t, x, y)=\limsup _{\left(t^{\prime}, x^{\prime}, y^{\prime}\right) \rightarrow(t, x, y)} V^{\varepsilon}\left(t^{\prime}, x^{\prime}, y^{\prime}\right),
\end{gathered}
$$

where $\left(t^{\prime}, x^{\prime}, y^{\prime}\right) \in\left([0, T] \times \overline{\mathbb{R}_{+}^{n}} \times \mathbb{R}^{m}\right)$. By definition, $V_{*}^{\varepsilon}(t, x, y) \leq V^{\varepsilon}(t, x, y) \leq\left(V^{\varepsilon}\right)^{*}(t, x, y)$ and, moreover, both $V_{*}^{\varepsilon}$ and $\left(V^{\varepsilon}\right)^{*}$ satisfy the growth condition (19). A standard argument in viscosity solution theory, based on the dynamic programming principle (see, e.g., [20, Chap. V, sect. 2]), gives that $V_{*}^{\varepsilon}$ and $\left(V^{\varepsilon}\right)^{*}$ are, respectively, a viscosity supersolution and a viscosity subsolution to (17) at every point $(t, x, y) \in(0, T) \times \mathbb{R}_{+}^{n} \times \mathbb{R}^{m}$.

Step 3 (behavior of $V_{*}^{\varepsilon}$ and $\left(V^{\varepsilon}\right)^{*}$ at time $T$ ). We show that the value function $V_{\varepsilon}$ attains continuously the final data (locally uniformly with respect to $(x, y)$ ). This means that $\lim _{t \rightarrow T} V^{\varepsilon}(t, x, y)=g(x, y)$ locally uniformly in $(x, y) \in \overline{\mathbb{R}}_{+}^{n} \times \mathbb{R}^{m}$. This result is well known and follows from (10), (19), and the continuity in the mean square of $X_{t}, Y_{t}$. Indeed for every
$K>0$ and $\delta>0$ there exists a constant $C(K, \delta)$ depending also on the Lipschitz constants of the coefficients of the equation (see [27, Thms. 1 and 4, Chap. 2] or [20, App. D]), such that

$$
\mathbf{P}\left(\left|X_{T}-x\right| \geq \delta \mid X_{t}=x, Y_{t}=y\right), \quad \mathbf{P}\left(\left|Y_{T}-y\right| \geq \delta \mid X_{t}=x, Y_{t}=y\right) \leq C(K, \delta)(T-t)
$$

for all $x \in \mathbb{R}_{+}^{n}, y \in \mathbb{R}^{m}$ such that $|x|,|y| \leq K$. Define $A:=\left\{\left|X_{T}-x\right| \geq \delta\right\} \cup\left\{\left|Y_{T}-y\right| \geq \delta\right\}$ so that

$$
\mathbf{P}\left(A \mid X_{t}=x, Y_{t}=y\right) \leq 2 C(K, \delta)(T-t) .
$$

Then for every $\eta>0$ there exists an admissible control $u$ such that

$$
\begin{gather*}
\left|V^{\varepsilon}(t, x, y)-V^{\varepsilon}(T, x, y)\right| \leq \mathbf{E}\left(\left|g\left(X_{T}^{u}, Y_{T}\right)-g(x, y)\right| \mid X_{t}=x, Y_{t}=y\right)+\eta \\
\leq \mathbf{E}\left(\chi_{\Omega \backslash A}\left|g\left(X_{T}^{u}, Y_{T}\right)-g(x, y)\right| \mid X_{t}=x, Y_{t}=y\right)+\eta  \tag{20}\\
+2^{1 / 2} C(K, \delta)^{1 / 2}(T-t)^{1 / 2}\left(\mathbf{E}\left(\left|g\left(X_{T}^{u}, Y_{T}\right)-g(x, y)\right|^{2} \mid X_{t}=x, Y_{t}=y\right)\right)^{1 / 2} \tag{21}
\end{gather*}
$$

Term (21) can be computed using (10) and the estimates on the mean square of $X_{T}$ and $Y_{T}$ in terms of the initial data:

$$
\begin{aligned}
(21) \leq & {\left[2 C(K, \delta)(T-t)\left(2 K_{g}\right)\right]^{1 / 2}\left[\left(1+|x|^{2}\right)+\left(\mathbf{E}\left(1+\left|X_{T}\right|^{2} \mid X_{t}=x, Y_{t}=y\right)\right)^{1 / 2}\right] } \\
& \leq 2 C(K, \delta)^{1 / 2}(T-t)^{1 / 2} K_{g}^{1 / 2} C\left(1+|x|^{2}\right) \leq H(K, \delta, g)(T-t)^{1 / 2} \rightarrow 0
\end{aligned}
$$

uniformly as $T \rightarrow t$. Term (20) can be estimated as follows:

$$
(20) \leq \mathbf{E}\left(\omega_{g, K}\left(\left|X_{T}^{u}-x\right|,\left|Y_{T}-y\right|\right) \mid X_{t}=x, Y_{t}=y\right)+\eta \rightarrow \eta
$$

uniformly as $T \rightarrow t$, where $\delta<K$ and $\omega_{g, K}$ is the continuity modulus of $g$ restricted to $\{(x, y)||x| \leq 2 K,|y| \leq 2 K\}$. We conclude by the arbitrariness of $\eta$.

Finally, using the definitions, it is easy to show that $V_{*}^{\varepsilon}(T, x, y)=\left(V^{\varepsilon}\right)^{*}(T, x, y)=g(x, y)$ for every $(x, y) \in \mathbb{R}_{+}^{n} \times \mathbb{R}^{m}$.

Step 4 (behavior of $V_{*}^{\varepsilon}$ and $\left(V^{\varepsilon}\right)^{*}$ at the boundary of $\mathbb{R}_{+}^{n}$ ). We check that all the points of the boundary of $\mathbb{R}_{+}^{n}$ are irrelevant, according to Fichera-type classification of boundary points for elliptic problems. This means the following. Suppose that $\phi$ is smooth and $\left(V^{\varepsilon}\right)^{*}-\phi$ has a local maximum (resp., $V_{*}^{\varepsilon}-\phi$ has a local minimum) relative to $(0, T) \times \overline{\mathbb{R}_{+}^{n}} \times \mathbb{R}^{m}$ at $(\bar{t}, \bar{x}, \bar{y})$ with the $i$ th coordinate $\bar{x}^{i}=0$ for some $i \in\{1, \ldots, n\}$ and $0<\bar{t}<T$. Then

$$
\begin{equation*}
-\phi_{t}+F\left(\bar{x}, \bar{y}, V, D_{x} \phi, \frac{D_{y} \phi}{\varepsilon}, D_{x x}^{2} \phi, \frac{D_{y y}^{2} \phi}{\varepsilon}, \frac{D_{x y}^{2} \phi}{\sqrt{\varepsilon}}\right) \leq 0 \quad(\text { resp., } \geq 0) \quad \text { at }(\bar{t}, \bar{x}, \bar{y}) \tag{22}
\end{equation*}
$$

We give the proof of this claim only for the subsolution inequality and for the case that only two components, say $\bar{x}^{1}$ and $\bar{x}^{2}$, are null. All the other cases can be proved in the same way with obvious changes.

Therefore we fix $(\bar{t}, \bar{x}, \bar{y})$ with $0<\bar{t}<T, \bar{x} \in \mathbb{R}^{n}$ with $\bar{x}^{1}=\bar{x}^{2}=0$ and $\bar{x}^{i}>0$, for $i \neq 1,2, \bar{y} \in \mathbb{R}^{m}$, and a smooth function $\psi$ such that the maximum of $\left(V^{\varepsilon}\right)^{*}-\psi$ in $\bar{B}=B((\bar{t}, \bar{x}, \bar{y}), r) \cap\left([0, T] \times \overline{\mathbb{R}_{+}^{n}} \times \mathbb{R}^{m}\right)$ is attained at $(\bar{t}, \bar{x}, \bar{y})$. Without loss of generality we
can assume that the maximum is strict, $\bar{x}^{i}>r$ for every $i=3, \ldots, n$, and $0<\bar{t}-r<\bar{t}+r<T$. For $\delta>0$ we define

$$
\psi_{\delta}(t, x, y):=\psi(t, x, y)+\frac{\delta}{x^{1}}+\frac{\delta}{x^{2}}
$$

and $\left(t_{\delta}, x_{\delta}, y_{\delta}\right)$ a maximum point of $\left(V^{\varepsilon}\right)^{*}-\psi_{\delta}$ in $\bar{B}$. Note that $x_{\delta} \in \mathbb{R}_{+}^{n}$ and $0<t_{\delta}<T$. By taking a subsequence we can assume that

$$
\left(t_{\delta}, x_{\delta}, y_{\delta}\right) \rightarrow(\tilde{t}, \tilde{x}, \tilde{y}) \in \bar{B} \quad \text { and } \quad\left(\left(V^{\varepsilon}\right)^{*}-\psi_{\delta}\right)\left(t_{\delta}, x_{\delta}, y_{\delta}\right) \rightarrow s \quad \text { as } \delta \rightarrow 0
$$

Observe that, since $\left(V^{\varepsilon}\right)^{*}-\psi_{\delta} \leq\left(V^{\varepsilon}\right)^{*}-\psi$ by definition, we get

$$
s \leq\left(\left(V^{\varepsilon}\right)^{*}-\psi\right)(\tilde{t}, \tilde{x}, \tilde{y}) \leq\left(\left(V^{\varepsilon}\right)^{*}-\psi\right)(\bar{t}, \bar{x}, \bar{y}) .
$$

Moreover, for $\delta<r^{2}$, we get

$$
\left(\left(V^{\varepsilon}\right)^{*}-\psi_{\delta}\right)\left(t_{\delta}, x_{\delta}, y_{\delta}\right) \geq\left(\left(V^{\varepsilon}\right)^{*}-\psi_{\delta}\right)\left(\bar{t}, \sqrt{\delta}, \sqrt{\delta}, \bar{x}^{3}, \ldots, \bar{x}^{n}, \bar{y}\right) .
$$

By letting $\delta \rightarrow 0$ we obtain $s \geq\left(\left(V^{\varepsilon}\right)^{*}-\psi\right)(\bar{t}, \bar{x}, \bar{y})$. Therefore,

$$
(\tilde{t}, \tilde{x}, \tilde{y})=(\bar{t}, \bar{x}, \bar{y}), \quad s=\left(\left(V^{\varepsilon}\right)^{*}-\psi\right)(\bar{t}, \bar{x}, \bar{y}) \quad \text { and } \quad \frac{\delta}{x_{\delta}^{1}}, \frac{\delta}{x_{\delta}^{2}} \rightarrow 0 \quad \text { as } \delta \rightarrow 0
$$

Now we use the fact that $\left(V^{\varepsilon}\right)^{*}$ is a subsolution to (17), that $\left(V^{\varepsilon}\right)^{*}-\psi_{\delta}$ has a maximum at ( $t_{\delta}, x_{\delta}, y_{\delta}$ ), and that $x_{\delta} \in \mathbb{R}_{+}^{n}$ and $0<t_{\delta}<T$, so the PDE holds at such a point. We get

$$
\begin{equation*}
-\psi_{t}+H\left(x_{\delta}, y_{\delta}, D_{x} \psi-\delta p_{\delta}, D_{x x}^{2} \psi+2 \delta X_{\delta}, \frac{D_{x y}^{2} \psi}{\sqrt{\varepsilon}}\right)-\frac{1}{\varepsilon} \mathcal{L}\left(y_{\delta}, D_{y} \psi, D_{y y}^{2} \psi\right)+\lambda\left(V^{\varepsilon}\right)^{*} \leq 0 \tag{23}
\end{equation*}
$$

where all the derivatives of $\psi$ and $\left(V^{\varepsilon}\right)^{*}$ are computed at $\left(t_{\delta}, x_{\delta}, y_{\delta}\right)$,

$$
p_{\delta}:=\left(\frac{1}{\left(x_{\delta}^{1}\right)^{2}}, \frac{1}{\left(x_{\delta}^{2}\right)^{2}}, 0, \ldots, 0\right),
$$

and $X_{\delta}$ is the diagonal matrix with

$$
\left(X_{\delta}\right)_{i i}=\frac{1}{\left(x_{\delta}^{i}\right)^{3}} \quad \text { for } i=1,2 ; \quad\left(X_{\delta}\right)_{i i}=0 \quad \text { for } i=3, \ldots, n
$$

By the definition of $H, \tilde{\phi}$, and $\tilde{\sigma}$, the second term on the left-hand side of (23) is

$$
\begin{align*}
& \min _{u \in U}\left\{-\tilde{\phi} D_{x} \psi+\frac{\delta}{x_{\delta}^{1}} \phi^{1}+\frac{\delta}{x_{\delta}^{2}} \phi^{2}-\operatorname{trace}\left(\tilde{\sigma} \tilde{\sigma}^{T} D_{x x}^{2} \psi\right)\right.  \tag{24}\\
& \left.-\frac{2 \delta}{x_{\delta}^{1}}\left|\sigma_{1}\right|^{2}-\frac{2 \delta}{x_{\delta}^{2}}\left|\sigma_{2}\right|^{2}-\frac{2}{\sqrt{\varepsilon}} \operatorname{trace}\left(\tau \tilde{\sigma}^{T} D_{x, y}^{2} \psi\right)\right\},
\end{align*}
$$

where $\tilde{\phi}, \phi^{i}, \tilde{\sigma}, \sigma_{i}$ are computed at $\left(x_{\delta}, y_{\delta}, u\right)$, the derivatives of $\psi$ at $\left(t_{\delta}, x_{\delta}, y_{\delta}\right)$, and $\tau$ at $y_{\delta}$. Since $\delta / x_{\delta}^{i} \rightarrow 0$ as $\delta \rightarrow 0$ for $i=1,2$, the quantity in (24) tends to

$$
H\left(\bar{x}, \bar{y}, D_{x} \psi, D_{x x}^{2} \psi, \frac{D_{x y}^{2} \psi}{\sqrt{\varepsilon}}\right)
$$

where all the derivatives are computed at $(\bar{t}, \bar{x}, \bar{y})$. Therefore the limit of (23) as $\delta \rightarrow 0$ gives (22) at $(\bar{t}, \bar{x}, \bar{y})$, as desired.

Step 5 (comparison principle and conclusion). We now use a recent comparison result between sub- and supersolutions to parabolic problems satisfying the quadratic growth condition

$$
|V(t, x, y)| \leq C\left(1+|x|^{2}+|y|^{2}\right)
$$

proved in [15, Thm. 2.1]. We already observed that the estimate (19) holds also for $V_{*}^{\varepsilon}$ and $\left(V^{\varepsilon}\right)^{*}$, so they both satisfy the appropriate growth condition. Moreover we proved in Step 3 that $\left(V^{\varepsilon}\right)^{*}(T, x, y)=V_{*}^{\varepsilon}(T, x, y)=g(x, y)$. The comparison result is stated in [15] for parabolic problems in the whole spaces $[0, T] \times \mathbb{R}^{k}$. Nevertheless, because of the fact that our suband supersolutions $\left(V^{\varepsilon}\right)^{*}$ and $V_{*}^{\varepsilon}$ satisfy the equation also on the boundary of $\mathbb{R}_{+}^{n}$ as proved in Step 4, their argument applies without relevant changes to our case. Therefore $\left(V^{\varepsilon}\right)^{*}(t, x, y) \leq$ $V_{*}^{\varepsilon}(t, x, y)$ for every $(t, x, y) \in\left([0, T] \times \overline{\mathbb{R}_{+}^{n}} \times \mathbb{R}^{m}\right)$. Using the definition of upper and lower envelopes and the comparison result in Step 5, we get $\left(V^{\varepsilon}\right)^{*}(t, x, y)=V_{*}^{\varepsilon}(t, x, y)=V^{\varepsilon}(t, x, y)$ for every $(t, x, y) \in\left([0, T] \times \overline{\mathbb{R}_{+}^{n}} \times \mathbb{R}^{m}\right)$. Then $V^{\varepsilon}$ is the unique continuous viscosity solution to (17) satisfying a quadratic growth condition.
4. Ergodicity of the fast variables and the effective Hamiltonian and initial data. In this section we consider an ergodic problem in $\mathbb{R}^{m}$ whose solution will be useful for defining the limit problem as $\varepsilon \rightarrow 0$ of the singularly perturbed HJB equation with terminal condition (17). We consider the diffusion process in $\mathbb{R}^{m}$

$$
\begin{equation*}
d Y_{t}=b\left(Y_{t}\right) d t+\sqrt{2} \tau\left(Y_{t}\right) d W_{t} \tag{25}
\end{equation*}
$$

and the infinitesimal generator $\mathcal{L}$ of the process $Y_{t}$. Our standing assumptions are those of section 2. It is well known that such conditions imply the existence of a unique global solution for (25) (see [27, Chap. 2, sect. 6, Thms. 3 and 4]).

The first result of this section is a Liouville property that replaces the standard strong maximum principle of the periodic case and is the key ingredient for extending some results of [3] to the nonperiodic setting.

Lemma 4.1. Consider the problem

$$
\begin{equation*}
-\mathcal{L}\left(y, D V(y), D^{2} V(y)\right)=0, \quad y \in \mathbb{R}^{m} \tag{26}
\end{equation*}
$$

under the assumption (16). Then the following hold:
(i) every bounded viscosity subsolution to (26) is constant;
(ii) every bounded viscosity supersolution to (26) is constant.

Remark 4.1. This result holds also under a weaker condition than (16), namely,

$$
\begin{equation*}
\exists w \in \mathcal{C}\left(\mathbb{R}^{m}\right) \quad \text { and } \quad R_{0}>0 \tag{27}
\end{equation*}
$$

such that $-\mathcal{L} w \geq 0$ for $|y|>R_{0} \quad$ and $\quad|w(y)| \rightarrow+\infty \quad$ as $|y| \rightarrow+\infty$.
Proof. This proof uses an argument borrowed from [34]. We start by proving (i). Let $V$ be a bounded subsolution to (26). We can assume, without loss of generality, that $V \geq 0$. Define, for every $\eta>0, V_{\eta}(y)=V(y)-\eta w(y)$, where $w$ is as in (16).

We fix $R>R_{0}$, and we claim that $V_{\eta}$ is a viscosity subsolution to (26) in $|y|>R$ for every $\eta>0$. Indeed consider $\bar{y} \in \mathbb{R}^{m},|\bar{y}|>R$, and a smooth function $\psi$ such that $V_{\eta}(\bar{y})=\psi(\bar{y})$ and $V_{\eta}-\psi$ has a strict maximum at $\bar{y}$.

Assume by contradiction that $-\mathcal{L}\left(\bar{y}, D \psi(\bar{y}), D^{2} \psi(\bar{y})\right)>0$. By the regularity of $\psi$ and of $\mathcal{L}$, there exists $0<k<R-R_{0}$ such that $-\mathcal{L}\left(y, D \psi(y), D^{2} \psi(y)>0\right.$ for every $y$ with $|y-\bar{y}| \leq k$. Now we prove that $\eta w+\psi$ is a supersolution to (26) in $B(\bar{y}, k)$. Take $\tilde{y} \in B(\bar{y}, k)$ and $\xi$ smooth such that $\eta w+\psi-\xi$ has a minimum at $\tilde{y}$. Using the fact that $w$ is a supersolution to (26) in $|y|>R_{0}$ and the linearity of the differential operator $\mathcal{L}$, we obtain

$$
\begin{gathered}
0 \leq-\mathcal{L}\left(\tilde{y}, \frac{1}{\eta} D(\xi-\psi)(\tilde{y}), \frac{1}{\eta} D^{2}(\xi-\psi)(\tilde{y})\right) \\
=-\frac{1}{\eta} \mathcal{L}\left(\tilde{y}, D \xi(\tilde{y}), D^{2} \xi(\tilde{y})\right)+\frac{1}{\eta} \mathcal{L}\left(\tilde{y}, D \psi(\tilde{y}), D^{2} \psi(\tilde{y})\right)<-\mathcal{L}\left(\tilde{y}, D \xi(\tilde{y}), D^{2} \xi(\tilde{y})\right),
\end{gathered}
$$

where in the last inequality we used that $\psi$ is a supersolution in $B(\bar{y}, k)$. Recall that by our assumption $V-(\eta w+\psi)$ has a strict maximum at $\bar{y}$ and $V(\bar{y})=(\eta w+\psi)(\bar{y})$. Then there exists $\alpha>0$ such that $V(y)-(\eta w+\psi)(y)<-\alpha$ on $\partial B(\bar{y}, k)$. A standard comparison principle gives that $V(y) \leq \eta w(y)+\psi(y)-\alpha$ on $\bar{B}(\bar{y}, k)$, a contradiction with our assumptions. This proves the claim: $V_{\eta}$ is a viscosity subsolution to (26) in $|y|>R$ for every $\eta>0$.

Now, observing that $V_{\eta}(y) \rightarrow-\infty$ as $|y| \rightarrow+\infty$, for every $\eta$ we fix $M_{\eta}>R$ such that $V_{\eta}(y) \leq \sup _{|z|=R} V_{\eta}(z)$ for every $y$ such that $|y| \geq M_{\eta}$. By the maximum principle applied in $\left\{y, R \leq|y| \leq M_{\eta}\right\}$,

$$
\begin{equation*}
V_{\eta}(y) \leq \sup _{|z|=R} V_{\eta}(z) \quad \forall|y| \geq R, \quad \forall \eta>0 . \tag{28}
\end{equation*}
$$

Next we let $\eta \rightarrow 0$ in (28) and obtain $V(y) \leq \sup _{|z|=R} V(z)$ for every $y$ such that $|y|>R$. Therefore $V$ attains its global maximum at some interior point, so it is a constant by the strong maximum principle (see [7] for its extension to viscosity subsolutions).

The proof of (ii) for bounded supersolutions $U$ is analogous, with minor changes. It is sufficient to define $U_{\eta}(y)$ as $U(y)+\eta w(y)$ and to prove that $U_{\eta} \rightarrow+\infty$ as $|y| \rightarrow+\infty$ and that it is a viscosity supersolution to (26) in $|y|>R$. So, the same argument holds exchanging the role of super- and subsolutions and using the strong minimum principle [7].

The second result is about the existence of an invariant measure.
Proposition 4.2. Under the standing assumptions, there exists a unique invariant probability measure $\mu$ on $\mathbb{R}^{m}$ for the process $Y_{t}$.

Proof. Hasminskii in [28, Chap. IV] proves that there exists an invariant probability measure for $Y_{t}$ (see Theorem IV.4.1 in [28]) if, besides the standing assumptions of section 2, the following condition is satisfied: there exists a bounded set $K$ with smooth boundary such that

$$
\begin{equation*}
\mathbf{E} \tau_{K}(y) \text { is locally bounded } \quad \text { for } y \in \mathbb{R}^{m} \backslash K \tag{29}
\end{equation*}
$$

where $\tau_{K}(y)$ is the first time at which the path of the process (25) issuing from $y$ reaches the set $K$. We claim that condition (16) implies (29), with $K=B(0, R)$, with $R>R_{0}$. We fix $w$ as in (16) and $R>R_{0}$ such that $w(y) \geq 0$ for $|y|>R$. A standard superoptimality principle for viscosity supersolutions to equation $-\mathcal{L} w \geq k$ (see, e.g., [20, sect. V.2]) implies that

$$
w(y) \geq k \mathbf{E} \tau_{K}(y)+\mathbf{E} w\left(Y_{\tau_{K}(y)}\right) \geq k \mathbf{E} \tau_{K}(y) \quad \text { for every } y \in \mathbb{R}^{m} \backslash K
$$

This gives our claim immediately, because $w$ is locally bounded.
The uniqueness of the invariant measure is a standard result under the current assumptions, because the diffusion is nondegenerate; see, e.g., [28, Cor. IV.5.2] or [16].

The previous two results-the Liouville property in Lemma 4.1 and the existence and uniqueness of the invariant measure in Proposition 4.2 - are the main tools used to define the candidate limit Cauchy problem of the singularly perturbed problem (17) as $\varepsilon \rightarrow 0$. The underlying idea is that Proposition 4.2 provides the ergodicity of the process $Y_{t}$. This property allows us to construct the effective Hamiltonian and the effective terminal data. In the following we will perform such constructions in Theorem 4.3 and Proposition 4.4 using mainly PDE methods; nevertheless it must be noted that the same results could also be obtained using direct probabilistic arguments (see Remark 4.2).

We start by showing the existence of an effective Hamiltonian giving the limit PDE. In principle, for each $(\bar{x}, \bar{p}, \bar{X})$ one expects the effective Hamiltonian $\bar{H}(\bar{x}, \bar{p}, \bar{X})$ to be the unique constant $c \in \mathbb{R}$ such that the cell problem

$$
\begin{equation*}
-\mathcal{L}\left(y, D \chi, D^{2} \chi\right)+H(\bar{x}, y, \bar{p}, \bar{X}, 0)=c \quad \text { in } \mathbb{R}^{m} \tag{30}
\end{equation*}
$$

has a viscosity solution $\chi$, called corrector (see $[35,18,1]$ ). Actually, for our approach, it is sufficient to consider, as in [2], a $\delta$-cell problem

$$
\begin{equation*}
\delta w_{\delta}-\mathcal{L}\left(y, D w_{\delta}, D^{2} w_{\delta}\right)+H(\bar{x}, y, \bar{p}, \bar{X}, 0)=0 \quad \text { in } \mathbb{R}^{m} \tag{31}
\end{equation*}
$$

whose solution $w_{\delta}$ is called approximate corrector. The next result states that $\delta w_{\delta}$ converges to $-\bar{H}$ and it is smooth.

Theorem 4.3. For any fixed $(\bar{x}, \bar{p}, \bar{X})$ and $\delta>0$ there exists a solution $w_{\delta}=w_{\delta ; \bar{x}, \bar{p}, \bar{X}}(y)$ in $\mathcal{C}^{2}\left(\mathbb{R}^{m}\right)$ of (31) such that

$$
\begin{equation*}
-\lim _{\delta \rightarrow 0} \delta w_{\delta}=\bar{H}(\bar{x}, \bar{p}, \bar{X}):=\int_{\mathbb{R}^{m}} H(\bar{x}, y, \bar{p}, \bar{X}, 0) d \mu(y) \quad \text { locally uniformly in } \mathbb{R}^{m} \tag{32}
\end{equation*}
$$

where $\mu$ is the invariant probability measure on $\mathbb{R}^{m}$ for the process $Y_{t}$.
Proof. We borrow some ideas from ergodic control theory in periodic environments; see [5].
The PDE (31) is linear with locally Lipschitz coefficients and forcing term

$$
f(y):=H(\bar{x}, y, \bar{p}, \bar{X}, 0)
$$

bounded and Lipschitz by the assumptions of section 2. The existence and uniqueness of a viscosity solution satisfying

$$
\begin{equation*}
\left|w_{\delta}(y)\right| \leq C\left(1+|y|^{2}\right) \tag{33}
\end{equation*}
$$

for some $C$ follow from the Perron-Ishii method and the comparison principle in [15] (here we are using the growth assumption (7) on the coefficients). Moreover $w_{\delta} \in \mathcal{C}^{2}\left(\mathbb{R}^{m}\right)$ by standard elliptic regularity theory.

By comparison with constant sub- and supersolutions we get the uniform bound

$$
\left|\delta w_{\delta}(y)\right| \leq \sup |f|=: C_{f}
$$

Then the functions $v_{\delta}:=\delta w_{\delta}$ are uniformly bounded and satisfy

$$
\left|\mathcal{L}\left(y, D v_{\delta}, D^{2} v_{\delta}\right)\right| \leq 2 \delta C_{f}
$$

By the Krylov-Safonov estimates for elliptic equations, in any compact set the family $\left\{v_{\delta}\right\}$ with $\delta \leq 1$ is equi-Hölder continuous for some exponent and constants depending only on $C_{f}$ and the coefficients of $\mathcal{L}$. Therefore by the Ascoli-Arzelà theorem there is a sequence $\delta_{n} \rightarrow 0$ such that $v_{\delta_{n}} \rightarrow v$ locally uniformly and

$$
\mathcal{L}\left(y, D v, D^{2} v\right)=0 \quad \text { in } \mathbb{R}^{m}
$$

in the viscosity sense. By Lemma $4.1 v$ is constant.
To complete the proof we show that on any subsequence the limit of $v_{\delta}:=\delta w_{\delta}$ is the same and it is given by the formula (32). We claim that

$$
\begin{equation*}
w_{\delta}(y)=\mathbf{E} \int_{0}^{+\infty} f\left(Y_{t}\right) e^{-\delta t} d t \tag{34}
\end{equation*}
$$

where $Y_{t}$ is the process defined by the fast subsystem (25) with initial condition $Y_{0}=y$. In fact, the right-hand side is a viscosity solution of (31) by Itô's rule and other standard arguments [20]. Moreover it is bounded by $C_{f} / \delta$, and so the growth assumption (33) is satisfied. Therefore it is the viscosity solution of (31) by the comparison principle in [15], which proves the claim. Next we recall that by definition of invariant measure

$$
\mathbf{E} \int_{\mathbb{R}^{m}} f\left(Y_{t}\right) d \mu(y)=\int_{\mathbb{R}^{m}} f(y) d \mu(y) \quad \forall t>0
$$

As a consequence, by integrating both sides of (34) with respect to $\mu$ and exchanging the order of integration we get

$$
\int_{\mathbb{R}^{m}} w_{\delta}(y) d \mu(y)=\int_{0}^{+\infty} \int_{\mathbb{R}^{m}} f(y) d \mu(y) e^{-\delta t} d t=\frac{\int_{\mathbb{R}^{m}} f(y) d \mu(y)}{\delta} .
$$

Therefore the constant limit $v$ of $\delta w_{\delta}$ must be $\int_{\mathbb{R}^{m}} f(y) d \mu(y)$.
We end this section by defining the effective terminal value for the limit as $\varepsilon \rightarrow 0$ of the singular perturbation problem (17). We fix $\bar{x}$ and consider the following Cauchy initial problem:

$$
\left\{\begin{array}{l}
w_{t}-\mathcal{L}\left(y, D w, D^{2} w\right)=0 \quad \text { in }(0,+\infty) \times \mathbb{R}^{m}  \tag{35}\\
w(0, y)=g(\bar{x}, y)
\end{array}\right.
$$

where $g$ satisfies assumption (10).
Proposition 4.4. Under our standing assumptions, for every $\bar{x}$ there exists a unique bounded classical solution $w(\cdot, \cdot ; \bar{x})$ to (35) and

$$
\begin{equation*}
\lim _{t \rightarrow+\infty} w(t, y ; \bar{x})=\int_{\mathbb{R}^{m}} g(\bar{x}, y) d \mu(y)=: \bar{g}(\bar{x}) \quad \text { locally uniformly in } y . \tag{36}
\end{equation*}
$$

Proof. The PDE in (35) is parabolic with coefficients which are locally Lipschitz and grow at most linearly, whereas the initial data are bounded and continuous, by the assumptions of section 2. Classical results on these equations give the existence of a bounded classical solution to the Cauchy problem (35) (see, e.g., Theorem 1.2.1 in [36] and the references therein), whereas uniqueness among viscosity solutions is given by Theorem 2.1 in [15]. This solution can be represented as $w(t, y ; \bar{x})=\mathbf{E} g\left(\bar{x}, Y_{t}\right)$, where $Y_{t}$ is the process starting at $y$ and satisfying (25). Moreover the function $w(t, y ; \bar{x})$ is uniformly continuous in every domain $\left[t_{0},+\infty\right) \times K$, where $K \subseteq \mathbb{R}^{m}$ is a compact set; see [26, Thm. 3.5] or [28, Lem. 4.6.2].

To complete the proof it is enough to show that $\bar{w}(y)=\limsup _{s \rightarrow+\infty} w(s, y ; \bar{x})$ and $\underline{w}(y)=$ $\liminf _{s \rightarrow+\infty} w(s, y ; \bar{x})$ are constants, i.e., $\bar{w}(y)=\bar{w}$ and $\underline{w}(y)=\underline{w}$ for every $y$, and that they both coincide with $\bar{g}(\bar{x})$, i.e., $\underline{w}=\bar{w}=\bar{g}(\bar{x})$.

The proof that $\bar{w}(y)$ and $\underline{w}(y)$ are constants is the same as in the periodic case, Theorem 4.2 in [3], once we replace the strong maximum (and minimum) principle with the Liouville property, Lemma 4.1.

To conclude we show that $\bar{w}=\bar{g}(\bar{x})=\underline{w}$. We detail the argument only for $\underline{w}$, since it is completely analogous for $\bar{w}$. We fix a subsequence such that $\underline{w}=\lim _{n} w\left(t_{n}, 0 ; \bar{x}\right)$ and define $w_{n}(t, y)=w\left(t+t_{n}, y ; \bar{x}\right)$. Since $w_{n}$ is equibounded and equicontinuous, by taking a subsequence we can assume that $w_{n}(t, y) \rightarrow \tilde{w}(t, y)$ locally uniformly. Note that by construction $\tilde{w}(t, y) \geq \underline{w}$ for every $(t, y)$ and $\tilde{w}(0,0)=\underline{w}$. By stability results of viscosity solutions, $\tilde{w}$ is a viscosity solution to $w_{t}-\mathcal{L}\left(y, D w, D^{2} w\right)=0$ in $(-\infty,+\infty) \times \mathbb{R}^{m}$. Then, by the strong minimum principle, we get that $\tilde{w}(0, y)=\underline{w}$ for every $y$. This means that $w\left(t_{n}, y ; \bar{x}\right)$ converges to $\underline{w}$ locally uniformly in $y$, in particular $w\left(t_{n}, y ; \bar{x}\right) \rightarrow \underline{w} \mu$-almost surely, where $\mu$ is the invariant probability measure for $Y_{t}$ (see Proposition 4.2). Moreover $\left|w\left(t_{n}, y\right)\right| \leq\|w\|_{\infty} \in L^{1}\left(\mathbb{R}^{m}, \mu\right)$ and then, by the Lebesgue theorem and the definition of invariant measure,

$$
\underline{w}=\int_{\mathbb{R}^{m}} \underline{w} d \mu(y)=\lim _{n} \int_{\mathbb{R}^{m}} \mathbf{E} g\left(\bar{x}, Y_{t_{n}}\right) d \mu(y)=\int_{\mathbb{R}^{m}} g(\bar{x}, y) d \mu(y) .
$$

Remark 4.2. The results in Theorem 4.3 and Proposition 4.4 could also be proved using direct probabilistic methods and semigroup theory.

We consider the infinitesimal generator $L$ of the Markov semigroup in $\mathcal{C}_{b}\left(\mathbb{R}^{m}\right)$ associated with the diffusion process $Y_{t}$. In this abstract setting, the cell problem (30) can be seen as the Poisson equation $L \chi=c-c(y)$, where $c(y):=H(\bar{x}, y, \bar{p}, \bar{X})$, and the $\delta$-cell problem (31) is the resolvent equation $(\delta-L) w_{\delta}=-c(y)$. Finally the initial layer problem (35) is the abstract Cauchy problem $w_{t}-L w=0, w(0, y)=g(\bar{x}, y)$ (for more details see the monograph [36]). In particular, thanks to the existence of a unique invariant probability measure $\mu$ (see Proposition 4.2), the solution of the Poisson equation $L \chi=c-c(y)$ is given by the representation formula

$$
w(y)=\int_{0}^{\infty} \int_{\mathbb{R}^{n}} f(z)(P(t, y, d z)-\mu(d z)) d t
$$

where $P(t, y, \cdot)$ are the transition probabilities associated with $Y_{t}$, provided the convergence of $P(t, y, \cdot)$ to $\mu$ is fast enough. Using the same approach and appropriate representation formulas, the convergence results (32) and (36) can be obtained as consequences of a sufficiently strong convergence result of the transition probabilities to the invariant measure.

Related results on the (exponential) convergence of the transition probabilities to the unique invariant measure were obtained in [17, Thm. 5.2] under a stronger condition than (16), namely, the existence of a positive function $w$ and positive constants $b, c$ such that $\lim _{|y| \rightarrow+\infty} w(y)=+\infty$ and $-\mathcal{L} w \geq c w-b$ in $\mathbb{R}^{m}$.
5. The convergence theorem. We state now the main result of the paper, namely, the convergence theorem for the singular perturbation problem. We will prove that the value function $V^{\varepsilon}(t, x, y)$, solution to (17), converges locally uniformly, as $\varepsilon \rightarrow 0$, to a function $V(t, x)$ which can be characterized as the unique solution of the limit problem

$$
\begin{cases}-V_{t}+\bar{H}\left(x, D_{x} V, D_{x x}^{2} V\right)+\lambda V(x)=0 & \text { in }(0, T) \times \mathbb{R}_{+}^{n},  \tag{37}\\ V(T, x)=\bar{g}(x) & \text { in } \overline{\mathbb{R}_{+}^{n}}\end{cases}
$$

The Hamiltonian $\bar{H}$ and the terminal data $\bar{g}$ have been defined, respectively, in (32) and in (36) as the averages of $H$ (see (13)) and $g$ with respect to the unique invariant measure $\mu$ for the process $Y_{t}$, defined in (15).

Theorem 5.1. The solution $V^{\varepsilon}$ to (17) converges uniformly on compact subsets of $[0, T) \times$ $\overline{\mathbb{R}_{+}^{n}} \times \mathbb{R}^{m}$ to the unique continuous viscosity solution to the limit problem (37) satisfying a quadratic growth condition in $x$, i.e.,

$$
\begin{equation*}
\exists K>0 \quad \text { s.t. } \forall(t, x) \in[0, T] \times \overline{\mathbb{R}_{+}^{n}} \quad|V(t, x)| \leq K\left(1+|x|^{2}\right) . \tag{38}
\end{equation*}
$$

Moreover, if $g$ is independent of $y$, then the convergence is uniform on compact subsets of $[0, T] \times \overline{\mathbb{R}_{+}^{n}} \times \mathbb{R}^{m}$ and $\bar{g}=g$.

Proof. The proof is divided into several steps.
Step 1 (relaxed semilimits). Recall that by (19) the functions $V^{\varepsilon}$ are locally equibounded in $[0, T] \times \overline{\mathbb{R}_{+}^{n}} \times \mathbb{R}^{m}$, uniformly in $\varepsilon$. We define the half-relaxed semilimits in $[0, T] \times \overline{\mathbb{R}_{+}^{n}} \times \mathbb{R}^{m}$ (see [6, Chap. V]):

$$
\underline{V}(t, x, y)=\liminf _{\substack{\varepsilon, \rightarrow 0 \\ t^{\prime} \rightarrow t, x^{\prime} \rightarrow x, y^{\prime} \rightarrow y}} V^{\varepsilon}\left(t^{\prime}, x^{\prime}, y^{\prime}\right), \quad \bar{V}(t, x, y)=\limsup _{\substack{\varepsilon \rightarrow 0 \\ t^{\prime} \rightarrow t, x^{\prime} \rightarrow x, y^{\prime} \rightarrow y}} V^{\varepsilon}\left(t^{\prime}, x^{\prime}, y^{\prime}\right)
$$

for $t<T, x \in \overline{\mathbb{R}_{+}^{n}}$ and $y \in \mathbb{R}^{d}$, and

$$
\underline{V}(T, x, y)=\liminf _{t^{\prime} \rightarrow T^{-}, x^{\prime} \rightarrow x, y^{\prime} \rightarrow y} \underline{V}\left(t^{\prime}, x^{\prime}, y^{\prime}\right), \quad \bar{V}(T, x, y)=\limsup _{t^{\prime} \rightarrow T^{-}, x^{\prime} \rightarrow x, y^{\prime} \rightarrow y} \bar{V}\left(t^{\prime}, x^{\prime}, y^{\prime}\right)
$$

It is immediate to get by definitions that the estimates (19) hold also for $\bar{V}$ and $\underline{V}$. This means that

$$
\begin{equation*}
|\underline{V}(t, x, y)|,|\bar{V}(t, x, y)| \leq K_{V}\left(1+|x|^{2}\right) \quad \forall t \in[0, T], \quad x \in \overline{\mathbb{R}_{+}^{n}}, \quad y \in \mathbb{R}^{m} \tag{39}
\end{equation*}
$$

Step $2(\bar{V}, \underline{V}$ do not depend on $y)$. We check that $\bar{V}(t, x, y), \underline{V}(t, x, y)$ do not depend on $y$ for every $t \in[0, T)$ and $x \in \mathbb{R}_{+}^{n}$. We claim that $\bar{V}(t, x, y)$ (resp., $\left.\underline{V}(t, x, y)\right)$ is, for every $t \in(0, T)$ and $x \in \mathbb{R}_{+}^{n}$, a viscosity subsolution (resp., supersolution) to

$$
\begin{equation*}
-\mathcal{L}\left(y, D_{y} V, D_{y y}^{2} V\right)=0 \quad \text { in } \mathbb{R}^{d}, \tag{40}
\end{equation*}
$$

where $\mathcal{L}$ is the differential operator defined in (14). If the claim is true, we can use Lemma 4.1, since $\bar{V}, \underline{V}$ are bounded in $y$ according to estimates (39), to conclude that the functions $y \rightarrow \bar{V}(t, x, y), y \rightarrow \underline{V}(t, x, y)$ are constants for every $(t, x) \in(0, T) \times \mathbb{R}_{+}^{n}$. Finally, using the definition it is immediate to see that this implies that also $\bar{V}(T, x, y)$ and $\underline{V}(T, x, y)$ do not depend on $y$. We prove the claim only for $\bar{V}$, since the other case is completely analogous.

First of all we show that the function $\bar{V}(t, x, y)$ is a viscosity subsolution to (40). To do this, we fix a point $(\bar{t}, \bar{x}, \bar{y})$ and a smooth function $\psi$ such that $\bar{V}-\psi$ has a maximum at $(\bar{t}, \bar{x}, \bar{y})$. Using the definition of weak relaxed semilimits it is possible to prove (see $[6$, Lem. V.1.6]) that there exist $\varepsilon_{n} \rightarrow 0$ and $\bar{B} \ni\left(t_{n}, x_{n}, y_{n}\right) \rightarrow(\bar{t}, \bar{x}, \bar{y})$ maxima for $V^{\varepsilon_{n}}-\psi$ in $\bar{B}$ such that $V^{\varepsilon_{n}}\left(t_{n}, x_{n}, y_{n}\right) \rightarrow \bar{V}(\bar{t}, \bar{x}, \bar{y})$. Therefore, recalling that $V^{\varepsilon}$ is a subsolution to (17), we get

$$
-\psi_{t}+H\left(x_{n}, y_{n}, D_{x} \psi, D_{x x}^{2} \psi, \frac{1}{\sqrt{\varepsilon_{n}}} D_{x y}^{2} \psi\right)-\frac{1}{\varepsilon_{n}} \mathcal{L}\left(y_{n}, D_{y} \psi, D_{y y}^{2} \psi\right)+\lambda V^{\varepsilon_{n}} \leq 0
$$

where $V^{\varepsilon_{n}}$ and all the derivatives of $\psi$ are computed in $\left(t_{n}, x_{n}, y_{n}\right)$. This implies

$$
\begin{equation*}
-\mathcal{L}\left(y_{n}, D_{y} \psi, D_{y y}^{2} \psi\right) \leq \varepsilon_{n}\left[\psi_{t}-H\left(x_{n}, y_{n}, D_{x} \psi, D_{x x}^{2} \psi, \frac{1}{\sqrt{\varepsilon_{n}}} D_{x y}^{2} \psi\right)-\lambda V^{\varepsilon_{n}}\right] \tag{41}
\end{equation*}
$$

We observe that the term in square brackets is uniformly bounded with respect to $n$ in $\bar{B}$, and using the regularity properties of $\psi$ and of the coefficients in the equation we get the desired conclusion as $\varepsilon_{n} \rightarrow 0$.

We show now that if $\bar{V}(t, x, y)$ is a subsolution to (40), then for every fixed $(\bar{t}, \bar{x})$ the function $y \mapsto \bar{V}(\bar{t}, \bar{x}, y)$ is a subsolution to (40), which was our claim. To do this, we fix $\bar{y}$ and a smooth function $\phi$ such that $\bar{V}(\bar{t}, \bar{x}, \cdot)-\phi$ has a strict local maximum at $\bar{y}$ in $B(\bar{y}, \delta)$ and such that $\phi(y) \geq 1$ for all $y \in B(\bar{y}, \delta)$. We define, for $\eta>0, \phi_{\eta}(t, x, y)=\phi(y)\left(1+\frac{|x-\bar{x}|^{2}+|t-\bar{t}|^{2}}{\eta}\right)$, and we consider $\left(t_{\eta}, x_{\eta}, y_{\eta}\right)$ a maximum point of $\bar{V}-\phi_{\eta}$ in $B((\bar{t}, \bar{x}, \bar{y}), \delta)$. Repeating the same argument as in [6, Lem. II.5.17], it is possible to prove, eventually passing to subsequences, that, as $\eta \rightarrow 0,\left(t_{\eta}, x_{\eta}, y_{\eta}\right) \rightarrow(\bar{t}, \bar{x}, \bar{y})$ and $K_{\eta}:=\left(1+\frac{\left|x_{\eta}-\bar{x}\right|^{2}+\left|t_{\eta}-\bar{t}\right|^{2}}{\eta}\right) \rightarrow K>0$. Moreover, using the fact that $\bar{V}$ is a subsolution to (40), we get $-\mathcal{L}\left(y_{\eta}, K_{\eta} D \phi\left(y_{\eta}\right), K_{\eta} D^{2} \phi\left(y_{\eta}\right) \geq 0\right.$, which gives, using the linearity of $\mathcal{L}$ and passing to the limit as $\eta \rightarrow 0,-\mathcal{L}\left(\bar{y}, D \phi(\bar{y}), D^{2} \phi(\bar{y})\right) \geq 0$.

Step 3 ( $\bar{V}$ and $\underline{V}$ are sub- and supersolutions of the limit PDE). First we claim that $\bar{V}$ and $\underline{V}$ are sub- and supersolutions to the PDE in $(37)$ in $(0, T) \times \mathbb{R}_{+}^{n}$. We prove the claim only for $\bar{V}$ since the other case is completely analogous. The proof adapts the perturbed test function method introduced in [18] for the periodic setting. We fix $(\bar{t}, \bar{x}) \in\left((0, T) \times \mathbb{R}_{+}^{n}\right)$, and we show that $\bar{V}$ is a viscosity subsolution at $(\bar{t}, \bar{x})$ of the limit problem. This means that if $\psi$ is a smooth function such that $\psi(\bar{t}, \bar{x})=\bar{V}(\bar{t}, \bar{x})$ and $\bar{V}-\psi$ has a maximum at $(\bar{t}, \bar{x})$, then

$$
\begin{equation*}
-\psi_{t}(\bar{t}, \bar{x})+\bar{H}\left(\bar{x}, D_{x} \psi(\bar{t}, \bar{x}), D_{x x}^{2} \psi(\bar{t}, \bar{x})\right)+\lambda \bar{V}(\bar{t}, \bar{x}) \leq 0 \tag{42}
\end{equation*}
$$

Without loss of generality we assume that the maximum is strict in $B((\bar{t}, \bar{x}), r) \cap\left([0, T] \times \overline{\mathbb{R}_{+}^{n}}\right)$ and that $\bar{x}^{i}>r$ for every $i$ and $0<\bar{t}-r<\bar{t}+r<T$. We fix $\bar{y} \in \mathbb{R}^{m}, \eta>0$ and consider a solution $\chi=w_{\delta} \in \mathcal{C}^{2}$ of the $\delta$-cell problem (31) at $\left(\bar{x}, D_{x} \psi(\bar{t}, \bar{x}), D_{x x}^{2} \psi(\bar{t}, \bar{x})\right)$ (see Theorem 4.3) such that

$$
\begin{equation*}
\left|\delta \chi(y)+\bar{H}\left(\bar{x}, D_{x} \psi(\bar{t}, \bar{x}), D_{x x}^{2} \psi(\bar{t}, \bar{x})\right)\right| \leq \eta \quad \forall y \in B(\bar{y}, r) \tag{43}
\end{equation*}
$$

We define the perturbed test function as

$$
\psi^{\varepsilon}(t, x, y):=\psi(t, x)+\varepsilon \chi(y)
$$

Observe that

$$
\lim _{\varepsilon \rightarrow 0, t^{\prime} \rightarrow \bar{t}, x^{\prime} \rightarrow \bar{x}, y^{\prime} \rightarrow \bar{y}} V^{\varepsilon}\left(t^{\prime}, x^{\prime}, y^{\prime}\right)-\psi^{\varepsilon}\left(t^{\prime}, x^{\prime}, y^{\prime}\right)=\bar{V}(t, x)-\psi(t, x) .
$$

By a standard argument in viscosity solution theory (see [6, Lem. V.1.6]) we get that there exist sequences $\varepsilon_{n} \rightarrow 0$ and $\left(t_{n}, x_{n}, y_{n}\right) \in \bar{B}:=B((\bar{t}, \bar{x}, \bar{y}), r) \cap\left([0, T] \times \overline{\mathbb{R}_{+}^{n}} \times \mathbb{R}^{m}\right)$ such that the following hold:

$$
\begin{aligned}
& \left(t_{n}, x_{n}, y_{n}\right) \rightarrow(\bar{t}, \bar{x}, y) \text { for some } y \in B(\bar{y}, r) ; \\
& V^{\varepsilon_{n}}\left(t_{n}, x_{n}, y_{n}\right)-\psi^{\varepsilon_{n}}\left(t_{n}, x_{n}, y_{n}\right) \rightarrow \bar{V}(\bar{t}, \bar{x})-\psi(\bar{t}, \bar{x}) ; \\
& \left(t_{n}, x_{n}, y_{n}\right) \text { is a strict maximum of } V^{\varepsilon_{n}}-\psi^{\varepsilon_{n}} \text { in } \bar{B} .
\end{aligned}
$$

Then, using the fact that $V^{\varepsilon}$ is a subsolution to (17), we get

$$
\begin{equation*}
-\psi_{t}+H\left(x_{n}, y_{n}, D_{x} \psi, D_{x x}^{2} \psi, 0\right)+\lambda V^{\varepsilon_{n}}\left(t_{n}, x_{n}, y_{n}\right)-\mathcal{L}\left(y_{n}, D_{y} \chi, D_{y y}^{2} \chi\right) \leq 0 \tag{44}
\end{equation*}
$$

where the derivatives of $\psi$ and $\chi$ are computed, respectively, in $\left(t_{n}, x_{n}\right)$ and in $y_{n}$. Using the fact that $\chi$ solves the $\delta$-cell problem (31), we obtain

$$
\begin{aligned}
& -\psi_{t}\left(t_{n}, x_{n}\right)+H\left(x_{n}, y_{n}, D_{x} \psi\left(t_{n}, x_{n}\right), D_{x x}^{2} \psi\left(t_{n}, x_{n}\right), 0\right)-\delta \chi\left(y_{n}\right) \\
& \quad-H\left(\bar{x}, y_{n}, D_{x} \psi(\bar{t}, \bar{x}), D_{x x}^{2} \psi(\bar{t}, \bar{x}), 0\right)+\lambda V^{\varepsilon_{n}}\left(t_{n}, x_{n}, y_{n}\right) \leq 0
\end{aligned}
$$

By taking the limit as $n \rightarrow+\infty$ the second and third terms of the left-hand side of this inequality cancel out. Next we use (43) to replace $-\delta \chi$ with $\bar{H}-\eta$ and get that the left-hand side of (42) is $\leq \eta$. Finally, by letting $\eta \rightarrow 0$ we obtain (42).

Now we claim that $\underline{V}$ and $\bar{V}$ are, respectively, a super- and a subsolution to (37) also at the boundary of $\mathbb{R}_{+}^{n}$. In this case it is sufficient to repeat exactly the same argument of Step 4 in the proof of Proposition 3.1 to get the conclusion, recalling that the Hamiltonian $\bar{H}$ is defined as

$$
\bar{H}(x, p, X)=\int_{\mathbb{R}^{m}} \min _{u \in U}\left\{-\operatorname{trace}\left(\tilde{\sigma} \tilde{\sigma}^{T}(x, y, u) X\right)-\tilde{\phi}(x, y, u) \cdot p\right\} d \mu(y)
$$

Step 4 (behavior of $\bar{V}$ and $\underline{V}$ at time $T$ ). The arguments in this step are based on analogous results given in [2, Thm. 3] in the periodic setting, with minor corrections due to the unboundedness of our domain. We repeat briefly the proof for the convenience of the reader. We prove only the statement for subsolution, since the proof for the supersolution is completely analogous.

We fix $\bar{x} \in \overline{\mathbb{R}_{+}^{n}}$ and consider the unique bounded solution $w^{r}$ to the Cauchy problem

$$
\left\{\begin{array}{l}
w_{t}-\mathcal{L}\left(y, D w, D^{2} w\right)=0  \tag{45}\\
w(0, y)=\sup _{\{|x-\bar{x}| \leq r, x \geq 0\}} g(x, y) .
\end{array} \quad \text { in }(0,+\infty) \times \mathbb{R}^{m},\right.
$$

Using stability properties of viscosity solutions it is not hard to see that $w^{r}$ converges, as $r \rightarrow 0$, to $w_{\bar{x}}$, solution to (35), uniformly on compact sets.

We fix $k>0$. Using the definition of $\bar{g}$ given in (36) and the uniform convergence of $w^{r}$ to $w_{\bar{x}}$, it is easy to see that for every $\eta>0$ there exist $t_{0}>0$ and $r_{0}$ such that $\left|w_{r}\left(t_{0}, y\right)-\bar{g}(\bar{x})\right| \leq \eta$ for every $r<r_{0}$ and $|y| \leq k$. Moreover, since $\mathcal{L}(y, 0,0)=0$, using a comparison principle, we get that

$$
\begin{equation*}
\left|w_{r}(t, y)-\bar{g}(\bar{x})\right| \leq \eta \quad \text { for every } r<r_{0}, t \geq t_{0},|y| \leq k . \tag{46}
\end{equation*}
$$

We now fix $r<r_{0}$ and a constant $M$ such that $V^{\varepsilon}(t, x, y) \leq M$ for every $\varepsilon>0$ and $x \in$ $\bar{B}:=\overline{B(\bar{x}, r)} \cap \overline{\mathbb{R}_{+}^{n}}$. Observe that this is possible by estimates (19). Moreover we fix a smooth nonnegative function $\psi$ such that $\psi(\bar{x})=0$ and $\psi(x)+\inf _{y} g(x, y) \geq M$ for every $x \in \partial B$ (using condition (10)). Let $C$ be a positive constant such that

$$
\left|H\left(y, x, D \psi(x), D^{2} \psi(x)\right)\right| \leq C \quad \text { for } x \in \bar{B} \quad \text { and } \quad y \in \mathbb{R}^{m}
$$

where $H$ is defined in (13). We define the function

$$
\psi^{\varepsilon}(t, x, y)=w_{r}\left(\frac{T-t}{\varepsilon}, y\right)+\psi(x)+C(T-t)
$$

and we claim that it is a supersolution to the parabolic problem

$$
\begin{cases}-V_{t}+F\left(x, y, V, D_{x} V, \frac{D_{y} V}{\varepsilon}, D_{x x}^{2} V, \frac{D_{y y}^{2} V}{\varepsilon}, \frac{D_{x y}^{2} V}{\sqrt{\varepsilon}}\right)=0 & \text { in }(T-r, T) \times B \times \mathbb{R}^{m}  \tag{47}\\ V(t, x, y)=M & \text { in }(T-r, T) \times \partial B \times \mathbb{R}^{m} \\ V(T, x, y)=g(x, y) & \text { in } \bar{B} \times \mathbb{R}^{m},\end{cases}
$$

where $F$ is defined in (18). Indeed if $w_{r}$ is smooth,

$$
\begin{gathered}
-\psi_{t}^{\varepsilon}+F\left(x, y, D_{x} \psi^{\varepsilon}, \frac{D_{y} \psi^{\varepsilon}}{\varepsilon}, D_{x x}^{2} \psi^{\varepsilon}, \frac{D_{y y}^{2} V \psi^{\varepsilon}}{\varepsilon}, \frac{D_{x y}^{2} \psi^{\varepsilon}}{\sqrt{\varepsilon}}\right) \\
=\frac{1}{\varepsilon}\left(w_{r}\right)_{t}+C+H\left(y, x, D \psi(x), D^{2} \psi(x)\right)-\frac{1}{\varepsilon} \mathcal{L}\left(y, D w_{r}, D^{2} w_{r}\right) \\
\geq \frac{1}{\varepsilon}\left(\left(w_{r}\right)_{t}-\mathcal{L}\left(y, D w_{r}, D^{2} w_{r}\right)\right) \geq 0
\end{gathered}
$$

This computation is made in the case where $w_{r}$ is smooth but can be easily generalized to $w_{r}$ continuous using test functions (see [2, Thm. 3]). Moreover

$$
\psi^{\varepsilon}(T, x, y)=\sup _{|x-\bar{x}| \leq r} g(x, y)+\psi(x) \geq g(x, y)
$$

Finally, recalling that by the comparison principle, $w_{r}(t, y) \geq \inf _{y} \sup _{|x-\bar{x}| \leq r} g(x, y)$, we get

$$
\psi^{\varepsilon}(t, x, y) \geq \inf _{y} \sup _{|x-\bar{x}| \leq r} g(x, y)+M-\inf _{y} g(x, y)+C(T-t) \geq M
$$

for every $x \in \bar{B}$. For our choice of $M$, we get that $V^{\varepsilon}$ is a subsolution to (47). Moreover note that both $V^{\varepsilon}$ and $\psi^{\varepsilon}$ are bounded in $[0, T] \times \bar{B} \times \mathbb{R}^{m}$, because of the estimate (19), the
boundedness of $w_{r}$, and the regularity of $\psi$. So, a standard comparison principle for viscosity solutions gives

$$
\begin{equation*}
V^{\varepsilon}(t, x, y) \leq \psi^{\varepsilon}(t, x, y)=w_{r}\left(\frac{T-t}{\varepsilon}, y\right)+\psi(x)+C(T-t) \tag{48}
\end{equation*}
$$

for every $\varepsilon>0,(t, x, y) \in\left([0, T] \times \bar{B} \times \mathbb{R}^{m}\right)$. We compute the upper limit of both sides of (48) as $\left(\varepsilon, t^{\prime}, x^{\prime}, y^{\prime}\right) \rightarrow(0, t, x, y)$ for $t \in\left(t_{0}, T\right), x \in B,|y|<k$ and get, recalling (46),

$$
\bar{V}(t, x) \leq \bar{g}(\bar{x})+\eta+\psi_{0}(x)+C(T-t) .
$$

This permits us to conclude, taking the upper limit for $(t, x) \rightarrow(T, \bar{x})$ and recalling that $\eta$ is arbitrary.

Step 5 (uniform convergence). Observe that by definition $\underline{V} \leq \bar{V}$ and that both $\underline{V}$ and $\bar{V}$ satisfy the same quadratic growth condition (39). Moreover the Hamiltonian $\bar{H}$ defined in (32) and the terminal data $\bar{g}$ in (36) inherit all the regularity properties of $H$ in (13), and $g$ in (10), as is easily seen by their definitions. Therefore we can again use the comparison result between sub- and supersolutions to parabolic problems satisfying a quadratic growth condition, given in [15, Thm. 2.1], to deduce $\underline{V} \geq \bar{V}$. Therefore $\underline{V}=\bar{V}=: V$. In particular $V$ is continuous, and by the definition of half-relaxed semilimits, this implies that $V^{\varepsilon}$ converges locally uniformly to $V$ (see [6, Lem. V.1.9]).

Remark 5.1. The result in Theorem 5.1 still holds if the fast variables $Y_{t}$ have an extra term such as $\Lambda(y) / \sqrt{\varepsilon}$ in the drift, with $\Lambda: \mathbb{R}^{m} \rightarrow \mathbb{R}^{m}$ bounded and Lipschitz continuous. This means that fast variables in the singularly perturbed system (6) satisfy

$$
d Y_{t}^{k}=\frac{1}{\varepsilon} b^{k}\left(Y_{t}\right) d t+\frac{1}{\sqrt{\varepsilon}} \Lambda^{k}\left(Y_{t}\right) d t+\sqrt{\frac{2}{\varepsilon}} \tau_{k}\left(Y_{t}\right) \cdot d W_{t}, \quad Y_{t_{o}}^{k}=y^{k}, k=1, \ldots, m
$$

and the singularly perturbed HJB equation is

$$
-V_{t}^{\varepsilon}+H\left(x, y, D_{x} V^{\varepsilon}, D_{x x}^{2} V^{\varepsilon}, \frac{D_{x y}^{2} V^{\varepsilon}}{\sqrt{\varepsilon}}\right)-\frac{1}{\varepsilon} \mathcal{L}\left(y, D_{y} V^{\varepsilon}, D_{y y}^{2} V^{\varepsilon}\right)-\frac{\Lambda \cdot D_{y} V^{\varepsilon}}{\sqrt{\varepsilon}}+\lambda V^{\varepsilon}=0 .
$$

The new term $\frac{1}{\sqrt{\varepsilon}} \Lambda(y) \cdot D_{y} V^{\varepsilon}$ appearing in the equation is a lower order term with respect to $\frac{1}{\varepsilon} \mathcal{L}\left(y, D_{y} V^{\varepsilon}, D_{y y}^{2} V^{\varepsilon}\right)$ and does not affect the convergence argument. In particular it is sufficient to check the validity of Steps 2, 3, and 4 in the proof of Theorem 5.1.

In Step 2, we substitute formula (41) with

$$
\begin{gathered}
-\mathcal{L}\left(y_{n}, D_{y} \psi, D_{y y}^{2} \psi\right) \\
\leq \varepsilon_{n}\left[\psi_{t}-H\left(x_{n}, y_{n}, D_{x} \psi, D_{x x}^{2} \psi, \frac{1}{\sqrt{\varepsilon_{n}}} D_{x y}^{2} \psi\right)-\lambda V^{\varepsilon_{n}}\right]+\sqrt{\varepsilon_{n}} \Lambda\left(y_{n}\right) \cdot D_{y} \psi
\end{gathered}
$$

and observe that the right-hand side is vanishing as $\varepsilon_{n} \rightarrow 0$ since $D_{y} \psi$ is locally bounded and $\Lambda$ is bounded.

In Step 3, we replace formula (44) with

$$
-\psi_{t}+H\left(x_{n}, y_{n}, D_{x} \psi, D_{x x}^{2} \psi, 0\right)+\lambda V^{\varepsilon_{n}}-\mathcal{L}\left(y_{n}, D_{y} \chi, D_{y y}^{2} \chi\right) \leq \sqrt{\varepsilon_{n}} \Lambda\left(y_{n}\right) \cdot D_{y} \chi
$$

and repeat the same argument since the right-hand side is vanishing as $\varepsilon_{n} \rightarrow 0$, due again to the boundedness of $\Lambda$ and the smoothness of the approximate corrector $\chi$.

Finally in Step 4, we substitute the Cauchy problem (45) with

$$
\left\{\begin{array}{l}
w_{t}-\mathcal{L}\left(y, D w, D^{2} w\right)-\sqrt{\varepsilon} \Lambda(y) \cdot D w=0 \quad \text { in }(0,+\infty) \times \mathbb{R}^{m}, \\
w(0, y)=\sup _{\{|x-\bar{x}| \leq r, x \geq 0\}} g(x, y)
\end{array}\right.
$$

and denote with $w^{r, \varepsilon}$ its unique bounded solution. Stability properties of viscosity solutions imply that $w^{r, \varepsilon}$ converges, as $r \rightarrow 0, \varepsilon \rightarrow 0$, to $w_{\bar{x}}$, solution to (35), uniformly on compact sets.

## 6. Examples and extensions.

6.1. The model problem: Risky assets with stochastic volatility. We consider $N$ underlying risky assets with price $X^{i}$ evolving according to the standard lognormal model:

$$
\begin{cases}d X_{t}^{i}=\alpha^{i} X_{t}^{i} d t+\sqrt{2} X_{t}^{i} f_{i}\left(Y_{t}\right) \cdot d \bar{W}_{t}, & X_{t_{o}}^{i}=x^{i} \geq 0,  \tag{49}\\ d Y_{t}^{j}=\frac{1}{\varepsilon} b^{j}\left(Y_{t}\right) d t+\sqrt{\frac{2}{\varepsilon}} \nu_{j}\left(Y_{t}\right) d \bar{Z}_{t}^{j}, & Y_{t_{o}}^{j}=y^{j} \in \mathbb{R}, \quad j=1, \ldots, m, \quad \varepsilon>0,\end{cases}
$$

where $f_{i}: \mathbb{R}^{m} \rightarrow \mathbb{R}^{k}$ is a bounded Lipschitz continuous function, with each component bounded away from 0 , and $b^{i}: \mathbb{R}^{m} \rightarrow \mathbb{R}$ and $\nu_{j}: \mathbb{R}^{m} \rightarrow \mathbb{R}$ are locally Lipschitz continuous functions with linear growth (see (7)). We assume that

$$
\begin{equation*}
\nu_{j}^{2}(y)>0 \quad \forall y \in \mathbb{R}^{m}, \quad j=1, \ldots, m \tag{50}
\end{equation*}
$$

The processes $\bar{W}_{t}$ and $\bar{Z}_{t}$ are, respectively, standard $k$ - and $m$-dimensional Brownian motions, and they are correlated. In particular we assume that there exists an $m$-dimensional standard Brownian motion $Z_{t}$ such that $W_{t}=\left(\bar{W}_{t}, Z_{t}\right)$ is a $(k+m)$-dimensional standard Brownian motion and

$$
\begin{equation*}
\bar{Z}_{t}^{j}=\sum_{i=1}^{k} \rho_{i j} \bar{W}_{t}^{i}+\left(1-\sum_{i=1}^{k} \rho_{i j}^{2}\right)^{\frac{1}{2}} Z_{t}^{j} \quad \forall j=1, \ldots, m, \forall t \geq 0 . \tag{51}
\end{equation*}
$$

This model problem is essentially the one described in [23, sect. 10.6], where $k=n=m$.
We denote with $\rho$ the correlation $(k \times m)$-matrix $\left(\rho_{i j}\right)$ and with $c^{j}$ the quantity

$$
\begin{equation*}
c^{j}:=\left(1-\sum_{i=1}^{k} \rho_{i j}^{2}\right)^{\frac{1}{2}} \tag{52}
\end{equation*}
$$

In the following proposition we describe the main properties of $\rho$.
Proposition 6.1.
(i) $-1 \leq \rho_{i j} \leq 1$ for every $i \in\{1, \ldots, k\}$ and $j \in\{1, \ldots, m\}$;
(ii) $\sum_{i=1}^{k} \rho_{i j}^{2} \leq 1$ for every $j \in\{1, \ldots, m\}$;
(iii) $\sum_{i=1}^{k} \rho_{i j} \rho_{i l}=0$ for every $l \neq j \in\{1, \ldots, m\}$.

Proof. Items (i) and (ii) can be easily proved by exploiting the definition of $\rho_{i j}$. To show (iii), we multiply $\sum_{i=1}^{k} \rho_{i j} \rho_{i l}$ by $t$, for fixed $l \neq j \in\{1, \ldots, m\}$, and use the properties of $\bar{W}$. to get

$$
\begin{equation*}
t \sum_{i=1}^{k} \rho_{i j} \rho_{i l}=\mathbf{E} \sum_{i=1}^{k} \rho_{i j} \bar{W}_{t}^{i} \rho_{i l} \bar{W}_{t}^{i}=\mathbf{E}\left(\sum_{i=1}^{k} \rho_{i j} \bar{W}_{t}^{i} \sum_{i=1}^{k} \rho_{i l} \bar{W}_{t}^{i}\right) \tag{53}
\end{equation*}
$$

since the components of $\bar{W}_{t}$ are independent. Substituting (51) in (53) we get

$$
\begin{gathered}
t \sum_{i=1}^{k} \rho_{i j} \rho_{i l}=\mathbf{E}\left[\left(\bar{Z}_{t}^{j}-c^{j} Z_{t}^{j}\right)\left(\bar{Z}_{t}^{l}-c^{l} Z_{t}^{l}\right)\right] \\
=\mathbf{E}\left(\bar{Z}_{t}^{j} \bar{Z}_{t}^{l}\right)-c^{j} \mathbf{E}\left(Z_{t}^{j} \bar{Z}_{t}^{l}\right)-c^{l} \mathbf{E}\left(\bar{Z}_{t}^{j} Z_{t}^{l}\right)+c^{j} c^{l} \mathbf{E}\left(Z_{t}^{j} Z_{t}^{l}\right)=0
\end{gathered}
$$

for $j \neq l$, since the components of the Brownian motions $Z_{t}$ and $\bar{Z}_{t}$ are independent and, moreover,

$$
\mathbf{E}\left(Z_{t}^{j} \bar{Z}_{t}^{l}\right)=0
$$

as can be easily obtained using (51) and the fact that $Z_{t}$ and $\bar{W}_{t}$ are independent Brownian motions.

Substituting (51) in (49) we get

$$
\left\{\begin{array}{l}
d X_{t}=\tilde{\phi}\left(X_{t}\right) d t+\sqrt{2} \tilde{\sigma}\left(X_{t}, Y_{t}\right) d W_{t}  \tag{54}\\
d Y_{t}=\frac{1}{\varepsilon} b\left(Y_{t}\right) d t+\sqrt{\frac{2}{\varepsilon}} \tau\left(Y_{t}\right) d W_{t}
\end{array}\right.
$$

where $\tilde{\phi}: \mathbb{R}^{n} \rightarrow \mathbb{R}^{n}$ and $\tilde{\sigma}: \mathbb{R}^{n} \times \mathbb{R}^{m} \rightarrow \mathbb{M}^{n, k+m}$ are defined as $\tilde{\phi}^{i}(x)=\alpha^{i} x^{i}$ and $\tilde{\sigma}_{i j}(x, y)=$ $x^{i} f_{i}^{j}(y)$ for $j=1, \ldots, k$ and $\tilde{\sigma}_{i j}(x, y)=0$ for $j=k+1, \ldots, k+m$, while $\tau: \mathbb{R}^{m} \rightarrow \mathbb{R}^{m \times(k+m)}$ is the $(m \times(k+m))$-matrix

$$
\tau(y)=\left(\begin{array}{ccccccc}
\rho_{11} \nu_{1}(y) & \cdots & \rho_{k 1} \nu_{1}(y) & c^{1} \nu_{1}(y) & 0 & \cdots & 0  \tag{55}\\
\rho_{12} \nu_{2}(y) & \cdots & \rho_{k 2} \nu_{2}(y) & 0 & c^{2} \nu_{2}(y) & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\rho_{1 m} \nu_{m}(y) & \cdots & \rho_{k m} \nu_{m}(y) & 0 & 0 & 0 & c^{m} \nu_{m}(y)
\end{array}\right)
$$

We consider now the matrix $\tau(y) \tau^{T}(y)$. An easy computation shows that the diagonal terms of this matrix are

$$
\left(\tau(y) \tau^{T}(y)\right)_{j j}=\nu_{j}^{2}(y)\left(\sum_{i=1}^{k} \rho_{i j}^{2}+\left(c^{j}\right)^{2}\right)=\nu_{j}^{2}(y)
$$

by the definition of $c^{j}$ in (52). The extra diagonal terms are given by

$$
\left(\tau(y) \tau^{T}(y)\right)_{j l}=\nu_{j}(y) \nu_{l}(y)\left(\sum_{i=1}^{k} \rho_{i j} \rho_{i l}\right)=0
$$

by item (iii) in Proposition 6.1. Then the matrix $\tau \tau^{T}$ is the diagonal matrix

$$
\tau(y) \tau^{T}(y)=\left(\begin{array}{ccc}
\nu_{1}^{2}(y) & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & \nu_{m}^{2}(y)
\end{array}\right)
$$

and in particular satisfies (9) by (50).
Observe that the system (54) fits in our basic assumptions of section 2. It includes as a special case the multidimensional option pricing model of [23, sect. 10.6], where each $Y_{t}^{i}$ is a standard one-dimensional Ornstein-Uhlenbeck processes. Here we are only assuming, besides standard regularity conditions on $b$ and $\tau$ and nondegeneracy (50), that the infinitesimal generator of the process satisfies the Lyapunov-like condition (16).

The problem we consider here is the pricing of a European option given by a nonnegative payoff function $g$ depending on the underlying $X^{i}$ and by a maturity time $T$. According to risk-neutral theory, to define a no-arbitrage derivative price we have to use an equivalent martingale measure $\mathbf{P}^{*}$ under which the discounted stock prices $e^{-r t} X_{t}^{i}$ are martingales, where $r$ is the instantaneous interest rate for lending or borrowing money. For a brief review of noarbitrage price theory in the context of stochastic volatility we refer the reader to [23, sect. 2.5]. The system (54) can be written, under a risk-neutral probability $\mathbf{P}^{*}$, as

$$
\left\{\begin{array}{l}
d X_{t}=r X_{t} d t+\sqrt{2} \tilde{\sigma}\left(X_{t}, Y_{t}\right) d W_{t}^{*}  \tag{56}\\
d Y_{t}=\frac{1}{\varepsilon}\left[b\left(Y_{t}\right)-\sqrt{\varepsilon} \Lambda\left(Y_{t}\right)\right] d t+\sqrt{\frac{2}{\varepsilon}} \tau\left(Y_{t}\right) d W_{t}^{*}
\end{array}\right.
$$

for some volatility risk premium $\Lambda(Y)$ chosen by the market and describing the relationship between the physical measure $\mathbf{P}$ under which the stock prices are observed and the riskneutral measure $\mathbf{P}^{*}$ (see [23, sect. 10.6] and [24]). In (56) $W^{*}$ is a $(k+m)$-dimensional standard Brownian motion obtained by an appropriate shift of $W$, and $\Lambda$ can be assumed bounded and smooth. In this setting, an European contract has no-arbitrage price given by the formula

$$
\begin{equation*}
V^{\varepsilon}(t, x, y):=\mathbf{E}^{*}\left[e^{\lambda(t-T)} g\left(X_{T}\right) \mid X_{t}=x, Y_{t}=y\right], \quad 0 \leq t \leq T \tag{57}
\end{equation*}
$$

where $\lambda>0$ and the payoff function $g$ satisfies (10). When there is only one asset $X_{t}$ (say $n=1$ in the system (56)), typically the payoff function $g$ is defined as $g(x)=\max \{(x-K), 0\}$ for call options and $g(x)=\max \{(K-x), 0\}$ for put options, where $K$ is the contracted strike price.

The (linear) HJB equation associated with the price function is

$$
\begin{aligned}
-V_{t}^{\varepsilon} & +H_{P}\left(x, y, D_{x} V^{\varepsilon}, D_{x x}^{2} V^{\varepsilon}, \frac{D_{x y}^{2} V^{\varepsilon}}{\sqrt{\varepsilon}}\right)+\lambda V^{\varepsilon} \\
& =\frac{1}{\varepsilon}\left[\mathcal{L}\left(y, D_{y} V^{\varepsilon}, D_{y y}^{2} V^{\varepsilon}\right)-\sqrt{\varepsilon} \Lambda(y) \cdot D_{y} V^{\varepsilon}\right]
\end{aligned}
$$

in $(0, T) \times \mathbb{R}_{+}^{n} \times \mathbb{R}^{m}$ complemented with the obvious terminal condition

$$
V^{\varepsilon}(T, x, y)=g(x),
$$

where

$$
H_{P}(x, y, p, X, Z):=-\operatorname{trace}\left(\tilde{\sigma} \tilde{\sigma}^{T} X\right)-\phi_{r} \cdot p-2 \operatorname{trace}\left(\tilde{\sigma} \tau^{T} Z^{T}\right)
$$

and $\mathcal{L}$ is defined in (14). The prices $V^{\varepsilon}(t, x, y)$ converge locally uniformly, as $\varepsilon \rightarrow 0$, to the unique viscosity solution $V$ of the limit equation (37), due to our convergence result, Theorem 5.1 (see also Remark 5.1 describing the slight modifications to the argument in the proof needed to treat this case). $V$ can be represented as

$$
V(t, x):=\mathbf{E}^{*}\left[e^{\lambda(t-T)} g\left(X_{T}\right) \mid X_{t}=x\right], \quad 0 \leq t \leq T
$$

where $\mu$ is the unique invariant measure associated with the fast subsystem (see section 4) and $X_{t}$ satisfies the averaged effective system

$$
\begin{equation*}
d X_{t}=r X_{t} d t+\sqrt{2} \bar{\sigma}\left(X_{t}\right) d W_{t}^{*} \tag{58}
\end{equation*}
$$

whose volatility is the so-called mean historical volatility

$$
\bar{\sigma}(x):=\sqrt{\int_{\mathbb{R}^{m}} \tilde{\sigma}(x, y) \tilde{\sigma}^{T}(x, y) d \mu(y)}
$$

Therefore the limit of the pricing problem as $\varepsilon \rightarrow 0$ is a new pricing problem for the effective system (58). This convergence result complements and extends a bit section 10.6 of [23] on multidimensional problems.

Let us recall also that $\mu(y)$ is explicitly known in some interesting cases, in particular when the fast variables are an Ornstein-Uhlenbeck process, as in [23]. For instance, if $Y_{t}$ and $\bar{Z}_{t}$ are scalar processes, the measure $\mu$ has the Gaussian density

$$
d \mu(y)=\frac{1}{\sqrt{2 \pi \tau^{2}}} e^{-(y-m)^{2} / 2 \tau^{2}} d y
$$

with the notations of Example 2.1.
6.2. Merton portfolio optimization problem. We consider now another classical problem in finance, the Merton optimal portfolio allocation, under the assumption of fast oscillating stochastic volatility.

We consider a financial market consisting of a nonrisky asset $X^{0}$ evolving according to the deterministic equation $d X_{t}^{0}=r X_{t}^{0} d t$, with $r>0$, and $n$ risky assets $X_{t}^{i}$ evolving according to the stochastic system (54). We denote by $\mathcal{W}$ the wealth of an investor. The investment policy-which will be the control input-is defined by a progressively measurable process $u$ taking values in a compact set $U$, and $u_{t}^{i}$ represents the proportion of wealth invested in the asset $X_{t}^{i}$ at time $t$. Then the wealth process evolves according to the following system:

$$
\left\{\begin{array}{l}
d \mathcal{W}_{t}=\mathcal{W}_{t}\left(r+\sum_{i=1}^{n}\left(\alpha^{i}-r\right) u_{t}^{i}\right) d t+\sqrt{2} \mathcal{W}_{t} \sum_{i=1}^{n} u_{t}^{i} f_{i}\left(Y_{t}\right) \cdot d \bar{W}_{t}, \quad \mathcal{W}_{t_{o}}=w>0  \tag{59}\\
d Y_{t}=\frac{1}{\varepsilon} b\left(Y_{t}\right) d t+\sqrt{\frac{2}{\varepsilon}} \nu\left(Y_{t}\right) d \bar{Z}_{t}
\end{array}\right.
$$

with the same notations and assumptions as in section 6.1. Also this system is a special case of (6), now with a one-dimensional slow state variable $\mathcal{W}_{t}$, and it satisfies the assumptions of section 2.

The Merton problem consists in choosing a strategy $u$. which maximizes a given utility function $g$ at some final time $T$. In particular the problem can be described in terms of the value function

$$
\begin{equation*}
V^{\varepsilon}(t, w, y):=\sup _{u \in \mathcal{U}} \mathbf{E}\left[g\left(\mathcal{W}_{T}, Y_{T}\right) \mid \mathcal{W}_{t}=w, Y_{t}=y\right] \tag{60}
\end{equation*}
$$

Typically the utility functions in financial applications are chosen in the class of HARA (hyperbolic absolute risk aversion) functions $g(w, y)=a(b w+c)^{\gamma}$, where $a, b, c$ are bounded and continuous given functions of $y$, and $\gamma \in(0,1)$ is a given coefficient called the relative risk premium coefficient. Observe that the function $g$ satisfies assumption (10).

We remark also that in the classical HARA functions typically $a, b, c$ are constants. We choose to consider $y$ dependent coefficients since our method also permits us to manage this general case and, moreover, utilities of such a form are employed in the pricing of derivatives with nontraded assets (see [44]).

The HJB equation associated with the Merton value function is

$$
\begin{equation*}
-V_{t}^{\varepsilon}+H_{M}\left(w, y, V_{w}^{\varepsilon}, V_{w w}^{\varepsilon}, \frac{D_{y} V_{w}^{\varepsilon}}{\sqrt{\varepsilon}}\right)-\frac{1}{\varepsilon} \mathcal{L}\left(y, D_{y} V^{\varepsilon}, D_{y y}^{2} V^{\varepsilon}\right)=0 \tag{61}
\end{equation*}
$$

in $(0, T) \times \mathbb{R}_{+} \times \mathbb{R}^{m}$ complemented with the terminal condition $V^{\varepsilon}(T, x, y)=g(x, y)$. In (61) $\mathcal{L}$ is as in (14) and $H_{M}(w, y, p, X, Z)$ is defined as

$$
\begin{gathered}
\inf _{u \in U}\left\{-\left[r+\sum_{i=1}^{n}\left(\alpha^{i}-r\right) u^{i}\right] w p-\sum_{j=1}^{k}\left|\sum_{i=1}^{n} u^{i} f_{i}^{j}(y)\right|^{2} w^{2} X\right. \\
\left.-2 \sum_{h=1}^{m} \sum_{j=1}^{k} \sum_{i=1}^{n} u^{i} f_{i}^{j}(y) \tau_{h j}(y) w Z_{h}\right\}
\end{gathered}
$$

with the matrix $\tau$ given by (55). Our main theorem (Theorem 5.1) applies also in this case and says that the value function $V^{\varepsilon}$ converges locally uniformly to the unique solution of the limit problem

$$
\begin{cases}-V_{t}+\int_{\mathbb{R}^{m}} H_{M}\left(w, y, V_{w}, V_{w w}, 0\right) d \mu(y)=0 & \text { for } t \in(0, T), \quad w>0  \tag{62}\\ V(T, w)=\int_{\mathbb{R}^{m}} g(w, y) d \mu(y) & \text { for } w>0\end{cases}
$$

where $\mu(y)$ is the invariant measure associated with the fast subsystem (15).
This convergence result is new even in the case of a single risky asset and $g$ independent of $y$ that is studied in [23]. Next we interpret it in terms of stochastic control.

For simplicity we restrict ourselves to the case of a single risky asset and a scalar fast process $Y_{t}$, i.e., $n=m=1$. The equation for the wealth becomes

$$
\begin{equation*}
d \mathcal{W}_{t}=\mathcal{W}_{t}\left(r+(\alpha-r) u_{t}\right) d t+\sqrt{2} \mathcal{W}_{t} u_{t} f\left(Y_{t}\right) \cdot d \bar{W}_{t}, \quad \alpha>r \tag{63}
\end{equation*}
$$

and the HJB equation for $V^{\varepsilon}$ is

$$
\begin{align*}
-\frac{\partial V^{\varepsilon}}{\partial t}-\sup _{u \in U}\left\{[r+(\alpha-r) u] w \frac{\partial V^{\varepsilon}}{\partial w}+u^{2}|f|^{2} w^{2} \frac{\partial^{2} V^{\varepsilon}}{\partial w^{2}}+\frac{2 u w}{\sqrt{\varepsilon}}\right. & \left.\sum_{j=1}^{k} \rho_{j} f^{j} \nu \frac{\partial^{2} V^{\varepsilon}}{\partial w \partial y}\right\}  \tag{64}\\
& =\frac{1}{\varepsilon} \mathcal{L}\left(y, \frac{\partial V^{\varepsilon}}{\partial y}, \frac{\partial^{2} V^{\varepsilon}}{\partial y^{2}}\right)
\end{align*}
$$

where $\rho_{j}$ is the correlation factor between $\bar{Z}_{t}^{j}$ and $\bar{W}_{t}$; see (51). The effective PDE is

$$
\begin{equation*}
-\frac{\partial V}{\partial t}-\int_{\mathbb{R}^{m}} \max _{u \in U}\left\{[r+(\alpha-r) u] w \frac{\partial V}{\partial w}+u^{2}|f(y)|^{2} w^{2} \frac{\partial^{2} V}{\partial w^{2}}\right\} d \mu(y)=0 \tag{65}
\end{equation*}
$$

Effective utility. Note that since the utility depends also on $y$, we have an initial boundary layer. The effective utility $\bar{g}$ can be interpreted as an averaged utility which is robust with respect to fast mean reverting fluctuations and uncertainty in the market (depending also, e.g., on nontraded assets). If $g$ is independent of $y$, then the convergence is uniform up to time $T$.

Solution of the effective Cauchy problem. In some cases the effective Cauchy problem (62) can be solved explicitly. As a constraint on the control $u_{t}$ we take the interval

$$
U:=\left[R_{1}, R\right], \quad \text { with }-R \leq R_{1} \leq 0<R
$$

We also assume that the terminal cost is the HARA function

$$
g(w, y)=a(y) \frac{w^{\gamma}}{\gamma}, \quad 0<\gamma<1, \quad a(y) \geq a_{o}>0
$$

Then the terminal condition in (62) is

$$
V(T, w)=\bar{a} \frac{w^{\gamma}}{\gamma}, \quad \bar{a}:=\int_{\mathbb{R}^{m}} a(y) d \mu(y)
$$

and we look for solutions of (62) of the form $V(t, w)=\frac{w^{\gamma}}{\gamma} v(t)$ with $v(t) \geq 0$. By plugging it into the Cauchy problem we get

$$
\dot{v}=-\gamma \bar{h} v, \quad v(T)=\bar{a}, \quad \bar{h}:=r+\int_{\mathbb{R}^{m}} \max _{u \in U}\left[(\alpha-r) u+(\gamma-1)|f(y)|^{2} u^{2}\right] d \mu(y)
$$

Therefore the uniqueness of solution to (62) gives

$$
\begin{equation*}
V(t, w)=\bar{a} e^{\gamma \bar{h}(T-t)} \frac{w^{\gamma}}{\gamma}, \quad 0<t<T \tag{66}
\end{equation*}
$$

We compute the rate of exponential increase $\bar{h}$ and get

$$
\begin{aligned}
\bar{h}=r+ & \int_{\left\{y: 2 R(1-\gamma)|f(y)|^{2}<\alpha-r\right\}}\left[(\alpha-r) R+(\gamma-1) R^{2}|f(y)|^{2}\right] d \mu(y) \\
& +\int_{\left\{y: 2 R(1-\gamma)|f(y)|^{2} \geq \alpha-r\right\}} \frac{(\alpha-r)^{2}}{4(1-\gamma)|f(y)|^{2}} d \mu(y)
\end{aligned}
$$

The limit is a Merton problem. It is interesting to compare this solution with the value function of the Merton problem with constant volatility $\sigma>0$, where the wealth dynamics is

$$
d \mathcal{W}_{t}=\mathcal{W}_{t}\left(r+(\alpha-r) u_{t}\right) d t+\sqrt{2} \mathcal{W}_{t} u_{t} \sigma d \bar{W}_{t}
$$

and the utility function is $a w^{\gamma} / \gamma$.
In the case $2 R(1-\gamma) \sigma \geq \alpha-r$ (in particular, for a large or no upper bound on the control) the value function is given by the classical Merton formula

$$
\begin{equation*}
a \exp \left[\gamma\left(r+\frac{(\alpha-r)^{2}}{4(1-\gamma) \sigma^{2}}\right)(T-t)\right] \frac{w^{\gamma}}{\gamma} . \tag{67}
\end{equation*}
$$

It coincides with the solution (66) of the effective HJB equation (65) with terminal condition $\bar{g}=\bar{a} w^{\gamma} / \gamma$ if and only if $a=\bar{a}$ and

$$
\sigma=\bar{\sigma}:=\frac{\alpha-r}{2 \sqrt{(1-\gamma)(\bar{h}-r)}} .
$$

Therefore these are the correct parameters to use in a Merton model with constant volatility if we consider it as an approximation of a model with fast and ergodic stochastic volatility. We can call it the effective Merton model.

The effective volatility. The preceding formula for the effective volatility $\bar{\sigma}$ simplifies considerably if the $\mu$-probability of the set $\left\{y: 2 R(1-\gamma)|f(y)|^{2} \geq \alpha-r\right\}$ is 1 , e.g., for large upper bound $R$ on the control. In fact we get

$$
\bar{\sigma}=\left(\int_{\mathbb{R}^{m}} \frac{1}{|f(y)|^{2}} d \mu(y)\right)^{-\frac{1}{2}}
$$

a formula derived in section 10.1.2 of [23] in the case of unconstrained controls $(R=+\infty)$.
We remark that $\bar{\sigma}$ for the Merton problem is the harmonically averaged long-run volatility that is smaller than the mean historical volatility derived in section 6.1 for uncontrolled systems. Therefore using the correct parameter in the model leads to an increase of the value function, i.e., of the optimal expected utility.

The limit of the optimal control. Consider the effective Merton problem ( $a=\bar{a}, \sigma=\bar{\sigma}$ ) and suppose the upper bound $R$ on the control large enough to allow all the usual calculations of the case $R=+\infty$. The control where the Hamiltonian attains the maximum is

$$
u^{*}:=\frac{\alpha-r}{2(1-\gamma) \bar{\sigma}^{2}}=\frac{\alpha-r}{2(1-\gamma)} \int_{\mathbb{R}^{m}} \frac{1}{|f(y)|^{2}} d \mu(y),
$$

which is then the optimal control. We want to compare it with the optimal control for the problem with $\varepsilon>0$. For the terminal condition $V^{\varepsilon}(T, w, y)=a(y) w^{\gamma} / \gamma$ we expect a solution of (64) of the form $V^{\varepsilon}(t, w, y)=v^{\varepsilon}(t, y) w^{\gamma} / \gamma$. Then we can compute the maximum in the Hamiltonian of (64) and get

$$
\begin{equation*}
u_{\varepsilon}^{*}(t, y)=\frac{\alpha-r}{2(1-\gamma)|f(y)|^{2}}+\frac{\Phi(y)}{\sqrt{\varepsilon} v^{\varepsilon}(t, y)} \frac{\partial v^{\varepsilon}}{\partial y}(t, y), \quad \Phi(y):=\frac{\sum_{j=1}^{k} \rho_{j} f^{j}(y) \nu(y)}{(1-\gamma)|f(y)|^{2}} . \tag{68}
\end{equation*}
$$

By our main theorem $v^{\varepsilon}(t, y) \rightarrow v(t)$ locally uniformly in $[0, T) \times \mathbb{R}$ as $\varepsilon \rightarrow 0$, so $\frac{\partial v^{\varepsilon}}{\partial y}(t, y) \rightarrow 0$ in the sense of distributions with respect to $y$, locally uniformly in $t<T$. Then we wonder if the second term of $u_{\varepsilon}^{*}$ vanishes in some sense, despite the $\sqrt{\varepsilon}$ at the denominator, therefore giving

$$
\begin{equation*}
\lim _{\varepsilon \rightarrow 0} u_{\varepsilon}^{*}(t, y)=\frac{\alpha-r}{2(1-\gamma)|f(y)|^{2}}=: u_{0}^{*}(y) \tag{69}
\end{equation*}
$$

Note that the candidate limit $u_{0}^{*}$ is different from $u^{*}$, but $u^{*}=\int_{\mathbb{R}^{m}} u_{0}^{*}(y) d \mu(y)$.
Let us assume for simplicity that

$$
\begin{equation*}
\mu \text { has a density } \varphi \in C^{1} \quad \text { and } \quad \lim _{|y| \rightarrow \infty} \varphi(y)=0 \tag{70}
\end{equation*}
$$

The former assumption is satisfied, for instance, if the coefficients $b, \nu$ of $\mathcal{L}$ are smooth, because $\mathcal{L}^{*} \mu=0$ in the sense of distributions and the regularity theory for elliptic equations applies $\left(\mathcal{L}^{*}\right.$ being the formal adjoint of $\left.\mathcal{L}\right)$. The latter assumption is natural for an integrable $\varphi$, and it is satisfied, for instance, by the Ornstein-Uhlenbeck process ( $\varphi$ is a Gaussian function). Then, when we take the integral of (68) with respect to $\mu$ and integrate by parts the second term, we get

$$
\int_{\mathbb{R}^{m}} u_{\varepsilon}^{*}(t, y) d \mu(y)=u^{*}+o\left(\frac{1}{\sqrt{\varepsilon}}\right) \quad \text { as } \varepsilon \rightarrow 0
$$

which again is not very insightful. To get some convergence we write an asymptotic expansion for $v_{\varepsilon}^{*}(t, y)$ in powers of $\sqrt{\varepsilon}$, in the spirit of section 10.1.2 of the book by Fouque, Papanicolaou, and Sircar [23] but under weaker assumptions and using different arguments.

Proposition 6.2. Besides the standing assumptions of the section and (70) suppose

$$
\begin{equation*}
v^{\varepsilon}(t, y)=v(t)+\sqrt{\varepsilon} v_{1}^{\varepsilon}(t, y), v_{1}^{\varepsilon}(t, y) \rightarrow v_{1}(t, y) \quad \text { locally uniformly, } v_{1} \text { bounded. } \tag{71}
\end{equation*}
$$

Then the following hold:
(i) $v_{1}=v_{1}(t)$, so $\frac{1}{\sqrt{\varepsilon}} \frac{\partial v^{\varepsilon}}{\partial y}=\frac{\partial v_{1}^{\varepsilon}}{\partial y}(t, y) \rightarrow 0$ in the sense of distributions with respect to $y$;
(ii) if, in addition,

$$
\begin{equation*}
\left|v_{1}^{\varepsilon}\right| \leq C, \quad \sqrt{\varepsilon} \int_{\mathbb{R}^{m}}\left|\frac{\partial v_{1}^{\varepsilon}}{\partial y}\right| d \mu(y) \rightarrow 0 \quad \forall t<T \tag{72}
\end{equation*}
$$

then

$$
\begin{equation*}
u^{*}=\lim _{\varepsilon \rightarrow 0} \int_{\mathbb{R}^{m}} u_{\varepsilon}^{*}(t, y) d \mu(y) \quad \forall t<T \tag{73}
\end{equation*}
$$

(iii) if, in addition,

$$
\begin{equation*}
v_{1}^{\varepsilon}(t, y)=v_{1}(t)+\omega(\varepsilon) v_{2}^{\varepsilon}(t, y), \quad \omega(\varepsilon) \rightarrow 0, \quad\left|\frac{\partial v_{2}^{\varepsilon}}{\partial y}(t, y)\right| \leq C(t, y) \tag{74}
\end{equation*}
$$

then $\frac{1}{\sqrt{\varepsilon}} \frac{\partial v^{\varepsilon}}{\partial y} \rightarrow 0$ and (69) holds uniformly on every set where $C(\cdot, \cdot)$ is bounded.

Proof. (i) By plugging the optimal control (68) into the HJB equation (64) we get

$$
-\frac{\partial v^{\varepsilon}}{\partial t}-\gamma r v^{\varepsilon}-F_{1}(y)\left((\alpha-r) v^{\varepsilon}+\frac{F_{2}(y)}{\sqrt{\varepsilon}} \frac{\partial v^{\varepsilon}}{\partial y}\right)^{2}=\frac{1}{\varepsilon} \mathcal{L}\left(y, \frac{\partial v^{\varepsilon}}{\partial y}, \frac{\partial^{2} v^{\varepsilon}}{\partial y^{2}}\right)
$$

for suitable continuous $F_{i}, i=1,2$. Using the expansion (71) the equation becomes

$$
-\mathcal{L}\left(y, \frac{\partial v_{1}^{\varepsilon}}{\partial y}, \frac{\partial^{2} v_{1}^{\varepsilon}}{\partial y^{2}}\right)=\sqrt{\varepsilon}\left[\frac{\partial v^{\varepsilon}}{\partial t}+\gamma r v^{\varepsilon}+F_{1}(y)\left((\alpha-r) v^{\varepsilon}+F_{2}(y) \frac{\partial v_{1}^{\varepsilon}}{\partial y}\right)^{2}\right] .
$$

Letting $\varepsilon \rightarrow 0$ we obtain, by standard properties of viscosity solutions,

$$
-\mathcal{L}\left(y, \frac{\partial v_{1}}{\partial y}, \frac{\partial^{2} v_{1}}{\partial y^{2}}\right)=0 \quad \text { in } \mathbb{R}
$$

so $v_{1}$ is constant with respect to $y$ by the Liouville property, Lemma 4.1.
(ii) First observe that $v^{\varepsilon}$ is uniformly bounded and bounded away from 0 . The upper bound follows from (19). The lower bound is obtained by using the definition (60) of $V^{\varepsilon}$ and computing the payoff of the control $u$. $\equiv 0$. We get

$$
V^{\varepsilon}(t, w, y) \geq \mathbf{E}\left[a\left(Y_{T}\right) \mid Y_{t}=y\right] e^{\gamma r(T-t)} \frac{w^{\gamma}}{\gamma}
$$

and therefore

$$
v^{\varepsilon}(t, y) \geq a_{o} e^{\gamma r(T-t)} \geq a_{o} \quad \forall t \leq T, \forall y .
$$

From (68) and the expansion (71) we get

$$
\int_{\mathbb{R}^{m}} u_{\varepsilon}^{*}(t, y) d \mu(y)=u^{*}+\int_{\mathbb{R}^{m}} \frac{\Phi(y)}{v^{\varepsilon}(t, y)} \frac{\partial v_{1}^{\varepsilon}}{\partial y}(t, y) \varphi(y) d y .
$$

Integrating by parts, the integral on the right-hand side becomes

$$
-\int_{\mathbb{R}^{m}} \frac{\partial}{\partial y}\left(\frac{\Phi \varphi}{v^{\varepsilon}}\right) v_{1}^{\varepsilon} d y+\left[\Phi \varphi \frac{v_{1}^{\varepsilon}}{v^{\varepsilon}}\right]_{y \rightarrow-\infty}^{y \rightarrow+\infty}
$$

and the second term is null by (70) and the uniform boundedness of $\Phi v_{1}^{\varepsilon} / v^{\varepsilon}$. The first term can be written as

$$
-\int_{\mathbb{R}^{m}} \frac{\partial(\Phi \varphi)}{\partial y} \frac{v_{1}^{\varepsilon}}{v^{\varepsilon}} d y+\int_{\mathbb{R}^{m}} \sqrt{\varepsilon} \frac{\partial v_{1}^{\varepsilon}}{\partial y} \frac{\Phi v_{1}^{\varepsilon}}{\left(v^{\varepsilon}\right)^{2}} \varphi d y
$$

and we let $\varepsilon \rightarrow 0$ : the second integral vanishes by (72) and the uniform boundedness of $\Phi v_{1}^{\varepsilon} /\left(v^{\varepsilon}\right)^{2}$, whereas the first converges to

$$
-\frac{v_{1}(t)}{v(t)} \int_{\mathbb{R}^{m}} \frac{\partial(\Phi \varphi)}{\partial y}(y) d y=0
$$

by (70). This completes the proof of (73).
(iii) By (74)

$$
\frac{1}{\sqrt{\varepsilon}} \frac{\partial v^{\varepsilon}}{\partial y}=\omega(\varepsilon) \frac{\partial v_{2}^{\varepsilon}}{\partial y}(t, y) \rightarrow 0
$$

uniformly on every set where $\partial v_{2}^{\varepsilon} / \partial y$ is uniformly bounded. By (68) $u_{\varepsilon}^{*}$ converges uniformly on every such set to $u_{0}^{*}$.

We can roughly summarize the preceding proposition by saying that an asymptotic expansion of $v^{\varepsilon}$ of the form

$$
v^{\varepsilon}=v+\sqrt{\varepsilon} v_{1}+o(\sqrt{\varepsilon}) v_{2}^{\varepsilon}
$$

implies that the optimal control $u^{*}$ of the effective Merton model is the limit of the averages and the average of the limit of the optimal controls for the models with $\varepsilon>0$, i.e.,

$$
u^{*}=\lim _{\varepsilon \rightarrow 0} \int_{\mathbb{R}^{m}} u_{\varepsilon}^{*}(t, y) d \mu(y)=\int_{\mathbb{R}^{m}} \lim _{\varepsilon \rightarrow 0} u_{\varepsilon}^{*}(t, y) d \mu(y)
$$

The financial interpretation of this statement is clear: the optimal control for the Merton problem with constant volatility $\bar{\sigma}$ approximates the expectation of the optimal control for the same problem with stochastic volatility, provided the volatility evolves much faster than the assets.
6.3. Periodic day effects and volatility with a slow component. Section 10.2 of [23] discusses a refinement of the model in section 6.1, where the volatilities of the prices depend on time on a fast periodic scale, thus modeling the daily oscillations. This amounts to replacing $f_{i}\left(Y_{t}\right)$ in (49) and (59) with

$$
f_{i}=f_{i}\left(\frac{t}{\varepsilon}, Y_{t}\right)
$$

where $f_{i}$ is 1-periodic in the first entry. We incorporate this in our setting by adding the new variable $s:=t / \varepsilon$ whose dynamics is $\dot{s}:=1 / \varepsilon$. The fast subsystem now has the additional variable $s_{t}$ that is trivially ergodic on the unit circle with respect to the Lebesgue measure. Now the effective Hamiltonian of the limit PDE is

$$
\bar{H}=\int_{0}^{1} \int_{\mathbb{R}^{m}} H(x, y, s, p, X, 0) d \mu(y) d s
$$

Another possible extension of the model in sections 6.1 and 6.2 is the addition of another stochastic quantity $Z_{t}$ affecting the volatilities of the prices and evolving on a slower time scale than the prices:

$$
\begin{gather*}
f_{i}=f_{i}\left(Y_{t}, Z_{t}\right), \\
d Z_{t}=\theta c\left(Z_{t}\right) d t+\sqrt{\theta} d\left(Z_{t}\right) d W_{t}, \quad Z_{0}=z, \tag{75}
\end{gather*}
$$

with $\theta$ small, and $c, d$ Lipschitz and growing at most linearly at infinity. This is done, for instance, in $[25,37]$. This modeling allows much more flexibility and is motivated by various empirical studies (see [25] and the references therein) which outline a volatility composed by one highly persistent factor and one quickly mean reverting factor. The slow volatility factor in particular is useful when considering options with longer maturities.

The value function now depends also on the initial position $z$ of the new variable $Z_{t}$, and the HJB equation (12) becomes

$$
\begin{aligned}
& \lambda V-V_{t}+H\left(x, y, z, D_{x} V, D_{x x}^{2} V, \frac{D_{x y}^{2} V}{\sqrt{\varepsilon}}, \sqrt{\theta} D_{x z}^{2} V\right)-\frac{1}{\varepsilon} \mathcal{L}\left(y, D_{y} V, D_{y y}^{2} V\right) \\
&-\theta\left[c \cdot D_{z} V+\theta \operatorname{trace}\left(d d^{T} D_{z z}^{2} V\right)\right]-\sqrt{\frac{\theta}{\varepsilon}} \operatorname{trace}\left[\tau d^{T} D_{y z}^{2} V+D_{y z}^{2} V \tau d^{T}(z)\right]=0 .
\end{aligned}
$$

In particular this can be seen as a regular perturbation of (12). If $\theta$ is independent of $\varepsilon$ and we let it tend to 0 , the basic properties of viscosity solutions give the convergence of the value function $V^{\varepsilon, \theta}(t, x, y, z)$ to the solution $V(t, x, z)$ of the same effective Cauchy problem as before, with the only difference that $\bar{H}$ now depends also on $z$ (but $z$ appears only as a fixed parameter in the limit PDE). It possible to check this result regardless of the order of taking the limits $\theta \rightarrow 0$ and $\varepsilon \rightarrow 0$. Indeed the term $-\sqrt{\frac{\theta}{\varepsilon}}$ trace $\left[\tau(y) d^{T}(z) D_{y z}^{2} V+D_{y z}^{2} V \tau(y) d^{T}(z)\right]$ is a lower order term with respect to $\frac{1}{\varepsilon} \mathcal{L}\left(y, D_{y} V, D_{y y}^{2} V\right)$, and then a similar argument as in Remark 5.1 holds. If, instead, $\theta=\theta(\varepsilon)$, the same conclusion follows with a much more delicate argument, following a theorem on regular perturbations of singular perturbation problems proved in [4].

Of course the periodic oscillations in time and the slow component of the volatility can also be treated simultaneously. As an example, we consider the scalar Merton problem (60), (63) with volatility and utility functions given by

$$
f_{i}=f_{i}\left(\frac{t}{\varepsilon}, Y_{t}, Z_{t}\right), \quad g=a\left(Y_{T}, Z_{T}\right) \frac{\mathcal{W}_{T}^{\gamma}}{\gamma}
$$

with $Z_{t}$ satisfying (75). Then the value function $V^{\varepsilon, \theta}(t, x, y, z)$ converges locally uniformly to the classical Merton formula (67) for the problem with constant volatility

$$
\sigma=\bar{\sigma}(z):=\left(\int_{0}^{1} \int_{\mathbb{R}^{m}} \frac{1}{|f(s, y, z)|^{2}} d \mu(y) d s\right)^{-\frac{1}{2}}
$$

at least when the upper bound $R$ on the controls is large enough, and

$$
a=\bar{a}(z):=\int_{\mathbb{R}^{m}} a(y, z) d \mu(y) .
$$

6.4. Worst case optimization under unknown disturbances. Assume that the general stochastic control system (8) is affected by an additional disturbance $\tilde{u}_{t}$ taking values in a compact set $\tilde{U}$, and suppose you want to maximize the payoff under the worst possible behavior of $\tilde{u}_{t}$. There are several possible reasons for this choice, such as the lack of statistical information on the disturbance or the desire to avoid with probability one some catastrophic events caused by a particularly nasty behavior of $\tilde{u}_{t}$. The mathematical framework for modeling these problems is the theory of two-person 0 -sum differential games, where the controller is the first player and the disturbance is considered as the control of a second player wishing to minimize the payoff.

For simplicity we suppose the following form of the drift and diffusion in (6):

$$
\phi^{i}=\phi_{1}^{i}(x, y, u)+\phi_{2}^{i}(x, y, \tilde{u}), \quad \sigma^{i}=\sigma_{1}^{i}(x, y, u)+\sigma_{2}^{i}(x, y, \tilde{u}),
$$

with $\phi_{j}^{i}, \sigma_{j}^{i}$ bounded, continuous, and Lipschitz in $(x, y)$ uniformly in $u, \tilde{u}$. For the system written in vector form (8) we then have $\tilde{\phi}^{i}=\tilde{\phi}_{1}^{i}(x, y, u)+\tilde{\phi}_{2}^{i}(x, y, \tilde{u})$ and $\tilde{\sigma}^{i}=\tilde{\sigma}_{1}^{i}(x, y, u)+$ $\tilde{\sigma}_{2}^{i}(x, y, \tilde{u})$ with the obvious definitions. The Isaacs equation associated with the game is again of the form (12), but now the Hamiltonian is $H=H_{1}+H_{2}$ with

$$
\begin{aligned}
& H_{1}(x, y, p, X, Z):=\min _{u \in U}\left\{-\operatorname{trace}\left(\tilde{\sigma}_{1} \tilde{\sigma}_{1}^{T} X\right)-\tilde{\phi}_{1} \cdot p-2 \operatorname{trace}\left(\tilde{\sigma}_{1} \tau^{T} Z^{T}\right)\right\}, \\
& H_{2}(x, y, p, X, Z):=\max _{\tilde{u} \in \tilde{U}}\left\{-\operatorname{trace}\left(\tilde{\sigma}_{2} \tilde{\sigma}_{2}^{T} X\right)-\tilde{\phi}_{2} \cdot p-2 \operatorname{trace}\left(\tilde{\sigma}_{2} \tau^{T} Z^{T}\right)\right\} .
\end{aligned}
$$

The precise definition of value function is more delicate for a stochastic differential game, as well as the proof that it is a viscosity solution of (12), and we refer the reader to [21]. We remark that the comparison principle of [15] still holds for the Cauchy problem (17) with the new convex-concave Hamiltonian, and therefore there is a unique viscosity solution $V^{\varepsilon}$. The convergence theorem (Theorem 5.1) of $V^{\varepsilon}(t, x, y)$ to $V(t, x)$ holds with no changes, because its proof never uses the convexity of $H$ with respect to $(p, X)$. The effective Hamiltonian now is

$$
\bar{H}=\int_{\mathbb{R}^{m}} \min _{u \in U}\left\{-\operatorname{trace}\left(\tilde{\sigma}_{1} \tilde{\sigma}_{1}^{T} X\right)-\tilde{\phi}_{1} \cdot p\right\}+\max _{\tilde{u} \in \tilde{U}}\left\{-\operatorname{trace}\left(\tilde{\sigma}_{2} \tilde{\sigma}_{2}^{T} X\right)-\tilde{\phi}_{2} \cdot p\right\} d \mu(y)
$$

6.5. Applications to problems with degenerate diffusion. We pointed out in the introduction that we do not make any nondegeneracy assumption on the diffusion matrix $\sigma \sigma^{T}$ for the slow variables $X_{t}$. This makes our methods applicable to a wide range of models, even in deterministic control, if one wants to study the sensitivity to random parameters evolving on a fast time scale. For instance, some differential games arising in marketing and advertising are under investigation.

Within mathematical finance, path-dependent models, such as Asian options, involve degenerate diffusion processes; see $[42,8]$ and the references therein. In these models one augments the state space by a new variable $A_{s}$ that is the time integral of some functions of a price $S_{s}$. Therefore an ODE is added to the system, such as $d A_{s}=S_{s} d s$ for problems involving the arithmetic mean of the prices or $d A_{s}=\log \left(S_{s}\right) d s$ for the geometric mean. Therefore the process $X_{s}=\left(S_{s}, A_{s}\right)$ is a degenerate diffusion. Models of Asian options with fast stochastic volatility are studied in Chapter 8.3 of [25] and in [22, 43].

Interest rate models are another area where the uniform nondegeneracy of the diffusion matrix would not be a reasonable assumption. The LIBOR models with stochastic volatility reviewed in Chapter 11 of [13] all have a volatility function $\sigma\left(X_{s}, Y_{s}\right)$ vanishing at $Y_{s}=0$. This event usually has null probability, by the choice of the dynamics for $Y_{s}$. So the associated PDE is parabolic but not uniformly parabolic. Some of these models with two time scales are studied in Chapter 11 of [25].

A stronger form of degeneracy occurs in the Heath-Jarrow-Morton (HJM) framework for forward rate models, where there are an infinite number of traded assets (one for each
maturity) and a finite number of sources of randomness (components of the Brownian motion); see, e.g., Chapter 23 of [10]. The possibility of arbitrage is ruled out by the HJM drift condition. If one considers a large but finite number of maturities, the assets evolve as a degenerate diffusion and our methods can be used for the asymptotics of the fast stochastic volatility problem. HJM models with stochastic volatility (with the same time scale as the prices) were studied in [11].
7. Conclusion. In this paper we study stochastic control problems with random parameters driven by a fast ergodic process. Our methods are based on viscosity solutions theory and the Hamilton-Jacobi approach to singular perturbations. The assumptions are chosen to fit problems of pricing derivative securities and optimizing the portfolio allocation in financial markets with fast mean reverting stochastic volatility.

The main steps of our HJB approach to singular perturbations are the following:

- Write the HJB equation for the value function $V^{\varepsilon}$ and characterize it as the unique viscosity solution of the Cauchy problem for such an equation (see section 3).
- Define a limit (effective) PDE and limit (effective) initial data resolving appropriate ergodic-type problems (see section 4).
- Prove the (locally) uniform convergence of $V^{\varepsilon}$ to a function $V$, which can be characterized as the unique solution of the effective Cauchy problem (see section 5).
- Interpret the effective PDE as the HJB equation for a limit (effective) control problem. Such a problem approximates the one with $\varepsilon>0$, and it has lower-dimensional state variables; therefore it is easier to solve. There is no general recipe for this step, and we do it in section 6 for a multidimensional option pricing model and for the Merton portfolio optimization problem.
The main contributions of the present paper are the following. On the mathematical side we extend the HJB approach from the setting of periodic fast variables (see $[1,2,3]$ and the references therein) to the case of unbounded fast variables. The probabilistic literature on singular perturbations in stochastic control (see the monographs [32,33] and the references therein) allows unbounded fast variables but makes other restrictive assumptions that rule out some financial models such as the Merton optimization problem (e.g., in [33] the diffusion matrix $\tilde{\sigma}$ is assumed uncontrolled).

On the side of financial models our approach complements the methods of Fouque, Papanicolaou, and Sircar [23]. They assume an asymptotic expansion for $V^{\varepsilon}$ of the form

$$
\begin{equation*}
V^{\varepsilon}=V+\sqrt{\varepsilon} V_{1}+\varepsilon V_{2}+\cdots, \tag{76}
\end{equation*}
$$

plug it into the HJB PDE for $V^{\varepsilon}$, set equal to 0 each term multiplying a power of $\varepsilon$, and solve iteratively such PDEs to compute the correctors $V_{i}$. This gives information not only on the limit but also for $\varepsilon$ positive with various orders of magnitude. The validity of the expansion can be proved in some problems without control; this is done, for instance, in [24] for the option pricing of a single asset. Our result in section 6.1 complements it by treating the multiasset problem but only up to the first term of the expansion. Since the PDE is linear we believe that the arguments can be carried on to study further terms, but we do not try to do it here.

For problems with controls, however, the validity of the asymptotic expansion (76) is not known, even for particular problems like Merton, and presumably it is not true in general. Section 10.1 of [23] assumes (76) for the Merton problem and gets some interesting insight on the correction of the optimal control. Our contribution in section 6.2 is a rigorous proof of the locally uniform convergence of the value function with stochastic volatility to the value of the Merton problem with constant effective volatility $\bar{\sigma}$ (instead of the historical volatility)

$$
\lim _{\varepsilon \rightarrow 0} V^{\varepsilon}(t, w, y)=V(t, w), \quad \bar{\sigma}^{2}=\int_{\mathbb{R}^{m}} \frac{1}{|f(y)|^{2}} d \mu(y)
$$

and also for utility functions depending on the fast variable $y$. The problem of justifying further terms of the asymptotic expansion is wide open in stochastic control and fully nonlinear PDEs, even for the first corrector $V_{1}$. The only related result we know is in the very recent paper by Camilli and Marchi [14] and concerns the rate of convergence in periodic homogenization. We plan to study this issue for particular models arising in applications. As for the convergence of the optimal control, at the end of section 6.2 we assume the expansion

$$
V^{\varepsilon}=V+\sqrt{\varepsilon} V_{1}+o(\sqrt{\varepsilon}) V_{2}^{\varepsilon}
$$

and prove that

$$
u^{*}=\lim _{\varepsilon \rightarrow 0} \mathbf{E}\left[u_{\varepsilon}^{*}(t, Y)\right]=\mathbf{E}\left[\lim _{\varepsilon \rightarrow 0} u_{\varepsilon}^{*}(t, Y)\right], \quad Y \sim \mu,
$$

which has a clear financial interpretation.
Finally, we remark that our method is very general and can be used for a number of models, financial or not, including 0 -sum differential games and degenerate diffusions. The case of controls appearing also in the fast variables was studied in $[1,2,3]$ and the references therein when the fast variables are bounded; see also [12]. We plan to push the methods of the present paper further and treat problems with controlled and unbounded fast variables.

Acknowledgments. We are grateful to Wolfgang Runggaldier and Tiziano Vargiolu for useful discussions and remarks and wish to thank the anonymous referees for their suggestions that helped us to improve the paper.

## REFERENCES

[1] O. Alvarez and M. Bardi, Viscosity solutions methods for singular perturbations in deterministic and stochastic control, SIAM J. Control Optim., 40 (2001), pp. 1159-1188.
[2] O. Alvarez and M. Bardi, Singular perturbations of nonlinear degenerate parabolic PDEs: A general convergence result, Arch. Ration. Mech. Anal., 170 (2003), pp. 17-61.
[3] O. Alvarez and M. Bardi, Ergodicity, stabilization, and singular perturbations for Bellman-Isaacs equations, Mem. Amer. Math. Soc., 204 (2010), article 960.
[4] O. Alvarez, M. Bardi, and C. Marchi, Multiscale problems and homogenization for second-order Hamilton-Jacobi equations, J. Differential Equations, 243 (2007), pp. 349-387.
[5] M. Arisawa and P.-L. Lions, On ergodic stochastic control, Comm. Partial Differential Equations, 23 (1998), pp. 2187-2217.
[6] M. Bardi and I. Capuzzo-Dolcetta, Optimal Control and Viscosity Solutions of Hamilton-JacobiBellman Equations, Birkhäuser, Boston, 1997.
[7] M. Bardi and F. Da Lio, On the strong maximum principle for fully nonlinear degenerate elliptic equations, Arch. Math. (Basel), 73 (1999), pp. 276-285.
[8] E. Barucci, S. Polidoro, and V. Vespri, Some results on partial differential equations and Asian options, Math. Models Methods Appl. Sci., 11 (2001), pp. 475-497.
[9] A. Bensoussan, Perturbation Methods in Optimal Control, John Wiley \& Sons, Montrouge, France, 1988.
[10] T. Bנörk, Arbitrage Theory in Continuous Time, Oxford University Press, Oxford, UK, 1998.
[11] T. Björk, C. Landén, and L. Svensson, Finite-dimensional Markovian realizations for stochastic volatility forward-rate models, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 460 (2004), pp. 53-83.
[12] V. S. Borkar and V. Gaitsgory, Singular perturbations in ergodic control of diffusions, SIAM J. Control Optim., 46 (2007), pp. 1562-1577.
[13] D. Brigo and F. Mercurio, Interest Rate Models-Theory and Practice, 2nd ed., Springer-Verlag, Berlin, 2006.
[14] F. Camilli and C. Marchi, Rates of convergence in periodic homogenization of fully nonlinear uniformly elliptic PDEs, Nonlinearity, 22 (2009), pp. 1481-1498.
[15] F. Da Lio and O. Ley, Uniqueness results for second-order Bellman-Isaacs equations under quadratic growth assumptions and applications, SIAM J. Control Optim., 45 (2006), pp. 74-106.
[16] G. Da Prato and J. Zabczyk, Ergodicity for Infinite-Dimensional Systems, Cambridge University Press, Cambridge, UK, 1996.
[17] D. Down, S. P. Meyn, and R. L. Tweedie, Exponential and uniform ergodicity of Markov processes, Ann. Probab., 23 (1995), pp. 1671-1691.
[18] L. C. Evans, The perturbed test function method for viscosity solutions of nonlinear PDE, Proc. Roy. Soc. Edinburgh Sect. A, 111 (1989), pp. 359-375.
[19] W. H. Fleming and D. Hernández-Hernández, An optimal consumption model with stochastic volatility, Finance Stoch., 7 (2003), pp. 245-262.
[20] W. H. Fleming and H. M. Soner, Controlled Markov Processes and Viscosity Solutions, 2nd ed., Springer-Verlag, New York, 2006.
[21] W. H. Fleming and P. E. Souganidis, On the existence of value functions of two-player, zero-sum stochastic differential games, Indiana Univ. Math. J., 38 (1989), pp. 293-314.
[22] J.-P. Fouque and C.-H. Han, Pricing Asian options with stochastic volatility, Quant. Finance, 3 (2003), pp. 353-362.
[23] J.-P. Fouque, G. Papanicolaou, and R. Sircar, Derivatives in Financial Markets with Stochastic Volatility, Cambridge University Press, Cambridge, UK, 2000.
[24] J.-P. Fouque, G. Papanicolaou, R. Sircar, and K. Solna, Singular perturbations in option pricing, SIAM J. Appl. Math., 63 (2003), pp. 1648-1665.
[25] J.-P. Fouque, G. Papanicolaou, R. Sircar, and K. Solna, Multiscale stochastic volatility asymptotics, Multiscale Model. Simul., 2 (2003), pp. 22-42.
[26] A. Friedman, Partial Differential Equations of Parabolic Type, Prentice-Hall, Englewood Cliffs, NJ, 1964.
[27] Ĭ. Ī. Gīhman and A. V. Skorohod, Stochastic Differential Equations, Springer-Verlag, New York, 1972.
[28] R. Z. HasminskiI, Stochastic Stability of Differential Equations, Sijthoff \& Noordhoff, Alphen aan den Rijn, The Netherlands, Germantown, MD, 1980.
[29] R. Z. Hasminskil and G. Yin, Uniform asymptotic expansions for pricing European options, Appl. Math. Optim., 52 (2005), pp. 279-296.
[30] A. Ilhan, M. Jonsson, and R. Sircar, Singular perturbations for boundary value problems arising from exotic options, SIAM J. Appl. Math., 64 (2004), pp. 1268-1293.
[31] M. Jonsson and R. Sircar, Partial hedging in a stochastic volatility environment, Math. Finance, 12 (2002), pp. 375-409.
[32] Y. Kabanov and S. Pergamenshchikov, Two-Scale Stochastic Systems. Asymptotic Analysis and Control, Springer-Verlag, Berlin, 2003.
[33] H. J. Kushner, Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems, Birkhäuser, Boston, 1990.
[34] P.-L. Lions and M. Musiela, Ergodicity of Diffusion Processes, manuscript, 2002.
[35] P.-L. Lions, G. Papanicolaou, and S. R. S. Varadhan, Homogenization of Hamilton-Jacobi Equations, manuscript, 1986.
[36] L. Lorenzi and M. Bertoldi, Analytical Methods for Markov Semigroups, Chapman \& Hall, Boca Raton, FL, 2007.
[37] E. Papageorgiou and R. Sircar, Multiscale intensity models for single name credit derivatives, Appl. Math. Finance, 15 (2008), pp. 73-105.
[38] E. Pardoux and A. Yu. Veretennikov, On the Poisson equation and diffusion approximation, I, Ann. Probab., 29 (2001), pp. 1061-1085.
[39] E. Pardoux and A. Yu. Veretennikov, On the Poisson equation and diffusion approximation, II, Ann. Probab., 31 (2003), pp. 1166-1192.
[40] E. Pardoux and A. Yu. Veretennikov, On the Poisson equation and diffusion approximation, III, Ann. Probab., 33 (2005), pp. 1111-1133.
[41] R. Sircar and T. Zariphopoulou, Bounds and asymptotic approximations for utility prices when volatility is random, SIAM J. Control Optim., 43 (2005), pp. 1328-1353.
[42] P. Wilmott, S. Howison, and J. Dewynne, The Mathematics of Financial Derivatives, Cambridge University Press, Cambridge, UK, 1995.
[43] H. Y. Wong and Y. L. Cheung, Geometric Asian options: Valuation and calibration with stochastic volatility, Quant. Finance, 4 (2004), pp. 301-314.
[44] T. Zariphopoulou, A solution approach to valuation with unhedgeable risks, Finance Stoch., 5 (2001), pp. 61-82.

# Maturity-Independent Risk Measures* 

Thaleia Zariphopoulou ${ }^{\dagger}$ and Gordan Žitković ${ }^{\ddagger}$


#### Abstract

The new notion of maturity-independent risk measures is introduced and contrasted with the existing risk measurement concepts. It is shown, by means of two examples, one set on a finite probability space and the other in a diffusion framework, that, surprisingly, some of the widely utilized risk measures cannot be used to build maturity-independent counterparts. We construct a large class of maturity-independent risk measures and give representative examples in both continuous- and discrete-time financial models.


Key words. risk measures, maturity independence, incomplete markets, forward performance processes, exponential utility

AMS subject classifications. 91B16, 91B28
DOI. 10.1137/080739732

1. Introduction. The abstract notion of a risk measure appeared first in $[2,1]$. The simple axioms set forth in [1] opened a venue for a rich field of research that shows no signs of fatigue. The main reason for such success is the fundamental need for quantification and measurement of risk. While the initial impetus came from the requirements of the financial and insurance industries, applications in a wide range of situations, together with a mathematical tractability and elegance of this theory, have promoted risk measurement to an independent field of interest and research. The early cornerstones include (but are not limited to) [11, 12, 14]; see also [13] for more information.

The first notions of risk measures were all static, meaning that the time of measurement as well as the time of resolution (maturity, expiry) of the risk were fixed. Soon afterwards, however, dynamic and conditional risk measures started to appear (see [3, 30, 5, 10, 6, 7, 31] as well as the book [13]).

Despite all the recent work in this wide area, there are still a number of theoretical, as well as practical, questions left unanswered. The one we focus on in the present paper deals with the problem one faces when the maturity (horizon, expiration date, etc.) associated with a particular risky position is not fixed. We take the view that the mechanism used to measure the risk content of a certain random variable should not depend on any a priori choice of the measurement horizon. This is, for example, the case in complete financial markets. Indeed, consider for simplicity the Samuelson (Black-Scholes) market model with zero interest rate

[^19]and the procedure one would follow to price a contingent claim therein. The fundamental theorem of asset pricing tells us to simply compute the expectation of the discounted claim under the unique martingale measure. There is no explicit mention of the maturity date of the contingent claim in this algorithm or, for that matter, any other prespecified horizon. Letting the claim's payoff stay unexercised for any amount of time after its expiry would not change its arbitrage-free price in any way.

It is exactly this property that, in our opinion, has not received sufficient attention in the literature. As one of the fundamental properties clearly exhibited under market completeness, it should be shared by any workable risk measurement and pricing procedure in arbitrary incomplete markets.

The incorporation of the maturity-independence property described above into the existing framework of risk measurement has been guided by the principle of minimal impact: we strove to keep new axioms as similar as possible to the existing ones for convex risk measures and to implement only minimally needed changes. This led us to the realization that it is the domain of the risk measure that inadvertently dictates the use of a specific time horizon, and if we replace it by a more general domain, the maturity independence would follow. Thus, our axioms are identical to the axioms of a replication-invariant convex risk measure, except for the choice of the domain which is not a subspace of a function space on $\mathcal{F}_{T}$ for a fixed time horizon $T$.

In addition to the novel axiom pertinent to maturity independence, a link to the notion of forward performance processes recently proposed by Musiela and the first author (see [ $25,29,27,26,28]$ ) is established. Indeed, focusing on the exponential case, it is shown that every forward performance process can be used to create an example of a maturity-independent risk measure. On one hand, this connection provides a useful and simple tool for (a nontrivial task of) constructing maturity-independent risk measures. On the other hand, we hope that it would give a firm decision-theoretic foundation to the theory of forward performances.

We start off by introducing the financial model, trading, and no-arbitrage conditions and recalling some well-known facts about risk measures. In section 3, we introduce the notion of a maturity-independent risk measure, argue for its feasibility and relevance, and give first examples. We also show, via two simple examples, that a naïve approach to the construction of maturity-independent risk measures can fail. Section 4 opens with the notion of a performance random field and goes on to describe the important class of forward performance processes. These objects are, in turn, used to produce a class of maturity-independent risk measures which we call forward entropic risk measures. Finally, several special cases of these measures are mentioned and interpreted in section 5, and an independent example, set in a binomialtype incomplete financial model, is presented.

## 2. Generalities on the financial market model and risk measures.

### 2.1. Market setup, absence of arbitrage, and admissible portfolios.

2.1.1. The model. Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a probability space, with the filtration $\mathbb{F}=\left(\mathcal{F}_{t}\right)_{t \in[0, \infty)}$ satisfying the usual assumptions. The evolution of the prices of risky assets is modeled by a $k$-dimensional locally bounded $\mathbb{F}$-semimartingale $S=\left(S_{t}^{1}, \ldots, S_{t}^{k}\right)_{t \in[0, \infty)}$. The existence of a liquid risk-free asset $S^{0}$ is also postulated. As usual, we quote all asset prices in the units
of $S^{0}$. Operationally, this amounts to the simplifying assumption $S_{t}^{0}=1, t \geq 0$, which will hold throughout.

Remark 2.1. We would like to point out that the theory presented in this paper would require little or no conceptual adjustment if the normalization $S^{0} \equiv 1$ were not introduced and a general predictable numéraire $S^{0}$ satisfying the usual regularity conditions were used. The reason we do not pursue such a generalization is that the notation would unnecessarily suffer and the important aspects of the theory would, consequently, be obscured.
2.1.2. Portfolio processes. An $\mathbb{F}$-predictable process $\pi=\left(\pi_{t}^{1}, \ldots, \pi_{t}^{k}\right)_{t \in[0, \infty)}$ is called a portfolio (process) if it is $S$-integrable in the sense of (vector) stochastic integration (see section 4d in [19]). A portfolio $\pi$ is called admissible if there exists a constant $a>0$ (possibly depending on $\pi$ but not on the state of the world) such that the gains process $X^{\pi}=\left(X_{t}^{\pi}\right)_{t \in[0, \infty)}$, defined as

$$
X_{t}^{\pi}=\int_{0}^{t} \pi_{s} d S_{s}=\sum_{i=1}^{k} \int_{0}^{t} \pi_{s}^{i} d S_{s}^{i}, \quad t \geq 0
$$

is bounded from below by $-a$, i.e., $X_{t}^{\pi} \geq-a$ for all $t \geq 0$ a.s. The set of all portfolio processes $\pi$ whose gains processes $X^{\pi}$ are admissible is denoted by $\mathcal{A}$. For technical reasons, which will be clear shortly, we introduce the set $\mathcal{A}_{\text {bd }}$ of all portfolio processes $\pi$ whose gains process $X^{\pi}$ is uniformly bounded from above as well as from below, i.e., $\mathcal{A}_{\mathrm{bd}}=\mathcal{A} \cap(-\mathcal{A})=$ $\{\pi \in \mathcal{A}:-\pi \in \mathcal{A}\}$.
2.1.3. No free lunch with vanishing risk. The natural assumption of no arbitrage is routinely replaced in the literature by the slightly stronger, but still economically feasible, assumption of no free lunch with vanishing risk (NFLVR). It was shown in the seminal paper [9] that, when postulated on finite time intervals $[0, t], t \in(0, \infty)$, NFLVR is equivalent to the following statement: for each $t \geq 0$, there exists a probability measure $\mathbb{Q}^{(t)}$, defined on $\mathcal{F}_{t}$, with the following properties:

1. $\left.\mathbb{Q}^{(t)} \sim \mathbb{P}\right|_{\mathcal{F}_{t}}$, where $\left.\mathbb{P}\right|_{\mathcal{F}_{t}}$ is the restriction of the probability measure $\mathbb{P}$ to $\mathcal{F}_{t}$; and
2. the stock-price process $S$ is a $\mathbb{Q}^{(t)}$-local martingale when restricted to the interval $[0, t]$. It will be a standing assumption that the condition of NFLVR and, thus, the equivalent statement above are satisfied by $\left(S_{t}\right)_{t \in[0, \infty)}$ on finite intervals $[0, t], t \in(0, \infty)$. Therefore, for $t \geq 0$, the set of all measures $\mathbb{Q}^{(t)}$ with the above properties is nonempty. We will denote this set by $\mathcal{M}_{t}^{e}$.
2.1.4. Closed market models. It is immediate that, for $0 \leq s<t$, we have the following relation:

$$
\mathcal{M}_{s}^{e}=\left\{\left.\mathbb{Q}^{(t)}\right|_{\mathcal{F}_{s}}: \mathbb{Q}^{(t)} \in \mathcal{M}_{t}^{e}\right\} .
$$

The restriction map turns the family $\left(\mathcal{M}_{t}^{e}\right)_{t \in[0, \infty)}$ into an inversely directed system. In general, such a system will not have an inverse limit in the family of probability measures equivalent to $\mathbb{P}$; i.e., there will exist no set $\mathcal{M}_{\infty}^{e}$ with the property that $\mathcal{M}_{t}^{e}=\left\{\mathbb{Q} \mid \mathcal{F}_{t}: \mathbb{Q} \in \mathcal{M}_{\infty}^{e}\right\}$ for all $t$. In other words, even though the market may admit no arbitrage (NFLVR) on any finite interval $[0, t]$, arbitrage opportunities might arise if we allow the trading horizon to be arbitrarily long. In order to differentiate those cases, we introduce the notion of a closed market model.

Definition 2.2. A market model $\left(S_{t}\right)_{t \in[0, \infty)}$ is said to be closed if there exists a set $\mathcal{M}_{\infty}^{e}$ of probability measures $\mathbb{Q} \sim \mathbb{P}$ such that, for every $t \geq 0, \mathbb{Q}^{(t)} \in \mathcal{M}_{t}^{e}$ if and only if $\mathbb{Q}^{(t)}=\left.\mathbb{Q}\right|_{\mathcal{F}_{t}}$ for some $\mathbb{Q} \in \mathcal{M}_{\infty}^{e}$.

Remark 2.3. Most market models used in practice are not closed. The simplest example is Samuelson's model, where the filtration is generated by a single Brownian motion $\left(W_{t}\right)_{t \in[0, \infty)}$, and the price of the risky asset satisfies $d S_{t}=S_{t}\left(\mu d t+\sigma d W_{t}\right)$ for some constants $\mu \in \mathbb{R}$, $\sigma>0$. For $t \geq 0$, the only element in $\mathcal{M}_{t}^{e}$ corresponds to a Girsanov transformation. However, as $t \rightarrow \infty$, this transformation becomes "more and more singular" with respect to $\left.\mathbb{P}\right|_{\mathcal{F}_{t}}$, and no $\mathbb{Q}$ as in Definition 2.2 can be found (see [21, Remark on p. 193]).

### 2.2. Convex risk measures.

2.2.1. Axioms of convex risk measures. One of the main reasons for the wide use and general acceptance of the theory of risk measures lies in its axiomatic nature. Only the most fundamental traits of an economic agent, such as risk aversion, are encoded parsimoniously into the axioms of risk measures. The resulting theory is nevertheless rich and relevant to the financial practice. The pioneering notion of a coherent risk measure (see [1]) has, soon after its conception, been replaced by a very similar, but more flexible, notion of a convex risk measure (introduced in [16, 11, 14, 17]).

Definition 2.4. A functional $\rho$ mapping $\mathbb{L}^{\infty}(\Omega, \mathcal{F}, \mathbb{P})$ into $\mathbb{R}$ is called a convex risk measure $i f$, for all $f, g \in \mathbb{L}^{\infty}$, we have

$$
\begin{align*}
& \rho(f) \leq 0 \text { if } f \geq 0 \text { a.s. }  \tag{1}\\
& \rho(f-m)=\rho(f)+m, \quad m \in \mathbb{R}  \tag{2}\\
& \rho(\lambda f+(1-\lambda) g) \leq \lambda \rho(f)+(1-\lambda) \rho(g), \quad \lambda \in[0,1] \tag{3}
\end{align*}
$$

(antipositivity), (cash translativity), (convexity).
2.2.2. Replication invariance. The idea that two risky positions which differ only by a quantity replicable in the market at no cost should have the same risk content has appeared very soon after the notion of a risk measure has been applied to the study of financial markets. In order to expand on this tenet, let us, temporarily, pick an arbitrary time $T>0$, and suppose that we are dealing with a finite-horizon financial market $\left(S_{t}\right)_{t \in[0, T]}$, where all finite-horizon analogues of the assumptions and definitions above hold. In such a situation, the investors will trade in the market in order to reduce the overall risk of the terminal position, as measured by the risk measure $\rho$ defined on $\mathbb{L}^{\infty}\left(\mathcal{F}_{T}\right)$. In other words, the combination of the financial market and the risk measure $\rho$ will give rise to a new risk measure, denoted herein by $\rho(\cdot ; T)$, given by

$$
\rho(f ; T)=\inf _{\pi \in \mathcal{A}_{\mathrm{bd}}} \rho\left(f+\int_{0}^{T} \pi_{s} d S_{s}\right) .
$$

We will use the $T$-notation to stress the dependence of this risk measure on the specific maturity date. In addition to axioms (1)-(3) from Definition 2.4 , the functional $\rho(\cdot ; T)$ satisfies the following property:

$$
\begin{equation*}
\rho(f ; T)=\rho\left(f+\int_{0}^{T} \pi_{s} d S_{s} ; T\right) \quad \forall f \in \mathbb{L}^{\infty}\left(\mathcal{F}_{T}\right), \pi \in \mathcal{A}_{\mathrm{bd}} \tag{4}
\end{equation*}
$$

Definition 2.5. A mapping $\rho(\cdot ; T): \mathbb{L}^{\infty}\left(\Omega, \mathcal{F}_{T},\left.\mathbb{P}\right|_{\mathcal{F}_{T}}\right) \rightarrow \mathbb{R}$ is called a replication-invariant convex risk measure if it satisfies axioms (1)-(3) of Definition 2.4 and (4) above.

The notion of replication invariance was introduced in [11] and further developed and generalized in [15]. An accessible discussion of coherent and convex risk measures, as well as the notion of replication invariance, can be found in Chapter 4 of [13].

Remark 2.6.

1. When the market model is complete, the restrictions imposed by adding the replication invariance axiom will necessarily force any replication-invariant risk measure to coincide with the replication price functional (the "Black-Scholes price"). It is only in the setting of incomplete markets that the interplay between risk measurement and trading in the market produces a nontrivial theory.
2. It may seem somewhat counterintuitive at first glance that a replication-invariant risk measure should assign the same risk content to the constant $S_{0}$ as to the random variable $S_{T}$ (where $\left(S_{t}\right)_{t \in[0, T]}$ is a price process of a traded risky asset). The resolution can be found in the fact that the risk contained in $S_{T}$ is virtual since it can be hedged away completely in the financial market. Replication-invariant measures are, however, typically not law-invariant; i.e., there are random variables with the same $\mathbb{P}$-distribution as $S_{T}$ whose risk content is possibly much larger.
The following examples of (maturity-specific) replication-invariant convex risk measures are very well known (see [13]). We use them as test cases for the notion of maturity independence introduced in Definition 3.1. One can easily show that all of them satisfy axioms (1)-(4).

## Example 2.7.

1. Superhedging. For $f \in \mathbb{L}^{\infty}\left(\mathcal{F}_{T}\right)$, let $\hat{\rho}(f ; T)$ be the superhedging price of $f$, i.e.,

$$
\hat{\rho}(f ; T)=\inf \left\{m \in \mathbb{R}: \exists \pi \in \mathcal{A}_{\mathrm{bd}}, \quad \int_{0}^{T} \pi_{s} d S_{s} \geq m+f \text { a.s. }\right\} .
$$

The risk measure $\hat{\rho}(\cdot ; T)$ is extremal in the sense that, for each replication-invariant convex risk measure $\rho(\cdot ; T)$, we have $\hat{\rho}(f ; T) \geq \rho(f ; T)$ for all $f \in \mathbb{L}^{\infty}\left(\mathcal{F}_{T}\right)$.
2. Entropic risk measures. For $f \in \mathbb{L}^{\infty}\left(\mathcal{F}_{T}\right)$, the entropic risk measure $\rho(f ; T)$, with risk aversion coefficient $\gamma>0$, is defined as the unique solution $\rho \in \mathbb{R}$ to the indifferencepricing equation

$$
\begin{align*}
& \sup _{\pi \in \mathcal{A}_{\mathrm{bd}}} \mathbb{E}\left[-\exp \left(-\gamma\left(x+\rho+f+\int_{0}^{T} \pi_{s} d S_{s}\right)\right)\right]  \tag{2.1}\\
= & \sup _{\pi \in \mathcal{A}_{\mathrm{bd}}} \mathbb{E}\left[-\exp \left(-\gamma\left(x+\int_{0}^{T} \pi_{s} d S_{s}\right)\right)\right], \quad x \in \mathbb{R} .
\end{align*}
$$

The value $\rho(-f ; T)$ at the negative of $f$ is also known as the exponential indifference price $\nu(f ; T)$ of $f$. The measure $\rho(\cdot ; T)$ admits a simple dual representation

$$
\begin{equation*}
\rho(f ; T)=\sup _{\mathbb{Q} \in \mathcal{M}_{T}^{e}}\left(\mathbb{E}^{\mathbb{Q}}[-f]-\frac{1}{\gamma} H(\mathbb{Q} \mid \mathbb{P} ; T)\right), \tag{2.2}
\end{equation*}
$$

where the relative entropy $H(\mathbb{Q} \mid \mathbb{P} ; T)$ of $\mathbb{Q} \in \mathcal{M}_{T}^{e}$ with respect to $\mathbb{P}$ is given by

$$
H(\mathbb{Q} \mid \mathbb{P} ; T)=\mathbb{E}^{\mathbb{Q}}\left[\ln \left(\frac{d \mathbb{Q}}{d\left(\left.\mathbb{P}\right|_{\mathcal{F}_{T}}\right)}\right)\right] \in[0, \infty] .
$$

3. General replication-invariant risk measures. Under appropriate topological regularity conditions replication-invariant convex risk measure $\rho(\cdot ; T): \mathbb{L}^{\infty}\left(\mathcal{F}_{T}\right) \rightarrow \mathbb{R}$ admits the following dual representation:

$$
\begin{equation*}
\rho(f ; T)=\sup _{\mathbb{Q} \in \mathcal{M}_{T}^{e}}\left(\mathbb{E}^{\mathbb{Q}}[-f]-\alpha(\mathbb{Q})\right) \tag{2.3}
\end{equation*}
$$

for some convex penalty function $\alpha: \mathcal{M}_{T}^{e} \rightarrow[0, \infty]$. See Theorem 17, p. 445 in [11] for the proof in the discrete-time case. The proof in our setting is similar.

## 3. Maturity-independent risk measures.

3.1. The need for maturity independence. The classical notion of a convex risk measure as well as its replication-invariant specialization are inextricably linked to a specific maturity date with respect to which risk measurement is taking place while ignoring all other time instances. On the other hand, a fundamental property of financial markets is that they facilitate transfers of wealth among different time points as well as between different states of the world. The notion of replication invariance, introduced above, abstracts the latter property and ties it to the decision-theoretic notion of a convex risk measure. The former property, however, has not yet been incorporated into the risk measurement framework in the same manner in the existing literature. One of the goals herein is to do exactly this. We, then, pose and address the following question:

Is there a class of risk measures that are not constructed in reference to a specific time instance and can be, thus, used to measure the risk content of claims of all (arbitrary) maturities?
Equivalently, we wish to avoid the case when two versions of the same risk measure (differing only in the choice of the maturity date) give different risk values to the same contingent claim. ${ }^{1}$

Before we proceed with formal definitions, let us recall some of the fundamental properties of the arbitrage-free pricing ("Black-Scholes") functional $P_{B S}$ in the context of a complete financial market. For a "regular-enough" contingent claim $f$, the value $P_{B S}(f)$ is defined as the capital needed at inscription to replicate it perfectly and is computed by taking the expectation $\mathbb{E}^{\mathbb{Q}}[f]$ with respect to the unique risk-neutral measure $\mathbb{Q}$. The functional $\rho_{B S}(f)=P_{B S}(-f)$ satisfies the axioms of convex risk measures and is replication-invariant. Moreover, it is per se unaffected by the expiration date of the generic claim $f$. Indeed, the risk-neutral

[^20]pricing measure corresponding to the earlier time horizon $s$ is the restriction of the time- $t$ riskneutral pricing measure to a smaller $\sigma$-algebra $\mathcal{F}_{s}$. Therefore, two claims, one with maturity $t$ and the other with maturity $s$, which give identical payoffs, measurable with respect to the $\sigma$-algebra $\mathcal{F}_{s}$, will be given identical risk-neutral prices.

The observation that risk-neutral prices in complete markets are maturity-independent (in the above sense) leads naturally to axiomatization and a search for examples of decisiontheoretic tools consistent with it, even in the absence of completeness. To give the reader a better feel for what exactly such an axiom entails, here is how it can be deduced from even more fundamental (but somewhat harder to make precise) postulates. Consider an economic agent who values the risks of two contingent claims with the same payoffs but different maturities in a simplified market environment where the time value of money is 0 . Let the dollar value of this payoff be modeled by a random variable $f \in \mathcal{F}_{s}$, and let the two maturities be $s$ and $t>s$. The agent's risk measurements of these two contingent claims are denoted $\rho(f, s)$ and $\rho(f, t)$.

To relate $\rho(f, s)$ and $\rho(f, t)$, we let the agent act according to the following strategy: at time 0 , he/she invests $\rho(f, s)$ dollars in an optimal way and ends up with an acceptable total wealth of $\rho(f, s)+\int_{0}^{s} \pi_{u} d S_{u}+f$ at time $s$. After that, he/she relaxes and does nothing up to time $t$. Therefore, the influx of $\rho(f, s)$ dollars at time 0 surely guarantees a trading strategy (namely, $\left.\pi 1_{[0, s]}+01_{(s, t]}\right)$ which makes $f$ acceptable at time $t$. Consequently, $\rho(f, t) \leq \rho(f, s)$. On the other hand, $\rho(f, t)$ dollars can be used to produce a trading strategy $\hat{\pi}$ such that $\rho(f, t)+\int_{0}^{t} \hat{\pi}_{u} d S_{u}+f$ is acceptable at time $t$. Then, $\rho(f, t)+\int_{0}^{s} \hat{\pi}_{u} d S_{u}+f$ must also be acceptable at time $s$, which would imply $\rho(f, t) \leq \rho(f, s)$, and, consequently, $\rho(f, s)=\rho(f, t)$. This is, indeed, the case; otherwise, the two time- $t$ positions

$$
X=\rho(f, t)+\int_{0}^{t} \hat{\pi}_{u} d S_{u}+f \quad \text { and } \quad \hat{X}=\rho(f, t)+\int_{0}^{s} \hat{\pi}_{u} d S_{u}+f
$$

which differ in a zero-cost-replicable quantity $\int_{s}^{t} \hat{\pi}_{u} d S_{u}$, would differ in acceptability.
If we analyze the assumptions made in the above argument, we reach the conclusion that whenever one

1. does not change his/her attitude towards risk from one time point to another in a qualitative way and
2. accepts replication invariance for claims with a fixed maturity, one must also accept the axiom of maturity independence. The above two properties also point towards situations when maturity independence is not appropriate. This is the case, for example, if there is a cost (or any other kind of friction) associated with transfer of wealth across time instances or if the agent's general attitude towards risk changes with time (due to effects of more/less stringent regulation, e.g.).
3.2. Definition of maturity independence. Let $\mathbb{L}$ denote the set of all bounded random variables with finite maturities, i.e.,

$$
\mathbb{L}=\bigcup_{t \geq 0} \mathbb{L}^{\infty}\left(\mathcal{F}_{t}\right)
$$

The set $\mathbb{L}$ will serve as a natural domain for the class of risk measures we propose in what follows. Note that $\mathbb{L}$ contains all $\mathcal{F}_{t}$-measurable bounded contingent claims for all times $t \geq 0$,
but it avoids the (potentially pathological) cases of random variables in $\mathbb{L}^{\infty}\left(\mathcal{F}_{\tau}\right)$, where $\tau$ is a finite, but possibly unbounded, stopping time.

We are now ready to define the class of maturity-independent risk measures. With a slight abuse of notation, we still use the symbol $\rho$. In contrast to their maturity-dependent counterparts $\rho(\cdot ; T)$, however, all maturity-specific notation has vanished.

Definition 3.1. A functional $\rho: \mathbb{L} \rightarrow \mathbb{R}$ is called $a$ maturity-independent convex risk measure if it has the following properties for all $f, g \in \mathbb{L}$, and $\lambda \in[0,1]$ :
(1) $\quad \rho(f) \leq 0 \quad \forall f \geq 0$
(4) $\forall t \geq 0$, and $\pi \in \mathcal{A}_{\mathrm{bd}}, \rho\left(f+\int_{0}^{t} \pi_{s} d S_{s}\right)=\rho(f)$ (replication and maturity independence).

We note that the properties which differentiate the maturity-independent risk measures from the existing notions are the choice of the domain $\mathbb{L}$ on the one hand and the validity of axiom (4) for all maturities $t \geq 0$ on the other.
3.3. Motivational examples. We start off our investigation of maturity-independent risk measures by giving three examples - one of an extremal such risk measure, one of a class of maturity-independent risk measures for closed markets, and one in which the maturity independence property fails.
3.3.1. Superhedging prices. The simplest example of a maturity-independent risk measure is the superhedging price, i.e., its risk-measure analogue $\hat{\rho}: \mathbb{L} \rightarrow \mathbb{R}$ given by

$$
\hat{\rho}(f)=\inf \left\{m \in \mathbb{R}: \exists \pi \in \mathcal{A}_{\mathrm{bd}}, m+\int_{0}^{\infty} \pi_{s} d S_{s}+f \geq 0 \text { a.s. }\right\}
$$

It is easy to see that it satisfies all axioms in Definition 3.1. As in the maturity-dependent case, $\hat{\rho}$ has the extremal property $\hat{\rho}(f) \geq \rho(f)$ for any $f \in \mathbb{L}$ and any maturity-independent risk measure $\rho$.
3.3.2. The case of closed markets. The dual characterization (2.3) of replicationinvariant risk measures for finite maturities can be used to construct maturity-independent risk measures when the market model is closed (see Definition 2.2 and subsection 2.1.4 for notation and terminology). Indeed, let $\alpha: \mathcal{M}_{\infty}^{e} \rightarrow[0, \infty]$ be a proper function (i.e., satisfying $\alpha(\mathbb{Q})<\infty$ for at least one $\left.\mathbb{Q} \in \mathcal{M}_{\infty}^{e}\right)$. It is not difficult to check that the functional $\rho: \mathbb{L} \rightarrow \mathbb{R}$, defined by

$$
\rho(f)=\sup _{\mathbb{Q} \in \mathcal{M}_{\infty}^{e}}\left(\mathbb{E}^{\mathbb{Q}}[-f]-\alpha(\mathbb{Q})\right),
$$

is a maturity-independent risk measure. We have already seen that many market models used in practice are not closed. The natural construction used above will clearly not be applicable in those cases, and, thus, an entirely different approach will be needed.


Figure 1. The market tree.
3.3.3. Risk measures lacking maturity independence. When markets are incompleteand no canonical ("Black-Scholes") pricing mechanism exists - some traditional and widely used risk measures are not maturity-independent. In other words, under these measures, indifference prices of the same contingent claim but calculated in terms of two distinct maturities will, in general, differ.

It is tempting to assume that a maturity-independent risk measure $\rho$ can always be constructed by identifying a maturity date $t$ associated with a contingent claim $f$ and setting $\rho(f)=\rho(f ; t)$ for some replication-invariant risk measure $\rho(\cdot ; t)$. As shown in the following two examples, this construction will not always be possible even if we restrict our attention to the well-explored class of entropic risk measures. Both examples are based on the entropic risk measure (see 2 in Example 2.7). Note that the first example can easily be fitted in the framework described in section 2 by extending its paths to be constant on intervals $[0,1)$, $[1,2)$, and $[2, \infty)$.
(a) A noncompliance example on a finite probability space. We present a simple two-period example in which entropic risk measurement gives different results for the same time-1-measurable contingent claim $f$ when considered at times 1 and 2 . The market structure is described by the simple tree in Figure 1, where the (physical) probability of each of the branches leaving the initial node is $\frac{1}{3}$, and the conditional probabilities of the two contingencies (leading to $S_{4}$ and $S_{5}$ ) after the node $S_{3}$ are equal to $\frac{1}{3}$ and $\frac{2}{3}$, respectively. One can implement the described situation on a 4 -element probability space $\Omega=\left\{\omega_{1}, \omega_{2}, \omega_{3}, \omega_{4}\right\}$, as in Figure 1, with $\mathbb{P}\left[\omega_{1}\right]=\mathbb{P}\left[\omega_{2}\right]=1 / 3, \mathbb{P}\left[\omega_{3}\right]=1 / 9$, and $\mathbb{P}\left[\omega_{4}\right]=2 / 9$.

There are two financial instruments: a riskless bond $S^{0} \equiv 1$, and a stock $S=S^{1}$ whose price is denoted by $S_{0}, \ldots, S_{5}$ for various nodes of the information tree, such that the following relations hold:

$$
S_{0}=S_{2}, \quad S_{2}=\frac{1}{2}\left(S_{1}+S_{3}\right), \quad S_{1} \neq S_{3}, \quad S_{3}=\frac{1}{2}\left(S_{4}+S_{5}\right), \quad S_{4} \neq S_{5} .
$$

This implies, in particular, that the market is arbitrage-free, and, due to its incompleteness, the set of equivalent martingale measures is larger that just a singleton. Next, we consider a family $\left\{f_{a}\right\}_{a>0}$ of contingent claims defined by

$$
f_{a}(\omega)= \begin{cases}0, & \omega=\omega_{1}, \omega_{2} \\ a, & \omega=\omega_{3}, \omega_{4}\end{cases}
$$

We are going to compare $\rho\left(f_{a} ; 1\right)$ and $\rho\left(f_{a} ; 2\right)$, where $\rho\left(f_{a} ; t\right), t=1,2$, is the value of the entropic risk measure (as defined in (2.1)) of the contingent claim $f_{a}$, seen as the time-t random variable (note that $f_{a}$ is $\mathcal{F}_{1}$-measurable for all $a$ ).

Let us first focus on $\rho(f ; 2)$. The set of all martingale measures is given by $\mathcal{M}=$ $\left\{\mathbb{Q}^{\nu}: \nu \in\left(-\frac{1}{6}, \frac{1}{3}\right)\right\}$, where

$$
\mathbb{Q}^{\nu}(\omega)= \begin{cases}\frac{1}{3}-\nu, & \omega=\omega_{1} \\ \frac{1}{3}+2 \nu, & \omega=\omega_{2} \\ \frac{1}{2}\left(\frac{1}{3}-\nu\right), & \omega=\omega_{3}, \omega_{4}\end{cases}
$$

By a finite-dimensional analogue of (2.2), we have

$$
\begin{align*}
\rho\left(f_{a} ; 2\right) & =\sup _{\nu \in(-1 / 6,1 / 3)}\left(\mathbb{E}^{\mathbb{Q}^{\nu}}\left[-f_{a}\right]-h_{2}(\nu)\right) \\
& =\sup _{\nu \in(-1 / 6,1 / 3)}\left(-a(1 / 3-\nu)-h_{2}(\nu)\right), \tag{3.1}
\end{align*}
$$

where, as one can easily check, the relative-entropy function $h_{2}$ is given by

$$
h_{2}(\nu)=\bar{h}_{2}(\nu)-\inf _{\mu} \bar{h}_{2}(\mu),
$$

where

$$
\bar{h}_{2}(\nu)=\frac{\mathbb{Q}^{\nu}\left[\omega_{1}\right]}{\mathbb{P}\left[\omega_{1}\right]} \ln \left(\frac{\mathbb{Q}^{\nu}\left[\omega_{1}\right]}{\mathbb{P}\left[\omega_{1}\right]}\right)+\frac{\mathbb{Q}^{\nu}\left[\omega_{2}\right]}{\mathbb{P}\left[\omega_{\omega}\right]} \ln \left(\frac{\mathbb{Q}^{\nu}\left[\omega_{2}\right]}{\mathbb{P}\left[\omega_{2}\right]}\right)+\frac{\mathbb{Q}^{\nu}\left[\omega_{3}\right]}{\mathbb{P}\left[\omega_{3}\right]} \ln \left(\frac{\mathbb{Q}^{\nu}\left[\omega_{3}\right]}{\mathbb{P}\left[\omega_{3}\right]}\right)+\frac{\mathbb{Q}^{\nu}\left[\omega_{4}\right]}{\mathbb{P}\left[\omega_{4}\right]} \ln \left(\frac{\mathbb{Q}^{\nu}\left[\omega_{4}\right]}{\mathbb{P}\left[\omega_{4}\right]}\right) .
$$

Similarly,

$$
\begin{align*}
\rho\left(f_{a} ; 1\right) & =\sup _{\nu \in(-1 / 6,1 / 3)}\left(\mathbb{E}^{\mathbb{Q}^{\nu}}\left[-f_{a}\right]-h_{1}(\nu)\right) \\
& =\sup _{\nu \in(-1 / 6,1 / 3)}\left(-a(1 / 3-\nu)-h_{1}(\nu)\right), \tag{3.2}
\end{align*}
$$

where the function $h_{1}$ is given by $h_{1}(\nu)=\bar{h}_{1}(\nu)-\inf _{\nu} \bar{h}_{1}(\nu)$, with

$$
\bar{h}_{1}(\nu)=\mathbb{Q}^{\nu}\left[\omega_{1}\right] \ln \left(\frac{\mathbb{Q}^{\nu}\left[\omega_{1}\right]}{\mathbb{P}\left[\omega_{1}\right]}\right)+\mathbb{Q}^{\nu}\left[\omega_{2}\right] \ln \left(\frac{\mathbb{Q}^{\nu}\left[\omega_{2}\right]}{\mathbb{P}\left[\omega_{2}\right]}\right)+\left(\mathbb{Q}^{\nu}\left[\omega_{3}\right]+\mathbb{Q}^{\nu}\left[\omega_{4}\right]\right) \ln \left(\frac{\mathbb{Q}^{\nu}\left[\omega_{3}\right]+\mathbb{Q}^{\nu}\left[\omega_{4}\right]}{\mathbb{P}\left[\omega_{3}\right]+\mathbb{P}\left[\omega_{4}\right]}\right) .
$$

The expressions (3.1) and (3.2) can be seen as the Legendre-Fenchel transforms of the translated entropy functions $h_{2}(1 / 3-\nu)$ and $h_{1}(1 / 3-\nu)$. Therefore, by the bijectivity of these transforms and the convexity of the functions $h_{1}$ and $h_{2}$, the equality $\rho\left(f_{a} ; 1\right)=\rho\left(f_{a} ; 2\right)$, for all $a>0$, would imply that $h_{1}=h_{2}$. It is now a matter of a straightforward computation to show that that is, in fact, not the case. Thus, the two values do not coincide; i.e., for at least one $a>0$,

$$
\rho\left(f_{a} ; 1\right) \neq \rho\left(f_{a} ; 2\right) .
$$

(b) A noncompliance example in a diffusion market model. We consider a financial market as in section 2 , with $k=1$ (one risky asset) and an augmentation of the filtration generated by two independent driving Brownian motions $\left(W_{t}^{1}\right)_{t \in[0, \infty)}$ and $\left(W_{t}^{2}\right)_{t \in[0, \infty)}$. It will be enough to consider a stock-price process with stochastic volatility of the form

$$
\begin{equation*}
d S_{s}=S_{s}\left(\mu d s+\sigma\left(B_{s}\right) d W_{s}\right) \tag{3.3}
\end{equation*}
$$

$s \geq 0$, where $B_{t}=\rho W_{t}^{1}+\sqrt{1-\rho^{2}} W_{t}^{2}$ is a Brownian motion correlated with $W^{1}$, with the correlation coefficient $\rho \in(0,1)$. It will be convenient to introduce the market price of risk $\lambda(y)=\mu / \sigma(y)$, assuming throughout that $\lambda: \mathbb{R} \rightarrow(0, \infty)$ is a strictly increasing $C^{1}$-function with range of the form $(\varepsilon, M)$ for some constants $0<\varepsilon<M<\infty$. In addition to its usefulness in what follows, this assumption will guarantee that the condition NFLVR holds on each finite time interval $[0, t]$. To facilitate the dynamic-programming approach, we assume that trading starts at time $t$, after which two maturities $T, \bar{T}$, with $T<\bar{T}$, are chosen.

Let $C_{T}=-B_{T}$ model the payoff of a contingent claim which is, clearly, nonreplicable. The value of the time- $t$ entropic $(\gamma=1)$ risk measure $\rho_{t}\left(C_{T} ; T\right)$ equals the indifference price $\nu_{t}\left(-C_{T} ; T\right)$ of the claim $B_{T}$ measured on the trading horizon $[t, T]$. According to [32], $\rho_{t}\left(C_{T} ; T\right)$ admits a representation in terms of a solution to a partial differential equation. More precisely, taking into account the fact that neither the payoff $C_{T}$ nor the dynamics of the volatility depends on the stock price, we have $\rho_{t}\left(C_{T} ; T\right)=p\left(t,-B_{t}\right)$ a.s., where the function $p:[0, T] \times \mathbb{R} \rightarrow \mathbb{R}$ is a classical solution of the quasi-linear equation

$$
\left\{\begin{array}{l}
p_{t}+\mathcal{L}^{f} p+\frac{1}{2}\left(1-\rho^{2}\right) p_{y}^{2}=0  \tag{3.4}\\
p(T, y)=y
\end{array}\right.
$$

where $\mathcal{L}^{f} p=\frac{1}{2} p_{y y}+\left(f_{y} / f-\rho \lambda(y)\right) p_{y}$. The function $f:[0, T] \times \mathbb{R} \rightarrow \mathbb{R}$ is the unique solution to the linear problem

$$
\left\{\begin{array}{l}
f_{t}+\mathcal{A} f=0  \tag{3.5}\\
f(T, y)=1
\end{array}\right.
$$

where $\mathcal{A} f=\frac{1}{2} f_{y y}-\rho \lambda(y) f_{y}-\frac{1}{2}\left(1-\rho^{2}\right) \lambda^{2}(y) f$. Standard arguments show that $f$ is of class $C^{1,3}$ and admits a representation in the manner of Feynman and Kac as

$$
\begin{equation*}
f(t, y)=\mathbb{E}\left[\left.e^{\int_{t}^{T} \frac{\left(1-\rho^{2}\right)}{2} \lambda^{2}\left(Y_{s}\right) d s} \right\rvert\, Y_{t}=y\right], \quad(t, y) \in[0, T] \times \mathbb{R}, \tag{3.6}
\end{equation*}
$$

where $\left\{Y_{s}\right\}_{s \in[t, \infty)}$ is the unique strong solution to $d Y_{s}=d B_{s}-\rho \lambda\left(Y_{s}\right) d s, Y_{t}=y$. In particular, there exists a constant $C>1$ such that $1 \leq f(t, y) \leq C$ for $(t, y) \in[0, T] \times \mathbb{R}$.

Similarly, the indifference price $\nu_{t}\left(-C_{T} ; \bar{T}\right)$ (which equals the value $\rho_{t}\left(C_{T} ; \bar{T}\right)$ of the maturity- $\bar{T}$ entropic risk measure $\rho_{t}(\cdot ; \bar{T})$ applied to the same contingent claim only on the longer horizon $[0, \bar{T}], \bar{T}>T)$ can be represented via $\bar{p}(t, y)$, where $\bar{p}$ solves

$$
\left\{\begin{array}{l}
\bar{p}_{t}+\mathcal{L}^{\bar{f}} \bar{p}+\frac{1}{2}\left(1-\rho^{2}\right) \bar{p}_{y}^{2}=0,  \tag{3.7}\\
\bar{p}(T, y)=y .
\end{array}\right.
$$

Herein, $\mathcal{L}^{\bar{f}}$ is given as in (3.5) with $f$ replaced by the function $\bar{f}$, which solves

$$
\left\{\begin{array}{l}
\bar{f}_{t}+\mathcal{A} \bar{f}=0  \tag{3.8}\\
\bar{f}(\bar{T}, y)=1
\end{array}\right.
$$

Just like $f$, the function $\bar{f}$ admits a representation analogous to (3.6) and a uniform bound $1 \leq \bar{f}(t, y) \leq \bar{C}$ for $(t, y) \in[0, T] \times \mathbb{R}$.

The goal of this example is to show that the indifference prices $\nu\left(B_{T} ; T\right)$ and $\nu\left(B_{T} ; \bar{T}\right)$, or, equivalently, the entropic risk measures $\rho_{t}\left(C_{T} ; T\right)$ and $\rho_{t}\left(C_{T} ; \bar{T}\right)$, do not always coincide, i.e., that $p(t, y)$ and $\bar{p}(t, y)$ differ for at least one choice of $(t, y) \in[0, T) \times \mathbb{R}$. We start with an auxiliary result, namely,

$$
\begin{equation*}
\frac{f_{y}(T, y)}{f(T, y)} \neq \frac{\bar{f}_{y}(T, y)}{\bar{f}(T, y)} \quad \text { for each } y \in \mathbb{R} \tag{3.9}
\end{equation*}
$$

In order to establish (3.9), we note that the function $g:[0, T] \times \mathbb{R} \rightarrow \mathbb{R}$, defined by $g=f_{y}$, is a classical solution to

$$
\left\{\begin{array}{l}
g_{t}+\mathcal{B} g=0  \tag{3.10}\\
g(T, y)=0
\end{array}\right.
$$

where

$$
\mathcal{B} g=\frac{1}{2} g_{y y}-\rho \lambda(y) g_{y}-A(y) g-B(t, y),
$$

with $A(y)=\rho \lambda^{\prime}(y)+\frac{1}{2}\left(1-\rho^{2}\right) \lambda^{2}(y)$ and $B(t, y)=\left(1-\rho^{2}\right) \lambda(y) \lambda^{\prime}(y) f(t, y)$.
Thanks to the assumptions placed on $\rho$ and $\lambda$, and the positivity of $f$, we have

$$
\begin{equation*}
A(y)>0 \quad \text { and } \quad B(t, y)>0 \quad \forall(t, y) \in[0, T] \times \mathbb{R} . \tag{3.11}
\end{equation*}
$$

The function $\bar{g}=\bar{f}_{y}$ is defined in an analogous fashion (only on the larger domain $[0, \bar{T}] \times \mathbb{R}$ ), and a similar set of properties can be derived. Since $f_{y}(T, y)=0$ for all $y \in \mathbb{R}$, it will be enough to show that $\bar{f}_{y}(T, y)>0$ for all $y \in \mathbb{R}$. This follows immediately from the strict inequalities in (3.11) and the Feynman-Kac representation

$$
\begin{equation*}
\bar{g}(T, y)=\bar{f}_{y}(T, y)=\mathbb{E}\left[\int_{T}^{\bar{T}} B\left(t, Y_{t}\right) e^{\int_{t}^{\bar{T}} A\left(Y_{s}\right) d s} d t \mid Y_{T}=y\right], \quad y \in \mathbb{R} \tag{3.12}
\end{equation*}
$$

Having established (3.9), we conclude that, thanks to the smoothness of the functions $f$ and $\bar{f}$, the operators $\mathcal{L}^{f}$ and $\mathcal{L}^{\bar{f}}$ differ in the $\frac{\partial}{\partial y}$-coefficient in some open neighborhood $\mathcal{N}$ of the line $\{T\} \times \mathbb{R}$ in $[0, T] \times \mathbb{R}$. Assuming that $\bar{p}$ and $p$ coincide in $\mathcal{N}$, subtracting the equations (3.4) and (3.7) yields

$$
\begin{equation*}
\left(\frac{f_{y}}{f}(t, y)-\frac{\bar{f}_{y}}{\bar{f}}(t, y)\right) \bar{p}_{y}(t, y)=0 \quad \text { for }(t, y) \in \mathcal{N} . \tag{3.13}
\end{equation*}
$$

Equation (3.9) now implies that $\bar{p}_{y}=0$ on $\mathcal{N}$, which is clearly in contradiction with the terminal condition $\bar{p}(T, y)=y, y \in \mathbb{R}$. Therefore, there exists $(t, y) \in \mathcal{N} \backslash\{T\} \times \mathbb{R} \subseteq[0, T) \times \mathbb{R}$ such that $p(t, y) \neq \bar{p}(t, y)$.
4. Forward entropic risk measures. In the previous section, we saw three examples of risk measures and their dependence on the specific choice of the maturity date. In particular, we pointed out that the superhedging risk measure in subsection 3.3.1, as well as the ones constructed in subsection 3.3.2, for the class of closed markets, are maturity-independent. However, both of these classes are rather restrictive. Indeed, the one associated with superhedging is extremely conservative, while the other requires the rather stringent assumption of market closedness.

In this section, we introduce a new family of convex risk measures that have the maturity independence property and, at the same time, are applicable to a wide range of settings. Their construction is based on the idea mentioned in the introductory paragraph of subsection 3.3.3 but avoids the pitfalls responsible for the failure of examples (a) and (b) following it.

The risk measures we are going to introduce are closely related to indifference prices. The novelty of the approach is that the underlying risk preference functionals are not tied down to a specific maturity, as has been the case in the standard expected utility formulation. Rather, they can be seen as specified at initiation and subsequently "generated" across all times. This approach was proposed by the first author and Musiela (see [25, 29, 27, 26, 28]) and is briefly reviewed below.
4.1. Forward exponential performances. The notion of a forward performance process has arisen from the search for ways to measure the performance of investment strategies across all times in $[0, \infty)$. In order to produce such a nontrivial object, we look for a random field $U=U_{t}(\omega, x)$ defined for all times $t \geq 0$ and parametrized by a wealth argument $x$ such that the mapping $x \mapsto U_{t}(\omega, x)$ admits the classical properties of utility functions. More precisely, we have the following definition.

Definition 4.1. A mapping $U:[0, \infty) \times \Omega \times \mathbb{R} \rightarrow \mathbb{R}$ is called a performance random field if the following hold:
(1) for each $(t, \omega) \in[0, \infty) \times \Omega$, the mapping $x \mapsto U_{t}(x, \omega)$ defines a utility function: it is strictly concave, strictly increasing, continuously differentiable, and satisfies the Inada conditions $\lim _{x \rightarrow \infty} U^{\prime}(x)=0$ and $\lim _{x \rightarrow-\infty} U^{\prime}(x)=+\infty$;
(2) $U .(\cdot, \cdot)$ is measurable with respect to the product of the progressive $\sigma$-algebra on $\Omega \times$ $[0, \infty)$ and the Borel $\sigma$-algebra on $\mathbb{R}$; and
(3) $\mathbb{E}\left|U_{t}(x)\right|<\infty$ for all $(t, x) \in[0, \infty) \times \mathbb{R}$.

Remark 4.2.

1. The last requirement in Definition 4.1 implies, in particular, that $\mathbb{E}\left|U_{t}(\xi)\right|<\infty$ for all random variables $\xi \in \mathbb{L}^{\infty}$.
2. It is possible to construct a parallel theory where the performance functions $U_{t}(\omega, \cdot)$ are defined on the positive semiaxis $(0, \infty)$. We choose the domain $\mathbb{R}$ for the wealth argument $x$ because it leads to a slightly simpler analysis and because the examples to follow will be based on the exponential function.
On an arbitrary trading horizon, say $[s, t], 0 \leq s<t<\infty$, the investor whose preferences are described by the random field $U$ seeks to maximize the expected investment performance:

$$
\begin{equation*}
V_{s}^{t}(x)=\underset{\pi \in \mathcal{A}_{\mathrm{bd}}}{\operatorname{esssup}} \mathbb{E}\left[U_{t}\left(X_{t}^{x, \pi}\right) \mid \mathcal{F}_{s}\right], \quad 0 \leq s \leq t \tag{4.1}
\end{equation*}
$$

Herein, $X^{x, \pi}$ denotes the investor's wealth process, $x \in \mathbb{R}$ the investor's initial wealth at
time $s$, and $\pi$ a generic investment strategy belonging to $\mathcal{A}_{\mathrm{bd}}$ (the set of admissible policies introduced in subsection 2.1.2). To concentrate on the new notions, we abstract throughout from control and state constraints as well as the most general specification of admissibility requirements.

It has been argued in [28] that the class of performance random fields with the additional property

$$
\begin{equation*}
V_{s}^{t}(x)=U_{s}(x) \text { a.s. } \quad \forall 0 \leq s \leq t<\infty, \quad x \in \mathbb{R} \tag{4.2}
\end{equation*}
$$

possesses several desirable properties and gives rise to an analytically tractable theory.
Definition 4.3. A random field $U$ satisfying (4.2), where $V$ is defined by (4.1), is called self-generating.

Remark 4.4. We remind the reader that a classical example of a self-generating performance random field (albeit only on the finite horizon $[0, T]$ ) is the traditional value function, defined as

$$
U_{t}(x)=\underset{\pi \in \mathcal{A}_{\mathrm{bd}}}{\operatorname{esssup}} \mathbb{E}\left[U_{T}\left(X_{T}^{x, \pi}\right) \mid \mathcal{F}_{t}\right], \quad t \in[0, T], \quad x \in \mathbb{R}
$$

where $T$ is a prespecified maturity beyond which no investment activity is measured, and $U_{T}(\cdot, \cdot): \Omega \times \mathbb{R} \rightarrow \mathbb{R}$ is a classical (state-dependent) utility function (see, for example, [20, $23,22,34]$ ). When the horizon is infinite, such a construction will not produce any results. Indeed, there is no appropriate time for the final datum to be given.

What (4.1) and (4.2) tell us is that (under additional regularity conditions) the sought-after criterion (performance random field) $U$ must have the property that the stochastic process $U_{t}\left(X_{t}^{x, \pi}\right)$ is a supermartingale for an arbitrary control $\pi \in \mathcal{A}_{\text {bd }}$ and becomes "closer and closer" to a martingale as the controls get "better and better." In the case when the class of control problems (4.1) actually admits an optimizer $\pi^{*} \in \mathcal{A}_{\mathrm{bd}}$ (or in some larger, appropriately chosen, class), the composition $U_{t}\left(X_{t}^{x, \pi^{*}}\right)$ becomes a martingale.

In the traditional framework, as already mentioned in Remark 4.4, the datum (terminal utility) is assigned at some fixed future time $T$. Alternatively, in the case of an infinite time horizon, it is more natural to think of the datum $u_{0}: \mathbb{R} \rightarrow \mathbb{R}$ as being assigned at time $t=0$ and a self-generating performance random field $U_{t}$ chosen so that $U_{0}(x)=u_{0}(x)$. It is because of this interpretation that the self-generating performance random fields may also be referred to as forward performances.

The notion of forward performance processes was first developed for binomial models in $[25,29]$ and later generalized to diffusion models with a stochastic factor [26] and, more recently, to models of Itô asset price dynamics (see, among others, [26, 28] as well as [4]). A related stochastic optimization problem that allows for semimartingale price processes and random horizons can be found in [8]. A similar notion of utilities without horizon preference was developed in [18]; therein, asset prices are taken to be lognormal, leading to deterministic forward solutions.

While traditional performance random fields on finite horizons are straightforward to construct and characterize, producing a "forward" performance random field on $[0, \infty)$ from a given initial datum $u_{0}$ is considerably more difficult. Several examples of such a construction, all based on the exponential initial datum, are given in the following subsection. These
random fields are the most important building blocks for the class of maturity-independent risk measures presented in subsection 4.2.

Definition 4.5. A performance random field $U$ is called a forward exponential performance if
(1) it is self-generating, and
(2) there exists a constant $\gamma>0$ such that

$$
\begin{equation*}
U_{0}(x)=-e^{-\gamma x}, \quad x \in \mathbb{R} . \tag{4.3}
\end{equation*}
$$

The construction presented below can be found in [27]. The assumptions and definitions from section 2 will be used in what follows without explicit mention. For the statement of Theorem 4.8 (and, also, for some of the later sections), we introduce an additional set of assumptions on the structure of $\left(S_{t}\right)_{t \in[0, \infty)}$. Before we do, we remind the reader that the Moore-Penrose pseudoinverse of a real $m \times n$ matrix $M$ is the unique real $n \times m$ matrix $A^{+}$ such that $A A^{+} A=A$ and $A^{+} A A^{+}=A^{+}$, and the matrices $A A^{+}$and $A^{+} A$ are symmetric.

Assumption 4.6. The filtration $\mathbb{F}$ is the usual augmentation of the filtration generated by a $d$-dimensional Brownian motion $\left(W_{t}\right)_{t \in[0, \infty)}=\left(W_{t}^{1}, \ldots, W_{t}^{d}\right)_{t \in[0, \infty)}$ and $\left(S_{t}\right)_{t \in[0, \infty)}=$ $\left(S_{t}^{1}, \ldots, S_{t}^{k}\right)_{t \in[0, \infty)}$ is an Itô process of the form

$$
\begin{equation*}
d S_{t}^{i}=S_{t}^{i}\left(\mu_{t}^{i} d t+\sum_{j=1}^{d} \sigma_{t}^{j i} d W_{t}^{j}\right) \tag{4.4}
\end{equation*}
$$

for $t \geq 0, i=1, \ldots, k$, and $j=1, \ldots, d$, where the processes $\left(\mu_{t}^{i}\right)_{t \in[0, \infty)}$ and $\left(\sigma_{t}^{j i}\right)_{t \in[0, \infty)}$ are $\mathbb{F}$ progressively measurable and bounded uniformly by a deterministic constant on each segment $[0, t], t>0$.

Moreover, the matrix $\sigma$ admits a progressively measurable and bounded Moore-Penrose pseudoinverse $\sigma^{+}$. Consequently, the $d$-dimensional bounded and progressively measurable process $\left(\lambda_{t}\right)_{t \in[0, \infty)}$, given by

$$
\begin{equation*}
\lambda_{t}^{j}=\sum_{i=1}^{k}\left(\sigma^{+}\right)_{t}^{j i} \mu_{t}^{i}, \quad t \geq 0, \text { a.s. } \tag{4.5}
\end{equation*}
$$

satisfies

$$
\begin{equation*}
\sum_{j=1}^{d} \sigma_{t}^{j i} \lambda_{t}^{j}=\mu_{t}^{i}, \quad i=1, \ldots, k, \quad t \geq 0, \text { a.s. } \tag{4.6}
\end{equation*}
$$

Remark 4.7. It is one of the main properties of the Moore-Penrose inverse that $\lambda_{t}$ is the smallest (in Euclidean norm) solution of the linear system $\sigma_{t} \lambda_{t}=\mu_{t}$. It would, therefore, be enough to ask in Assumption 4.6 that some progressively measurable solution to (4.6) be bounded uniformly on each segment $[0, t]$.

Theorem 4.8 (see Theorem 4 in [27]). Under conditions and notation of Assumption 4.6, let $\left(\delta_{t}\right)_{t \in[0, \infty)}$ and $\left(\phi_{t}\right)_{t \in[0, \infty)}$ be $k$-dimensional $\mathbb{F}$-progressive processes such that $\sigma_{t} \sigma_{t}^{+} \delta_{t}=\delta_{t}$ for all $t \geq 0$ a.s. Define two continuous (one-dimensional) processes $\left(Y_{t}\right)_{t \in[0, \infty)},\left(Z_{t}\right)_{t \in[0, \infty)}$ by

$$
\begin{equation*}
d Y_{t}=Y_{t} \delta_{t}\left(\lambda_{t} d t+d W_{t}\right), \quad Y_{0}=1 / \gamma \tag{4.7}
\end{equation*}
$$

and

$$
\begin{equation*}
d Z_{t}=Z_{t} \phi_{t} d W_{t}, \quad Z_{0}=1 \tag{4.8}
\end{equation*}
$$

where we assume that $\delta$ and $\phi$ are regular enough for the integrals in (4.7) and (4.8) to be defined. Moreover, we assume that $Z$ is a positive martingale and that, when restricted to any finite interval $[0, t], Y$ is uniformly bounded from above and away from zero.

With the process $\left(A_{t}\right)_{t \in[0, \infty)}$ defined as

$$
\begin{equation*}
A_{t}=\int_{0}^{t}\left\|\sigma_{s} \sigma_{s}^{+}\left(\lambda_{s}+\phi_{s}\right)-\delta_{s}\right\|^{2} d s \tag{4.9}
\end{equation*}
$$

the random field $U$, given by

$$
\begin{equation*}
U_{t}(x ; \omega)=-Z_{t} \exp \left(-\frac{x}{Y_{t}}+\frac{A_{t}}{2}\right) \tag{4.10}
\end{equation*}
$$

is a forward exponential performance. In particular, for $0 \leq s \leq t$ and $\xi \in \mathbb{L}^{\infty}\left(\mathcal{F}_{s}\right)$, we have

$$
\begin{equation*}
U_{s}(\xi)=\underset{\pi \in \mathcal{A}_{\mathrm{bd}}}{\operatorname{esssup}} \mathbb{E}\left[U_{t}\left(\xi+\int_{s}^{t} \pi_{u} d S_{u}\right) \mid \mathcal{F}_{s}\right] \text { a.s. } \tag{4.11}
\end{equation*}
$$

Remark 4.9. In (4.10), one can give a natural financial interpretation to the processes $Y$ (which normalizes the wealth argument) and $Z$ (which appears as a multiplicative factor). One might think of $Y$ as a benchmark (or a numéraire) in relation to which we wish to measure the performance of our investment strategies. The values of the process $Z$, on the other hand, can be thought of as Radon-Nikodym derivatives of the investor's subjective probability measure with respect to the measure $\mathbb{P}$.
4.2. Forward entropic risk measures. We are now ready to introduce the forward entropic risk measures (FERMs). We start with an auxiliary object, denoted by $\rho(C ; t)$.

Definition 4.10. Let $U$ be the forward exponential performance defined in (4.10), and let $t \geq 0$ be arbitrary but fixed. For a contingent claim written at time $s=0$ and yielding a payoff $C \in \mathbb{L}^{\infty}\left(\mathcal{F}_{t}\right)$, we define $\rho(C ; t) \in \mathbb{R}$ as the unique solution of

$$
\begin{equation*}
\sup _{\pi \in \mathcal{A}_{\mathrm{bd}}} \mathbb{E}\left[U_{t}\left(x+\int_{0}^{t} \pi_{s} d S_{s}\right)\right]=\sup _{\pi \in \mathcal{A}_{\mathrm{bd}}} \mathbb{E}\left[U_{t}\left(x+\rho(C ; t)+C+\int_{0}^{t} \pi_{s} d S_{s}\right)\right] \quad \forall x \in \mathbb{R} \tag{4.12}
\end{equation*}
$$

The mapping $\rho(\cdot ; t): \mathbb{L}^{\infty}\left(\mathcal{F}_{t}\right) \rightarrow \mathbb{R}$ is called the $t$-normalized forward entropic measure.
One can readily check that (4.12) indeed admits a unique solution (independent of the initial wealth $x$ ), so that the $t$-normalized forward entropic measures are well defined. The reader can convince him-/herself of the validity of the following result.

Proposition 4.11. The t-normalized FERMs are replication-invariant convex risk measures on $\mathbb{L}^{\infty}\left(\mathcal{F}_{t}\right)$ for each $t \geq 0$.

The fundamental property in which FERMs differ from a generic replication-invariant risk measure (see examples in subsection 3.3.2) is the following.

Proposition 4.12. For $0 \leq s<t<\infty$, and $C^{(s)} \in \mathbb{L}^{\infty}\left(\mathcal{F}_{s}\right)$, consider the $s$ - and $t$ normalized forward entropic measures $\rho\left(C^{(s)} ; s\right)$ and $\rho\left(C^{(s)} ; t\right)$ applied to the contingent claim $C^{(s)}$. Then,

$$
\begin{equation*}
\rho\left(C^{(s)} ; s\right)=\rho\left(C^{(s)} ; t\right) \tag{4.13}
\end{equation*}
$$

More generally, for $C^{(r)} \in \mathbb{L}^{\infty}\left(\mathcal{F}_{r}\right)$, where $0 \leq r<s<t<\infty$, we have

$$
\begin{equation*}
\rho\left(C^{(r)} ; s\right)=\rho\left(C^{(r)} ; t\right) \tag{4.14}
\end{equation*}
$$

Proof. We are going to only establish (4.13) since (4.14) follows from similar arguments. To this end, note that a self-financing policy $\pi \in \mathcal{A}_{\mathrm{bd}}$ if and only if $\pi \mathbf{1}_{[0, t]} \in \mathcal{A}_{\mathrm{bd}}$ and $\pi \mathbf{1}_{(t, \infty)} \in$ $\mathcal{A}_{\mathrm{bd}}$. Using Definition 4.10 at $x=0$, we obtain

$$
\begin{aligned}
U_{0}(0) & =\sup _{\pi \in \mathcal{A}_{\mathrm{bd}}} \mathbb{E}\left[U_{t}\left(\rho\left(C^{(s)} ; t\right)+C^{(s)}+\int_{0}^{t} \pi_{u} d S_{u}\right)\right] \\
& =\sup _{\pi, \pi^{\prime} \in \mathcal{A}_{\mathrm{bd}}} \mathbb{E}\left[\mathbb{E}\left[U_{t}\left(\rho\left(C^{(s)} ; t\right)+C^{(s)}+\int_{0}^{s} \pi_{u} d S_{u}+\int_{s}^{t} \pi_{u}^{\prime} d S_{u}\right) \mid \mathcal{F}_{s}\right]\right] \\
& =\sup _{\pi \in \mathcal{A}_{\mathrm{bd}}} \mathbb{E}\left[\operatorname{esssup}_{\pi^{\prime} \in \mathcal{A}_{\mathrm{bd}}} \mathbb{E}\left[U_{t}\left(\rho\left(C^{(s)} ; t\right)+C^{(s)}+\int_{0}^{s} \pi_{u} d S_{u}+\int_{s}^{t} \pi_{u}^{\prime} d S_{u}\right) \mid \mathcal{F}_{s}\right]\right] \\
& =\sup _{\pi \in \mathcal{A}_{\mathrm{bd}}} \mathbb{E}\left[U_{s}\left(\rho\left(C^{(s)} ; t\right)+C^{(s)}+\int_{0}^{s} \pi_{u} d S_{u}\right)\right]
\end{aligned}
$$

where we used the semigroup property (4.11) of $U$ and the fact that the random variable $\rho\left(C^{(s)} ; t\right)+C^{(s)}+\int_{0}^{s} \pi_{u} d S_{u}$ is an element of $\mathbb{L}^{\infty}\left(\mathcal{F}_{s}\right)$ for all $\pi \in \mathcal{A}_{\mathrm{bd}}$. We compare the obtained expression with the defining equation (4.12) to conclude that $\rho\left(C^{(s)} ; t\right)=\rho\left(C^{(s)} ; s\right)$.

We are now ready to define the FERMs.
Definition 4.13. For $C \in \mathbb{L}$, define the earliest maturity $t_{C} \in[0, \infty)$ of $C$ as

$$
\begin{equation*}
t_{C}=\inf \left\{t \geq 0: C \in \mathcal{F}_{t}\right\} . \tag{4.15}
\end{equation*}
$$

The $\operatorname{FERM} \nu: \mathbb{L} \rightarrow \mathbb{R}$ is defined as

$$
\begin{equation*}
\rho(C)=\rho\left(C ; t_{C}\right), \tag{4.16}
\end{equation*}
$$

where $\rho\left(C ; t_{C}\right)$ is the value of the $t_{C}$-normalized $F E R M$, defined in (4.12), applied to the contingent claim $C$.

The focal point of the present section is the following theorem.
Theorem 4.14. The mapping $\rho: \mathbb{L} \rightarrow \mathbb{R}$ is a maturity-independent risk measure.
Proof. We need to verify axioms (1)-(4) of Definition 3.1. Axioms (1) and (3) follow directly from elementary properties of the $t$-normalized forward risk measures. To show axiom (2) we take $\lambda \in(0,1)$ and $C_{1}, C_{2} \in \mathbb{L}$. Then, since $\lambda C_{1}+(1-\lambda) C_{2} \in \mathcal{F}_{\max \left(t_{C_{1}}, t_{C_{2}}\right)}$, we have $\max \left(t_{C_{1}}, t_{C_{2}}\right) \geq t_{\lambda C_{1}+(1-\lambda) C_{2}}$. Therefore,

$$
\begin{aligned}
\rho\left(\lambda C_{1}+(1-\lambda) C_{2}\right) & =\rho\left(\lambda C_{1}+(1-\lambda) C_{2} ; t_{\lambda C_{1}+(1-\lambda) C_{2}}\right) \\
& =\rho\left(\lambda C_{1}+(1-\lambda) C_{2} ; \max \left(t_{C_{1}}, t_{C_{2}}\right)\right),
\end{aligned}
$$

where we used (4.13). Using property (4.13) and the fact that the $t$-FERMs are convex risk measures, we get

$$
\begin{aligned}
\rho\left(\lambda C_{1}+(1-\lambda) C_{2}\right) & \leq \lambda \rho\left(C_{1} ; \max \left(t_{C_{1}}, t_{C_{2}}\right)\right)+(1-\lambda) \rho\left(C_{2} ; \max \left(t_{C_{1}}, t_{C_{2}}\right)\right) \\
& =\lambda \rho\left(C_{1} ; t_{C_{1}}\right)+(1-\lambda) \rho\left(C_{2} ; t_{C_{2}}\right) \\
& =\lambda \rho\left(C_{1}\right)+(1-\lambda) \rho\left(C_{2}\right) .
\end{aligned}
$$

It remains to check the replication and maturity independence axiom (4). To this end, we let $\xi=\int_{0}^{\infty} \pi_{u} d S_{u}$ for some portfolio process $\pi \in \mathcal{A}_{\mathrm{bd}}$. We need to show that

$$
\rho(C+\xi)=\rho(C)
$$

for any $C \in \mathbb{L}$. Observe that $\max \left(t_{C}, t_{\xi}\right) \geq t_{C+\xi}$, and, therefore, by (4.13) and (4.16), we have

$$
\rho(C+\xi)=\rho\left(C+\xi ; t_{C+\xi}\right)=\rho\left(C+\xi ; \max \left(t_{C}, t_{\xi}\right)\right)
$$

On the other hand, Proposition 4.11, the form of $\xi$, and (4.13) yield

$$
\rho\left(C+\xi ; \max \left(t_{C}, t_{\xi}\right)\right)=\rho\left(C ; \max \left(t_{C}, t_{\xi}\right)\right)=\rho\left(C ; t_{C}\right)=\rho(C),
$$

establishing axiom (4).
Next, we provide an explicit representation of the FERMs.
Theorem 4.15. Let $Y, Z, A$, and $U_{t}(\cdot)$ be as in Theorem 4.8. For $C \in \mathbb{L}$, its $F E R M$ is given by

$$
\begin{equation*}
\rho(C)=\inf _{\pi \in \mathcal{A}_{\mathrm{bd}}}\left(\frac{1}{\gamma} \ln \mathbb{E}\left[-U_{t}\left(C+\int_{0}^{t} \pi_{s} d S_{s}\right)\right]\right) \quad \text { for any } t \geq t_{C} \tag{4.17}
\end{equation*}
$$

where $t_{C}$ is defined in (4.15).
Proof. Equation (4.12) (with $x=-\rho(C ; t)$ and $t \geq t_{C}$ ) and the property (4.11) of the random field $U$ yield that

$$
\begin{equation*}
-\exp (\gamma \rho(C))=\sup _{\pi \in \mathcal{A}_{\mathrm{bd}}} \mathbb{E}\left[U_{t}\left(C+\int_{0}^{t} \pi_{s} d S_{s}\right)\right] \quad \text { for any } t \geq t_{C} \tag{4.18}
\end{equation*}
$$

By (4.11), the right-hand side of (4.18) is independent of $t$ for $t \geq t_{C}$.
4.3. Relationship with dynamic risk measures. Before we present concrete examples of maturity-independent risk measures in section 5 , let us briefly discuss their relationship with the dynamic risk measures (see the introduction for references). A family of mappings $\rho_{s}(\cdot ; t): \mathbb{L}^{\infty}\left(\mathcal{F}_{t}\right) \rightarrow \mathbb{L}^{\infty}\left(\mathcal{F}_{s}\right)$, where $0 \leq s \leq t \leq T$, with $T \in[0, \infty]$, is said to be a dynamic (time-consistent) risk measure if each $\rho_{s}(\cdot ; t)$ satisfies the analogues of the axioms of convex risk measures and the semigroup property

$$
\rho_{s}\left(-\rho_{t}(f ; u) ; t\right)=\rho_{s}(f ; u), \quad 0 \leq s \leq t \leq u \leq T,
$$

holds. Using a version of Definition 4.13 and Theorem 4.14, the reader can readily check that each replication-invariant dynamic risk measure defined on the whole positive semiaxis
$[0, \infty)$ (i.e., when $T=\infty$ ) gives rise to a maturity-independent risk measure. Under certain conditions, the reverse construction can be carried out as well (details will be presented in [33]).

The philosophies of the two approaches are quite different, though. Perhaps the best way to illustrate this point is through the analogy with the expected utility theory. Dynamic risk measures correspond to the traditional utility framework where a system of decisions relating various maturity dates is interlaced together through a consistency criterion. The maturity-independent risk measures take the opposite point of view and correspond to forward performances. While the dynamic risk measures are natural in the case $T<\infty$, the maturityindependent risk measures fit well with infinite or unprespecified maturities.
5. Examples. In this section, we provide two representative classes of FERMs. For the first one, we adopt the setting and notation of Assumption 4.6 and single out some of the special cases obtained when specific choices for the processes $Z$ and $Y$ (of Theorem 4.8) are used in conjunction with Definition 4.13 of the FERMs. Then, we illustrate the versatility of the general notion of maturity-independent risk measures by constructing an example in an incomplete binomial-type model which can be seen as a special case of the locally bounded semimartingale setup of section 2. Some background and technical details pertaining to this example can be found in [24].
5.1. Itô-process-driven markets. This example is set in a financial market described in Assumption 4.6, with $k=1$ (one risky asset) and $d=2$ (two driving Brownian motions). Without loss of generality, we assume that $\sigma_{t}^{12} \equiv 0$, and $\sigma_{t}=\sigma_{t}^{11}>0$, i.e., that the second Brownian motion does not drive the tradeable asset. In this case, we have $\lambda_{t}=\left(\lambda_{t}^{1}, \lambda_{t}^{2}\right)$, where $\lambda_{t}^{1}=\mu_{t} / \sigma_{t}$ and $\lambda_{t}^{2}=0$. Therefore, the stock-price process satisfies

$$
d S_{t}=S_{t}\left(\mu_{t} d t+\sigma_{t} d W_{t}^{1}\right)
$$

on an augmented filtration generated by a two-dimensional Brownian motion $\left(W^{1}, W^{2}\right)$. The processes $Z, Y, A$ from Theorem 4.8 can be written as

$$
\begin{equation*}
d Y_{t}=Y_{t} \delta_{t}\left(\lambda_{t}^{1} d t+d W_{t}^{1}\right), \quad Y_{0}=1 / \gamma>0, \quad d Z_{t}=Z_{t} \phi_{t} d W_{t}^{1}, \quad Z_{0}=1 \tag{5.1}
\end{equation*}
$$

and

$$
\begin{equation*}
A_{t}=\int_{0}^{t}\left(\lambda_{s}^{1}+\phi_{s}-\delta_{s}\right)^{2} d s, \quad A_{0}=0 \tag{5.2}
\end{equation*}
$$

subject to a choice of two processes $\phi$ and $\delta$, under the regularity conditions stated in Theorem 4.8.
(a) $\phi \equiv \delta \equiv 0$. In this case, $Z_{t} \equiv 1, Y_{t} \equiv 1 / \gamma, A_{t} \equiv \int_{0}^{t}\left(\lambda_{s}^{1}\right)^{2} d s$, and the random field $U$ of (4.10) becomes

$$
U_{t}(x)=-\exp \left(-\gamma x+\frac{A_{t}}{2}\right)
$$

Using the indifference-pricing equation (4.12) and the self-generation property (4.11) of $U_{t}$, we deduce that for $C \in \mathbb{L}$ the value $\rho(C)$ satisfies

$$
-\exp (\gamma \rho(C))=\sup _{\pi \in \mathcal{A}_{\mathrm{bd}}} \mathbb{E}\left[-\exp \left(-\gamma\left(C+\int_{0}^{t} \pi_{s} d S_{s}\right)+\frac{A_{t}}{2}\right)\right] \quad \text { for any } t \geq t_{C}
$$

On the other hand, the classical (exponential) indifference price, $\nu\left(C-\frac{A_{t}}{2 \gamma} ; t\right.$ ), of the contingent claim $C-\frac{A_{t}}{2 \gamma}$, maturing at time $t$, satisfies

$$
\begin{aligned}
& \sup _{\pi \in \mathcal{A}_{\mathrm{bd}}} \mathbb{E}\left[-\exp \left(-\gamma\left(\nu\left(C-\frac{A_{t}}{2 \gamma} ; t\right)+\int_{0}^{t} \pi_{s} d S_{s}\right)\right)\right] \\
= & \sup _{\pi \in \mathcal{A}_{\mathrm{bd}}} \mathbb{E}\left[-\exp \left(-\gamma\left(C-\frac{A_{t}}{2 \gamma}+\int_{0}^{t} \pi_{s} d S_{s}\right)\right)\right] .
\end{aligned}
$$

With $H_{t}=\ln \sup _{\pi \in \mathcal{A}_{\mathrm{bd}}} \mathbb{E}\left[-\exp \left(-\gamma \int_{0}^{t} \pi_{s} d S_{u}\right)\right]$ (which will be recognized by the reader familiar with exponential utility maximization as the aggregate relative entropy), we now have

$$
\begin{equation*}
\rho(C)=-\nu\left(C-\frac{A_{t}}{2 \gamma} ; t\right)-\frac{1}{\gamma} H_{t} \quad \text { for any } t \geq t_{C} . \tag{5.3}
\end{equation*}
$$

(b) $\delta \equiv 0$. Then $Y_{t} \equiv 1 / \gamma, A_{t} \equiv \int_{0}^{t}\left(\lambda_{s}^{1}+\phi_{s}\right)^{2} d s$, and the random field $U$ of (4.10) takes the form

$$
U_{t}(x)=-Z_{t} \exp \left(-\gamma x+\frac{A_{t}}{2}\right)
$$

The risk measure $\rho(C)$ can be represented as in (5.3), with one important difference. Specifically, the (physical) probability measure $\mathbb{P}$ has to be replaced by the probability $\tilde{\mathbb{P}}$ whose Radon-Nikodym derivative with respect to $\mathbb{P}$ is given by $Z_{t}$ on $\mathcal{F}_{t}$ for any $t \geq 0$.

We leave the discussion of further examples in this setting - in particular for the case $\delta \neq 0$ - for the upcoming work of the second author [33].
5.2. The binomial case. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which two sequences $\left\{\xi_{t}\right\}_{t \in \mathbb{N}}$ and $\left\{\eta_{t}\right\}_{t \in \mathbb{N}}$ of random variables are defined. The stochastic processes $\left\{S_{t}\right\}_{t \in \mathbb{N}_{0}}$ and $\left\{Y_{t}\right\}_{t \in \mathbb{N}_{0}}$ are defined, in turn, as follows:

$$
S_{t}=\prod_{k=1}^{t} \xi_{k}, \quad Y_{t}=\prod_{k=1}^{t} \eta_{k}, \quad t \in \mathbb{N}, \quad S_{0}=Y_{0}=1
$$

The process $S$ models the evolution of a (traded) risky asset, and $Y$ is a (nontraded) factor. We assume, for simplicity, that the agents are allowed to invest in a zero-interest riskless bond $S^{0} \equiv 1$. The following two filtrations are naturally defined on $(\Omega, \mathcal{F}, \mathbb{P})$ :

$$
\begin{aligned}
\mathcal{F}_{t}^{S} & =\sigma\left(S_{0}, S_{1}, \ldots, S_{t}\right)=\sigma\left(\xi_{1}, \ldots, \xi_{t}\right), \quad t \in \mathbb{N}_{0}, \text { and } \\
\mathcal{F}_{t} & =\sigma\left(S_{0}, Y_{0}, S_{1}, Y_{1}, \ldots, S_{t}, Y_{t}\right)=\sigma\left(\xi_{1}, \ldots, \xi_{t}, \eta_{1}, \ldots, \eta_{t}\right), \quad t \in \mathbb{N}_{0}
\end{aligned}
$$

We assume that for each $t \in \mathbb{N}$ there exist $\xi_{t}^{u}, \xi_{t}^{d}, \eta_{t}^{u}, \eta_{t}^{d} \in \mathbb{R}$ with $0<\xi_{t}^{d}<1<\xi_{t}^{u}$ and $0<\eta_{t}^{d}<$ $\eta_{t}^{u}$ such that $\mathbb{P}\left[\xi_{t}=\xi_{t}^{u} \mid \mathcal{F}_{t-1}\right]=1-\mathbb{P}\left[\xi_{t}=\xi_{t}^{d} \mid \mathcal{F}_{t-1}\right]>0$ a.s., and $\mathbb{P}\left[\eta_{t}=\eta_{t}^{u}\right]=1-\mathbb{P}\left[\eta_{t}=\eta_{t}^{d}\right]$.

The agent starts with initial wealth $x \in \mathbb{R}$ and trades in the market by holding $\alpha_{t+1}$ shares of the asset $S$ in the interval $(t, t+1], t \in \mathbb{N}_{0}$, financing his/her purchases by borrowing (or lending to) the risk-free bond $S^{0}$. Therefore, the wealth process $\left\{X_{t}\right\}_{t \in \mathbb{N}_{0}}$ is given by

$$
X_{t}=x+\sum_{k=0}^{t-1} \alpha_{k+1}\left(S_{k+1}-S_{k}\right), \quad t \in \mathbb{N},
$$

with $X_{0}=x$. It can be shown that, for each $t \in \mathbb{N}$, there exists a unique minimal martingale measure $\mathbb{Q}^{(t)}$ on $\mathcal{F}_{t}$ (see [24] for details).

Define the $\mathcal{F}_{t}$-predictable ( $\mathcal{F}_{t-1}$-adapted) process $\left\{h_{t}\right\}_{t \in \mathbb{N}}$ given by

$$
h_{t}=q_{t} \ln \left(\frac{q_{t}}{\mathbb{P}\left[A_{t} \mid \mathcal{F}_{t-1}\right]}\right)+\left(1-q_{t}\right) \ln \left(\frac{1-q_{t}}{1-\mathbb{P}\left[A_{t} \mid \mathcal{F}_{t-1}\right]}\right), \quad t \in \mathbb{N}_{0}
$$

with

$$
A_{t}=\left\{\omega: \xi_{t}(\omega)=\xi_{t}^{u}\right\} \quad \text { and } \quad q_{t}=\frac{1-\xi_{t}^{d}}{\xi_{t}^{u}-\xi_{t}^{d}}=\mathbb{Q}^{(t)}\left[A_{t} \mid \mathcal{F}_{t-1}\right] .
$$

In [24] (see also [25]) it is shown that the random field $U: \Omega \times \mathbb{N}_{0} \times \mathbb{R} \rightarrow \mathbb{R}$ defined by

$$
U_{t}(x)=-\exp \left(-x+\sum_{k=1}^{t} h_{k}\right)
$$

is a forward exponential performance. We also consider the inverse $U^{-1}$ of $U$ given by

$$
U_{t}^{-1}(y)=-\ln (-y)-\sum_{k=1}^{t} h_{k}
$$

for $y \in(-\infty, 0)$ and $\left\{h_{t}\right\}_{t \in \mathbb{N}_{0}}$ as above.
For $t \in \mathbb{N}_{0}$, we define the (single-period) iterative forward price functional $\mathcal{E}^{(t, t+1)}$ : $\mathbb{L}^{\infty}\left(\mathcal{F}_{t+1}\right) \rightarrow \mathbb{L}^{\infty}\left(\mathcal{F}_{t}\right)$, given by

$$
\mathcal{E}^{(t, t+1)}(C)=\mathbb{E}_{\mathbb{Q}^{(t+1)}}\left[-U_{t+1}^{-1}\left(\mathbb{E}_{\mathbb{Q}^{(t+1)}}\left[U_{t+1}(-C) \mid \mathcal{F}_{t} \vee \mathcal{F}_{t+1}^{S}\right]\right) \mid \mathcal{F}_{t}\right]
$$

for any $C \in \mathbb{L}^{\infty}\left(\mathcal{F}_{t+1}\right)$. Similarly, for $t<t^{\prime}$ and $C \in \mathbb{L}^{\infty}\left(\mathcal{F}_{t^{\prime}}\right)$ we define the (multistep) forward pricing functional $\mathcal{E}^{\left(t, t^{\prime}\right)}: \mathbb{L}^{\infty}\left(\mathcal{F}_{t^{\prime}}\right) \rightarrow \mathbb{L}^{\infty}\left(\mathcal{F}_{t}\right)$ by

$$
\mathcal{E}^{\left(t, t^{\prime}\right)}(C)=\mathcal{E}^{(t, t+1)}\left(\mathcal{E}^{(t+1, t+2)}\left(\ldots\left(\mathcal{E}^{\left(t^{\prime}-1, t^{\prime}\right)}(C)\right)\right)\right) .
$$

Proposition 5.1. Let $\rho(\cdot ; t): \mathbb{L}^{\infty}\left(\mathcal{F}_{t}\right) \rightarrow \mathbb{R}$ be defined by

$$
\rho(C ; t)=\mathcal{E}^{(0, t)}(C) .
$$

Then, the mapping $\rho: \mathcal{L}=\bigcup_{t \in \mathbb{N}_{0}} \mathbb{L}^{\infty}\left(\mathcal{F}_{t}\right) \rightarrow \mathbb{R}$ defined by

$$
\rho(C)=\rho\left(C ; t_{C}\right)
$$

for $t_{C}=\inf \left\{t \geq 0: C \in \mathcal{F}_{t}\right\}$ is a maturity-independent convex risk measure.
The statement of the proposition follows from an argument analogous to the one in the proof of Proposition 4.12. For a detailed exposition of all steps, see [24].
6. Summary and future research. The goals of the present paper are twofold:

1. we want to bring forth and illustrate the concept of maturity-independent risk measures; and
2. we want to provide a class of such measures.

Two examples - one defined on a finite probability space and the other in an Itô-process setting - are given. Their analysis shows that, while plausible and simple from the decisiontheoretic point of view, the notion of maturity independence is nontrivial and reveals an interesting structure.

One of the major sources of appeal of the theory of maturity-independent risk measures is, in our opinion, the fact that it opens a venue for a wide variety of research opportunities both from the mathematical as well as the financial points of view. One of these directions, which we intend to pursue in a forthcoming work (see [33]), follows the link between maturity independence and forward performance processes in the direction opposite to the one explored here: while FERMs provide a wide class of examples of maturity-independent risk measures, it is natural to ask whether there are any others. In other words, we would like to give a full characterization of maturity-independent risk measures arising from performance random fields. Such a characterization would not only complete the outlined theory from the mathematical point of view, but it would also provide a firm decision-theoretic foundation for the sister theory of forward performance processes.

Acknowledgment. We thank the participants of the "Workshop on Financial Engineering and Actuarial Mathematics," University of Michigan, May 2007, the meeting on "Stochastic Analysis in Finance and Insurance," Oberwolfach, January 2008, and "AMS Sectional Meeting," Baton Rouge, March 2008, for engaging discussions and insightful comments.

## REFERENCES

[1] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, Coherent measures of risk, Math. Finance, 9 (1999), pp. 203-228.
[2] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, Thinking coherently, RISK, 10 (1997), pp. 68-71.
[3] P. Artzner, F. Delbaen, J. M. Eber, D. Heath, and H. Ku, Coherent Multiperiod Risk Measurement, preprint, 2003.
[4] F. P. Y. S. Berrier, L. C. G. Rogers, and M. R. Tehranchi, A Characterization of Forward Utility Functions, preprint, 2007.
[5] P. Cheridito, F. Delbaen, and M. Kupper, Coherent and convex risk measures for bounded càdlàg processes, Stochastic Process. Appl., 112 (2004), pp. 1-22.
[6] P. Cheridito, F. Delbaen, and M. Kupper, Coherent and convex risk measures for unbounded càdlàg processes, Finance Stoch., 9 (2005), pp. 369-387.
[7] P. Cheridito, F. Delbaen, and M. Kupper, Dynamic monetary risk measures for bounded discretetime processes, Electron. J. Probab., 11 (2006), pp. 57-106.
[8] T. Choulli, C. Stricker, and J. Li, Minimal Hellinger martingale measures of order q, Finance Stoch., 11 (2007), pp. 399-427.
[9] F. Delbaen and W. Schachermayer, A general version of the fundamental theorem of asset pricing, Math. Ann., 300 (1994), pp. 463-520.
[10] K. Detlefsen and G. Scandolo, Conditional and dynamic convex risk measures, Finance Stoch., 9 (2005), pp. 539-561.
[11] H. Föllmer and A. Schied, Convex measures of risk and trading constraints, Finance Stoch., 6 (2002), pp. 429-447.
[12] H. Föllmer and A. Schied, Robust preferences and convex measures of risk, in Advances in Finance and Stochastics, Springer, Berlin, 2002, pp. 39-56.
[13] H. Föllmer and A. Schied, Stochastic Finance. An Introduction in Discrete Time, de Gruyter Stud. Math. 27, Walter de Gruyter \& Co., Berlin, 2004.
[14] M. Frittelli and E. Rosazza Gianin, Putting order in risk measures, J. Banking Finance, 26 (2002), pp. 1473-1486.
[15] M. Frittelli and G. Scandolo, Risk measures and capital requirements for processes, Math. Finance, 16 (2006), pp. 589-612.
[16] D. Heath, Back to the future, Plenary lecture, First World Congress of the Bachelier Finance Society, Paris, 2000.
[17] D. Heath and E. Ku, Pareto equilibria with coherent measures of risk, Math. Finance, 14 (2004), pp. 163-172.
[18] V. Henderson and D. Hobson, Horizon-unbiased utility functions, Stochastic Process. Appl., 117 (2007), pp. 1621-1641.
[19] J. Jacod and A. N. Shiryaev, Limit Theorems for Stochastic Processes, 2nd ed., Grundlehren Math. Wiss. 288, Springer, Berlin, 2003.
[20] I. Karatzas, J. P. Lehoczky, S. E. Shreve, and G.-L. Xu, Optimality conditions for utility maximization in an incomplete market, in Analysis and Optimization of Systems (Antibes, 1990), Lecture Notes in Control and Inform. Sci. 144, Springer, Berlin, 1990, pp. 3-23.
[21] I. Karatzas and S. E. Shreve, Brownian Motion and Stochastic Calculus, 2nd ed., Grad. Texts in Math. 113, Springer, New York, 1991.
[22] I. Karatzas and G. Žitković, Optimal consumption from investment and random endowment in incomplete semimartingale markets, Ann. Probab., 31 (2003), pp. 1821-1858.
[23] D. Kramkov and W. Schachermayer, The asymptotic elasticity of utility functions and optimal investment in incomplete markets, Ann. Appl. Probab., 9 (1999), pp. 904-950.
[24] M. Musiela, E. Sokolova, and T. Zariphopoulou, Indifference Prices under Forward Investment Performance Criteria: The Case Study of the Binomial Model, preprint, 2007.
[25] M. Musiela and T. Zariphopoulou, Backward and Forward Utilities and the Associated Pricing Systems: The Case Study of the Binomial Model, preprint, 2003.
[26] M. Musiela and T. Zariphopoulou, Investment and valuation under backward and forward dynamic exponential utilities in a stochastic factor model, in Advances in Mathematical Finance, Birkhäuser, Boston, 2007, pp. 303-334.
[27] M. Musiela and T. Zariphopoulou, Optimal asset allocation under forward exponential criteria, in Markov Processes and Related Topics: A Festschrift for T. G. Kurtz, Institute for Mathematical Statistics, Beachwood, OH, 2008.
[28] M. Musiela and T. Zariphopoulou, Portfolio choice under dynamic investment performance criteria, Quant. Finance, 9 (2009), pp. 161-170.
[29] M. Musiela and T. Zariphopoulou, The single period binomial model, in Indifference Pricing, R. Carmona, ed., Princeton University Press, Princeton, NJ, 2009, pp. 3-43.
[30] F. Riedel, Dynamic coherent risk measures, Stochastic Process. Appl., 112 (2004), pp. 185-200.
[31] B. Roorda, J. M. Schumacher, and J. Engwerda, Coherent acceptability measures in multiperiod models, Math. Finance, 15 (2005), pp. 589-612.
[32] R. Sircar and T. Zariphopoulou, Bounds and asymptotic approximations for utility prices when volatility is random, SIAM J. Control Optim., 43 (2005), pp. 1328-1353.
[33] G. Žitković, A dual characterization of self-generation and exponential forward performances, Ann. Appl. Probab., 19 (2009), pp. 2176-2210.
[34] G. Žitković, Utility maximization with a stochastic clock and an unbounded random endowment, Ann. Appl. Probab., 15 (2005), pp. 748-777.

# Time Dependent Heston Model* 

E. Benhamou ${ }^{\dagger}$, E. Gobet ${ }^{\ddagger}$, and M. Miri ${ }^{\dagger}$


#### Abstract

The use of the Heston model is still challenging because it has a closed formula only when the parameters are constant [S. Heston, Rev. Financ. Stud., 6 (1993), pp. 327-343] or piecewise constant [S. Mikhailov and U. Nogel, Wilmott Magazine, July (2003), pp. 74-79]. Hence, using a small volatility of volatility expansion and Malliavin calculus techniques, we derive an accurate analytical formula for the price of vanilla options for any time dependent Heston model (the accuracy is less than a few bps for various strikes and maturities). In addition, we establish tight error estimates. The advantage of this approach over Fourier-based methods is its rapidity (gain by a factor 100 or more) while maintaining a competitive accuracy. From the approximative formula, we also derive some corollaries related first to equivalent Heston models (extending some work of Piterbarg on stochastic volatility models [V. Piterbarg, Risk Magazine, 18 (2005), pp. 71-75]) and second, to the calibration procedure in terms of ill-posed problems.


Key words. asymptotic expansion, Malliavin calculus, small volatility of volatility, time dependent Heston model

AMS subject classifications. $60 \mathrm{~J} 75,60 \mathrm{HXX}$
DOI. 10.1137/090753814

1. Introduction. Stochastic volatility modeling emerged in the late nineties as a way to manage the smile. In this work, we focus on the Heston model, which is a lognormal model where the square of volatility follows a Cox-Ingersoll-Ross (CIR) ${ }^{1}$ process. The call (and put) price has a closed formula in this model thanks to a Fourier inversion of the characteristic function (see Heston [22], Lewis [27], and Lipton [29]). When the parameters are piecewise constant, one can still derive a recursive closed formula using a PDE method (see Mikhailov and Nogel [31]) or a Markov argument in combination with affine models (see Elices [16]), but formula evaluation becomes increasingly time consuming. However, for general time dependent parameters there is no analytical formula and one usually has to perform Monte Carlo simulations. This explains the interest of recent works for designing more efficient Monte Carlo simulations: see Broadie and Kaya [13] for an exact simulation and bias-free scheme based on Fourier integral inversion; see Andersen [4] based on a Gaussian moment matching method and a user friendly algorithm; see Smith [40] relying on an almost exact scheme;

[^21]see Alfonsi [2] using higher order schemes and a recursive method for the CIR process. For numerical PDEs, we refer the reader to Kluge's doctoral dissertation [25].

Comparison with the literature. A more recent trend in the quantitative literature has been the use of the so-called approximation method to derive analytical formulas. This has led to an impressive number of papers, with many original ideas. For instance, Alòs, León, and Vives [3] have been studying the short time behavior of implied volatility for stochastic volatility using an extension of Itô's formula. Another trend has focused on analytical techniques to derive the asymptotic expansion of the implied volatility near expiry (see, for instance, Berestycki, Busca, and Florent [10], Labordère [26], Hagan et al. [21], Lewis [28], Osajima [34], or Forde [17]). But in these works the implied volatility near expiry does not have a closed formula because the related geodesic distance is not explicit. It can, however, be approximated by a series expansion [28]. The drawback to these methods is their inability to handle nonhomogeneous (that is, time dependent) parameters. For long maturities, another approach has been the asymptotic expansion w.r.t. the mean reversion parameter of the volatility as shown in [18]. In the case of zero correlation, averaging techniques as exposed in [36] and [35] can be used. Antonelli and Scarlatti take another view in [5] and have suggested price expansion w.r.t. correlation. For all of these techniques, the domain of availability of the expansion is restricted to short or long maturities, to zero correlation, or to homogeneous parameters. However, in [19] regular and singular expansions are generalized to the case of time dependent parameters (with volatilities of Ornstein-Uhlenbeck type). The PDE approach used in [19] could also be used in the present situation and would lead formally to the same expansion we derive. In our work, we aim to give an analytical formula which covers both short and long maturities that also handles time inhomogeneous parameters as well as nonnull correlations. As a difference with several previously quoted papers, our purpose consists also of justifying our approximation mathematically.

The results closest to ours are probably those based on an expansion w.r.t. the volatility of volatility by Lewis [27]: it is based on formal analytical arguments and is restricted to constant parameters. Our formula can be viewed as an extension of Lewis' formula in order to address a time dependent Heston model, using a direct probabilistic approach. In addition, we prove an error estimate which shows that our approximation formula for call/put is of order 2 w.r.t. the volatility of volatility. The advantage of this current approximation is that the evaluation is about 100 to 1000 times quicker than a Fourier-based method (see our numerical tests).

Comparison with our previous works [8] and [7]. Our approach here consists of expanding the price w.r.t. the volatility of volatility and of computing the correction terms using Malliavin calculus. In these respects, the current approach is similar to our previous works [8] and [7]; however, the techniques for estimating error are different. Indeed, we use the fact that the price of vanilla options can be expressed as an expectation of a smooth price function for stochastic volatility models. This is based on a conditioning argument as in [38]. Consequently, the smoothness hypotheses $\left(H_{1}, H_{2}, H_{3}\right)$ of our previous papers are no longer required. Note also that the square root function arising in the martingale part of the CIR process is not Lipschitz continuous. Hence, the Heston model does not fit the smoothness framework used previously. Therefore, to overcome this difficulty, we derive new technical results in order to prove the accuracy of the formula.

Contribution of the paper. We give an explicit analytical formula for the price of vanilla options in a time dependent Heston model. Our approach is based on an expansion w.r.t. a small volatility of volatility. This is practically justified by the fact that this parameter is usually quite small (of order 1 or less; see [6], [27], or [13], for instance). The resulting formula is the sum of two terms: the leading term is the Black-Scholes price for the model without volatility of volatility, while the correction term is a combination of Greeks of the leading term with explicit weights depending only on the model parameters. Proving the accuracy of the expansion is far from straightforward, but with some technicalities and a relevant analysis of error we succeed in giving tight error estimates. Our expansion enables us to obtain averaged parameters for the dynamic Heston model.

Formulation of the problem. We consider the solution of the stochastic differential equation (SDE)

$$
\begin{align*}
\mathrm{d} X_{t} & =\sqrt{v_{t}} d W_{t}-\frac{v_{t}}{2} \mathrm{~d} t, \quad X_{0}=x_{0},  \tag{1.1}\\
\mathrm{~d} v_{t} & =\kappa\left(\theta_{t}-v_{t}\right) \mathrm{d} t+\xi_{t} \sqrt{v_{t}} \mathrm{~d} B_{t}, \quad v_{0},  \tag{1.2}\\
\mathrm{~d}\langle W, B\rangle_{t} & =\rho_{t} \mathrm{~d} t,
\end{align*}
$$

where $\left(B_{t}, W_{t}\right)_{0 \leq t \leq T}$ is a two-dimensional correlated Brownian motion on a filtered probability space $\left(\Omega, \mathcal{F},\left(\mathcal{F}_{t}\right)_{0 \leq t \leq T}, \mathbb{P}\right)$ with the usual assumptions on filtration $\left(\mathcal{F}_{t}\right)_{0 \leq t \leq T}$. In our setting, $\left(X_{t}\right)_{t}$ is the log of the forward price and $\left(v_{t}\right)_{t}$ is the square of the volatility which follows a CIR process with an initial value $v_{0}>0$, a positive mean reversion $\kappa$, a positive long-term level $\left(\theta_{t}\right)_{t}$, a positive volatility of volatility $\left(\xi_{t}\right)_{t}$, and a correlation $\left(\rho_{t}\right)_{t}$. These time dependent parameters are assumed to be measurable and bounded on $[0, T]$.

To develop our approximation method, we will examine the following perturbed process w.r.t. $\epsilon \in[0,1]$ :

$$
\begin{align*}
\mathrm{d} X_{t}^{\epsilon} & =\sqrt{v_{t}^{\epsilon}} d W_{t}-\frac{v_{t}^{\epsilon}}{2} \mathrm{~d} t, \quad X_{0}^{\epsilon}=x_{0}, \\
\mathrm{~d} v_{t}^{\epsilon} & =\kappa\left(\theta_{t}-v_{t}^{\epsilon}\right) \mathrm{d} t+\epsilon \xi_{t} \sqrt{v_{t}^{\epsilon}} \mathrm{d} B_{t}, \quad v_{0}^{\epsilon}=v_{0}, \tag{1.3}
\end{align*}
$$

so that our perturbed process coincides with the initial one for $\epsilon=1: X_{t}^{1}=X_{t}, v_{t}^{1}=v_{t}$. For the existence of the solution $v^{\epsilon}$, we refer the reader to Chapter IX in [39] (moreover, the process is nonnegative for $k \theta_{t} \geq 0$; see also the proof of Lemma 4.2). Our main purpose is to give an accurate analytic approximation, in a certain sense, of the expected payoff of a put option:

$$
\begin{equation*}
g(\epsilon)=e^{-\int_{0}^{T} r_{t} \mathrm{~d} t} \mathbb{E}\left[\left(K-e^{\int_{0}^{T}\left(r_{t}-q_{t}\right) \mathrm{d} t+X_{T}^{\epsilon}}\right)_{+}\right], \tag{1.4}
\end{equation*}
$$

where $r$ (resp., $q$ ) is the risk-free rate (resp., the dividend yield), $T$ is the maturity, and $\epsilon=1$. Extensions to call options and other payoffs are discussed later.

Outline of the paper. In section 2, we explain the methodology of the small volatility of volatility expansion. An approximation formula is then derived in Theorem 2.2 and its accuracy stated in Theorem 2.3. This section ends by explicitly expressing the formula's coefficients for general time dependent parameters (constant, smooth, and piecewise constant).

Our expansion allows us to give equivalent constant parameters for the time dependent Heston model (see subsection 2.5). As a second corollary, the options calibration for Heston's model using only one maturity becomes an ill-posed problem; we give numerical results to confirm this situation. In section 3, we provide numerical tests to benchmark our formula with the closed formula in the case of constant and piecewise constant parameters. In section 4, we prove the accuracy of the approximation stated in Theorem 2.3: this section is the technical core of the paper. In section 5, we establish lemmas used to make the calculation of the correction terms explicit (those derived in Theorem 2.2). In section 6, we conclude this work and give a few extensions. In the appendix, we recall details about the closed formula (of Heston [22] and Lewis [27]) in the case of constant (and piecewise constant) parameters.

## 2. Smart Taylor expansion.

### 2.1. Notations.

Notation 2.1 (extremes of deterministic functions). For a càdlàg function $l:[0, T] \rightarrow \mathbb{R}$, we denote $l_{\text {Inf }}=\inf _{t \in[0, T]} l_{t}$ and $l_{\text {Sup }}=\sup _{t \in[0, T]} l_{t}$.

Notation 2.2 (differentiation).
(i) For a smooth function $x \mapsto l(x)$, we denote by $l^{(i)}(x)$ its ith derivative.
(ii) Given a fixed time $t$ and for a function $\epsilon \rightarrow f_{t}^{\epsilon}$, we denote (if it has a meaning) the ith derivative at $\epsilon=0$ by $f_{i, t}=\left.\frac{\partial^{i} f_{\epsilon}^{\epsilon}}{\partial \epsilon^{i}}\right|_{\epsilon=0}$.

### 2.2. About the CIR process.

Assumptions. In order to bound the approximation errors, we need a positivity assumption for the CIR process.

Assumption P. The parameters of the CIR process (1.2) verify the following conditions:

$$
\xi_{I n f}>0, \quad\left(\frac{2 \kappa \theta}{\xi^{2}}\right)_{I n f} \geq 1
$$

This assumption is crucial to ensure the positivity of the process on $[0, T]$, which is stated in detail in Lemma 4.2 (remember that $v_{0}>0$ ). We have

$$
\mathbb{P}\left(\forall t \in[0, T]: v_{t}>0\right)=1 .
$$

When the functions $\theta$ and $\xi$ are constant, Assumption (P) coincides with the usual Feller test condition $\frac{2 \kappa \theta}{\xi^{2}} \geq 1$ (see [24]).

Note that the above assumption ensures that the positivity property also holds for the perturbed CIR process (1.3): for any $\epsilon \in[0,1]$, we have

$$
\mathbb{P}\left(\forall t \in[0, T]: v_{t}^{\epsilon}>0\right)=1
$$

(see Lemma 4.2). We also need a uniform bound of the correlation in order to preserve the nondegeneracy of the $\mathrm{SDE}(1.1)$ conditionally on $\left(B_{t}\right)_{0 \leq t \leq T}$.

Assumption R . The correlation is bounded away from -1 and +1 :

$$
|\rho|_{\text {Sup }}<1 .
$$

2.3. Taylor development. In this section, we present the main steps leading to our results. Complete proofs are given later.

If $\left(\mathcal{F}_{t}^{B}\right)_{t}$ denotes the filtration generated by the Brownian motion $B$, the distribution of $X_{T}^{\epsilon}$ conditionally to $\mathcal{F}_{T}^{B}$ is a Gaussian distribution with mean $x_{0}+\int_{0}^{T} \rho_{t} \sqrt{v_{t}^{\epsilon}} \mathrm{d} B_{t}-\frac{1}{2} \int_{0}^{T} v_{t}^{\epsilon} \mathrm{d} t$ and variance $\int_{0}^{T}\left(1-\rho_{t}^{2}\right) v_{t}^{\epsilon} \mathrm{d} t(\epsilon \in[0,1])$. Therefore, the function (1.4) can be expressed as follows:

$$
\begin{equation*}
g(\epsilon)=\mathbb{E}\left[P_{B S}\left(x_{0}+\int_{0}^{T} \rho_{t} \sqrt{v_{t}^{\epsilon}} \mathrm{d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2} v_{t}^{\epsilon} \mathrm{d} t, \int_{0}^{T}\left(1-\rho_{t}^{2}\right) v_{t}^{\epsilon} \mathrm{d} t\right)\right] \tag{2.1}
\end{equation*}
$$

where the function $(x, y) \rightarrow P_{B S}(x, y)$ is the put function price in a Black-Scholes model with spot $e^{x}$, strike $K$, total variance $y$, risk-free rate $r_{e q}=\frac{\int_{0}^{T} r(t) \mathrm{d} t}{T}$, dividend yield $q_{e q}=\frac{\int_{0}^{T} q(t) \mathrm{d} t}{T}$, and maturity $T$. For the sake of completeness, we recall that $P_{B S}(x, y)$ has the following explicit expression:

$$
K e^{-r_{e q} T} \mathcal{N}\left(\frac{1}{\sqrt{y}} \log \left(\frac{K e^{-r_{e q} T}}{e^{x} e^{-q_{e q} T}}\right)+\frac{1}{2} \sqrt{y}\right)-e^{x} e^{-q_{e q} T} \mathcal{N}\left(\frac{1}{\sqrt{y}} \log \left(\frac{K e^{-r_{e q} T}}{e^{x} e^{-q_{e q} T}}\right)-\frac{1}{2} \sqrt{y}\right) .
$$

In the following, we expand $P_{B S}(.,$.$) w.r.t. its two arguments. For this, we note that P_{B S}$ is a smooth function (for $y>0$ ). In addition, there is a simple relation between its partial derivatives:

$$
\begin{equation*}
\frac{\partial P_{B S}}{\partial y}(x, y)=\frac{1}{2}\left(\frac{\partial^{2} P_{B S}}{\partial x^{2}}(x, y)-\frac{\partial P_{B S}}{\partial x}(x, y)\right) \quad \forall x \in \mathbb{R}, \forall y>0 \tag{2.2}
\end{equation*}
$$

which can be proved easily by a standard calculation left to the reader.
Under Assumption (P), for any $t, v_{t}^{\epsilon}$ is $C^{2}$ w.r.t $\epsilon$ at $\epsilon=0$ (differentiation in the $L_{p}$ sense). This result will be shown later. In addition, $v^{\epsilon}$ does not vanish (for any $\epsilon \in[0,1]$ ). Hence, by putting $v_{i, t}^{\epsilon}=\frac{\partial^{i} v_{t}^{\epsilon}}{\partial \epsilon^{2}}$, we get

$$
\begin{array}{ll}
\mathrm{d} v_{1, t}^{\epsilon}=-\kappa v_{1, t}^{\epsilon} \mathrm{d} t+\xi_{t} \sqrt{v_{t}^{\epsilon}} \mathrm{d} B_{t}+\epsilon \xi_{t} \frac{v_{1, t}^{\epsilon}}{2 \sqrt{v_{t}^{\epsilon}}} \mathrm{d} B_{t} & v_{1,0}^{\epsilon}=0 \\
\mathrm{~d} v_{2, t}^{\epsilon}=-\kappa v_{2, t}^{\epsilon} \mathrm{d} t+\xi_{t} \frac{v_{1, t}^{\epsilon}}{\sqrt{v_{t}^{\epsilon}}} \mathrm{d} B_{t}+\epsilon \xi_{t} \frac{v_{2, t}^{\epsilon}}{2 \sqrt{v_{t}^{\epsilon}}} \mathrm{d} B_{t}-\epsilon \xi_{t} \frac{\left[v_{1, t}^{\epsilon}\right]^{2}}{4\left[v_{t}^{\epsilon}\right]^{3 / 2}} \mathrm{~d} B_{t}, & v_{2,0}^{\epsilon}=0
\end{array}
$$

From the definitions $\left.v_{i, t} \equiv \frac{\partial^{i} v_{t}^{\epsilon}}{\partial \epsilon^{i}}\right|_{\epsilon=0}$, we easily deduce

$$
\begin{align*}
& v_{0, t}=e^{-\kappa t}\left(v_{0}+\int_{0}^{t} \kappa e^{\kappa s} \theta_{s} \mathrm{~d} s\right) \\
& v_{1, t}=e^{-\kappa t} \int_{0}^{t} e^{\kappa s} \xi_{s} \sqrt{v_{0, s}} \mathrm{~d} B_{s}  \tag{2.3}\\
& v_{2, t}=e^{-\kappa t} \int_{0}^{t} e^{\kappa s} \xi_{s} \frac{v_{1, s}}{\left(v_{0, s}\right)^{\frac{1}{2}}} \mathrm{~d} B_{s} \tag{2.4}
\end{align*}
$$

Note that $v_{0, t}$ coincides also with the expected variance $\mathbb{E}\left(v_{t}\right)$ because of the linearity of the drift coefficient of $\left(v_{t}\right)_{t}$. Now, to expand $g(\epsilon)$, we use the Taylor formula twice, first applied to $\epsilon \rightarrow v_{t}^{\epsilon}$ and $\sqrt{v_{t}^{\epsilon}}$ at $\epsilon=1$ using derivatives computed at $\epsilon=0$ :

$$
\begin{aligned}
v_{t}^{1} & =v_{0, t}+v_{1, t}+\frac{v_{2, t}}{2}+\cdots \\
\sqrt{v_{t}^{1}} & =\sqrt{v_{0, t}}+\frac{v_{1, t}}{2\left(v_{0, t}\right)^{\frac{1}{2}}}+\frac{v_{2, t}}{4\left(v_{0, t}\right)^{\frac{1}{2}}}-\frac{v_{1, t}^{2}}{8\left(v_{0, t}\right)^{\frac{3}{2}}}+\cdots
\end{aligned}
$$

second, it is applied to the smooth function $P_{B S}$ at the second order w.r.t. the first and second variables around $\left(x_{0}+\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2} v_{0, t} \mathrm{~d} t, \int_{0}^{T}\left(1-\rho_{t}^{2}\right) v_{0, t} \mathrm{~d} t\right)$. For convenience, we simply write

$$
\begin{align*}
\tilde{P}_{B S} & =P_{B S}\left(x_{0}+\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2} v_{0, t} \mathrm{~d} t, \int_{0}^{T}\left(1-\rho_{t}^{2}\right) v_{0, t} \mathrm{~d} t\right)  \tag{2.5}\\
\frac{\partial^{i+j} \tilde{P}_{B S}}{\partial x^{i} y^{j}} & =\frac{\partial^{i+j} P_{B S}}{\partial x^{i} y^{j}}\left(x_{0}+\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2} v_{0, t} \mathrm{~d} t, \int_{0}^{T}\left(1-\rho_{t}^{2}\right) v_{0, t} \mathrm{~d} t\right)
\end{align*}
$$

Then, one gets

$$
\begin{equation*}
+\frac{1}{2} \mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial x^{2}}\left(\int_{0}^{T} \rho_{t} \frac{v_{1, t}}{2\left(v_{0, t}\right)^{\frac{1}{2}}} \mathrm{~d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2} v_{1, t} \mathrm{~d} t\right)^{2}\right] \tag{2.9}
\end{equation*}
$$

$$
\begin{align*}
g(1)= & \mathbb{E}\left[\tilde{P}_{B S}\right]  \tag{2.6}\\
+ & {\left[\mathbb { E } \left[\frac { \partial \tilde { P } _ { B S } } { \partial x } \left(\int_{0}^{T} \rho_{t}\left(\frac{v_{1, t}}{2\left(v_{0, t}\right)^{\frac{1}{2}}}+\frac{v_{2, t}}{4\left(v_{0, t}\right)^{\frac{1}{2}}}-\frac{v_{1, t}^{2}}{8\left(v_{0, t}\right)^{\frac{3}{2}}}\right) \mathrm{d} B_{t}\right.\right.\right.}  \tag{2.7}\\
& \left.\left.-\int_{0}^{T} \frac{\rho_{t}^{2}}{2}\left(v_{1, t}+\frac{v_{2, t}}{2}\right) \mathrm{d} t\right)\right] \\
+\mathbb{E} & {\left[\frac{\partial \tilde{P}_{B S}}{\partial y} \int_{0}^{T}\left(1-\rho_{t}^{2}\right)\left(v_{1, t}+\frac{v_{2, t}}{2}\right) \mathrm{d} t\right] } \tag{2.8}
\end{align*}
$$

$$
\begin{equation*}
+\frac{1}{2} \mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial y^{2}}\left(\int_{0}^{T}\left(1-\rho_{t}^{2}\right) v_{1, t} \mathrm{~d} t\right)^{2}\right] \tag{2.10}
\end{equation*}
$$

$$
\begin{equation*}
+\mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial x y}\left(\int_{0}^{T}\left(1-\rho_{t}^{2}\right) v_{1, t} \mathrm{~d} t\right)\left(\int_{0}^{T} \rho_{t} \frac{v_{1, t}}{2\left(v_{0, t}\right)^{\frac{1}{2}}} \mathrm{~d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2} v_{1, t} \mathrm{~d} t\right)\right] \tag{2.11}
\end{equation*}
$$

$$
\begin{equation*}
+\mathcal{E} \tag{2.12}
\end{equation*}
$$

where $\mathcal{E}$ is the error in our Taylor expansion. In fact, we notice that

$$
\begin{aligned}
\mathbb{E}\left[\tilde{P}_{B S}\right] & =\mathbb{E}\left[\mathbb{E}\left[\left.e^{-\int_{0}^{T} r_{t} \mathrm{~d} t}\left(K-e^{x_{0}+\int_{0}^{T}\left(r_{t}-q_{t}-\frac{v_{0, t}}{2}\right) \mathrm{d} t+\int_{0}^{T} \sqrt{v_{0, t}}\left(\rho_{t} \mathrm{~d} B_{t}+\sqrt{1-\rho_{t}^{2}} \mathrm{~d} B_{t}^{\perp}\right)}\right)_{+} \right\rvert\, \mathcal{F}_{T}^{B}\right]\right] \\
& =P_{B S}\left(x_{0}, \int_{0}^{T} v_{0, t} \mathrm{~d} t\right),
\end{aligned}
$$

where $B^{\perp}$ is a Brownian motion independent on $\mathcal{F}_{T}^{B}$. Furthermore, the relation (2.2) remains the same for $\tilde{P}_{B S}$, and this enables us to simplify the expansion above. This gives the following proposition.

Proposition 2.1. The approximation (2.12) is equivalent to
$g(1)=P_{B S}\left(x_{0}, \int_{0}^{T} v_{0, t} \mathrm{~d} t\right)+\mathbb{E}\left[\frac{\partial \tilde{P}_{B S}}{\partial y} \int_{0}^{T}\left(v_{1, t}+v_{2, t}\right) \mathrm{d} t\right]+\frac{1}{2} \mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial y^{2}}\left(\int_{0}^{T} v_{1, t} \mathrm{~d} t\right)^{2}\right]+\mathcal{E}$.
The details of the proof are given in subsection 5.3. At first sight, the above formula looks like a Taylor formula of $P_{B S}$ w.r.t. the cumulated variance. In fact, it is different; note that the coefficient of $v_{2, t}$ is not $1 / 2$ but 1 . We do not have any direct interpretation of this formula.

The next step consists of making explicit the correction terms as a combination of Greeks of the Black-Scholes price.

Theorem 2.2. Under Assumptions $(\mathrm{P})$ and $(\mathrm{R})$, the put ${ }^{2}$ price is approximated by

$$
\begin{align*}
e^{-\int_{0}^{T} r_{t} \mathrm{~d} t} \mathbb{E}\left[\left(K-e^{\int_{0}^{T}\left(r_{t}-q_{t}\right) \mathrm{d} t+X_{T}^{1}}\right)_{+}\right]= & P_{B S}\left(x_{0}, v a r_{T}\right)+\sum_{i=1}^{2} a_{i, T} \frac{\partial^{i+1} P_{B S}}{\partial x^{i} y}\left(x_{0}, v a r_{T}\right) \\
& +\sum_{i=0}^{1} b_{2 i, T} \frac{\partial^{2 i+2} P_{B S}}{\partial x^{2 i} y^{2}}\left(x_{0}, v a r_{T}\right)+\mathcal{E} \tag{2.13}
\end{align*}
$$

where

$$
\begin{array}{rlrl}
v a r_{T} & =\int_{0}^{T} v_{0, t} \mathrm{~d} t, & a_{1, T}=\int_{0}^{T} e^{\kappa s} \rho_{s} \xi_{s} v_{0, s} d s \int_{s}^{T} e^{-\kappa u} d u \\
a_{2, T} & =\int_{0}^{T} e^{\kappa s} \rho_{s} \xi_{s} v_{0, s} d s \int_{s}^{T} \rho_{t} \xi_{t} d t \int_{t}^{T} e^{-\kappa u} d u \\
b_{0, T} & =\int_{0}^{T} e^{2 \kappa s} \xi_{s}^{2} v_{0, s} d s \int_{s}^{T} e^{-\kappa t} d t \int_{t}^{T} e^{-\kappa u} d u, & b_{2, T}=\frac{a_{1, T}^{2}}{2}
\end{array}
$$

The proof is postponed to subsection 5.4. Finally, we give an estimate regarding the error $\mathcal{E}$ arising in the above theorem.

Theorem 2.3. Under Assumptions ( P ) and ( R ), the error in the approximation (2.13) is estimated as follows:

$$
\mathcal{E}=O\left(\xi_{\text {Sup }}^{3} T^{2}\right) .
$$

In view of Theorem 2.3, we may refer to the formula (2.13) as a second order approximation formula w.r.t. the volatility of volatility.

[^22]
### 2.4. Computation of coefficients.

Constant parameters. The case of constant parameters $(\theta, \xi, \rho)$ gives us the coefficients $a$ and $b$ explicitly. Using Mathematica, we derive the following explicit expressions.

Proposition 2.4 (explicit computations). For constant parameters, one has

$$
\begin{aligned}
v a r_{T} & =m_{0} v_{0}+m_{1} \theta, & a_{1, T} & =\rho \xi\left(p_{0} v_{0}+p_{1} \theta\right) \\
a_{2, T} & =(\rho \xi)^{2}\left(q_{0} v_{0}+q_{1} \theta\right), & b_{0, T} & =\xi^{2}\left(r_{0} v_{0}+r_{1} \theta\right)
\end{aligned}
$$

where

$$
\begin{aligned}
m_{0} & =\frac{e^{-\kappa T}\left(-1+e^{\kappa T}\right)}{\kappa} \\
p_{0} & =\frac{e^{-\kappa T}\left(-\kappa T+e^{\kappa T}-1\right)}{\kappa^{2}} \\
q_{0} & =\frac{e^{-\kappa T}\left(-\kappa T(\kappa T+2)+2 e^{\kappa T}-2\right)}{2 \kappa^{3}} \\
r_{0} & =\frac{e^{-2 \kappa T}\left(-4 e^{\kappa T} \kappa T+2 e^{2 \kappa T}-2\right)}{4 \kappa^{3}}
\end{aligned}
$$

$$
m_{1}=T-\frac{e^{-\kappa T}\left(-1+e^{\kappa T}\right)}{\kappa}
$$

$$
p_{1}=\frac{e^{-\kappa T}\left(\kappa T+e^{\kappa T}(\kappa T-2)+2\right)}{\kappa^{2}}
$$

$$
q_{1}=\frac{e^{-\kappa T}\left(2 e^{\kappa T}(\kappa T-3)+\kappa T(\kappa T+4)+6\right)}{2 \kappa^{3}}
$$

$$
r_{1}=\frac{e^{-2 \kappa T}\left(4 e^{\kappa T}(\kappa T+1)+e^{2 \kappa T}(2 \kappa T-5)+1\right)}{4 \kappa^{3}}
$$

Remark 2.1. In the case of constant parameters $(\theta, \xi, \rho)$, we retrieve the usual Heston model. In this particular case, our expansion coincides exactly with Lewis's volatility of volatility series expansion (see equation (3.4), p. 84 in [27] for Lewis's expansion formula and p. 93 in [27] for the explicit calculation of the coefficients $J^{(i)}$ with $\left.\varphi=\frac{1}{2}\right)$. Using his notation, we have $a_{1, T}=J^{(1)}, a_{2, T}=J^{(4)}$, and $b_{0, T}=J^{(3)}$.

Smooth parameters. In this case, we may use a Gauss-Legendre quadrature formula for the computation of the terms $a$ and $b$.

Piecewise constant parameters. The computation of the variance $\operatorname{var}_{T}$ is straightforward. Thus, it remains to provide explicit expressions of $a$ and $b$ as a function of the piecewise constant data. Let $T_{0}=0 \leq T_{1} \leq \cdots \leq T_{n}=T$ such that $\theta, \rho, \xi$ are constant on each interval $] T_{i}, T_{i+1}$ [ and are equal, respectively, to $\theta_{T_{i+1}}, \rho_{T_{i+1}}, \xi_{T_{i+1}}$. Before giving the recursive relation, we need to introduce the following functions: $\tilde{\omega}_{1, t}=\int_{0}^{t} e^{\kappa s} \rho_{s} \xi_{s} v_{0, s} d s, \tilde{\omega}_{2, t}=\int_{0}^{t} e^{2 \kappa s} \xi_{s}^{2} v_{0, s} d s$, $\alpha_{t}=\int_{0}^{t} e^{\kappa s} \rho_{s} \xi_{s} v_{0, s} d s \int_{s}^{t} \rho_{u} \xi_{u} d u, \beta_{t}=\int_{0}^{t} e^{2 \kappa s} \xi_{s}^{2} v_{0, s} d s \int_{s}^{t} e^{-\kappa u} d u$.

Proposition 2.5 (recursive calculations). For piecewise constant coefficients, one has

$$
\begin{aligned}
a_{1, T_{i+1}} & =a_{1, T_{i}}+\tilde{\omega}_{T_{i}, T_{i+1}}^{-\kappa} \tilde{\omega}_{1, T_{i}}+\rho_{T_{i+1}} \xi_{T_{i+1}} f_{\kappa, v_{0, T_{i}}}^{1}\left(\theta_{T_{i+1}}, T_{i}, T_{i+1}\right) \\
a_{2, T_{i+1}} & =a_{2, T_{i}}+\tilde{\omega}_{T_{i}, T_{i+1}}^{-\kappa} \alpha_{T_{i}}+\rho_{T_{i+1}} \xi_{T_{i+1}} \tilde{\omega}_{T_{i}, T_{i+1}}^{0,-\kappa} \tilde{\omega}_{1, T_{i}}+\left(\rho_{T_{i+1}} \xi_{T_{i+1}}\right)^{2} f_{\kappa, v_{0, T_{i}}}^{2}\left(\theta_{T_{i+1}}, T_{i}, T_{i+1}\right), \\
b_{0, T_{i+1}} & =b_{0, T_{i}}+\tilde{\omega}_{T_{i}, T_{i+1}}^{-\kappa} \beta_{T_{i}}+\tilde{\omega}_{T_{i}, T_{i+1}}^{-\kappa, \kappa} \tilde{\omega}_{2, T_{i}}+\xi_{T_{i+1}}^{2} f_{\kappa, v_{0, T}}^{0}\left(\theta_{T_{i+1}}, T_{i}, T_{i+1}\right) \\
\alpha_{T_{i+1}} & =\alpha_{T_{i}}+\rho_{T_{i+1}} \xi_{T_{i+1}}\left(T_{i+1}-T_{i}\right) \tilde{\omega}_{1, T_{i}}+\rho_{T_{i+1}}^{2} \xi_{T_{i+1}}^{2} g_{\kappa, v_{0, T_{i}}}^{1}\left(\theta_{T_{i+1}}, T_{i}, T_{i+1}\right), \\
\beta_{T_{i+1}} & =\beta_{T_{i}}+\tilde{\omega}_{T_{i}, T_{i+1}}^{-\kappa} \tilde{\omega}_{2, T_{i}}+\xi_{T_{i+1}}^{2} g_{\kappa, v_{0, T}}^{2}\left(\theta_{T_{i+1}}, T_{i}, T_{i+1}\right) \\
\tilde{\omega}_{1, T_{i+1}} & =\tilde{\omega}_{1, T_{i}}+\rho_{T_{i+1}} \xi_{T_{i+1}} h_{\kappa, v_{0, T_{i}}}^{1}\left(\theta_{T_{i+1}}, T_{i}, T_{i+1}\right), \\
\tilde{\omega}_{2, T_{i+1}} & =\tilde{\omega}_{2, T_{i}}+\xi_{T_{i+1}}^{2} h_{\kappa, v_{0, T i}}^{2}\left(\theta_{T_{i+1}}, T_{i}, T_{i+1}\right) \\
v_{0, T_{i+1}} & =e^{-\kappa\left(T_{i+1}-T_{i}\right)}\left(v_{0, T_{i}}-\theta_{T_{i+1}}\right)+\theta_{T_{i+1}}
\end{aligned}
$$

where

$$
\begin{aligned}
& \text { where } \\
& f_{\kappa, v_{0}}^{0}(\theta, t, T)=\frac{e^{-2 \kappa T}\left(e^{2 \kappa t}\left(\theta-2 v_{0}\right)+e^{2 \kappa T}\left((-2 \kappa t+2 \kappa T-5) \theta+2 v_{0}\right)+4 e^{\kappa(t+T)}\left((-\kappa t+\kappa T+1) \theta+\kappa(t-T) v_{0}\right)\right)}{4 \kappa^{3}}, \\
& f_{\kappa, v_{0}}^{1}(\theta, t, T)=\frac{e^{-\kappa T}\left(e^{\kappa T}\left((-\kappa t+\kappa T-2) \theta+v_{0}\right)-e^{\kappa t}\left((\kappa t-\kappa T-2) \theta-\kappa t v_{0}+\kappa T v_{0}+v_{0}\right)\right)}{\kappa^{2}}, \\
& f_{\kappa, v_{0}}^{2}(\theta, t, T)=\frac{e^{-\kappa(t+3 T)}\left(2 e^{\kappa(t+3 T)}\left((\kappa(T-t)-3) \theta+v_{0}\right)+e^{2 \kappa(t+T)}\left((\kappa(\kappa(t-T)-4)(t-T)+6) \theta-(\kappa(\kappa(t-T)-2)(t-T)+2) v_{0}\right)\right.}{2 \kappa^{3}} \\
& g_{\kappa, v_{0}}^{1}(\theta, t, T)=\frac{2 e^{\kappa T} \theta+e^{\kappa t}\left(\kappa^{2}(t-T)^{2} v_{0}-(\kappa(\kappa(t-T)-2)(t-T)+2) \theta\right)}{2 \kappa^{2}}, \\
& g_{\kappa, v_{0}}^{2}(\theta, t, T)=\frac{e^{-\kappa T}\left(e^{2 \kappa T} \theta-e^{2 \kappa t}\left(\theta-2 v_{0}\right)+2 e^{\kappa(t+T)}\left(\kappa(t-T)\left(\theta-v_{0}\right)-v_{0}\right)\right)}{2 \kappa^{2}}, \\
& h_{\kappa, v_{0}}^{1}(\theta, t, T)=\frac{e^{\kappa T} \theta+e^{\kappa t}\left((\kappa t-\kappa T-1) \theta+\kappa(T-t) v_{0}\right)}{\kappa}, \\
& h_{\kappa, v_{0}}^{2}(\theta, t, T)=\frac{\left(e^{\kappa t}-e^{\kappa T}\right)\left(e^{\kappa t}\left(\theta-2 v_{0}\right)-e^{\kappa T} \theta\right)}{2 \kappa}, \\
& \text { and } \tilde{\omega}_{t, T}^{u}=\frac{-e^{t u}+e^{T u}}{u}, \tilde{\omega}_{t, T}^{0, u}=\frac{e^{T u}(-t u+T u-1)+e^{t u}}{u^{2}}, \tilde{\omega}_{t, T}^{u, u}=\frac{\left(e^{t u}-e^{T u}\right)^{2}}{2 u^{2}} . \\
& \text { Proof. According to Theorem 2.2, one has }
\end{aligned}
$$

$$
\begin{aligned}
a_{1, T_{i+1}} & =\int_{0}^{T_{i}} e^{\kappa t} \rho_{t} \xi_{t} v_{0, t}\left(\int_{t}^{T_{i+1}} e^{-\kappa s} d s\right) \mathrm{d} t+\int_{T_{i}}^{T_{i+1}} e^{\kappa t} \rho_{t} \xi_{t} v_{0, t}\left(\int_{t}^{T_{i+1}} e^{-\kappa s} d s\right) \mathrm{d} t \\
& =a_{1, T_{i}}+\int_{0}^{T_{i}} e^{\kappa t} \rho_{t} \xi_{t} v_{0, t}\left(\int_{T_{i}}^{T_{i+1}} e^{-\kappa s} d s\right) \mathrm{d} t+\int_{T_{i}}^{T_{i+1}} e^{\kappa t} \rho_{t} \xi_{t} v_{0, t}\left(\int_{t}^{T_{i+1}} e^{-\kappa s} d s\right) \mathrm{d} t \\
& =a_{1, T_{i}}+\left(\int_{T_{i}}^{T_{i+1}} e^{-\kappa s} d s\right) \int_{0}^{T_{i}} e^{\kappa t} \rho_{t} \xi_{t} v_{0, t} \mathrm{~d} t+\int_{T_{i}}^{T_{i+1}} e^{\kappa t} \rho_{t} \xi_{t} v_{0, t}\left(\int_{t}^{T_{i+1}} e^{-\kappa s} d s\right) \mathrm{d} t \\
& =a_{1, T_{i}}+\tilde{\omega}_{T_{i}, T_{i+1}}^{-\kappa} \tilde{\omega}_{1, T_{i}}+\rho_{T_{i+1}} \xi_{T_{i+1}} f_{\kappa, v_{0, T_{i}}}^{1}\left(\theta_{i+1}, T_{i}, T_{i+1}\right)
\end{aligned}
$$

where the functions $f_{\kappa, v_{0}}^{1}$ and $\tilde{\omega}^{-\kappa}$ are calculated analytically using Mathematica. The other terms are calculated analogously.

### 2.5. Corollaries of the approximation formula (2.13).

Averaging Heston's model parameters. We derive a first corollary of the approximation formula in terms of equivalent Heston models. As explained in [36], this averaging principle may facilitate efficient calibration. Namely, we search for equivalent constant parameters $\bar{\kappa}, \bar{\theta}, \bar{\xi}, \bar{\rho}$ for the Heston model ${ }^{3}$

$$
\begin{aligned}
d \bar{X}_{t} & =\sqrt{\bar{v}_{t}} d W_{t}-\frac{\bar{v}_{t}}{2} \mathrm{~d} t, \quad \bar{X}_{0}=x_{0}, \\
d \bar{v}_{t} & =\bar{\kappa}\left(\bar{\theta}_{t}-\bar{v}_{t}\right) \mathrm{d} t+\bar{\xi} \sqrt{\bar{v}_{t}} \mathrm{~d} B_{t}, \quad \bar{v}_{0}=v_{0} \\
\mathrm{~d}\langle W, B\rangle_{t} & =\bar{\rho} \mathrm{d} t
\end{aligned}
$$

that equalize the price of call/put options maturing at $T$ in the time dependent model (equality up to the approximation error $\mathcal{E}$ ). The following rules give the equivalent parameters as a function of the variance $\operatorname{var}_{T}$ and the coefficients $a_{1, T}, a_{2, T}, b_{0, T}$ that are computed in the time dependent model.

[^23]Averaging rule in the case of zero correlation. If $\rho_{t} \equiv 0$, the equivalent constant parameters (for maturity $T$ ) are

$$
\bar{\kappa}=\kappa, \quad \bar{\theta}=\frac{v a r_{T}-m_{0} v_{0}}{m_{1}}, \quad \bar{\xi}=\sqrt{\frac{b_{0, T}}{r_{0} v_{0}+r_{1} \bar{\theta}}}, \quad \bar{\rho}=0 .
$$

Proof. Two sets of prices coincide at maturity $T$ if they have the same approximation formula (2.13). In this case $a_{1, T}=a_{2, T}=b_{2, T}=0$; thus the approximation formula depends only on two quantities $\operatorname{var}_{T}$ and $b_{0, T}$. It is quite clear that there is not a single choice of parameters to fit these two quantities. A simple solution results from the choices of $\bar{\kappa}=\kappa$ and $\bar{\rho}=0$ : then, using Proposition 2.4, we obtain the announced parameters $\bar{\theta}$ and $\bar{\xi}$.

Remark 2.2. In this case of zero correlation and $\theta=v_{0}=\bar{\theta}$, we exactly retrieve Piterbarg's results for the averaged volatility of volatility $\bar{\xi}$ (see [36]).

Averaging rule in the case of nonzero correlation. We follow the same arguments as before. Now the approximation formula also depends on the four quantities $\operatorname{var}_{T}, a_{1, T}$, $a_{2, T}$, and $b_{2, T}$. Thus, equalizing call/put prices at maturity $T$ is equivalent to equalizing these four quantities in both models, by adjusting $\bar{\kappa}, \bar{\theta}, \bar{\xi}$, and $\bar{\rho}$. Unfortunately, we have not found a closed expression for these equivalent parameters. An alternative and simpler way of proceeding consists of modifying the unobserved initial value $\bar{v}_{0}$ of the variance process while keeping $\bar{\kappa}=\kappa$. For nonvanishing correlation $\left(\rho_{t}\right)_{t}$, it leads to two possibilities

$$
\begin{array}{ll}
\overline{v_{0}}=b \frac{\left(b \pm \sqrt{b^{2}-4 a c}\right)}{2 a}-\frac{p_{1} v a r_{T}}{m_{1} p_{0}-m_{0} p_{1}}, & \bar{\theta}=\frac{v a r_{T}-m_{0} \overline{v_{0}}}{m_{1}}, \\
\bar{\xi}=\sqrt{\frac{b_{0, T}}{r_{0} \overline{v_{0}}+r_{1} \bar{\theta}}}, & \bar{\rho}=-\frac{2 a}{\bar{\xi}\left(b \pm \sqrt{b^{2}-4 a c}\right)},
\end{array}
$$

such that

$$
a=\frac{a_{2, T} m_{1}}{m_{1} q_{0}-m_{0} q_{1}}, \quad b=-\frac{a_{1, T} m_{1}}{m_{1} p_{0}-m_{0} p_{1}}, \quad c=\operatorname{var}_{T}\left(\frac{p_{1}}{m_{1} p_{0}-m_{0} p_{1}}-\frac{q_{1}}{m_{1} q_{0}-m_{0} q_{1}}\right),
$$

where $m_{0}, m_{1}, p_{0}, p_{1}, q_{0}, q_{1}, r_{0}$, and $r_{1}$ are given in Proposition 2.4.
In practice, only one solution gives realistic parameters. However, this rule is heuristic since there is a priori no guarantee that these averaged parameters satisfy Assumption (P), which is the basis for the argument's correctness.

Proof. Using Proposition 2.4, one has to solve the following system of equations:

$$
\begin{aligned}
\operatorname{var}_{T} & =m_{0} \bar{v}_{0}+m_{1} \bar{\theta}, & a_{1, T} & =\bar{\rho} \bar{\xi}\left(p_{0} \bar{v}_{0}+p_{1} \bar{\theta}\right), \\
a_{2, T} & =(\bar{\rho} \bar{\xi})^{2}\left(q_{0} \bar{v}_{0}+q_{1} \bar{\theta}\right), & b_{0, T} & =\bar{\xi}^{2}\left(r_{0} \bar{v}_{0}+r_{1} \bar{\theta}\right) .
\end{aligned}
$$

The first equation gives $\bar{\theta}=\frac{v a r_{T}-m_{0} \bar{v}_{0}}{m_{1}}$. Replacing this identity in $a_{1, T}$ and $a_{2, T}$ gives

$$
\bar{v}_{0}=\left(\frac{a_{1, T}}{(\bar{\rho} \bar{\xi})}-\frac{p_{1} v a r_{T}}{m_{1}}\right) \frac{m_{1}}{p_{0} m_{1}-p_{1} m_{0}}, \quad \bar{v}_{0}=\left(\frac{a_{2, T}}{(\bar{\rho} \bar{\xi})^{2}}-\frac{q_{1} v a r_{T}}{m_{1}}\right) \frac{m_{1}}{q_{0} m_{1}-q_{1} m_{0}} .
$$

It readily leads to a quadratic equation $a x^{2}+b x+c=0$ with $x=\frac{1}{\bar{\rho} \xi}$. By solving this equation, we easily complete the proof of the result.

Collinearity effect in the Heston model. Another corollary of the approximation formula (2.13) is that we can obtain the same vanilla prices at time $T$ with different sets of parameters. For instance, take on the one hand $v_{0}=\theta=4 \%, \kappa_{1}=2$, and $\xi_{1}=30 \%$ (model $M_{1}$ ) and on the other hand $v_{0}=\theta=4 \%, \kappa_{2}=3$, and $\xi_{2}=38.042 \%$ (model $M_{2}$ ), both models having zero correlation. The resulting errors between implied volatilities within the two models are presented in Table 1: they are so small that prices can be considered as equal. Actually, this kind of example is easy to create even with nonnull correlation: as before, in view of the approximation formula (2.13), it is sufficient to equalize the four quantities $\mathrm{var}_{T}$, $a_{1, T}, a_{2, T}$, and $b_{2, T}$.

## Table 1

Errors in implied Black-Scholes volatilities (in bps) between the closed formulas (see the appendix) of the two models $M_{1}$ and $M_{2}$ expressed as relative strikes. Maturity is equal to one year.

| Strikes $K$ | $80 \%$ | $90 \%$ | $100 \%$ | $110 \%$ | $120 \%$ |
| :--- | :--- | :--- | :--- | :--- | :--- |
| Model $M_{1}$ | $20.12 \%$ | $19.64 \%$ | $19.50 \%$ | $19.62 \%$ | $19.92 \%$ |
| Model $M_{2}$ | $20.11 \%$ | $19.65 \%$ | $19.51 \%$ | $19.62 \%$ | $19.92 \%$ |
| Errors | 0.69 | -0.35 | -0.81 | -0.42 | 0.34 |
| (bps) |  |  |  |  |  |

As a consequence, calibrating a Heston model using options with a single maturity is an ill-posed problem, which is not a surprising fact. From the work [16] by Elices, we know that it may be possible to compute in a unique manner the Heston parameters by adding other options in the set of calibrated instruments.
3. Numerical accuracy of the approximation. We give numerical results of the performance of our method. In what follows, the spot $S_{0}$, the risk-free rate $r$, and the dividend yield $q$ are set, respectively, to $100,0 \%$, and $0 \%$. The initial value of the variance process is set to $v_{0}=4 \%$ (initial volatility equal to $20 \%$ ). Then we study the numerical accuracy w.r.t. $K, T, \kappa, \theta, \xi$, and $\rho$ by testing different values for these parameters.

In order to present more interesting results for various relevant maturities and strikes, we allow the range of strikes to vary over the maturities. The strike values evolve approximately as $S_{0} \exp (c \sqrt{\theta T})$ for some real numbers $c$ and $\theta=6 \%$. The extreme values of $c$ are chosen to be equal to $\pm 2.57$, which represents the $1 \%-99 \%$ quantile of the standard normal distribution. This corresponds to very out-of-the-money options or very deep-in-the-money options. The set of pairs (maturity, strike) chosen for the tests is given in Table 2.

Constant parameters. In Table 3, we report the numerical results when $\theta=6 \%, \kappa=3$, $\xi=30 \%$, and $\rho=0 \%$, giving the errors of implied Black-Scholes volatilities between our approximation formula (see (2.13)) and the price calculated using the closed formula (see the appendix) for the maturities and strikes of Table 2. The table should be read as follows: for example, for one year maturity and strike equal to 170 , the implied volatility is equal to $24.14 \%$ using the closed formula and $24.20 \%$ with the approximation formula, giving an error of -6.33 bps . In Table 3, we observe that the errors do not exceed 7 bps for a large range of strikes and maturities. We notice that the errors are surprisingly higher for short maturities. At first sight, it is counterintuitive, as one would expect our perturbation method to work better for short maturities and worse for long maturities, since the difference between our

Table 2
Set of maturities and relative strikes (in \%) used for the numerical tests.

| $T / K$ |  |  |  |  |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 3 M | 70 | 80 | 90 | 100 | 110 | 120 | 125 | 130 |
| 6 M | 60 | 70 | 80 | 100 | 110 | 130 | 140 | 150 |
| 1 Y | 50 | 60 | 80 | 100 | 120 | 150 | 170 | 180 |
| 2 Y | 40 | 50 | 70 | 100 | 130 | 180 | 210 | 240 |
| 3 Y | 30 | 40 | 60 | 100 | 140 | 200 | 250 | 290 |
| 5 Y | 20 | 30 | 60 | 100 | 150 | 250 | 320 | 400 |
| 7 Y | 10 | 30 | 50 | 100 | 170 | 300 | 410 | 520 |
| 10 Y | 10 | 20 | 50 | 100 | 190 | 370 | 550 | 730 |

Table 3
Implied Black-Scholes volatilities of the closed formula and of the approximation formula and related errors (in bps), expressed as a function of maturities in fractions of years and relative strikes. Parameters: $\theta=6 \%$, $\kappa=3, \xi=30 \%$, and $\rho=0 \%$.

| 3 M | $23.24 \%$ | $22.14 \%$ | $21.43 \%$ | $21.19 \%$ | $21.39 \%$ | $21.86 \%$ | $22.14 \%$ | $22.44 \%$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
|  | $23.06 \%$ | $22.19 \%$ | $21.42 \%$ | $21.19 \%$ | $21.38 \%$ | $21.88 \%$ | $22.19 \%$ | $22.49 \%$ |
|  | $\mathbf{1 8 . 0 1}$ | $-\mathbf{4 . 8 6}$ | $\mathbf{0 . 5 3}$ | $\mathbf{0 . 3 8}$ | $\mathbf{0 . 6 5}$ | $-\mathbf{2 . 6 8}$ | $-\mathbf{4 . 8 6}$ | $-\mathbf{4 . 7 1}$ |
| 6 M | $24.32 \%$ | $23.29 \%$ | $22.55 \%$ | $21.99 \%$ | $22.10 \%$ | $22.75 \%$ | $23.17 \%$ | $23.60 \%$ |
|  | $24.12 \%$ | $23.36 \%$ | $22.57 \%$ | $21.98 \%$ | $22.09 \%$ | $22.79 \%$ | $23.24 \%$ | $23.65 \%$ |
|  | $\mathbf{1 9 . 6 9}$ | $-\mathbf{7 . 1 7}$ | $-\mathbf{1 . 8 9}$ | $\mathbf{0 . 9 3}$ | $\mathbf{1 . 0 5}$ | $-\mathbf{3 . 9 7}$ | $-\mathbf{7 . 1 2}$ | $-\mathbf{4 . 5 7}$ |
| 1 Y | $24.85 \%$ | $24.06 \%$ | $23.14 \%$ | $22.90 \%$ | $23.06 \%$ | $23.66 \%$ | $24.14 \%$ | $24.38 \%$ |
|  | $24.78 \%$ | $24.12 \%$ | $23.14 \%$ | $22.89 \%$ | $23.06 \%$ | $23.71 \%$ | $24.20 \%$ | $24.42 \%$ |
|  | $\mathbf{7 . 7 2}$ | $-\mathbf{6 . 4 9}$ | $\mathbf{0 . 2 6}$ | $\mathbf{1 . 1 2}$ | $\mathbf{0 . 7 2}$ | -4.54 | $-\mathbf{6 . 3 3}$ | $-\mathbf{4 . 2 7}$ |
| 2 Y | $24.86 \%$ | $24.36 \%$ | $23.82 \%$ | $23.61 \%$ | $23.73 \%$ | $24.16 \%$ | $24.46 \%$ | $24.76 \%$ |
|  | $24.86 \%$ | $24.40 \%$ | $23.82 \%$ | $23.61 \%$ | $23.72 \%$ | $24.19 \%$ | $24.50 \%$ | $24.78 \%$ |
|  | $-\mathbf{0 . 2 1}$ | $-\mathbf{3 . 5 1}$ | $-\mathbf{0 . 1 2}$ | $\mathbf{0 . 6 8}$ | $\mathbf{0 . 3 7}$ | $-\mathbf{2 . 5 4}$ | $-\mathbf{3 . 6 2}$ | $-\mathbf{1 . 7 1}$ |
| 3 Y | $24.95 \%$ | $24.53 \%$ | $24.10 \%$ | $23.89 \%$ | $23.98 \%$ | $24.27 \%$ | $24.53 \%$ | $24.74 \%$ |
|  | $24.94 \%$ | $24.55 \%$ | $24.10 \%$ | $23.89 \%$ | $23.98 \%$ | $24.28 \%$ | $24.55 \%$ | $24.75 \%$ |
|  | $\mathbf{1 . 8 0}$ | $-\mathbf{2 . 1 2}$ | $-\mathbf{0 . 3 3}$ | $\mathbf{0 . 3 9}$ | $\mathbf{0 . 1 9}$ | $-\mathbf{1 . 2 7}$ | $-\mathbf{- 2 . 1 2}$ | $-\mathbf{1 . 2 6}$ |
| 5 Y | $24.88 \%$ | $24.56 \%$ | $24.20 \%$ | $24.12 \%$ | $24.17 \%$ | $24.38 \%$ | $24.53 \%$ | $24.69 \%$ |
|  | $24.86 \%$ | $24.57 \%$ | $24.20 \%$ | $24.12 \%$ | $24.17 \%$ | $24.39 \%$ | $24.54 \%$ | $24.70 \%$ |
| 7 Y | $\mathbf{1 . 3 8}$ | $-\mathbf{0 . 9 6}$ | $\mathbf{0 . 0 3}$ | $\mathbf{0 . 1 7}$ | $\mathbf{0 . 1 0}$ | $-\mathbf{0 . 5 8}$ | $-\mathbf{0 . 9 5}$ | $-\mathbf{0 . 5 9}$ |
|  | $25.03 \%$ | $24.46 \%$ | $24.30 \%$ | $24.23 \%$ | $24.27 \%$ | $24.42 \%$ | $24.54 \%$ | $24.65 \%$ |
|  | $24.97 \%$ | $24.46 \%$ | $24.30 \%$ | $24.22 \%$ | $24.27 \%$ | $24.42 \%$ | $24.55 \%$ | $24.66 \%$ |
| 10 Y | $\mathbf{5 . 7 2}$ | $-\mathbf{0 . 4 3}$ | $-\mathbf{0 . 0 2}$ | $\mathbf{0 . 0 9}$ | $\mathbf{0 . 0 4}$ | $-\mathbf{0 . 3 3}$ | $-\mathbf{0 . 5 4}$ | $-\mathbf{0 . 3 5}$ |
|  | $24.72 \%$ | $24.51 \%$ | $24.34 \%$ | $24.30 \%$ | $24.34 \%$ | $24.44 \%$ | $24.54 \%$ | $24.62 \%$ |
|  | $24.71 \%$ | $24.51 \%$ | $24.34 \%$ | $24.30 \%$ | $24.34 \%$ | $24.44 \%$ | $24.54 \%$ | $24.62 \%$ |
|  | $\mathbf{0 . 4 2}$ | $-\mathbf{0 . 2 8}$ | $\mathbf{0 . 0 2}$ | $\mathbf{0 . 0 5}$ | $\mathbf{0 . 0 2}$ | $-\mathbf{0 . 1 7}$ | $-\mathbf{0 . 2 9}$ | $-\mathbf{0 . 1 9}$ |

proxy model (BS with volatility $\left.\left(v_{0, t}\right)_{t}\right)$ and the original one is increasing w.r.t. time. In fact, this intuition is true for prices but not for implied volatilities. When we compare the price errors (in price $\mathrm{bp}^{4}$ ) for the same data, we observe in Table 4 that the error terms are not any bigger for short maturities but vary slightly over time with two observed effects. The error term first increases over time as the error between the proxy and the original model increases

[^24]Table 4
Put prices of the closed formulas and of the approximation formula and related errors (in bps), expressed as a function of maturities in fractions of years and relative strikes. Parameters: $\theta=6 \%, \kappa=3, \xi=30 \%$, and $\rho=0 \%$.

| 3 M | 30.00 | 20.08 | 10.87 | 4.22 | 1.14 | 0.24 | 0.10 | 0.04 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
|  | 30.00 | 20.08 | 10.87 | 4.22 | 1.14 | 0.24 | 0.10 | 0.04 |
|  | $\mathbf{0 . 0 3}$ | $-\mathbf{0 . 1 1}$ | $\mathbf{0 . 0 6}$ | $\mathbf{0 . 0 8}$ | $\mathbf{0 . 0 9}$ | $-\mathbf{0 . 1 5}$ | $-\mathbf{0 . 1 4}$ | $-\mathbf{0 . 0 7}$ |
| 6 M | 40.01 | 30.07 | 20.52 | 6.20 | 2.72 | 0.40 | 0.14 | 0.05 |
|  | 40.01 | 30.08 | 20.52 | 6.19 | 2.71 | 0.40 | 0.14 | 0.05 |
|  | $\mathbf{0 . 0 5}$ | $-\mathbf{0 . 1 6}$ | $-\mathbf{0 . 1 8}$ | $\mathbf{0 . 2 6}$ | $\mathbf{0 . 2 6}$ | $-\mathbf{0 . 3 4}$ | $-\mathbf{0 . 2 9}$ | $-\mathbf{0 . 0 8}$ |
| 1 Y | 50.01 | 40.11 | 21.84 | 9.12 | 3.08 | 0.51 | 0.15 | 0.09 |
|  | 50.01 | 40.11 | 21.84 | 9.11 | 3.07 | 0.52 | 0.16 | 0.09 |
|  | $\mathbf{0 . 0 4}$ | $-\mathbf{0 . 2 1}$ | $\mathbf{0 . 0 6}$ | $\mathbf{0 . 4 4}$ | $\mathbf{0 . 2 3}$ | $-\mathbf{0 . 5 1}$ | $-\mathbf{0 . 2 9}$ | $-\mathbf{0 . 1 2}$ |
| 2 Y | 60.03 | 50.20 | 32.08 | 13.26 | 4.71 | 0.79 | 0.28 | 0.11 |
|  | 60.03 | 50.20 | 32.08 | 13.26 | 4.71 | 0.79 | 0.29 | 0.11 |
|  | $\mathbf{0 . 0 0}$ | $-\mathbf{0 . 1 8}$ | $-\mathbf{0 . 0 3}$ | $\mathbf{0 . 3 8}$ | $\mathbf{0 . 1 7}$ | $-\mathbf{0 . 4 3}$ | $-\mathbf{0 . 2 9}$ | $-\mathbf{0 . 0 6}$ |
| 3 Y | 70.02 | 60.15 | 41.70 | 16.39 | 5.73 | 1.21 | 0.36 | 0.15 |
|  | 70.02 | 60.15 | 41.70 | 16.39 | 5.73 | 1.21 | 0.37 | 0.15 |
|  | $\mathbf{0 . 0 1}$ | $-\mathbf{0 . 0 9}$ | $-\mathbf{0 . 0 8}$ | $\mathbf{0 . 2 7}$ | $\mathbf{0 . 1 1}$ | $-\mathbf{0 . 3 1}$ | $-\mathbf{0 . 2 2}$ | $-\mathbf{0 . 0 7}$ |
| 5 Y | 80.01 | 70.15 | 43.80 | 21.26 | 8.50 | 1.61 | 0.58 | 0.21 |
|  | 80.01 | 70.15 | 43.80 | 21.26 | 8.50 | 1.61 | 0.58 | 0.21 |
|  | $\mathbf{0 . 0 1}$ | $-\mathbf{0 . 0 4}$ | $\mathbf{0 . 0 1}$ | $\mathbf{0 . 1 5}$ | $\mathbf{0 . 0 8}$ | $-\mathbf{0 . 1 9}$ | $-\mathbf{0 . 1 5}$ | $-\mathbf{0 . 0 4}$ |
| 7 Y | 90.00 | 70.42 | 53.15 | 25.14 | 9.32 | 1.97 | 0.66 | 0.26 |
|  | 90.00 | 70.42 | 53.15 | 25.14 | 9.32 | 1.97 | 0.67 | 0.26 |
|  | $\mathbf{0 . 0 0}$ | $-\mathbf{0 . 0 4}$ | $-\mathbf{0 . 0 1}$ | $\mathbf{0 . 0 9}$ | $\mathbf{0 . 0 4}$ | $-\mathbf{0 . 1 4}$ | $-\mathbf{0 . 1 0}$ | $-\mathbf{0 . 0 3}$ |
| 10 Y | 90.01 | 80.23 | 55.22 | 29.92 | 11.49 | 2.62 | 0.84 | 0.33 |
|  | 90.01 | 80.23 | 55.22 | 29.92 | 11.49 | 2.62 | 0.84 | 0.33 |
|  | $\mathbf{0 . 0 0}$ | $-\mathbf{0 . 0 2}$ | $\mathbf{0 . 0 1}$ | $\mathbf{0 . 0 6}$ | $\mathbf{0 . 0 3}$ | $-\mathbf{0 . 0 9}$ | $-\mathbf{0 . 0 7}$ | $-\mathbf{0 . 0 2}$ |

over time, as forecasted. But for long maturities, presumably because the volatility converges to its stationary regime, errors decrease. The mean absolute error over the 64 prices is 0.13 bps. When we convert these prices to implied Black-Scholes volatilities, these error terms are dramatically amplified for short maturities due to very small vega. Finally, note that for fixed maturity, price errors are quite uniform w.r.t. strike $K$.

Impact of the correlation. Results are analogous for various values of correlation. For instance for $\rho=20 \%$ (resp., $-20 \%$ and $-50 \%$ ), the mean absolute error for the price is 0.28 bps (resp., 0.33 bps and 0.84 bps ). We refer to Table 5 for the prices when $\rho=-50 \%$ (for other values of $\rho$, results can be found in the first version of this work [9]). We notice that the errors are smaller for a correlation close to zero and become larger when the absolute value of the correlation increases. However, for realistic correlation values ( $-50 \%$, for instance), the accuracy for the usual maturities and strikes remains excellent, except for very extreme strikes.

Impact of the volatility of volatility. In view of Theorem 2.3, the smaller the volatility of volatility, the more accurate the approximation. In the following numerical tests, we increase $\xi$. We consider as Heston parameters the calibrated parameters obtained in [6, Table III]: $\kappa=1.15, \theta=3.48 \%, \xi=39 \%$, and $\rho=-64 \%$. Moreover, we set $v_{0}=4 \%$. We vary the value of $\xi$ in the numerical tests from $0 \%$ to $100 \%$. There are two important values for $\xi$ :

- The positivity value $\sqrt{2 \kappa \theta}$. For this value, the so-called positivity ratio is $\frac{2 \kappa \theta}{\xi^{2}}=1$.

Table 5
Put prices of the closed formula and of the approximation formula and related errors (in bp), expressed as a function of maturities in fractions of years and relative strikes. Parameters: $\theta=6 \%, \kappa=3, \xi=30 \%$, and $\rho=-50 \%$.

| 3 M | 30.01 | 20.14 | 11.01 | 4.21 | 0.95 | 0.12 | 0.03 | 0.01 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
|  | 30.01 | 20.15 | 11.02 | 4.21 | 0.94 | 0.11 | 0.03 | 0.01 |
|  | $\mathbf{0 . 2 1}$ | $-\mathbf{0 . 4 7}$ | $-\mathbf{0 . 3 1}$ | $\mathbf{0 . 0 4}$ | $\mathbf{0 . 5 7}$ | $\mathbf{0 . 8 2}$ | $\mathbf{0 . 1 6}$ | $-\mathbf{0 . 3 6}$ |
| 6 M | 40.02 | 30.15 | 20.70 | 6.16 | 2.43 | 0.19 | 0.04 | 0.01 |
|  | 40.02 | 30.15 | 20.71 | 6.15 | 2.42 | 0.17 | 0.04 | 0.02 |
|  | $\mathbf{0 . 3 7}$ | $-\mathbf{0 . 5 9}$ | $-\mathbf{1 . 3 3}$ | $\mathbf{0 . 2 3}$ | $\mathbf{0 . 8 1}$ | $\mathbf{1 . 5 9}$ | $-\mathbf{0 . 0 9}$ | $-\mathbf{1 . 0 5}$ |
| 1 Y | 50.04 | 40.21 | 22.11 | 9.03 | 2.59 | 0.22 | 0.03 | 0.01 |
|  | 50.04 | 40.22 | 22.12 | 9.02 | 2.57 | 0.21 | 0.05 | 0.03 |
|  | $\mathbf{0 . 3 6}$ | $-\mathbf{0 . 8 8}$ | $-\mathbf{1 . 1 7}$ | $\mathbf{0 . 6 1}$ | $\mathbf{2 . 2 7}$ | $\mathbf{1 . 6 7}$ | $-\mathbf{1 . 0 5}$ | $-\mathbf{1 . 6 9}$ |
| 2 Y | 60.08 | 50.33 | 32.38 | 13.11 | 4.06 | 0.39 | 0.08 | 0.02 |
|  | 60.08 | 50.34 | 32.39 | 13.10 | 4.03 | 0.37 | 0.09 | 0.04 |
|  | $\mathbf{0 . 0 9}$ | $-\mathbf{1 . 0 0}$ | $-\mathbf{1 . 3 2}$ | $\mathbf{0 . 8 0}$ | $\mathbf{2 . 4 7}$ | $\mathbf{1 . 5 9}$ | $-\mathbf{0 . 8 4}$ | $-\mathbf{2 . 0 0}$ |
| 3 Y | 70.05 | 60.25 | 41.99 | 16.20 | 4.98 | 0.69 | 0.13 | 0.03 |
|  | 70.05 | 60.25 | 42.00 | 16.19 | 4.96 | 0.67 | 0.13 | 0.05 |
|  | $\mathbf{0 . 1 7}$ | $-\mathbf{0 . 5 4}$ | $-\mathbf{1 . 2 1}$ | $\mathbf{0 . 7 2}$ | $\mathbf{2 . 2 0}$ | $\mathbf{1 . 7 3}$ | $-\mathbf{0 . 7 4}$ | $-\mathbf{1 . 8 0}$ |
| 5 Y | 80.03 | 70.23 | 44.06 | 21.01 | 7.65 | 0.99 | 0.25 | 0.06 |
|  | 80.03 | 70.23 | 44.07 | 21.00 | 7.64 | 0.98 | 0.26 | 0.07 |
|  | $\mathbf{0 . 1 1}$ | $-\mathbf{0 . 3 0}$ | $-\mathbf{0 . 5 3}$ | $\mathbf{0 . 5 4}$ | $\mathbf{1 . 5 4}$ | $\mathbf{1 . 2 9}$ | $-\mathbf{0 . 3 8}$ | $-\mathbf{1 . 5 0}$ |
| 7 Y | 90.00 | 70.54 | 53.40 | 24.84 | 8.36 | 1.28 | 0.31 | 0.09 |
|  | 90.00 | 70.55 | 53.40 | 24.84 | 8.35 | 1.27 | 0.32 | 0.10 |
|  | $\mathbf{0 . 0 6}$ | $-\mathbf{0 . 4 1}$ | $-\mathbf{0 . 4 4}$ | $\mathbf{0 . 4 3}$ | $\mathbf{1 . 3 2}$ | $\mathbf{1 . 0 4}$ | $-\mathbf{0 . 4 5}$ | $-\mathbf{1 . 3 2}$ |
| 10 Y | 90.02 | 80.30 | 55.42 | 29.57 | 10.43 | 1.82 | 0.44 | 0.13 |
|  | 90.02 | 80.30 | 55.42 | 29.57 | 10.42 | 1.81 | 0.44 | 0.14 |
|  | $\mathbf{0 . 0 3}$ | $-\mathbf{0 . 1 8}$ | $-\mathbf{0 . 2 0}$ | $\mathbf{0 . 3 4}$ | $\mathbf{1 . 0 4}$ | $\mathbf{0 . 8 9}$ | $-\mathbf{0 . 4 2}$ | $-\mathbf{1 . 1 7}$ |

Hence, the positivity assumption (Assumption (P)) is maintained while $\xi$ is smaller than $\sqrt{2 \kappa \theta}$. For the current example, the positivity value is $28.28 \%$.

- The calibrated value. For this example, it is $39 \%$. Remark that the related positivity ratio is $0.53<1$. Hence, the positivity assumption (Assumption (P)) is not satisfied for the calibrated parameters.
We plot in Figure 1 the mean absolute error for the prices ${ }^{5}$ at each maturity (in bps) according to the volatility of volatility. As expected from Theorem 2.3, we observe that the mean absolute error is increasing w.r.t. the volatility of volatility and the maturity. For $\xi \leq \sqrt{2 \kappa \theta}$ (when the positivity assumption (Assumption (P)) is satisfied), the accuracy is excellent. For the calibrated value, this is very satisfactory as well. Even for larger values of $\xi$, it is very accurate for nonlarge maturities.

Piecewise constant parameters. Heston's constant parameters have been set to $v_{0}=$ $4 \%, \kappa=3$. In addition, the piecewise constant functions $\theta, \xi$, and $\rho$ are equal, respectively, at each interval of the form $] \frac{i}{4}, \frac{i+1}{4}$ [ to $4 \%+i \times 0.05 \%, 30 \%+i \times 0.5 \%$, and $-20 \%+i \times 0.35 \%$. In Tables 7 and 8, we report values using three different formulas. For a given maturity, the first row is obtained using the closed formula with piecewise constant parameters (see the appendix), the second row uses our approximation formula (2.13), and the third row uses the

[^25]

Figure 1. The mean absolute error for the prices (in bps), expressed as a function of the volatility of volatility and computed for each maturity.

Table 6
Equivalent averaged parameters.

| $T$ | $\bar{v}_{0}$ | $\bar{\theta}$ | $\bar{\xi}$ | $\bar{\rho}$ |
| :--- | :--- | :--- | :--- | :--- |
| 3 M | $4 \%$ | $4 \%$ | $30 \%$ | $-20 \%$ |
| 6M | $3.97 \%$ | $4.04 \%$ | $30.12 \%$ | $-19.93 \%$ |
| 1Y | $3.28 \%$ | $4.38 \%$ | $30.89 \%$ | $-19.72 \%$ |
| 2Y | $4.64 \%$ | $4.02 \%$ | $31.12 \%$ | $-18.95 \%$ |
| 3Y | $56.24 \%$ | $4.04 \%$ | $32.10 \%$ | $-18.20 \%$ |
| 5Y | $28.58 \%$ | $2.68 \%$ | $33.63 \%$ | $-16.52 \%$ |
| 7Y | $84.92 \%$ | $0.59 \%$ | $35.41 \%$ | $-14.80 \%$ |
| 10 Y | $14.54 \%$ | $4.57 \%$ | $39.98 \%$ | $-12.32 \%$ |

closed formula with constant parameters computed by averaging (see subsection 2.5). In order to give complete information on our tests, we also report in Table 6 the values used for the averaging parameters (following subsection 2.5).

Of course, the quickest approach is the use of the approximation formula (2.13). As before, its accuracy is very good, except for very extreme strikes. It is quite interesting to observe that the averaging rules that we propose are extremely accurate.

Calibration to real market data. In this section, we will show how the calibration using time dependent parameters strongly reduces the calibration error.

As a benchmark for the pricing method (i.e., the Lewis formula for constant coefficients or our formula for time dependent coefficients), we choose a Monte Carlo approach using the very accurate QE scheme of Andersen [4] (with 100000 simulations and a monthly time step). Then, we consider the implied Black-Scholes volatility surface of the S\&P 500 Index (see Table 9). The spot value is 1360.14 . The equivalent risk-free rate is computed from the bond price curve on June 6, 2008. First, we calibrate the constant Heston model using the Lewis formula. The calibrated parameters are $\kappa=41.95 \%, \rho=-38.27 \%, \xi=59.34 \%, \theta=12.40 \%$,

Table 7
Implied Black-Scholes volatilities of the closed formula, of the approximation formula, and of the averaging formula, expressed as a function of maturities in fractions of years and relative strikes. Piecewise constant parameters.

| 3 M | $23.45 \%$ | $21.88 \%$ | $20.58 \%$ | $19.70 \%$ | $19.39 \%$ | $19.55 \%$ | $19.74 \%$ | $19.97 \%$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $22.73 \%$ | $21.96 \%$ | $20.60 \%$ | $19.69 \%$ | $19.35 \%$ | $19.53 \%$ | $19.84 \%$ | $20.28 \%$ |
|  | $23.45 \%$ | $21.88 \%$ | $20.58 \%$ | $19.70 \%$ | $19.39 \%$ | $19.55 \%$ | $19.74 \%$ | $19.97 \%$ |
| 6 M | $24.09 \%$ | $22.59 \%$ | $21.30 \%$ | $19.63 \%$ | $19.33 \%$ | $19.58 \%$ | $19.92 \%$ | $20.31 \%$ |
|  | $23.09 \%$ | $22.60 \%$ | $21.43 \%$ | $19.61 \%$ | $19.30 \%$ | $19.58 \%$ | $20.19 \%$ | $20.93 \%$ |
|  | $24.09 \%$ | $22.59 \%$ | $21.30 \%$ | $19.63 \%$ | $19.33 \%$ | $19.58 \%$ | $19.92 \%$ | $20.31 \%$ |
| 1 Y | $23.95 \%$ | $22.66 \%$ | $20.76 \%$ | $19.70 \%$ | $19.37 \%$ | $19.69 \%$ | $20.12 \%$ | $20.36 \%$ |
|  | $23.12 \%$ | $22.66 \%$ | $20.81 \%$ | $19.68 \%$ | $19.32 \%$ | $19.78 \%$ | $20.62 \%$ | $21.05 \%$ |
|  | $23.95 \%$ | $22.66 \%$ | $20.76 \%$ | $19.70 \%$ | $19.37 \%$ | $19.69 \%$ | $20.12 \%$ | $20.35 \%$ |
| 2 Y | $23.26 \%$ | $22.30 \%$ | $21.01 \%$ | $19.99 \%$ | $19.66 \%$ | $19.83 \%$ | $20.09 \%$ | $20.37 \%$ |
|  | $22.84 \%$ | $22.33 \%$ | $21.04 \%$ | $19.96 \%$ | $19.62 \%$ | $19.90 \%$ | $20.43 \%$ | $21.02 \%$ |
| 3 Y | $23.26 \%$ | $22.30 \%$ | $21.01 \%$ | $19.98 \%$ | $19.66 \%$ | $19.83 \%$ | $20.09 \%$ | $20.37 \%$ |
|  | $23.28 \%$ | $22.40 \%$ | $21.27 \%$ | $20.26 \%$ | $19.96 \%$ | $20.02 \%$ | $20.23 \%$ | $20.43 \%$ |
|  | $22.81 \%$ | $22.38 \%$ | $21.33 \%$ | $20.24 \%$ | $19.93 \%$ | $20.04 \%$ | $20.47 \%$ | $20.90 \%$ |
| 5 Y | $23.28 \%$ | $22.40 \%$ | $21.27 \%$ | $20.26 \%$ | $19.96 \%$ | $20.02 \%$ | $20.23 \%$ | $20.42 \%$ |
|  | $23.22 \%$ | $22.46 \%$ | $21.34 \%$ | $20.77 \%$ | $20.54 \%$ | $20.54 \%$ | $20.65 \%$ | $20.80 \%$ |
|  | $22.88 \%$ | $22.44 \%$ | $21.35 \%$ | $20.77 \%$ | $20.52 \%$ | $20.55 \%$ | $20.76 \%$ | $21.09 \%$ |
| 7 Y | $23.22 \%$ | $22.46 \%$ | $21.34 \%$ | $20.77 \%$ | $20.54 \%$ | $20.54 \%$ | $20.64 \%$ | $20.79 \%$ |
|  | $23.86 \%$ | $22.36 \%$ | $21.81 \%$ | $21.26 \%$ | $21.06 \%$ | $21.06 \%$ | $21.16 \%$ | $21.27 \%$ |
|  | $23.25 \%$ | $22.39 \%$ | $21.82 \%$ | $21.26 \%$ | $21.05 \%$ | $21.07 \%$ | $21.23 \%$ | $21.45 \%$ |
| 10 Y | $23.86 \%$ | $22.37 \%$ | $21.81 \%$ | $21.26 \%$ | $21.06 \%$ | $21.06 \%$ | $21.15 \%$ | $21.26 \%$ |
|  | $23.59 \%$ | $22.96 \%$ | $22.30 \%$ | $21.97 \%$ | $21.82 \%$ | $21.83 \%$ | $21.92 \%$ | $22.02 \%$ |
|  | $23.46 \%$ | $22.98 \%$ | $22.30 \%$ | $21.97 \%$ | $21.81 \%$ | $21.84 \%$ | $21.96 \%$ | $22.12 \%$ |
|  | $23.59 \%$ | $22.96 \%$ | $22.30 \%$ | $21.97 \%$ | $21.82 \%$ | $21.83 \%$ | $21.92 \%$ | $22.01 \%$ |

and $v_{0}=4.40 \%$. We give in Table 10 the error of implied Black-Scholes volatilities between the calibrated price computed using the Monte Carlo scheme and the market price. The mean absolute error is $0.72 \%$. Now, we calibrate the implied Black-Scholes volatility surface using our approximation formula (2.13) for piecewise constant parameters. The time break points are the surface time break points. The calibrated parameters are given in Table 11. For the reader interested in practical considerations in the calibration procedure, we refer him/her to [16] for interesting suggestions defining constraints on the parameters that ensure a more robust calibration. We give in Table 12 the error of implied Black-Scholes volatilities between the calibrated price computed using the Monte Carlo scheme and the market price. The mean absolute error is $0.19 \%$. Hence, the accuracy is improved by a factor 3.8 , which demonstrates the potential interest of using the time dependent Heston model. However, in this example, we observe that the time parameters vary greatly and the use of the Heston model (on these data) is questionable.

Computational time. Regarding the computational time, the approximation formula (2.13) yields essentially the same computational cost as the Black-Scholes formula, while the closed formula requires an additional space integration involving many exponential and trigonometric functions for which evaluation costs are higher. For instance, using a 2.6 GHz Pentium PC, the computations of the 64 numerical values in Table 3 (or 5) take 4.71 ms using the approximation formula and 301 ms using the closed formula. For the example with time

Table 8
Put prices of the closed formula, of the approximation formula, and of the averaging formula, expressed as a function of maturities in fractions of years and relative strikes. Piecewise constant parameters.

| 3 M | 30.00 | 20.07 | 10.78 | 3.93 | 0.87 | 0.13 | 0.05 | 0.02 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
|  | 30.00 | 20.08 | 10.78 | 3.93 | 0.87 | 0.13 | 0.05 | 0.02 |
|  | 30.00 | 20.07 | 10.78 | 3.93 | 0.87 | 0.13 | 0.05 | 0.02 |
| 6 M | 40.01 | 30.06 | 20.41 | 5.53 | 2.06 | 0.18 | 0.05 | 0.01 |
|  | 40.00 | 30.06 | 20.42 | 5.53 | 2.05 | 0.18 | 0.05 | 0.02 |
|  | 40.01 | 30.06 | 20.41 | 5.53 | 2.06 | 0.18 | 0.05 | 0.01 |
| 1 Y | 50.01 | 40.07 | 21.33 | 7.85 | 1.97 | 0.17 | 0.03 | 0.02 |
|  | 50.01 | 40.07 | 21.35 | 7.84 | 1.95 | 0.18 | 0.04 | 0.02 |
|  | 50.01 | 40.07 | 21.33 | 7.85 | 1.97 | 0.17 | 0.03 | 0.02 |
| 2 Y | 60.02 | 50.11 | 31.38 | 11.23 | 2.92 | 0.24 | 0.06 | 0.01 |
|  | 60.01 | 50.11 | 31.39 | 11.23 | 2.90 | 0.25 | 0.07 | 0.02 |
|  | 60.02 | 50.11 | 31.38 | 11.23 | 2.92 | 0.24 | 0.06 | 0.01 |
| 3 Y | 70.01 | 60.07 | 41.07 | 13.92 | 3.55 | 0.41 | 0.08 | 0.02 |
|  | 70.01 | 60.07 | 41.08 | 13.92 | 3.54 | 0.42 | 0.09 | 0.03 |
|  | 70.01 | 60.07 | 41.07 | 13.92 | 3.55 | 0.41 | 0.08 | 0.02 |
| 5 Y | 80.01 | 70.07 | 42.64 | 18.37 | 5.74 | 0.61 | 0.15 | 0.04 |
|  | 80.01 | 70.07 | 42.64 | 18.36 | 5.72 | 0.61 | 0.16 | 0.04 |
|  | 80.01 | 70.07 | 42.64 | 18.37 | 5.74 | 0.61 | 0.15 | 0.04 |
| 7 Y | 90.00 | 70.24 | 52.22 | 22.15 | 6.46 | 0.86 | 0.21 | 0.06 |
|  | 90.00 | 70.24 | 52.22 | 22.15 | 6.45 | 0.86 | 0.21 | 0.07 |
| 10 Y | 90.00 | 70.24 | 52.22 | 22.15 | 6.46 | 0.86 | 0.21 | 0.06 |
|  | 90.01 | 80.14 | 54.13 | 27.17 | 8.71 | 1.42 | 0.35 | 0.11 |
|  | 90.01 | 80.14 | 54.13 | 27.16 | 8.70 | 1.42 | 0.36 | 0.12 |
|  | 90.01 | 80.14 | 54.13 | 27.17 | 8.71 | 1.42 | 0.35 | 0.11 |

Table 9
Implied volatility surface for the SGP 500 Index (on June 16, 2008), expressed as a function of maturities in fractions of years and relative strikes.

| $T / K$ | $90.00 \%$ | $95.00 \%$ | $97.50 \%$ | $100.00 \%$ | $102.50 \%$ | $105.00 \%$ | $110.00 \%$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 3 M | $22.48 \%$ | $21.61 \%$ | $21.11 \%$ | $20.60 \%$ | $20.18 \%$ | $19.76 \%$ | $18.88 \%$ |
| 1Y | $24.62 \%$ | $23.25 \%$ | $22.57 \%$ | $21.87 \%$ | $21.22 \%$ | $20.60 \%$ | $19.30 \%$ |
| $1 \mathrm{Y}+6 \mathrm{M}$ | $24.66 \%$ | $23.51 \%$ | $22.95 \%$ | $22.38 \%$ | $21.80 \%$ | $21.29 \%$ | $20.22 \%$ |
| 2 Y | $24.83 \%$ | $23.81 \%$ | $23.33 \%$ | $22.84 \%$ | $22.34 \%$ | $21.86 \%$ | $20.93 \%$ |

dependent coefficients (reported in Table 7), the computational time for the 64 prices is about 40.2 ms using the approximation formula and 2574 ms using the closed formula. Roughly speaking, the use of the approximation formula enables us to speed up the valuation (and thus the calibration) by a factor 100 to 600 .
4. Proof of Theorem 2.3. The proof is divided into several steps. In subsection 4.1 we give the upper bounds for derivatives of the put function $P_{B S}$, in subsection 4.2 the conditions for positivity of the squared volatility process $v$, in subsection 4.3 the upper bounds for the negative moments of the integrated squared volatility $\int_{0}^{T} v_{t} \mathrm{~d} t$, and in subsection 4.4 the upper bounds for derivatives of functionals of the squared volatility process $v$. Finally, in subsection 4.5 , we complete the proof of Theorem 2.3 using the previous subsections.

Table 10
Error of implied volatility surface (in \%) when calibrating the constant Heston model.

| $T / K$ | $90.00 \%$ | $95.00 \%$ | $97.50 \%$ | $100.00 \%$ | $102.50 \%$ | $105.00 \%$ | 1.1 |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 3 M | $-1.71 \%$ | $-0.63 \%$ | $-0.26 \%$ | $-0.01 \%$ | $0.17 \%$ | $0.18 \%$ | $-0.36 \%$ |
| 1 Y | $1.27 \%$ | $1.05 \%$ | $0.88 \%$ | $0.64 \%$ | $0.40 \%$ | $0.12 \%$ | $-0.71 \%$ |
| $1 \mathrm{Y}+6 \mathrm{M}$ | $1.57 \%$ | $1.12 \%$ | $0.89 \%$ | $0.62 \%$ | $0.31 \%$ | $0.03 \%$ | $-0.68 \%$ |
| 2 Y | $2.12 \%$ | $1.42 \%$ | $1.10 \%$ | $0.76 \%$ | $0.39 \%$ | $0.04 \%$ | $-0.69 \%$ |

Table 11
The piecewise constant calibrated parameters.

| $T /$ Calibrated parameters | $\kappa$ | $\rho$ | $\xi$ | $v_{0}$ | $\theta$ |
| ---: | ---: | ---: | ---: | ---: | ---: |
| 3 M | $8.63 \%$ | $-99.30 \%$ | $15.63 \%$ | $3.49 \%$ | $83.35 \%$ |
| 1 Y | $8.63 \%$ | $-99.84 \%$ | $37.16 \%$ | $3.49 \%$ | $19.92 \%$ |
| $1 \mathrm{Y}+6 \mathrm{M}$ | $8.63 \%$ | $-10.65 \%$ | $16.85 \%$ | $3.49 \%$ | $15.83 \%$ |
| 2 Y | $8.63 \%$ | $40.91 \%$ | $0.13 \%$ | $3.49 \%$ | $53.15 \%$ |

Notation. In order to alleviate the proofs, we introduce some notation specific to this section.

Differentiation. For every process $Z^{\epsilon}$, we write (if these derivatives have a meaning)
(i) $Z_{i, t}=\left.\frac{\partial^{i} Z_{t}}{\partial \epsilon^{i}}\right|_{\epsilon=0}$,
(ii) the $i$ th Taylor residual by $R_{i, t}^{Z^{\epsilon}}=Z_{t}^{\epsilon}-\sum_{j=0}^{i} \frac{\epsilon^{j}}{j!} Z_{j, t}$.

Generic constants. We keep the same notation $C$ for all nonnegative constants
(i) depending on universal constants, on a number $p \geq 1$ arising in $L_{p}$ estimates, on $\theta_{\text {Inf }}, v_{0}$, and $K$;
 We write $A=O(B)$ when $|A| \leq C B$ for a generic constant.

## Miscellaneous.

(i) We write $\sigma_{t}^{\epsilon}=\sqrt{v_{t}^{\epsilon}}$ for the volatility for the perturbed process.
(ii) If $(Z)_{t \in[0, T]}$ is a càdlàg process, we denote by $Z^{*}$ its running extremum: $Z_{t}^{*}=$ $\sup _{s \leq t}\left|Z_{s}\right|$ for all $t \in[0, T]$.
(iii) The $L_{p}$ norm of a random variable is denoted, as usual, by $\|Z\|_{p}=\mathbb{E}\left[|Z|^{p}\right]^{1 / p}$.

### 4.1. Upper bounds for put derivatives.

Lemma 4.1. For every $(i, j) \in \mathbb{N}^{2}$, there exists a polynomial $P$ with positive coefficients such that

$$
\sup _{x \in \mathbb{R}}\left|\frac{\partial^{i+j} P_{B S}}{\partial x^{i} y^{j}}(x, y)\right| \leq \frac{P(\sqrt{y})}{y^{\frac{(2 j+i-1)_{+}}{2}}}
$$

Proof. Note that it is enough to prove the estimates for $j=0$, owing to the relation (2.2). We now take $j=0$. For $i=0$, the inequality holds because $P_{B S}$ is bounded. Thus consider

Table 12
Error of implied volatility surface (in \%) when calibrating the time dependent Heston model.

| $T / K$ | $90.00 \%$ | $95.00 \%$ | $97.50 \%$ | $100.00 \%$ | $102.50 \%$ | $105.00 \%$ | $110.00 \%$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 3 M | $0.41 \%$ | $0.24 \%$ | $0.13 \%$ | $0.03 \%$ | $0.04 \%$ | $0.05 \%$ | $0.07 \%$ |
| 1 Y | $-0.20 \%$ | $-0.25 \%$ | $-0.25 \%$ | $-0.25 \%$ | $-0.18 \%$ | $-0.07 \%$ | $0.16 \%$ |
| $1 \mathrm{Y}+6 \mathrm{M}$ | $-0.27 \%$ | $-0.18 \%$ | $-0.11 \%$ | $-0.05 \%$ | $0.02 \%$ | $0.17 \%$ | $0.46 \%$ |
| 2 Y | $-0.35 \%$ | $-0.32 \%$ | $-0.28 \%$ | $-0.25 \%$ | $-0.24 \%$ | $-0.20 \%$ | $-0.13 \%$ |

$i \geq 1$. Then, by differentiating the payoff, one gets

$$
\begin{aligned}
\frac{\partial^{i} P_{B S}}{\partial x^{i}}(x, y) & =\partial_{x}^{i} \mathbb{E}\left[e^{-\int_{0}^{T} r_{t} \mathrm{~d} t}\left(K-e^{x+\int_{0}^{T}\left(r_{t}-q_{t}\right) \mathrm{d} t-\frac{y}{2}+\sqrt{\frac{y}{T}} W_{T}}\right)_{+}\right] \\
& =-\partial_{x}^{i-1} \mathbb{E}\left[\mathbb{1}_{\left(e^{x+\int_{0}^{T}\left(r_{t}-q_{t}\right) \mathrm{d} t-\frac{y}{2}+\sqrt{\frac{y}{T}} W_{T}} \leq K\right)} e^{x-\int_{0}^{T} q_{t} \mathrm{~d} t-\frac{y}{2}+\sqrt{\frac{y}{T}} W_{T}}\right] \\
& =-\partial_{x}^{i-1} \mathbb{E}[\Psi(x+G)]
\end{aligned}
$$

where $\Psi$ is a bounded function (by $K$ ) and $G$ is a Gaussian variable with zero mean and variance equal to $y$. For such a function, we write $\mathbb{E}[\Psi(x+G)]=\int_{\mathbb{R}} \Psi(z) \frac{e^{-(z-x)^{2} /(2 y)}}{\sqrt{2 \pi y}} \mathrm{~d} z$, and from this it follows by a direct computation that

$$
\left|\partial_{x}^{i-1} \mathbb{E}[\Psi(x+G)]\right| \leq \frac{C}{y^{\frac{i-1}{2}}}
$$

for any $x$ and $y$. We have proved the estimate for $j=0$ and $i \geq 1$.
4.2. Positivity of the squared volatility process $v$. For a complete review related to time homogeneous CIR processes, we refer the reader to [20]. For time dependent CIR processes, see [30], where the existence and representation using squared Bessel processes are provided.

To prove the positivity of the process $v$, we show that it can be bounded from below by a suitable time homogeneous CIR process, time scale being the only difference (see Definition 5.1.2 in [39]). The arguments are quite standard, but since we need a specific statement that is not available in the literature, we detail the result and its proof. The time change $t \mapsto A_{t}$ is defined by

$$
t=\int_{0}^{A_{t}} \xi_{s}^{2} \mathrm{~d} s
$$

Because $\xi_{\text {Inf }}>0, A$ is a continuous, strictly increasing time change and its inverse $A^{-1}$ enjoys the same properties.

Lemma 4.2. Assume Assumption (P) and $v_{0}>0$. Denote by $\left(y_{s}\right)_{0 \leq s \leq A_{T}^{-1}}$ the CIR process defined by

$$
\mathrm{d} y_{t}=\left(\frac{1}{2}-\frac{\kappa}{\xi_{\text {Inf }}^{2}} y_{t}\right) \mathrm{d} t+\sqrt{y_{t}} d \tilde{B}_{t}, \quad y_{0}=v_{0}
$$

where $\tilde{B}$ is the Brownian motion given by

$$
\begin{equation*}
\tilde{B}_{t}=\int_{0}^{A_{T}} \xi_{s} \mathrm{~d} B_{s} \tag{4.1}
\end{equation*}
$$

Then, a.s. one has $v_{t} \geq y_{A_{t}^{-1}}$ for any $t \in[0, T]$. In particular, $\left(v_{t}\right)_{0 \leq t \leq T}$ is a.s. positive.
Proof. Note that $\left(\tilde{B}_{t}\right)_{0 \leq t \leq A_{T}^{-1}}$ is really a Brownian motion because by Lévy's characterization theorem it is a continuous local martingale with $\langle\tilde{B}, \tilde{B}\rangle_{t}=t$ (see Proposition 5.1.5 in [39] for the computation of the bracket). Now that we have set $\tilde{v}_{t}=v_{A_{t}}$, our aim is to prove that $\tilde{v}_{t} \geq y_{t}$ for $t \in\left[0, A_{T}^{-1}\right]$. Using Propositions 5.1.4 and 5.1.5 in [39], we write

$$
\tilde{v}_{t}=v_{0}+\int_{0}^{A_{t}}\left(\kappa\left(\theta_{s}-v_{s}\right) \mathrm{d} s+\xi_{s} \sqrt{v_{s}} \mathrm{~d} B_{s}\right)=v_{0}+\int_{0}^{t}\left(\frac{\kappa}{\xi_{A_{s}}^{2}}\left(\theta_{A_{s}}-\tilde{v}_{s}\right) \mathrm{d} s+\sqrt{\tilde{v}_{s}} \mathrm{~d} \tilde{B}_{s}\right) .
$$

Now we apply a comparison result for SDEs twice (see Proposition 5.2.18 in [24]).

1. First, one gets $\tilde{v}_{t} \geq n_{t}$, where $\left(n_{s}\right)_{s}$ is the (unique) solution of

$$
n_{t}=0+\int_{0}^{t}-\frac{\kappa}{\xi_{A_{s}}^{2}} n_{s} \mathrm{~d} s+\sqrt{n_{s}} \mathrm{~d} \tilde{B}_{s}
$$

because $v_{0} \geq 0$ and $\frac{\kappa}{\xi_{A_{s}}^{2}}\left(\theta_{A_{s}}-x\right) \geq-\frac{\kappa}{\xi_{A_{s}}^{2}} x$ for all $x \in \mathbb{R}$ and $s \in\left[0, A_{T}^{-1}\right]$. Of course $n_{t}=0$; thus $\tilde{v}_{t}$ is nonnegative.
2. Second, using the nonnegativity of $\tilde{v}$, we need only compare drift coefficients for the nonnegative variable $x$. Under Assumption (P), since

$$
\frac{\kappa}{\xi_{A_{s}}^{2}}\left(\theta_{A_{s}}-x\right) \geq \frac{1}{2}-\frac{\kappa}{\xi_{I n f}^{2}} x \quad \forall x \geq 0, \forall s \in\left[0, A_{T}^{-1}\right]
$$

we obtain $\tilde{v}_{t} \geq y_{t}$ for $t \in\left[0, A_{T}^{-1}\right]$ a.s.
Moreover, the positivity of $y$ (and consequently that of $v$ ) is standard: indeed, $y$ is a two-dimensional squared Bessel process with a time/space scale change (see [20] or the proof of Lemma 4.3).
4.3. Upper bound for negative moments of the integrated squared volatility process $\int_{0}^{T} v_{t} \mathrm{~d} t$.

Lemma 4.3. Assume Assumption (P). Then, for every $p>0$, one has

$$
\sup _{0 \leq \epsilon \leq 1} \mathbb{E}\left[\left(\int_{0}^{T} v_{t}^{\epsilon} \mathrm{d} t\right)^{-p}\right] \leq \frac{C}{T^{p}} .
$$

Before proving the result, we mention that analogous estimates appear in [12] (Lemmas A. 1 and A.2): some exponential moments are stated under stronger conditions than those in Assumption (P). In addition, the uniformity of the estimates w.r.t. $\xi$ (or equivalently w.r.t. $\epsilon$ ) is not emphasized. In our study, it is crucial to get uniform estimates w.r.t. $\epsilon$.

Proof. Fix $p \geq \frac{1}{2}$ (for $0<p<\frac{1}{2}$, we derive the result from the case $p=\frac{1}{2}$ using the Hölder inequality). The proof is divided into two steps. We first prove the estimates in the case of constant coefficients $\kappa, \theta$, and $\xi$ with $\kappa \theta=\frac{1}{2}, \epsilon=1$, and $\xi=1$. Then, using the time change of Lemma 4.2, we derive the result for $\left(v_{t}^{\epsilon}\right)_{t}$. The critical point is to get estimates that are uniform w.r.t. $\epsilon$.

Step 1. Take $\theta_{t} \equiv \theta, \xi_{t} \equiv 1, \kappa \theta=\frac{1}{2}, \epsilon=1$ and consider

$$
\mathrm{d} y_{t}=\left(\frac{1}{2}-\kappa y_{t}\right) \mathrm{d} t+\sqrt{y_{t}} \mathrm{~d} B_{t}, \quad y_{0}=v_{0}
$$

for a standard Brownian motion $B$. We represent $y$ as a time/space transformed squared Bessel process (see [20])

$$
y_{t}=e^{-\kappa t} z_{\frac{\left(e^{\kappa t}-1\right)}{4 \kappa}}
$$

where $z$ is a two-dimensional squared Bessel process. Therefore, using a change of variable and the explicit expression of the Laplace transform for the integral of $z$ (see [11, p. 377]), one obtains for any $u \geq 0$

$$
\begin{aligned}
\mathbb{E}\left[\exp \left(-u \int_{0}^{T} y_{t} \mathrm{~d} t\right)\right] & \leq \mathbb{E}\left[\exp \left(-4 u e^{-2 \kappa T} \int_{0}^{\frac{\left(e^{\kappa T}-1\right)}{4 \kappa}} z_{s} \mathrm{~d} s\right)\right] \\
& \leq \cosh \left(\frac{\sqrt{2 u}\left(1-e^{-\kappa T}\right)}{2 \kappa}\right)^{-1} \exp \left(-\sqrt{2 u} e^{-\kappa T} v_{0} \tanh \left(\frac{\sqrt{2 u}\left(1-e^{-\kappa T}\right)}{2 \kappa}\right)\right)
\end{aligned}
$$

Combining this with the identity $x^{-p}=\frac{1}{\Gamma(p)} \int_{0}^{\infty} u^{p-1} e^{-u x} \mathrm{~d} u$ for $x=\int_{0}^{T} y_{t} \mathrm{~d} t$, one gets

$$
\begin{aligned}
\mathbb{E}\left[\left(\int_{0}^{T} y_{t} \mathrm{~d} t\right)^{-p}\right] & \leq \frac{1}{\Gamma(p)} \int_{0}^{\infty} u^{p-1} \cosh \left(\frac{\sqrt{2 u}\left(1-e^{-\kappa T}\right)}{2 \kappa}\right)^{-1} \\
& \times \exp \left(-\sqrt{2 u} e^{-\kappa T} v_{0} \tanh \left(\frac{\sqrt{2 u}\left(1-e^{-\kappa T}\right)}{2 \kappa}\right)\right) \mathrm{d} u
\end{aligned}
$$

Define the parameter $\lambda^{2}=\frac{\left(e^{\kappa T}-1\right)}{2 \kappa v_{0}}$ and the new variable $n=\frac{\sqrt{2 u}\left(1-e^{-\kappa T}\right)}{2 \kappa}=v_{0} e^{-\kappa T} \lambda^{2} \sqrt{2 u}$. It readily follows that

$$
\mathbb{E}\left[\left(\int_{0}^{T} y_{t} \mathrm{~d} t\right)^{-p}\right] \leq C\left(\frac{e^{\kappa T}}{\lambda^{2}}\right)^{2 p} \int_{0}^{\infty} n^{2 p-1} \cosh (n)^{-1} \exp \left(-\frac{\tanh (n) n}{\lambda^{2}}\right) \mathrm{d} n
$$

where $C$ is a constant depending only on $v_{0}$ and $p$. We upper bound the above integral differently according to the value of $\lambda$.
(i) If $\lambda \geq 1$, then

$$
\begin{equation*}
\mathbb{E}\left[\left(\int_{0}^{T} y_{t} \mathrm{~d} t\right)^{-p}\right] \leq C\left(\frac{e^{\kappa T}}{\lambda^{2}}\right)^{2 p} \int_{0}^{\infty} n^{2 p-1} \cosh (n)^{-1} \mathrm{~d} n \leq C e^{2 p \kappa T} \tag{4.2}
\end{equation*}
$$

(ii) If $\lambda \leq 1$, split the integral into two parts, $n \leq \operatorname{arctanh}(\lambda)$ and $n \geq \operatorname{arctanh}(\lambda)$. For the first part, simply use $n \geq \tanh (n)$ for any $n$. For the second part, use $\tanh (n) \geq \lambda$ and
$\cosh (n)^{-1} \leq 1$. This gives

$$
\begin{aligned}
\mathbb{E}\left[\left(\int_{0}^{T} y_{t} \mathrm{~d} t\right)^{-p}\right] \leq C & {\left[\left(\frac{e^{\kappa T}}{\lambda^{2}}\right)^{2 p} \int_{0}^{\operatorname{arctanh}(\lambda)} n^{2 p-1} \cosh (n)^{-1} \exp \left(-\frac{\tanh ^{2}(n)}{\lambda^{2}}\right) \mathrm{d} n\right.} \\
& \left.+\left(\frac{e^{\kappa T}}{\lambda^{2}}\right)^{2 p} \int_{\operatorname{arctanh}(\lambda)}^{\infty} n^{2 p-1} \exp \left(-\frac{n}{\lambda}\right) \mathrm{d} n\right]:=C\left[\mathcal{T}_{1}+\mathcal{T}_{2}\right]
\end{aligned}
$$

We upper bound the two terms separately.

1. Term $\mathcal{T}_{1}$. Using the change of variable $m=\frac{\tanh (n)}{\lambda}$, one has

$$
\mathcal{T}_{1} \leq e^{2 p \kappa T} \lambda^{-4 p+1} \int_{0}^{1} \operatorname{arctanh}(\lambda m)^{2 p-1} \cosh (\operatorname{arctanh}(\lambda m)) \exp \left(-m^{2}\right) \mathrm{d} m
$$

Because of $\lambda \leq 1$, we have the following inequalities for $m \in[0,1[$ :

$$
\operatorname{arctanh}(\lambda m) \leq \lambda \operatorname{arctanh}(m), \quad \cosh (\operatorname{arctanh}(\lambda m)) \leq \cosh (\operatorname{arctanh}(m))
$$

Using $2 p-1 \geq 0$, it readily follows that

$$
\begin{equation*}
\mathcal{T}_{1} \leq\left(\frac{e^{2 \kappa T}}{\lambda^{2}}\right)^{p} \int_{0}^{1} \operatorname{arctanh}(m)^{2 p-1} \cosh (\operatorname{arctanh}(m)) \exp \left(-m^{2}\right) \mathrm{d} m \tag{4.4}
\end{equation*}
$$

2. Term $\mathcal{T}_{2}$. Clearly, we have

$$
\begin{equation*}
\mathcal{T}_{2} \leq\left(\frac{e^{\kappa T}}{\lambda^{2}}\right)^{2 p} \int_{0}^{\infty} n^{2 p-1} \exp \left(-\frac{n}{\lambda}\right) \mathrm{d} n=\left(\frac{e^{2 \kappa T}}{\lambda^{2}}\right)^{p} \int_{0}^{\infty} v^{2 p-1} e^{-v} \mathrm{~d} v \tag{4.5}
\end{equation*}
$$

Combining (4.3), (4.4), and (4.5), we obtain $\mathbb{E}\left[\left(\int_{0}^{T} y_{t} \mathrm{~d} t\right)^{-p}\right] \leq C\left(\frac{e^{2 \kappa T}}{\lambda^{2}}\right)^{p}$. In view of the inequality $\left(e^{x}-1 \geq x, x \geq 0\right)$, we have $\lambda^{2}=\frac{\left(e^{\kappa T}-1\right)}{2 \kappa v_{0}} \geq \frac{T}{2 v_{0}}$, which gives

$$
\begin{equation*}
\mathbb{E}\left[\left(\int_{0}^{T} y_{t} \mathrm{~d} t\right)^{-p}\right] \leq C \frac{e^{2 p \kappa T}}{T^{p}} \tag{4.6}
\end{equation*}
$$

available when $\lambda \leq 1$.
To sum up (4.2) and (4.6), we have proved that

$$
\begin{equation*}
\mathbb{E}\left[\left(\int_{0}^{T} y_{t} \mathrm{~d} t\right)^{-p}\right] \leq C e^{2 p \kappa T}\left(1+\frac{1}{T^{p}}\right) \tag{4.7}
\end{equation*}
$$

for a constant $C$ depending only on $p$ and $v_{0}$.
Step 2. Take $\epsilon \in] 0,1]$. We apply Lemma 4.2 to $v^{\epsilon}$, in order to write $v_{t}^{\epsilon} \geq y_{A_{\epsilon, t}}^{\epsilon-1}$, where $t=\int_{0}^{A_{\epsilon, t}}\left(\epsilon \xi_{s}\right)^{2} \mathrm{~d} s$ and $\mathrm{d} y_{t}^{\epsilon}=\left(\frac{1}{2}-\frac{\kappa}{\left(\epsilon \xi_{\text {Inf }}\right)^{2}} y_{t}^{\epsilon}\right) \mathrm{d} t+\sqrt{y_{t}^{\epsilon}} d \tilde{B}_{t}^{\epsilon}, y_{0}^{\epsilon}=y_{0}$. Thus, we get $\int_{0}^{T} v_{t}^{\epsilon} \mathrm{d} t \geq$
$\left(\int_{0}^{A_{\epsilon, t}^{-1}} y_{s}^{\epsilon} \mathrm{d} s\right) /\left(\epsilon \xi_{\text {Sup }}\right)^{2}$ and, in view of (4.7), it follows that

$$
\begin{aligned}
\mathbb{E}\left[\left(\int_{0}^{T} v_{t}^{\epsilon} \mathrm{d} t\right)^{-p}\right] & \leq\left(\epsilon \xi_{S u p}\right)^{2 p} \mathbb{E}\left(\int_{0}^{A_{\epsilon, T}^{-1}} y_{s}^{\epsilon} \mathrm{d} s\right)^{-p} \\
& \leq C\left(\epsilon \xi_{S u p}\right)^{2 p} e^{2 p \frac{\kappa}{\left(\epsilon \xi_{\text {Inf }}\right)^{2}} A_{\epsilon, T}^{-1}}\left(1+\frac{1}{\left[A_{\epsilon, T}^{-1}\right]^{p}}\right) \\
& \leq C e^{2 p \kappa \frac{\xi_{\text {Sup }}^{2}}{\xi_{\text {Inf }}^{2} T}}\left(\xi_{\text {Sup }}^{2 p}+\frac{\xi_{S u p}^{2 p}}{\xi_{\text {Inf }}^{2 p}} \frac{1}{T^{p}}\right)
\end{aligned}
$$

where we have used $\epsilon^{2} \xi_{\text {Inf }}^{2} T \leq A_{\epsilon, T}^{-1} \leq \epsilon^{2} \xi_{\text {Sup }}^{2} T$.
Note that the upper bound does not depend on $\epsilon \in] 0,1]$. For $\epsilon=0$, the upper bound in Lemma 4.3 is also true because $\left(v_{t}^{0}\right)_{t}$ is deterministic and

$$
\begin{equation*}
\max \left(v_{0}, \theta_{\text {Sup }}\right) \geq v_{t}^{0} \geq \min \left(v_{0}, \theta_{\text {Inf }}\right)>0 \tag{4.8}
\end{equation*}
$$

4.4. Upper bound for residuals of the Taylor development of $g(\epsilon)$ defined in (1.4). Throughout the following section, we assume that Assumption (P) is in force. We define the variables:

$$
P_{T}^{\epsilon}=\int_{0}^{T} \rho_{t}\left(\sigma_{t}^{\epsilon}-\sigma_{0, t}\right) \mathrm{d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2}\left(v_{t}^{\epsilon}-v_{0, t}\right) \mathrm{d} t, \quad Q_{T}^{\epsilon}=\int_{0}^{T}\left(1-\rho_{t}^{2}\right)\left(v_{t}^{\epsilon}-v_{0, t}\right) \mathrm{d} t
$$

Notice that $\left(x_{0}+\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2} v_{0, t} \mathrm{~d} t, \int_{0}^{T}\left(1-\rho_{t}^{2}\right) v_{0, t} \mathrm{~d} t\right)+\left(P_{T}^{1}, Q_{T}^{1}\right)=\left(x_{0}+\int_{0}^{T} \rho_{t} \sqrt{v_{t}^{1}} \mathrm{~d} B_{t}-\right.$ $\left.\int_{0}^{T} \frac{\rho_{t}^{2}}{2} v_{t}^{1} \mathrm{~d} t, \int_{0}^{T}\left(1-\rho_{t}^{2}\right) v_{t}^{1} \mathrm{~d} t\right)$.

The main result of this subsection is the following proposition, the statement of which uses the notation introduced at the beginning of section 4 .

Proposition 4.4. One has the following estimates for every $p \geq 1$ :

$$
\begin{aligned}
\left\|P_{T}^{1}\right\|_{p} & \leq C\left(\xi_{S u p} \sqrt{T}\right) \sqrt{T} \\
\left\|R_{2, T}^{P^{1}}\right\|_{p} & \leq C\left(\xi_{S u p} \sqrt{T}\right)^{3} \sqrt{T} \\
\left\|R_{2, T}^{\left(P^{1}\right)^{2}}\right\|_{p} & \leq C\left(\xi_{S u p} \sqrt{T}\right)^{3} T \\
\left\|Q_{T}^{1}\right\|_{p} & \leq C\left(\xi_{S u p} \sqrt{T}\right) T \\
\left\|R_{2, T}^{Q^{1}}\right\|_{p} & \leq C\left(\xi_{S u p} \sqrt{T}\right)^{3} T \\
\left\|R_{2, T}^{\left(Q^{1}\right)^{2}}\right\|_{p} & \leq C\left(\xi_{S u p} \sqrt{T}\right)^{3} T^{2} \\
\left\|R_{2, T}^{P^{1} Q^{1}}\right\|_{p} & \leq C\left(\xi_{S u p} \sqrt{T}\right)^{3} T^{\frac{3}{2}}
\end{aligned}
$$

To estimate the derivatives and the residuals for the variables $P_{T}^{\epsilon}$ and $Q_{T}^{\epsilon}$, we first need to prove the existence of the derivatives and the residuals of the volatility process $\sigma_{t}^{\epsilon}=\sqrt{v_{t}^{\epsilon}}$ and its square $v^{\epsilon}$. Finally we prove Proposition 4.4.
4.4.1. Upper bounds for derivatives of $\sigma^{\epsilon}$ and $\boldsymbol{v}^{\epsilon}$. Under Assumption ( P ), the volatility process $\sigma_{t}^{\epsilon}$ is governed by the SDE

$$
\begin{equation*}
d \sigma_{t}^{\epsilon}=\left(\left(\frac{\kappa \theta_{t}}{2}-\frac{\epsilon^{2} \xi_{t}^{2}}{8}\right) \frac{1}{\sigma_{t}^{\epsilon}}-\frac{\kappa}{2} \sigma_{t}^{\epsilon}\right) \mathrm{d} t+\frac{\epsilon \xi_{t}}{2} \mathrm{~d} B_{t}, \quad \sigma_{0}^{\epsilon}=\sqrt{v_{0}} \tag{4.9}
\end{equation*}
$$

where we have used Itô's lemma and the positivity of $v_{t}^{\epsilon}$ (see Lemma 4.2).
In order to estimate $R_{0, t}^{\sigma^{\epsilon}}$, we are going to prove that it verifies a linear equation (Lemma 4.5) from which we deduce an a priori upper bound (Proposition 4.6). We iterate the same analysis for the residuals $R_{1, t}^{\sigma^{\epsilon}}$ (Proposition 4.7) and $R_{2, t}^{\sigma^{\epsilon}}$ (Proposition 4.8). Analogously, we give upper bounds for the residuals of $v_{t}^{\epsilon}$ (Corollary 4.9).

Lemma 4.5. Under Assumption (P), the process $\left(R_{0, t}^{\sigma^{\epsilon}}=\sigma_{t}^{\epsilon}-\sigma_{t}^{0}\right)_{0 \leq t \leq T}$ is given by

$$
R_{0, t}^{\sigma^{\epsilon}}=U_{t}^{\epsilon} \int_{0}^{t}\left(U_{s}^{\epsilon}\right)^{-1}\left(-\frac{\epsilon^{2} \xi_{s}^{2}}{8 \sigma_{0, s}} \mathrm{~d} s+\frac{\epsilon \xi_{s}}{2} \mathrm{~d} B_{s}\right)
$$

where

$$
\begin{aligned}
\mathrm{d} U_{t}^{\epsilon} & =-\alpha_{t}^{\epsilon} U_{t}^{\epsilon} \mathrm{d} t, \quad U_{t}^{\epsilon}=1 \\
\alpha_{t}^{\epsilon} & =\left(\frac{\kappa \theta_{t}}{2}-\frac{\epsilon^{2} \xi_{t}^{2}}{8}\right) \frac{1}{\sigma_{t}^{\epsilon} \sigma_{0, t}}+\frac{\kappa}{2}
\end{aligned}
$$

Proof. From the definition $\left(\sigma_{0, t}\right)_{t}=\left(\sigma_{t}^{0}\right)_{t}$ and (4.9), one obtains the SDE

$$
d \sigma_{0, t}=\left(\frac{\kappa \theta_{t}}{2 \sigma_{0, t}}-\frac{\kappa}{2} \sigma_{0, t}\right) \mathrm{d} t, \quad \sigma_{0,0}=\sqrt{v_{0}} .
$$

Substitute this equation in (4.9) to obtain

$$
\begin{equation*}
\mathrm{d} R_{0, t}^{\sigma^{\epsilon}}=-\alpha_{t}^{\epsilon} R_{0, t}^{\sigma^{\epsilon}} \mathrm{d} t-\frac{\epsilon^{2} \xi_{t}^{2}}{8 \sigma_{0, t}} \mathrm{~d} t+\frac{\epsilon \xi_{t}}{2} \mathrm{~d} B_{t}, \quad R_{0,0}^{\sigma^{\epsilon}}=0 \tag{4.10}
\end{equation*}
$$

Note that $R_{0, \text {, }}^{\sigma^{\epsilon}}$ is the solution of a linear SDE. Hence, it can be explicitly represented using the process $U^{\epsilon}$ (see Theorem 52 in [37]):

$$
R_{0, t}^{\sigma^{\epsilon}}=U_{t}^{\epsilon} \int_{0}^{t}\left(U_{s}^{\epsilon}\right)^{-1}\left(-\frac{\epsilon^{2} \xi_{s}^{2}}{8 \sigma_{0, s}} \mathrm{~d} s+\frac{\epsilon \xi_{s}}{2} \mathrm{~d} B_{s}\right)
$$

Proposition 4.6. Under Assumption (P), for every $p \geq 1$ one has

$$
\left\|\left(R_{0, .,}^{\sigma^{\epsilon}}\right)_{t}^{*}\right\|_{p} \leq C \epsilon \xi_{S u p} \sqrt{t}
$$

In particular, the application $\epsilon \rightarrow \sigma_{t}^{\epsilon}$ is continuous ${ }^{6}$ at $\epsilon=0$ in $L_{p}$.

[^26]Proof. At first sight, the proof seems to be straightforward from Lemma 4.5. But actually, the difficulty lies in the fact that one cannot uniformly in $\epsilon$ upper bound $U_{t}^{\epsilon}$ in $L_{p}$ (because of the term with $1 / \sigma_{t}^{\epsilon}$ in $\alpha_{t}^{\epsilon}$ ).

Using Lemma 4.5 and Itô's formula for the product $\left(U_{t}^{\epsilon}\right)^{-1}\left(\int_{0}^{t} \frac{\epsilon \xi_{s}}{2} \mathrm{~d} B_{s}\right)$, one has

$$
R_{0, t}^{\sigma^{\epsilon}}=U_{t}^{\epsilon} \int_{0}^{t}\left(U_{s}^{\epsilon}\right)^{-1}\left(-\frac{\epsilon^{2} \xi_{s}^{2}}{8 \sigma_{0, s}}\right) \mathrm{d} s+\int_{0}^{t} \frac{\epsilon \xi_{s}}{2} \mathrm{~d} B_{s}-U_{t}^{\epsilon} \int_{0}^{t}\left(\int_{0}^{s} \frac{\epsilon \xi_{u}}{2} \mathrm{~d} B_{u}\right) d\left(U_{s}^{\epsilon}\right)^{-1}
$$

Under Assumption (P), one has $\alpha_{t}^{\epsilon} \geq \kappa / 2>0$, which implies that $t \mapsto U_{t}^{\epsilon}$ is decreasing and $t \mapsto\left(U_{t}^{\epsilon}\right)^{-1}$ is increasing. Thus, $0 \leq U_{t}^{\epsilon}\left(U_{s}^{\epsilon}\right)^{-1} \leq 1$ for $s \in[0, t]$. Consequently, we deduce

$$
\begin{align*}
\left|R_{0, t}^{\sigma^{\epsilon}}\right| & \leq \int_{0}^{t} \frac{\epsilon^{2} \xi_{s}^{2}}{8 \sigma_{0, s}} \mathrm{~d} s+\left(\int_{0} \frac{\epsilon \xi_{s}}{2} \mathrm{~d} B_{s}\right)_{t}^{*}+\left(\int_{0} \frac{\epsilon \xi_{s}}{2} \mathrm{~d} B_{s}\right)_{t}^{*}\left(1-U_{t}^{\epsilon}\right) \\
& \leq \int_{0}^{t} \frac{\epsilon^{2} \xi_{s}^{2}}{8 \sigma_{0, s}} \mathrm{~d} s+\left(\int_{0} \epsilon \xi_{s} \mathrm{~d} B_{s}\right)_{t}^{*} \tag{4.11}
\end{align*}
$$

Now we easily complete the proof by observing that $\sigma_{0, s} \geq \min \left(\sqrt{\theta_{I n f}}, \sqrt{v_{0}}\right)$ and $\left\|\left(\int_{0} \xi_{s} \mathrm{~d} B_{s}\right)_{t}^{*}\right\|_{p}$ $\leq C \xi_{\text {Sup }} \sqrt{t}$.

We define $\left(\sigma_{1, t}\right)_{0 \leq t \leq T}$ as the solution of the linear equation (4.12), which is obtained by taking derivatives w.r.t. $\epsilon$ of (4.9) and setting $\epsilon$ equal to zero thereafter:

$$
\begin{equation*}
d \sigma_{1, t}=-\left(\frac{\kappa \theta_{t}}{2\left(\sigma_{0, t}\right)^{2}}+\frac{\kappa}{2}\right) \sigma_{1, t} \mathrm{~d} t+\frac{\xi_{t}}{2} \mathrm{~d} B_{t}, \quad \sigma_{1,0}=0 \tag{4.12}
\end{equation*}
$$

The solution of this SDE is obtained similarly as was done with (4.10) and is given by the equation

$$
\sigma_{1, t}=U_{t}^{0} \int_{0}^{t}\left(U_{s}^{0}\right)^{-1} \frac{\xi_{s}}{2} \mathrm{~d} B_{s}
$$

For every $p \geq 1$, taking into account that $0 \leq U_{t}^{0}\left(U_{s}^{0}\right)^{-1} \leq 1$ for $s \in[0, t]$, the upper bound is given by (4.13) (see now that there is a uniform bound w.r.t. $\epsilon$ ):

$$
\begin{equation*}
\left\|\left(\sigma_{1, .}\right)_{t}^{*}\right\|_{p} \leq C \xi_{S u p} \sqrt{t} \tag{4.13}
\end{equation*}
$$

Proposition 4.7. Under Assumption (P), the process $\left(R_{1, t}^{\sigma^{\epsilon}}=\sigma_{t}^{\epsilon}-\sigma_{t}^{0}-\epsilon \sigma_{1, t}\right)_{0 \leq t \leq T}$ fulfills the equality

$$
R_{1, t}^{\sigma^{\epsilon}}=U_{t}^{\epsilon} \int_{0}^{t}\left(U_{s}^{\epsilon}\right)^{-1}\left(-\frac{\epsilon^{2} \xi_{s}^{2}}{8 \sigma_{0, s}}+\epsilon \sigma_{1, s}\left(\left(\frac{\alpha_{s}^{\epsilon}}{\sigma_{0, s}}-\frac{\kappa}{2 \sigma_{0, s}}\right) R_{0, s}^{\sigma^{\epsilon}}+\frac{\epsilon^{2} \xi_{s}^{2}}{8 \sigma_{0, s}}\right)\right) \mathrm{d} s
$$

Moreover, for every $p \geq 1$, one has

$$
\left\|\left(R_{1, .}^{\sigma^{\epsilon}}\right)_{t}^{*}\right\|_{p} \leq C\left(\epsilon \xi_{S u p} \sqrt{t}\right)^{2}
$$

In particular, the application $\epsilon \rightarrow \sigma_{t}^{\epsilon}$ is $\mathcal{C}^{1}$ at $\epsilon=0$ in the $L_{p}$ sense with the first derivative at $\epsilon=0$ equal to $\sigma_{1, t}$ (justifying a posteriori the definition $R_{1, .}^{\sigma^{\epsilon}}$ ).

Proof. From (4.10) and (4.12), it readily follows that

$$
\mathrm{d} R_{1, t}^{\sigma^{\epsilon}}=-\alpha_{t}^{\epsilon} R_{1, t}^{\sigma^{\epsilon}} \mathrm{d} t-\epsilon \sigma_{1, t}\left(\alpha_{t}^{\epsilon}-\frac{\kappa \theta_{t}}{2\left(\sigma_{0, t}\right)^{2}}-\frac{\kappa}{2}\right) \mathrm{d} t-\frac{\epsilon^{2} \xi_{t}^{2}}{8 \sigma_{0, t}} \mathrm{~d} t, \quad R_{1,0}^{\sigma^{\epsilon}}=0
$$

Because of the identity

$$
-\left(\alpha_{t}^{\epsilon}-\frac{\kappa \theta_{t}}{2\left(\sigma_{0, t}\right)^{2}}-\frac{\kappa}{2}\right)=\left(\left(\frac{\alpha_{t}^{\epsilon}}{\sigma_{0, t}}-\frac{\kappa}{2 \sigma_{0, t}}\right) R_{0, t}^{\sigma^{\epsilon}}+\frac{\epsilon^{2} \xi_{t}^{2}}{8\left(\sigma_{0, t}\right)^{2}}\right),
$$

one deduces the equality

$$
R_{1, t}^{\sigma^{\epsilon}}=U_{t}^{\epsilon} \int_{0}^{t}\left(U_{s}^{\epsilon}\right)^{-1}\left(-\frac{\epsilon^{2} \xi_{s}^{2}}{8 \sigma_{0, s}}+\epsilon \sigma_{1, s}\left(\left(\frac{\alpha_{s}^{\epsilon}}{\sigma_{0, s}}-\frac{\kappa}{2 \sigma_{0, s}}\right) R_{0, s}^{\sigma^{\epsilon}}+\frac{\epsilon^{2} \xi_{s}^{2}}{8\left(\sigma_{0, s}\right)^{2}}\right)\right) \mathrm{d} s
$$

Then

$$
\begin{aligned}
\left|R_{1, t}^{\sigma^{\epsilon}}\right| \leq & \int_{0}^{t} U_{t}^{\epsilon}\left(U_{s}^{\epsilon}\right)^{-1}\left(\frac{\epsilon^{2} \xi_{s}^{2}}{8 \sigma_{0, s}}+\epsilon\left|\sigma_{1, s}\right|\left(\left(\frac{\alpha_{s}^{\epsilon}}{\sigma_{0, s}}+\frac{\kappa}{2 \sigma_{0, s}}\right)\left|R_{0, s}^{\sigma^{\epsilon}}\right|+\frac{\epsilon^{2} \xi_{s}^{2}}{8\left(\sigma_{0, s}\right)^{2}}\right)\right) \mathrm{d} s \\
\leq & \int_{0}^{t} U_{t}^{\epsilon}\left(U_{s}^{\epsilon}\right)^{-1}\left(\frac{\epsilon^{2} \xi_{s}^{2}}{8 \sigma_{0, s}}+\epsilon\left|\sigma_{1, s}\right|\left(\frac{\kappa}{2 \sigma_{0, s}}\left|R_{0, s}^{\sigma^{\epsilon}}\right|+\frac{\epsilon^{2} \xi_{s}^{2}}{8\left(\sigma_{0, s}\right)^{2}}\right)\right) \mathrm{d} s \\
& \left.+\epsilon \int_{0}^{t} U_{t}^{\epsilon}\left(U_{s}^{\epsilon}\right)^{-1} \frac{\alpha_{s}^{\epsilon}}{\sigma_{0, s}}\left|\sigma_{1, s}\right| \| R_{0, s}^{\sigma^{\epsilon}} \right\rvert\, \mathrm{d} s \\
\leq & \int_{0}^{t}\left(\frac{\epsilon^{2} \xi_{s}^{2}}{8 \sigma_{0, s}}+\epsilon\left|\sigma_{1, s}\right|\left(\frac{\kappa}{2 \sigma_{0, s}}\left|R_{0, s}^{\epsilon^{\epsilon}}\right|+\frac{\epsilon^{2} \xi_{s}^{2}}{8\left(\sigma_{0, s}\right)^{2}}\right)\right) \mathrm{d} s+\epsilon\left(\frac{\sigma_{1, .} R_{0, .}^{\sigma^{\epsilon}}}{\sigma_{0, .}}\right)_{t}^{*}
\end{aligned}
$$

where we have used $U_{t}^{\epsilon}\left(U_{s}^{\epsilon}\right)^{-1} \leq 1$ for every $s \in[0, t]$ and $U_{t}^{\epsilon} \int_{0}^{t} \alpha_{s}^{\epsilon}\left(U_{s}^{\epsilon}\right)^{-1} \mathrm{~d} s=1-U_{t}^{\epsilon} \leq 1$ for the third inequality. Apply Proposition 4.6 and inequality (4.13) to complete the proof of the estimate of $\left\|\left(R_{1, .}^{\sigma^{\epsilon}}\right)_{t}^{*}\right\|_{p}$.

We define $\left(\sigma_{2, t}\right)_{0 \leq t \leq T}$ as the solution of the linear equation (4.14), which is obtained by differentiating twice (4.9) w.r.t. $\epsilon$ and setting $\epsilon$ equal to zero:

$$
\begin{equation*}
d \sigma_{2, t}=\left(-\left(\frac{\kappa \theta_{t}}{2\left(\sigma_{0, t}\right)^{2}}+\frac{\kappa}{2}\right) \sigma_{2, t}+\kappa \theta_{t} \frac{\left(\sigma_{1, t}\right)^{2}}{\left(\sigma_{0, t}\right)^{3}}-\frac{\xi_{t}^{2}}{4 \sigma_{0, t}}\right) \mathrm{d} t, \quad \sigma_{2,0}=0 . \tag{4.14}
\end{equation*}
$$

Clearly, for $p \geq 1$, we have

$$
\begin{equation*}
\left\|\left(\sigma_{2, .}\right)_{t}^{*}\right\|_{p} \leq C\left(\xi_{\text {Sup }} \sqrt{t}\right)^{2} \tag{4.15}
\end{equation*}
$$

Proposition 4.8. Under Assumption (P), the process $\left(R_{2, t}^{\sigma^{\epsilon}}=\sigma_{t}^{\epsilon}-\sigma_{t}^{0}-\epsilon \sigma_{1, t}-\frac{\epsilon^{2}}{2} \sigma_{2, t}\right)_{0 \leq t \leq T}$ fulfills the equality

$$
\begin{aligned}
R_{2, t}^{\sigma^{\epsilon}}=U_{t}^{\epsilon} \int_{0}^{t}\left(U_{s}^{\epsilon}\right)^{-1} & {\left[\epsilon^{2}\left(\left(\frac{\alpha_{s}^{\epsilon}}{\sigma_{0, s}}-\frac{\kappa}{2 \sigma_{0, s}}\right) R_{0, s}^{\sigma^{\epsilon}}+\frac{\epsilon^{2} \xi_{s}^{2}}{8\left(\sigma_{0, s}\right)^{2}}\right)\left(\frac{\sigma_{2, s}}{2}-\frac{\left(\sigma_{1, s}\right)^{2}}{\sigma_{0, s}}\right)\right.} \\
& \left.+\epsilon\left(\left(\frac{\alpha_{s}^{\epsilon}}{\sigma_{0, s}}-\frac{\kappa}{2 \sigma_{0, s}}\right) R_{1, s}^{\sigma^{\epsilon}}+\frac{\epsilon^{2} \xi_{s}^{2}}{8\left(\sigma_{0, s}\right)^{2}}\right) \sigma_{1, s}\right] \mathrm{d} s
\end{aligned}
$$

Moreover, for every $p \geq 1$, one has

$$
\left\|\left(R_{2, .}^{\sigma^{\epsilon}}\right)_{t}^{*}\right\|_{p} \leq C\left(\epsilon \xi_{\text {Sup }} \sqrt{t}\right)^{3} .
$$

In particular, the application $\epsilon \rightarrow \sigma_{t}^{\epsilon}$ is $\mathcal{C}^{2}$ at $\epsilon=0$ in the $L_{p}$ sense with the second derivative at $\epsilon=0$ equal to $\sigma_{2, t}$.

Proof. The equality is easy to check. The estimate is proved in the same way as in the proof of Proposition 4.7; we therefore skip the details.

Corollary 4.9. The application $\epsilon \rightarrow v_{t}^{\epsilon}$ is $\mathcal{C}^{2}$ at $\epsilon=0$ in the $L_{p}$ sense. The residuals for the squared volatility satisfy the following inequalities: for every $p \geq 1$, one has

$$
\begin{aligned}
& \left\|\left(R_{0, .,}^{v^{\epsilon}}\right)_{t}\right\|_{p} \leq C \epsilon \xi_{\text {Sup }} \sqrt{t}, \\
& \left\|\left(R_{1, .,}^{v^{\epsilon}}\right)_{t}^{*}\right\|_{p} \leq C\left(\epsilon \xi_{\text {Sup }} \sqrt{t}\right)^{2}, \\
& \left\|\left(R_{2, .}^{v^{\epsilon}}\right)_{t}^{*}\right\|_{p} \leq C\left(\epsilon \xi_{\text {Sup }} \sqrt{t}\right)^{3} .
\end{aligned}
$$

Proof. Note that $v_{t}^{\epsilon}=\left(\sigma_{t}^{\epsilon}\right)^{2}=\left(\sigma_{0, t}+R_{0, t}^{\sigma^{\epsilon}}\right)^{2}=v_{0, t}+2 \sigma_{0, t} R_{0, t}^{\sigma^{\epsilon}}+\left(R_{0, t}^{\sigma^{\epsilon}}\right)^{2}$. Thus, we have $R_{0, t}^{v^{\epsilon}}=2 \sigma_{0, t} R_{0, t}^{\sigma^{\epsilon}}+\left(R_{0, t}^{\sigma^{\epsilon}}\right)^{2}$, which leads to the required estimate using $\sigma_{0, t} \leq \max \left(\sqrt{v_{0}}, \sqrt{\theta_{\text {Sup }}}\right)$ and Proposition 4.6. The other estimates are proved analogously using Propositions 4.7 and 4.8 and inequalities (4.13) and (4.15).
4.4.2. Proof of Proposition 4.4. We can write

$$
P_{T}^{1}=\int_{0}^{T} \rho_{t} R_{0, t}^{\sigma^{1}} \mathrm{~d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2} R_{0, t}^{v^{1}} \mathrm{~d} t, \quad R_{2, T}^{P^{1}}=\int_{0}^{T} \rho_{t} R_{2, t}^{\sigma^{1}} \mathrm{~d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2} R_{2, t}^{v^{1}} \mathrm{~d} t .
$$

Then, using Propositions 4.6 and 4.8 and Corollary 4.9, we prove the two first estimates of Proposition 4.4. The others inequalities are proved in the same way.
4.5. Proof of Theorem 2.3. For convenience, we introduce the following notation for $\lambda \in[0,1]:$

$$
\begin{aligned}
& \bar{P}_{B S}(\lambda)=P_{B S}\left(x_{0}+\right. \int_{0}^{T} \rho_{t}\left((1-\lambda) \sqrt{v_{0, t}}+\lambda \sqrt{v_{t}^{1}}\right) \mathrm{d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2}\left((1-\lambda) v_{0, t}+\lambda v_{t}^{1}\right) \mathrm{d} t, \\
&\left.\int_{0}^{T}\left(1-\rho_{t}^{2}\right)\left((1-\lambda) v_{0, t}+\lambda v_{t}^{1}\right) \mathrm{d} t\right), \\
& \frac{\partial^{i+j} \bar{P}_{B S}}{\partial x^{i} y^{j}}(\lambda)=\frac{\partial^{i+j} P_{B S}}{\partial x^{i} y^{j}}\left(x_{0}+\int_{0}^{T} \rho_{t}\left((1-\lambda) \sqrt{v_{0, t}}+\lambda \sqrt{v_{t}^{1}}\right) \mathrm{d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2}\left((1-\lambda) v_{0, t}+\lambda v_{t}^{1}\right) \mathrm{d} t,\right. \\
&\left.\int_{0}^{T}\left(1-\rho_{t}^{2}\right)\left((1-\lambda) v_{0, t}+\lambda v_{t}^{1}\right) \mathrm{d} t\right) .
\end{aligned}
$$

Notice that $\tilde{P}_{B S}$ (see (2.5)) is a particular case of $\bar{P}_{B S}$ for $\lambda=0$ :

$$
\tilde{P}_{B S}=\bar{P}_{B S}(0), \quad \frac{\partial^{i+j} \tilde{P}_{B S}}{\partial x^{i} y^{j}}=\frac{\partial^{i+j} \bar{P}_{B S}}{\partial x^{i} y^{j}}(0) .
$$

Now, we represent the error $\mathcal{E}$ in (2.12) using the previous notation. A second order Taylor expansion leads to

$$
g(1)=\mathbb{E}\left(\bar{P}_{B S}(1)\right)=\mathbb{E}\left(\bar{P}_{B S}(0)+\partial_{\lambda} \bar{P}_{B S}(0)+\frac{1}{2} \partial_{\lambda}^{2} \bar{P}_{B S}(0)+\int_{0}^{1} \mathrm{~d} \lambda \frac{(1-\lambda)^{2}}{2} \partial_{\lambda}^{3} \bar{P}_{B S}(\lambda)\right) .
$$

The first term $\mathbb{E}\left(\bar{P}_{B S}(0)\right)$ is equal to (2.6). Approximations of the three above derivatives contribute to the error $\mathcal{E}$.

1. We have $\mathbb{E}\left(\partial_{\lambda} \bar{P}_{B S}(0)\right)=\mathbb{E}\left(\frac{\partial \tilde{P}_{B S}}{\partial x} P_{T}^{1}+\frac{\partial \tilde{P}_{B S}}{\partial y} Q_{T}^{1}\right)$. These two terms are equal to (2.7) and (2.8) plus an error equal to

$$
\mathbb{E}\left(\frac{\partial \tilde{P}_{B S}}{\partial x} R_{2, T}^{P^{1}}+\frac{\partial \tilde{P}_{B S}}{\partial y} R_{2, T}^{Q^{1}}\right) .
$$

2. W.r.t. the second derivatives, we have $\mathbb{E}\left(\frac{1}{2} \partial_{\lambda}^{2} \bar{P}_{B S}(0)\right)=\mathbb{E}\left(\frac{1}{2} \frac{\partial^{2} \tilde{P}_{B S}}{\partial x^{2}}\left(P_{T}^{1}\right)^{2}+\frac{1}{2} \frac{\partial^{2} \tilde{P}_{B S}}{\partial y^{2}}\left(Q_{T}^{1}\right)^{2}+\right.$ $\left.\frac{\partial^{2} \tilde{P}_{B S}}{\partial x y} P_{T}^{1} Q_{T}^{1}\right)$. These terms are equal to (2.9), (2.10), and (2.11) plus an error equal to

$$
\mathbb{E}\left(\frac{1}{2} \frac{\partial^{2} \tilde{P}_{B S}}{\partial x^{2}} R_{2, T}^{\left(P^{1}\right)^{2}}+\frac{1}{2} \frac{\partial^{2} \tilde{P}_{B S}}{\partial y^{2}} R_{2, T}^{\left(Q^{1}\right)^{2}}+\frac{\partial^{2} \tilde{P}_{B S}}{\partial x y} R_{2, T}^{P^{1} Q^{1}}\right)
$$

3. The last term with $\partial_{\lambda}^{3} \bar{P}_{B S}$ is neglected and thus is considered as an error. To sum up, we have shown that

$$
\begin{aligned}
\mathcal{E}= & \sum_{i=0}^{1} \mathbb{E}\left[\frac{\partial^{1} \bar{P}_{B S}}{\partial x^{i} y^{1-i}}(0) R_{2, T}^{\left(P^{1}\right)^{i}\left(Q^{1}\right)^{1-i}}\right]+\sum_{i=0}^{2} \frac{C_{2}^{i}}{2} \mathbb{E}\left[\frac{\partial^{2} \bar{P}_{B S}}{\partial x^{i} y^{2-i}}(0) R_{2, T}^{\left(P^{1}\right)^{i}\left(Q^{1}\right)^{2-i}}\right] \\
& +\int_{0}^{1} \frac{(1-\lambda)^{2}}{2} \sum_{i=0}^{3} C_{3}^{i} \mathbb{E}\left[\frac{\partial^{3} \bar{P}_{B S}}{\partial x^{i} y^{3-i}}(\lambda)\left(P_{T}^{1}\right)^{i}\left(Q_{T}^{1}\right)^{3-i}\right] \mathrm{d} \lambda .
\end{aligned}
$$

Using Lemma 4.1 and Assumption ( R ), one obtains for all $\lambda \in[0,1]$

$$
\begin{aligned}
\left\|\frac{\partial^{i+j} \bar{P}_{B S}}{\partial x^{i} y^{j}}(\lambda)\right\|_{2} & \leq C\left\|\left(\int_{0}^{T}\left((1-\lambda) v_{0, t}+\lambda v_{t}^{1}\right) \mathrm{d} t\right)^{\frac{-(2 j+i-1)+}{2}}\right\|_{4} \\
& \leq C\left((1-\lambda)\left\|\left(\int_{0}^{T} v_{0, t} \mathrm{~d} t\right)^{\frac{-(2 j+i-1)+}{2}}\right\|_{4}+\lambda\left\|\left(\int_{0}^{T} v_{t}^{1} \mathrm{~d} t\right)^{\frac{-(2 j+i-1)+}{2}}\right\|_{4}\right),
\end{aligned}
$$

where we have applied a convexity argument. Finally, apply Lemma 4.3 with $\epsilon=0$ and $\epsilon=1$ to conclude that

$$
\left\|\frac{\partial^{i+j} \bar{P}_{B S}}{\partial x^{i} y^{j}}(\lambda)\right\|_{2} \leq \frac{C}{(\sqrt{T})^{(2 j+i-1)_{+}}}
$$

uniformly w.r.t. $\lambda \in[0,1]$. Combining this with Proposition 4.4 yields that

$$
\begin{aligned}
|\mathcal{E}| & \leq C\left(\sum_{i=0}^{1}\left(\xi_{\text {Sup }} \sqrt{T}\right)^{3} \frac{T^{1-i / 2}}{(\sqrt{T})^{1-i}}+\sum_{i=0}^{2}\left(\xi_{\text {Sup }} \sqrt{T}\right)^{3} \frac{T^{2-i / 2}}{(\sqrt{T})^{3-i}}+\sum_{i=0}^{3}\left(\xi_{\text {Sup }} \sqrt{T}\right)^{3} \frac{T^{3-i / 2}}{(\sqrt{T})^{5-i}}\right) \\
& \leq C \xi_{\text {Sup }}^{3} T^{2} .
\end{aligned}
$$

Theorem 2.3 is proved.

## 5. Proofs of Proposition 2.1 and Theorem 2.2.

5.1. Definitions. In order to make the approximation explicit, we introduce the following family of operators indexed by maturity $T$.

Definition 5.1 (integral operator). We define the integral operator $\omega_{., T}^{(., .)}$as follows:
(i) For any real number $k$ and any integrable function $l$, we set

$$
\omega_{t, T}^{(k, l)}=\int_{t}^{T} e^{k u} l_{u} \mathrm{~d} u \quad \forall t \in[0, T] .
$$

(ii) For any real numbers $\left(k_{1}, \ldots, k_{n}\right)$ and for any integrable functions $\left(l_{1}, \ldots, l_{n}\right)$, the $n$-times iteration is given by

$$
\omega_{t, T}^{\left(k_{1}, l_{1}\right), \ldots,\left(k_{n}, l_{n}\right)}=\omega_{t, T}^{\left(k_{1}, l_{1} \omega_{, T}^{\left(k_{2}, l_{2}\right), \ldots,\left(k_{n}, l_{n}\right)}\right)} \quad \forall t \in[0, T] .
$$

(iii) When the functions $\left(l_{1}, \ldots, l_{n}\right)$ are equal to the unity constant function 1 , we simply write

$$
\tilde{\omega}_{t, T}^{k_{1}, \ldots, k_{n}}=\omega_{t, T}^{\left(k_{1}, 1\right), \ldots,\left(k_{n}, 1\right)} \quad \forall t \in[0, T] .
$$

Note that with this short notation used in Theorem 2.2, we have $a_{1, T}=\omega_{0, T}^{\left(\kappa, \rho \xi v_{0 . .}\right),(-\kappa, 1)}$, $a_{2, T}=\omega_{0, T}^{\left(\kappa, \rho \xi v_{0, .}\right),(0, \rho \xi),(-\kappa, 1)}$, and $b_{0, T}=\omega_{0, T}^{\left(2 \kappa, \xi^{2} v_{0,},\right),(-\kappa, 1),(-\kappa, 1)}$.
5.2. Preliminary results. In this section, we bring together the results (and their proofs) which allow us to derive the explicit terms in the formula (2.13). Before stating the lemmas, we give some guidance about the way to express the correction terms as derivatives of the leading price (2.13):

- Lemma 5.2 is the equivalent for random variables of the integration by parts formula. It allows expressing the expectation of an Itô integral multiplied by a random variable.
This will lead to taking derivatives of $\tilde{P}_{B S}$ (the objective of the reduction).
- Lemma 5.3 is a simpler version of Lemma 5.2 in order to simplify some calculus.
- Lemma 5.4 is a simple application of the integration by parts formula w.r.t. time.
- Lemma 5.5 computes the expectation of integrals of derivatives w.r.t. $\epsilon$ of the stochastic process $v_{t}$ in terms of the expectation of the derivatives of $\tilde{P}_{B S}$.
- Lemma 5.6 computes the expectation of the derivatives of $\tilde{P}_{B S}$.

In the following, $\alpha_{t}$ (resp., $\beta_{t}$ ) is a square integrable and predictable process (resp., deterministic) and $l$ is a smooth function with derivatives having, at most, exponential growth.

For the next Malliavin calculus computations, we freely use standard notation from [33].

Lemma 5.2 (see [33, Lemma 1.2.1]). Let $G \in \mathbb{D}^{1, \infty}(\Omega)$. One has

$$
\mathbb{E}\left[G \int_{0}^{t} \alpha_{s} \mathrm{~d} B_{s}\right]=\mathbb{E}\left[\int_{0}^{t} \alpha_{s} D_{s}^{B}(G) \mathrm{d} s\right]
$$

where $D^{B}(G)=\left(D_{s}^{B}(G)\right)_{s \geq 0}$ is the first Malliavin derivative of $G$ w.r.t. B.
Taking $G=l\left(\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}\right)$ gives $D_{s}^{B}(G)=l^{(1)}\left(\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}\right) \rho_{s} \sqrt{v_{0, s}} \mathbf{1}_{s \leq T}$. Hence, we obtain immediately the following result.

Lemma 5.3. One has

$$
\mathbb{E}\left[\left(\int_{0}^{T} \alpha_{t} \mathrm{~d} B_{t}\right) l\left(\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}\right)\right]=\mathbb{E}\left[\left(\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \alpha_{t} \mathrm{~d} t\right) l^{(1)}\left(\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}\right)\right] .
$$

Lemma 5.4. For any deterministic integrable function $f$ and any continuous semimartingale $Z$ vanishing at $t=0$, one has

$$
\int_{0}^{T} f(t) Z_{t} \mathrm{~d} t=\int_{0}^{T} \omega_{t, T}^{(0, f)} \mathrm{d} Z_{t}
$$

Proof. This is an application of the Itô formula to the product $\omega_{t, T}^{(0, f)} Z_{t}$.
Lemma 5.5. One has
$\mathbb{E}\left[l\left(\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}\right) \int_{0}^{T} \beta_{t} v_{1, t} \mathrm{~d} t\right]=\omega_{0, T}^{\left(\kappa, p \xi v_{0, ~},(-\kappa, \beta)\right.} \mathbb{E}\left[l^{(1)}\left(\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}\right)\right]$,
$\mathbb{E}\left[l\left(\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}\right) \int_{0}^{T} \beta_{t} v_{1, t}^{2} \mathrm{~d} t\right]=\omega_{0, T}^{\left(2 \kappa, \xi^{2} v_{0, .},(-2 \kappa, \beta)\right.} \mathbb{E}\left[l\left(\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}\right)\right]$

$$
+2 \omega_{0, T}^{\left(\kappa, \rho \xi v_{0,},\right),\left(\kappa, \rho \xi v_{0,},\right),(-2 \kappa, \beta)} \mathbb{E}\left[l^{(2)}\left(\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}\right)\right],
$$

$\mathbb{E}\left[l\left(\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}\right) \int_{0}^{T} \beta_{t} v_{2, t} \mathrm{~d} t\right]=\omega_{0, T}^{\left(\kappa, \rho \xi v_{0, .}\right),(0, \rho \xi),(-\kappa, \beta)} \mathbb{E}\left[l^{(2)}\left(\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}\right)\right]$.
Proof. Using Lemmas $5.3\left(f(t)=e^{-\kappa t} \beta_{t}, Z_{t}=\int_{0}^{t} e^{\kappa s} \xi_{s} \sqrt{v_{0, s}} \mathrm{~d} B_{s}\right)$ and 5.4, one has

$$
\begin{aligned}
\mathbb{E}\left[l\left(\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}\right) \int_{0}^{T} \beta_{t} v_{1, t} \mathrm{~d} t\right] & =\mathbb{E}\left[l\left(\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}\right) \int_{0}^{T} e^{-\kappa t} \beta_{t} \int_{0}^{t} e^{\kappa s} \xi_{s} \sqrt{v_{0, s}} \mathrm{~d} B_{s} \mathrm{~d} t\right] \\
& =\mathbb{E}\left[l\left(\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}\right) \int_{0}^{T} \omega_{t, T}^{(-\kappa, \beta)} e^{\kappa t} \xi_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}\right] \\
& =\mathbb{E}\left[l^{(1)}\left(\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}\right)\right] \int_{0}^{T} \omega_{t, T}^{(-\kappa, \beta)} e^{\kappa t} \rho_{t} \xi_{t} v_{0, t} \mathrm{~d} t,
\end{aligned}
$$

which gives the first equality. The second and the third equalities are proved in the same way.

Lemma 5.6. One has

$$
\mathbb{E}\left[\frac{\partial^{i+j} \tilde{P}_{B S}}{\partial x^{i} y^{j}}\right]=\frac{\partial^{i+j} P_{B S}}{\partial x^{i} y^{j}}\left(x_{0}, \int_{0}^{T} v_{0, t} \mathrm{~d} t\right) .
$$

Proof. One has

$$
\begin{aligned}
\mathbb{E}\left[\frac{\partial^{i} \tilde{P}_{B S}}{\partial x^{i}}\right] & =\partial_{x=x_{0}}^{i} \mathbb{E}\left[P_{B S}\left(x_{0}+\int_{0}^{T} \rho_{t} \sqrt{v_{0, t}} \mathrm{~d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2} v_{0, t} \mathrm{~d} t, \int_{0}^{T}\left(1-\rho_{t}^{2}\right) v_{0, t} \mathrm{~d} t\right)\right] \\
& =\frac{\partial^{i} P_{B S}}{\partial x^{i}}\left(x_{0}, \int_{0}^{T} v_{0, t} \mathrm{~d} t\right) .
\end{aligned}
$$

Since $\tilde{P}_{B S}$ verifies the relation

$$
\begin{equation*}
\frac{\partial \tilde{P}_{B S}}{\partial y}=\frac{1}{2}\left(\frac{\partial^{2} \tilde{P}_{B S}}{\partial x^{2}}-\frac{\partial \tilde{P}_{B S}}{\partial x}\right) \tag{5.1}
\end{equation*}
$$

we immediately obtain the result.
5.3. Proof of Proposition 2.1. One has

$$
\begin{aligned}
& \mathbb{E}\left[\frac{\partial \tilde{P}_{B S}}{\partial x}\left(\int_{0}^{T} \rho_{t}\left(\frac{v_{1, t}}{2\left(v_{0, t}\right)^{\frac{1}{2}}}+\frac{v_{2, t}}{4\left(v_{0, t}\right)^{\frac{1}{2}}}-\frac{v_{1, t}^{2}}{8\left(v_{0, t}\right)^{\frac{3}{2}}}\right) \mathrm{d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2}\left(v_{1, t}+\frac{v_{2, t}}{2}\right) \mathrm{d} t\right)\right] \\
& =\mathbb{E}\left[\frac{1}{2}\left(\frac{\partial^{2} \tilde{P}_{B S}}{\partial x^{2}}-\frac{\partial \tilde{P}_{B S}}{\partial x}\right) \int_{0}^{T} \rho_{t}^{2}\left(v_{1, t}+\frac{v_{2, t}}{2}\right) \mathrm{d} t\right]-\mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial x^{2}} \int_{0}^{T} \frac{\rho_{t}^{2} v_{1, t}^{2}}{8 v_{0, t}} \mathrm{~d} t\right] \\
& =\mathbb{E}\left[\frac{\partial \tilde{P}_{B S}}{\partial y} \int_{0}^{T} \rho_{t}^{2}\left(v_{1, t}+\frac{v_{2, t}}{2}\right) \mathrm{d} t\right]-\mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial x^{2}} \int_{0}^{T} \frac{\rho_{t}^{2} v_{1, t}^{2}}{8 v_{0, t}} \mathrm{~d} t\right]
\end{aligned}
$$

where we have used Lemma 5.3 at the first equality and identity (5.1) at the second one. Plugging this relation into the approximation (2.12) and summing the second and third lines, one has

$$
\begin{align*}
g(1) & =\mathbb{E}\left[\tilde{P}_{B S}\right]+\mathbb{E}\left[\frac{\partial \tilde{P}_{B S}}{\partial y} \int_{0}^{T}\left(v_{1, t}+\frac{v_{2, t}}{2}\right) \mathrm{d} t\right] \\
& -\mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial x^{2}} \int_{0}^{T} \frac{\rho_{t}^{2} v_{1, t}^{2}}{8 v_{0, t}} \mathrm{~d} t\right]+\frac{1}{2} \mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial x^{2}}\left(\int_{0}^{T} \rho_{t} \frac{v_{1, t}}{2\left(v_{0, t}\right)^{\frac{1}{2}}} \mathrm{~d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2} v_{1, t} \mathrm{~d} t\right)^{2}\right] \\
& +\frac{1}{2} \mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial y^{2}}\left(\int_{0}^{T}\left(1-\rho_{t}^{2}\right) v_{1, t} \mathrm{~d} t\right)^{2}\right] \\
& +\mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial x y}\left(\int_{0}^{T}\left(1-\rho_{t}^{2}\right) v_{1, t} \mathrm{~d} t\right)\left(\int_{0}^{T} \rho_{t} \frac{v_{1, t}}{2\left(v_{0, t} t^{\frac{1}{2}}\right.} \mathrm{d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2} v_{1, t} \mathrm{~d} t\right)\right]+\mathcal{E} \tag{5.2}
\end{align*}
$$

In addition, one has

$$
\begin{aligned}
& -\mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial x^{2}} \int_{0}^{T} \frac{\rho_{t}^{2} v_{1, t}^{2}}{8 v_{0, t}} \mathrm{~d} t\right]+\frac{1}{2} \mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial x^{2}}\left(\int_{0}^{T} \rho_{t} \frac{v_{1, t}}{2\left(v_{0, t}\right)^{\frac{1}{2}}} \mathrm{~d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2} v_{1, t} \mathrm{~d} t\right)^{2}\right] \\
& =\mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial x^{2}} \int_{0}^{T}\left(\int_{0}^{t} \rho_{s} \frac{v_{1, s}}{2\left(v_{0, s}\right)^{\frac{1}{2}}} \mathrm{~d} B_{s}-\int_{0}^{t} \frac{\rho_{s}^{2}}{2} v_{1, s} \mathrm{~d} s\right)\left(\rho_{t} \frac{v_{1, t}}{2\left(v_{0, t}\right)^{\frac{1}{2}}} \mathrm{~d} B_{t}-\frac{\rho_{t}^{2}}{2} v_{1, t} \mathrm{~d} t\right)\right] \\
& =\mathbb{E}\left[\frac{1}{2}\left(\frac{\partial^{3} \tilde{P}_{B S}}{\partial x^{3}}-\frac{\partial^{2} \tilde{P}_{B S}}{\partial x^{2}}\right) \int_{0}^{T}\left(\int_{0}^{t} \rho_{s} \frac{v_{1, s}}{2\left(v_{0, s}\right)^{\frac{1}{2}}} \mathrm{~d} B_{s}-\int_{0}^{t} \frac{\rho_{s}^{2}}{2} v_{1, s} \mathrm{~d} s\right) \rho_{t}^{2} v_{1, t} \mathrm{~d} t\right] \\
& =\mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial x y} \int_{0}^{T}\left(\int_{0}^{t} \rho_{s} \frac{v_{1, s}}{2\left(v_{0, s}\right)^{\frac{1}{2}}} \mathrm{~d} B_{s}-\int_{0}^{t} \frac{\rho_{s}^{2}}{2} v_{1, s} \mathrm{~d} s\right) \rho_{t}^{2} v_{1, t} \mathrm{~d} t\right]
\end{aligned}
$$

where we have used Itô's lemma for the square at the first equality, Lemma 5.3 at the second one, and identity (5.1) at the third one. Substituting this relation in the approximation (5.2) and summing the second and fourth lines, one gets

$$
\begin{align*}
g(1)= & \mathbb{E}\left[\tilde{P}_{B S}\right]+\mathbb{E}\left[\frac{\partial \tilde{P}_{B S}}{\partial y} \int_{0}^{T}\left(v_{1, t}+\frac{v_{2, t}}{2}\right) \mathrm{d} t\right] \\
+ & \mathbb{E}\left[\frac { \partial ^ { 2 } \tilde { P } _ { B S } } { \partial x y } \left(\int_{0}^{T}\left(\int_{0}^{t} \rho_{s} \frac{v_{1, s}}{2\left(v_{0, s}\right)^{\frac{1}{2}}} \mathrm{~d} B_{s}-\int_{0}^{t} \frac{\rho_{s}^{2}}{2} v_{1, s} \mathrm{~d} s\right) \rho_{t}^{2} v_{1, t} \mathrm{~d} t\right.\right. \\
& \left.\left.+\left(\int_{0}^{T}\left(1-\rho_{t}^{2}\right) v_{1, t} \mathrm{~d} t\right)\left(\int_{0}^{T} \rho_{t} \frac{v_{1, t}}{2\left(v_{0, t}\right)^{\frac{1}{2}}} \mathrm{~d} B_{t}-\int_{0}^{T} \frac{\rho_{t}^{2}}{2} v_{1, t} \mathrm{~d} t\right)\right)\right] \\
+ & \frac{1}{2} \mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial y^{2}}\left(\int_{0}^{T}\left(1-\rho_{t}^{2}\right) v_{1, t} \mathrm{~d} t\right)^{2}\right]+\mathcal{E} \tag{5.3}
\end{align*}
$$

First, we define

$$
H_{t}=\int_{0}^{t} \rho_{s} \frac{v_{1, s}}{2\left(v_{0, s}\right)^{\frac{1}{2}}} \mathrm{~d} B_{s}-\int_{0}^{t} \frac{\rho_{s}^{2}}{2} v_{1, s} \mathrm{~d} s, \quad d H_{t}=\rho_{t} \frac{v_{1, t}}{2\left(v_{0, t}\right)^{\frac{1}{2}}} \mathrm{~d} B_{t}-\frac{\rho_{t}^{2}}{2} v_{1, t} \mathrm{~d} t
$$

We now study the second term of (5.3). In the computations below, we use Itô's lemma for the second equality, Lemma 5.3 and identity (5.1) for the third equality, and Lemma 5.2
( $\left.G=\frac{\partial^{2} \tilde{P}_{B S}}{\partial x y} v_{1, t}\right)$ for the fourth equality; it gives

$$
\begin{aligned}
A= & \mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial x y}\left(\int_{0}^{T} H_{t} \rho_{t}^{2} v_{1, t} \mathrm{~d} t+\left(\int_{0}^{T}\left(1-\rho_{t}^{2}\right) v_{1, t} \mathrm{~d} t\right) H_{T}\right)\right] \\
= & \mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial x y}\left(\int_{0}^{T} H_{t}\left(\rho_{t}^{2}+1-\rho_{t}^{2}\right) v_{1, t} \mathrm{~d} t+\int_{0}^{T}\left(\int_{0}^{t}\left(1-\rho_{s}^{2}\right) v_{1, s} \mathrm{~d} s\right) d H_{t}\right)\right] \\
= & \int_{0}^{T} \mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial x y} v_{1, t} H_{t}\right] \mathrm{d} t+\mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial y^{2}} \int_{0}^{T}\left(\int_{0}^{t}\left(1-\rho_{s}^{2}\right) v_{1, s} \mathrm{~d} s\right) \rho_{t}^{2} v_{1, t} \mathrm{~d} t\right] \\
= & \int_{0}^{T} \mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial x y}\left(v_{1, t}\left(-\int_{0}^{t} \frac{\rho_{s}^{2}}{2} v_{1, s} \mathrm{~d} s\right)+\int_{0}^{t} \rho_{s} \frac{v_{1, s}}{2 \sqrt{v_{0, s}}} D_{s}^{B} v_{1, t} \mathrm{~d} s\right)\right. \\
& \left.\quad+\frac{\partial^{3} \tilde{P}_{B S}}{\partial x^{2} y} v_{1, t} \int_{0}^{t} \frac{\rho_{s}^{2}}{2} v_{1, s} \mathrm{~d} s\right] \mathrm{d} t+\mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial y^{2}} \int_{0}^{T}\left(\int_{0}^{t}\left(1-\rho_{s}^{2}\right) v_{1, s} \mathrm{~d} s\right) \rho_{t}^{2} v_{1, t} \mathrm{~d} t\right]
\end{aligned}
$$

From (2.3), one has $D_{s}^{B} v_{1, t}=e^{-k t} e^{k s} \xi_{s} \sqrt{v_{0, s}}$. Hence it is deterministic. Thus, using identity (5.1) and Lemma 5.3 for the first equality and (2.4) for the second equality, one has

$$
\begin{aligned}
A= & \mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial y^{2}}\left(\int_{0}^{T}\left(\int_{0}^{t} \rho_{s}^{2} v_{1, s} \mathrm{~d} s\right) v_{1, t} \mathrm{~d} t+\left(\int_{0}^{t}\left(1-\rho_{s}^{2}\right) v_{1, s} \mathrm{~d} s\right) \rho_{t}^{2} v_{1, t} \mathrm{~d} t\right)\right] \\
& +\mathbb{E}\left[\frac{\partial \tilde{P}_{B S}}{\partial y} \int_{0}^{T}\left(\int_{0}^{t} \frac{v_{1, s}}{2 v_{0, s}} e^{-k t} e^{k s} \xi_{s} \sqrt{v_{0, s}} \mathrm{~d} B_{s}\right) \mathrm{d} t\right] \\
= & \mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial y^{2}}\left(\int_{0}^{T}\left(\int_{0}^{t} \rho_{s}^{2} v_{1, s} \mathrm{~d} s\right) v_{1, t} \mathrm{~d} t+\left(\int_{0}^{t}\left(1-\rho_{s}^{2}\right) v_{1, s} \mathrm{~d} s\right) \rho_{t}^{2} v_{1, t} \mathrm{~d} t\right)\right] \\
& +\mathbb{E}\left[\frac{\partial \tilde{P}_{B S}}{\partial y} \int_{0}^{T} \frac{v_{2, t}}{2} \mathrm{~d} t\right]
\end{aligned}
$$

Now, plug this last equality into (5.3) and use the easy identity

$$
\begin{aligned}
& \int_{0}^{T}\left(\left(\int_{0}^{t} \rho_{s}^{2} v_{1, s} \mathrm{~d} s\right) v_{1, t} \mathrm{~d} t+\left(\int_{0}^{t}\left(1-\rho_{s}^{2}\right) v_{1, s} \mathrm{~d} s\right) \rho_{t}^{2} v_{1, t} \mathrm{~d} t\right)+\frac{1}{2}\left(\int_{0}^{T}\left(1-\rho_{t}^{2}\right) v_{1, t} \mathrm{~d} t\right)^{2} \\
& =\int_{0}^{T}\left(\left(\int_{0}^{t} \rho_{s}^{2} v_{1, s} \mathrm{~d} s\right) v_{1, t} \mathrm{~d} t+\left(\int_{0}^{t}\left(1-\rho_{s}^{2}\right) v_{1, s} \mathrm{~d} s\right)\left(\rho_{t}^{2}+1-\rho_{t}^{2}\right) v_{1, t} \mathrm{~d} t\right) \\
& =\int_{0}^{T}\left(\left(\int_{0}^{t}\left(\rho_{s}^{2}+1-\rho_{s}^{2}\right)\right) v_{1, s} \mathrm{~d} s\right) v_{1, t} \mathrm{~d} t=\frac{1}{2}\left(\int_{0}^{T} v_{1, t} \mathrm{~d} t\right)^{2}
\end{aligned}
$$

it immediately gives the result.

### 5.4. Proof of Theorem 2.2.

Proof. Step 1. We show the equality

$$
\mathbb{E}\left[\frac{\partial \tilde{P}_{B S}}{\partial y} \int_{0}^{T}\left(v_{1, t}+v_{2, t}\right) \mathrm{d} t\right]=\sum_{i=1}^{2} a_{i, T} \frac{\partial^{i+1} P_{B S}\left(x_{0}, \int_{0}^{T} v_{0, t} \mathrm{~d} t\right)}{\partial x^{i} y}
$$

where

$$
a_{1, T}=\omega_{0, T}^{\left(\kappa, \rho \xi v_{0, .}\right),(-\kappa, 1)}, \quad a_{2, T}=\omega_{0, T}^{\left(\kappa, \rho \xi v_{0, .}\right),(0, \rho \xi),(-\kappa, 1)}
$$

Actually, the result is an immediate application of Lemmas 5.5 and 5.6.
Step 2 . We show the equality

$$
\frac{1}{2} \mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial y^{2}}\left(\int_{0}^{T} v_{1, t} \mathrm{~d} t\right)^{2}\right]=\sum_{i=0}^{1} b_{2 i, T} \frac{\partial^{2 i+2} P_{B S}\left(x_{0}, \int_{0}^{T} v_{0, t} \mathrm{~d} t\right)}{\partial x^{2 i} y^{2}}
$$

where

$$
\begin{aligned}
& b_{0, T}=\omega_{0, T}^{\left(2 \kappa, \xi^{2} v_{0, .}\right),(-\kappa, 1),(-\kappa, 1)} \\
& b_{2, T}=\omega_{0, T}^{\left(\kappa, \rho \xi v_{0, .}\right),(-\kappa, 1),\left(\kappa, \rho \xi v_{0, .}\right),(-\kappa, 1)}+2 \omega_{0, T}^{\left(\kappa, \rho \xi v_{0, .}\right),\left(\kappa, \rho \xi v_{0, .}\right),(-\kappa, 1),(-\kappa, 1)}=\frac{a_{1, T}^{2}}{2}
\end{aligned}
$$

Indeed, one has

$$
\begin{aligned}
\frac{1}{2} \mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial y^{2}}\right. & \left.\left(\int_{0}^{T} v_{1, t} \mathrm{~d} t\right)^{2}\right]=\mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial y^{2}} \int_{0}^{T}\left(\int_{0}^{t} v_{1, s} \mathrm{~d} s\right) v_{1, t} \mathrm{~d} t\right] \\
& =\mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial y^{2}} \int_{0}^{T}\left(\int_{t}^{T} e^{-\kappa s} \mathrm{~d} s\right)\left(e^{\kappa t} v_{1, t}^{2} \mathrm{~d} t+\xi_{t} \sqrt{v_{0, t}} e^{\kappa t}\left(\int_{0}^{t} v_{1, s} \mathrm{~d} s\right) \mathrm{d} B_{t}\right)\right] \\
& =\mathbb{E}\left[\frac{\partial^{2} \tilde{P}_{B S}}{\partial y^{2}} \int_{0}^{T}\left(\int_{t}^{T} e^{-\kappa s} \mathrm{~d} s\right) e^{\kappa t} v_{1, t}^{2} \mathrm{~d} t\right]+\mathbb{E}\left[\frac{\partial^{3} \tilde{P}_{B S}}{\partial x y^{2}} \int_{0}^{T} \omega_{t, T}^{\left(\kappa, \rho \xi v_{0, .}\right),(-\kappa, 1)} v_{1, t} \mathrm{~d} t\right]
\end{aligned}
$$

where we have used Lemma $5.4\left(f(t)=e^{-\kappa t}, Z_{t}=\left(\int_{0}^{t} v_{1, s} \mathrm{~d} s\right)\left(e^{\kappa t} v_{1, t}\right)\right)$ for the second equality and Lemmas 5.3 and $5.4\left(f(t)=\left(\int_{t}^{T} e^{-\kappa s} \mathrm{~d} s\right) \rho_{t} \xi_{t} v_{0, t} e^{\kappa t}, Z_{t}=\int_{0}^{t} v_{1, s} \mathrm{~d} s\right)$ for the third equality.

An application of the first and second equalities in Lemma 5.5 gives the announced result. Actually, it remains to show that $b_{2, T}=a_{1, T}^{2} / 2$. Indeed, consider two càdlàg functions $f$ and $g:[0, T] \rightarrow \mathbb{R}$. Then

$$
\begin{aligned}
\frac{\left(\int_{0}^{T} f_{t}\left(\int_{t}^{T} g_{s} \mathrm{~d} s\right) \mathrm{d} t\right)^{2}}{2}= & \frac{\int_{0}^{T} \int_{0}^{T} f_{t_{1}}\left(\int_{t_{1}}^{T} g_{t_{3}} \mathrm{~d} t_{3}\right) f_{t_{2}}\left(\int_{t_{2}}^{T} g_{t_{4}} \mathrm{~d} t_{4}\right) \mathrm{d} t_{2} \mathrm{~d} t_{1}}{2} \\
= & \int_{0}^{T} f_{t_{1}}\left(\int_{t_{1}}^{T} \int_{t_{1}}^{T} g_{t_{3}} f_{t_{2}}\left(\int_{t_{2}}^{T} g_{t_{4}} \mathrm{~d} t_{4}\right) \mathrm{d} t_{3} \mathrm{~d} t_{2}\right) \mathrm{d} t_{1} \\
= & \int_{0}^{T} f_{t_{1}}\left(\int_{t_{1}}^{T} f_{t_{2}} \int_{t_{2}}^{T} \int_{t_{2}}^{T} g_{t_{3}} g_{t_{4}} \mathrm{~d} t_{3} \mathrm{~d} t_{4} \mathrm{~d} t_{2}\right. \\
& \left.\quad+\int_{t_{1}}^{T} g_{t_{3}} \int_{t_{3}}^{T} f_{t_{2}} \int_{t_{2}}^{T} g_{t_{4}} \mathrm{~d} t_{4} \mathrm{~d} t_{2} \mathrm{~d} t_{3}\right) \mathrm{d} t_{1} \\
= & 2 \int_{0}^{T} f_{t_{1}} \int_{t_{1}}^{T} f_{t_{2}} \int_{t_{2}}^{T} g_{t_{3}} \int_{t_{3}}^{T} g_{t_{4}} \mathrm{~d} t_{3} \mathrm{~d} t_{4} \mathrm{~d} t_{2} \mathrm{~d} t_{1} \\
& \quad+\int_{0}^{T} f_{t_{1}} \int_{t_{1}}^{T} g_{t_{3}} \int_{t_{3}}^{T} f_{t_{2}} \int_{t_{2}}^{T} g_{t_{4}} \mathrm{~d} t_{4} \mathrm{~d} t_{2} \mathrm{~d} t_{3} \mathrm{~d} t_{1}
\end{aligned}
$$

Putting $f(t)=\rho_{t} \xi_{t} v_{0, t} e^{k t}$ and $g(t)=e^{-k t}$ in the previous equality readily gives $b_{2, T}=\frac{a_{1, T}^{2}}{2}$, which finishes the proof.
6. Conclusion. We have established an approximation pricing formula for call/put options in the time dependent Heston models. We prove that the error is of order 3 w.r.t. the volatility of volatility and 2 w.r.t. the maturity. In practice, taking the Fourier method as a benchmark, the accuracy is excellent for a large range of strikes and maturities. In addition, the computational time is about 100 to 1000 times smaller than using an efficient Fourier method.

Following the arguments in [8], our formula extends immediately to other payoffs depending on $S_{T}$ (note that the identities (2.2) and (5.1) are valid for any payoff of this type). As explained in [8], the smoother the payoff, the higher the error order w.r.t. $T$; the less smooth the payoff, the lower the error order w.r.t. $T$. For digital options, the error order w.r.t. $T$ becomes $3 / 2$ instead of 2 .

Extensions to exotic options and to the third order expansion formula w.r.t. the volatility of volatility are left for further research.
7. Appendix: Closed formulas in the Heston model. There are few closed representations for the call/put prices written on the asset $S_{t}=e^{\int_{0}^{t}\left(r_{s}-q_{s}\right) d s} e^{X_{t}}$ in the Heston model (defined in (1.1) and (1.2)). We focus on the Heston formula [22] and on the Lewis formula [27]. Both of them rely on the knowledge of the characteristic function of the log-asset price $\left(X_{t}\right)_{t}$ and on Fourier transform-based approaches.
(i) In [22], Heston obtains a representation in a Black-Scholes form

$$
\operatorname{Call}_{\text {Heston }}\left(t, S_{t}, v_{t} ; T, K\right)=S_{t} e^{-\int_{t}^{T} q_{s} d s} P_{1}-K e^{-\int_{t}^{T} r_{s} d s} P_{2}
$$

where both probabilities $P_{1}$ and $P_{2}$ are equal to a one-dimensional integral of characteristic functions.
(ii) In [27], Lewis takes advantage of the generalized Fourier transform, by using an integration along a straight line in the complex plane parallel to the real axis. It is important to detect the strip where the integration is safe. Lewis suggests the use of complex numbers $z$ such that $\operatorname{Im}(z)=\frac{1}{2}$. His formula writes

$$
\operatorname{Call}_{\text {Heston }}\left(t, S_{t}, v_{t} ; T, K\right)=S_{t} e^{-\int_{t}^{T} q_{s} d s}-\frac{K e^{-\int_{t}^{T} r_{s} d s}}{2 \pi} \int_{\frac{i}{2}-\infty}^{\frac{i}{2}+\infty} e^{-i z X} \phi_{T}(-z) \frac{d z}{z^{2}-i z},
$$

where $X=\log \left(\frac{S_{t}-\int_{t}^{T} q_{s} d s}{K e^{-\int_{t}^{T} r_{s} d s}}\right)$ and $\phi_{T}(z)=\mathbb{E}\left(e^{z\left(X_{T}-X_{t}\right)} \mid \mathcal{F}_{t}\right)$. Then, the above integral is evaluated by numerical integration.

Using PDE arguments in combination with affine models, we can obtain an explicit formula for $\phi_{T}(z)$ in the case of constant Heston parameters. In addition, it can be computed without discontinuities in $z$, following the arguments in [23]. For piecewise constant parameters, the characteristic function $\phi_{T}(z)$ can be computed recursively using nested Riccati equations with constant coefficients: we refer the reader to the work by Mikhailov and Nogel [31].

In our numerical tests, we prefer the Lewis formula, which gives better numerical results, in particular for very small or very large strikes, compared to the Heston formula.

## REFERENCES

[1] A. Alfonsi, On the discretization schemes for the CIR (and Bessel squared) processes, Monte Carlo Methods Appl., 11 (2005), pp. 355-384.
[2] A. Alfonsi, High order discretization schemes for the CIR process: Application to affine term structure and Heston models, Math. Comp., 79 (2010), pp. 209-237.
[3] E. Alòs, J. A. León, and J. Vives, On the short-time behaviour of the implied volatility for jumpdiffusion models with stochastic volatility, Finance Stoch., 11 (2007), pp. 571-589.
[4] L. Andersen, Simple and efficient simulation of the Heston stochastic volatility model, J. Comput. Finance, 11 (2008), pp. 1-42.
[5] F. Antonelli and S. Scarlatti, Pricing options under stochastic volatility: A power series approach, Finance Stoch., 13 (2009), pp. 269-303.
[6] G. Baski, C. Cao, and Z. Chen, Empirical performance of alternative option pricing models, J. Finance, 52 (1997), pp. 2003-2049.
[7] E. Benhamou, E. Gobet, and M. Miri, Expansion formulas for European options in a local volatility model, Int. J. Theoret. Appl. Finance, to appear.
[8] E. Benhamou, E. Gobet, and M. Miri, Smart expansion and fast calibration for jump diffusion, Finance Stoch., 13 (2009), pp. 563-589.
[9] E. Benhamou, E. Gobet, and M. Miri, Time Dependent Heston Model, http://ssrn.com/abstract= 1367955 (2009).
[10] H. Berestycki, J. Busca, and I. Florent, Computing the implied volatility in stochastic volatility models, Comm. Pure Appl. Math., 10 (2004), pp. 1352-1373.
[11] A. Borodin and P. Salminen, Handbook of Brownian Motion - Facts and Formulae, 2nd ed., Probab. Appl., Birkhäuser Verlag, Basel, 2002.
[12] M. Bossy and A. Diop, An Efficient Discretisation Scheme for One Dimensional SDEs with a Diffusion Coefficient Function of the Form $|x|^{\alpha}$, $\alpha$ in $[1 / 2,1)$, Technical report, INRIA, Sophia Antipolis, France, 2007; available online from http://hal.inria.fr/docs/00/16/47/45/PDF/RR-5396_V2.pdf.
[13] M. Broadie and O. Kaya, Exact simulation of stochastic volatility and other affine jump diffusion processes, Oper. Res., 54 (2006), pp. 217-231.
[14] A. Diop, Sur la discrétisation et le comportement à petit bruit d'EDS multidimensionnelles dont les coefficients sont à dérivées singulières, Ph.D. thesis, INRIA, Sophia Antipolis, France, 2003; available online from http://www.inria.fr/rrrt/tu-0785.html.
[15] D. Dufresne, The Integrated Square-Root Process, Working paper, University of Montreal, Montreal, Canada, 2001.
[16] A. Elices, Models with Time-Dependent Parameters Using Transform Methods: Application to Heston's Model, http://arxiv.org/abs/0708.2020 (2008).
[17] M. Forde, The Small-Time Behavior of Diffusion and Time-Changed Diffusion Processes on the Line, http://arxiv.org/abs/math/0609117 (2008).
[18] J. Fouque, G. Papanicolaou, and R. Sircar, Derivatives in Financial Markets with Stochastic Volatility, Cambridge University Press, Cambridge, UK, 2000.
[19] J. Fouque, G. Papanicolaou, R. Sircar, and K. Solna, Maturity cycles in implied volatility, Finance Stoch., 8 (2004), pp. 451-477.
[20] A. Göing-Jaeschke and M. Yor, A survey and some generalizations of Bessel processes, Bernoulli, 9 (2003), pp. 313-349.
[21] P. Hagan, D. Kumar, A. Lesniewski, and D. Woodward, Managing smile risk, Willmott Magazine, 15 (2002), pp. 84-108.
[22] S. Heston, A closed-form solutions for options with stochastic volatility with applications to bond and currency options,, Rev. Financ. Stud., 6 (1993), pp. 327-343.
[23] P. Jackel and C. Kahl, Not-so-complex logarithms in the Heston model, Wilmott Magazine, 19 (2005), pp. 94-103.
[24] I. Karatzas and S. Shreve, Brownian Motion and Stochastic Calculus, Springer-Verlag, New York, 1988.
[25] T. Kluge, Pricing Derivatives in Stochastic Volatility Models Using the Finite Difference Method, Ph.D. thesis, Technical University, Chemnitz, Germany, 2002.
[26] P. H. Labordère, A General Asymptotic Implied Volatility for Stochastic Volatility Models, http://arxiv. org/abs/cond-mat/0504317 (2005).
[27] A. Lewis, Option Valuation under Stochastic Volatility with Mathematica Code, Finance Press, Newport Beach, CA, 2000.
[28] A. Lewis, Geometries and Smile Asymptotics for a Class of Stochastic Volatility Models, http://www. optioncity.net (2007).
[29] A. Lipton, The vol smile problem, Risk Magazine, 15 (2002), pp. 61-65.
[30] Y. Maghsoodi, Solution of the extended CIR term structure and bond option valuation, Math. Finance, 6 (1996), pp. 89-109.
[31] S. Mikhailov and U. Nogel, Heston's stochastic volatility model implementation, calibration and some extensions, Wilmott Magazine, July (2003), pp. 74-79.
[32] M. Miri, Stochastic Expansions and Closed Pricing Formulas of European Options, Ph.D. thesis, Grenoble Institute of Technology, Grenoble, France; available online from http://tel.archives-ouvertes.fr/tel-00452857/fr/ (2009).
[33] D. Nualart, Malliavin Calculus and Related Topics, 2nd ed., Springer-Verlag, Berlin, 2006.
[34] Y. Osajima, The Asymptotic Expansion Formula of Implied Volatility for Dynamic SABR Model and FX Hybrid Model, http://ssrn.com/abstract=965265 (2007).
[35] V. Piterbarg, Stochastic volatility model with time-dependent skew, Appl. Math. Finance, 12 (2005), pp. 147-185.
[36] V. Piterbarg, Time to smile, Risk Magazine, 18 (2005), pp. 71-75.
[37] P. Protter, Stochastic Integration and Differential Equations, Springer-Verlag, Heidelberg, 1990.
[38] E. Renault and N. Touzi, Option hedging and implied volatilities in a stochastic volatility model, Math. Finance, 6 (1996), pp. 279-302.
[39] D. Revuz and M. Yor, Continuous Martingales and Brownian Motion, 3rd ed., Springer-Verlag, Berlin, 1999.
[40] R. Smith, An almost exact simulation method for the Heston model, J. Comput. Finance, 11 (2008), pp. 115-125.

# Portfolio Choice under Space-Time Monotone Performance Criteria* 

M. Musiela ${ }^{\dagger}$ and T. Zariphopoulou ${ }^{\ddagger}$


#### Abstract

The class of time-decreasing forward performance processes is analyzed in a portfolio choice model of Itô-type asset dynamics. The associated optimal wealth and portfolio processes are explicitly constructed and their probabilistic properties are discussed. These formulae are, in turn, used in analyzing how the investor's preferences can be calibrated to the market, given his desired investment targets.


Key words. portfolio choice, forward investment performance, heat equation
AMS subject classifications. Primary, 91B16, 91B28
DOI. 10.1137/080745250

1. Introduction. This paper is a contribution to portfolio management from the perspective of investor preferences and, hence, in its spirit is related to the classical expected utility maximization problem introduced by Merton [8]. Therein, one first chooses an investment horizon and assigns a utility function at the end of it and, in turn, seeks an investment strategy which delivers the maximal expected (indirect) utility of terminal wealth. Recently, we proposed an alternative approach to optimal portfolio choice which is based on the so-called forward performance criterion (see, among others, [10] and [9]). In this approach, the investor does not choose her risk preferences at a single point in time, as is the case in the Merton model, but has the flexibility to revise them dynamically.

Herein, we focus on a specific case of a forward performance criterion, originally introduced in [12]. This criterion is a composition of deterministic and stochastic inputs. The deterministic input corresponds to the investor's preferences, or alternatively, to her tolerance towards risk. It is investor specific, represented by a function $u(x, t)$, which is increasing and concave in $x$ and decreasing in $t$. The stochastic input, however, is universal for all investors and is given by $A_{t}=\int_{0}^{t}\left|\lambda_{s}\right|^{2} d s, t \geq 0$, with $\lambda_{t}$ being the Sharpe ratio of the securities available for trading. The performance criterion is, then, given by the process $U_{t}(x)=u\left(x, A_{t}\right)$, $t \geq 0$. Because of its form and the properties of the involved inputs, the performance process is monotone in wealth and time.

[^27]Our contribution is threefold. First, we provide a general characterization of the differential input function, $u(x, t)$. A space-time harmonic function plays a pivotal role in achieving this. This function is fully characterized by a positive measure which, in turn, becomes the underlying element in the specification of all quantities of interest. An important ingredient is the support of this measure, as it directly affects the domain of the differential input. We provide a detailed study of this interplay.

The second contribution is the explicit construction of the optimal investment strategy and the associated optimal wealth. The specification of these processes is rather general, for it does not rely on any Markovian assumptions on asset dynamics or on any specific structure of the investor's input. To our knowledge, this is one of the very rare cases in which such explicit formulae can be derived in a model as general as the one considered herein.

The third contribution is the initiation of a study on how we can learn about the investor's risk preferences from her investment goals. For example, the investor may want to specify the average level of wealth she could generate in future times, in the particular market she chooses. This information is then used to deduce risk preferences which are consistent with this investment target. Such inference problems are, in general, very hard to solve due to the lack of closed form formulae, a difficulty that is overcome herein due to the availability of explicit solutions. It is important to notice that the assessment of market movements is implicitly embedded in the investor's desired investment goals. In many aspects, this approach can be compared with the calibration of derivative pricing models. Indeed, therein, one also needs to make a statement about the market under the historical measure. Then, assuming no arbitrage, derivatives are valued under the risk neutral measure, with valuation requiring calibration of the model to the observable market prices of the relevant assets. There is, however, an important difference between derivative pricing and the portfolio selection problem. In the latter, we cannot rely on the market to give information about the investor's individual preferences. However, we show how to acquire information about them by asking the investor to specify the desirable properties of the wealth process she wishes to generate.

The criterion studied in this paper does not allow for arbitrary stochastic evolution of preferences as the performance process is monotonically decreasing, and hence its quadratic variation is equal to zero. To incorporate more flexibility, one needs to work with selection criteria in their full generality. Preliminary results in this direction can be found in [11].

The paper is organized as follows. The model and the general portfolio selection criteria are defined in the next section. In section 3, we present the monotone performance criterion and provide explicit solutions for the associated optimal wealth and portfolio processes. In section 4, we provide a detailed construction of the differential input via the associated spacetime harmonic function. In section 5, we analyze the case of deterministic market price of risk and discuss the distributional properties of the optimal wealth. We finish by discussing how the investor-specific input can be inferred from targeted properties of her future expected wealth.
2. The model and portfolio selection criteria. The market environment consists of one riskless and $k$ risky securities. The prices of risky securities are modelled as Itô processes. Namely, the price $S_{t}^{i}, t>0$, of the $i$ th risky asset follows

$$
d S_{t}^{i}=S_{t}^{i}\left(\mu_{t}^{i} d t+\sum_{j=1}^{d} \sigma_{t}^{j i} d W_{t}^{j}\right)
$$

with $S_{0}^{i}>0$ for $i=1, \ldots, k$. The process $W_{t}=\left(W_{t}^{1}, \ldots, W_{t}^{d}\right), t \geq 0$, is a standard $d$ dimensional Brownian motion, defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The coefficients $\mu_{t}^{i}$ and $\sigma_{t}^{i}=\left(\sigma_{t}^{1 i}, \ldots, \sigma_{t}^{d i}\right), i=1, \ldots, k, t \geq 0$, are $\mathcal{F}_{t}$-adapted processes with values in $\mathbb{R}$ and $\mathbb{R}^{d}$, respectively. For the sake of brevity, we write $\sigma_{t}$ to denote the volatility matrix, i.e., the $d \times k$ random matrix $\left(\sigma_{t}^{j i}\right)$, whose $i$ th column represents the volatility $\sigma_{t}^{i}$ of the $i$ th risky asset. We may, then, alternatively write the above equation as

$$
\begin{equation*}
d S_{t}^{i}=S_{t}^{i}\left(\mu_{t}^{i} d t+\sigma_{t}^{i} \cdot d W_{t}\right) \tag{1}
\end{equation*}
$$

The riskless asset, the savings account, has the price process $B_{t}, t>0$, satisfying

$$
d B_{t}=r_{t} B_{t} d t
$$

with $B_{0}=1$ and for a nonnegative $\mathcal{F}_{t}$-adapted interest rate process $r_{t}$. Also, we denote by $\mu_{t}$ the $k$-dimensional vector with coordinates $\mu_{t}^{i}$, and by 1 the $k$-dimensional vector with every component equal to one. The processes, $\mu_{t}, \sigma_{t}$, and $r_{t}$ satisfy the appropriate integrability conditions.

We assume that the volatility vectors are such that

$$
\begin{equation*}
\mu_{t}-r_{t} \mathbf{1} \in \operatorname{Lin}\left(\sigma_{t}^{T}\right) \tag{2}
\end{equation*}
$$

where $\operatorname{Lin}\left(\sigma_{t}^{T}\right)$ denotes the linear space generated by the columns of $\sigma_{t}^{T}$. This implies that $\sigma_{t}^{T}\left(\sigma_{t}^{T}\right)^{+}\left(\mu_{t}-r_{t} \mathbf{1}\right)=\mu_{t}-r_{t} \mathbf{1}$, and therefore the vector

$$
\begin{equation*}
\lambda_{t}=\left(\sigma_{t}^{T}\right)^{+}\left(\mu_{t}-r_{t} \mathbf{1}\right) \tag{3}
\end{equation*}
$$

is a solution to the equation $\sigma_{t}^{T} x=\mu_{t}-r_{t} \mathbf{1}$. The matrix $\left(\sigma_{t}^{T}\right)^{+}$is the Moore-Penrose pseudoinverse ${ }^{1}$ of the matrix $\sigma_{t}^{T}$.

Occasionally, we will be referring to $\lambda_{t}$ as the market price of risk. It easily follows that

$$
\begin{equation*}
\sigma_{t} \sigma_{t}^{+} \lambda_{t}=\lambda_{t} \tag{4}
\end{equation*}
$$

and, hence, $\lambda_{t} \in \operatorname{Lin}\left(\sigma_{t}\right)$. We assume throughout that the process $\lambda_{t}$ is bounded by a deterministic constant $c>0$; i.e., for all $t \geq 0$,

$$
\begin{equation*}
\left|\lambda_{t}\right| \leq c \tag{5}
\end{equation*}
$$

Starting at $t=0$ with an initial endowment $x \in \mathbb{R}$, the investor invests at any time $t>0$ in the riskless and risky assets. The present values of the amounts invested are denoted by $\pi_{t}^{0}$ and $\pi_{t}^{i}, i=1, \ldots, k$, respectively.

The present value of her investment is, then, given by $X_{t}^{\pi}=\sum_{i=0}^{k} \pi_{t}^{i}, t>0$. We will refer to $X_{t}^{\pi}$ as the discounted wealth. The investment strategies will play the role of control

[^28]processes and are taken to satisfy the standard assumption of being self-financing. Using (1), we then deduce that the discounted wealth satisfies
\[

$$
\begin{equation*}
d X_{t}^{\pi}=\sum_{i=1}^{k} \pi_{t}^{i} \sigma_{t}^{i} \cdot\left(\lambda_{t} d t+d W_{t}\right)=\sigma_{t} \pi_{t} \cdot\left(\lambda_{t} d t+d W_{t}\right), \tag{6}
\end{equation*}
$$

\]

where the (column) vector $\pi_{t}=\left(\pi_{t}^{i} ; i=1, \ldots, k\right)$. The set of admissible strategies, $\mathcal{A}$, is defined as

$$
\begin{equation*}
\mathcal{A}=\left\{\pi: \text { self-financing with } \pi_{t} \in \mathcal{F}_{t} \text { and } E\left(\int_{0}^{t}\left|\sigma_{s} \pi_{s}\right|^{2} d s\right)<\infty, t>0\right\} . \tag{7}
\end{equation*}
$$

The problem we propose to address is that of a choice of an investment strategy from the set $\mathcal{A}$. To this aim, we introduce below a process which measures the performance of any admissible portfolio and gives us a selection criterion. Specifically, a strategy is deemed optimal if it generates a wealth process whose average performance is maintained over time. In other words, the average performance of this strategy at any future date, conditional on today's information, preserves the performance of this strategy up until today. Any strategy that fails to maintain the average performance over time is, then, suboptimal.

We present the definition of the forward performance next. It first appeared in [10] and is given herein for completeness. We note that this definition is slightly different than the original one, introduced by the authors in [12], in that the initial condition is not explicitly included. As the analysis in section 4 will show, not all strictly increasing and concave solutions can serve as initial conditions, even for the special classes of monotone processes we examine herein. Characterizing the set of appropriate initial conditions is challenging and is currently being investigated by the authors.

Definition 1. An $\mathcal{F}_{t}$-adapted process $U_{t}(x)$ is a forward performance if for $t \geq 0$ and $x \in \mathbb{R}$ :
(i) the mapping $x \rightarrow U_{t}(x)$ is strictly concave and increasing;
(ii) for each $\pi \in \mathcal{A}, E\left(U_{t}\left(X_{t}^{\pi}\right)\right)^{+}<\infty$, and

$$
\begin{equation*}
E\left(U_{s}\left(X_{s}^{\pi}\right) \mid \mathcal{F}_{t}\right) \leq U_{t}\left(X_{t}^{\pi}\right), \quad s \geq t ; \tag{8}
\end{equation*}
$$

(iii) there exists $\pi^{*} \in \mathcal{A}$, for which

$$
\begin{equation*}
E\left(U_{s}\left(X_{s}^{\pi^{*}}\right) \mid \mathcal{F}_{t}\right)=U_{t}\left(X_{t}^{\pi^{*}}\right), \quad s \geq t \tag{9}
\end{equation*}
$$

The intuition behind the above definition comes from the analogous martingale and supermartingale properties that the traditional maximal expected utility (value function) has (see, among others, [8], [5], and [14]). Indeed, we recall that the latter is defined in a finite trading horizon, say $[0, T]$, by

$$
\begin{equation*}
v(x, t ; T)=\sup _{\mathcal{A}_{T}} E\left(V\left(X_{T}^{\pi}\right) \mid \mathcal{F}_{t}, X_{t}^{\pi}=x\right), \tag{10}
\end{equation*}
$$

with $(x, t) \in \mathbb{R} \times[0, T]$, where $\mathcal{A}_{T}$ is the set of admissible policies defined similarly to $\mathcal{A}$ herein, and $V$ is the investor's utility, given by an increasing, concave, and smooth function. Under
rather general model assumptions, the value function satisfies the dynamic programming principle (DDP), namely, for $0 \leq t \leq s \leq T$,

$$
\begin{equation*}
v(x, t ; T)=\sup _{\mathcal{A}_{T}} E\left(v\left(X_{s}^{\pi}, s ; T\right) \mid \mathcal{F}_{t}, X_{t}^{\pi}=x\right) . \tag{11}
\end{equation*}
$$

One then sees that if the above supremum is achieved and certain integrability conditions hold, the processes $v\left(X_{s}^{\pi}, s ; T\right)$ and $v\left(X_{s}^{*}, s ; T\right)$ are, respectively, a supermartingale and a martingale on $[0, T]$.

We stress that the analogous equivalence in the forward formulation of the problem has not yet been established. Specifically, one could define the forward performance process via the (forward) stochastic optimization problem

$$
U_{t}(x)=\sup _{\mathcal{A}} E\left(U_{s}\left(X_{s}^{\pi}\right) \mid \mathcal{F}_{t}, X_{t}^{\pi}=x\right),
$$

for all $0 \leq t \leq s$, and for an appropriately defined initial condition. Characterizing its solutions poses a number of challenging questions, some of them being currently investigated by the authors. ${ }^{2}$ From a different perspective, one could seek an axiomatic construction of a forward performance process. Results in this direction, as well as on the dual formulation of the problem, can be found in [19] for the exponential case (see, also, [1] for a constrained case). We refer the reader to [10] for further discussion on the forward performance and its similarities to and differences from the classical value function.

The definition of the forward performance process requires the integrability of $\left(U_{t}\left(X_{t}^{\pi}\right)\right)^{+}$. This allows us to define the conditional expectations $E\left(U_{s}\left(X_{s}^{\pi}\right) \mid \mathcal{F}_{t}\right)$ for $s \geq t$ and, in turn, obtain a more practical intuition for our criteria. This leads, however, to additional integrability assumptions which further constrain the class of forward solutions the investor may employ. On the other hand, from the applications perspective, this may help in the calibration process of the investor's initial risk preferences.

Alternatively, and simpler from the mathematical viewpoint, we could replace conditions (ii) and (iii) above with corresponding local statements, as proposed next. To this end, we first relax the set of admissible strategies to

$$
\begin{equation*}
\mathcal{A}^{l}=\left\{\pi: \text { self-financing with } \pi_{t} \in \mathcal{F}_{t} \text { and } \mathbb{P}\left(\int_{0}^{t}\left|\sigma_{s} \pi_{s}\right|^{2} d s<\infty\right)=1, t>0\right\} \tag{12}
\end{equation*}
$$

Definition 2. An $\mathcal{F}_{t}$-adapted process $U_{t}(x)$ is a local forward performance if for $t \geq 0$ and $x \in \mathbb{R}$ :
(i) the mapping $x \rightarrow U_{t}(x)$ is strictly concave and increasing,
(ii) for each $\pi \in \mathcal{A}^{l}$ the process $U_{t}\left(X_{t}^{\pi}\right)$ is a local supermartingale, and
(iii) there exists $\pi^{*} \in \mathcal{A}^{l}$ such that the process $U_{t}\left(X_{t}^{\pi^{*}}\right)$ is a local martingale.

Herein, we do not analyze the relaxed formulation of the problem but only present an example of a local forward performance (see Example 13).

Other modifications of the definition of the forward performance process are possible, all based on the same principle, namely, to choose an investment strategy that keeps the

[^29]expected investment performance constant across time. For example, one can relax or modify the assumption on monotonicity and (strict) concavity. This is desirable, in particular, for the development of time consistent behavioral portfolio selection models. A natural modification would be to assume the existence of a reference point for the investor's wealth that defines gains and losses (see, for example, [4]). The mapping $x \rightarrow U_{t}(x)$ should then be concave for gains and convex for losses. The supermartingality condition in the above definitions would have to be replaced by a statement about the sign of the drift in the semimartingale decomposition of $U_{t}\left(X_{t}^{\pi}\right)$. Specifically, the drift would be negative when the wealth is above the reference point and positive when below. However, such modifications and extensions are beyond the scope of this paper and are mentioned here only to expose the flexibility of our definition.

We conclude by mentioning that the classical Markowitz portfolio selection problem (see, among others, [3] and [6]) could also be incorporated into our framework. Indeed, one would need to choose the mean level of wealth and find the portfolio that deviates from it the least, in the variance sense. Note that not all mean functions would be admissible, as is already demonstrated in this paper (however, with respect to criteria not covering the case of variance). The variance-based criteria deserve a separate treatment which will be carried out in a future study.

## 3. Monotone performance processes and their optimal wealth and portfolio processes.

 We focus on the class of time-decreasing performance processes introduced by the authors in [10] (see also [12] and [9]) and provide a full characterization of the associated optimal wealth and portfolio processes.It was shown in [10] (see Theorems 4 and 8) that the performance process, $U_{t}(x)$, is constructed by compiling market related input with a deterministic function of space and time. Specifically, for $t \geq 0$, we have

$$
\begin{equation*}
U_{t}(x)=u\left(x, A_{t}\right) \tag{13}
\end{equation*}
$$

where $u(x, t)$ is increasing and strictly concave in $x$ and satisfies

$$
\begin{equation*}
u_{t}=\frac{1}{2} \frac{u_{x}^{2}}{u_{x x}} \tag{14}
\end{equation*}
$$

with $A_{t}$ as in (20) below. It was also shown that the optimal wealth and the associated investment process, denoted respectively by $X_{t}^{*}$ and $\pi_{t}^{*}, t \geq 0$, are constructed via an autonomous system of stochastic differential equations whose coefficients depend functionally on the spatial derivatives of $u$. Specifically, let $R: \mathbb{R} \times[0,+\infty) \rightarrow \mathbb{R}^{+}$be defined as

$$
\begin{equation*}
R(x, t)=-\frac{u_{x}(x, t)}{u_{x x}(x, t)} \tag{15}
\end{equation*}
$$

with $u$ satisfying (14), and define (with slight abuse of notation) the process

$$
\begin{equation*}
R_{t}^{*}=R\left(X_{t}^{*}, A_{t}\right) \tag{16}
\end{equation*}
$$

with $A_{t}, t \geq 0$, as in (20). Consider the system

$$
\left\{\begin{array}{c}
d X_{t}^{*}=R\left(X_{t}^{*}, A_{t}\right) \lambda_{t}\left(\lambda_{t} d t+d W_{t}\right)  \tag{17}\\
d R_{t}^{*}=R_{x}\left(X_{t}^{*}, A_{t}\right) d X_{t}^{*}
\end{array}\right.
$$

and its solution $\left(X_{t}^{*}, R_{t}^{*}\right), t \geq 0$. Then, the process $\pi_{t}^{*}$ defined by

$$
\begin{equation*}
\pi_{t}^{*}=R_{t}^{*} \sigma_{t}^{+} \lambda_{t} \tag{18}
\end{equation*}
$$

is optimal and generates the optimal wealth process $X_{t}^{*}$.
The main contribution of this section is the explicit construction of the optimal processes. We establish that, in analogy to the forward performance process, $X_{t}^{*}$ and $\pi_{t}^{*}$ are also given as a compilation of market input and deterministic functions of space and time. Namely, we show that

$$
X_{t}^{*}=h\left(h^{(-1)}(x, 0)+A_{t}+M_{t}, A_{t}\right) \quad \text { and } \quad \pi_{t}^{*}=h_{x}\left(h^{(-1)}\left(X_{t}^{*}, A_{t}\right), A_{t}\right) \sigma_{t}^{+} \lambda_{t}
$$

where $h(x, t)$ is strictly increasing in $x$ and solves the (backward) heat equation

$$
\begin{equation*}
h_{t}+\frac{1}{2} h_{x x}=0 \tag{19}
\end{equation*}
$$

for $(x, t) \in \mathbb{R} \times[0,+\infty)$. The function $h^{(-1)}$ stands for the spatial inverse of $h$.
The market input processes $A_{t}$ and $M_{t}, t \geq 0$, are defined as

$$
\begin{equation*}
A_{t}=\int_{0}^{t}\left|\lambda_{s}\right|^{2} d s \quad \text { and } \quad M_{t}=\int_{0}^{t} \lambda_{s} \cdot d W_{s} \tag{20}
\end{equation*}
$$

with $\lambda_{t}$ as in (3).
The above formulae demonstrate that all quantities of interest can be fully specified as long as the market price of risk is chosen and the functions $u$ and $h$ are known. A considerable part of this paper is, thus, dedicated to the study of these functions and, especially, their representation and connection with each other. For the reader's convenience, we choose to present the detailed results separately. We do this because different cases for the domain and range of the functions $u$ and $h$ require appropriately modified and computationally tedious arguments which, if presented at this point, would obscure the clarity of the presentation. It is shown in section 4 (see Propositions $10,14,15$, and 19) that there exists a one-to-one correspondence (modulo normalization constants) between increasing and strictly concave solutions to (14) with strictly increasing solutions to (19). It is also shown that the latter can be represented in terms of the bilateral Laplace transform of a positive finite Borel measure, denoted throughout by $\nu$. This measure then emerges as the defining element in the entire analysis of the problem at hand. Its presence originates from the classical results of Widder (see Chapter XIV in [17]) on the construction of positive solutions to the heat equation (19). In the investment model we consider, the solution $h$ of (19) is used to construct the optimal wealth which, however, can take arbitrary values. As a result, a more detailed study is required depending on the range of $h$.

In order to present the general ideas and provide some insights for the upcoming main theorem, we present the following representative case. We stress that the results below are not complete but are presented in this form in order to build intuition. The complete arguments are presented in Propositions 9 and 10.

To this end, we introduce the set of measures $\mathcal{B}^{+}(\mathbb{R})$ defined by

$$
\begin{equation*}
\mathcal{B}^{+}(\mathbb{R})=\left\{\nu \in \mathcal{B}(\mathbb{R}): \forall B \in \mathcal{B}, \nu(B) \geq 0 \text { and } \int_{\mathbb{R}} e^{y x} \nu(d y)<\infty, x \in \mathbb{R}\right\} . \tag{21}
\end{equation*}
$$

Proposition 3. (i) Let $\nu \in \mathcal{B}^{+}(\mathbb{R})$. Then the function $h$ defined, for $(x, t) \in \mathbb{R} \times[0,+\infty)$, by

$$
h(x, t)=\int_{\mathbb{R}} \frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y} \nu(d y)+C
$$

is a strictly increasing solution to (19).
(ii) Assume that $h$ above is of full range for each $t \geq 0$, and let $h^{(-1)}: \mathbb{R} \times[0,+\infty) \rightarrow \mathbb{R}$ be its spatial inverse. Then, the function $u$ defined for $(x, t) \in \mathbb{R} \times[0,+\infty)$ and given by

$$
u(x, t)=-\frac{1}{2} \int_{0}^{t} e^{-h^{(-1)}(x, s)+\frac{s}{2}} h_{x}\left(h^{(-1)}(x, s), s\right) d s+\int_{0}^{x} e^{-h^{(-1)}(z, 0)} d z
$$

is an increasing and strictly concave solution of (14).
We proceed with the main theorem in which we provide closed form expressions for the optimal wealth, the associated optimal investment strategy, and the space-time monotone forward performance process. We state the result without making specific reference to the range of $h$, or to the domain and range of $u$, as the different cases are analyzed in detail later. We also do not make any reference to the regularity of these functions, since the required smoothness follows trivially from their representation.

We stress, however, that we introduce the integrability condition (22). This condition is stronger than the one needed for the representations of $h$ (cf. (21)), and in turn of $u$, but sufficient in order to guarantee the admissibility of the candidate optimal policy (24). It may be relaxed if, for example, one chooses to work instead with local forward performance processes, introduced in Definition 2. For additional comments on condition (22), see the discussion after Example 13.

Theorem 4. (i) Let $h$ be a strictly increasing solution to (19), for $(x, t) \in \mathbb{R} \times[0,+\infty)$, and assume that the associated measure $\nu$ satisfies

$$
\begin{equation*}
\int_{\mathbb{R}} e^{y x+\frac{1}{2} y^{2} t} \nu(d y)<\infty . \tag{22}
\end{equation*}
$$

Let also $A_{t}$ and $M_{t}, t \geq 0$, be as in (20), and define the processes $X_{t}^{*}$ and $\pi_{t}^{*}$ by

$$
\begin{equation*}
X_{t}^{*}=h\left(h^{(-1)}(x, 0)+A_{t}+M_{t}, A_{t}\right) \tag{23}
\end{equation*}
$$

and

$$
\begin{equation*}
\pi_{t}^{*}=h_{x}\left(h^{(-1)}\left(X_{t}^{*}, A_{t}\right), A_{t}\right) \sigma_{t}^{+} \lambda_{t} \tag{24}
\end{equation*}
$$

$t \geq 0, x \in \mathbb{R}$, with $h$ as above and $h^{(-1)}$ standing for its spatial inverse. Then, the portfolio $\pi_{t}^{*}$ is admissible and generates $X_{t}^{*}$, i.e.,

$$
\begin{equation*}
X_{t}^{*}=x+\int_{0}^{t} \sigma_{s} \pi_{s}^{*} \cdot\left(\lambda_{s} d s+d W_{s}\right) \tag{25}
\end{equation*}
$$

(ii) Let $u$ be the associated with $h$ increasing and strictly concave solution to (14). Then, the process $u\left(X_{t}^{*}, A_{t}\right), t \geq 0$, satisfies

$$
\begin{equation*}
d u\left(X_{t}^{*}, A_{t}\right)=u_{x}\left(X_{t}^{*}, A_{t}\right) \sigma_{t} \pi_{t}^{*} \cdot d W_{t} \tag{26}
\end{equation*}
$$

with $X_{t}^{*}$ and $\pi_{t}^{*}$ as in (23) and (24).
(iii) Let $U_{t}(x), t \geq 0, x \in \mathbb{R}$, be given by

$$
\begin{equation*}
U_{t}(x)=u\left(x, A_{t}\right) . \tag{27}
\end{equation*}
$$

Then, $U_{t}(x)$ is a forward performance process and the processes $X_{t}^{*}$ and $\pi_{t}^{*}$ are optimal.
Proof. We provide the proof only when $h$ is of infinite range since the cases of semi-infinite range can be worked out by analogous arguments.

As was mentioned earlier, the representation of $h$ is established in section 4. When $h$ is of infinite range, it is given in Proposition 9 (cf. (39)), rewritten below for convenience, namely,

$$
h(x, t)=\int_{\mathbb{R}} \frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y} \nu(d y)
$$

for $(x, t) \in \mathbb{R} \times[0,+\infty)$ (for simplicity we take $C=0$ in (39)).
For $x \in \mathbb{R}$ and $A_{t}$ and $M_{t}$ as in (20), we then define the process

$$
N_{t}=h^{(-1)}(x, 0)+A_{t}+M_{t},
$$

where $h^{(-1)}$ is the spatial inverse of $h$. Applying Itô's formula to $X_{t}^{*}$, given in (23), and using (19) yields

$$
\begin{equation*}
d X_{t}^{*}=h_{x}\left(N_{t}, A_{t}\right) d N_{t} \tag{28}
\end{equation*}
$$

On the other hand, (23) and (24) imply

$$
\pi_{t}^{*}=h_{x}\left(N_{t}, A_{t}\right) \sigma_{t}^{+} \lambda_{t}
$$

$t \geq 0$, and (25) follows from the above and (4).
To establish that $\pi_{t}^{*} \in \mathcal{A}$, it suffices to show that the integrability condition in (7) is satisfied. Using that

$$
h_{x}(x, t)=\int_{\mathbb{R}} e^{y x-\frac{1}{2} y^{2} t} \nu(d y)
$$

(cf. (81)) and (23), we have

$$
\begin{gathered}
\left(h_{x}\left(h^{(-1)}\left(X_{t}^{*}, A_{t}\right), A_{t}\right)\right)^{2} \\
=\int_{\mathbb{R}} \int_{\mathbb{R}} e^{\left(y_{1}+y_{2}\right)\left(h^{(-1)}(x, 0)+A_{t}+M_{t}\right)-\frac{1}{2}\left(y_{1}^{2}+y_{2}^{2}\right) A_{t}} \nu\left(d y_{1}\right) \nu\left(d y_{2}\right) .
\end{gathered}
$$

From (4), Fubini's theorem, and (20), we deduce

$$
\begin{aligned}
& E\left(\int_{0}^{t}\left|\sigma_{s} \pi_{s}^{*}\right|^{2} d s\right)=E\left(\int_{0}^{t}\left|h_{x}\left(h^{(-1)}\left(X_{s}^{*}, A_{s}\right), A_{s}\right) \sigma_{s} \sigma_{s}^{+} \lambda_{s}\right|^{2} d s\right) \\
&=E\left(\int_{0}^{t}\left(h_{x}\left(h^{(-1)}\left(X_{s}^{*}, A_{s}\right), A_{s}\right)\right)^{2} d A_{s}\right) \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} E\left(\int_{0}^{t} e^{\left(y_{1}+y_{2}\right)\left(h^{(-1)}(x, 0)+A_{s}+M_{s}\right)-\frac{1}{2}\left(y_{1}^{2}+y_{2}^{2}\right) A_{s}} d A_{s}\right) \nu\left(d y_{1}\right) \nu\left(d y_{2}\right) \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} E\left(\int_{0}^{A_{t}} e^{\left(y_{1}+y_{2}\right)\left(h^{(-1)}(x, 0)+s+\beta_{s}\right)-\frac{1}{2}\left(y_{1}^{2}+y_{2}^{2}\right) s} d s\right) \nu\left(d y_{1}\right) \nu\left(d y_{2}\right),
\end{aligned}
$$

where $\beta_{t}=M_{A_{t}^{(-1)}}$ and $A_{t}^{(-1)}$ stands for the inverse of $A_{t}, t \geq 0$. Using that $\beta_{t}, t \geq 0$, is normally distributed with mean 0 and variance $t$, we obtain

$$
\begin{gathered}
E\left(\int_{0}^{t}\left|\sigma_{s} \pi_{s}^{*}\right|^{2} d s\right) \\
\leq \int_{\mathbb{R}} \int_{\mathbb{R}} E\left(\int_{0}^{c^{2} t} e^{\left(y_{1}+y_{2}\right)\left(h^{(-1)}(x, 0)+s+\beta_{s}\right)-\frac{1}{2}\left(y_{1}^{2}+y_{2}^{2}\right) s} d s\right) \nu\left(d y_{1}\right) \nu\left(d y_{2}\right) \\
=\int_{0}^{c^{2} t} \int_{\mathbb{R}} \int_{\mathbb{R}} e^{\left(y_{1}+y_{2}\right)\left(h^{(-1)}(x, 0)+s\right)+y_{1} y_{2} s} \nu\left(d y_{1}\right) \nu\left(d y_{2}\right) d s \\
\leq \int_{0}^{c^{2} t}\left(\int_{\mathbb{R}} e^{y\left(h^{(-1)}(x, 0)+s\right)+\frac{1}{2} y^{2} s} \nu(d y)\right)^{2} d s
\end{gathered}
$$

and using (22), we conclude.
(ii) The facts that $u$ satisfies (14) and has the claimed monotonicity and strict concavity properties are established separately, in Proposition 10, where a detailed construction of this function is presented.

To show (26), we apply Itô's formula to $u\left(X_{t}^{*}, A_{t}\right), t \geq 0$. To this end, using (28) yields

$$
\begin{aligned}
d u\left(X_{t}^{*}, A_{t}\right)= & u_{x}\left(X_{t}^{*}, A_{t}\right) d X_{t}^{*}+u_{t}\left(X_{t}^{*}, A_{t}\right) d A_{t}+\frac{1}{2} u_{x x}\left(X_{t}^{*}, A_{t}\right) d\left\langle X^{*}\right\rangle_{t} \\
= & u_{x}\left(X_{t}^{*}, A_{t}\right) h_{x}\left(h^{(-1)}\left(X_{t}^{*}, A_{t}\right), A_{t}\right) \lambda_{t} \cdot d W_{t} \\
& +u_{x}\left(X_{t}^{*}, A_{t}\right) h_{x}\left(h^{(-1)}\left(X_{t}^{*}, A_{t}\right), A_{t}\right) d A_{t} \\
& +u_{t}\left(X_{t}^{*}, A_{t}\right) d A_{t}+\frac{1}{2} u_{x x}\left(X_{t}^{*}, A_{t}\right) d\left\langle X^{*}\right\rangle_{t}
\end{aligned}
$$

From (91) in the proof of Proposition 10, we deduce that

$$
\begin{equation*}
-\frac{u_{x}\left(X_{t}^{*}, A_{t}\right)}{u_{x x}\left(X_{t}^{*}, A_{t}\right)}=h_{x}\left(h^{(-1)}\left(X_{t}^{*}, A_{t}\right), A_{t}\right) \tag{29}
\end{equation*}
$$

which combined with the above yields

$$
\begin{gathered}
d u\left(X_{t}^{*}, A_{t}\right)=u_{x}\left(X_{t}^{*}, A_{t}\right) h_{x}\left(h^{(-1)}\left(X_{t}^{*}, A_{t}\right), A_{t}\right) \lambda_{t} \cdot d W_{t} \\
-\frac{\left(u_{x}\left(X_{t}^{*}, A_{t}\right)\right)^{2}}{u_{x x}\left(X_{t}^{*}, A_{t}\right)} d A_{t}+u_{t}\left(X_{t}^{*}, A_{t}\right) d A_{t}+\frac{1}{2} u_{x x}\left(X_{t}^{*}, A_{t}\right) d\left\langle X^{*}\right\rangle_{t} .
\end{gathered}
$$

On the other hand, (28) gives

$$
\begin{aligned}
u_{x x}\left(X_{t}^{*}, A_{t}\right) d\left\langle X^{*}\right\rangle_{t}= & u_{x x}\left(X_{t}^{*}, A_{t}\right)\left(h_{x}\left(h^{(-1)}\left(X_{t}^{*}, A_{t}\right), A_{t}\right)\right)^{2} d A_{t} \\
& =\frac{\left(u_{x}\left(X_{t}^{*}, A_{t}\right)\right)^{2}}{u_{x x}\left(X_{t}^{*}, A_{t}\right)} d A_{t} .
\end{aligned}
$$

Therefore,

$$
\begin{gathered}
d u\left(X_{t}^{*}, A_{t}\right)=u_{x}\left(X_{t}^{*}, A_{t}\right) h_{x}\left(h^{(-1)}\left(X_{t}^{*}, A_{t}\right), A_{t}\right) \lambda_{t} \cdot d W_{t} \\
+\left(u_{t}\left(X_{t}^{*}, A_{t}\right)-\frac{1}{2} \frac{\left(u_{x}\left(X_{t}^{*}, A_{t}\right)\right)^{2}}{u_{x x}\left(X_{t}^{*}, A_{t}\right)}\right) d A_{t},
\end{gathered}
$$

and (26) follows since $u$ satisfies (14).
(iii) We need to establish that $U_{t}(x)$ satisfies all conditions in Definition 1. The facts that $u\left(x, A_{t}\right)$ is $\mathcal{F}_{t}$-adapted and that the mapping $x \rightarrow u\left(x, A_{t}\right)$ is increasing and strictly concave follow trivially from the properties of $u$ and $A_{t}$.

To establish the integrability condition $E\left(U_{t}\left(X_{t}^{\pi}\right)\right)^{+}<\infty$, we work as follows. We first observe that the strict concavity of $u$ together with (14) yields $u_{t}<0$ and, hence, $u(x, t) \leq$ $u(x, 0) \leq a x^{+}+b$, for some positive constants $a$ and $b$. Also, (6) implies

$$
\left(X_{t}^{\pi}\right)^{+} \leq x^{+}+\frac{1}{2} \int_{0}^{t}\left|\sigma_{s} \pi_{s}\right|^{2} d s+\frac{1}{2} \int_{0}^{t}\left|\lambda_{s}\right|^{2} d s+\left|\int_{0}^{t} \sigma_{s} \pi_{s} \cdot d W_{s}\right| .
$$

The integrability of $E\left(U_{t}\left(X_{t}^{\pi}\right)\right)^{+}$then follows, using (5) and that $\pi_{t} \in \mathcal{A}$.
To show (8), we observe that for $\pi_{t} \in \mathcal{A}$ and $X_{t}^{\pi}$ as in (6), Itô's formula yields

$$
\begin{aligned}
d u\left(X_{t}^{\pi}, A_{t}\right)=\left(u_{x}\left(X_{t}^{\pi}, A_{t}\right)\right. & \left.\sigma_{t} \pi_{t} \cdot \lambda_{t}+u_{t}\left(X_{t}^{\pi}, A_{t}\right)\left|\lambda_{t}\right|^{2}+\frac{1}{2} u_{x x}\left(X_{t}^{\pi}, A_{t}\right)\left|\sigma_{t} \pi_{t}\right|^{2}\right) d t \\
& +u_{x}\left(X_{t}^{\pi}, A_{t}\right) \sigma_{t} \pi_{t} \cdot d W_{t} \\
=\left(u_{x}\left(X_{t}^{\pi}, A_{t}\right) \sigma_{t} \pi_{t} \cdot \lambda_{t}+\right. & \left.\frac{1}{2} \frac{\left(u_{x}\left(X_{t}^{\pi}, A_{t}\right)\right)^{2}}{u_{x x}\left(X_{t}^{\pi}, A_{t}\right)}\left|\lambda_{t}\right|^{2}+\frac{1}{2} u_{x x}\left(X_{t}^{\pi}, A_{t}\right)\left|\sigma_{t} \pi_{t}\right|^{2}\right) d t \\
& +u_{x}\left(X_{t}^{\pi}, A_{t}\right) \sigma_{t} \pi_{t} \cdot d W_{t} \\
\left.=\frac{1}{2} u_{x x}\left(X_{t}^{\pi}, A_{t}\right) \right\rvert\, \sigma_{t} \pi_{t} & +\left.\frac{u_{x}\left(X_{t}^{\pi}, A_{t}\right)}{u_{x x}\left(X_{t}^{\pi}, A_{t}\right)} \lambda_{t}\right|^{2} d t+u_{x}\left(X_{t}^{\pi}, A_{t}\right) \sigma_{t} \pi_{t} \cdot d W_{t}
\end{aligned}
$$

where we used that $u$ solves (14). Using the concavity of $u$, we conclude.

To show (9) we use the form of the above drift, (29), and (24).
We remind the reader that the forward performance process in [10] is more general than the one in (13); namely, it is given by

$$
\begin{equation*}
U_{t}(x)=u\left(\frac{x}{Y_{t}}, \tilde{A}_{t}\right) Z_{t} \tag{30}
\end{equation*}
$$

where the processes $\left(Y_{t}, Z_{t}\right)$ represent, respectively, a benchmark (or numeraire) and alternative market views. They solve

$$
d Y_{t}=Y_{t} \delta_{t} \cdot\left(\lambda_{t} d t+d W_{t}\right) \quad \text { and } \quad d Z_{t}=Z_{t} \phi_{t} \cdot d W_{t}
$$

with $Y_{0}=Z_{0}=1$ and $\delta_{t}, \phi_{t}$ being $\mathcal{F}_{t}$-adapted processes, satisfying $\sigma_{t} \sigma_{t}^{+} \delta_{t}=\delta_{t}$ and $\sigma_{t} \sigma_{t}^{+} \phi_{t}=$ $\phi_{t}, t \geq 0$. The process $\tilde{A}_{t}$ has a form similar to (20),

$$
\tilde{A}_{t}=\int_{0}^{t}\left|\lambda_{s}+\phi_{s}-\delta_{s}\right|^{2} d s
$$

Herein, we assume $\delta_{t}=\phi_{t}=0, t \geq 0$, throughout, as we focus on monotone in time forward performance processes. It is immediate, as (30) shows, that the more general form of the forward process can be readily constructed once the function $u$ is specified and the market input processes $A_{t}, Y_{t}$, and $Z_{t}$ (which are independent of $u$ ) are chosen.
3.1. Dependence on the initial wealth. The explicit formulae (23) and (24) enable us to analyze the mappings $x \rightarrow X_{t}^{*}(\omega)$ and $x \rightarrow \pi_{t}^{*}(\omega)$ for fixed $t$ and $\omega$. We study this dependence next. To ease the presentation, we discuss only the case Range $(h)=(-\infty,+\infty)$. We also use the notation $X_{t}^{*, x}(\omega)$ and $\pi_{t}^{*, x}$ and introduce the function $r: \mathbb{R} \times[0,+\infty) \rightarrow(0,+\infty)$, defined as

$$
\begin{equation*}
r(x, t)=h_{x}\left(h^{(-1)}(x, t), t\right) . \tag{31}
\end{equation*}
$$

A detailed discussion on its role, representation, and differential properties is provided in section 4.5. Using (31), the optimal portfolio (cf. (24)) can then be written, for $t \geq 0$, as

$$
\begin{equation*}
\pi_{t}^{*, x}=r\left(X_{t}^{*, x}, A_{t}\right) \sigma_{t}^{+} \lambda_{t} . \tag{32}
\end{equation*}
$$

Proposition 5. Let $X_{t}^{*, x}$ be given in (23), $t \geq 0$, and $r$ be as in (31). Then,

$$
\begin{equation*}
\frac{\partial}{\partial x} X_{t}^{*, x}=\frac{r\left(X_{t}^{*, x}, A_{t}\right)}{r(x, 0)} \tag{33}
\end{equation*}
$$

and

$$
\begin{equation*}
\frac{\partial}{\partial x} \pi_{t}^{*, x}=r_{x}\left(X_{t}^{*, x}, A_{t}\right) \frac{r\left(X_{t}^{*, x}, A_{t}\right)}{r(x, 0)} \sigma_{t}^{+} \lambda_{t} . \tag{34}
\end{equation*}
$$

Proof. Differentiating (23) with respect to $x$ yields

$$
\frac{\partial}{\partial x} X_{t}^{*, x}=h_{x}\left(h^{(-1)}(x, 0)+A_{t}+M_{t}, A_{t}\right) \frac{\partial}{\partial x} h^{(-1)}(x, 0),
$$

and (33) follows from (31). To establish (34) we differentiate (32) and use (33).
The above result implies that the mapping $x \rightarrow X_{t}^{*, x}$ is increasing. This is to be expected because the larger the initial endowment, the larger the future wealth should be. It also shows that the mapping $x \rightarrow \pi_{t}^{*, x}$ is increasing (or decreasing) depending on the monotonicity of the function $r$ and the sign of $\lambda_{t}$. In general, the latter is not monotone, and therefore nothing specific can be said about the dependence of the optimal allocation in terms of the initial endowment. ${ }^{3}$ The monotonicity holds, however, in a special but frequently considered case, namely when there is no bankruptcy, or more generally when the wealth stays always above a certain threshold. This case is considered in Proposition 23 herein, where it is shown that $r_{x} \geq 0$ (see (65)). As a result, the mapping $x \rightarrow \pi_{t}^{*, x}$ is always increasing. Respectively, the other results in the same proposition show that the mapping $x \rightarrow \pi_{t}^{*, x}$ is always decreasing if the wealth stays below a threshold.

The optimal wealth formula (23) enables us to calculate higher order derivatives. For example, the second order derivative is given below.

Proposition 6. Let $X_{t}^{*, x}, t \geq 0$, be given in (23), and let $r$ be as in (31). Then,

$$
\frac{\partial^{2}}{\partial x^{2}} X_{t}^{*, x}=\frac{r_{x}\left(X_{t}^{*, x}, A_{t}\right)-r_{x}(x, 0)}{r(x, 0)} \frac{\partial}{\partial x} X_{t}^{*, x}
$$

Proof. Differentiating (33) yields

$$
\frac{\partial^{2}}{\partial x^{2}} X_{t}^{*, x}=\frac{r_{x}\left(X_{t}^{*, x}, A_{t}\right)}{r(x, 0)} \frac{\partial}{\partial x} X_{t}^{*, x}-\frac{r_{x}(x, 0)}{r(x, 0)} \frac{r\left(X_{t}^{*, x}, A_{t}\right)}{r(x, 0)}
$$

and we easily conclude using (33) once more.
Representation (23) reveals how the market input processes, $A_{t}$ and $M_{t}, t \geq 0$, interact with the deterministic input, $h$, to generate the optimal wealth process. The function $h$ is, on the other hand, fully specified by the measure $\nu$. It is, then, natural to ask how the function $h$ and, in turn, the process $X_{t}^{*}, t \geq 0$, depend on the total mass $\nu(\mathbb{R})$.

The result below shows an interesting scaling property which allows us to normalize the function $h$ and assume that $\nu$ is a probability measure. For simplicity, we discuss only the case Range $(h)=(-\infty,+\infty)$.

Let $h_{0}=\nu(\mathbb{R})$, and denote, with a slight abuse of notation, the associated wealth process by $X_{t}^{*}\left(x ; h_{0}\right), t \geq 0$.

Proposition 7. For $h_{0}=\nu(\mathbb{R})$, the optimal wealth process (cf. (23)) satisfies, for $t \geq 0$,

$$
\frac{1}{h_{0}} X_{t}^{*}\left(x ; h_{0}\right)=X_{t}^{*}\left(\frac{x}{h_{0}} ; 1\right) .
$$

Proof. Let $\bar{h}(x, t)=\frac{h(x, t)}{h_{0}}$. Then,

$$
X_{t}^{*}\left(x ; h_{0}\right)=h_{0} \bar{h}\left(h^{(-1)}(x, 0)+A_{t}+M_{t}, A_{t}\right) .
$$

[^30]On the other hand, $h^{(-1)}(x, 0)=\bar{h}^{(-1)}\left(\frac{x}{h_{0}}, 0\right)$, and, hence,

$$
X_{t}^{*}\left(x ; h_{0}\right)=h_{0} \bar{h}\left(\bar{h}^{(-1)}\left(\frac{x}{h_{0}}, 0\right)+A_{t}+M_{t}, A_{t}\right)=h_{0} X_{t}^{*}\left(\frac{x}{h_{0}} ; 1\right) .
$$

4. Representation of the functions $u$ and $h$. The functions $u$ and $h$ were instrumental in the construction of the forward performance and the associated optimal wealth and portfolio processes (Theorem 4). In this section, we focus on the representation of these functions and their connection with each other. We recall that they satisfy (14) and (19), respectively, and that we are interested in solutions of (14) that are increasing and strictly concave in their spatial argument. We will show that there is a one-to-one correspondence (modulo normalization constants) between these functions and strictly increasing solutions to the (backward) heat equation (19).

As was discussed in the previous section (see representative results in Proposition 3), the key idea is to represent $h$ in terms of a finite positive Borel measure $\nu$ and, in turn, construct $u$ from $h$. This measure, then, emerges as the defining element in the construction of any object of interest. The main assumption about $\nu$ is that its bilateral Laplace transform exists. ${ }^{4}$ Namely, we will be working throughout this section with measures belonging to $\mathcal{B}^{+}(\mathbb{R})$, given in (21). The connection between $\nu$ and $h$ originates from the classical result of Widder (see [17]) for nonnegative solutions of (19). For completeness and motivation we present this result below.

Theorem 8 (Widder). Let $g(x, t),(x, t) \in \mathbb{R} \times[0,+\infty)$, be a positive solution of (19). Then, there exists $\mu \in \mathcal{B}^{+}(\mathbb{R})$ such that $g$ is represented as

$$
\begin{equation*}
g(x, t)=\int_{\mathbb{R}} e^{y x-\frac{1}{2} y^{2} t} \mu(d y) . \tag{35}
\end{equation*}
$$

This result cannot be applied directly herein because, for the investment applications we consider, the wealth may not be assumed to remain always positive or, more generally, stay above (or below) a given threshold. As a consequence, different choices for the range of $h$, which represents the optimal wealth (cf. (23)), require different analysis. However, Widder's theorem will be applied to the function $h_{x}$ which is positive (due to the assumed monotonicity of $h$ ) and also solves (19).

We start with the general theorem which gives us the representation of strictly increasing solutions to the heat equation (19). Its proof as well as all other proofs in this section are presented in the appendix.

We introduce the following sets:

$$
\begin{align*}
& \mathcal{B}_{0}^{+}(\mathbb{R})=\left\{\nu \in \mathcal{B}^{+}(\mathbb{R}) \text { and } \nu(\{0\})=0\right\},  \tag{36}\\
& \mathcal{B}_{+}^{+}(\mathbb{R})=\left\{\nu \in \mathcal{B}_{0}^{+}(\mathbb{R}): \nu((-\infty, 0))=0\right\} \tag{37}
\end{align*}
$$

and

$$
\begin{equation*}
\mathcal{B}_{-}^{+}(\mathbb{R})=\left\{\nu \in \mathcal{B}_{0}^{+}(\mathbb{R}): \nu((0,+\infty))=0\right\} . \tag{38}
\end{equation*}
$$

[^31]It is assumed throughout that the trivial case $\nu(\mathbb{R})=0$ is excluded.
In what follows, $C$ represents a generic constant. Special choices for it are discussed later.
Proposition 9. (i) Let $\nu \in \mathcal{B}^{+}(\mathbb{R})$. Then, the function $h$ defined, for $(x, t) \in \mathbb{R} \times[0,+\infty)$, by

$$
\begin{equation*}
h(x, t)=\int_{\mathbb{R}} \frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y} \nu(d y)+C \tag{39}
\end{equation*}
$$

is a strictly increasing solution to (19).
Moreover, if $\nu(\{0\})>0$, or $\nu(-\infty, 0)>0$ and $\nu(0,+\infty)>0$, or $\nu \in \mathcal{B}_{+}^{+}(\mathbb{R})$ and $\int_{0^{+}}^{+\infty} \frac{\nu(d y)}{y}=+\infty$, or $\nu \in \mathcal{B}_{-}^{+}(\mathbb{R})$ and $\int_{-\infty}^{0^{-}} \frac{\nu(d y)}{y}=-\infty$, then Range $(h)=(-\infty,+\infty)$ for $t \geq 0$.

On the other hand, if $\nu \in \mathcal{B}_{+}^{+}(\mathbb{R})$ with $\int_{0^{+}}^{+\infty} \frac{\nu(d y)}{y}<+\infty$ (resp., $\nu \in \mathcal{B}_{-}^{+}(\mathbb{R})$ with $\left.\int_{-\infty}^{0^{-}} \frac{\nu(d y)}{y}>-\infty\right)$, then Range $(h)=\left(C-\int_{0^{+}}^{+\infty} \frac{\nu(d y)}{y},+\infty\right)$ (resp., Range $(h)=(-\infty, C-$ $\left.\left.\int_{-\infty}^{0^{-}} \frac{\nu(d y)}{y}\right)\right)$ for $t \geq 0$.
(ii) Conversely, let $h: \mathbb{R} \times[0,+\infty) \rightarrow \mathbb{R}$ be a strictly increasing solution to (19). Then, there exists $\nu \in \mathcal{B}^{+}(\mathbb{R})$ such that $h$ is given by (39).

Moreover, if Range $(h)=(-\infty,+\infty), t \geq 0$, then it must be either that $\nu(\{0\})>0$, or $\nu(-\infty, 0)>0$ and $\nu(0,+\infty)>0$, or $\nu \in \mathcal{B}_{+}^{+}(\mathbb{R})$ and $\int_{0^{+}}^{+\infty} \frac{\nu(d y)}{y}=+\infty$, or $\nu \in \mathcal{B}_{-}^{+}(\mathbb{R})$ and $\int_{-\infty}^{0^{-}} \frac{\nu(d y)}{y}=-\infty$.

On the other hand, if Range $(h)=\left(x_{0},+\infty\right)$ (resp., Range $(h)=\left(-\infty, x_{0}\right)$ ), $t \geq 0$ and $x_{0} \in \mathbb{R}$, then it must be that $\nu \in \mathcal{B}_{+}^{+}(\mathbb{R})$ with $\int_{0^{+}}^{+\infty} \frac{\nu(d y)}{y}<+\infty$ (resp., $\nu \in \mathcal{B}_{-}^{+}(\mathbb{R})$ with $\left.\int_{-\infty}^{0^{-}} \frac{\nu(d y)}{y}>-\infty\right)$.

We continue with the representation of increasing and strictly concave solutions to (14). As mentioned earlier, we will show that there is a one-to-one correspondence (modulo normalization constants) between this class and that of strictly increasing solutions to (19).
4.1. Range $(h)=(-\infty,+\infty)$. We recall that $h$ is given by (39), for $(x, t) \in \mathbb{R} \times$ $[0,+\infty)$. For convenience, we choose $C=0$, and thus

$$
\begin{equation*}
h(0,0)=0 . \tag{40}
\end{equation*}
$$

We show how to construct from such an $h$ a globally defined, increasing, and strictly concave solution $u$ to (14). We also show the converse construction.

Note that from the properties of $u$ we would have $u_{x}(x, t) \neq 0$ and $|u(x, t)|<+\infty, t \geq 0$, for $(x, t) \in \mathbb{R} \times[0,+\infty)$. In addition, solutions of (14) are invariant with respect to affine transformations. Therefore, if $u$ is a solution, the function

$$
\hat{u}(x, t)=\frac{1}{u_{x}\left(x_{0}, 0\right)} u(x, t)-\frac{u\left(x_{0}, 0\right)}{u_{x}\left(x_{0}, 0\right)}
$$

is also a solution, for each $x_{0} \in \mathbb{R}$. Without loss of generality, we may choose, as in (40), $x_{0}=0$ to be a reference point. We then assume that

$$
\begin{equation*}
u(0,0)=0 \quad \text { and } \quad u_{x}(0,0)=1 \tag{41}
\end{equation*}
$$

Note, however, that while the first equality is imposed in an ad hoc way, the second one is in accordance with (40) (see (88) in the proof of next proposition).

Proposition 10. (i) Let $\nu \in \mathcal{B}^{+}(\mathbb{R})$ and $h: \mathbb{R} \times[0,+\infty) \rightarrow \mathbb{R}$ be as in (39) with the measure $\nu$ being used. Assume that $h$ is of full range, for each $t \geq 0$, and let $h^{(-1)}: \mathbb{R} \times[0,+\infty) \rightarrow \mathbb{R}$ be its spatial inverse. Then, the function $u$ defined for $(x, t) \in \mathbb{R} \times[0,+\infty)$ and given by

$$
\begin{equation*}
u(x, t)=-\frac{1}{2} \int_{0}^{t} e^{-h^{(-1)}(x, s)+\frac{s}{2}} h_{x}\left(h^{(-1)}(x, s), s\right) d s+\int_{0}^{x} e^{-h^{(-1)}(z, 0)} d z \tag{42}
\end{equation*}
$$

is an increasing and strictly concave solution of (14) satisfying (41).
Moreover, for $t \geq 0$, the Inada conditions,

$$
\begin{equation*}
\lim _{x \rightarrow-\infty} u_{x}(x, t)=+\infty \quad \text { and } \quad \lim _{x \rightarrow+\infty} u_{x}(x, t)=0 \tag{43}
\end{equation*}
$$

are satisfied.
(ii) Conversely, let $u$ be an increasing and strictly concave function satisfying, for $(x, t) \in$ $\mathbb{R} \times[0,+\infty)$, (14) and (41) and the Inada conditions (43), for $t \geq 0$. Then, there exists $\nu \in$ $\mathcal{B}^{+}(\mathbb{R})$ such that $u$ admits representation (42) with $h$ given by (39), for $(x, t) \in \mathbb{R} \times[0,+\infty)$. Moreover, $h$ is of full range, for each $t \geq 0$, and satisfies (40).

Example 11. Let $\nu=\delta_{0}$, where $\delta_{0}$ is a Dirac measure at 0 . Then, (39) yields

$$
h(x, t)=x,
$$

and therefore (42) implies

$$
u(x, t)=-\frac{1}{2} \int_{0}^{t} e^{-x+\frac{s}{2}} d s+\int_{0}^{x} e^{-z} d z=1-e^{-x+\frac{t}{2}}
$$

This class of forward performance processes is analyzed in detail in [9].
Example 12. Let $\nu(d y)=\frac{b}{2}\left(\delta_{a}+\delta_{-a}\right), a, b>0$, and let $\delta_{ \pm a}$ be Dirac measures at $\pm a$. We then have

$$
h(x, t)=\frac{b}{a} e^{-\frac{1}{2} a^{2} t} \sinh (a x) .
$$

Thus,

$$
h^{(-1)}(x, t)=\frac{1}{a} \ln \left(\frac{a}{b} x e^{\frac{1}{2} a^{2} t}+\sqrt{\frac{a^{2}}{b^{2}} x^{2} e^{a^{2} t}+1}\right)
$$

and, in turn,

$$
\begin{gathered}
h_{x}\left(h^{(-1)}(x, t), t\right)=b e^{-\frac{1}{2} a^{2} t} \cosh \left(\ln \left(\frac{a}{b}{ }^{\frac{1}{2} a^{2} t}+\sqrt{\frac{a^{2}}{b^{2}} x^{2} e^{a^{2} t}+1}\right)\right) \\
=\sqrt{a^{2} x^{2}+b^{2} e^{-a^{2} t}} .
\end{gathered}
$$

If, $a=1$, then (42) yields

$$
u(x, t)=\frac{1}{2}\left(\ln \left(x+\sqrt{x^{2}+b^{2} e^{-t}}\right)-\frac{e^{t}}{b^{2}} x\left(x-\sqrt{x^{2}+b^{2} e^{-t}}\right)-\frac{t}{2}\right)-\frac{1}{2} \ln b,
$$

while, if $a \neq 1$,

$$
u(x, t)=\frac{\sqrt[a]{a}}{a^{2}-1} e^{\frac{1-a}{2} t} \frac{b^{2} e^{-a^{2} t}+a(1+a)\left(a x^{2}+x \sqrt{a^{2} x^{2}+b^{2} e^{-a^{2} t}}\right)}{\left(a x+\sqrt{a^{2} x^{2}+b^{2} e^{-a^{2} t}}\right)^{1+\frac{1}{a}}}-\frac{\sqrt[a]{a}}{a^{2}-1} b^{1-\frac{1}{a}}
$$

The calculations involved are cumbersome and, for this, omitted. A complete description of this class of solutions can be found in [18].

It is worth mentioning that the above functions provide an interesting extension of the traditional power and logarithmic utilities, most frequently used in portfolio choice. Note, however, that the latter utilities are not globally defined, while the above are.

Example 13. Let $\nu(d y)=\frac{1}{\sqrt{2 \pi}} e^{-\frac{1}{2} y^{2}} d y$. Then,

$$
\begin{equation*}
h(x, t)=F\left(\frac{x}{\sqrt{t+1}}\right) \quad \text { with } \quad F(x)=\int_{0}^{x} e^{\frac{1}{2} z^{2}} d z, x \in \mathbb{R} . \tag{44}
\end{equation*}
$$

Therefore, $h^{(-1)}(x, t)=\sqrt{t+1} F^{(-1)}(x)$ and thus,

$$
\begin{equation*}
h_{x}\left(h^{(-1)}(x, t), t\right)=\frac{1}{\sqrt{t+1}} f\left(F^{(-1)}(x)\right) \tag{45}
\end{equation*}
$$

with $f(x)=F^{\prime}(x)$. Then, (42) becomes

$$
u(x, t)=-\frac{1}{2} \int_{0}^{t} \frac{1}{\sqrt{s+1}} f\left(F^{(-1)}(x)\right) e^{-\sqrt{s+1} F^{(-1)}(x)+\frac{s}{2}} d s+\int_{0}^{x} e^{-F^{(-1)}(z)} d z
$$

It turns out that

$$
\begin{equation*}
u(x, t)=k_{1} F\left(F^{(-1)}(x)-\sqrt{t+1}\right)+k_{2} \tag{46}
\end{equation*}
$$

with $k_{1}=e^{-\frac{1}{2}}$ and $k_{2}=e^{-\frac{1}{2}} \int_{-1}^{0} e^{\frac{1}{2} z^{2}} d z$.
The calculations are rather tedious, but one can verify that $u$ satisfies (41) and solves (42). Indeed,

$$
u_{t}(x, t)=-k_{1} \frac{f\left(F^{(-1)}(x)-\sqrt{t+1}\right)}{2 \sqrt{t+1}}, \quad u_{x}(x, t)=k_{1} \frac{f\left(F^{(-1)}(x)-\sqrt{t+1}\right)}{f\left(F^{(-1)}(x)\right)},
$$

and

$$
u_{x x}(x, t)=-k_{1} \sqrt{t+1} \frac{f\left(F^{(-1)}(x)-\sqrt{t+1}\right)}{\left(f\left(F^{(-1)}(x)\right)\right)^{2}}
$$

and (14) follows. The equalities in (41) also follow from the form of $u$ and the choice of the constants $k_{1}, k_{2}$. Note, moreover, that the above yields

$$
u_{x}\left(F\left(\frac{x}{\sqrt{t+1}}\right), t\right)=e^{-x+\frac{t}{2}}
$$

and (44) follows from (89).
From (24) and (45), we deduce that the optimal policy of the above example turns out to be

$$
\begin{equation*}
\pi_{t}^{*}=\frac{1}{\sqrt{A_{t}+1}} f\left(F^{(-1)}\left(X_{t}^{*}\right)\right) \sigma_{t}^{+} \lambda_{t} \tag{47}
\end{equation*}
$$

with $A_{t}, t \geq 0$, as in (20) and

$$
X_{t}^{*}=F\left(\frac{F^{(-1)}(x)+A_{t}+M_{t}}{\sqrt{A_{t}+1}}\right)
$$

with the latter following from (23) and (44).
We can see that the above measure, $\nu(d y)=\frac{1}{\sqrt{2 \pi}} e^{-\frac{1}{2} y^{2}} d y$, violates condition (22) (for $t>0)$ and satisfies only (21). In turn, straightforward calculations show that $\pi_{t}^{*}, t \geq 0$, is admissible but only in the local sense, i.e., $\pi_{t}^{*} \in \mathcal{A}^{l}$ but $\pi_{t}^{*} \notin \mathcal{A}$. We then deduce that the process $U_{t}(x)=u\left(x, A_{t}\right)$, with $u$ as in (46), satisfies Definition 2 of a local forward performance process.
4.2. Range $(h)=\left(x_{0},+\infty\right), x_{0} \in \mathbb{R}$. We recall that in this case, $h$ is given by (39), where $\nu$ satisfies

$$
\begin{equation*}
\nu \in \mathcal{B}_{+}^{+}(\mathbb{R}) \quad \text { and } \quad \int_{0^{+}}^{+\infty} \frac{\nu(d y)}{y}<+\infty \tag{48}
\end{equation*}
$$

with $\mathcal{B}_{+}^{+}(\mathbb{R})$ given in (37). For convenience, we set $C=\int_{0^{+}}^{+\infty} \frac{1}{y} \nu(d y)$ in (39), yielding ${ }^{5}$ that Range $(h)=(0,+\infty)$, as

$$
\begin{equation*}
h(x, t)=\int_{0^{+}}^{+\infty} \frac{e^{y x-\frac{1}{2} y^{2} t}}{y} \nu(d y) . \tag{49}
\end{equation*}
$$

We can easily see, using the above and (14), that $h$ is convex in its spatial argument and decreasing with regards to time,

$$
\begin{equation*}
h_{x x}(x, t)>0 \quad \text { and } \quad h_{t}(x, t)<0 . \tag{50}
\end{equation*}
$$

Next, we obtain the representation of the differential input $u$ analogous to (42). As (42) shows, $h$ plays the role of the space argument of $u$. Thus, the latter is now defined on the half-line. Consideration, then, needs to be given to $\lim _{x \rightarrow 0} u(x, t), t \geq 0$. The results below demonstrate that, depending on where the measure $\nu$ is concentrated, this limit can be finite or infinite. For the case of finite limit we have the following result.

Proposition 14. (i) Let $\nu$ satisfy (48) and, in addition, $\nu((0,1])=0$ and $\int_{1+}^{+\infty} \frac{\nu(d y)}{y-1}<+\infty .{ }^{6}$ Let also $h: \mathbb{R} \times[0,+\infty) \rightarrow(0,+\infty)$ be as in (49) and $h^{(-1)}:(0,+\infty) \times[0,+\infty) \rightarrow \mathbb{R}$ be its

[^32]spatial inverse. Then, the function $u$ defined, for $(x, t) \in(0,+\infty) \times[0,+\infty)$, by
\[

$$
\begin{equation*}
u(x, t)=-\frac{1}{2} \int_{0}^{t} e^{-h^{(-1)}(x, s)+\frac{s}{2}} h_{x}\left(h^{(-1)}(x, s), s\right) d s+\int_{0}^{x} e^{-h^{(-1)}(z, 0)} d z \tag{51}
\end{equation*}
$$

\]

is an increasing and strictly concave solution of (14) with

$$
\begin{equation*}
\lim _{x \rightarrow 0} u(x, t)=0 \quad \text { for } t \geq 0 \tag{52}
\end{equation*}
$$

Moreover, for $t \geq 0$, the Inada conditions

$$
\begin{equation*}
\lim _{x \rightarrow 0} u_{x}(x, t)=+\infty \quad \text { and } \quad \lim _{x \rightarrow+\infty} u_{x}(x, t)=0 \tag{53}
\end{equation*}
$$

are satisfied.
(ii) Conversely, let $u$, defined for $(x, t) \in(0,+\infty) \times[0,+\infty)$, be an increasing and strictly concave function satisfying (14), (52), and the Inada conditions (53). Then, there exists $\nu \in$ $\mathcal{B}^{+}(\mathbb{R})$ satisfying $(48), \nu((0,1])=0$, and $\int_{1+}^{+\infty} \frac{\nu(d y)}{y-1}<+\infty$ such that $u$ admits representation (51) with $h$ given by (49), for $(x, t) \in \mathbb{R} \times[0,+\infty)$.

Working along similar arguments, we obtain the result covering the case of an infinite limit.

Proposition 15. (i) Let $\nu$ satisfy (48) and, in addition, either $\nu((0,1])>0$ or $\nu((0,1])=0$ and $\int_{1+}^{+\infty} \frac{\nu(d y)}{y-1}=+\infty$. Let also $h: \mathbb{R} \times[0,+\infty) \rightarrow(0,+\infty)$ be as in (49) and let $h^{(-1)}$ : $(0,+\infty) \times[0,+\infty) \rightarrow \mathbb{R}$ be its spatial inverse. Then, the function $u$ defined, for $(x, t) \in$ $(0,+\infty) \times[0,+\infty)$, by

$$
\begin{equation*}
u(x, t)=-\frac{1}{2} \int_{0}^{t} e^{-h^{(-1)}(x, s)+\frac{s}{2}} h_{x}\left(h^{(-1)}(x, s), s\right) d s+\int_{x_{0}}^{x} e^{-h^{(-1)}(z, 0)} d z, \tag{54}
\end{equation*}
$$

for $x_{0}>0$, is an increasing and strictly concave solution of (14) with

$$
\begin{equation*}
\lim _{x \rightarrow 0} u(x, t)=-\infty \text { for } t \geq 0 \tag{55}
\end{equation*}
$$

Moreover, for each $t \geq 0$, the Inada conditions

$$
\begin{equation*}
\lim _{x \rightarrow 0} u_{x}(x, t)=+\infty \quad \text { and } \quad \lim _{x \rightarrow+\infty} u_{x}(x, t)=0 \tag{56}
\end{equation*}
$$

are satisfied.
(ii) Conversely, let $u$, defined for $(x, t) \in(0,+\infty) \times[0,+\infty)$, be an increasing and strictly concave function satisfying (14), (55), and the Inada conditions (56). Then, there exists $\nu \in \mathcal{B}^{+}(\mathbb{R})$ satisfying (48) and $\nu((0,1])>0$, or (48) and $\nu((0,1])=0$ and $\int_{1+}^{+\infty} \frac{\nu(d y)}{y-1}=+\infty$, such that $u$ admits representation (54) with $h$ given, for $(x, t) \in \mathbb{R} \times[0,+\infty)$, by (49).

Example 16. Let $\nu=\delta_{\gamma}, \gamma>1$. Then, (49) yields

$$
h(x, t)=\frac{1}{\gamma} e^{\gamma x-\frac{1}{2} \gamma^{2} t},
$$

for $(x, t) \in \mathbb{R} \times[0,+\infty)$. We, then, have $h^{(-1)}(x, t)=\ln (\gamma x)^{\frac{1}{\gamma}}+\frac{1}{2} \gamma t,(x, t) \in(0,+\infty) \times[0,+\infty)$, and, thus, $h_{x}\left(h^{(-1)}(x, t), t\right)=\gamma x$.

Since $\nu((0,1])=0, u$ is given by (51) and, therefore,

$$
\begin{gathered}
u(x, t)=-\frac{1}{2} \int_{0}^{t} \gamma x e^{-\left(\ln (\gamma x)^{\frac{1}{\gamma}}+\frac{1}{2} \gamma s\right)+\frac{s}{2}} d s+\int_{0}^{x}(\gamma z)^{-\frac{1}{\gamma}} d z \\
=\frac{\gamma^{\frac{\gamma-1}{\gamma}}}{\gamma-1} x^{\frac{\gamma-1}{\gamma}} e^{-\frac{\gamma-1}{2} t}
\end{gathered}
$$

Example 17. Let $\nu=\delta_{\gamma}, \gamma=1$. Then, (49) yields

$$
h(x, t)=e^{x-\frac{1}{2} t}
$$

for $(x, t) \in \mathbb{R} \times[0,+\infty)$. We, then, have $h^{(-1)}(x, t)=\ln x+\frac{1}{2} t,(x, t) \in(0,+\infty) \times[0,+\infty)$, and, thus, $h_{x}\left(h^{(-1)}(x, t), t\right)=x$.

Since $\nu((0,1]) \neq 0, u$ is given by (54). Therefore, for $(x, t) \in(0,+\infty) \times[0,+\infty)$ and $x_{0}>0$,

$$
u(x, t)=-\frac{1}{2} \int_{0}^{t} x e^{-\left(\ln x+\frac{1}{2} s\right)+\frac{s}{2}} d s+\int_{x_{0}}^{x} \frac{1}{z} d z=\ln \frac{x}{x_{0}}-\frac{t}{2}
$$

Example 18. Let $\nu=\delta_{\gamma}, \gamma \in(0,1)$. Using Example 16 and that $\nu((0,1]) \neq 0$, we easily deduce, using (54), that $u$ is given by

$$
\begin{aligned}
u(x, t)=- & \frac{1}{2} \int_{0}^{t} \gamma x e^{-\left(\ln (\gamma x)^{\frac{1}{\gamma}}+\frac{\gamma}{2} s\right)+\frac{s}{2}} d s+\int_{x_{0}}^{x}(\gamma z)^{-\frac{1}{\gamma}} d z \\
= & -\frac{\gamma^{\frac{\gamma-1}{\gamma}}}{1-\gamma} x^{\frac{\gamma-1}{\gamma}} e^{\frac{1-\gamma}{2} t}+\frac{\gamma^{\frac{\gamma-1}{\gamma}}}{1-\gamma} x_{0}^{\frac{\gamma-1}{\gamma}}
\end{aligned}
$$

for $x_{0}>0$.
4.3. Range $(h)=\left(-\infty, x_{0}\right), x_{0} \in \mathbb{R}$. We recall that in this case, $h$ is given by (39), where $\nu$ satisfies

$$
\begin{equation*}
\nu \in \mathcal{B}_{-}^{+}(\mathbb{R}) \quad \text { and } \quad \int_{-\infty}^{0^{-}} \frac{\nu(d y)}{y}>-\infty \tag{57}
\end{equation*}
$$

with $\mathcal{B}_{-}^{+}(\mathbb{R})$ given in (38). For convenience, we set $C=\int_{-\infty}^{0^{-}} \frac{\nu(d y)}{y}$ in (39), yielding Range $(h)=$ $(-\infty, 0)$, as

$$
\begin{equation*}
h(x, t)=\int_{-\infty}^{0^{-}} \frac{e^{y x-\frac{1}{2} y^{2} t}}{y} \nu(d y) \tag{58}
\end{equation*}
$$

In analogy to (50), one can show that $h$ is concave in its spatial argument and increasing with regards to time,

$$
\begin{equation*}
h_{x x}(x, t)<0 \quad \text { and } \quad h_{t}(x, t)>0 \tag{59}
\end{equation*}
$$

The next proposition follows from a modification of the arguments used to prove Proposition 14.

Proposition 19. (i) Let $\nu$ be as in (57). Let also $h: \mathbb{R} \times[0,+\infty) \rightarrow(-\infty, 0)$ be as in (58) and $h^{(-1)}:(-\infty, 0) \times[0,+\infty) \rightarrow \mathbb{R}$ be its spatial inverse. Then, the function $u$ defined, for $(x, t) \in(-\infty, 0) \times[0,+\infty)$, by

$$
\begin{equation*}
u(x, t)=-\frac{1}{2} \int_{0}^{t} e^{-h^{(-1)}(x, s)+\frac{s}{2}} h_{x}\left(h^{(-1)}(x, s), s\right) d s-\int_{x}^{0} e^{-h^{(-1)}(z, 0)} d z \tag{60}
\end{equation*}
$$

is an increasing and strictly concave solution of (14) with

$$
\begin{equation*}
\lim _{x \rightarrow 0} u(x, t)=0 \quad \text { for } t \geq 0 \tag{61}
\end{equation*}
$$

Moreover, for each $t \geq 0$, the Inada conditions

$$
\begin{equation*}
\lim _{x \rightarrow-\infty} u_{x}(x, t)=+\infty \quad \text { and } \quad \lim _{x \rightarrow 0} u_{x}(x, t)=0 \tag{62}
\end{equation*}
$$

are satisfied.
(ii) Conversely, let $u$ be an increasing and strictly concave function satisfying, for $(x, t) \in$ $(-\infty, 0) \times[0,+\infty),(14),(61)$, and the Inada conditions (62), for $t \geq 0$. Then, there exists $\nu$ as in (57) such that $u$ admits representation (60) with $h$ given by (58).

Example 20. We take $\nu=\delta_{\gamma}, \gamma=-\frac{1}{2 k+1}, k>0$. Then, (58) yields, for $(x, t) \in(-\infty, 0) \times$ $[0,+\infty)$,

$$
h(x, t)=-\frac{1}{\gamma} e^{\gamma x-\frac{1}{2} \gamma^{2} t} .
$$

Working as in Example 16, we deduce that, for $(x, t) \in(-\infty, 0) \times[0,+\infty)$,

$$
u(x, t)=\frac{\gamma^{\frac{\gamma-1}{\gamma}}}{\gamma-1} x^{\frac{\gamma-1}{\gamma}} e^{-\frac{\gamma-1}{2} t}=-\frac{(2 k+1)^{-2 k-1}}{2(k+1)} x^{2(k+1)} e^{\frac{k+1}{2 k+1} t} .
$$

4.4. Range $(h)=\left(x_{1}, x_{2}\right), x_{1}, x_{2} \in \mathbb{R}$. The case of finite range is not considered since it does not yield a meaningful solution. Indeed, we recall the following result derived by Widder (see [17]).

Proposition 21. Let $h$ be a solution to (19) such that for $(x, t) \in \mathbb{R} \times[0, \infty),-M \leq h(x, t) \leq$ $M$, for some constant $M$. Then, $h(x, t)$ is constant.

It then easily follows that in this case the problem degenerates, as there is no strictly increasing solution to (19) or, in turn, to (14).
4.5. The local risk tolerance. In the previous section we introduced the function $r$ in (31). This function facilitates the representation of the optimal portfolio policy (cf. (32)) and, as is shown below, is represented in terms of the spatial derivatives of $u$. In the traditional maximal expected utility models, a similar quantity is used, known as the risk tolerance. We keep an analogous terminology herein.

For the generic spatial domain $\mathbb{D}$ appearing below, we have $\mathbb{D}=\mathbb{R},(0,+\infty)$ or $(-\infty, 0)$. To ease the presentation, we omit any reference to the specific range $h$ (and, thus, to the domain of $u$ ).

The following result is a direct consequence of (31) and (89) (for an alternative proof, see [10]).

Proposition 22. Let $r: \mathbb{D} \times[0, \infty) \rightarrow(0,+\infty)$ be given by (31), i.e.,

$$
\begin{equation*}
r(x, t)=h_{x}\left(h^{(-1)}(x, t), t\right) \tag{63}
\end{equation*}
$$

and let $u$ be the associated with $h$ differential utility input. Then, for $(x, t) \in \mathbb{D} \times[0, \infty)$,

$$
r(x, t)=-\frac{u_{x}(x, t)}{u_{x x}(x, t)} .
$$

Therefore, $r(x, t)=R(x, t)$ with $R(x, t)$ as in (15).
In (34) we saw that the monotonicity of the optimal investment strategy with regards to the initial endowment depends directly on the sign of the partial derivative $r_{x}(x, t)$. As was mentioned in section 3 , when the risk tolerance is defined on $\mathbb{R} \times[0,+\infty)$, very little if anything can be established for its monotonicity or limiting behavior. When, however, its domain is semi-infinite, we have the following results.

Proposition 23. Let $r: \mathbb{D} \times[0,+\infty) \rightarrow(0,+\infty)$ with $\mathbb{D}=(0,+\infty)$ or $(-\infty, 0)$. Then, for $t \geq 0$,

$$
\begin{equation*}
\lim _{x \rightarrow 0} r(x, t)=0 . \tag{64}
\end{equation*}
$$

If $\mathbb{D}=(0,+\infty)$, then

$$
\begin{equation*}
r_{x}(x, t) \geq 0, \tag{65}
\end{equation*}
$$

while if $\mathbb{D}=(-\infty, 0)$,

$$
\begin{equation*}
r_{x}(x, t) \leq 0, \tag{66}
\end{equation*}
$$

for $t \geq 0$.
Proof. We establish (64) only when $\mathbb{D}=(0,+\infty)$. Recalling that

$$
h_{x}(x, t)=\int_{0^{+}}^{+\infty} e^{y x-\frac{1}{2} y^{2} t} \nu(d y)
$$

(cf. (49)), (63) yields

$$
r(x, t)=\int_{0^{+}}^{\infty} e^{y h^{(-1)}(x, t)-\frac{1}{2} y^{2} t} \nu(d y) .
$$

Passing to the limit and using the monotone convergence theorem and (93), we conclude.
Next, we show (65). Differentiating (63) yields

$$
r_{x}(x, t)=\left(\frac{\partial}{\partial x} h^{(-1)}(x, t)\right) h_{x x}\left(h^{(-1)}(x, t), t\right) .
$$

When $\mathbb{D}=(0,+\infty)$ (resp., $\mathbb{D}=(-\infty, 0)$ ), then (65) (resp., (66)) follows from (50) (resp., (59)).
5. Deterministic market prices of risk. In this section we assume that the process $\lambda_{t}, t \geq$ 0 , (cf. (3)) is deterministic. This, in turn, yields that $A_{t}, t \geq 0$, (see (20)) is deterministic.

The goal is twofold. First, we study distributional properties of the optimal wealth and compute its cumulative distribution, density, and moments. Second, we explore some inverse problems, namely, how the investor's preferences could be inferred from information about the targeted mean of his optimal wealth.
5.1. Distribution of the optimal wealth process. Recalling $A_{t}$ and $M_{t}$ from (20), we have $\langle M\rangle_{t}=A_{t}, t \geq 0$, and thus, by Levy's theorem, the process $M_{t}$ is a Gaussian martingale. This leads to the following properties of the distribution of the investor's optimal wealth process. The functions $N$ and $n$ below stand, respectively, for the cumulative distribution and the density functions of a standard normal variable. We recall that $h$ solves (19), $h^{(-1)}$ stands for its spatial inverse, and $r$ is given in (31).

Proposition 24. (i) The cumulative distribution and probability density functions of the optimal wealth $X_{t}^{*, x}, t>0$, are given, respectively, by

$$
\begin{equation*}
\mathbb{P}\left(X_{t}^{*, x} \leq y\right)=N\left(\frac{h^{(-1)}\left(y, A_{t}\right)-h^{(-1)}(x, 0)-A_{t}}{\sqrt{A_{t}}}\right) \tag{67}
\end{equation*}
$$

and

$$
f_{X_{t}^{*, x}}(y)=n\left(\frac{h^{(-1)}\left(y, A_{t}\right)-h^{(-1)}(x, 0)-A_{t}}{\sqrt{A_{t}}}\right) \frac{1}{r\left(y, A_{t}\right) \sqrt{A_{t}}}
$$

with $A_{t}$ as in (20).
(ii) For all $p \in[0,1]$ and $t>0$, the quantile of order $p$, i.e., the point $y_{p}(t)$ for which $\mathbb{P}\left(X_{t}^{*, x} \leq y_{p}(t)\right)=p$, is given by

$$
y_{p}(t)=h\left(h^{(-1)}(x, 0)+A_{t}+\sqrt{A_{t}} N^{(-1)}(p), A_{t}\right)
$$

Proof. The first statement follows directly from (23). Indeed,

$$
\begin{gathered}
\mathbb{P}\left(X_{t}^{*, x} \leq y\right)=\mathbb{P}\left(h\left(h^{(-1)}(x, 0)+A_{t}+M_{t}, A_{t}\right) \leq y\right) \\
\quad=\mathbb{P}\left(h^{(-1)}(x, 0)+A_{t}+M_{t} \leq h^{(-1)}\left(y, A_{t}\right)\right),
\end{gathered}
$$

and we easily conclude. The other two statements are also immediate.
Properties of the multivariate distributions may be analyzed along similar arguments.
Next, we study the expected value of the optimal wealth and portfolio processes.
Proposition 25. Let $X_{t}^{*, x}$ and $\pi_{t}^{*, x}$ be as in (23) and (24). Then, for $t>0$,

$$
\begin{equation*}
E\left(X_{t}^{*, x}\right)=h\left(h^{(-1)}(x, 0)+A_{t}, 0\right) . \tag{68}
\end{equation*}
$$

Moreover,

$$
\begin{equation*}
\frac{\partial}{\partial x} E\left(X_{t}^{*, x}\right)=\frac{r\left(E\left(X_{t}^{*, x}\right), 0\right)}{r(x, 0)} \tag{69}
\end{equation*}
$$

and

$$
\begin{equation*}
E\left(r\left(X_{t}^{*, x}, A_{t}\right)\right)=r\left(E\left(X_{t}^{*, x}\right), 0\right) \tag{70}
\end{equation*}
$$

Proof. We establish the above when the function $h$ is as in (39), with the other cases, exhibited in (49) and (58), following along similar arguments.

From (23) we have

$$
\begin{aligned}
& E\left(X_{t}^{*, x}\right)=E\left(h\left(h^{(-1)}(x, 0)+A_{t}+M_{t}, A_{t}\right)\right) \\
&= E\left(\int_{\mathbb{R}} \frac{e^{y\left(h^{(-1)}(x, 0)+A_{t}+M_{t}\right)-\frac{1}{2} y^{2} A_{t}}-1}{y} \nu(d y)\right) \\
&= \int_{\mathbb{R}} E\left(\frac{e^{y\left(h^{(-1)}(x, 0)+A_{t}+M_{t}\right)-\frac{1}{2} y^{2} A_{t}}-1}{y}\right) \nu(d y) \\
&=\int_{\mathbb{R}} \frac{e^{y\left(h^{(-1)}(x, 0)+A_{t}\right)}-1}{y} \nu(d y)=h\left(h^{(-1)}(x, 0)+A_{t}, 0\right) .
\end{aligned}
$$

Above we used that the two integrals can be interchanged. For this, it suffices to have

$$
\begin{equation*}
E\left(\int_{\mathbb{R}}\left|\frac{e^{y\left(h^{(-1)}(x, 0)+A_{t}+M_{t}\right)-\frac{1}{2} y^{2} A_{t}}-1}{y}\right| \nu(d y)\right)<+\infty . \tag{71}
\end{equation*}
$$

Indeed, inequality (79) yields

$$
\int_{\mathbb{R}}\left|\frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y}\right| \nu(d y) \leq \nu(\mathbb{R})\left(e^{x+\frac{t}{2}}+e^{-x+\frac{t}{2}}\right)+\int_{\mathbb{R}} e^{y x-\frac{1}{2} y^{2} t} \nu(d y) .
$$

Therefore,

$$
\begin{gathered}
E\left(\int_{\mathbb{R}}\left|\frac{e^{y\left(h^{(-1)}(x, 0)+A_{t}+M_{t}\right)-\frac{1}{2} y^{2} A_{t}}-1}{y}\right| \nu(d y)\right) \\
\leq \nu(\mathbb{R}) E\left(e^{h^{(-1)}(x, 0)+A_{t}+M_{t}+\frac{A_{t}}{2}}+e^{-\left(h^{(-1)}(x, 0)+A_{t}+M_{t}\right)+\frac{A_{t}}{2}}\right) \\
+E\left(\int_{\mathbb{R}} e^{y\left(h^{(-1)}(x, 0)+A_{t}+M_{t}\right)-\frac{1}{2} y^{2} A_{t}} \nu(d y)\right) .
\end{gathered}
$$

On the other hand,

$$
\begin{gathered}
E\left(e^{h^{(-1)}(x, 0)+A_{t}+M_{t}+\frac{A_{t}}{2}}+e^{-\left(h^{(-1)}(x, 0)+A_{t}+M_{t}\right)+\frac{A_{t}}{2}}\right) \\
\leq e^{h^{(-1)(x, 0)+\frac{3}{2} c^{2} t}} E\left(e^{M_{t}}\right)+e^{-h^{(-1)}(x, 0)},
\end{gathered}
$$

where we used (5) and (20). Similarly,

$$
\begin{gathered}
E\left(\int_{\mathbb{R}} e^{y\left(h^{(-1)}(x, 0)+A_{t}+M_{t}\right)-\frac{1}{2} y^{2} A_{t}} \nu(d y)\right) \\
=\int_{\mathbb{R}} E\left(e^{y\left(h^{(-1)}(x, 0)+A_{t}+M_{t}\right)-\frac{1}{2} y^{2} A_{t}}\right) \nu(d y) \\
=\int_{\mathbb{R}} e^{y\left(h^{(-1)}(x, 0)+A_{t}\right)} \nu(d y)<\int_{\mathbb{R}} e^{y\left(h^{(-1)}(x, 0)+c^{2} t\right)} \nu(d y)<+\infty
\end{gathered}
$$

and we easily obtain (71). Assertion (69) follows from (68) and (31).
To show (70), we recall (23), (31), and (81), yielding

$$
\begin{aligned}
E & \left(r\left(X_{t}^{*, x}, A_{t}\right)\right)=E\left(\int_{\mathbb{R}} e^{y\left(h^{(-1)}(x, 0)+A_{t}+M_{t}\right)-\frac{1}{2} y^{2} A_{t}} \nu(d y)\right) \\
& =\int_{\mathbb{R}} e^{y\left(h^{(-1)}(x, 0)+A_{t}\right)} \nu(d y)=h_{x}\left(h^{(-1)}(x, 0)+A_{t}, 0\right)
\end{aligned}
$$

and we easily conclude.
5.2. Inferring the investor's preferences. The investment performance criterion (27) combines the investor's preferences with the market input. As a consequence, the optimal portfolio and the associated wealth (see (24) and (23), respectively) contain implicit information about these preferences. In this section, we discuss how to learn about the individual's risk attitude by analyzing distributional characteristics of his optimal wealth. One can say, using the language of the derivatives industry, that our aim is to calibrate the investor's preferences, given the market dynamics and his desirable distributional outcomes for his wealth process.

This idea is relatively new. To the best of our knowledge, the authors of [15] were the first to propose a model and show how information about an investor's marginal utility of wealth can be inferred from her choice of a distribution. Other, more recent relevant references, are [2] and [16].

We discuss two examples in which we infer the investor's preferences using information about the behavior of her average future wealth. For simplicity, we concentrate only on the no-bankruptcy case, Range $(h)=(0,+\infty)$ (see section 4.2).

We remind the reader that the market price of risk is taken to be deterministic. As a result, $A_{t}, t \geq 0$, (cf. (20)) is also deterministic.

Proposition 26. Let the mapping $x \rightarrow E\left(X_{t}^{*, x}\right)$ be linear, for all $x>0$ and $t \geq 0$. Then, there exists a positive constant $\gamma>0$ such that the investor's forward performance process is given by

$$
\begin{equation*}
U_{t}(x)=\frac{\gamma^{\frac{\gamma-1}{\gamma}}}{\gamma-1} x^{\frac{\gamma-1}{\gamma}} e^{-\frac{1}{2}(\gamma-1) A_{t}} \tag{72}
\end{equation*}
$$

if $\gamma \neq 1$, and by

$$
\begin{equation*}
U_{t}(x)=\ln x-\frac{1}{2} A_{t} \tag{73}
\end{equation*}
$$

if $\gamma=1$. Moreover,

$$
\begin{equation*}
E\left(X_{t}^{*, x}\right)=x e^{\gamma A_{t}} \tag{74}
\end{equation*}
$$

Proof. Differentiating (69), we deduce

$$
\begin{aligned}
\frac{\partial^{2}}{\partial x^{2}} E\left(X_{t}^{*, x}\right) & =\frac{r_{x}\left(E\left(X_{t}^{*, x}\right), 0\right)}{r(x, 0)} \frac{\partial}{\partial x} E\left(X_{t}^{*, x}\right)-r\left(E\left(X_{t}^{*, x}\right), 0\right) \frac{r_{x}(x, 0)}{r^{2}(x, 0)} \\
& =\frac{r_{x}\left(E\left(X_{t}^{*, x}\right), 0\right)-r_{x}(x, 0)}{r^{2}(x, 0)} r\left(E\left(X_{t}^{*, x}\right), 0\right) .
\end{aligned}
$$

By assumption, $\frac{\partial^{2}}{\partial x^{2}} E\left(X_{t}^{*, x}\right)=0$. Moreover, $r\left(E\left(X_{t}^{*, x}\right), 0\right)>0$, as it follows from (70). Therefore, we must have

$$
r_{x}\left(E\left(X_{t}^{*, x}\right), 0\right)=r_{x}(x, 0)
$$

and, in turn,

$$
\frac{\partial}{\partial t} r_{x}\left(E\left(X_{t}^{*, x}\right), 0\right)=r_{x x}\left(E\left(X_{t}^{*, x}\right), 0\right) \frac{\partial}{\partial t} E\left(X_{t}^{*, x}\right)=0
$$

However, (68) implies $\frac{\partial}{\partial t} E\left(X_{t}^{*, x}\right) \neq 0$, and thus we deduce that

$$
r_{x x}\left(E\left(X_{t}^{*, x}\right), 0\right)=0 .
$$

Therefore, the function $r(x, 0)$ must be linear in $E\left(X_{t}^{*, x}\right)$ and, in turn, in $x$, per our assumption. Using (64), we obtain

$$
r(x, 0)=\gamma x,
$$

for some $\gamma>0$. From (31) and (49) we then deduce that, for all $x>0$,

$$
\int_{0^{+}}^{+\infty} e^{y h^{(-1)}(x, 0)} \nu(d y)=\gamma x
$$

and, in turn,

$$
\int_{0^{+}}^{+\infty} e^{y x} \nu(d y)=\gamma \int_{0^{+}}^{+\infty} \frac{e^{y x}}{y} \nu(d y)
$$

Therefore, we must have

$$
\nu(d y)=\delta_{\gamma},
$$

where $\delta_{\gamma}$ is a Dirac measure at $\gamma>0$. This yields

$$
h(x, t)=\frac{1}{\gamma} e^{\gamma x-\frac{1}{2} \gamma^{2} t},
$$

and assertions (72) and (73) follow from Examples 16, 17, 18, and 20.
Equality (74) follows from (68) and the form of $h$.
From the above analysis, we see that calibrating the investor's preferences consists of choosing a time horizon and the level of the mean of her optimal wealth, say $t_{0}$ and $m x$
$(m>1)$, respectively. Then, (74) implies that the corresponding $\gamma$ must satisfy $x e^{\gamma A_{t_{0}}}=m x$, or, equivalently,

$$
\gamma=\frac{\ln m}{A_{t_{0}}}
$$

Note that, under the linearity assumption, the investor can calibrate her expectations only for a single time horizon. The model interpolates for all other trading horizons, giving

$$
E\left(X_{t}^{* x}\right)=x m^{\frac{A_{t}}{A_{t}}}
$$

We easily deduce that the distribution of the optimal wealth $X_{t}^{*, x}$ is lognormal for all $(x, t)$.
The linearity of the mapping $x \rightarrow E\left(X_{t}^{*, x}\right)$ is a very strong assumption. Indeed, it allows for calibration of only a single parameter, namely, the slope, and, moreover, for a single time horizon. Therefore, if one intends to calibrate the investor's preferences to more refined information, then one needs to accept a more complicated dependence of $E\left(X_{t}^{*, x}\right)$ on $x$. We discuss this case next.

To this end, let us fix the level of initial wealth at $x=1$ and consider calibration to $E\left(X_{t}^{*, 1}\right)$ for $t>0$. The investor then chooses an increasing function $m(t)$ (with $\left.m(t)>1\right)$ to represent the latter; i.e., for $t>0$,

$$
\begin{equation*}
E\left(X_{t}^{*, 1}\right)=m(t) \tag{75}
\end{equation*}
$$

What does this choice reveal about her preferences? Moreover, can she choose an arbitrary increasing function $m(t)$ ? We give answers to these questions below.

In analogy to the previous proposition, we consider only the no-bankruptcy case, which corresponds to $h$ given in (49). Using arguments similar to those used in Proposition 7, we may assume, without loss of generality, that $\int_{0}^{+\infty} \frac{\nu(d y)}{y}=1$. We then have $h^{(-1)}(1,0)=0$, which, combined with (68), yields

$$
E\left(X_{t}^{*, 1}\right)=h\left(A_{t}, 0\right)=\int_{0^{+}}^{\infty} \frac{e^{y A_{t}}}{y} \nu(d y) .
$$

We easily see that the investor may only choose a function $m(t), t \geq 0$, which can be represented in the form

$$
\begin{equation*}
m(t)=\int_{0^{+}}^{\infty} \frac{e^{y A_{t}}}{y} \nu(d y) \tag{76}
\end{equation*}
$$

Therefore, the choice of a feasible $m(t)$ reduces to the choice of the appropriate measure $\nu$. Specifically, we must have

$$
\begin{equation*}
m\left(A_{t}^{(-1)}\right)=\int_{0^{+}}^{\infty} \frac{e^{y t}}{y} \nu(d y)=h(t, 0) \tag{77}
\end{equation*}
$$

where $A_{t}^{(-1)}$ stands for the inverse of the function $A_{t}, t \geq 0$. Note that the right-hand side above is the moment generating function of the probability measure $\mu(d y)=\frac{\nu(d y)}{y}$.

Assume now that the mean $m(t)$ (cf. (75)) was chosen so that (77) holds for some measure $\nu$. Then, for other values of $x>0, x \neq 1$, we have, using (68),

$$
\begin{aligned}
& E\left(X_{t}^{*, x}\right)=m\left(A^{(-1)}\left(h^{(-1)}(x, 0)+A(t)\right)\right) \\
& \quad=m\left(A^{(-1)}\left(A\left(m^{(-1)}(x)\right)+A(t)\right)\right)
\end{aligned}
$$

where, for notational convenience, $A_{t}$ (resp., $A_{t}^{(-1)}$ ) is denoted by $A(t)$ (resp., $A^{(-1)}(t)$ ).
In summary, the market related input $A_{t}$, coupled with the investor's targeted mean $m(t)$ (at initial wealth $x=1$ ), yields the investor's preferences by choosing the measure $\nu$, appearing in (76). Note that the function $A_{t}$ determines the distribution of the martingale $M_{t}$, and the measure $\nu$ defines the function $h$, which, in turn, determines the functions $u$ and $r$. Once these quantities are specified, we are able to construct the optimal portfolio process that generates the optimal wealth satisfying (75). The distribution of the optimal wealth process $X_{t}^{*, x}$, for $x>0, x \neq 1$, is, in turn, deduced from the specification of the targeted mean, $m(t)$, the market input $A_{t}$, and (67).

We conclude by mentioning that one may prefer to calibrate the distribution of the optimal wealth at a given time, say, $X_{t_{0}}^{*, 1}$, rather than the mean, $E\left(X_{t}^{*, 1}\right)$, for $t \geq 0$. It is easy to see what distributions are attainable. Indeed, Proposition 24 would give

$$
\mathbb{P}\left(X_{t_{0}}^{*, 1} \leq y\right)=N\left(\frac{h^{(-1)}\left(y, A_{t_{0}}\right)-A_{t_{0}}}{\sqrt{A_{t_{0}}}}\right)
$$

with

$$
h\left(y, A_{t_{0}}\right)=\int_{0^{+}}^{\infty} \frac{e^{z y-\frac{1}{2} z^{2} A_{t_{0}}}}{y} \nu(d z) .
$$

Further analysis of this and other calibration issues is left for future research.

## 6. Appendix.

Proof of Proposition 9. (i) Without loss of generality, we take $C=0$. We first establish that $h(x, t)$ is well defined. Indeed, for $(x, t) \in \mathbb{R} \times[0,+\infty)$, we have

$$
\begin{equation*}
\int_{\mathbb{R}}\left|\frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y}\right| \nu(d y)=\int_{|y|>1}\left|\frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y}\right| \nu(d y)+\int_{|y| \leq 1}\left|\frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y}\right| \nu(d y) . \tag{78}
\end{equation*}
$$

On the other hand, one can show that, for fixed $(x, t)$ and $|y| \leq 1$, the inequality

$$
\begin{equation*}
\left|\frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y}\right| \leq e^{|x|+\frac{t}{2}}-1 \tag{79}
\end{equation*}
$$

holds. ${ }^{7}$ Combining the above, we deduce

$$
\begin{equation*}
\int_{\mathbb{R}}\left|\frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y}\right| \nu(d y) \leq \int_{|y|>1} e^{y x} \nu(d y)+\nu(\mathbb{R}) e^{|x|+\frac{t}{2}}<+\infty . \tag{80}
\end{equation*}
$$

Differentiating under the integral yields

$$
\begin{equation*}
h_{x}(x, t)=\int_{\mathbb{R}} e^{y x-\frac{1}{2} y^{2} t} \nu(d y), \tag{81}
\end{equation*}
$$

and the claimed monotonicity of $h$ follows. Note that $h_{x}(x, t)$ is well defined because $0 \leq$ $h_{x}(x, t)<h_{x}(x, 0)<+\infty$, as $\nu \in \mathcal{B}^{+}(\mathbb{R})$. Further differentiation yields

$$
h_{x x}(x, t)=\int_{\mathbb{R}} y e^{y x-\frac{1}{2} y^{2} t} \nu(d y) \quad \text { and } \quad h_{t}(x, t)=-\frac{1}{2} \int_{\mathbb{R}} y e^{y x-\frac{1}{2} y^{2} t} \nu(d y) .
$$

The fact that $h$ solves (19) would follow, provided that the above integrals are well defined. For $x \neq 0$, we have

$$
\begin{gathered}
\left|\int_{\mathbb{R}} y e^{y x-\frac{1}{2} y^{2} t} \nu(d y)\right| \leq \frac{1}{|x|} \int_{\mathbb{R}}|y x| e^{y x-\frac{1}{2} y^{2} t} \nu(d y) \\
\leq \frac{1}{|x|} \int_{\mathbb{R}}\left(e^{|y x|}-1\right) e^{y x-\frac{1}{2} y^{2} t} \nu(d y) \\
\leq \frac{1}{|x|}\left(\int_{y x \leq 0}\left(e^{|y x|}-1\right) e^{y x-\frac{1}{2} y^{2} t} \nu(d y)+\int_{y x>0}\left(e^{|y x|}-1\right) e^{y x-\frac{1}{2} y^{2} t} \nu(d y)\right) \\
\leq \frac{1}{|x|}\left(\int_{y x \leq 0}\left(1-e^{y x}\right) e^{-\frac{1}{2} y^{2} t} \nu(d y)+\int_{y x>0} e^{2 y x-\frac{1}{2} y^{2} t} \nu(d y)\right) \\
\leq \frac{1}{|x|} \int_{y x \leq 0} \nu(d y)+\frac{1}{|x|} \int_{y x>0} e^{2 y x} \nu(d y)
\end{gathered}
$$

and the assertion follows, using that $\nu \in \mathcal{B}^{+}(\mathbb{R})$. The case $x=0$ follows trivially.
Next, we establish that if $\nu$ has the above-mentioned properties, then, for each $t \geq 0, h$ is of full range. Given that $h$ is continuous, we need to show that, for $t \geq 0$,

$$
\begin{equation*}
\lim _{x \rightarrow-\infty} h(x, t)=-\infty \quad \text { and } \quad \lim _{x \rightarrow+\infty} h(x, t)=+\infty . \tag{82}
\end{equation*}
$$

From (39) we have

$$
\begin{equation*}
h(x, t)=\int_{-\infty}^{0^{-}} \frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y} \nu(d y)+x \nu(\{0\})+\int_{0^{+}}^{+\infty} \frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y} \nu(d y) . \tag{83}
\end{equation*}
$$

${ }^{7}$ Indeed, we have $\frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y}=\sum_{n=1}^{\infty} \frac{y^{n-1}\left(x-\frac{1}{2} y t\right)^{n}}{n!}$ and, therefore,

$$
\left|\frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y}\right| \leq \sum_{n=1}^{\infty} \frac{|y|^{n-1}\left|x-\frac{1}{2} y t\right|^{n}}{n!} \leq \sum_{n=1}^{\infty} \frac{\left(|x|+\frac{1}{2}|y| t\right)^{n}}{n!} \leq e^{|x|+\frac{t}{2}}-1 .
$$

We first look at the case $\nu(\{0\})>0$. If both $\nu((-\infty, 0))=0$ and $\nu((0,+\infty))=0$, (82) follows directly. If $\nu((-\infty, 0))=0$ and $\nu((0,+\infty))>0$, the monotone convergence theorem yields

$$
\lim _{x \rightarrow \pm \infty}\left(x \nu(\{0\})+\int_{0^{+}}^{+\infty} \frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y} \nu(d y)\right)= \pm \infty .
$$

The case $\nu((-\infty, 0))>0$ and $\nu((0,+\infty))=0$ follows similarly.
Next, we assume that $\nu((-\infty, 0)) \times \nu((0,+\infty))>0$. Then, (83) yields

$$
h(x, t)=\int_{-\infty}^{0^{-}} \frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y} \nu(d y)+\int_{0^{+}}^{+\infty} \frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y} \nu(d y),
$$

and we easily deduce (82).
If $\nu \in \mathcal{B}_{0}^{+}(\mathbb{R})$ and it also satisfies $\nu((-\infty, 0))=0$ and $\int_{0^{+}}^{+\infty} \frac{1}{y} \nu(d y)=+\infty$, then the monotone convergence theorem yields

$$
\lim _{x \rightarrow+\infty} h(x, t)=\lim _{x \rightarrow+\infty} \int_{0^{+}}^{+\infty} \frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y} \nu(d y)=+\infty
$$

and

$$
\lim _{x \rightarrow-\infty} h(x, t)=-\int_{0^{+}}^{+\infty} \frac{1}{y} \nu(d y)=-\infty .
$$

The case $\nu \in \mathcal{B}_{-}^{+}(\mathbb{R})$ with $\int_{-\infty}^{0^{-}} \frac{1}{y} \nu(d y)=-\infty$ follows similarly, as well as the cases $\nu \in \mathcal{B}_{+}^{+}(\mathbb{R})$ with $\int_{0^{+}}^{+\infty} \frac{1}{y} \nu(d y)<+\infty$, and $\nu \in \mathcal{B}_{-}^{+}(\mathbb{R})$ with $\int_{-\infty}^{0^{-}} \frac{1}{y} \nu(d y)>-\infty$.
(ii) Let $h$ be a strictly increasing solution to (19). Then, its spatial derivative satisfies $h_{x}(x, t) \geq 0$ and solves (19). Thus, Widder's theorem implies the existence of $\nu \in \mathcal{B}^{+}(\mathbb{R})$ such that the representation

$$
\begin{equation*}
h_{x}(x, t)=\int_{\mathbb{R}} e^{y x-\frac{1}{2} y^{2} t} \nu(d y) \tag{84}
\end{equation*}
$$

holds. We then have $h_{x x}(x, t)=\int_{\mathbb{R}} y e^{y x-\frac{1}{2} y^{2} t} \nu(d y)$ (its finiteness follows easily), which, combined with (19), yields

$$
h_{t}(x, t)=-\frac{1}{2} \int_{\mathbb{R}} y e^{y x-\frac{1}{2} y^{2} t} \nu(d y) .
$$

If Range $(h)=(-\infty,+\infty), t \geq 0$, integrating yields

$$
\begin{equation*}
h(x, t)=\int_{0}^{t} h_{t}(x, s) d s+\int_{x_{0}}^{x} h_{x}(z, 0) d z+h\left(x_{0}, 0\right) \tag{85}
\end{equation*}
$$

for any $x_{0} \in \mathbb{R}$. Combining the above, we obtain

$$
\begin{equation*}
h(x, t)=-\frac{1}{2} \int_{0}^{t} \int_{\mathbb{R}} y e^{y x-\frac{1}{2} y^{2} s} \nu(d y) d s+\int_{x_{0}}^{x} \int_{\mathbb{R}} e^{y z} \nu(d y) d z+h\left(x_{0}, 0\right) . \tag{86}
\end{equation*}
$$

Note that, for $\nu \in \mathcal{B}^{+}(\mathbb{R})$,

$$
\int_{0}^{t} \int_{\mathbb{R}}\left|y e^{y x-\frac{1}{2} y^{2} s}\right| \nu(d y) d s \leq t \int_{\mathbb{R}}|y| e^{y x} \nu(d y)<\infty
$$

and thus Fubini's theorem yields

$$
\begin{gathered}
-\frac{1}{2} \int_{0}^{t} \int_{\mathbb{R}} y e^{y x-\frac{1}{2} y^{2} s} \nu(d y) d s=-\frac{1}{2} \int_{\mathbb{R}} \int_{0}^{t} y e^{y x-\frac{1}{2} y^{2} s} d s \nu(d y) \\
=\int_{\mathbb{R}} \frac{e^{y x-\frac{1}{2} y^{2} t}-e^{y x}}{y} \nu(d y)
\end{gathered}
$$

Moreover, Tonelli's theorem yields

$$
\int_{x_{0}}^{x} \int_{\mathbb{R}} e^{y z} \nu(d y) d z=\int_{\mathbb{R}} \int_{x_{0}}^{x} e^{y z} d z \nu(d y)=\int_{\mathbb{R}} \frac{e^{y x}-e^{y x_{0}}}{y} \nu(d y) .
$$

Observe that both integrals above are well defined, as was shown in the proof of part (i). Using (86) gives

$$
h(x, t)=\int_{\mathbb{R}} \frac{e^{y x-\frac{1}{2} y^{2} t}-e^{y x_{0}}}{y} \nu(d y)+h\left(x_{0}, 0\right)
$$

Without loss of generality we choose $x_{0}=0$, and we easily conclude.
Next, we establish that if $h$ is of full range, for each $t \geq 0$, it must be that $\nu(\{0\})>0$, or, otherwise, either $\nu(-\infty, 0)>0$ and $\nu(0,+\infty)>0$, or $\nu \in \mathcal{B}_{+}^{+}(\mathbb{R})$ with $\int_{0^{+}}^{+\infty} \frac{1}{y} \nu(d y)=+\infty$, or $\nu \in \mathcal{B}_{-}^{+}(\mathbb{R})$ with $\int_{-\infty}^{0^{-}} \frac{1}{y} \nu(d y)=-\infty$. Note that (82) must hold, for each $t \geq 0$, as $h$ is continuous.

Let us assume that $\nu \in \mathcal{B}_{+}^{+}(\mathbb{R})$ and $\int_{0^{+}}^{+\infty} \frac{1}{y} \nu(d y)<+\infty$. Then, (83) would give

$$
\lim _{x \rightarrow-\infty} h(x, t)=\lim _{x \rightarrow-\infty} \int_{0^{+}}^{+\infty} \frac{e^{y x-\frac{1}{2} y^{2} t}-1}{y} \nu(d y)=-\int_{0^{+}}^{+\infty} \frac{1}{y} \nu(d y)>-\infty,
$$

contradicting (82). All other cases follow along similar arguments, and their proof is thus omitted.

The following auxiliary result will be used in what follows. Because we will examine the various cases of the range of $h$ separately, we state the result without making specific reference to the domain of the spatial inverse, $h^{(-1)}$.

Lemma 27. A strictly increasing function, say h, satisfies (19) if and only if its spatial inverse, $h^{(-1)}$, satisfies

$$
\begin{equation*}
\frac{\partial}{\partial t} h^{(-1)}(x, t)+\frac{1}{2} \frac{\frac{\partial^{2}}{\partial x^{2}} h^{(-1)}(x, t)}{\left(\frac{\partial}{\partial x} h^{(-1)}(x, t)\right)^{2}}=0 . \tag{87}
\end{equation*}
$$

We continue with the proofs of Propositions 10, 14, and 15.
Proof of Proposition 10. (i) First, we establish that the integrals in (42) are well defined. Using (84) and Tonelli's theorem, we have

$$
\begin{gathered}
\int_{0}^{t} e^{-h^{(-1)}(x, s)+\frac{s}{2}} h_{x}\left(h^{(-1)}(x, s), s\right) d s=\int_{0}^{t} \int_{\mathbb{R}} e^{(y-1) h^{(-1)}(x, s)+\frac{s}{2}-\frac{1}{2} y^{2} s} \nu(d y) d s \\
\leq e^{\frac{t}{2}} \int_{\mathbb{R}} \int_{0}^{t} e^{(y-1) h^{(-1)}(x, s)} d s \nu(d y) \\
=e^{\frac{t}{2}} \int_{y \geq 1} \int_{0}^{t} e^{(y-1) h^{(-1)}(x, s)} d s \nu(d y)+e^{\frac{t}{2}} \int_{y<1} \int_{0}^{t} e^{(y-1) h^{(-1)}(x, s)} d s \nu(d y) \\
\leq t e^{\frac{t}{2}} \int_{y \geq 1} e^{(y-1) \max _{0 \leq s \leq t} h^{(-1)}(x, s)} \nu(d y)+t e^{\frac{t}{2}} \int_{y<1} e^{(y-1) \min _{0 \leq s \leq t} h^{(-1)}(x, s)} \nu(d y) .
\end{gathered}
$$

Using Tonelli's theorem once more, we have that the second integral in (42) satisfies

$$
\begin{gathered}
\int_{0}^{x} e^{-h^{(-1)}(z, 0)} d z=\int_{h^{(-1)}(0,0)}^{h^{(-1)}(x, 0)} e^{-z^{\prime}} h_{x}\left(z^{\prime}, 0\right) d z^{\prime} \\
=\int_{h^{(-1)}(0,0)}^{h^{(-1)}(x, 0)} \int_{\mathbb{R}} e^{(y-1) z^{\prime}} \nu(d y) d z^{\prime}=\int_{\mathbb{R}} \frac{e^{(y-1) h^{(-1)}(x, 0)}-e^{(y-1) h^{(-1)}(0,0)}}{y-1} \nu(d y)
\end{gathered}
$$

and its finiteness follows from arguments similar to those used in the proof of Proposition 9.
Differentiating (42) and using that $h$ solves (19) yields

$$
\begin{aligned}
u_{x}(x, t) & =\frac{1}{2} \int_{0}^{t}\left(1-\frac{h_{x x}\left(h^{(-1)}(x, s), s\right)}{h_{x}\left(h^{(-1)}(x, s), s\right)}\right) e^{-h^{(-1)}(x, s)+\frac{s}{2}} d s+e^{-h^{(-1)}(x, 0)} \\
= & \frac{1}{2} \int_{0}^{t}\left(1+2 \frac{h_{t}\left(h^{(-1)}(x, s), s\right)}{h_{x}\left(h^{(-1)}(x, s), s\right)}\right) e^{-h^{(-1)}(x, s)+\frac{s}{2}} d s+e^{-h^{(-1)}(x, 0)} \\
= & \int_{0}^{t}\left(\frac{1}{2}-\frac{\partial}{\partial s} h^{(-1)}(x, s)\right) e^{-h^{(-1)}(x, s)+\frac{s}{2}} d s+e^{-h^{(-1)}(x, 0)}
\end{aligned}
$$

and, therefore,

$$
\begin{equation*}
u_{x}(x, t)=e^{-h^{(-1)}(x, t)+\frac{t}{2}} \tag{88}
\end{equation*}
$$

Further differentiation yields

$$
\frac{u_{x}^{2}(x, t)}{u_{x x}(x, t)}=-\frac{e^{-h^{(-1)}(x, t)+\frac{t}{2}}}{\frac{\partial}{\partial x} h^{(-1)}(x, t)}
$$

On the other hand, (42) implies

$$
u_{t}(x, t)=-\frac{1}{2} e^{-h^{(-1)}(x, t)+\frac{t}{2}} h_{x}\left(h^{(-1)}(x, t), t\right)=-\frac{1}{2} \frac{e^{-h^{(-1)}(x, t)+\frac{t}{2}}}{\frac{\partial}{\partial x} h^{(-1)}(x, t)}
$$

Combining the above two equalities, we deduce that $u$ satisfies (14).

To establish (41) and (43), we first observe that the assumption of full range yields, for each $t \geq 0, \lim _{x \rightarrow \pm \infty} h^{(-1)}(x, t)= \pm \infty$. Both assertions then follow from (88).
(ii) Let $u$ be an increasing and strictly concave function, defined for $(x, t) \in \mathbb{R} \times[0,+\infty)$ and satisfying (14), (41), and (43). Using that $u_{x}$ is invertible, with $\left(u_{x}\right)^{(-1)}:(0,+\infty) \times[0,+\infty) \rightarrow$ $\mathbb{R}$, we define, for $(x, t) \in \mathbb{R} \times[0,+\infty)$, the function $h$ by

$$
\begin{equation*}
h(x, t)=\left(u_{x}\right)^{(-1)}\left(e^{-x+\frac{t}{2}}, t\right) . \tag{89}
\end{equation*}
$$

Note that $h$ is invertible in the space variable since

$$
h_{x}(x, t)=-\frac{e^{-x+\frac{t}{2}}}{u_{x x}(h(x, t), t)}>0 .
$$

Differentiating (14) yields

$$
\begin{equation*}
u_{x t}=u_{x}-\frac{1}{2} \frac{u_{x}^{2} u_{x x x}}{u_{x x}^{2}} . \tag{90}
\end{equation*}
$$

In turn,

$$
\begin{gathered}
u_{x t}(x, t)=\left(-\frac{\partial}{\partial t} h^{(-1)}(x, t)+\frac{1}{2}\right) u_{x}(x, t), \\
u_{x x}(x, t)=-\left(\frac{\partial}{\partial x} h^{(-1)}(x, t)\right) u_{x}(x, t)
\end{gathered}
$$

and

$$
u_{x x x}(x, t)=\left(-\frac{\partial^{2}}{\partial x^{2}} h^{(-1)}(x, t)+\left(\frac{\partial}{\partial x} h^{(-1)}(x, t)\right)^{2}\right) u_{x}(x, t) .
$$

Combining the above, we obtain that $h^{(-1)}$ satisfies

$$
\frac{\partial}{\partial t} h^{(-1)}(x, t)+\frac{1}{2} \frac{\frac{\partial^{2}}{\partial x^{2}} h^{(-1)}(x, t)}{\left(\frac{\partial}{\partial x} h^{(-1)}(x, t)\right)^{2}}=0
$$

and using Lemma 27, we deduce that its spatial inverse, $h$, solves (19). On the other hand, (89) and (41) yield that $h(0,0)=0$. Finally, (89) and the Inada conditions yield

$$
\lim _{x \rightarrow-\infty} h(x, t)=-\infty \quad \text { and } \quad \lim _{x \rightarrow+\infty} h(x, t)=+\infty .
$$

Therefore, $h$ solves (19) and is strictly increasing and of full range, for each $t \geq 0$. Using Proposition 9(ii), we obtain (39) for some $\nu \in \mathcal{B}^{+}(\mathbb{R})$ with the appropriate properties.

It remains to show that $u$ is given by (42). Using (14), (89), and the form of $u_{x x}(x, t)$, we obtain

$$
u_{t}(x, t)=-\frac{1}{2} e^{-h^{(-1)}(x, t)+\frac{t}{2}} h_{x}\left(h^{(-1)}(x, t), t\right)
$$

Integrating and using (41) yields

$$
u(x, t)=\int_{0}^{t} u_{t}(x, s) d s+\int_{0}^{x} u_{x}(z, 0) d z
$$

and (42) follows from direct integration. Note that the above two integrals are well defined, as follows from arguments used in the proof of part (i). We easily conclude.

The next result will be used in the proofs that follow.
Lemma 28. Let $h$ be such that Range $(h)=(0,+\infty)$ (resp., Range $(h)=(-\infty, 0)$ ). Then, for each $x, h^{(-1)}(x, t)$ is increasing (resp., decreasing) in $t$.

Proof. We look only at the case Range $(h)=(0,+\infty)$. Using (50) and differentiating the identity $h\left(h^{(-1)}(x, t), t\right)=x$ with respect to time yields the claimed monotonicity of $h^{(-1)}(x, t)$.

Proof of Proposition 14. (i) We first establish that $u$ in (51) is well defined for $x>0$, $t \geq 0$. From (49) and the assumptions on the measure $\nu$, we easily deduce that

$$
\begin{equation*}
h_{x}(x, t)=\int_{1^{+}}^{+\infty} e^{y x-\frac{1}{2} y^{2} t} \nu(d y) . \tag{92}
\end{equation*}
$$

Therefore,

$$
\begin{aligned}
& \int_{0}^{t} e^{-h^{(-1)}(x, s)+\frac{s}{2}} h_{x}\left(h^{(-1)}(x, s), s\right) d s \\
= & \int_{0}^{t} \int_{1^{+}}^{+\infty} e^{(y-1) h^{(-1)}(x, s)+\frac{s}{2}-\frac{1}{2} y^{2} s} \nu(d y) d s \\
\leq & t e^{-h^{(-1)}(x, 0)+\frac{t}{2}} \int_{1^{+}}^{+\infty} e^{y h h^{(-1)}(x, t)} \nu(d y),
\end{aligned}
$$

where we used Lemma 27. The finiteness of the integral then follows from the assumptions on the measure $\nu$.

The finiteness of the second integral in (51) also follows. Indeed, first observe that (49) yields

$$
\begin{equation*}
\lim _{x \rightarrow-\infty} h(x, t)=0 \quad \text { and } \quad \lim _{x \rightarrow+\infty} h(x, t)=+\infty \tag{93}
\end{equation*}
$$

Using the above, (92), and Tonelli's theorem, we obtain

$$
\begin{gathered}
\int_{0}^{x} e^{-h^{(-1)}(z, 0)} d z=\int_{-\infty}^{h^{(-1)}(x, 0)} e^{-z^{\prime}} h_{x}\left(z^{\prime}, 0\right) d z^{\prime} \\
=\int_{-\infty}^{h^{(-1)}(x, 0)} \int_{1^{+}}^{+\infty} e^{(y-1) z^{\prime}} \nu(d y) d z^{\prime}=\int_{1^{+}}^{+\infty} \int_{-\infty}^{h^{(-1)}(x, 0)} e^{(y-1) z^{\prime}} d z^{\prime} \nu(d y) \\
=\int_{1^{+}}^{+\infty} \frac{1}{y-1}\left(e^{(y-1) h^{(-1)}(x, 0)}-\lim _{z^{\prime} \rightarrow-\infty} e^{(y-1) z^{\prime}}\right) \nu(d y)
\end{gathered}
$$

Using that $\lim _{z^{\prime} \rightarrow-\infty} e^{(y-1) z^{\prime}}=0$ for $y>1$ and that $\int_{1^{+}}^{+\infty} \frac{\nu(d y)}{y-1}<+\infty$, we deduce that

$$
\begin{equation*}
\int_{0}^{x} e^{-h^{(-1)}(z, 0)} d z=\int_{1^{+}}^{+\infty} \frac{1}{y-1} e^{(y-1) h^{(-1)}(x, 0)} \nu(d y) \tag{94}
\end{equation*}
$$

For $\varepsilon>0$, we then have

$$
\begin{gathered}
\int_{0}^{x} e^{-h^{(-1)}(z, 0)} d z=\int_{1^{+}}^{1+\varepsilon} \frac{1}{y-1} e^{(y-1) h^{(-1)}(x, 0)} \nu(d y)+\int_{1+\varepsilon}^{+\infty} \frac{1}{y-1} e^{(y-1) h^{(-1)}(x, 0)} \nu(d y) \\
\quad \leq \max \left(1, e^{\varepsilon h^{(-1)}(x, 0)}\right) \int_{1^{+}}^{1+\varepsilon} \frac{\nu(d y)}{y-1}+\frac{e^{-h^{(-1)}(x, 0)}}{\varepsilon} \int_{1+\varepsilon}^{+\infty} e^{y h^{(-1)}(x, 0)} \nu(d y) .
\end{gathered}
$$

Using the assumptions on the measure $\nu$, we easily conclude.
The fact that $u$ solves (14) and has the claimed monotonicity and concavity properties follows from arguments similar to those used in the proof of Proposition 10(i).

Next, we establish (52). We first show that

$$
\begin{equation*}
\lim _{x \rightarrow 0} \int_{0}^{t} e^{-h^{(-1)}(x, s)+\frac{s}{2}} h_{x}\left(h^{(-1)}(x, s), s\right) d s=0 \tag{95}
\end{equation*}
$$

Indeed, note that the above integrand is monotone in $x$. This follows easily from its representation,

$$
e^{-h^{(-1)}(x, t)+\frac{t}{2}} h_{x}\left(h^{(-1)}(x, t), t\right)=\int_{1^{+}}^{+\infty} e^{(y-1) h^{(-1)}(x, t)-\frac{1}{2} y^{2} t+\frac{t}{2}} \nu(d y),
$$

combined with the monotonicity of $h^{(-1)}$. Using the monotone convergence theorem and (93), we obtain (95).

On the other hand, (94) yields

$$
\lim _{x \rightarrow 0} \int_{0}^{x} e^{-h^{(-1)}(z, 0)} d z=\lim _{x \rightarrow 0} \int_{1^{+}}^{+\infty} \frac{1}{y-1} e^{(y-1) h^{(-1)}(x, 0)} \nu(d y) .
$$

Using the monotone convergence theorem, (93) (for $t=0$ ), and that $\int_{1^{+}}^{+\infty} \frac{\nu(d y)}{y-1}<+\infty$, we conclude.
(ii) Using arguments similar to those in the proof of Proposition 10(ii), we deduce that the function $h$ given, for $(x, t) \in R \times[0,+\infty)$, by

$$
\begin{equation*}
h(x, t)=\left(u_{x}\right)^{(-1)}\left(e^{-x+\frac{t}{2}}, t\right) \tag{96}
\end{equation*}
$$

is well defined and solves (19). Moreover, the assumptions on $u$ imply that $h(x, t) \geq 0$ and $h_{x}(x, t) \geq 0$. Therefore, from Proposition 10 , we have that there exists $\nu \in \mathcal{B}^{+}(\mathbb{R})$ satisfying (48) and such that representation (49) holds. The Inada conditions (53) then yield that the normalization constant must be chosen as $C=\int_{0^{+}}^{+\infty} \frac{1}{y} \nu(d y)$.

Using (52) and working along arguments similar to those used in the proof of Proposition 10, we deduce the representation (51).

It remains to establish that $\nu$ satisfies $\nu((0,1])=0$ and $\int_{1^{+}}^{+\infty} \frac{\nu(d y)}{y-1}<+\infty$. We argue by contradiction. To this end, we first note that, because of (52), we have, for $x>0$,

$$
u(x, t)=\int_{0}^{x} u_{x}(z, t) d z=\int_{-\infty}^{h^{(-1)}(x, t)} \int_{0}^{+\infty} e^{(y-1) z^{\prime}+\frac{t}{2}\left(1-y^{2}\right)} \nu(d y) d z^{\prime}
$$

where we used (89) and (93). We then observe that $\nu$ cannot include a Dirac measure at $y=1$, as this would yield

$$
u(x, t) \geq \int_{-\infty}^{h^{(-1)}(x, t)} d z^{\prime}=+\infty
$$

contradicting the finiteness of $u(x, t)$. Therefore, we must have

$$
\begin{gathered}
u(x, t)=\int_{-\infty}^{h^{(-1)}(x, t)} \int_{0^{+}}^{1^{-}} e^{(y-1) z^{\prime}+\frac{t}{2}\left(1-y^{2}\right)} \nu(d y) d z^{\prime} \\
+\int_{-\infty}^{h^{(-1)}(x, t)} \int_{1^{+}}^{+\infty} e^{(y-1) z^{\prime}+\frac{t}{2}\left(1-y^{2}\right)} \nu(d y) d z^{\prime}
\end{gathered}
$$

and, in turn, for $x>0$,

$$
\begin{equation*}
\int_{-\infty}^{h^{(-1)}(x, t)} \int_{0^{+}}^{1^{-}} e^{(y-1) z^{\prime}+\frac{t}{2}\left(1-y^{2}\right)} \nu(d y) d z^{\prime}<+\infty \tag{97}
\end{equation*}
$$

and

$$
\begin{equation*}
\int_{-\infty}^{h^{(-1)}(x, t)} \int_{1^{+}}^{+\infty} e^{(y-1) z^{\prime}+\frac{t}{2}\left(1-y^{2}\right)} \nu(d y) d z^{\prime}<+\infty \tag{98}
\end{equation*}
$$

However, using Tonelli's theorem, we deduce

$$
\begin{gathered}
\int_{-\infty}^{h^{(-1)}(x, t)} \int_{0^{+}}^{1^{-}} e^{(y-1) z^{\prime}+\frac{t}{2}\left(1-y^{2}\right)} \nu(d y) d z^{\prime} \\
=\int_{0^{+}}^{1^{-}} \int_{-\infty}^{h^{(-1)}(x, t)} e^{(y-1) z^{\prime}+\frac{t}{2}\left(1-y^{2}\right)} d z^{\prime} \nu(d y) \\
=\int_{0^{+}}^{1^{-}} \frac{1}{y-1}\left(e^{(y-1) h^{(-1)}(x, t)}-\lim _{z^{\prime} \rightarrow-\infty} e^{(y-1) z^{\prime}}\right) e^{\frac{t}{2}\left(1-y^{2}\right)} \nu(d y),
\end{gathered}
$$

and we easily get a contradiction to (97) if $\nu((0,1)) \neq 0$.
Similarly, for all $x>0$, we must have

$$
\begin{gathered}
\int_{-\infty}^{h^{(-1)}(x, t)} \int_{1^{+}}^{+\infty} e^{(y-1) z^{\prime}+\frac{t}{2}\left(1-y^{2}\right)} \nu(d y) d z^{\prime} \\
=\int_{1^{+}}^{+\infty} \frac{1}{y-1}\left(e^{(y-1) h^{(-1)}(x, t)}-\lim _{z \rightarrow-\infty} e^{(y-1) z}\right) e^{\frac{t}{2}\left(1-y^{2}\right)} \nu(d y) \\
=\int_{1^{+}}^{+\infty} \frac{1}{y-1} e^{(y-1) h^{(-1)}(x, t)+\frac{t}{2}\left(1-y^{2}\right)} \nu(d y)<+\infty
\end{gathered}
$$

By assumption, for each $t \geq 0$, Range $\left(h^{(-1)}\right)=(-\infty,+\infty)$. Therefore, for $t=0$, there exists $x_{0}$ such that $h^{(-1)}\left(x_{0}, 0\right)=0$. We easily conclude.

Proof of Proposition 15. We prove only some of the main points, for the rest of the proof follows along arguments similar to those in the previous proof. To this end, we first show that the function given in (54) is well defined. Indeed,

$$
\begin{aligned}
& \int_{0}^{t} e^{-h^{(-1)}(x, s)+\frac{s}{2}} h_{x}\left(h^{(-1)}(x, s), s\right) d s \\
= & \int_{0}^{t} \int_{0+}^{+\infty} e^{(y-1) h^{(-1)}(x, s)+\frac{s}{2}-\frac{1}{2} y^{2} s} \nu(d y) d s \\
\leq & t e^{-h^{(-1)}(x, 0)+\frac{t}{2}} \int_{0^{+}}^{+\infty} e^{y h^{(-1)}(x, t)} \nu(d y)
\end{aligned}
$$

where we used Lemma 27. The finiteness of the integral follows from the assumptions on the measure $\nu$.

Moreover, for $x>x_{0}$ (the case $x<x_{0}$ follows similarly),

$$
\begin{gathered}
\int_{x_{0}}^{x} e^{-h^{(-1)}(z, 0)} d z=\int_{h^{(-1)}\left(x_{0}, 0\right)}^{h^{(-1)}(x, 0)} e^{-z^{\prime}} h_{x}\left(z^{\prime}, 0\right) d z^{\prime} \\
=\int_{h^{(-1)}\left(x_{0}, 0\right)}^{h^{(-1)}(x, 0)} \int_{0+}^{+\infty} e^{(y-1) z^{\prime}} \nu(d y) d z^{\prime}=\int_{0+}^{+\infty} \int_{h^{(-1)}\left(x_{0}, 0\right)}^{h^{(-1)}(x, 0)} e^{(y-1) z^{\prime}} d z^{\prime} \nu(d y) \\
=\int_{0+}^{+\infty} \frac{1}{y-1}\left(e^{(y-1) h^{(-1)}(x, 0)}-e^{(y-1) h^{(-1)}\left(x_{0}, 0\right)}\right) \nu(d y)
\end{gathered}
$$

We have

$$
\begin{aligned}
& \int_{0+}^{+\infty}\left|\frac{1}{y-1}\left(e^{(y-1) h^{(-1)}(x, 0)}-e^{(y-1) h^{(-1)}\left(x_{0}, 0\right)}\right)\right| \nu(d y) \\
&=\int_{0+}^{2}\left|\frac{1}{y-1}\left(e^{(y-1) h^{(-1)}(x, 0)}-e^{(y-1) h^{(-1)}\left(x_{0}, 0\right)}\right)\right| \nu(d y) \\
&+ \int_{2}^{+\infty}\left|\frac{1}{y-1}\left(e^{(y-1) h^{(-1)}(x, 0)}-e^{(y-1) h^{(-1)}\left(x_{0}, 0\right)}\right)\right| \nu(d y) \\
& \leq \int_{0+}^{2}\left|\frac{1}{y-1}\left(e^{(y-1) h^{(-1)}(x, 0)}-e^{(y-1) h^{(-1)}\left(x_{0}, 0\right)}\right)\right| \nu(d y) \\
&+e^{-h^{(-1)}(x, 0)} \int_{2}^{+\infty} e^{y h^{(-1)}(x, 0)} \nu(d y)+e^{-h^{(-1)}\left(x_{0}, 0\right)} \int_{2}^{+\infty} e^{y h^{(-1)}\left(x_{0}, 0\right)} \nu(d y)
\end{aligned}
$$

On the other hand,

$$
\begin{gathered}
\frac{1}{y-1}\left(e^{(y-1) h(-1)(x, 0)}-e^{(y-1) h^{(-1)}\left(x_{0}, 0\right)}\right) \\
=\frac{1}{y-1}\left(\sum_{n=1}^{+\infty}\left(\frac{(y-1)^{n}\left(h^{(-1)}(x, 0)\right)^{n}}{n!}-\frac{(y-1)^{n}\left(h^{(-1)}\left(x_{0}, 0\right)\right)^{n}}{n!}\right)\right) \\
=\sum_{n=1}^{+\infty}\left(\frac{(y-1)^{n-1}\left(h^{(-1)}(x, 0)\right)^{n}}{n!}-\frac{(y-1)^{n-1}\left(h^{(-1)}\left(x_{0}, 0\right)\right)^{n}}{n!}\right) .
\end{gathered}
$$

For $0 \leq y \leq 2$,

$$
\begin{gathered}
\left|\frac{1}{y-1}\left(e^{(y-1) h^{(-1)}(x, 0)}-e^{(y-1) h^{(-1)}\left(x_{0}, 0\right)}\right)\right| \\
\leq \sum_{n=1}^{+\infty}\left(\frac{|y-1|^{n-1}\left|h^{(-1)}(x, 0)\right|^{n}}{n!}+\frac{|y-1|^{n-1}\left|h^{(-1)}\left(x_{0}, 0\right)\right|^{n}}{n!}\right) \\
\leq e^{\left|h^{(-1)}(x, 0)\right|}+e^{\left|h^{(-1)}(x, 0)\right|}-2
\end{gathered}
$$

Combining the above yields

$$
\begin{gathered}
\int_{0+}^{+\infty}\left|\frac{1}{y-1}\left(e^{(y-1) h^{(-1)}(x, 0)}-e^{(y-1) h^{(-1)}\left(x_{0}, 0\right)}\right)\right| \nu(d y) \\
\leq\left(e^{\left|h^{(-1)}(x, 0)\right|}+e^{\left|h^{(-1)}(x, 0)\right|}-2\right) \nu([0,2]) \\
+e^{-h^{(-1)}(x, 0)} \int_{2}^{+\infty} e^{y h^{(-1)}(x, 0)} \nu(d y)+e^{-h^{(-1)}\left(x_{0}, 0\right)} \int_{2}^{+\infty} e^{y h^{(-1)}\left(x_{0}, 0\right)} \nu(d y),
\end{gathered}
$$

and we easily conclude.
Next we show that, under the assumptions on the measure $\nu$, (55) holds.
First we assume that $\nu((0,1])>0$. Observe that for $x$ sufficiently small,

$$
\begin{gathered}
\int_{x_{0}}^{x} e^{-h^{(-1)}(z, 0)} d z=-\int_{0^{+}}^{+\infty} \int_{h^{(-1)}(x, 0)}^{h^{(-1)}\left(x_{0}, 0\right)} e^{(y-1) z^{\prime}} d z^{\prime} \nu(d y) \\
=-\left(\int_{0+}^{1} \int_{h^{(-1)}(x, 0)}^{h^{(-1)}\left(x_{0}, 0\right)} e^{(y-1) z^{\prime}} d z^{\prime} \nu(d y)+\int_{1^{+}}^{+\infty} \int_{h^{(-1)}(x, 0)}^{h^{(-1)}\left(x_{0}, 0\right)} e^{(y-1) z^{\prime}} d z^{\prime} \nu(d y)\right) \\
\leq-\int_{0+}^{1} \frac{1}{y-1}\left(e^{(y-1) h^{(-1)}\left(x_{0}, 0\right)}-e^{(y-1) h^{(-1)}(x, 0)}\right) \nu(d y) .
\end{gathered}
$$

Passing to the limit and using the monotone convergence theorem yields

$$
\lim _{x \rightarrow 0} \int_{0+}^{1} \frac{1}{y-1}\left(e^{(y-1) h^{(-1)}\left(x_{0}, 0\right)}-e^{(y-1) h^{(-1)}(x, 0)}\right) \nu(d y)=+\infty .
$$

Next we look at the case $\nu((0,1])=0$ and $\int_{1+}^{+\infty} \frac{\nu(d y)}{y-1}=+\infty$. We then have

$$
\begin{aligned}
& \int_{x_{0}}^{x} e^{-h^{(-1)}(z, 0)} d z=-\int_{1^{+}}^{+\infty} \int_{h^{(-1)}(x, 0)}^{h^{(-1)}\left(x_{0}, 0\right)} e^{(y-1) z^{\prime}} d z^{\prime} \nu(d y) \\
= & -\int_{1^{+}}^{+\infty} \frac{1}{y-1}\left(e^{(y-1) h^{(-1)}\left(x_{0}, 0\right)}-e^{(y-1) h^{(-1)}(x, 0)}\right) \nu(d y) .
\end{aligned}
$$

Using the monotone convergence theorem and that

$$
\lim _{x \rightarrow 0}\left(e^{(y-1) h^{(-1)}\left(x_{0}, 0\right)}-e^{(y-1) h^{(-1)}(x, 0)}\right)=e^{(y-1) h^{(-1)}\left(x_{0}, 0\right)}
$$

yields

$$
\lim _{x \rightarrow 0} \int_{x_{0}}^{x} e^{-h^{(-1)}(z, 0)} d z=-\int_{1^{+}}^{+\infty} \frac{1}{y-1}\left(e^{(y-1) h^{(-1)}\left(x_{0}, 0\right)}\right) \nu(d y) .
$$

The elementary inequality $e^{(y-1) h^{(-1)}\left(x_{0}, 0\right)} \geq 1+(y-1) h^{(-1)}\left(x_{0}, 0\right)$ in turn implies

$$
\begin{aligned}
& -\int_{1^{+}}^{+\infty} \frac{1}{y-1}\left(e^{(y-1) h^{(-1)}\left(x_{0}, 0\right)}\right) \nu(d y) \\
\leq & -\int_{1^{+}}^{+\infty} \frac{1}{y-1}\left(1+(y-1) h^{(-1)}\left(x_{0}, 0\right)\right) \nu(d y) \\
=- & \int_{1^{+}}^{+\infty} \frac{1}{y-1} \nu(d y)-h^{(-1)}\left(x_{0}, 0\right) \nu((1,+\infty)),
\end{aligned}
$$

and we easily conclude.
Acknowledgments. We would like to thank P. Carr and M. Yor for their fruitful and insightful suggestions. We also thank two anonymous referees for their valuable comments.

## REFERENCES

[1] F. Berrier, L. C. Rogers, and M. Tehranchi, A Characterization of Forward Utility Functions, preprint, 2009.
[2] D. G. Goldstein, E. J. Johnson, and W. F. Sharpe, A New Way to Measure Consumer Risk Preferences, preprint, 2008.
[3] Y. Hu and X. Y. Zhou, Constrained stochastic $L Q$ control with random coefficients, and application to portfolio selection, SIAM J. Control Optim., 44 (2005), pp. 444-466.
[4] H. Jin and X. Y. Zhou, Behavioral portfolio selection in continuous time, Math. Finance, 18 (2008), pp. 385-426.
[5] I. Karatzas and G. Zitkovic, Optimal consumption from investment and random endowment in incomplete semimartingale markets, Ann. Appl. Probab., 31 (2003), pp. 1821-1858.
[6] X. Li and X. Y. Zhou, Continuous-time mean-variance efficiency: The $80 \%$ rule, Ann. Appl. Probab., 16 (2006), pp. 1751-1763.
[7] R. S. Lipster and A. N. Shiryaev, Statistics of Random Processes, II: Applications, 2nd ed., Springer, Berlin, 2000.
[8] R. Merton, Lifetime portfolio selection under uncertainty: The continuous-time case, Rev. Econ. Statist., 51 (1969), pp. 247-257.
[9] M. Musiela and T. Zariphopoulou, Optimal asset allocation under forward exponential criteria, in Markov Processes and Related Topics: A Festschrift for Thomas G. Kurtz, IMS Collections 4, Institute of Mathematical Statistics, Bethesda, MD, 2008, pp. 285-300.
[10] M. Musiela and T. Zariphopoulou, Portfolio choice under dynamic investment performance criteria, Quant. Finance, 9 (2009), pp. 161-170.
[11] M. Musiela and T. Zariphopoulou, Stochastic partial differential equations in portfolio choice, in Contemporary Quantitative Finance: Essays in Honour of Eckhard Platen, C. Chiarella and A. Novikov, eds., Springer, New York, 2010, to appear.
[12] M. Musiela and T. Zariphopoulou, Investments and Forward Utilities, preprint, 2006.
[13] R. Penrose, A generalized inverse for matrices, Proc. Cambridge Philos. Soc., 51 (1955), pp. 406-413.
[14] W. Schachermayer, A super-martingale property of the optimal portfolio process, Finance Stoch., 7 (2003), pp. 433-456.
[15] W. F. Sharpe, D. G. Goldstein, and P. W. Blythe, The Distribution Builder: A Tool for Inferring Investor's Preferences, preprint, 2000.
[16] W. F. Sharpe, Expected utility asset allocation, Financial Analysts J., 63 (2007), pp. 18-30.
[17] D. V. Widder, The Heat Equation, Academic Press, New York, 1975.
[18] T. Zariphopoulou and T. Zhou, Investment performance measurement under asymptotically linear local risk tolerance, in Mathematical Modeling and Numerical Methods in Finance, Handbook of Numerical Analysis, A. Bensoussan and Q. Zhang, eds., North-Holland, Amsterdam, 2009, pp. 227254.
[19] G. Zitkovic, A dual characterization of self-generation and log-affine forward performances, Ann. Appl. Probab., 19 (2009), pp. 2176-2210.

# Merton Problem with Taxes: Characterization, Computation, and Approximation* 

Imen Ben Tahar ${ }^{\dagger}$, H. Mete Soner ${ }^{\ddagger}$, and Nizar Touzi ${ }^{\S}$


#### Abstract

We formulate a computationally tractable extension of the classical Merton optimal consumptioninvestment problem to include the capital gains taxes. This is the continuous-time version of the model introduced by Dammon, Spatt, and Zhang [Rev. Financ. Stud., 14 (2001), pp. 583-616]. In this model the tax basis is computed as the average cost of the stocks in the investor's portfolio. This average rule introduces only one additional state variable, namely the tax basis. Since the other tax rules such as the first in first out rule require the knowledge of all past transactions, the average model is computationally much easier. We emphasize the linear taxation rule, which allows for tax credits when capital gains losses are experienced. In this context wash sales are optimal, and we prove it rigorously. Our main contributions are a first order explicit approximation of the value function of the problem and a unique characterization by means of the corresponding dynamic programming equation. The latter characterization builds on technical results isolated in the accompanying paper [I. Ben Tahar, H. M. Soner, and N. Touzi, SIAM J. Control Optim., 46 (2007), pp. 1779-1801]. We also suggest a numerical computation technique based on a combination of finite differences and the Howard iteration algorithm. Finally, we provide some numerical results on the welfare consequences of taxes and the quality of the first order approximation.


Key words. optimal consumption and investment in continuous time, transaction costs, capital gains taxes, finite differences

AMS subject classifications. 91B28, 35B37, 35K20
DOI. 10.1137/080742178

1. Introduction. Since the seminal papers of Merton [26, 27], there has been extensive literature on the problem of optimal consumption and investment decision in financial markets subject to imperfections. We refer the reader to Cox and Huang [10] and Karatzas, Lehoczky, and Shreve [21] for the case of incomplete markets, Cvitanić and Karatzas [11] for the case of portfolio constraints, and Constantinides and Magill [9], Davis and Norman [13], Shreve and Soner [28], Barles and Soner [3], and Duffie and Sun [15] for the case of transaction costs.

However, the problem of taxes on capital gains received limited attention, although taxes represent a much higher percentage than transaction costs in real securities markets. Compared to ordinary income, capital gains are taxed only when the investor sells the security, allowing for a deferral option. One may think that the taxes on capital gains have an appreciable impact on an individual's consumption and investment decisions. Indeed, under taxation

[^33]of capital gains, the portfolio rebalancing implies additional charges, therefore altering the available wealth for future consumption. This possibly induces a depreciation of consumption opportunities compared to a tax-free market. On the other hand, since taxes are paid only when embedded capital gains are actually realized, the investor may choose to defer the realization of capital gains and liquidate his/her position in case of a capital loss, particularly when the tax code allows for tax credits.

The first relevant work in the previous literature is due to Constantinides [8], who shows that the investment and consumption decisions are separable and that the optimal strategy consists in realizing losses and deferring gains. These results rely heavily on the possibility of short selling the risky asset. Since capital gains realizations are observed in real securities markets, the subsequent literature considers the problem under the no-short-sales constraint.

In a multiperiod context, many challenging difficulties appear because of the path dependency of the problem. The taxation code specifies the basis to which the price of a security has to be compared in order to evaluate the capital gains (or losses). The tax basis is defined as either (i) the specific purchase price of the asset to be sold, (ii) the purchase price of a freely chosen share held in the portfolio (of course the number of chosen shares must be more than the ones to be sold), or (iii) the weighted average of past purchase prices. In some countries, investors can choose any one of the above definitions of the tax basis.

A deterministic model with the above definition (i) of the tax basis, together with the first in first out priority rule for the stock to be sold, is introduced by Jouini, Koehl, and Touzi $[20,19]$. An existence result is proved, and the first order conditions of optimality are derived under some conditions. However, the numerical complexity due to the path dependency of the problem is not solved in the context of this model.

A financial model with the above definition (ii) of the taxation rule was considered by Dybvig and Koo [16] in the context of a four-period binomial model. Some numerical progress was achieved later by DeMiguel and Uppal [14], who were able to consider more periods in the binomial model and/or more stocks. This numerical progress is limited, as these authors were not able to go beyond 10 periods in the single-asset framework.

The taxation rule (iii), where the tax basis is the weighted average of past purchase prices, was first considered by Dammon, Spatt, and Zhang [12] in the context of a binomial model with short-sales constraints and the linear taxation rule. The average tax basis is actually used in Canada. Dammon, Spatt, and Zhang [12] considered the problem of maximizing the expected discounted utility from future consumption and provided a numerical analysis of this model based on the dynamic programming principle. The important technical feature of this model is that the path dependency of the problem is seriously reduced, as the dynamics of the tax basis is Markov. This implies a significant advantage of this model in comparison to [16]. This advantage was further justified by DeMiguel and Uppal [14], who provided numerical evidence that the certainty equivalent loss from using the average tax basis (iii) instead of the exact tax basis (ii) is typically less than $1 \%$ for a large choice of parameter values. The analysis of [12] was further extended to the multiasset framework by Gallmeyer, Kaniel, and Tompaidis [18].

In this paper, we formulate a continuous-time version of the Dammon-Spatt-Zhang utility maximization problem under capital gains taxes. Our model is similar to that of Leland [24], who instead considered the problem of minimizing the tracking error to some benchmark index.

The financial market consists of a tax-free riskless asset and a risky one. The holdings in the risky asset are subject to the no-short-sales constraint, and the total wealth is restricted by the no-bankruptcy condition. The risky asset is subject to taxes on capital gains. As in [12], the tax basis is defined as the weighted average of past purchase prices, and the taxation rule is linear, thus allowing for tax credits. However, we differ from [12] by considering an infinite horizon problem, as our main goal is to provide analytical tools for this class of problems. In particular, our model does not allow for tax forgiveness at death. Clearly, one should keep this difference in mind when interpreting results for investors with a short horizon.

The investor preferences are described by a power utility with a constant relative risk aversion factor. This assumption is needed only to reduce the computational complexity of the problem. However, with this reduction explicit descriptions of the value function and the optimal strategy are still not available.

This model enables us to rigorously prove several interesting properties observed in practice. Although these results are sometimes intuitively clear, their proofs require careful analysis and the use of the tractability of the model. The first of these results is the optimality of the wash sales. Namely, in Proposition 3.5 we prove that it is always optimal to realize capital losses whenever the tax basis exceeds the spot price. This property is observed in practice and is stated and embedded directly in the definition of the tax basis in [12]. We also prove the continuity of the value function (and even Lipschitz continuity, up to a change of variables). We recall that, in the tax-free models of [26, 9, 13, 28], this property follows from the obvious concavity of the value function. Under capital gains taxes, the concavity argument fails, and the numerical results of section 6 suggest that the value function is indeed not concave!

The first main result of this paper is to provide an explicit approximation of the value function which follows from an upper and a lower bound proved in section 4. In view of the absence of closed form solutions, such an approximation is useful for understanding the model better. Although this explicit approximation holds for small interest rate and tax parameters, our numerical experiments indicate that this approximation is satisfactory with realistic values of interest rate and tax parameters, as it leads to a relative error within $10 \%$. These findings are reported in section 6. This first order approximation allows one to draw the following observations:

- The lower bound is derived as the limit of the value implied by a sequence of strategies which mimics the Merton optimal strategy in a Merton-type fictitious frictionless financial market with tax-deflated drift and volatility coefficients. The risk premium of this fictitious financial market is smaller than that of the original market. So, even if the optimal strategy in our problem is not available in explicit form, our first order expansion is accompanied by an explicit strategy which achieves "the first order maximal utility value."
- In a situation of a capital loss, our first order approximation is increasing in the tax rate. For small interest rate and tax parameters, the advantage taken from an initial tax credit is never compensated by the increase of tax over the lifetime horizon. This is in agreement with Cadenillas and Pliska [7], who found that "sometimes investors are better off with a positive tax rate."
- Finally, the investment component of this approximation sequence exhibits a smaller exposition to the risky asset. This is in line with the risk premium puzzle highlighted
by Mehra and Prescott [25]. However, one should note that this model is only a partial equilibrium model and that the level of the equity premium is determined by general equilibrium considerations.
Our analysis of the optimal consumption-investment problem relies on a numerical approach based on dynamic programming and partial differential equations. Therefore, the second main result of this paper is a characterization of the value function as the limit (uniformly on compact subsets) of an approximating sequence defined by a slight perturbation of the "natural" dynamic programming equation of our problem. The financial interpretation of our perturbation is, on the one hand, to introduce a small transaction cost parameter and, on the other hand, to modify simultaneously the taxation rule when the tax basis approaches the critical point zero. Our analysis relies on the technical results in our accompanying paper [5], which shows that the perturbed dynamic programming equation has a unique continuous viscosity solution within the class of polynomially growing functions.

Finally, based on our dynamic programming characterization, we suggest a numerical approximation method combining finite differences with the Howard iterations. Unfortunately, we have no theoretical convergence result for our algorithm. Indeed, establishing such convergence results for Hamilton-Jacobi-Bellman equations corresponding to singular control problems is an open question in numerical analysis, and the existing results, based on the monotone scheme method of Barles and Souganidis [4], are restricted to the bounded control context; see Bonnans and Zidani [6], Krylov [22, 23], Barles and Jakobsen [2], and Fahim, Touzi, and Warin [17]. This difficulty was already observed in the related literature on transaction costs; see Akian, Menaldi, and Sulem [1] and Tourin and Zariphopoulou [31]. Following the latter papers, we therefore concentrate our effort on realizing the empirical convergence of the algorithm. The numerical scheme is implemented to obtain the qualitative behavior of the solution and to understand the welfare consequences of the taxation. In particular, the numerical approximation of the optimal strategy displays a bang-bang behavior, as expected in our singular control problem. As in the transaction cost context of [9, 13, 28], the state space is partitioned into three regions: the no-transaction region NT , the buy region B , and the sell region S ; but in contrast with the transaction cost framework these regions are not cones.

Notation. For a domain $\mathbf{D}$ in $\mathbb{R}^{n}$, we denote by $\operatorname{USC}(\mathbf{D})$ (resp., $\operatorname{LSC}(\mathbf{D})$ ) the collection of all upper semicontinuous (resp., lower semicontinuous) functions from $\mathbf{D}$ to $\mathbb{R}$. The set of continuous functions from $\mathbf{D}$ to $\mathbb{R}$ is denoted by $\mathrm{C}^{0}(\mathbf{D}):=\operatorname{USC}(\mathbf{D}) \cap \operatorname{LSC}(\mathbf{D})$. For a parameter $\delta>0$, we say that a function $f: \mathbf{D} \longrightarrow \mathbb{R}$ has $\delta$-polynomial growth if

$$
\sup _{x \in \mathbf{D}} \frac{|f(x)|}{1+|x|^{\delta}}<\infty .
$$

We finally denote $\operatorname{USC}_{\delta}(\mathbf{D}):=\{f \in \operatorname{USC}(\mathbf{D}): f$ has $\delta$-polynomial growth $\}$. The sets $\operatorname{LSC}_{\delta}(\mathbf{D})$ and $\mathrm{C}_{\delta}^{0}(\mathbf{D})$ are defined similarly.

## 2. Consumption-investment models with capital gains taxes.

2.1. The financial assets. Throughout this paper, we consider a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, endowed with a standard scalar Brownian motion $W=\left\{W_{t}, 0 \leq t\right\}$, and we denote by $\mathbb{F}$ the $\mathbb{P}$-completion of the natural filtration of the Brownian motion. We consider
a financial market consisting of one bank account with constant interest rate $r>0$ and one risky asset whose price process evolves according to the Black-Scholes model:

$$
\begin{equation*}
d P_{t}=P_{t}\left[(r+\theta \sigma) d t+\sigma d W_{t}\right] \tag{2.1}
\end{equation*}
$$

where $\theta>0$ is a constant risk premium, and $\sigma>0$ is a constant volatility parameter. The positivity restriction on the risk premium coefficient ensures that positive investment in the risky asset is interesting.
2.2. Taxation rule on capital gains. The sales of the stock are subject to taxes on capital gains. The amount of tax to be paid for each sale of risky asset, at time $t$, is computed by comparison of the current price $P_{t}$ to an index $B_{t}$ defined as the weighted average price of the shares purchased by the investor up to time $t$. When $P_{t} \geq B_{t}$, i.e., the current price of the risky asset is greater than the weighted average price, the investor would realize a capital gain by selling the risky asset. Similarly, when $P_{t} \leq B_{t}$, the sale of the risky asset corresponds to the realization of a capital loss.

In order to better explain the definition of the tax basis $B$, we provide the following example derived from the official Canadian tax code.

Table 1 reports transactions performed by an individual on shares of STU Ltd and how the tax basis of the individual changes over time.

Table 1
Extracted from Capital Gains 2007; http://www.cra.gc.ca.

| Transaction | Price $P$ <br> (dollars) | Number of shares <br> (unitless) | Portfolio composition <br> (unitless) | Tax basis $B$ <br> (dollars) |
| :---: | :---: | :---: | :---: | :---: |
| Purchase at $t_{1}$ | 15.00 | 100 | $100: \$ 15.00 /$ share | $\mathbf{1 5 . 0 0}$ |
| Purchase at $t_{2}$ | 20.00 | 150 | $100: \$ 15.00 /$ share |  |
|  |  |  | $150: \$ 20.00 /$ share | $\mathbf{1 8 . 0 0}$ |
| Sale at $t_{3}$ | - | 200 | $20: \$ 15.00 /$ share |  |
|  |  | $=\frac{4}{5}(100+150)$ | $30: \$ 20.00 /$ share | $\mathbf{1 8 . 0 0}$ |
| Purchase at $t_{4}$ | 21.00 | 350 | $20: \$ 15.00 /$ share |  |
|  |  |  | $30: \$ 20.00 /$ share |  |

Just after a sale transaction, the tax basis is not changed. However, sales do alter the tax basis starting from the date of the next purchase. Notice, however, that the tax basis is affected only by the number of shares sold and not by the sale price.

The sale of a unit share of stock at some time $t$ is subject to the payment of an amount of tax computed according to the tax basis of the portfolio at time $t$. In this paper, we consider a linear taxation rule, i.e., this amount of tax is given by

$$
\begin{equation*}
\ell\left(P_{t}-B_{t}\right):=\alpha\left(P_{t}-B_{t}\right) \tag{2.2}
\end{equation*}
$$

where $\alpha \in[0,1)$ is a constant tax rate coefficient. Our interest is of course in the case $\alpha>0$. When the tax basis is smaller than the spot price, the investor realizes a capital gain. Then, by selling one unit of risky asset at the spot price $P_{t}$, the amount of tax to be paid is $\alpha\left(P_{t}-B_{t}\right)$.

When the tax basis is larger than the spot price, the investor receives the tax credit $\alpha\left(B_{t}-P_{t}\right)$ for each unit of asset sold at time $t$.

Remark 2.1. In practice, the realized capital losses are deduced from the total amount of taxes that the investor has to pay, and the annual deductible capital losses amount may be limited by the tax code. In our model, we follow Dammon, Spatt, and Zhang [12] by adopting the simplifying assumption that capital losses are credited immediately without any limit.

Remark 2.2. Our definition of the tax basis $B$ is slightly different from that of Dammon, Spatt, and Zhang [12], who set the tax basis to be equal to the spot price whenever the average purchase price exceeds the current price. This does not affect the results, as Proposition 3.5 shows that wash sales are optimal.
2.3. Consumption-investment strategies. We denote by $X_{t}$ the (cash) position on the bank, $Y_{t}$ the amount invested in the risky assets, and

$$
\begin{equation*}
K_{t}:=B_{t} \frac{Y_{t}}{P_{t}}, \quad t \geq 0 \tag{2.3}
\end{equation*}
$$

the position on the risky asset account evaluated at the basis price. The trading in risky assets is subject the no-short-sales constraint

$$
\begin{equation*}
Y_{t} \geq 0 \quad \mathbb{P} \text {-a.s. for all } t \geq 0 \tag{2.4}
\end{equation*}
$$

and the position of the investor is required to satisfy the solvency condition

$$
\begin{equation*}
Z_{t}:=X_{t}+Y_{t}-\ell\left(P_{t}-B_{t}\right) \frac{Y_{t}}{P_{t}}=X_{t}+(1-\alpha) Y_{t}+\alpha K_{t} \geq 0 \quad \mathbb{P} \text {-a.s.; } \tag{2.5}
\end{equation*}
$$

i.e., the after-tax liquidation value of the portfolio is nonnegative at any point in time.

Trading on the financial market is described by means of the transfers between the two investment opportunities defined by two $\mathbb{F}$-adapted, right-continuous, and nondecreasing processes $L=\left\{L_{t}, t \geq 0\right\}$ and $M=\left\{M_{t}, t \geq 0\right\}$ with $L_{0^{-}}=M_{0^{-}}=0$. The amount transferred from the bank to the nonrisky asset account at time $t$ is given by $d L_{t}$ and corresponds to a purchase of risky asset. The amount transferred from the risky asset account to the bank at time $t$, corresponding to a sale of risky asset, is given by $Y_{t-} d M_{t}$ and is expressed in terms of proportions of the total holdings in risky asset as in the example of Table 1.

To force the short-sales constraint (2.4) to hold, we restrict the jumps of $M$ by

$$
\begin{equation*}
\Delta M_{t} \leq 1 \quad \text { for } \quad t \geq 0 \quad \mathbb{P} \text {-a.s. } \tag{2.6}
\end{equation*}
$$

With these notations, the evolution of the wealth on the risky asset account is given by

$$
\begin{equation*}
d Y_{t}=Y_{t} \frac{d P_{t}}{P_{t}}+d L_{t}-Y_{t-} d M_{t} \tag{2.7}
\end{equation*}
$$

and, by definition of the tax basis $B$ and (2.3), we have

$$
\begin{equation*}
d K_{t}=d L_{t}-K_{t-} d M_{t} . \tag{2.8}
\end{equation*}
$$

Observe that the contribution of the sales in the dynamics of $K_{t}$ is evaluated at the basis price. For any given initial condition $\left(Y_{0-}, K_{0-}\right),(2.7)-(2.8)$ define a unique $\mathbb{F}$-adapted process $(Y, K)$ taking values in $\mathbb{R}_{+}^{2}$, the nonnegative orthant of $\mathbb{R}^{2}$.

In addition to the trading activities, the investor consumes in continuous time at the rate $C=\left\{C_{t}, t \geq 0\right\}$. Here, $C$ is an $\mathbb{F}$-progressively measurable process with

$$
\begin{equation*}
C \geq 0 \text { and } \int_{0}^{T} C_{t} d t<\infty \quad \mathbb{P} \text {-a.s. for all } T>0 \tag{2.9}
\end{equation*}
$$

Then, the bank component of the wealth process satisfies the dynamics

$$
\begin{align*}
d X_{t} & =\left(r X_{t}-C_{t}\right) d t-d L_{t}+Y_{t-} d M_{t}-\ell\left(P_{t}-B_{t-}\right) \frac{Y_{t-} d M_{t}}{P_{t}} \\
& =\left(r X_{t}-C_{t}\right) d t-d L_{t}+\left[(1-\alpha) Y_{t-}+\alpha K_{t-}\right] d M_{t} . \tag{2.10}
\end{align*}
$$

Since the processes $Y$ and $K$ have been defined previously, the above dynamics uniquely defines an $\mathbb{F}$-adapted process $X$ valued in $\mathbb{R}$ for any given initial condition $X_{0-}$.

Remark 2.3. In Dammon, Spatt, and Zhang [12], the nonrisky asset is also subject to a constant proportional taxation rule. This is obviously caught by our model by interpreting $r$ as the after-tax instantaneous interest rate.

For later use, we report the dynamics of the corresponding liquidation value process defined in (2.5), which follows from (2.7), (2.8), (2.9), and (2.10):

$$
\begin{equation*}
d Z_{t}=\left(r Z_{t}-C_{t}\right) d t+(1-\alpha) Y_{t}\left(\frac{d P_{t}}{P_{t}}-r d t\right)-r \alpha K_{t} d t \tag{2.11}
\end{equation*}
$$

Definition 2.1. (i) A consumption investment strategy is a triple of $\mathbb{F}$-adapted processes $\nu=(C, L, M)$, where $C$ satisfies (2.9), $L$ and $M$ are nondecreasing and right continuous, $L_{0-}=M_{0-}=0$, and the jumps of $M$ satisfy (2.6).
(ii) Given an initial condition $s=(x, y, k) \in \mathbb{R} \times \mathbb{R}_{+} \times \mathbb{R}_{+}$and a consumption-investment strategy $\nu$, we denote by $S^{s, \nu}=\left(X^{s, \nu}, Y^{s, \nu}, K^{s, \nu}\right)$ the unique strong solution of (2.7), (2.8), (2.9), and (2.10) with initial condition $S_{0-}^{s, \nu}=s$.
(iii) Given an initial condition $s=(x, y, k) \in \mathbb{R} \times \mathbb{R}_{+} \times \mathbb{R}_{+}$, a consumption-investment strategy $\nu$ is said to be s-admissible if the corresponding state process $S^{s, \nu}$ satisfies the nobankruptcy constraint (2.5). We shall denote by $\mathcal{A}(s)$ the collection of all s-admissible con-sumption-investment strategies.

The admissibility conditions imply that the process $S^{s, \nu}$ is valued in the closure $\overline{\mathcal{S}}$ of

$$
\begin{equation*}
\mathcal{S}=\left\{(x, y, k) \in \mathbb{R}^{3}: x+(1-\alpha) y+\alpha k>0, y>0, k>0\right\} . \tag{2.12}
\end{equation*}
$$

We partition the boundary of $\mathcal{S}$ into $\partial \mathcal{S}=\partial^{z} \mathcal{S} \cup \partial^{y} \mathcal{S} \cup \partial^{k} \mathcal{S}$ with

$$
\partial^{y} \mathcal{S}:=\{(x, y, k) \in \overline{\mathcal{S}}: y=0\}, \quad \partial^{k} \mathcal{S}:=\{(x, y, k) \in \overline{\mathcal{S}}: k=0\}
$$

and

$$
\partial^{z} \mathcal{S}=\{(x, y, k) \in \overline{\mathcal{S}}: z:=x+(1-\alpha) y+\alpha k=0\} .
$$

2.4. The consumption-investment problem. The investor preferences are characterized by a power utility function with constant relative risk aversion coefficient $1-p \in(0,1)$ :

$$
U(c):=\frac{c^{p}}{p}, \quad c \geq 0, \quad \text { for some } \quad p \in(0,1)
$$

The restriction on the relative risk aversion coefficient to $(0,1)$ allows us to simplify the analysis of this paper, as the boundary condition on $\partial^{z} \mathcal{S}$ is easily obtained; see Proposition 3.2. However, several of our results hold for a general parameter $p<0$, and we will indicate whenever it is the case.

For every initial data $s \in \overline{\mathcal{S}}$ and any admissible strategy $\nu \in \mathcal{A}(s)$, we introduce the consumption-investment criterion

$$
\begin{equation*}
J_{T}(s, \nu):=\mathbb{E}\left[\int_{0}^{T} e^{-\beta t} U\left(C_{t}\right) d t+e^{-\beta T} U\left(Z_{T}^{s, \nu}\right) \mathbf{1}_{\{T<\infty\}}\right], \quad T \in \mathbb{R}_{+} \cup\{+\infty\} . \tag{2.13}
\end{equation*}
$$

The consumption-investment problem is defined by

$$
\begin{equation*}
V(s):=\sup _{\nu \in \mathcal{A}(s)} J_{\infty}(s, \nu), \quad s \in \overline{\mathcal{S}} . \tag{2.14}
\end{equation*}
$$

We shall assume that the parameters $r, \theta, \sigma, p$, and $\beta$ satisfy the condition

$$
\begin{equation*}
\bar{c}(r, \theta):=\frac{\beta-p r}{1-p}-\frac{p \theta^{2}}{2(1-p)^{2}}>0 \tag{2.15}
\end{equation*}
$$

which has been pointed out as a sufficient condition for the finiteness of the value function in the context of a financial market without taxes in [26] and [28].
2.5. Review of the tax-free model. In this section, we briefly review the solution of the consumption-investment problem when the financial market is free from taxes on capital gains. The properties of the corresponding value function are going to be useful to state relevant bounds for the maximal utility achieved in a financial market with taxes.

In the classical formulation of the tax-free consumption-investment problem [26], the investment control variable is described by means of a unique process $\pi$ which represents the proportion of wealth invested in risky assets at each time, and the consumption process $C$ is expressed as a proportion $c$ of the total wealth:

$$
\begin{equation*}
d \bar{Z}_{t}=\bar{Z}_{t}\left[\left(r-c_{t}\right) d t+\pi_{t} \sigma\left(\theta d t+d W_{t}\right)\right] . \tag{2.16}
\end{equation*}
$$

In this context, a consumption-investment admissible strategy is a pair of adapted processes $(c, \pi)$ such that $c$ is nonnegative and

$$
\int_{0}^{T} c_{t} d t+\int_{0}^{T}\left|\pi_{t}\right|^{2} d t<\infty \quad \mathbb{P} \text {-a.s. for all } T>0
$$

We shall denote by $\overline{\mathcal{A}}$ the collection of all such consumption-investment strategies. For every initial condition $z \geq 0$ and strategy $(c, \pi) \in \overline{\mathcal{A}}$, there is a unique strong solution to (2.16) that we denote by $\bar{Z}^{z, c, \pi}$. The frictionless consumption-investment problem is

$$
\begin{equation*}
\bar{V}(z):=\sup _{(c, \pi) \in \overline{\mathcal{A}}} \mathbb{E}\left[\int_{0}^{\infty} e^{-\beta t} U\left(c_{t} \bar{Z}_{t}^{z, c, \pi}\right) d t\right] . \tag{2.17}
\end{equation*}
$$

Theorem 2.1 (see [26]). Let condition (2.15) hold. Then, for all $z \geq 0$,

$$
\bar{V}(z)=\bar{c}(r, \theta)^{p-1} \frac{z^{p}}{p}
$$

and the constant consumption-investment strategy $\bar{c}(r, \theta), \bar{\pi}(\sigma, \theta):=\frac{\theta}{(1-p) \sigma}$ is optimal.
Remark 2.4. The reduction of the model of section 2 to the frictionless case, i.e., $\alpha=0$, does not alter the value function. This can be proved by approximating any investment strategy by a sequence of bounded variation strategies. However, the investment strategies in our formulation are constrained to have bounded variation. This is needed because sales and purchases have different impacts on the bank component of the wealth process (2.10). Since the Merton optimal strategy is well known to be unique and has unbounded variation, it follows that existence fails to hold in our formulation.
3. First properties of the value function. We first show that the optimal consumptioninvestment problem under taxes reduces to the Merton problem when the interest rate is zero. Let $\bar{c}(r, \theta)$ be as in Theorem 2.1.

Proposition 3.1. For $r=0, V(s)=\frac{\bar{c}(0, \theta)^{p-1}}{p}(x+(1-\alpha) y+\alpha k)^{p}, s=(x, y, k) \in \overline{\mathcal{S}}$.
Proof. Notice that the optimal consumption-investment problem (2.14) can be expressed equivalently in terms of the state processes $(Z, Y, K)$ instead of $(X, Y, K)$, and observe that the interest rate parameter $r$ is involved only in the dynamics of $Z$. When $r=0$, the dynamics of the process $Z$ in (2.11)

$$
d Z_{t}=-C_{t} d t+(1-\alpha) Y_{t} \frac{d P_{t}}{P_{t}}
$$

is independent of the tax basis $K_{t}$. Since the dynamics of $Y$ in (2.7) is independent of $K$, it follows that the value function does not depend on the variable $k$. Next, for $Z_{t}>0$, defining $c_{t}:=C_{t} / Z_{t}$ and $\pi_{t}:=(1-\alpha) Y_{t} / Z_{t}$, we see that the solution $Z$ of the above equation is the same as the solution $\bar{Z}^{z, c, \pi}$ of (2.16). In view of the state constraint $Z \geq 0$, our state dynamics in the context $r=0$ is then equivalent to (2.16). Then, the only difference between the control problem $V$ and the corresponding Merton problem $\bar{V}$ is the class of admissible trading strategies, which does not induce any difference on the value function; see Remark 2.4.

The argument in the above proof clearly does not involve the specific nature of the utility function. Therefore an analogous result holds for any utility function.

Since the tax basis is not inflated by the interest rate $r$, for nonzero values of $r$ the tax basis plays a role in the solution. This explains the importance of the point $r=0$.

We next discuss the value function on the boundary of the state space $\mathcal{S}$. Observe that there is no a priori information on the boundary components $\partial^{y} \mathcal{S}$ and $\partial^{k} \mathcal{S}$. This is one source of difficulty in the numerical part of this paper, as this state constraint problem needs special treatment; see [5].

Proposition 3.2. For every $s \in \partial^{z} \mathcal{S}$, we have $V(s)=0$.
Proof. Let $s$ be in $\partial^{z} \mathcal{S}$, and let $\nu$ be in $\mathcal{A}(s)$. By the definition of the set admissible controls, the process $Z^{s, \nu}$ is nonnegative. By Itô's lemma, together with the nonnegativity of $C$ and $K$ and the nondecrease of $L$, this provides

$$
0 \leq e^{-r t} Z_{t}^{s, \nu} \leq(1-\alpha) \int_{0}^{t} e^{-r u} Y_{u}^{s, \nu} \sigma\left[\theta d u+d W_{u}\right]
$$

Let $\mathbb{Q}$ be the probability measure equivalent to $\mathbb{P}$ under which the process $\left\{\theta u+W_{u}, u \geq 0\right\}$ is a Brownian motion. The process appearing on the right-hand side of the last inequality is a $\mathbb{Q}$-supermartingale because it is a nonnegative $\mathbb{Q}$-local martingale. By taking expected values under $\mathbb{Q}$, it then follows from the last inequalities that $Z^{s, \nu}=Y^{s, \nu}=K^{s, \nu}=C=L \equiv 0$. We have then proved that, for $s \in \partial^{z} \mathcal{S}$, any admissible strategy $\nu=(C, L, M) \in \mathcal{A}(s)$ is such that $C=L \equiv 0$, implying that $V(s)=0$.

Proposition 3.3. The value function $V$ is nondecreasing with respect to each of the variables $x, y$, and $k$. Moreover, for $(x, y, k) \in \overline{\mathcal{S}}$ with $z:=x+(1-\alpha) y+\alpha k>0$,

$$
\begin{equation*}
V(x, y, k)=z^{p} \mathcal{V}\left(\frac{y}{z}, \frac{k}{z}\right), \quad \text { where } \quad \mathcal{V}(\xi, \zeta):=V(1-(1-\alpha) \xi-\alpha \zeta, \xi, \zeta) . \tag{3.1}
\end{equation*}
$$

Proof. 1. The monotonicity property with respect to $x, y$, and $k$ follows immediately from the dynamics of the problem and the bound $\Delta M \leq 1$.
2. Let $\nu=(C, L, M)$ be an arbitrary strategy in $\mathcal{A}(s)$, and define the strategy $\nu^{\prime}:=$ $(\delta C, \delta L, M)$. We easily verify that $S^{\delta s, \nu^{\prime}}=\delta S^{s, \nu} \in \mathcal{\mathcal { S }}$, which implies that $\nu^{\prime}$ is in $\mathcal{A}(\delta s)$, and therefore

$$
V(\delta s) \geq \mathbb{E}\left[\int_{0}^{\infty} e^{-\beta u} U\left(\delta C_{u}\right) d u\right]=\delta^{p} J_{\infty}(\delta, \nu),
$$

where the last equality follows from the homogeneity property of the utility function $U$. By the arbitrariness of $\nu$ in $\mathcal{A}(s)$, this shows that $V(\delta s) \geq \delta^{p} V(s)$.
3. By writing $V(s)=V\left(\delta^{-1} \delta s\right) \geq \delta^{-p} V(\delta s)$, it follows from the previous step that we in fact have $V(\delta s)=\delta^{p} V(s)$, and the required result follows immediately from this homotheticity property.

In the absence of taxes on capital gains, i.e., $\alpha=0$, it is easy to deduce from the concavity of $U$ that the value function $V$ is concave and therefore continuous. The numerical results exhibited in section 6 reveal that this property is no longer valid when $\alpha>0$. The proof of the following continuity result is obtained by first reducing the continuity problem to the ray $\left\{(x, 0,0), x \in \mathbb{R}_{+}\right\}$. This is achieved by means of a comparison result in the sense of viscosity solutions. Then the continuity on the latter ray is proved by a direct argument.

Proposition 3.4. The function $\mathcal{V}$ of (3.1) is Lipschitz continuous on $\overline{\mathcal{S}}$.
Proof. See Appendix A.
We now show that it is always worth realizing capital losses whenever the tax basis exceeds the spot price of the risky asset. In other words, given $s=(x, y, k) \in \overline{\mathcal{S}}$, every admissible strategy $\nu \in \mathcal{A}(s)$, with $K_{\tau}^{s, \nu}>Y_{\tau}^{s, \nu}$ (i.e., $B_{\tau}^{s, \nu}>P_{\tau}$ ) for some stopping time $\tau$, can be improved strictly by realizing the capital loss on the entire portfolio at time $\tau$. This property is observed in practice and is known as a wash sale. It was stated in [12] and embedded directly in the definition of the tax basis. This result is independent of the choice of the utility function.

Proposition 3.5. Consider some $s \in \overline{\mathcal{S}}$ and $\nu=(C, L, M) \in \mathcal{A}(s)$. Assume that $K_{\tau}^{s, \nu}>$ $Y_{\tau}^{s, \nu}$ a.s. for some finite stopping time $\tau$. Then, there exists an admissible strategy $\tilde{\nu}=$ $(\tilde{C}, \tilde{L}, \tilde{M}) \in \mathcal{A}(s)$ such that, for any utility function,

$$
Y^{\tilde{\nu}}=Y^{\nu}, \quad \Delta \tilde{M}-\Delta M=\mathbf{1}_{\{\tau\}}, \quad \text { and } \quad J_{\infty}(s, \tilde{\nu})>J_{\infty}(s, \nu) ;
$$

i.e., a wash sale is optimal.

Proof. We organize the proof in two steps.

1. Set $\left(L^{\prime}, M^{\prime}\right):=(L, M)+\left(Y_{\tau}, 1\right)\left(1-\Delta M_{\tau}\right) 1_{t \geq \tau}$. We shall prove that $\nu^{\prime}=\left(C, L^{\prime}, M^{\prime}\right) \in$ $\mathcal{A}(s)$ and that the resulting state process satisfies

$$
Y^{s, \nu^{\prime}}=Y^{s, \nu}, \quad Z^{s, \nu^{\prime}} \geq Z^{s, \nu}, \quad K^{s, \nu^{\prime}} \leq K^{s, \nu} \quad \text { a.s. } \quad \text { and } \quad Z_{t}^{\nu^{\prime}}>Z_{t}^{\nu} \quad \text { a.s. } \quad \text { on } \quad\{t>\tau\} .
$$

To see this, observe that since $\nu$ and $\nu^{\prime}$ differ only by the jump at the stopping time $\tau$, and $\Delta Y_{\tau}^{\nu^{\prime}}=\Delta Y_{\tau}^{\nu}$, we have

$$
Y^{s, \nu^{\prime}}=Y^{s, \nu}, \quad\left(Z_{t}^{s, \nu^{\prime}}, K_{t}^{s, \nu^{\prime}}\right)=\left(Z_{t}^{s, \nu}, K_{t}^{s, \nu}\right) \quad \text { for } \quad t<\tau \quad \text { and } \quad Z_{\tau}^{s, \nu^{\prime}}=Z_{\tau}^{s, \nu}
$$

by the continuity of the process $Z$. Observe that the newly defined strategy $\nu^{\prime}$ consists in selling out the whole portfolio at time $\tau$ as $\Delta M_{\tau}^{\prime}=1$. Hence $K_{\tau}^{s, \nu^{\prime}}=Y_{\tau}^{s, \nu^{\prime}}=Y_{\tau}^{s, \nu}$, and we compute directly from (2.8) that

$$
K_{t}^{s, \nu^{\prime}}-K_{t}^{s, \nu}=\left(Y_{\tau}^{s, \nu}-K_{\tau}^{s, \nu}\right) e^{-M_{t}^{c}+M_{\tau}^{c}} \prod_{\tau<u \leq t}\left(1-\Delta M_{u}\right)>0 \quad \text { for } \quad t \geq \tau
$$

since $Y_{\tau}^{s, \nu}-K_{\tau}^{s, \nu}<0$. By (2.11), this provides

$$
e^{-r t}\left(Z_{t}^{s, \nu^{\prime}}-Z_{t}^{s, \nu}\right)=-r \alpha \int_{\tau}^{t} e^{-r u}\left(K_{u}^{s, \nu^{\prime}}-K_{u}^{s, \nu}\right) d u>0 \quad \text { for } \quad t>\tau
$$

which shows that $Z^{s, \nu^{\prime}} \geq 0$ and $\nu^{\prime} \in \mathcal{A}(s)$.
2. Define the strategy $\tilde{\nu}=(\tilde{C}, \tilde{L}, \tilde{M})$ by

$$
\begin{equation*}
\tilde{C}_{t}:=C_{t}+\xi\left(Z_{t}^{s, \tilde{\nu}}-Z_{t}^{s, \nu}\right) 1_{t \geq \tau} \quad \text { and } \quad(\tilde{L}, \tilde{M}):=\left(L^{\prime}, M^{\prime}\right), \tag{3.2}
\end{equation*}
$$

where $\xi$ is an arbitrary positive constant. Observe that $\left(Y^{s, \tilde{\nu}}, K^{s, \tilde{\nu}}\right)=\left(Y^{s, \nu^{\prime}}, K^{s, \nu^{\prime}}\right)$, and $Z_{t}^{s, \tilde{\nu}}=Z_{t}^{s, \nu^{\prime}}=Z_{t}^{s, \nu}$ for $t \leq \tau$. In particular, $K^{s, \tilde{\nu}}-K^{s, \nu}=K^{s, \nu^{\prime}}-K^{s, \nu} \leq 0$. Set $\Delta K:=$ $K^{s, \tilde{\nu}}-K^{s, \nu}$ and $\Delta Z:=Z^{s, \tilde{\nu}}-Z^{s, \nu}$. In order to check the admissibility of the strategy $\tilde{\nu}$, we directly compute that

$$
\begin{aligned}
e^{-r(t-\tau)} \Delta Z_{t} & =\Delta Z_{\tau}-r \alpha \int_{\tau}^{t} e^{-r(u-\tau)} \Delta K_{u} d u+\xi \int_{\tau}^{t} e^{-r(u-\tau)} \Delta Z_{u} d u \\
& \geq \xi \int_{\tau}^{t} e^{-r(u-\tau)} \Delta Z_{u} d u
\end{aligned}
$$

 $\{t>\tau\}$ with positive Lebesgue $\otimes P$ measure. Hence $J_{\infty}(s ; \tilde{\nu})>J_{\infty}(s ; \nu)$.

Remark 3.1. It follows from the previous proposition that $V(x, y, k)=V(x+y+\alpha(k-y)$, $0,0)$ whenever $k>y$. Then, we may restrict our analysis of the value function to the set $\{(x, y, k) \in \overline{\mathcal{S}}: k \leq y\}$. We could not find any benefit from this reduction. Even the numerical implementation is not simplified by this domain restriction because we have no natural boundary condition on $\{k=y\}$. We therefore continue our analysis of the value function $V$ on the total domain $\overline{\mathcal{S}}$.

Remark 3.2. The previous proposition highlights the difficulty in characterizing a solution of the optimal consumption-investment problem under taxes. The optimality of wash sales suggests that the optimal trading strategy has a local time type of behavior. We have no theoretical result to support this intuition. Our very limited information on the regularity of the value function is the main obstacle in developing a verification argument similar to that of Davis and Norman [13].
4. The first order approximation. In this section, we provide upper and lower bounds for the value function. The upper bound expresses that there is no way for the investor to take advantage of tax credits in order to do better than in the tax-free financial market, and this holds for any utility function. Our derivation of the lower bound explicitly uses the power utility but without any restriction on the risk aversion factor $p$. These bounds will be used in order to obtain a first order approximation result for $p<1$.

Proposition 4.1. For any $s=(x, y, k) \in \overline{\mathcal{S}}$, we have $V(s) \leq \bar{V}(x+(1-\alpha) y+\alpha k)$.
Proof. Let $s=(x, y, k)$ be in $\overline{\mathcal{S}}$. Consider some consumption-investment strategy $\nu=$ $(C, L, M)$ in $\mathcal{A}(s)$. Define a consumption-investment strategy $\tilde{\nu}=(C,(1-\alpha) L, M)$, and denote by $(\tilde{X}, \tilde{Y})$ the corresponding tax-free bank and risky asset account processes with the initial endowment $(x+\alpha k,(1-\alpha) y)$. Clearly, $\tilde{Y}=(1-\alpha) Y^{s, \nu} \geq 0$. To see that $\tilde{\nu}$ is admissible in the tax-free financial market, observe that the process $\tilde{Z}:=\tilde{X}+\tilde{Y}$ satisfies

$$
\tilde{Z}_{0-}-Z_{0-}^{s, \nu}=0 \quad \text { and } \quad \tilde{Z}_{t}-Z_{t}^{s, \nu} \geq e^{r t} \int_{0}^{t} e^{-r u} r \alpha K_{u}^{s, \nu} d u \geq 0
$$

so that $\tilde{Z}^{s, \nu} \geq Z^{s, \nu} \geq 0$. Hence, $\bar{V}(x+(1-\alpha) y+\alpha k) \geq J_{\infty}(s, \nu)$; see Remark 2.4. The required result follows from the arbitrariness of $\nu \in \mathcal{A}(s)$.

Proposition 4.2. For $s=(x, y, k)$ in $\overline{\mathcal{S}}$ and $z=x+(1-\alpha) y+\alpha k$, there exists a sequence of admissible strategies $\left(\nu^{n}\right)_{n \geq 1} \subset \mathcal{A}(s)$ such that

$$
V(s) \geq \bar{c}\left(r, \tilde{\theta}^{\alpha}\right)^{p-1} \frac{z^{p}}{p}=\lim _{n \rightarrow \infty} J_{\infty}\left(s, \nu^{n}\right), \quad \text { where } \quad \tilde{\theta}^{\alpha}:=\theta-\frac{r \alpha}{\sigma(1-\alpha)}
$$

i.e., the value function of the Merton frictionless problem with the smaller risk premium $\tilde{\theta}^{\alpha}$ can be approached as close as possible in the context of the financial market with taxes.

This result is proved by producing a sequence of admissible strategies $\left(C_{n}, L_{n}, M_{n}\right)_{n \geq 1} \subset$ $\mathcal{A}(s)$ which approximates Merton's value function with the smaller risk premium $\tilde{\theta}^{\alpha}$. To give an intuitive justification of this result, we rewrite (2.11) as

$$
\begin{equation*}
d Z_{t}=\left(r Z_{t}-C_{t}\right) d t+Y_{t} \tilde{\sigma}^{\alpha}\left(d W_{t}+\tilde{\theta}^{\alpha} d t\right)+r \alpha\left(Y_{t}-K_{t}\right) d t \tag{4.1}
\end{equation*}
$$

where $\tilde{\theta}^{\alpha}$ is defined as in the statement of Proposition 4.2 and $\tilde{\sigma}^{\alpha}:=(1-\alpha) \sigma$. Compare the above $\bar{Z}$ dynamics to (2.16) with modified parameters $\left(\tilde{\sigma}^{\alpha}, \tilde{\theta}^{\alpha}\right)$ and with $c_{t}=C_{t} / \bar{Z}_{t}$, $\pi_{t}=Y_{t} / \bar{Z}_{t}$. The only difference is the term $r \alpha(Y-K)$. However, in view of Proposition 3.5, we expect this term to be nonnegative for the optimal strategy (if it exists). This hints that the liquidation value process $\bar{Z}$ (with the above choices $C$ and $Y$ ) is larger than the wealth process in the fictitious tax-free financial market with a modified risk premium. This formally justifies the inequality of Proposition 4.2.

The proof reported in Appendix B exhibits an explicit sequence of strategies which mimics the optimal consumption-investment strategy in the Merton frictionless model while keeping the difference $Y-K$ small or, equivalently, the tax basis close to the spot price of the risky asset.

Remark 4.1. Let $b:=r+\theta \sigma$ be the instantaneous mean return coefficient in our financial market. Then, the modified risk premium $\tilde{\theta}^{\alpha}$ can be easily interpreted in terms of the modified volatility coefficient $\sigma^{\alpha}=(1-\alpha) \sigma$ and a similarly modified instantaneous mean return coefficient $b^{\alpha}:=(1-\alpha) b$ as $\tilde{\theta}^{\alpha}=\left(b^{\alpha}-r\right) / \sigma^{\alpha}$. This fictitious financial market with such modified coefficients corresponds to the situation where the investor is forced to realize the capital gains or losses, at each time $t$, before adjusting the portfolio.

Propositions 4.1 and 4.2 provide the following bounds on the value function $V$ :

$$
\begin{equation*}
\frac{\bar{c}\left(r, \tilde{\theta}^{\alpha}\right)^{p-1}}{p}(x+(1-\alpha) y+\alpha k)^{p} \leq V(x, y, k) \leq \frac{\bar{c}(r, \theta)^{p-1}}{p}(x+(1-\alpha) y+\alpha k)^{p}, \tag{4.2}
\end{equation*}
$$

where $\tilde{\theta}^{\alpha}$ is defined as in the statement of Proposition 4.2, and $\bar{c}$ is defined as in Theorem 2.1. Observe that $\tilde{\theta}^{\alpha}=\theta$ whenever $\alpha=0$ or $r=0$. Therefore, we might expect that these bounds are tight for small interest rate or tax parameters.

Corollary 4.1. For $s=(x, y, k) \in \mathcal{S}$, we have

$$
V(s)=V^{\mathrm{app}}(s)+\mathrm{o}(\alpha+r),
$$

where $\mathrm{o}(\xi)$ is a function on $\mathbb{R}$ with $\mathrm{o}(\xi) / \xi \longrightarrow 0$ as $\xi \rightarrow 0$, and

$$
V^{\mathrm{app}}(s):=\frac{\bar{c}(0, \theta)^{p-1}}{p}\left(1+\frac{r p}{\bar{c}(0, \theta)}\right)(x+y)^{p}+\alpha \bar{c}(0, \theta)^{p-1}(k-y)(x+y)^{p-1} .
$$

Proof. It is sufficient to observe that the bounds on the value function $V$ in (4.2) are smooth functions with the identical partial gradient with respect to $(r, \alpha)$ at the origin. This follows from the fact that $\left(\partial \tilde{\theta}^{\alpha} / \partial \alpha\right)=\left(\partial \tilde{\theta}^{\alpha} / \partial r\right)=0$ at $(r, \alpha)=(0,0)$.

Remark 4.2. Observe that the function $\bar{c}$ defined in Theorem 2.1 is increasing in the $r$ variable. Then, the above first order expansion shows that the value function $V$ is also increasing in the interest rate variable (for small interest rate and tax parameters). This is intuitively clear, as the larger interest rate provides the investor a better opportunity set.

The dependence of the value function on the tax rate $\alpha$ is more complex, and it depends on the initial position of the tax basis. If the initial tax basis is larger than the spot price, i.e., in a situation of capital gain loss, the investor takes immediate advantage of the tax credit, as stated in Proposition 3.5, and the value function $V$ is increasing in $\alpha$ (for small $\alpha$ ). In the opposite situation, i.e., when the initial tax basis is smaller than the spot price, the value function is decreasing in $\alpha$. Finally, when the initial tax basis coincides with the spot price, the value function is not sensitive to the tax rate in the first order.

This variation of the value function (up to the first order) in terms of the tax rate $\alpha$ is somehow surprising. Indeed, in a capital loss situation, an increase of the tax parameter implies two opposing results:

- an increase of the tax credit is received initially by the agent;
- a larger amount of tax is paid during the infinite lifetime of the agent.

Our first order expansion shows that, for small interest rate and tax parameters, the increase of initial tax credit is never compensated by the increase of tax over the infinite lifetime. This is in agreement with Cadenillas and Pliska [7], who found that "sometimes investors are better off with a positive tax rate."

The same reasoning also shows that when there are no initial embedded capital gains (i.e., when $y=k=0$ ) the effect of the tax parameter is only second order.

Remark 4.3. Since the lower bound in (4.2) has the same first order Taylor expansion as the value function $V$, we can view the corresponding strategy as nearly optimal. From the discussion following Proposition 4.2, the portfolio allocation defining the lower bound is by definition an approximation of the constant portfolio allocation

$$
\bar{\pi}\left(\tilde{\sigma}^{\alpha}, \tilde{\theta}^{\alpha}\right)=\frac{1}{(1-p) \sigma^{2}}\left[\frac{b}{1-\alpha}-\frac{r}{(1-\alpha)^{2}}\right]
$$

where $b:=\sigma \theta+r$ is the instantaneous mean return of the risky asset. Direct computation shows that $\bar{\pi}\left(\tilde{\sigma}^{\alpha}, \tilde{\theta}^{\alpha}\right) \leq \bar{\pi}(\sigma, \theta)$ if and only if $r \geq(1-\alpha)(b-(1-\alpha)(b-r))$. Using the data set of Dammon, Spatt, and Zhang [12] ( $r=6 \%, b=9 \%, \alpha=36 \%$ ), we see that $\bar{\pi}\left(\tilde{\sigma}^{\alpha}, \tilde{\theta}^{\alpha}\right) \leq \bar{\pi}(\sigma, \theta)$. Since this nearly optimal strategy exhibits a smaller exposition to the risky asset, the presence of taxes on capital gains contributes to explaining the equity premium puzzle highlighted by Mehra and Prescott [25].

Notice that this observation is in contradiction with the numerical results of Dammon, Spatt, and Zhang [12], who found that the exposition to the risky asset is increased by the presence of taxes. This is due to the fact that the bank account in their model is also subject to taxes with the same tax rate as for the risky asset, which implies that the optimal portfolio strategy in the forced realization case is given by

$$
\hat{\pi}^{\alpha}=\frac{b(1-\alpha)-r(1-\alpha)}{(1-p) \sigma^{2}(1-\alpha)^{2}}=\frac{\bar{\pi}(\sigma, \theta)}{1-\alpha}
$$

which is increasing in $\alpha$.
5. Characterization by the dynamic programming equation. The chief goal of this section is to provide a characterization of $V$ by means of a second order partial differential equation for which we shall provide a numerical solution in the subsequent section. Unfortunately, we are unable to obtain a characterization of $V$ by the corresponding dynamic programming equation. Therefore, we shall exhibit a consistent approximation $V^{\varepsilon}$ as the unique solution of an approximating second order partial differential equation.

For $s$ in $\overline{\mathcal{S}}$ and $\nu=(C, L, M)$ in $\mathcal{A}$, the jumps of the state processes $S$ are given by

$$
\Delta S_{t}^{s, \nu}=-\Delta L_{t} \mathbf{g}^{\mathbf{b}}-\Delta M_{t}\left[(1-\alpha) Y_{t-}^{s, \nu}+\alpha K_{t-}^{s, \nu}\right] \mathbf{g}^{\mathbf{s}}\left(S_{t-}^{s, \nu}\right)
$$

where the vector fields $\mathbf{g}^{\mathbf{b}}$ and $\mathbf{g}^{\mathbf{s}}(x, y, k)$ are defined by

$$
\mathbf{g}^{\mathbf{b}}:=\left(\begin{array}{c}
1 \\
-1 \\
-1
\end{array}\right) \quad \text { and } \quad \mathbf{g}^{\mathbf{s}}(s):=\left(\begin{array}{c}
-1 \\
\frac{1}{1-\alpha} \\
0
\end{array}\right)+\left(\begin{array}{c}
0 \\
\frac{-\alpha}{1-\alpha} \\
1
\end{array}\right) \frac{k}{(1-\alpha) y+\alpha k} \mathbf{1}_{(y, k) \neq 0}
$$

The value function of our consumption-investment problem is formally expected to solve the corresponding dynamic programming equation:

$$
\begin{equation*}
\min \left\{-\mathcal{L} v, \mathbf{g}^{\mathbf{b}} \cdot D v, \mathbf{g}^{\mathbf{s}} \cdot D v\right\}=0 \quad \text { on } \quad \overline{\mathcal{S}} \backslash \partial^{z} \mathcal{S} \quad \text { and } \quad v=0 \quad \text { on } \quad \partial^{z} \mathcal{S} \tag{5.1}
\end{equation*}
$$

where $\mathcal{L}$ is the second order differential operator defined by

$$
\mathcal{L} v=-\beta v+r x v_{x}+b y v_{y}+\frac{1}{2} \sigma^{2} y^{2} v_{y y}+\tilde{U}\left(v_{x}\right), \quad \text { with } \quad \tilde{U}(q)=\sup _{c \geq 0}(U(c)-c q) .
$$

Observe that we have no information on the regularity of the value function $V$; hence we cannot prove that $V$ is a classical solution to (5.1). Moreover, the value function $V$ is known only on the boundary $\partial^{z} \mathcal{S}$ (see Proposition 3.2), but there is no possible knowledge of $V$ on $\partial^{y} \mathcal{S} \cup \partial^{k} \mathcal{S}$. We then need to use the notion of viscosity solutions which allows for a weak formulation of solutions to partial differential equations and boundary conditions.

Because $\mathbf{g}^{\mathbf{s}}$ is not locally Lipschitz continuous, it is not clear that there is a unique characterization of the value function $V$ as the constrained viscosity solution of (5.1). Due to this technical difficulty, we isolated in the accompanying paper [5] some viscosity results for a slightly modified equation. The objective of this section is to build on these results in order to characterize the value function $V$ as a limit of an approximating function defined as the unique viscosity solution of a conveniently perturbed equation.

For every $\varepsilon>0$, we define the function

$$
f^{\varepsilon}(x, y, k):=1 \wedge\left(\frac{k}{\varepsilon z}-1\right)^{+}, \quad \text { with } \quad z:=x+(1-\alpha) y+\alpha k
$$

together with the approximation of $\mathbf{g}^{\mathbf{b}}$ and $\mathbf{g}^{\mathbf{s}}$ :

$$
\mathbf{g}_{\varepsilon}^{\mathbf{b}}:=\left(\begin{array}{c}
1+\varepsilon \\
-1 \\
-1
\end{array}\right) \quad \text { and } \quad \mathbf{g}_{\varepsilon}^{\mathbf{s}}(x, y, k):=\mathbf{g}^{\mathbf{s}}\left(x, y, k f^{\varepsilon}(s)\right)
$$

for $s \in \mathcal{S} \backslash \partial^{z} \mathcal{S}$. Notice that $\mathbf{g}_{\varepsilon}^{\mathbf{s}}$ is locally Lipschitz continuous on $\overline{\mathcal{S}} \backslash \partial^{z} \mathcal{S}$, and $\mathbf{g}_{\varepsilon}^{\mathbf{s}}(s)=\mathbf{g}^{\mathbf{s}}(s)$ whenever $k \geq 2 \varepsilon z$. The main result of this section provides a characterization of the value function $V$ by means of the approximating equation:

$$
\begin{equation*}
\min \left\{-\mathcal{L} v, \mathbf{g}_{\varepsilon}^{\mathbf{b}} \cdot D v, \mathbf{g}_{\varepsilon}^{\mathbf{s}} \cdot D v\right\}=0 \quad \text { on } \quad \overline{\mathcal{S}} \backslash \partial^{z} \mathcal{S} \quad \text { and } \quad v=0 \quad \text { on } \quad \partial^{z} \mathcal{S} \tag{5.2}
\end{equation*}
$$

We first recall the notion of a constrained viscosity solution first introduced in [29, 30]. For a locally bounded function $u: \overline{\mathcal{S}} \longrightarrow \mathbb{R}$, we shall use the classical notation in viscosity theory for the corresponding upper semicontinuous and lower semicontinuous envelopes:

$$
u^{*}(s):=\limsup _{\mathcal{S} \ni s^{\prime} \rightarrow s} u\left(s^{\prime}\right) \quad \text { and } \quad u_{*}(s):=\liminf _{\mathcal{S} \ni s^{\prime} \rightarrow s} u\left(s^{\prime}\right)
$$

Definition 5.1. (i) A locally bounded function $u$ is a constrained viscosity subsolution of (5.2) if $u^{*} \leq 0$ on $\partial^{z} \mathcal{S}$, and for all $s \in \overline{\mathcal{S}} \backslash \partial^{z} \mathcal{S}$ and $\varphi \in C^{2}(\overline{\mathcal{S}})$ with $0=\left(u^{*}-\varphi\right)(s)=$ $\max _{\overline{\mathcal{S}} \backslash \partial^{z} \mathcal{S}}\left(u^{*}-\varphi\right)$ we have $\min \left\{-\mathcal{L} \varphi, \mathbf{g}_{\varepsilon}^{\mathbf{b}} \cdot D \varphi, \mathbf{g}_{\varepsilon}^{\mathbf{s}} \cdot D \varphi\right\} \leq 0$.
(ii) A locally bounded function $u$ is a constrained viscosity supersolution of (5.2) if $u_{*} \geq 0$ on $\partial^{z} \mathcal{S}$, and for all $s \in \mathcal{S}$ and $\varphi \in C^{2}(\mathcal{S})$ with $0=\left(u_{*}-\varphi\right)(s)=\min _{\mathcal{S}}\left(u_{*}-\varphi\right)$ we have $\min \left\{-\mathcal{L} \varphi, \mathbf{g}_{\varepsilon}^{\mathbf{b}} \cdot D \varphi, \mathbf{g}_{\varepsilon}^{\mathbf{s}} \cdot D \varphi\right\} \geq 0$.
(iii) A locally bounded function $u$ is a constrained viscosity solution of (5.2) if it is a constrained viscosity subsolution and supersolution.

In the above definition, there is no boundary value assigned to the value function on $\partial^{y} \mathcal{S} \cup$ $\partial^{k} \mathcal{S}$. Instead, the subsolution property holds on this boundary. Notice that the supersolution property is satisfied only in the interior of the domain $\mathcal{S}$.

Theorem 5.1. For each $\varepsilon>0$, the boundary value problem (5.2) has a unique constrained viscosity solution $V^{\varepsilon}$ in the class $C_{p}^{0}(\overline{\mathcal{S}})$. Moreover, the following hold:
(i) the family $\left(V^{\varepsilon}\right)_{\varepsilon>0}$ is nonincreasing and converges to the value function $V$ uniformly on compact subsets of $\overline{\mathcal{S}}$ as $\varepsilon \searrow 0$;
(ii) for every $s \in \mathcal{\mathcal { S }}$ and $\delta \geq 0$, we have $V^{\varepsilon}(\delta s)=\delta^{p} V^{\varepsilon}(s)$.

Proof. The existence of $V^{\varepsilon}$ as the unique constrained viscosity solution of (5.2) in the class $C_{p}^{0}(\overline{\mathcal{S}})$ is shown in Theorem 3.2 of [5], where we introduced the value function $v^{\varepsilon, \lambda}$ of a consumption-investment problem with transaction costs $\lambda>0$ and an $\varepsilon$-modified taxation rule near the ray $\left\{(x, 0,0), x \in \mathbb{R}_{+}\right\}$. Here $V^{\varepsilon}=v^{\varepsilon, \varepsilon}$.

We next use Theorem 5.1 of [5], which provides a stochastic control representation of $v^{\varepsilon, \lambda}$. In particular, it follows directly from this representation that $v^{\varepsilon, \lambda}(\delta s)=\delta^{p} v^{\varepsilon, \lambda}(s)$ for all $\delta \geq 0$. This homotheticity property is obviously inherited by $V^{\varepsilon}=v^{\varepsilon, \varepsilon}$, as announced in (ii).

We now prove (i) in the next two steps.
Step 1. The monotonicity of the sequence $\left(V^{\varepsilon}\right)_{\varepsilon>0}$ is inherited from the nonincrease of the sequence $\left(v^{\varepsilon, \lambda}\right)_{\varepsilon>0}$, proved in Proposition 6.2 of [5], together with the decrease of the sequence $\left(v^{\varepsilon, \lambda}\right)_{\lambda>0}$. It follows from this monotonicity property that $V^{0}:=\lim _{\varepsilon \rightarrow 0} V^{\varepsilon}$ is well defined. Now observe that $v^{\varepsilon, \lambda} \leq V^{\varepsilon} \leq v^{0, \varepsilon} \leq V$ for every $\lambda \geq \varepsilon$. Then $\lim _{\lambda \rightarrow 0} \lim _{\varepsilon \rightarrow 0} v^{\varepsilon, \lambda} \leq V^{0} \leq V$. By Proposition 6.3 of [5], this implies that

$$
\begin{equation*}
\lim _{\lambda \rightarrow 0} v^{\lambda} \leq v^{0, \varepsilon} \leq V, \quad \text { where } \quad v^{\lambda}:=\lim _{\varepsilon \rightarrow 0} v^{\varepsilon, \lambda} \tag{5.3}
\end{equation*}
$$

is the value function of the optimal consumption-investment problem with taxes and proportional transaction cost $\lambda>0$. In order to conclude that $V^{0}=V$, it remains to show that

$$
\lim _{\lambda \rightarrow 0} v^{\lambda}=V
$$

Step 2. For fixed $s \in \overline{\mathcal{S}}$, let $\nu^{n}=\left(C^{n}, L^{n}, M^{n}\right)$ be such that

$$
V(s)-\frac{1}{n} \leq J_{\infty}\left(s, \nu^{n}\right) \quad \text { for all } \quad n \geq 1
$$

Denote by $Z^{\lambda, n}$ the after-tax liquidation value process with consumption-investment strategy $\nu^{n}$ in a financial market subject to taxes and constant proportional transaction cost parameter $\lambda$. See Appendix A for the precise formulation. We also denote by $Z^{n}$ the corresponding aftertax liquidation value process without transaction costs. Then, it follows immediately from the dynamics of these processes that

$$
Z_{t}^{\lambda, n}=Z_{t}^{n}-\lambda L_{t}^{n} \quad \text { for all } \quad t \geq 0
$$

Then the stopping times

$$
\tau^{\lambda}:=\inf \left\{t>0: Z^{\lambda, n} \leq \frac{1}{2} Z_{t}^{n}\right\} \quad \text { satisfy } \quad \tau^{\lambda} \longrightarrow \infty \quad \text { as } \quad \lambda \searrow 0 \quad \mathbb{P} \text {-a.s. }
$$

Define the strategies $\tilde{\nu}^{n}$ by

$$
\tilde{\nu}^{n}:=\nu^{n} \mathbf{1}_{\left[0, \tau^{\lambda}\right)}+\left(0, L_{\tau^{\lambda}-}^{n}, M_{\tau^{\lambda}-}^{n}+1\right) \mathbf{1}_{\left[\tau^{\lambda}, \infty\right)} .
$$

Clearly, $\tilde{\nu}^{n}$ is admissible for the problem with transaction costs, i.e., $\tilde{\nu}^{n} \in \mathcal{A}^{\lambda}(s)$ in the notation of Appendix A. Then

$$
v^{\lambda}(s) \geq J_{\infty}\left(s, \tilde{\nu}^{n}\right)=\mathbb{E}\left[\int_{0}^{\tau^{\lambda}} U\left(C_{t}^{n}\right) d t\right] \rightarrow \mathbb{E}\left[\int_{0}^{\infty} U\left(C_{t}^{n}\right) d t\right] \geq V(s)-\frac{1}{n},
$$

where we use the monotone convergence theorem. By the arbitrariness of $n$ and (5.3), this shows that $V^{0}=V$. Finally the convergence holds uniformly on compact subsets by the monotonicity of $\left(V^{\varepsilon}\right)$ and the continuity of the limit $V$.
6. Numerical estimate for $\boldsymbol{V}$. We have stated in the previous section that the value function $V$ is approximated by the functions $\left(V^{\varepsilon}\right)_{\varepsilon>0}$, where, for each $\varepsilon>0, V^{\varepsilon}$ can be computed as the unique viscosity solution of the boundary value problem (5.2). In this section, we provide a numerical estimate for $V$, based on a numerical scheme for (5.2). Unfortunately, we have no theoretical convergence result for our algorithm. Indeed, as discussed in the introduction, establishing convergence results for Hamilton-Jacobi-Bellman equations corresponding to singular control problems is an open question in numerical analysis. We therefore follow previous related works such as [1] and [31] by concentrating our effort on realizing the empirical convergence of the algorithm.

Following Akian, Menaldi, and Sulem [1], we adopt a numerical scheme based on the finite difference discretization and the classical Howard algorithm. For the convenience of the reader we briefly describe it hereafter, and we refer the reader to [1] for a detailed discussion.
6.1. Change of variables and reduction of the state dimension. By the homotheticity property of $V^{\varepsilon}$ (Theorem 5.1(ii)) we have for $s=(x, y, k) \in \overline{\mathcal{S}} \backslash \partial^{z} \mathcal{S}$ and $z:=x+[(1-\alpha) y+\alpha k]$

$$
V^{\varepsilon}(s)=z^{p} \mathcal{V}^{\varepsilon}\left(\frac{y}{z}, \frac{k}{z}\right), \quad \text { where } \quad \mathcal{V}^{\varepsilon}\left(\xi_{1}, \xi_{2}\right):=V^{\varepsilon}\left(1-(1-\alpha) \xi_{1}-\alpha \xi_{2}, \xi_{1}, \xi_{2}\right)
$$

Next, for a vector $\xi \in \mathbb{R}_{+}^{2}$, we define the vector $\zeta \in[0,1)^{2}$ by

$$
\zeta_{i}:=\xi_{i} /\left(1+\xi_{i}\right), \quad i=1,2, \quad \text { and } \quad \Psi^{\varepsilon}(\zeta):=\mathcal{V}^{\varepsilon}(\xi)
$$

This reduces the domain of $\mathcal{V}^{\varepsilon}$ from $\mathbb{R}_{+}^{2}$ to the bounded domain $[0,1)^{2}$. By changing variables, it is immediately checked that $\Psi^{\varepsilon}$ is a continuous constrained viscosity solution on $[0,1) \times[0,1)$ of

$$
\begin{equation*}
\min _{a \in \mathbb{A}}\left\{\beta(a) \Psi^{\varepsilon}(\zeta)-\sum_{i=1}^{2} b_{i}(a, \zeta) \cdot D_{i} \Psi^{\varepsilon}(\zeta)-\frac{1}{2} \sum_{i, j=1}^{2} \eta_{i j}(a, \zeta) D_{i j}^{2} \Psi^{\varepsilon}(\zeta)-g(a)\right\}=0 \tag{6.1}
\end{equation*}
$$

where the control set $\mathbb{A}$ and the expressions of $\beta,\left(b_{i}\right)_{i=1,2},\left(\eta_{i, j}\right)_{i, j=1,2}$, and $g$ are obtained by immediate calculation.
6.2. Finite differences for (6.1). We adopt a classical finite difference discretization in order to obtain a numerical scheme for (6.1). Let $N$ be a positive integer, and set $h:=\frac{1}{N}$, the finite difference step; we set $e_{1}:=(1,0)$ and $e_{2}:=(0,1)$, and we define the uniform grid $\overline{\mathcal{S}}_{h}:=[0,1]^{2} \cap(h \mathbb{Z})^{2}$. We denote by $\zeta^{h}:=\left(\zeta_{1}^{h}, \zeta_{2}^{h}\right)$ a point of the grid $\overline{\mathcal{S}}_{h}$, and we set $\mathcal{S}_{h}:=(0,1) \times[0,1) \cap(h \mathbb{Z})^{2}$. In order to define a discretization of (6.1), we approximate the partial derivatives of $\Psi^{\varepsilon}$ by the corresponding backward and forward finite differences

$$
\begin{aligned}
b_{i}(a, \zeta) \partial_{i} \Psi^{\varepsilon}(\zeta) & \approx \begin{cases}b_{i}(a, \zeta) D_{i}^{+} \Psi^{\varepsilon}(\zeta) & \text { if } b_{i}(a, \zeta) \geq 0, \\
b_{i}(a, \zeta) D_{i}^{-} \Psi^{\varepsilon}(\zeta) & \text { if } b_{i}(a, \zeta)<0,\end{cases} \\
\partial_{i i} \Psi^{\varepsilon}(\zeta) & \approx D_{i}^{2} \Psi^{\varepsilon}(\zeta), \\
\eta_{i j}(a, \zeta) \partial_{i j} \Psi^{\varepsilon}(\zeta) & \approx \begin{cases}\eta_{i j}(a, \zeta) D_{i j}^{+} \Psi^{\varepsilon}(\zeta) & \text { if } \eta_{i j}(a, \zeta) \geq 0, \\
\eta_{i j}(a, \zeta) D_{i j}^{-} \Psi^{\varepsilon}(\zeta) & \text { if } \eta_{i j}(a, \zeta)<0,\end{cases}
\end{aligned}
$$

where the finite difference operators are defined for $i \neq j \in\{1,2\}$ by

$$
\begin{aligned}
& D_{i}^{+} \Psi^{\varepsilon}(\zeta)=\frac{\Psi^{\varepsilon}\left(\zeta+h e_{i}\right)-\Psi^{\varepsilon}(\zeta)}{h}, \quad D_{i}^{-} \Psi^{\varepsilon}(\zeta)=\frac{\Psi^{\varepsilon}(\zeta)-\Psi^{\varepsilon}\left(\zeta-h e_{i}\right)}{h} \\
& D_{i}^{2} \Psi^{\varepsilon}(\zeta)= \\
& D_{i j}^{ \pm} \Psi^{\varepsilon}(\zeta)=\frac{\Psi^{\varepsilon}\left(\zeta+h e_{i}\right)-2 \Psi^{\varepsilon}(\zeta)+\Psi^{\varepsilon}\left(\zeta-h e_{i}\right)}{h^{2}}, \\
& \\
& \quad \begin{array}{l}
2 h^{2} \\
\\
\quad-\Psi^{\varepsilon}(\zeta)+\Psi^{\varepsilon}\left(\zeta+h e_{i} \pm h e_{j}\right)+\Psi^{\varepsilon}\left(\zeta-h e_{i} \mp h e_{j}\right)
\end{array} \\
& \left.\quad \Psi^{\varepsilon}\left(\zeta-h e_{i}\right)-\Psi^{\varepsilon}\left(\zeta+h e_{j}\right)-\Psi^{\varepsilon}\left(\zeta-h e_{j}\right)\right\} .
\end{aligned}
$$

In order to compute these differences at every point of $\mathcal{S}_{h}$, we extend $\Psi^{\varepsilon}$ as follows:

$$
\Psi^{\varepsilon}\left(\zeta_{0}^{h}\right)=\Psi^{\varepsilon}\left(\zeta_{0}^{h}+h e_{1}\right), \quad \Psi^{\varepsilon}\left(\zeta_{1}^{h}\right)=\Psi^{\varepsilon}\left(\zeta_{1}^{h}-h e_{1}\right)
$$

for $\zeta_{0}^{h} \in\{0\} \times[0,1]$ and $\zeta_{1}^{h} \in\{1\} \times[0,1]$ and

$$
\Psi^{\varepsilon}\left(\zeta_{0}^{h}-h e_{2}\right)=\Psi^{\varepsilon}\left(\zeta_{0}^{h}\right), \quad \Psi^{\varepsilon}\left(\zeta_{1}^{h}\right)=\Psi^{\varepsilon}\left(\zeta_{1}^{h}-h e_{2}\right)
$$

for $\zeta_{0}^{h} \in[0,1] \times\{0\}$ and $\zeta_{1}^{h} \in[0,1] \times\{1\}$. This provides a system of $(N-1) N$ nonlinear equations with the $(N-1) N$ unknowns $\Psi_{h}^{\varepsilon}\left(\zeta^{h}\right), \zeta^{h} \in \mathcal{S}_{h}$ :

$$
\begin{equation*}
\min _{a \in \mathbb{A}}\left\{A_{h}^{a} \Psi_{h}^{\varepsilon}-g(a)\right\}=0 \tag{6.2}
\end{equation*}
$$

6.3. The classical Howard algorithm. To solve (6.2) we adopt the classical Howard algorithm, which can be described as follows:

Step 0: $\quad$ start from an initial value for the control $a^{0} \in \mathbb{A}$, $\Psi_{h}^{0}$ solution of $A_{h}^{a^{0}} \varphi-g\left(a^{0}\right)=0$,
Step $k+1, k \geq 0$ : find $a^{k+1} \in \operatorname{argmin}_{a \in \mathbb{A}}\left\{A_{h}^{a} \Psi_{h}^{k}-g(a)\right\}$,

$$
\Psi_{h}^{k+1} \text { solution of } A_{h}^{a^{k+1}} \varphi-g\left(a^{k+1}\right)=0
$$

7. Numerical results. We implement the above numerical algorithm with the following parameters:

$$
p=0.3, \quad \sigma=0.3, \quad \text { and } \quad \beta=0.1
$$

We also fix the instantaneous mean return of the risky asset to

$$
b:=\theta \sigma+r=0.11
$$

Our numerical experiments showed that by taking a finite difference step $1 / 20 \leq h \leq 1 / 40$ our algorithm converges within a reasonable computation time: convergence error $\mid \Psi_{h}^{k+1}-$ $\Psi_{h}^{k} \mid \sim 10^{-6}$, computation time $\sim 1-5$ minutes.
7.1. Accuracy of the first order Taylor expansion. It is clear that one should not expect this algorithm to be reliable near the boundary of the grid. However, realistic initial points are far from the boundary, and we expect the error to be small for these points. The main purpose of this subsection is to examine the accuracy of the first order approximation for different sets of parameters $r$ and $\alpha$ :

$$
r \in\{0.001, .01, .07\} \quad \text { and } \quad \alpha \in\{.001, .01, .05, .1, .2, .3, .36\}
$$

Figure 1 plots the mean relative error between the results of the first order expansion and the numerical algorithm over all points of the grid:

$$
\frac{1}{N(N-1)} \sum_{i, j} \frac{\left|\mathcal{V}_{\varepsilon}^{h}\left(\zeta_{i j}^{h}\right)-\mathcal{V}^{\text {app }}\left(\zeta_{i j}^{h}\right)\right|}{\mathcal{V}^{\operatorname{app}}\left(\zeta_{i j}^{h}\right)}
$$

where $N(N-1)$ is the total number of points in the grid, $\mathcal{V}_{\varepsilon}^{h}$ is the approximation of $\mathcal{V}_{\varepsilon}$ obtained by our numerical scheme, and

$$
\mathcal{V}^{\text {app }}\left(\xi_{1}, \xi_{2}\right):=V^{\text {app }}\left(1-(1-\alpha) \xi_{1}-\alpha \xi_{2}, \xi_{1}, \xi_{2}\right)
$$

As expected, the relative error is zero at the origin and increases when the values of the parameters $r$ and $\alpha$ increase. The error size is large due to the boundary effects.

Indeed, in Figure 2 we focus our attention on a region away from the boundary and concentrate on $(y, k) \in[0,1]^{2}$ (i.e., a region with small initial embedded capital gains). We observe that the average relative error is remarkably small and is of the order of $4 \%$ for realistic values of $r$ and $\alpha$. This figure is our main numerical result, as it shows the reasonable accuracy of the first order Taylor approximation $V^{\text {app }}$ of the value function $V$.
7.2. Welfare analysis. In view of Remark 4.3 , an $\varepsilon$-maximizing strategy is given by the constant portfolio allocation $\bar{\pi}^{\alpha}$ and the constant consumption-wealth ratio $\bar{c}\left(r, \tilde{\theta}^{\alpha}\right)$. The expected utility realized by following this approximating strategy corresponds to the lower bound $\tilde{V}(z)=\bar{c}\left(r, \tilde{\theta}^{\alpha}\right) z^{p} / p$ of Proposition 4.2.

In order to compare this approximating strategy to the optimal one, we report in Figures 3 and 4 the welfare cost, $z^{*}$ such that $V\left(1-(1-\alpha) \xi_{1}-\alpha \xi_{2}, \xi_{1}, \xi_{2}\right)=\tilde{V}\left(1+z^{*}\right)$, with the following parameters:

$$
p=.3, \quad \beta=.1, \quad b:=r+\theta \sigma=.11, \quad \sigma=.3, \quad \text { and } \quad r=.07 .
$$



Figure 1. Mean relative error on $[0,10]^{2}$.


Figure 3. Welfare cost for $\alpha=0.20$.


Figure 2. Mean relative error on $[0,1]^{2}$.


Figure 4. Welfare cost for $\alpha=0.36$.

The welfare cost is nonincreasing with respect to the tax basis and remains relatively small for reasonable values of the parameters $\alpha$ : it reaches a maximum of $8 \%$ for $\alpha=0.2$ and of $12 \%$ for $\alpha=0.36$.
7.3. Optimal consumption-investment strategies. Throughout this subsection we implement our numerical algorithm with the following parameters: $p=0.3, \beta=0.1, b:=r+\theta \sigma=$ $0.11, \sigma=0.3$, and $r=.07$.

The tax-free model. For $\alpha=0.0$, our algorithm produces the well-known results of the Merton frictionless model. Given the above values of the parameters, Merton's optimal strategy is given by $\bar{\pi}=0.6349$ and $\bar{c}=0.1074$.

Figure 5 reports the numerical solution for the function $\mathcal{V}_{\varepsilon}^{h}$. We verify that the function $\mathcal{V}_{\varepsilon}^{h}$ in this tax-free context does not depend on the variable $\xi_{2}$, so that the value function $V_{\varepsilon}^{h}$ does not depend on the $k$ component. We also see that the value function is concave. Figure 6 reports the optimal investment strategy and produces the expected partition of the state space into three regions:

- The region of no transaction (NT) corresponds to positions such that the proportion of wealth allocated to the risky asset $y /(x+y)$ is equal to $\bar{\pi}$. In this region no position adjustment is considered by the investor.
- The Sell region is where the investor immediately sells risky assets so as to attain the region NT by moving along the ray $(1,-1)$.


Figure 5. $\mathcal{V}_{\varepsilon}^{h}$ for $\alpha=0.0$.



Figure 6. Partition for $\alpha=0.0$.


Figure 7. $\mathcal{V}_{\varepsilon}^{h}$ for $\alpha=0.20$.



Figure 8. $\mathcal{V}_{\varepsilon}^{h}$ for $\alpha=0.36$.

- The Buy region is where the investor immediately purchases risky assets so as to attain the region NT by moving along the ray $(-1,1)$.
We numerically verify again that this partition is independent of the variable $\xi_{2}$.
The value function approximation with taxes. We next concentrate on the case where the tax coefficient is positive. Figures 7 and 8 report the numerical solution for the function $\mathcal{V}_{\varepsilon}^{h}$ for $\alpha=0.2$ and 0.36 . The main observation out of these numerical results is that, for a positive tax parameter, the value function is no longer concave. This surprising feature leads


Figure 9. Partition for $\alpha=0.20$.


Figure 10. Partition for $\alpha=0.36$.
to mathematical difficulties, as we had to derive the dynamic programming equation without any a priori knowledge that the value function is continuous.

Optimal investment strategy under taxes. Figures 9 and 10 show that, for positive $\alpha$, the domain is again partitioned into three nonintersecting regions:

- The no-transaction region NT is where no portfolio adjustment is performed by the optimal investor.
- The Sell region is where the investor immediately sells risky assets so as to attain the region NT by moving towards the origin along the ray $((1-\alpha) y+\alpha k,-y,-k)=$ $-[(1-\alpha) y+\alpha k] \mathbf{g}^{\mathbf{s}}$.
- The Buy region is where the investor immediately purchases risky assets so as to attain the region NT by moving along the ray $(-1,1,1)=-\mathbf{g}^{\mathbf{b}}$.
For positive $\alpha$, the boundaries of the no-transaction region depend on the tax basis. The range of the proportion of wealth allocated to the risky asset, $(y / z)$, for which no-transaction is optimal, is very sensible to the values of the tax basis $(k / z)$. Indeed, we observe that the Buy region is limited from the left side by the wash-sales region which is part of the Sell region, exactly according to the statement of Proposition 3.5.

We also observe that, for small values of the $k$ variable, the no-transaction region NT contains the Merton optimal portfolio proportions $\bar{\pi}(\sigma, \theta)$ and $\bar{\pi}\left(\tilde{\sigma}^{\alpha}, \tilde{\theta}^{\alpha}\right)$ corresponding, respectively, to our financial market and to the fictitious financial market with modified parameters.

Optimal consumption strategy under taxes. Figures 11 and 12 report the consumptionwealth ratio for $\alpha=.2$ and .36 . We notice that this ratio depends on the value of the basis as well as on proportion of wealth allocated to the risky asset. Moreover, in the presence of taxes, on each point of the grid this ratio is higher than Merton's optimal consumption-wealth ratio.

Appendix A. Proof of Proposition 3.4. In order to prove the continuity of $V$, we follow [5] by introducing the approximation $v^{\lambda}$ defined as the value function of the control problem

$$
\begin{equation*}
v^{\lambda}(s):=\sup _{\nu \in \mathcal{A}^{\lambda}(s)} J^{\lambda}(s, \nu), \quad \text { where } \quad J^{\lambda}(s, \nu):=\mathbb{E}\left[\int_{0}^{\infty} e^{-\beta t} U\left(C_{t}\right) d t\right] \tag{A.1}
\end{equation*}
$$

and $\mathcal{A}^{\lambda}(s)$ is the collection of all $\mathbb{F}$-adapted processes $\nu=(C, L, M)$, where $C$ satisfies (2.9),


Figure 11. Consumption for $\alpha=0.20$.


Figure 12. Consumption for $\alpha=0.36$.
$L$ and $M$ are nondecreasing and right continuous, $L_{0-}=M_{0-}=0$, the jumps of $M$ satisfy (2.6), and the process $S^{s, \nu}=\left(X^{s, \nu}, Y^{s, \nu}, K^{s, \nu}\right)$ defined by $S_{0-}^{s, \nu}=s$ and the dynamics (2.7), (2.8),

$$
d X_{t}=\left(r X_{t}-C_{t}\right) d t-(1+\lambda) d L_{t}+\left[(1-\alpha) Y_{t-}+\alpha K_{t-}\right] d M_{t}
$$

takes values in $\overline{\mathcal{S}}$.
The above control problem corresponds to an optimal consumption investment problem with capital gains taxes and proportional transaction cost $\lambda>0$ on purchased risky assets. Clearly,

$$
\begin{equation*}
v^{\lambda} \searrow V \quad \text { as } \quad \lambda \searrow 0 \tag{A.2}
\end{equation*}
$$

For later use, we recall the following results from [5].
Theorem A.1. For $\lambda \geq 0$, the function $v^{\lambda}$ is a constrained viscosity solution of

$$
\begin{equation*}
\min \left\{-\mathcal{L} v^{\lambda} ; \mathbf{g}_{\lambda}^{\mathbf{b}} \cdot D v^{\lambda} ; \mathbf{g}^{\mathbf{s}} \cdot D v^{\lambda}\right\}=0 \quad \text { on } \quad \overline{\mathcal{S}} \backslash \partial^{z} S \quad \text { and } \quad v^{\lambda}=0 \quad \text { on } \quad \partial^{z} \mathcal{S} \tag{A.3}
\end{equation*}
$$

Theorem A.2. For $\lambda>0$, let $u$ be an upper semicontinuous viscosity subsolution of (A.3) and $v$ be a lower semicontinuous viscosity supersolution of (A.3), with $(u-v)^{+} \in \operatorname{USC}_{p}(\overline{\mathcal{S}})$. Assume further that $(u-v)(x, 0,0) \leq 0$ for all $x \geq 0$. Then $u \leq v$ on $\overline{\mathcal{S}}$.

We first need to prove the continuity of $v^{\lambda}$.
Lemma A.1. The function $v^{\lambda}$ is continuous on $\overline{\mathcal{S}}$.
Proof. By Proposition 4.1, the semicontinuous envelopes $v^{\lambda *}$ and $v_{*}^{\lambda}$ satisfy the polynomial growth condition $\left(v^{\lambda *}-v_{*}^{\lambda}\right)^{+} \in \operatorname{USC}_{p}(\overline{\mathcal{S}})$. We also know from Theorem A. 1 that they are, respectively, a constrained subsolution and supersolution of (5.1). We now claim that

$$
\begin{equation*}
\left(v_{*}^{\lambda}-v^{\lambda *}\right)(x, 0,0)=0 \quad \text { for all } \quad x \geq 0 \tag{A.4}
\end{equation*}
$$

so that $v_{*}^{\lambda} \geq v^{\lambda *}$ by the comparison result of Theorem A.2, and therefore $v_{*}^{\lambda}=v^{\lambda *}$ since the reverse inequality holds by definition.

It remains to prove (A.4). Notice that for all $s=(x, y, k) \in \overline{\mathcal{S}}$ and $z:=x+(1-\alpha) y+\alpha k$

$$
\begin{equation*}
v^{\lambda}(z, 0,0) \leq v^{\lambda}(s) \leq v^{\lambda}(z+y, 0,0) \tag{A.5}
\end{equation*}
$$

Before proving these inequalities, let us complete the proof of $v_{*}^{\lambda}=v^{\lambda *}$ on $\{(x, 0,0): x \geq 0\}$. For an arbitrary $x \in \mathbb{R}_{+}$, let $\left\{s_{n}=\left(x_{n}, y_{n}, k_{n}\right), n \geq 1\right\}$, $\left\{s_{n}^{\prime}=\left(x_{n}^{\prime}, y_{n}^{\prime}, k_{n}^{\prime}\right), n \geq 1\right\}$ be two sequences in $\overline{\mathcal{S}}$ such that

$$
s_{n}, s_{n}^{\prime} \underset{n \rightarrow \infty}{\longrightarrow}(x, 0,0), \quad v^{\lambda}\left(s_{n}\right) \xrightarrow[n \rightarrow \infty]{\longrightarrow} v_{*}^{\lambda}(x, 0,0), \quad \text { and } \quad v^{\lambda}\left(s_{n}^{\prime}\right) \xrightarrow[n \rightarrow \infty]{\longrightarrow} v^{\lambda *}(x, 0,0)
$$

By (A.5), together with the homotheticity property of Proposition 3.3, we see that

$$
\begin{aligned}
v^{\lambda}\left(s_{n}^{\prime}\right) \leq v^{\lambda}\left(z_{n}^{\prime}+y_{n}^{\prime}, 0,0\right) & =\left(z_{n}^{\prime}+y_{n}^{\prime}\right)^{p} v^{\lambda}(1,0,0) \\
v^{\lambda}\left(s_{n}\right) \geq v^{\lambda}\left(z_{n}, 0,0\right) & =\left(z_{n}\right)^{p} v^{\lambda}(1,0,0)
\end{aligned}
$$

where $z_{n}=x_{n}+(1-\alpha) y_{n}+\alpha k_{n}$ and $z_{n}^{\prime}=x_{n}^{\prime}+(1-\alpha) y_{n}^{\prime}+\alpha k_{n}^{\prime}$. Letting $n \rightarrow \infty$ in the above inequalities and recalling that $z_{n}, z_{n}^{\prime}+y_{n}^{\prime} \rightarrow x$, we get the required result.

We now turn to the proof of (A.5).

- The left-hand side of (A.5) holds since for each consumption-investment strategy $\nu=$ $(C, L, M) \in \mathcal{A}(z, 0,0)$ the strategy $\bar{\nu}:=\nu+\left\{1-\Delta M_{0}\right\}(0,0,1) \in \mathcal{A}(s)$.
- The right-hand side of (A.5) holds since for each $\nu=(C, L, M) \in \mathcal{A}(s)$ the strategy $\bar{\nu}:=\nu+\left\{y\left(1-\Delta M_{0}\right)\right\}(0,1,0) \in \mathcal{A}(\bar{s})$, where $\bar{s}:=(z+y, 0,0)$.
Indeed, since $\nu$ and $\bar{\nu}$ differ only by the jump at time $t=0$, the dynamics of the state processes $S^{s, \nu}$ and $S^{\bar{s}, \bar{\nu}}$ are such that $Y_{0}^{s, \nu}=Y_{0}^{\bar{s}, \bar{\nu}}$ and therefore $Y_{t}^{s, \nu}=Y_{t}^{\bar{s}, \bar{\nu}}$ for $t \geq 0$, and

$$
\begin{aligned}
K_{t}^{\bar{s}, \bar{\nu}}-K_{t}^{s, \nu} & =\left(K_{0}^{\bar{s}, \bar{\nu}}-K_{0}^{s, \nu}\right) e^{-M_{t}^{c}} \prod_{0 \leq s \leq t}\left(1-\Delta M_{s}\right) \\
& \leq\left(K_{0}^{\bar{s}, \bar{\nu}}-K_{0}^{s, \nu}\right)^{+}=(y-k)^{+}\left(1-\Delta M_{0}\right) .
\end{aligned}
$$

Then the corresponding liquidation value processes $Z^{s, \nu}$ and $Z^{\bar{s}, \bar{\nu}}$ are such that

$$
\begin{aligned}
Z_{t}^{\bar{s}, \bar{\nu}}-Z_{t}^{s, \nu} & =e^{r t}\left\{Z_{0}^{\bar{s}, \bar{\nu}}-Z_{0}^{s, \nu}-\alpha \int_{0}^{t} e^{-r s}\left(K^{s, \nu}-K^{\bar{s}, \bar{\nu}}\right)_{s} d s\right\} \\
& \geq e^{r t}\left\{Z_{0}^{\bar{s}, \bar{\nu}}-Z_{0}^{s, \nu}-\left(K_{0}^{s, \nu}-K_{0}^{\bar{s}, \bar{\nu}}\right)^{+}\right\} \\
& =e^{r t}\left\{y-(y-k)^{+}\left(1-\Delta M_{0}\right)\right\} \geq 0
\end{aligned}
$$

It follows that $Z^{\bar{s}, \nu} \geq 0$; hence $\bar{\nu} \in \mathcal{A}(\bar{s})$.
Proof of Proposition 3.4. Since $\left(v^{\lambda}\right)_{\lambda>0}$ is a nonincreasing sequence of continuous functions (Lemma A.1) converging to $V$ as $\lambda \searrow 0$, it follows that the function $V$ is lower semicontinuous.

Let $\mathcal{V}$ be the lower semicontinuous function defined on $\mathbb{R}_{+}^{2}$ by

$$
\begin{equation*}
\mathcal{V}(\xi, \zeta):=V(1-(1-\alpha) \xi+\alpha \zeta, \xi, \zeta) \tag{A.6}
\end{equation*}
$$

so that

$$
\begin{equation*}
V(x, y, k)=z^{p} \mathcal{V}\left(\frac{y}{z}, \frac{k}{z}\right), \quad \text { with } \quad z=x+(1-\alpha) y+\alpha k, \tag{A.7}
\end{equation*}
$$

by the homotheticity property of $V$ stated in Proposition 3.3. By Theorem A.1, we have $\mathbf{g}^{\mathbf{b}} \cdot D V \geq 0$ and $\mathbf{g}^{\mathbf{s}} \cdot D V \geq 0$ in the viscosity sense. By a direct change of variables, this implies that $\mathcal{V}$ is a lower semicontinuous viscosity supersolution of the equation

$$
p \mathcal{V}-\left(\xi \mathcal{V}_{\xi}+\zeta \mathcal{V}_{\zeta}\right)-\varepsilon^{-1}\left(\mathcal{V}_{\xi}+\mathcal{V}_{\zeta}\right) \geq 0 \text { and } \xi \mathcal{V}_{\xi}+\zeta \mathcal{V}_{\zeta} \geq 0 .
$$

Also, from the monotonicity of $V$ in $x, y$, and $k$, it follows that $\mathcal{V}$ is a lower semicontinuous viscosity supersolution of the equation

$$
p \mathcal{V}-\left(\xi \mathcal{V}_{\xi}+\zeta \mathcal{V}_{\zeta}\right)+\min \left\{0, \frac{1}{1-\alpha} \mathcal{V}_{\xi}, \frac{1}{\alpha} \mathcal{V}_{\zeta}\right\} \geq 0
$$

Observe that $\mathcal{V}$ is bounded as a consequence of the upper bound provided in Proposition 4.1. We then deduce from the above viscosity supersolution properties that $-|\nabla \mathcal{V}| \geq-A$ on $(0, \infty)^{2}$, in the viscosity sense, for some constant $A$. Hence $\mathcal{V}$ is Lipschitz continuous.

## Appendix B. Proof of Proposition 4.2.

Preliminaries and notation. For $s \in \overline{\mathcal{S}}$ and $\nu \in \mathcal{A}(s)$, the process $Z^{s, \nu}$ is defined by the initial condition $Z_{0}^{s, \nu}=z:=x+(1-\alpha) y+\alpha k$ and the dynamics

$$
d Z_{t}^{s, \nu}=\left(r Z_{t}^{s, \nu}-C_{t}\right) d t+Y_{t}^{s, \nu} \tilde{\sigma}^{\alpha}\left(\tilde{\theta}^{\alpha} d t+d W_{t}\right)+r \alpha Y_{t}^{s, \nu}\left(1-\frac{B_{t}^{s, \nu}}{P_{t}}\right) d t
$$

where

$$
\tilde{\sigma}^{\alpha}:=(1-\alpha) \sigma \quad \text { and } \quad \tilde{\theta}^{\alpha}:=\theta-\frac{r \alpha}{\tilde{\sigma}^{\alpha}} .
$$

Our purpose is to show that the value function $V$ outperforms the maximal utility achieved in a frictionless financial market consisting of one bank account with the constant interest rate $r$ and one risky asset with price process $P^{\alpha}$ given by

$$
d P_{t}^{\alpha}=P_{t}^{\alpha}\left[r d t+\tilde{\sigma}^{\alpha}\left(\tilde{\theta}^{\alpha} d t+d W_{t}\right)\right] \quad \text { and } \quad \tilde{P}_{0}^{\alpha}=P_{0}
$$

From Theorem 2.1, the solution of the optimal consumption-investment problem with price process $P^{\alpha}$ is given by the constant controls $\bar{c}^{\alpha}:=\bar{c}\left(r, \tilde{\theta}^{\alpha}\right)$ and $\bar{\pi}^{\alpha}:=\bar{\pi}\left(\tilde{\sigma}^{\alpha}, \tilde{\theta}^{\alpha}\right)$ and the corresponding optimal wealth process is defined by

$$
\bar{Z}^{\alpha}=z \quad \text { and } \quad d \bar{Z}_{t}^{\alpha}=\bar{Z}^{\alpha}\left[\left(r-\bar{c}^{\alpha}\right) d t+\bar{\pi}^{\alpha} \tilde{\sigma}^{\alpha}\left(\tilde{\theta}^{\alpha} d t+d W_{t}\right)\right] .
$$

In order to prove the required result, we shall fix an arbitrary maturity $T>0$ and construct a sequence of admissible strategies $\hat{\nu}^{T, n}$ such that

$$
V(s) \geq \lim _{n \rightarrow \infty} J_{T}\left(s, \hat{\nu}^{T, n}\right)=\mathbb{E}\left[\int_{0}^{T} e^{-\beta t} U\left(\bar{c} \bar{Z}_{t}^{\alpha}\right) d t\right]
$$

Then, the required result follows by sending $T$ to infinity in this inequality.
A sequence of strategies tracking the Merton optimal policy. Let $T>0$ be a fixed maturity. We construct a sequence of consumption-investment strategies $\hat{\nu}^{T, n, k}$ by forcing the tax basis $B$ to be close to the spot price and by tracking Merton's optimal strategy, i.e., keeping the proportion of wealth invested in the risky asset and the proportion of wealth dedicated for consumption

$$
\pi_{t}:=\frac{Y_{t}}{Z_{t}} \mathbf{1}_{\left\{Z_{t} \neq 0\right\}} \quad \text { and } \quad c_{t}:=\frac{C_{t}}{Z_{t}} \mathbf{1}_{\left\{Z_{t} \neq 0\right\}}, \quad 0 \leq t \leq T,
$$

close to the pair $\left(\bar{\pi}^{\alpha}, \bar{c}^{\alpha}\right)$.

To do this, we define a convenient sequence $\left(\nu^{T, n, m}\right)_{n, m \geq 1}:=\left(C^{T, n, m}, L^{T, n, m}, M^{T, n, m}\right)_{n, m \geq 1}$ for all $s=(x, y, k) \in \overline{\mathcal{S}}$. We shall denote by

$$
\left(X^{T, n, m}, Y^{T, n, m}, K^{T, n, m}\right)=\left(X^{\nu^{T, n, m}}, Y^{\nu^{T, n, m}}, K^{\nu^{T, n, m}}\right)
$$

the corresponding state processes and by $Z^{T, n, m}=Z^{\nu^{T, n, m}}$ the corresponding after-tax liquidation value process. For all integers $n \geq 1$ and $m \geq 1$, the consumption-investment strategy $\nu^{T, n, m}$ is defined as follows:

1. The consumption strategy is defined by $C_{t}^{T, n, m}:=\bar{c}^{\alpha} Z_{t}^{T, n, m}$ for $0 \leq t \leq T$. The investment strategy is piecewise constant:

$$
d L_{t}^{T, n, m}=d M_{t}^{T, n, m}=0 \quad \text { for all } \quad t \in(0, T] \backslash\left\{\tau_{j}^{T, n, m}, j \leq m\right\}
$$

where the sequence of stopping times $\left(\tau_{j}^{T, n, m}\right)_{j \leq m}$ is defined in step 3 .
2. At time 0 , set $\Delta L_{0}^{T, n, m}:=\bar{\pi}^{\alpha} z$ and $\Delta M_{0}^{T, n, m}:=1$, so that

$$
K_{0}^{T, n, m}=Y_{0}^{T, n, m}, \quad \pi_{0}^{T, n, m}=\bar{\pi}^{\alpha}, \quad \text { and } \quad Z_{0}^{T, n, m}=z .
$$

3. We now introduce the sequence of stopping times $\tau_{j}^{T, n, m}$ as the hitting times of the pair process $\left(\pi^{T, n, m}, \frac{K^{T, n, m}}{Y^{T, n, m}}\right)$ of some barrier close to $(\bar{\pi}, 1)$. Set

$$
\tau_{0}^{T, n, m}:=0 \quad \text { and } \quad \tau_{j}^{T, n, m}:=T \wedge \tau_{j}^{\pi} \wedge \tau_{j}^{B} \quad \text { for } \quad 1 \leq j \leq m,
$$

where

$$
\begin{aligned}
\tau_{j}^{\pi} & :=\inf \left\{t \geq \tau_{j-1}^{T, n, m}:\left|\pi_{t}^{T, n, m}-\bar{\pi}^{\alpha}\right|>n^{-1} \bar{\pi}^{\alpha}\right\}, \\
\tau_{j}^{B} & :=\inf \left\{t \geq \tau_{j-1}^{T, n, m}:\left|1-\frac{K^{T, n, m}}{Y^{T, n, m}}\right|>n^{-1}\right\} .
\end{aligned}
$$

4. Finally, we specify the jumps $\left(\Delta L^{T, n, m}, \Delta M^{T, n, m}\right)$ at each time $\tau_{j}^{T, n, m}, j \geq 1$, by

$$
\Delta L_{t}^{T, n, m}:=\bar{\pi}^{\alpha} Z_{t}^{T, n, m} \quad \text { and } \quad \Delta M_{t}^{T, n, m}:=1 \quad \text { for } \quad t \in\left\{\tau_{j}^{T, n, m}, j<m\right\}
$$

so that

$$
\pi_{t}^{T, n, m}=\bar{\pi}^{\alpha} \quad \text { and } \quad B_{t}^{T, n, m}=P_{t} \quad \text { for } \quad t \in\left\{\tau_{j}^{T, n, m}, j<m\right\} .
$$

And at time $t=\tau_{m}^{T, n, m}$, set $\Delta L_{t}^{T, n, m}:=0$ and $\Delta M_{t}^{T, n, m}=1$, so that all the wealth is transferred to the bank:

$$
Y_{t}^{T, n, m}=K_{t}^{T, n, m}=0 \quad \text { and } \quad X_{t}^{T, n, m}=Z_{\tau_{m}^{T, n, m}}^{T, n, m} e^{r\left(T-\tau_{m}^{T, n, m}\right)} \quad \text { for } \quad \tau_{m}^{T, n, m} \leq t \leq T
$$

Lemma B.1. For each $n \geq 1$, the sequence $\left(\tau_{m}^{T, n, m}\right)_{m \geq 0}$ converges to $T \mathbb{P}$-a.s.
Proof. Let $n \geq 1$. In order to alleviate the notation, we shall denote $\tau_{m}:=\tau_{m}^{T, n, m}$, $\pi:=\pi^{T, n, m}$, and $(Y, Z, K):=\left(Y^{T, n, m}, Z^{T, n, m}, K^{T, n, m}\right)$.

1. In a first step, we shall prove that $\tau_{m}^{\pi} \rightarrow \infty$ a.s. For sufficiently large $n$, observe that our sequence of stopping times $\left(\tau_{m}^{\pi}\right)_{m \geq 1}$ is equivalently defined by

$$
\tau_{m+1}^{\pi}:=\inf \left\{t \geq \tau_{m}:\left|\tilde{\pi}_{t}-\bar{\pi}^{\alpha}\right| \geq \bar{\pi}^{\alpha} / n\right\}
$$

where $\tilde{\pi}$ is defined by

$$
\tilde{\pi}_{t}=\pi_{t \wedge \theta}, \quad \text { with } \quad \theta:=\inf \left\{t \geq \tau_{m}:\left|\pi_{t}-\bar{\pi}^{\alpha}\right| \geq \frac{\bar{\pi}^{\alpha}}{2} \text { and }\left|\frac{K_{t}}{Y_{t}}-1\right| \geq \frac{1}{2}\right\} .
$$

Notice that $d \tilde{\pi}_{t}=f(u) d u+g(u) d W_{u}, t \geq \tau_{m}$, with uniformly bounded processes $f$ and $g$. We then estimate

$$
\begin{aligned}
\mathbb{P}\left[\tau_{m+1}^{\pi} \leq \tau_{m}+\frac{1}{m}\right] & \leq \mathbb{P}\left[\sup _{\tau_{m} \leq t \leq \tau_{m}+\frac{1}{m}}\left|\tilde{\pi}_{t}-\bar{\pi}^{\alpha}\right| \geq \bar{\pi}^{\alpha} n^{-1}\right] \\
& \leq\left(\frac{n}{\bar{\pi}^{\alpha}}\right)^{4} \mathbb{E}\left[\left(\sup _{\tau_{m} \leq t \leq \tau_{m}+\frac{1}{m}}\left|\tilde{\pi}_{t}-\bar{\pi}^{\alpha}\right|\right)^{4}\right] \\
& \leq C\left(\frac{1}{m^{4}}+\frac{1}{m^{2}}\right) \text { for some positive constant } C
\end{aligned}
$$

where we used the Chebyshev and the Burkholder-Davis-Gundy inequalities.
2. Following the same reasoning, we prove that $\mathbb{P}\left[\tau_{m+1}^{B} \leq \tau_{m}+\frac{1}{m}\right] \leq C\left(\frac{1}{m^{4}}+\frac{1}{m^{2}}\right)$, and we conclude that $\sum_{m<\infty} \mathbb{P}\left[\tau_{m+1} \leq \tau_{m}+\frac{1}{m}\right]<\infty$. Then, by the Borel-Cantelli lemma, $\mathbb{P}\left[\limsup _{m}\left\{\tau_{m+1} \leq \tau_{m}+\frac{1}{m}\right\}\right]<0$, which implies that $\lim _{m \rightarrow \infty} \tau_{m}=\infty$.

Lemma B.2. For each integer $n$, we have $\nu^{T, n, m} \in \mathcal{A}(s)$.
Proof. By (4.1), we have

$$
d Z_{t}^{T, n, m}=Z_{t}^{T, n, m}\left[\left(r-\bar{c}^{\alpha}\right) d t+\pi_{t}^{T, \nu, m} \tilde{\sigma}^{\alpha}\left(\tilde{\theta}^{\alpha} d t+d W_{t}\right)+r \alpha \pi_{t}^{T, n, m}\left(1-B_{t}^{T, n}\right) d t\right] .
$$

Also, $0<\left(1-n^{-1}\right) \bar{\pi}^{\alpha} \leq \pi_{t}^{T, n, m} \leq\left(1+n^{-1}\right) \bar{\pi}^{\alpha}$. In particular, the process $\pi^{T, n, m}$ is bounded, so that the above dynamics implies that the process $Z^{T, n, m}$ is positive, and $Y_{t}^{T, n, m}=$ $\pi_{t}^{T, n, m} Z_{t}^{T, n, m}>0 \mathbb{P}$-a.s.

## The convergence result.

Lemma B.3. There is a constant $A$ depending on $T$ such that

$$
\mathbb{E}\left[\sup _{0 \leq t \leq T}\left|Z_{t}^{T, n, m}-\bar{Z}_{t}^{\alpha}\right|^{2}\right] \leq\left(n^{-2}+\mathbb{E}\left[T-\tau_{m}^{T, n, m}\right]\right) A e^{A T}
$$

Proof. By definition of the sequence of consumption-investment strategies $\left(\nu^{T, n, m}\right)$, we have

$$
\begin{equation*}
\sup _{0 \leq t \leq T \wedge \tau^{T, n, m}}\left|\pi_{t}^{T, n}-\bar{\pi}^{\alpha}\right| \leq \frac{1}{n} \bar{\pi}^{\alpha} \quad \text { and } \quad \sup _{0 \leq t \leq T \wedge \tau^{T, n, m}}\left|1-\frac{K_{t}^{T, n, m}}{Y_{t}^{T, n, m}}\right| \leq \frac{1}{n} . \tag{B.1}
\end{equation*}
$$

By direct computation, we decompose the difference $Z^{T, n, m}-\bar{Z}^{\alpha}$ into

$$
D_{t}:=Z_{t}^{T, n, m}-\bar{Z}_{t}^{\alpha}=F_{t}+G_{t}+H_{t},
$$

where

$$
\begin{aligned}
F_{t} & :=\int_{0}^{t} D_{t}\left[\left(r-\bar{c}^{\alpha}\right) d u+\pi_{u}^{T, n, m}\left(\tilde{\sigma}^{\alpha} \tilde{\theta}^{\alpha} d u+\alpha r\left(1-\frac{K_{u}^{T, n, m}}{Y_{u}^{T, n, m}}\right) d u+\tilde{\sigma}^{\alpha} d W_{u}\right)\right] \\
G_{t} & :=\int_{0}^{t} \bar{Z}_{u}^{\alpha} \tilde{\sigma}^{\alpha}\left(\pi_{t}^{T, n, m}-\bar{\pi}^{\alpha}\right)\left(\theta^{\alpha} d u+d W_{u}\right) \\
H_{t} & :=\alpha r \int_{0}^{t} \pi_{u}^{T, n, m} \bar{Z}_{u}^{\alpha}\left(1-\frac{K_{u}^{T, n, m}}{Y_{u}^{T, n, m}}\right) d u
\end{aligned}
$$

In the subsequent calculation, $A$ will denote a generic ( $T$-dependent) constant whose value may change from line to line. We shall also denote $V_{t}^{*}:=\sup _{0 \leq u \leq t}\left|V_{u}\right|$ for all processes $\left(V_{t}\right)_{t}$.

We first start by estimating the first component $F$. Observe that the process $\pi^{T, n, m}$ is bounded by $2 \bar{\pi}$. Then

$$
\left|F_{t}\right|^{2} \leq A \int_{0}^{t}\left|D_{u}^{*}\right|^{2} d u+2\left(\int_{0}^{t} D_{u} \pi_{u}^{t, n, m} \tilde{\kappa} d W_{u}\right)^{2}
$$

By the Burkholder-Davis-Gundy inequality, this provides

$$
\mathbb{E}\left|F_{t}^{*}\right|^{2} \leq A \int_{0}^{t} \mathbb{E}\left|D_{u}^{*}\right|^{2} d u
$$

By a similar calculation, it follows from the boundedness of $\pi^{T, n, m}$ and from (B.1) that

$$
\mathbb{E}\left|G_{t}^{*}\right|^{2} \leq A\left(n^{-2}+\mathbb{E}\left[T-\tau_{m}^{T, n, m}\right]\right) \quad \text { and } \quad \mathbb{E}\left|H_{t}^{*}\right|^{2} \leq A\left(n^{-2}+\mathbb{E}\left[T-\tau_{m}^{T, n, m}\right]\right)
$$

Collecting the above estimates, we see that

$$
E\left|D_{t}^{*}\right|^{2} \leq A\left(n^{-2}+\mathbb{E}\left[T-\tau_{m}^{T, n, m}\right]\right)+K \int_{0}^{t} \mathbb{E}\left|D_{u}^{*}\right|^{2} d u \quad \text { for all } \quad t \leq T
$$

and we obtain the required result by the Gronwall inequality.
B.1. Proof of Proposition 4.2. For $s=(x, y, k) \in \overline{\mathcal{S}}$, and $T>0$,

$$
\begin{aligned}
\left|J_{T}\left(s, \nu^{T, n, m}\right)-\int_{0}^{T} e^{-\beta t} U\left(\bar{c}^{\alpha} \bar{Z}_{t}^{\alpha}\right) d t\right| & =\left|\int_{0}^{T} e^{-\beta t}\left(U\left(\bar{c}^{\alpha} Z_{t}^{T, n, m}\right)-U\left(\bar{c}^{\alpha} \bar{Z}_{t}^{\alpha}\right) d t\right)\right| \\
& \leq A \int_{0}^{T} e^{-\beta t}\left|Z_{t}^{T, n, m}-\bar{Z}_{t}^{\alpha}\right|^{p} d t
\end{aligned}
$$

for some positive constant $A$. Now, by Lemma B. 1 and the estimate of Lemma B.3, it follows that

$$
\lim _{n, m \rightarrow \infty} J_{T}\left(s, \nu^{T, n, m}\right)=\int_{0}^{T} e^{-\beta t} U\left(\bar{c}^{\alpha} \bar{Z}_{t}^{\alpha}\right) d t
$$

Since $V(s) \geq J_{T}\left(s, \nu^{T, n, m}\right)$ for every $T>0$, this implies that

$$
V(s) \geq \lim _{T \rightarrow \infty} \int_{0}^{T} e^{-\beta t} U\left(\bar{c}^{\alpha} \bar{Z}_{t}^{\alpha}\right) d t=\gamma\left(r, \tilde{\theta}^{\alpha}\right) \frac{z^{p}}{p}
$$

Acknowledgments. We are grateful to Stathis Tompaidis for numerous discussions on the modeling issue of this paper. We have also benefited from interesting comments and discussions with Victor DeMiguel, Chris Rogers, Jose Scheinkman, and Raman Uppal.

## References

[1] M. Akian, J. L. Menaldi, and A. Sulem, On an investment-consumption model with transaction costs, SIAM J. Control Optim., 34 (1996), pp. 329-364.
[2] G. Barles and E. R. Jakobsen, On the convergence rate of approximation schemes for Hamilton-Jacobi-Bellman equations, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 33-54.
[3] G. Barles and H. M. Soner, Option pricing with transaction costs and a nonlinear Black-Scholes equation, Finance Stoch., 2 (1998), pp. 369-397.
[4] G. Barles and P. E. Souganidis, Convergence of approximation schemes for fully nonlinear equations, Asymptotic Anal., 4 (1991), pp. 271-283.
[5] I. Ben Tahar, H. M. Soner, and N. Touzi, The dynamic programming equation for the problem of optimal investment under capital gains taxes, SIAM J. Control Optim., 46 (2007), pp. 1779-1801.
[6] J. F. Bonnans and H. Zidani, Consistency of generalized finite difference schemes for the stochastic HJB equation, SIAM J. Numer. Anal., 41 (2003), pp. 1008-1021.
[7] A. Cadenillas and S. Pliska, Optimal trading of a security when there are taxes and transaction costs, Finance Stoch., 3 (1999), pp. 137-165.
[8] G. M. Constantinides, Capital market equilibrium with personal taxes, Econometrica, 51 (1983), pp. 611-636.
[9] G. M. Constantinides and M. J. P. Magill, Portfolio selection with transaction costs, J. Econom. Theory, 13 (1976), pp. 245-263.
[10] J. Cox and C. F. Huang, Optimal consumption and portfolio policies when asset prices follow a diffusion process, J. Econom. Theory, 49 (1989), pp. 33-83.
[11] J. Cvitanić and I. Karatzas, Convex duality in constrained portfolio optimization, Ann. Appl. Probab., 2 (1992), pp. 767-818.
[12] R. M. Dammon, C. S. Spatt, and H. H. Zhang, Optimal consumption and investment with capital gains taxes, Rev. Financ. Stud., 14 (2001), pp. 583-616.
[13] M. H. A. Davis and A. R. Norman, Portfolio selection with transaction costs, Math. Oper. Res., 15 (1990), pp. 676-713.
[14] V. DeMiguel and R. Uppal, Portfolio investment with the exact tax basis via nonlinear programming, Management Sci., 51 (2005), pp. 277-290.
[15] D. Duffie and T. Sun, Transaction costs and portfolio choice in a discrete-continuous-time setting, J. Econom. Dynam. Control, 14 (1990), pp. 35-51.
[16] P. Dybvig and H. Koo, Investment with Taxes, Working paper, Washington University, St. Louis, MO, 1996.
[17] A. Fahim, N. Touzi, and X. Warin, A Probabilistic Numerical Method for Fully Nonlinear Parabolic PDEs, preprint, 2009.
[18] M. Gallmeyer, R. Kaniel, and S. Tompaidis, Tax management strategies with multiple risky assets, J. Finan. Econ., 80 (2006), pp. 243-291.
[19] E. Jouini, P.-F. Koehl, and N. Touzi, Optimal investment with taxes: An existence result, J. Math. Econom., 33 (2000), pp. 373-388.
[20] E. Jouini, P.-F. Koehl, and N. Touzi, Optimal investment with taxes: An optimal control problem with endogenous delay, Nonlinear Anal., 37 (1999), pp. 31-56.
[21] I. Karatzas, J. P. Lehoczky, and S. E. Shreve, Optimal portfolio and consumption decisions for a "small investor" on a finite horizon, SIAM J. Control Optim., 25 (1987), pp. 1557-1586.
[22] N. V. Krylov, On the rate of convergence of finite-difference approximations for Bellman's equations, St. Petersburg Math. J., 9 (1997), pp. 245-256.
[23] N. V. Krylov, On the rate of convergence of finite-difference approximations for Bellman's equations with Lipschitz coefficients, Appl. Math. Optim., 3 (2005), pp. 365-399.
[24] H. E. Leland, Optimal Portfolio Management with Transaction Costs and Capital Gains Taxes, preprint, 1999.
[25] R. Mehra and E. Prescott, The equity premium: A puzzle, J. Monet. Econ., 15 (1985), pp. 145-161.
[26] R. C. Merton, Lifetime portfolio selection under uncertainty: The continuous-time model, Rev. Econ. Statist., 51 (1969), pp. 247-257.
[27] R. C. Merton, Optimum consumption and portfolio rules in a continuous-time model, J. Econom. Theory, 3 (1971), pp. 373-413.
[28] S. E. Shreve and H. M. Soner, Optimal investment and consumption with transaction costs, Ann. Appl. Probab., 4 (1994), pp. 609-692.
[29] H. M. Soner, Optimal control with state-space constraint I, SIAM J. Control Optim., 24 (1986), pp. 552-561.
[30] H. M. Soner, Optimal control with state-space constraint II, SIAM J. Control Optim., 24 (1986), pp. 1110-1122.
[31] A. Tourin and T. Zariphopoulou, Numerical schemes for investment models with singular transactions, Comput. Econom., 7 (1994), pp. 287-307.

# Multivariate Extension of Put-Call Symmetry* 

llya Molchanov ${ }^{\dagger}$ and Michael Schmutz ${ }^{\dagger}$


#### Abstract

Multivariate analogues of the put-call symmetry can be expressed as certain symmetry properties of basket options and options on the maximum of several assets with respect to some (or all) permutations of the weights and the strike. The so-called self-dual distributions satisfying these symmetry conditions are completely characterized and their properties explored. It is also shown how to relate some multivariate asymmetric distributions to symmetric ones by a power transformation that is useful to adjust for carrying costs. Particular attention is devoted to the case of asset prices driven by Lévy processes. Based on this, semistatic hedging techniques for multiasset barrier options are suggested.


Key words. barrier option, dual market, Lévy process, multiasset option, put-call symmetry, self-dual distribution, semistatic hedging

AMS subject classifications. 60E05, 60G51, 91B28, 91B70
DOI. 10.1137/090754194

1. Introduction. Consider European options on $S_{T}=S_{0} e^{r T} \eta$ being the price of a (say nondividend paying) asset at the maturity time $T$, where $S_{0}$ is the spot price and $e^{r T} \eta$ is the factor by which the price changes, $r$ is the (constant) risk-free interest rate, and $\eta$ is an almost surely positive random variable. In arbitrage-free and complete markets, the option price equals the discounted expected payoff, where the expectation can be taken with respect to the unique equivalent martingale measure. In this case $\mathbf{E} \eta=1$ and the discounted price process $\left(S_{t} e^{-r t}\right)_{t \in[0, T]}$ becomes a martingale. Unless indicated by a different subscript, all expectations in this paper are understood with respect to the probability measure $\mathbf{Q}$, which is not necessarily a martingale measure. In this paper we do not address the choice of a martingale measure in incomplete markets.

One of the most basic relationships between options in arbitrage-free markets is the European call-put parity. This parity can be expressed by

$$
\begin{equation*}
c(k, F)=e^{-r T}(F-k)+p(k, F), \tag{1.1}
\end{equation*}
$$

where $c(k, F)$ (resp., $p(k, F)$ ) denotes the price of the European call (resp., put) option with strike $k$ on the asset with forward price $F$. Deterministic dividends or income until maturity can be incorporated into the forward price, e.g., by setting $F=S_{0} e^{(r-q) T}$ in the case of a continuous dividend yield $q$. The maturity time $T$ is supposed to be the same for all instruments.

[^34]Recall that $\eta$ defined from $S_{T}=F \eta$ is almost surely positive. If $\eta$ is distributed according to a martingale measure $\mathbf{Q}$, then a new probability measure $\tilde{\mathbf{Q}}$ can be defined from

$$
\frac{d \tilde{\mathbf{Q}}}{d \mathbf{Q}}=\eta
$$

Since $\eta$ is usually represented as $e^{H_{T}}$ for a semimartingale $\left(H_{t}\right)_{t \in[0, T]}$ and $\left(-H_{t}\right)_{t \in[0, T]}$ is called the dual to $\left(H_{t}\right)_{t \in[0, T]}$, the random variable $\tilde{\eta}=\eta^{-1}$ is said to be the dual of $\eta$.

The duality principle in option pricing deals with this change of measure and traces its roots to $[3,7,20,26,27]$. This principle has been studied extensively in recent years; e.g., [8] and [15] discuss the American version of duality. For a detailed presentation of the duality principle in a general exponential semimartingale setting and for its various applications, see [17] and the references therein. The multiasset case has been studied in [16].

If the distribution of $\eta$ under $\mathbf{Q}$ coincides with the distribution of $\tilde{\eta}$ under $\tilde{\mathbf{Q}}$, we call the positive random variable $\eta$ self-dual. This symmetry property clearly depends on the probability measure used to define the expectation. The following result is well known.

Theorem 1.1. Assume that the asset price at maturity of an asset with deterministic dividend payments is $S_{T}=F \eta$ with self-dual $\eta$. Then

$$
\begin{equation*}
p(k, F)=c(F, k) . \tag{1.2}
\end{equation*}
$$

For instance, the log-normal distribution in the risk-neutral setting is self-dual, which implies the put-call symmetry (or the Bates rule) for the Black-Scholes economy. The above relation and its variants obtained by combining it with the put-call parity are known in the literature as put-call symmetry; see, e.g., [4, 5, 7, 10] and more recently [19] for log-infinitely divisible models. Further recent developments are presented by Carr and Lee [12].

There are various applications of the put-call symmetry, especially in connection with hedging exotic options. Following [12], semistatic hedging is the replication of contracts by trading European-style claims at no more than two times after inception. In the univariate case, $[9,10,12]$ and also $[1,2,31]$ developed a machinery for replicating barrier contracts having fundamental relevance for other path-dependent contracts.

Theorem 1.2 (cf. [12, Thm. 2.5]). An integrable random variable $\eta$ is self-dual if and only if, for any payoff function $f: \mathbb{R}_{+} \mapsto \mathbb{R}_{+}$and any $F>0$,

$$
\begin{equation*}
\mathbf{E} f(F \eta)=\mathbf{E}\left[f\left(F \eta^{-1}\right) \eta\right] . \tag{1.3}
\end{equation*}
$$

By (1.2) and writing the call payoff as

$$
\begin{equation*}
\left(S_{T}-k\right)_{+}=\max \left(S_{T}, k\right)-k, \quad S_{T}, k \geq 0 \tag{1.4}
\end{equation*}
$$

it is immediately seen that (1.2) is equivalent to

$$
\begin{equation*}
\mathbf{E} \max (F \eta, k)=\mathbf{E} \max (F, k \eta) \quad \text { for all } k \geq 0 \tag{1.5}
\end{equation*}
$$

This paper explores extensions of the put-call symmetry property for the multiasset case. Section 2 characterizes random vectors that possess symmetry properties generalizing the
classical put-call symmetry for basket options and options on the maximum of several assets. The symmetry (or self-duality) is understood with respect to each particular asset or for all assets simultaneously. It is also shown that the joint distribution of assets is completely determined by the prices of basket calls, options on the maximum of weighted asset prices, or calls and puts written on the weighted maximum or minimum of risky assets, complementing the result of [11] formulated in the absolutely continuous case for basket options, where it also yields an explicit formula for the probability density function. Relationships between the self-duality property and the swap invariance in Margrabe-type options have been studied in [29]. In currency markets, the self-duality results can be interpreted with respect to real existing markets, yielding the basis for further applications; see [36].

It is shown in section 3 that the joint self-duality implies that expected payoffs from basket options are symmetric with respect to the weights of all assets and the strike price. The new effect in the multivariate setting is that the independence of asset prices prevents them from being jointly self-dual. In other words, symmetry properties for several assets enforce a certain dependency structure between them, which is explored in this paper.

Section 4 deals with multivariate log-infinitely divisible distributions, exhibiting the multivariate put-call symmetry. In order to extend the application range of the self-duality property and also in view of incorporating the carrying costs, we define in section 5 quasi-self-dual random vectors and characterize their distributions. These random vectors become self-dual if their components are normalized by constants (representing carrying costs) and raised to a certain power. The related power transformation was used by Carr and Lee [12, sec. 6.2] in the one-dimensional case. Here we establish an explicit relationship between carrying costs and the required power of transformation for rather general price models based on Lévy processes.

These results are then used in section 6 to obtain several new results for self-dual random variables, thereby complementing the results from [12]. In particular, self-dual random variables have been characterized in terms of their distribution functions; it is shown that self-dual random variables always have nonnegative skewness, and several examples of self-dual random variables are given.

As in the univariate case, in the multiasset case there are also various applications of symmetry results. First, symmetry results may be used for validating models or analyzing market data, e.g., similarly as in [4, 18] for the univariate case. Furthermore, they could be used for deriving certain investment strategies; see, e.g., section 7.4. Probably the most important application will potentially be found in the area of hedging, especially in developing semistatic replicating strategies of multiasset barrier contracts and possibly more complicated path-dependent contracts. As far as the relevance of this application is concerned, we should mention that there has been a liquid market in structured products, particularly in Europe. At the moment the majority of the trades still occur over the counter, but more and more trades are also organized at exchanges, especially at the quite new European exchange for structured products, Scoach. Structured products often involve equity indices, sometimes several purpose-built shares, and quite often have barriers. Hence, developing robust hedging strategies for multiasset path-dependent products seems to be of a certain importance. The importance of developing robust hedging strategies for multiasset path-dependent products is also particularly stressed by Carr and Laurence [11]. Section 7 contains the first applications of the multivariate symmetry properties, especially for hedging complex barrier options, thereby
extending results from $[9,10,12]$ for some multiasset options. The development of a more general multivariate semistatic hedging machinery is left for future research.
2. Characterization of multivariate distributions with symmetry properties. When working with $n$ assets, we write $\eta$ for an $n$-dimensional random vector ( $\eta_{1}, \ldots, \eta_{n}$ ) such that the price $S_{T i}$ of the $i$ th asset at time $T$ equals $F_{i} \eta_{i}$ with $F_{i}$ being the corresponding forward price. We denote this in short by

$$
\begin{equation*}
S_{T}=F \circ \eta=\left(F_{1} \eta_{1}, \ldots, F_{n} \eta_{n}\right) . \tag{2.1}
\end{equation*}
$$

We often extend the $n$-dimensional random vector $\eta=\left(\eta_{1}, \ldots, \eta_{n}\right)$ by adding to it the first coordinate, which is 1 . In the financial setting this extra coordinate represents a riskless bond. Because of this, we number the coordinates of $(n+1)$-dimensional vectors as $0,1, \ldots, n$ and write these vectors as $\left(u_{0}, u\right)$ for $u_{0} \in \mathbb{R}$ and $u \in \mathbb{R}^{n}$ or as $\left(u_{0}, u_{1}, \ldots, u_{n}\right)$. The scalar product of $n$-dimensional vectors $u$ and $\eta$ is denoted by $\langle u, \eta\rangle$.

Let $S_{T}=\left(S_{T 1}, \ldots, S_{T n}\right)$ and integrable $\eta=\left(\eta_{1}, \ldots, \eta_{n}\right)$ be as defined in (2.1). Assume that all coordinates of $\eta$ are positive, i.e., $\eta \in \mathbb{E}^{n}=(0, \infty)^{n}$ and $\eta=e^{\xi}$ for a random vector $\xi=\left(\xi_{1}, \ldots, \xi_{n}\right)$, where the exponential function is applied componentwise. For simplicity of notation, we do not write time $T$ as a subscript of $\eta$ and incorporate the forward prices $F_{j}$, $j=1, \ldots, n$, into payoff functions; i.e., payoffs will be real-valued functions of $\eta$.

In what follows two specific payoff functions are of particular importance, namely

$$
f_{b}\left(u_{0}, u_{1}, \ldots, u_{n}\right)=\left(\sum_{l=1}^{n} u_{l} \eta_{l}+u_{0}\right)_{+}, \quad u_{0}, u_{1}, \ldots, u_{n} \in \mathbb{R},
$$

for a European basket option and

$$
f_{m}\left(u_{0}, u_{1}, \ldots, u_{n}\right)=u_{0} \vee \bigvee_{l=1}^{n} u_{l} \eta_{l}, \quad u_{0}, u_{1}, \ldots, u_{n} \geq 0
$$

for a European derivative on the maximum of $n$ weighted risky assets together with a riskless bond, where $\vee$ denotes the maximum operation. Despite the fact that the payoff functions $f_{b}$ and $f_{m}$ depend on $\eta$, we stress their dependence on the coefficients, since it is crucial for symmetry properties. By (1.4), call and put options on the maximum of several weighted assets can be written by means of the payoff function $f_{m}$, e.g.,

$$
\left(\bigvee_{l=1}^{n} u_{l} \eta_{l}-k\right)_{+}=f_{m}\left(k, u_{1}, \ldots, u_{n}\right)-k, \quad k \geq 0
$$

It is well known (see, e.g., $[6,32]$ ) that prices of vanilla put or call options with all possible strikes uniquely determine the asset distribution under the martingale measure and thereupon the prices of all other European options, while Carr and Madan [13] present an explicit decomposition of general smooth payoff functions as integrals of vanilla options and riskless bonds. Carr and Laurence [11] proved that the joint probability density function of the assets is determined by the prices of basket options on these assets. In the multivariate
case the following result holds without assuming that the distribution of the assets possesses a density for (i); cf. [22].

Theorem 2.1. The distribution of an integrable $\eta \in \mathbb{E}^{n}$ is uniquely determined by either
(i) the expected values $\mathbf{E} f_{b}\left(u_{0}, u_{1}, \ldots, u_{n}\right)$ for all $u_{0} \in \mathbb{R}$ and $u \in \mathbb{R}^{n}$ or
(ii) the expected values $\mathbf{E} f_{m}\left(u_{0}, u_{1}, \ldots, u_{n}\right)$ for all $u_{0}, u_{1}, \ldots, u_{n} \geq 0$.

Remark 2.2. In view of the positive homogeneity of payoff functions $f_{b}$ and $f_{m}$, it suffices to assume that the expected values from (i) and (ii) are known for parameter vectors ( $u_{0}, u$ ) with norm one in $\mathbb{R}^{n+1}$. By also invoking the put-call parity, it suffices to take in (i) any fixed $u_{0} \neq 0$ and any $u \in \mathbb{R}^{n}$ and in (ii) any fixed $u_{0}>0$ and $u$ with strictly positive coordinates.

Proof of Theorem 2.1. (i) Note that $\mathbf{E} f_{b}\left(u_{0}, u_{1}, \ldots, u_{n}\right)$ with $u_{0}<0$ becomes a call option on the basket $\langle u, \eta\rangle$ and thus uniquely determines the distribution of $\langle u, \eta\rangle$ for each $u \in$ $\mathbb{R}^{n}$. The latter uniquely determines the distribution of $\eta$ by the Cramér-Wold device. This uniqueness result can be also derived geometrically by noticing that $\mathbf{E} f_{b}\left(u_{0}, u_{1}, \ldots, u_{n}\right)$ is the support function of the lift zonoid of $\eta$, which uniquely identifies the distribution of $\eta$; see [30].
(ii) Note that $\mathbf{E} f_{m}\left(u_{0}, u_{1}, \ldots, u_{n}\right)=\mathbf{E} u_{0} \vee \zeta$ for $\zeta=u_{1} \eta_{1} \vee \cdots \vee u_{n} \eta_{n}$ and thus uniquely determines the distribution of $\zeta$. The cumulative distribution function of $\zeta$ is given by

$$
F_{\zeta}(t)=\mathbf{P}\left\{\eta_{1} \leq \frac{t}{u_{1}}, \ldots, \eta_{n} \leq \frac{t}{u_{n}}\right\},
$$

and so these functions for all $u_{1}, \ldots, u_{n}$ uniquely determine the joint cumulative distribution function of $\eta_{1}, \ldots, \eta_{n}$. Note in this relation that $\mathbf{E} f_{m}\left(u_{0}, u_{1}, \ldots, u_{n}\right), u_{0}, u_{1}, \ldots, u_{n} \geq 0$, is the support function of a convex body called the lift max-zonoid of $\eta$; see [28].

Remark 2.3. Consider call and put options that are written on the weighted maximum $\max \left(u_{1} \eta_{1}, \ldots, u_{n} \eta_{n}\right)$ or minimum $\min \left(u_{1} \eta_{1}, \ldots, u_{n} \eta_{n}\right)$ of $n$ integrable assets. By the univariate version (with $n=1$ ) of Theorem 2.1(i), the prices of such calls and puts determine the distribution of the weighted maximum and minimum. By the argument similar to the proof of statement (ii) of Theorem 2.1, it is easy to see that the families of such prices for all weights $u_{1}, \ldots, u_{n} \geq 0$ uniquely determine the distribution of $\eta=\left(\eta_{1}, \ldots, \eta_{n}\right)$.

Fix an arbitrary asset number $i \in\{1, \ldots, n\}$ and assume that $\mathbf{Q}$ is a probability measure that makes $\eta$ integrable. Recall that $\mathbf{E}$ without the subscript denotes the expectation with respect to $\mathbf{Q}$; otherwise the subscript is used to indicate the relevant probability measure. Since $\eta^{1 / 2}=\left(\eta_{1}^{1 / 2}, \ldots, \eta_{n}^{1 / 2}\right)=e^{\frac{1}{2} \xi}$ is integrable, we can define new probability measures $\mathbf{Q}^{i}$ and $\mathcal{E}^{i}$ by

$$
\frac{d \mathbf{Q}^{i}}{d \mathbf{Q}}=\frac{\eta_{i}}{\mathbf{E} \eta_{i}}, \quad \frac{d \mathcal{E}^{i}}{d \mathbf{Q}}=\frac{e^{\frac{1}{2} \xi_{i}}}{\mathbf{E}^{\frac{1}{2} \xi_{i}}} .
$$

Hence, $\mathcal{E}^{i}$ is the Esscher (exponential) transform of $\mathbf{Q}$ with parameter $\frac{1}{2} e_{i}$, where $e_{i}$ is the $i$ th standard basis vector in $\mathbb{R}^{n}$; see [33] and [35, Ex. 7.3] for the Esscher transform in the context of multivariate Lévy processes. Since $\mathbf{E}_{\mathcal{E}^{i}} e^{-\frac{1}{2} \xi_{i}}=\left(\mathbf{E} e^{\frac{1}{2} \xi_{i}}\right)^{-1}$, we see that $\mathbf{Q}$ is the Esscher transform of $\mathcal{E}^{i}$ with parameter $-\frac{1}{2} e_{i}$.

For simplicity of notation, define families of functions $\varkappa_{i}: \mathbb{E}^{n} \mapsto \mathbb{E}^{n}$ and linear mappings
$K_{i}: \mathbb{R}^{n} \mapsto \mathbb{R}^{n}$ acting as

$$
\begin{align*}
\varkappa_{i}(x) & =\left(\frac{x_{1}}{x_{i}}, \ldots, \frac{x_{i-1}}{x_{i}}, \frac{1}{x_{i}}, \frac{x_{i+1}}{x_{i}}, \ldots, \frac{x_{n}}{x_{i}}\right) \\
K_{i} x & =\left(x_{1}-x_{i}, \ldots, x_{i-1}-x_{i},-x_{i}, x_{i+1}-x_{i}, \ldots, x_{n}-x_{i}\right) \tag{2.2}
\end{align*}
$$

for $i=1, \ldots, n$. The linear mapping $K_{i}$ can be represented by the matrix $K_{i}=\left(k_{l m}\right)_{l m=1}^{n}$ with $k_{l l}=1$ for all $l \neq i, k_{l i}=-1$ for $l=1, \ldots, n$ with all remaining entries being 0 . Note that $\varkappa_{i}$ and $K_{i}$ are self-inverse, i.e., $\varkappa_{i}\left(\varkappa_{i}(x)\right)=x$ and $K_{i} K_{i} x=x$, and that the $i$ th coordinate of $K_{i} x$ is $-x_{i}$. The transpose of $K_{i}$ is denoted by $K_{i}^{\top}$. In the following we consider vectors as rows or columns depending on the situation.

The permutation of the zero coordinate with the $i$ th coordinate of a vector $\left(u_{0}, u\right) \in \mathbb{R}^{n+1}$ is denoted by

$$
\pi_{i}\left(u_{0}, u\right)=\left(u_{i}, u_{1}, \ldots, u_{i-1}, u_{0}, u_{i+1}, \ldots, u_{n}\right) \quad \text { for } i=1, \ldots, n \text {. }
$$

Finally, $\varphi_{\xi}^{\mathbf{Q}}$ (resp., $\varphi_{\xi}^{\mathcal{E}^{i}}$ ) denotes the characteristic function of the random vector $\xi$ under the probability measure $\mathbf{Q}$ (resp., $\mathcal{E}^{i}$ ). Univariate versions of nearly all statements (apart from (ii), which implies the equivalence between (1.3) and (1.5)) of the following theorem are already known from [12, Th. 2.5, Cor. 2.10].

Theorem 2.4. Let $\eta=e^{\xi}$ be an $n$-dimensional $\mathbf{Q}$-integrable random vector with positive components and let $i$ be a fixed number from $\{1, \ldots, n\}$. The following conditions are equivalent:
(i) For all $u_{0} \in \mathbb{R}$ and $u \in \mathbb{R}^{n}$,

$$
\mathbf{E} f_{b}\left(u_{0}, u\right)=\mathbf{E} f_{b}\left(\pi_{i}\left(u_{0}, u\right)\right)
$$

(ii) For all $u_{0} \geq 0$ and $u \in \mathbb{R}_{+}^{n}$,

$$
\mathbf{E} f_{m}\left(u_{0}, u\right)=\mathbf{E} f_{m}\left(\pi_{i}\left(u_{0}, u\right)\right) .
$$

(iii) For any payoff function $f: \mathbb{R}_{+}^{n} \mapsto \mathbb{R}_{+}$, we have

$$
\mathbf{E} f(\eta)=\mathbf{E}\left[f\left(\varkappa_{i}(\eta)\right) \eta_{i}\right] .
$$

(iv) The distribution of $\eta$ under $\mathbf{Q}$ is identical to the distribution of $\tilde{\eta}^{i}=\varkappa_{i}(\eta)$ under $\mathbf{Q}^{i}$.
(v) The distributions of $\xi$ and $K_{i} \xi$ under $\mathcal{E}^{i}$ coincide.
(vi) For every $u \in \mathbb{R}^{n}$,

$$
\varphi_{\xi}^{\mathcal{E}^{i}}(u)=\varphi_{\xi}^{\mathcal{E}^{i}}\left(K_{i}^{\top} u\right)
$$

or, equivalently,

$$
\varphi_{\xi}^{\mathbf{Q}}\left(u-\frac{1}{2} \imath e_{i}\right)=\varphi_{\xi}^{\mathbf{Q}}\left(K_{i}^{\top} u-\frac{1}{2} \imath e_{i}\right),
$$

where $\boldsymbol{\imath}=\sqrt{-1}$ is the imaginary unit and $e_{i}$ is the ith standard basis vector in $\mathbb{R}^{n}$.

Since (iv) corresponds to the duality transform in the univariate setting, we say that $\eta$ satisfying one of the above conditions is self-dual with respect to the ith numeraire and write in short $\eta \in \mathrm{SD}_{i}$. If $\eta$ is self-dual with respect to all numeraires $i=1, \ldots, n$, we call $\eta$ jointly self-dual.

Remark 2.5 (martingale property). If $\eta \in \mathrm{SD}_{i}$, then (iii) applied to $f$ identically equal to one implies that $\mathbf{E} \eta_{i}=1$, which is the case if the $i$ th component of $\eta$ arises from a Q-martingale sampled at maturity. However, $\mathbf{E} \eta_{j}$ is not necessarily one for $i \neq j$, quite differently from the univariate case; see [12]. This property has to be imposed additionally, if needed.

Remark 2.6. If the forward prices are not included in the payoff function, then (iii) for the self-duality of $\eta$ with respect to the $i$ th numeraire can be equivalently expressed as

$$
\mathbf{E} f\left(S_{T 1}, \ldots, S_{T n}\right)=\mathbf{E}\left[f\left(\frac{S_{T 1} F_{i}}{S_{T i}}, \ldots, \frac{S_{T(i-1)} F_{i}}{S_{T i}}, \frac{\left(F_{i}\right)^{2}}{S_{T i}}, \frac{S_{T(i+1)} F_{i}}{S_{T i}}, \ldots, \frac{S_{T n} F_{i}}{S_{T i}}\right) \frac{S_{T i}}{F_{i}}\right]
$$

by applying (iii) to $\tilde{f}(\eta)=f(F \circ \eta)$. Note also that the self-duality with respect to the $i$ th numeraire implies (iii) for not necessarily positive integrable payoff functions.

Remark 2.7 (conditioning in Theorem 2.4). All conditions of Theorem 2.4 can be written conditionally on a fixed event or conditionally on a $\sigma$-algebra. This has the following application for stochastic processes. Consider a family $\{\eta(t), t \geq 0\}$ of random vectors being self-dual with respect to the $i$ th numeraire. If $\tau$ is a nonnegative random variable, which is independent of $\{\eta(t), t \geq 0\}$, then $\eta(\tau)$ satisfies all statements in Theorem 2.4 with the expectations taken conditionally on the $\sigma$-algebra generated by $\tau$; i.e., $\eta(\tau)$ is conditionally self-dual with respect to the $i$ th numeraire.

Proof of Theorem 2.4. We will establish all equivalences in several steps. (i) $\Rightarrow$ (iv) By (i), $\mathbf{E} \eta_{i}=1$, so that for all $\left(u_{0}, u\right) \in \mathbb{R}^{n+1}$

$$
\begin{aligned}
\mathbf{E}_{\mathbf{Q}^{i}}\left(\sum_{l=1}^{n} u_{l} \tilde{\eta}_{l}^{i}+u_{0}\right)_{+} & =\mathbf{E}_{\mathbf{Q}^{i}}\left(\sum_{l=1, l \neq i}^{n} u_{l} \frac{\eta_{l}}{\eta_{i}}+\frac{u_{i}}{\eta_{i}}+u_{0}\right)_{+} \\
& =\mathbf{E}\left[\left(\sum_{l=1, l \neq i}^{n} u_{l} \frac{\eta_{l}}{\eta_{i}}+\frac{u_{i}}{\eta_{i}}+u_{0}\right)_{+} \eta_{i}\right] \\
& =\mathbf{E}\left(\sum_{l=1, l \neq i}^{n} u_{l} \eta_{l}+u_{i}+u_{0} \eta_{i}\right)_{+}=\mathbf{E}\left(\sum_{l=1}^{n} u_{l} \eta_{l}+u_{0}\right)_{+}
\end{aligned}
$$

where the last equality follows from (i). Theorem 2.1(i) yields (iv).
(iv) $\Rightarrow$ (iii) $\Rightarrow$ (i) The definition of $\mathbf{Q}^{i}$, the self-inverse property of $\varkappa_{i}$, and (iv) (also implying $\mathbf{E} \eta_{i}=1$ ) yield that

$$
\mathbf{E} f(\eta)=\mathbf{E}_{\mathbf{Q}^{i}}\left[f(\eta) \eta_{i}^{-1}\right]=\mathbf{E}_{\mathbf{Q}^{i}}\left[f\left(\varkappa_{i}\left(\tilde{\eta}^{i}\right)\right) \tilde{\eta}_{i}^{i}\right]=\mathbf{E}\left[f\left(\varkappa_{i}(\eta)\right) \eta_{i}\right],
$$

so that (iii) holds. By applying (iii) for the payoff function $f_{b}$ of a basket option, we arrive at (i).
(iii) $\Rightarrow$ (ii) $\Rightarrow$ (iv) By applying (iii) to the payoff function $f_{m}$, we obtain (ii). The next implication is proved similarly to (i) $\Rightarrow$ (iv) above.
$($ iii $) \Leftrightarrow($ v $) \Leftrightarrow($ vi $)$ Since $m_{i}=\mathbf{E}\left(e^{\frac{1}{2} \xi_{i}}\right)$ is finite,

$$
\begin{aligned}
\mathbf{E} f(\eta) & =\mathbf{E} f\left(e^{\xi}\right)=m_{i} \mathbf{E}_{\mathcal{E}^{i}}\left[f\left(e^{\xi}\right) e^{-\frac{1}{2} \xi_{i}}\right], \\
\mathbf{E}\left[f\left(\varkappa_{i}(\eta)\right) \eta_{i}\right] & =\mathbf{E}\left[f\left(e^{K_{i} \xi}\right) e^{\xi_{i}}\right]=m_{i} \mathbf{E}_{\mathcal{E}^{i}}\left[f\left(e^{K_{i} \xi}\right) e^{\frac{1}{2} \xi_{i}}\right] .
\end{aligned}
$$

Thus, (iii) yields that

$$
\begin{equation*}
\mathbf{E}_{\mathcal{E}^{i}}\left[f\left(e^{\xi}\right) e^{-\frac{1}{2} \xi_{i}}\right]=\mathbf{E}_{\mathcal{E}^{i}}\left[f\left(e^{K_{i} \xi}\right) e^{\frac{1}{2} \xi_{i}}\right]=\mathbf{E}_{\mathcal{E}^{i}}\left[f\left(e^{K_{i} \xi}\right) e^{-\frac{1}{2}\left(K_{i} \xi\right)_{i}}\right] \tag{2.3}
\end{equation*}
$$

for any payoff function $f$. Recall that the $i$ th coordinate of $K_{i} \xi$ is $-\xi_{i}$. Choosing $f\left(e^{x}\right)=$ $g\left(e^{x}\right) e^{\frac{1}{2} x_{i}}$, we see that the $\mathcal{E}^{i}$-expectations of $g\left(e^{\xi}\right)$ and $g\left(e^{K_{i} \xi}\right)$ coincide for all nonnegative continuous functions $g$ with bounded support, whence $\xi$ coincides in distribution with $K_{i} \xi$ under $\mathcal{E}^{i}$. Conversely, if $\xi$ and $K_{i} \xi$ share the same distribution under $\mathcal{E}^{i}$, then (2.3) holds and implies (iii).

Furthermore, (v) is equivalent to

$$
\varphi_{\xi}^{\mathcal{E}^{i}}(u)=\varphi_{K_{i} \xi}^{\mathcal{E}^{i}}(u)=\varphi_{\xi}^{\mathcal{E}^{i}}\left(K_{i}^{\top} u\right)
$$

for every $u \in \mathbb{R}^{n}$. Writing the characteristic functions as $\mathcal{E}^{i}$-expectations and referring to the change of measure, the latter condition is equivalent to

$$
\mathbf{E}\left[e^{\imath\langle u, \xi\rangle} \frac{e^{\frac{1}{2} \xi_{i}}}{\mathbf{E} e^{\frac{1}{2} \xi_{i}}}\right]=\mathbf{E}\left[e^{\imath\left\langle K_{i}^{\top} u, \xi\right\rangle} \frac{e^{\frac{1}{2} \xi_{i}}}{\mathbf{E} e^{\frac{1}{2} \xi_{i}}}\right], \quad u \in \mathbb{R}^{n},
$$

so that

$$
\varphi_{\xi}^{\mathrm{Q}}\left(u-\frac{1}{2} \imath e_{i}\right)=\varphi_{\xi}^{\mathbf{Q}}\left(K_{i}^{\top} u-\frac{1}{2} \imath e_{i}\right)
$$

for all $u \in \mathbb{R}^{n}$.
The following result can be helpful for constructing distributions of self-dual random vectors. Its univariate version is stated in [37, Ex. 8].

Theorem 2.8. Consider an integrable random vector $\eta \in \mathbb{E}^{n}$ with distribution $\mathbf{Q}$.
(a) If $\eta$ is absolutely continuous with probability density $p_{\eta}$, then $\eta \in \mathrm{SD}_{i}$ if and only if

$$
\begin{equation*}
p_{\eta}(x)=x_{i}^{-(n+2)} p_{\eta}\left(\varkappa_{i}(x)\right) \quad \text { for almost all } x \in \mathbb{E}^{n} ; \tag{2.4}
\end{equation*}
$$

equivalently, the density $p_{\xi}$ of $\xi=\log \eta$ satisfies

$$
\begin{equation*}
p_{\xi}(x)=e^{-x_{i}} p_{\xi}\left(K_{i} x\right) \quad \text { for almost all } x \in \mathbb{R}^{n} . \tag{2.5}
\end{equation*}
$$

(b) If $\eta$ is discrete, then $\eta \in \mathrm{SD}_{i}$ if and only if $\mathbf{Q}\left(\eta=\varkappa_{i}(x)\right)=x_{i} \mathbf{Q}(\eta=x)$ for each atom $x$ of $\eta$.
Proof. (a) Condition (iii) of Theorem 2.4 can be written in integral form as

$$
\int_{\mathbb{R}_{+}^{n}} f(x) p_{\eta}(x) d x=\int_{\mathbb{R}_{+}^{n}} f\left(\varkappa_{i}(y)\right) y_{i} p_{\eta}(y) d y=\int_{\mathbb{R}_{+}^{n}} f(x) \frac{1}{x_{i}^{n+2}} p_{\eta}\left(\varkappa_{i}(x)\right) d x
$$

where the last equality is obtained by changing variables $x=\varkappa_{i}(y)$ and noticing that $\varkappa_{i}\left(\varkappa_{i}(x)\right)$ $=x$. Consider the function $f(x)=\mathbb{1}_{x \in\left[a_{1}, b_{1}\right] \times \cdots \times\left[a_{n}, b_{n}\right]}$ for any parameters $a_{1}, \ldots, a_{n}, b_{1}, \ldots, b_{n}$ $\in \mathbb{R}$. Differentiating both sides with respect to $b_{1}, \ldots, b_{n}$ we get (2.4) almost everywhere. For the converse, write the right-hand side of (iii) as an integral, refer to (2.4), and change variables. The equivalence between (2.4) and (2.5) can be seen by the classical density transformation.
(b) In the discrete case the necessity relies on writing Theorem 2.4(iii) for $f(x)=\mathbb{1}_{x=x_{*}}$, where $x_{*}$ is any atom of $\mathbf{Q}$. The sufficiency follows from writing the expected payoff $f_{b}\left(u_{0}, u\right)$ as the sum over the atoms of $\mathbf{Q}$.

Now we give a result about the marginal distribution of $\eta_{i}$ for the random vector $\eta$ being self-dual with respect to this numeraire.

Lemma 2.9. If $\eta \in \mathrm{SD}_{i}$, then $\eta_{i}$ is a self-dual random variable.
Proof. Choose in Theorem 2.4(ii) vector $u$ with all coordinates being zero apart from $u_{0}$ and $u_{i}$. Then Theorem 2.4(ii) reads $\mathbf{E}\left(u_{0} \vee u_{i} \eta_{i}\right)=\mathbf{E}\left(u_{i} \vee u_{0} \eta_{i}\right)$ for every $u_{0}, u_{i} \geq 0$. This is exactly the symmetry condition in the one-dimensional version of Theorem 2.4(ii).
3. Jointly self-dual random vectors. Recall that random vector $\eta$ is called jointly selfdual if it is self-dual with respect to all numeraires. Since permutations of coordinate 0 and an arbitrary $i \in\{1, \ldots, n\}$ generate by successive applications the transpositions of any two $i, j \in\{0,1, \ldots, n\}$, the expected payoff functions $f_{b}$ and $f_{m}$ for jointly self-dual $\eta$ are invariant with respect to any permutation of their arguments; e.g.,

$$
\begin{equation*}
\mathbf{E} f_{b}\left(u_{0}, u_{1}, \ldots, u_{n}\right)=\mathbf{E} f_{b}\left(u_{l_{0}}, u_{l_{1}}, \ldots, u_{l_{n}}\right) \tag{3.1}
\end{equation*}
$$

for each permutation $i \mapsto l_{i}$. In view of this, Theorem 2.1 implies the following result.
Corollary 3.1. If $\eta$ is jointly self-dual, then all its components are identically distributed self-dual random variables with expectation one and $\eta$ is exchangeable; i.e., its distribution does not change after any permutation of its coordinates.

It should be noted that the converse statement to Corollary 3.1 does not hold; i.e., the exchangeability of $\eta$ does not imply joint self-duality. This is easily seen as a consequence of the following result, which says that any nontrivial random vector $\eta$ with independent coordinates cannot be jointly self-dual.

Theorem 3.2. Assume that $n \geq 2$.
(a) If $\eta \in \mathrm{SD}_{i}$ and $\eta_{i}$ and $\eta_{j}$ are independent for some $j \neq i$, then $\eta_{i}$ equals 1 almost surely.
(b) If $\eta$ is a jointly self-dual random vector with independent coordinates, then all coordinates of $\eta$ are deterministic and equal 1 almost surely.
Proof. It suffices to prove only (a). By Theorem 2.4(ii), letting $u_{l}=0$ for $l \neq 0, i, j$,

$$
\mathbf{E}\left(u_{0} \vee u_{i} \eta_{i} \vee u_{j} \eta_{j}\right)=\mathbf{E}\left(u_{i} \vee u_{0} \eta_{i} \vee u_{j} \eta_{j}\right)
$$

for all $u_{0}, u_{i}, u_{j} \in \mathbb{R}_{+}$. In particular, if $u_{i}=0$, then

$$
\mathbf{E}\left(u_{0} \vee u_{j} \eta_{j}\right)=\mathbf{E}\left(u_{0} \eta_{i} \vee u_{j} \eta_{j}\right) \quad \text { for all }\left(u_{0}, u_{j}\right) \in \mathbb{R}_{+}^{2} .
$$

Since $\eta_{i}$ is self-dual by Lemma 2.9, the conditioning on $\eta_{j}$ yields that $\mathbf{E}\left(u_{0} \eta_{i} \vee u_{j} \eta_{j}\right)=$ $\mathbf{E}\left(u_{0} \vee u_{j} \eta_{i} \eta_{j}\right)$. Hence,

$$
\mathbf{E}\left(u_{0} \vee u_{j} \eta_{j}\right)=\mathbf{E}\left(u_{0} \vee u_{j} \eta_{i} \eta_{j}\right) \quad \text { for all }\left(u_{0}, u_{j}\right) \in \mathbb{R}_{+}^{2},
$$

whence $\eta_{j}$ coincides in distribution with $\eta_{i} \eta_{j}$; see Theorem 2.1(ii). If $\xi_{l}=\log \eta_{l}$ for $l=i, j$, then $\xi_{j}$ and $\xi_{i}+\xi_{j}$ share the same distribution. Therefore, the characteristic function of $\xi_{i}$ identically equals one for some neighborhood of the origin, whence $\xi_{i}=0$ almost surely.

Remark 3.3 (random vectors sampled from Lévy processes). Assume that $\zeta_{1}, \ldots, \zeta_{n}$ are independent integrable random variables. Consider the vector

$$
\eta=\left(\zeta_{1}, \zeta_{1} \zeta_{2}, \zeta_{1} \zeta_{2} \zeta_{3}, \ldots, \zeta_{1} \cdots \zeta_{n}\right)
$$

This construction is important, since then $\xi=\log \eta$ is a vector whose components form a random walk. However, $\eta$ cannot be jointly self-dual as a vector, except in the trivial deterministic case. Indeed, setting $u_{0}, u_{1}, u_{2} \geq 0$ and $u_{3}=\cdots=u_{n}=0$, writing the expected payoff $f_{m}$ conditionally on $\zeta_{2}$, and using the self-duality of $\zeta_{1}$ (which follows from $\eta \in \mathrm{SD}_{1}$ ), we see that

$$
\mathbf{E}\left(u_{0} \vee u_{1} \eta_{1} \vee u_{2} \eta_{2}\right)=\mathbf{E}\left(u_{0} \zeta_{1} \vee u_{1} \vee u_{2} \zeta_{2}\right)
$$

is symmetric in $u_{0}, u_{1}, u_{2}$ if $\eta$ is jointly self-dual. Thus, $\left(\zeta_{1}, \zeta_{2}\right)$ is a jointly self-dual vector with independent components, which is necessarily trivial by Theorem 3.2(b). An extension of this argument shows that $\zeta_{1}=\cdots=\zeta_{n}=1$ almost surely. Therefore, it is not possible to obtain jointly self-dual random vectors by taking exponentials of the values of a Lévy process at different time points.

Example 3.4. Let $\zeta_{0}, \zeta_{1}, \ldots, \zeta_{n}$ be independent and identically distributed self-dual random variables (their examples are provided in section 6). Define $\eta_{i}=\zeta_{0} \zeta_{i}$ for $i=1, \ldots, n$. Conditioning on $\zeta_{1}, \ldots, \zeta_{n}$ yields that

$$
\begin{aligned}
\mathbf{E}\left(u_{0} \vee u_{1} \eta_{1} \vee \cdots \vee u_{n} \eta_{n}\right) & =\mathbf{E}\left[\mathbf{E}\left(u_{0} \vee \zeta_{0}\left(u_{1} \zeta_{1} \vee \cdots \vee u_{n} \zeta_{n}\right) \mid \zeta_{1}, \ldots, \zeta_{n}\right)\right] \\
& =\mathbf{E}\left(u_{0} \zeta_{0} \vee u_{1} \zeta_{1} \vee \cdots \vee u_{n} \zeta_{n}\right)
\end{aligned}
$$

is symmetric in $u_{0}, u_{1}, \ldots, u_{n}$; i.e., $\eta=\left(\eta_{1}, \ldots, \eta_{n}\right)$ is jointly self-dual. Note that $\eta_{1}, \ldots, \eta_{n}$ are all self-dual random variables but are no longer independent. In particular if the $\zeta$ 's are $\log$-normally distributed with $\mu=-\frac{1}{2}$ and $\sigma=1$, then $\log \eta$ is normally distributed with mean $\left(-\frac{1}{2}, \ldots,-\frac{1}{2}\right)$ and the covariance matrix having diagonal elements 1 and all others $\frac{1}{2}$. We will return to this situation in Example 4.5.
4. Exponentially self-dual infinitely divisible random vectors. A random vector $\xi$ has an infinitely divisible distribution if and only if $\xi=L_{1}$ for a Lévy process $\left(L_{t}\right)_{t \geq 0}$; see [34]. In view of the widespread use of Lévy models for derivative pricing, we aim to characterize infinitely divisible random vectors $\xi=\log \eta$ for $\eta$ being self-dual with respect to the $i$ th numeraire or all numeraires. If $\eta \in \mathrm{SD}_{i}$, then $\xi$ is said to be exponentially self-dual with respect to the $i$ th numeraire and we write in short $\xi \in \mathrm{ESD}_{i}$.

The Euclidean norm $\|\cdot\|$ is not invariant with respect to the transformation $x \mapsto K_{i} x$ defined by (2.2). For simplifying the formulation of the results, we introduce the following
norm on $\mathbb{R}^{n}$ :

$$
\begin{equation*}
\|u\|^{2}=\frac{1}{2}\left(\|u\|^{2}+\left\|K_{i} u\right\|^{2}\right), \quad u \in \mathbb{R}^{n} \tag{4.1}
\end{equation*}
$$

where the number $i \in\{1, \ldots, n\}$ is fixed in what follows. It is easy to see that $\|\cdot \cdot\|$ is indeed a norm, which is equivalent to the Euclidean norm on $\mathbb{R}^{n}$. Since $K_{i}$ is self-inverse, $\|u\|=\left\|K_{i} u\right\|$ for every $u \in \mathbb{R}^{n}$.

We use the following formulation of the Lévy-Khintchine formula (see [34, Chap. 2]) for the characteristic function of $\xi$ :

$$
\begin{align*}
\varphi_{\xi}^{\mathbf{Q}}(u)=\mathbf{E} e^{\imath\langle u, \xi\rangle}=\exp \{\boldsymbol{\imath}\langle\gamma, u\rangle & -\frac{1}{2}\langle u, A u\rangle  \tag{4.2}\\
& \left.+\int_{\mathbb{R}^{n}}\left(e^{\imath\langle u, x\rangle}-1-\boldsymbol{\imath}\langle u, x\rangle \mathbb{1}_{\|x\| \leq 1}\right) d \nu(x)\right\}
\end{align*}
$$

for $u \in \mathbb{R}^{n}$, where $A$ is a symmetric nonnegative definite $n \times n$ matrix, $\gamma \in \mathbb{R}^{n}$ is a constant vector, and $\nu$ is a measure on $\mathbb{R}^{n}$ (called the Lévy measure) satisfying $\nu(\{0\})=0$ and

$$
\begin{equation*}
\int_{\mathbb{R}^{n}} \min \left(\|x\|^{2}, 1\right) d \nu(x)<\infty \tag{4.3}
\end{equation*}
$$

Note that the latter condition can be equivalently written in the new norm $\|\cdot\|$.
Theorem 4.1. Let $\eta$ be an integrable random vector under the probability measure $\mathbf{Q}$ such that $\xi=\log \eta$ is infinitely divisible under $\mathbf{Q}$. Then $\xi \in \mathrm{ESD}_{i}$ if and only if for the generating triplet $(A, \nu, \gamma)$ the following three conditions hold:
(1) The matrix $A=\left(a_{l j}\right)_{l j=1}^{n}$ satisfies $a_{i j}=a_{j i}=\frac{1}{2} a_{i i}$ for all $j=1, \ldots, n, j \neq i$.
(2) The Lévy measure satisfies

$$
\begin{equation*}
d \nu(x)=e^{-x_{i}} d \nu\left(K_{i} x\right) \quad \text { almost everywhere } \tag{4.4}
\end{equation*}
$$

meaning that $\nu(B)=\int_{K_{i} B} e^{x_{i}} d \nu(x)$ for all Borel $B$.
(3) The ith coordinate of $\gamma$ satisfies

$$
\begin{equation*}
\gamma_{i}=\int_{\|x\| \leq 1} x_{i}\left(1-e^{\frac{1}{2} x_{i}}\right) d \nu(x)-\frac{1}{2} a_{i i} \tag{4.5}
\end{equation*}
$$

Proof. Since $\eta$ is positive integrable, $0<\mathbf{E} e^{\frac{1}{2} \xi_{i}}<\infty$, so that the Esscher transform $\mathcal{E}^{i}$ of $\mathbf{Q}$ with parameter $\frac{1}{2} e_{i}$ and the inverse transform are well defined. According to [33] or [35, Ex. 7.3], $\xi$ under $\mathcal{E}^{i}$ also has an infinitely divisible distribution, so that

$$
\varphi_{\xi}^{\mathcal{E}^{i}}(u)=\exp \left\{\boldsymbol{\imath}\left\langle\gamma^{\mathcal{E}^{i}}, u\right\rangle-\frac{1}{2}\langle u, A u\rangle+\int_{\mathbb{R}^{n}}\left(e^{\imath\langle u, x\rangle}-1-\boldsymbol{\imath}\langle u, x\rangle \mathbb{1}_{\|x\| \leq 1}\right) d \nu^{\mathcal{E}^{i}}(x)\right\}
$$

for a new vector $\gamma^{\mathcal{E}^{i}}$ and Lévy measure $\nu^{\mathcal{E}}$. Note that the matrix $A$ is invariant under the Esscher transform (see [33] or [35, Ex. 7.3]), as is the case under any absolutely continuous change of measure; see, e.g., [24].

By Theorem 2.4(vi), $\xi \in \mathrm{ESD}_{i}$ if and only if

$$
\begin{equation*}
\varphi_{\xi}^{\mathcal{E}^{i}}(u)=\varphi_{\xi}^{\mathcal{E}^{i}}\left(K_{i}^{\top} u\right) \quad \text { for all } u \in \mathbb{R}^{n} \tag{4.6}
\end{equation*}
$$

Noticing that $\left\langle K_{i}^{\top} u, x\right\rangle=\left\langle u, K_{i} x\right\rangle$, changing the variable $x$ to $K_{i} x$, using the $K_{i}$-invariance of $\|\cdot\|$ and the self-inverse property of $K_{i}$, the Lévy-Khintchine formula yields that

$$
\begin{aligned}
\varphi_{\xi}^{\mathcal{E}^{i}}\left(K_{i}^{\top} u\right)=\exp \left\{\imath\left\langle K_{i} \mathcal{\gamma}^{\mathcal{E}}, u\right\rangle\right. & -\frac{1}{2}\left\langle u, K_{i} A K_{i}^{\top} u\right\rangle \\
& \left.+\int_{\mathbb{R}^{n}}\left(e^{\imath\langle u, x\rangle}-1-\imath\langle u, x\rangle \mathbb{1}\|x\| \leq 1\right) d \nu^{\mathcal{E}^{i}}\left(K_{i} x\right)\right\}
\end{aligned}
$$

The uniqueness of the parameters $A, \nu^{\mathcal{E}^{i}}$, and $\gamma^{\mathcal{E}^{i}}$ of the Lévy-Khintchine representation for $\varphi_{\xi}^{\mathcal{E}^{i}}$ [34, Thm. 8.1] implies that (4.6) holds if and only if the Lévy measure $\nu^{\mathcal{E}^{i}}$ is $K_{i}$-invariant, $A=K_{i} A K_{i}^{\top}$, and $\gamma^{\mathcal{E}^{i}}=K_{i} \gamma^{\mathcal{E}^{i}}$. The condition on $\gamma^{\mathcal{E}^{i}}$ holds if and only if the $i$ th component of $\gamma^{\mathcal{E}^{i}}$ is 0 . A direct computation of matrix products leads to (1).

Since the norm $\|\cdot\|$ does not change integrability properties in the Lévy-Khintchine representation, it is possible to replicate the proof from [33] to show that the Esscher transform with parameter $-\frac{1}{2} e_{i}$ leaves $A$ invariant, while other parts of the Lévy triplet are transformed as

$$
\begin{aligned}
d \nu(x) & =e^{-\frac{1}{2} x_{i}} d \nu^{\mathcal{E}^{i}}(x), \\
\gamma & =\gamma^{\mathcal{E}^{i}}+\int_{\|x\| \leq 1} x\left(e^{-\frac{1}{2} x_{i}}-1\right) d \nu^{\mathcal{E}^{i}}(x)+A\left(-\frac{1}{2} e_{i}\right) .
\end{aligned}
$$

The latter condition is equivalent to (4.5), noticing that the $i$ th component of $A\left(\frac{1}{2} e_{i}\right)$ is $a_{i i} / 2$, $d \nu^{\mathcal{E}^{i}}(x)=e^{\frac{1}{2} x_{i}} d \nu(x)$, and $\gamma^{\mathcal{E}^{i}}$ has zero as its $i$ th component, while other components are arbitrary.

Furthermore, for almost all $x$,

$$
d \nu(x)=e^{-\frac{1}{2} x_{i}} d \nu^{\mathcal{E}^{i}}(x)=e^{-x_{i}} e^{-\frac{1}{2}\left(-x_{i}\right)} d \nu^{\mathcal{E}^{i}}\left(K_{i} x\right)=e^{-x_{i}} d \nu\left(K_{i} x\right),
$$

where again we used the fact that the $i$ th component of $K_{i} x$ is $-x_{i}$ and that $\nu^{\mathcal{E}^{i}}$ is $K_{i}$-invariant.
Conversely, the integrability of $\eta$ ensures the existence of the Esscher transform of $\mathbf{Q}$ with parameter $\frac{1}{2} e_{i}$. By doing this transform and the converse calculations, it is easy to verify that Theorem 2.4(vi) applies, i.e., $\eta=e^{\xi} \in \mathrm{SD}_{i}$.

Since an infinitely divisible random variable $\xi$ is symmetric if and only if $\gamma$ vanishes and the Lévy measure is symmetric, the above proof is very short in the univariate case and immediately yields the corresponding univariate result stated in [19, 12].

The $\mathrm{SD}_{i}$-property of $\eta$ implies that the $i$ th component of $\eta$ has expectation one. If this holds for other components, e.g., if $\eta$ forms a martingale, this imposes further restrictions on the coordinates of $\gamma$, namely

$$
\gamma_{j}+\frac{1}{2} a_{j j}+\int_{\mathbb{R}^{n}}\left(e^{x_{j}}-1-x_{j} \mathbb{\mathbb { }}\|x\| \leq 1\right) d \nu(x)=0, \quad j=1, \ldots, n
$$

Remark 4.2 (the role of the norm). If we use the Euclidean norm to define the truncation in (4.2), then this change affects only the value of $\gamma$, while $A$ and $\nu$ remain the same. If $\gamma_{\|\cdot\|}$ denotes the "drift" calculated for the Euclidean norm, then

$$
\gamma_{\|\cdot\|}=\gamma+\int_{\mathbb{R}^{n}} x\left(\mathbb{1}_{\|x\| \leq 1}-\mathbb{1}_{\|x\| \leq 1}\right) d \nu
$$

so that (4.5) transforms into

$$
\gamma_{\|\cdot\| i}=\int_{\mathbb{R}^{n}} x_{i}\left(\mathbb{1}_{\|x\| \leq 1}-\mathbb{1}_{\|x\| \leq 1} e^{\frac{1}{2} x_{i}}\right) d \nu(x)-\frac{1}{2} a_{i i} .
$$

Remark 4.3 (jointly self-dual case). Assume that the conditions of Theorem 4.1 hold for each $i=1, \ldots, n$. The first condition implies that $A$ equals up to a constant factor the matrix having all diagonal elements 1 and the others $\frac{1}{2}$. By applying (4.4) consecutively to coordinates $i \neq j$ and noticing that $K_{i} K_{j} K_{i}$ defines the transposition of the $i$ th and $j$ th coordinates of $n$-dimensional vectors, we see that in this case the Lévy measure $\nu$ is invariant under permutations and all components of $\gamma$ coincide.

Remark 4.4 (finite mean case). Now we also assume that $\xi$ has a finite mean, which is the case if and only if $\int_{\|x\|>1}\|x\| d \nu(x)<\infty$; see [34, Cor. 25.8]. Then we can rewrite (4.2) in the following form:

$$
\begin{equation*}
\varphi_{\xi}^{\mathbf{Q}}(u)=\exp \left\{\imath\langle\mu, u\rangle-\frac{1}{2}\langle u, A u\rangle+\int_{\mathbb{R}^{n}}\left(e^{\imath\langle u, x\rangle}-1-\imath\langle u, x\rangle\right) d \nu(x)\right\} \tag{4.7}
\end{equation*}
$$

for $u \in \mathbb{R}^{n}$, where $\mu$ is the $\mathbf{Q}$-expectation of $\xi$. Replicating the proof of Theorem 4.1 (or by adjusting $\gamma$ and using $\left.d \nu(x)=e^{-x_{i}} d \nu\left(K_{i} x\right)\right)$ we obtain that $\xi \in \operatorname{ESD}_{i}$ if and only if conditions (1) and (2) of Theorem 4.1 hold, while (4.5) is replaced by

$$
\begin{equation*}
\mu_{i}=\int_{\mathbb{R}^{n}} x_{i}\left(1-e^{\frac{1}{2} x_{i}}\right) d \nu(x)-\frac{1}{2} a_{i i} . \tag{4.8}
\end{equation*}
$$

Example 4.5 (log-normal distribution, Black-Scholes setting). Assume that $\eta$ is $\log$-normal with underlying normal vector $\xi=\log \eta$, so that

$$
\varphi_{\xi}^{\mathbf{Q}}(u)=\exp \left\{\imath\langle\mu, u\rangle-\frac{1}{2}\langle u, A u\rangle\right\}, \quad u \in \mathbb{R}^{n}
$$

Then $\eta \in \mathrm{SD}_{i}$ if and only if the covariance matrix $A=\left(a_{l m}\right)_{l m=1}^{n}$ satisfies $a_{l i}=a_{i l}=\frac{1}{2} a_{i i}$ for $l=1, \ldots, n, l \neq i$, and $\mu_{i}=-\frac{1}{2} a_{i i}$; see (4.8). For this, it is necessary (but not sufficient) that the $i$ th component of $\eta$ be sampled from a $\mathbf{Q}$-martingale.

Finally, $\eta$ is jointly self-dual if and only if $a_{l l}=\sigma^{2}$ for all $l=1, \ldots, n, a_{l m}=\frac{1}{2} \sigma^{2}$ for all $l \neq m$, i.e., for $\sigma>0$ the correlations between the components of $\xi$ are $\frac{1}{2}$, and the mean is $-\frac{\sigma^{2}}{2}(1, \ldots, 1)$.

Remark 4.6 (square integrable case and covariance). As a consequence of Example 4.5, for log-normal $\eta=e^{\xi} \in \mathrm{SD}_{i}$ the correlations between the $i$ th and other components of $\xi$ are $\frac{1}{2} \sqrt{a_{i i} / a_{l l}}$ (assuming $a_{i i}, a_{l l}>0$ ), while other correlations are not affected. In order to relax
this correlation structure between the $i$ th and other components, it is useful to introduce a jump component. For doing that, we assume $\int_{\|x\|>1}\|x\|^{2} d \nu(x)<\infty$; i.e., $\xi$ is square integrable. Then the elements of the covariance matrix of $\xi$ are given by

$$
v_{l j}=a_{l j}+\int_{\mathbb{R}^{n}} x_{l} x_{j} d \nu(x)
$$

(see [34, Ex. 25.12]); i.e., despite the constraints on the Lévy measure given in (4.4) there are various possibilities for the covariance and correlation structures. Simple examples can be constructed as in the following remark; see also Remark 5.6 and Example 5.10 (for $\alpha=1$ ).

Remark 4.7 (Lévy measures). Assume that $\xi \in \mathrm{ESD}_{i}$ is infinitely divisible with the Lévy measure $\nu$. If $\nu$ is finite, then the second condition of Theorem 4.1 means that random vector $\zeta$ distributed according to the normalized $\nu$ is itself $\mathrm{ESD}_{i}$. In particular, if $\nu$ is absolutely continuous, its density satisfies (2.5). An immediate example of $\nu$ is the Gaussian law with the mean and positive definite covariance matrix from Example 4.5, so that $e^{\zeta}$ is log-normally distributed as in Example 4.5. Since this $\nu$ is finite, the non-Gaussian part of $\xi$ corresponds to the compound Poisson law with Gaussian jumps.
5. Quasi-self-dual vectors. As we have seen, the symmetry properties of random price changes (interpretation of $\eta$ in a risk-neutral case) are considered separately from the forward prices of the assets. In some cases, notably for semistatic hedging of barrier options with carrying costs (see [9, 12]), the symmetry is imposed on price changes adjusted with carrying costs $a=e^{\lambda} \in \mathbb{R}^{n}$, where usually $\lambda_{j}=r-q_{j}$ for the risk-free interest rate $r$ and the dividend yield $q_{j}$ of the $j$ th asset, $j=1, \ldots, n$.

In view of applications to derivative pricing, it is natural to assume that all components of $\eta$ have expectation one, which is the case if $\mathbf{Q}$ is a martingale measure. Then the random vector

$$
e^{\lambda} \circ \eta=\left(e^{\lambda_{1}} \eta_{1}, \ldots, e^{\lambda_{n}} \eta_{n}\right)
$$

cannot be self-dual with respect to the $i$ th numeraire (resp., for all numeraires) unless $\lambda_{i}=0$ (resp., all components of $\lambda$ vanish), since the multiplication by $e^{\lambda}$ moves the expectation away from one. One can, however, relate $e^{\lambda} \circ \eta$ to a self-dual random vector by means of a power transformation.

Definition 5.1. A random vector $\eta \in \mathbb{E}^{n}$ is said to be quasi-self-dual (of order $\alpha$ ) if there exist $\lambda \in \mathbb{R}^{n}$ and $\alpha \neq 0$ such that $\left(e^{\lambda} \circ \eta\right)^{\alpha}$ is integrable and self-dual with respect to the $i t h$ numeraire. We then write $\eta \in \operatorname{QSD}_{i}(\lambda, \alpha)$.

If $\eta \in \operatorname{QSD}_{i}(\lambda, \alpha)$, then $\mathbf{E}\left(e^{\lambda_{i}} \eta_{i}\right)^{\alpha}=1$ by Lemma 2.9 , so that the values of $\alpha$ and $\lambda_{i}$ are closely related to each other. Later in this section, we discuss this relation for a special case of quasi-self-dual Lévy models. If useful, $\lambda$ can also have interpretations other than being the pure carrying costs, and one can also drop the assumption that $\eta$ stems from a martingale. If imposed, the martingale assumption will be mentioned explicitly.

By Theorem 2.4(iii), $\eta \in \mathrm{QSD}_{i}(\lambda, \alpha)$ yields that

$$
\begin{equation*}
\mathbf{E} f\left(e^{\lambda} \circ \eta\right)=\mathbf{E} f\left(\left(\left(e^{\lambda} \circ \eta\right)^{\alpha}\right)^{\frac{1}{\alpha}}\right)=\mathbf{E}\left[f\left(\varkappa_{i}\left(e^{\lambda} \circ \eta\right)\right)\left(a_{i} \eta_{i}\right)^{\alpha}\right] \tag{5.1}
\end{equation*}
$$

Define random vector $\zeta=\lambda+\xi$, where $\eta=e^{\xi}$ for $\xi=\left(\xi_{1}, \ldots, \xi_{n}\right)$. Then $e^{\lambda} \circ \eta=e^{\zeta}$. If we consider the payoff function as a function of asset prices $S_{T}=\left(S_{T 1}, \ldots, S_{T n}\right)$ with
$S_{T j}=S_{0 j} e^{\zeta_{j}}$ for $j=1, \ldots, n$, then (5.1) can be written as

$$
\mathbf{E} f\left(S_{T}\right)=\mathbf{E}\left[f\left(\frac{S_{0 i}}{S_{T i}}\left(S_{T 1}, \ldots, S_{T(i-1)}, S_{0 i}, S_{T(i+1)}, \ldots, S_{T n}\right)\right)\left(\frac{S_{T i}}{S_{0 i}}\right)^{\alpha}\right]
$$

Fix an asset number $i \in\{1, \ldots, n\}$ and assume now that $\mathbf{Q}$ is a probability measure such that $\eta \in \operatorname{QSD}_{i}(\lambda, \alpha)$. Since $\eta^{\alpha}$ is positive integrable, $0<\mathbf{E} e^{\frac{\alpha}{2} \zeta_{i}}<\infty$.

Hence, we can define probability measure $\tilde{\mathcal{E}}^{i}$ by

$$
\frac{d \tilde{\mathcal{E}}^{i}}{d \mathbf{Q}}=\frac{e^{\frac{\alpha}{2} \zeta_{i}}}{\mathbf{E} e^{\frac{\alpha}{2} \zeta_{i}}}, \quad \frac{d \mathbf{Q}}{d \tilde{\mathcal{L}}^{i}}=\frac{e^{-\frac{\alpha}{2} \zeta_{i}}}{\mathbf{E}_{\tilde{\mathcal{E}}^{i}} e^{-\frac{\alpha}{2} \zeta_{i}}},
$$

i.e., the Esscher transform of $\mathbf{Q}$ with parameter $\frac{\alpha}{2} e_{i}$ and the corresponding inverse transform.

It is obvious that $\eta \in \operatorname{QSD}_{i}(\lambda, \alpha)$ is equivalent to any of the conditions of Theorem 2.4 for $\left(e^{\lambda} \circ \eta\right)^{\alpha}=e^{\alpha \zeta}$. The following theorem yields a more direct characterization.

Theorem 5.2. Let $\eta^{\alpha}$ be integrable for some $\alpha \neq 0$. Then $\eta \in \operatorname{QSD}_{i}(\lambda, \alpha)$ is equivalent to one of the following conditions for $\zeta$ defined from $e^{\zeta}=e^{\lambda} \circ \eta=e^{\lambda+\xi}$ :
(i) For any payoff function $f: \mathbb{R}_{+}^{n} \mapsto \mathbb{R}_{+}$,

$$
\begin{equation*}
\mathbf{E} f\left(e^{\zeta}\right)=\mathbf{E}\left[f\left(e^{K_{i} \zeta}\right) e^{\alpha \zeta_{i}}\right] \tag{5.2}
\end{equation*}
$$

(ii) The distributions of $\zeta$ and $K_{i} \zeta$ under $\tilde{\mathcal{E}}^{i}$ coincide.
(iii) For every $u \in \mathbb{R}^{n}$,

$$
\varphi_{\zeta}^{\tilde{\mathcal{E}}^{i}}(u)=\varphi_{\zeta}^{\tilde{\mathcal{E}}^{i}}\left(K_{i}^{\top} u\right)
$$

or, equivalently,

$$
\varphi_{\xi}^{\mathbf{Q}}\left(u-\frac{\alpha}{2} \boldsymbol{\imath} e_{i}\right)=\varphi_{\xi}^{\mathbf{Q}}\left(K_{i}^{\top} u-\frac{\alpha}{2} \boldsymbol{\imath} e_{i}\right) e^{-\imath \lambda_{i}\left(\sum_{l=1}^{n} u_{l}+u_{i}\right)} .
$$

Moreover, if additionally $\eta$ is integrable, we have that $\eta \in \operatorname{QSD}_{i}(\lambda, \alpha)$ if and only if (5.2) holds for $f$ being payoffs from basket options with arbitrary strikes and weights of assets.

Note also that all conditions of Theorem 5.2 can be written conditionally on a fixed event or conditionally on a $\sigma$-algebra; cf. Remark 2.7. The joint quasi-self-duality can be achieved by raising the components of $\eta$ to different powers.

Proof of Theorem 5.2. For (i) it suffices to note that $\eta$ is quasi-self-dual if and only if $\alpha \zeta \in \mathrm{ESD}_{i}$ and to refer to (5.1) and Theorem 2.4(iii).

Replace $\eta$ by $e^{\lambda} \circ \eta, \mathbf{E}\left[f\left(\varkappa_{i}(\eta)\right) \eta_{i}\right]$ by $\mathbf{E}\left[f\left(\varkappa_{i}\left(e^{\lambda} \circ \eta\right)\right)\left(e^{\lambda_{i}} \eta_{i}\right)^{\alpha}\right], \mathcal{E}^{i}$ by $\tilde{\mathcal{E}}^{i}, \xi$ by $\zeta$, and $\frac{1}{2}$ by $\frac{\alpha}{2}$ in the proof of the equivalence (iii) $\Leftrightarrow$ (v) in Theorem 2.4 to see that (i) is equivalent to (ii). A similar argument yields the equivalence of (ii) and $\varphi_{\zeta}^{\tilde{\mathcal{E}}^{i}}(u)=\varphi_{\zeta}^{\tilde{\mathcal{E}}^{i}}\left(K_{i}^{\top} u\right)$ for all $u \in \mathbb{R}^{n}$ as well as the equivalence of this equation with

$$
\begin{equation*}
\varphi_{\zeta}^{\mathbf{Q}}\left(u-\frac{\alpha}{2} \imath e_{i}\right)=\varphi_{\zeta}^{\mathbf{Q}}\left(K_{i}^{\top} u-\frac{\alpha}{2} \imath e_{i}\right) \tag{5.3}
\end{equation*}
$$

for all $u \in \mathbb{R}^{n}$. Writing the characteristic functions as $\mathbf{Q}$-expectations and using that $\zeta=\lambda+\xi$ yields that

$$
\begin{aligned}
\mathbf{E} \exp \left\{\boldsymbol{\imath}\left\langle u-\frac{\alpha}{2} \imath e_{i}, \xi\right\rangle\right. & \left.+\boldsymbol{\imath}\left\langle u-\frac{\alpha}{2} \boldsymbol{\imath} e_{i}, \lambda\right\rangle\right\} \\
& =\mathbf{E} \exp \left\{\boldsymbol{\imath}\left\langle K_{i}^{\top} u-\frac{\alpha}{2} \boldsymbol{\imath} e_{i}, \xi\right\rangle+\boldsymbol{\imath}\left\langle K_{i}^{\top} u-\frac{\alpha}{2} \boldsymbol{\imath} e_{i}, \lambda\right\rangle\right\}
\end{aligned}
$$

Dividing by $\exp \left\{\boldsymbol{\imath}\left\langle u-\frac{\alpha}{2} \imath e_{i}, \lambda\right\rangle\right\}$ yields the equivalence of the second statement in (iii) and (5.3).
If (5.2) holds for $f$ being payoffs from basket options with arbitrary strikes and weights of assets and integrable $\eta=e^{\zeta-\lambda}$, we first have that $\mathbf{E} e^{\alpha \zeta_{i}}=1$ by letting the strike be one and other weights vanish. Hence, we can define the measure $\mathbf{P}$ by

$$
\frac{d \mathbf{P}}{d \mathbf{Q}}=e^{\alpha \zeta_{i}}
$$

so that

$$
\mathbf{E}\left(u_{0}+\left\langle u, e^{\zeta}\right\rangle\right)_{+}=\mathbf{E}\left[\left(u_{0}+\left\langle u, \varkappa_{i}\left(e^{\zeta}\right)\right\rangle\right)_{+} e^{\alpha \zeta_{i}}\right]=\mathbf{E}_{\mathbf{P}}\left(u_{0}+\left\langle u, \varkappa_{i}\left(e^{\zeta}\right)\right\rangle\right)_{+}
$$

for every $\left(u_{0}, u\right) \in \mathbb{R}^{n+1}$; i.e., by Theorem 2.1(i), $e^{\zeta}$ under $\mathbf{Q}$ and $\varkappa_{i}\left(e^{\zeta}\right)$ under $\mathbf{P}$ share the same distribution. Hence, for every payoff function we have

$$
\mathbf{E} f\left(e^{\zeta}\right)=\mathbf{E}_{\mathbf{P}} f\left(\varkappa_{i}\left(e^{\zeta}\right)\right)=\mathbf{E}\left[f\left(\varkappa_{i}\left(e^{\zeta}\right)\right) e^{\alpha \zeta_{i}}\right] ;
$$

i.e., we arrive at (5.2). The other implication is obvious.

We now use Theorem 5.2 to characterize all quasi-self-dual $\eta$ such that $\xi=\log \eta$ is infinitely divisible with the Lévy-Khintchine representation (4.2).

Theorem 5.3. Let the random vector $\xi=\log \eta$ be infinitely divisible under $\mathbf{Q}$ with the generating triplet $(A, \nu, \gamma)$ and let $\eta^{\alpha}$ be integrable for some $\alpha \neq 0$. Then $\eta \in \operatorname{QSD}_{i}(\lambda, \alpha)$ if and only if the following three conditions hold:
(1) The matrix $A=\left(a_{l j}\right)_{l j=1}^{n}$ satisfies $a_{i j}=a_{j i}=\frac{1}{2} a_{i i}$ for all $j=1, \ldots, n, j \neq i$.
(2) The Lévy measure satisfies

$$
\begin{equation*}
d \nu(x)=e^{-\alpha x_{i}} d \nu\left(K_{i} x\right) \quad \text { almost everywhere }, \tag{5.4}
\end{equation*}
$$

meaning that $\nu(B)=\int_{K_{i} B} e^{\alpha x_{i}} d \nu(x)$ for all Borel B.
(3) The ith coordinate of $\gamma$ satisfies

$$
\begin{equation*}
\gamma_{i}=\int_{\|x\| \leq 1} x_{i}\left(1-e^{\frac{\alpha}{2} x_{i}}\right) d \nu(x)-\frac{\alpha}{2} a_{i i}-\lambda_{i} . \tag{5.5}
\end{equation*}
$$

Proof. Denote $\zeta=\lambda+\xi$. Since $0<\mathbf{E} e^{\frac{\alpha}{2} \zeta_{i}}<\infty$, the Esscher transform $\tilde{\mathcal{E}}^{i}$ of $\mathbf{Q}$ with parameter $\frac{\alpha}{2} e_{i}$ and the inverse transform are well defined. Therefore, $\zeta$ under $\tilde{\mathcal{E}}^{i}$ also has an infinitely divisible distribution. By using Theorem 5.2(iii) instead of Theorem 2.4(vi) and replacing $\mathcal{E}^{i}$ by $\tilde{\mathcal{E}}^{i}, \xi$ by $\zeta$, and $\frac{1}{2}$ by $\frac{\alpha}{2}$ in the proof of Theorem 4.1, we obtain (1), (2), and

$$
\gamma_{i}=\int_{\|x\| \leq 1} x_{i}\left(1-e^{\frac{\alpha}{2} x_{i}}\right) d \nu(x)-\frac{\alpha}{2} a_{i i}
$$

for the generating triplet of $\zeta$ under $\mathbf{Q}$. Since $\xi=\zeta-\lambda$, we need only adjust $\gamma_{i}$ by $-\lambda_{i}$ to finish the proof of the first implication.

The integrability of $\eta^{\alpha}$ implies the existence of the Esscher transform of $\mathbf{Q}$ with parameter $\frac{\alpha}{2} e_{i}$. By doing this transform and the converse calculations, it is easy to verify that Theorem 5.2 (iii) applies, i.e., $\eta=e^{\xi} \in \operatorname{QSD}_{i}(\lambda, \alpha)$.

Note that condition (1) is identical to Theorem 4.1(1). As a consequence of Theorem 4.1(2) and $5.3(2)$, we immediately get the following results.

Corollary 5.4. Let the random vector $\xi=\log \eta$ be infinitely divisible under $\mathbf{Q}$ with nonvanishing Lévy measure $\nu$. Then $\eta$ cannot be quasi-self-dual of two different orders with respect to the same numeraire.

Corollary 5.5. If $\left(\xi_{t}\right)_{t \geq 0}$ is the Lévy process with generating triplet $(A, \nu, \gamma)$ that satisfies the conditions of Theorem 5.3, then $e^{\xi_{t}} \in \operatorname{QSD}_{i}(\lambda t, \alpha)$ for all $t \geq 0$.

Proof. It suffices to note that $\varphi_{\xi_{t}}^{\mathbf{Q}}(u)=\left(\varphi_{\xi_{1}}^{\mathbf{Q}}(u)\right)^{t}$ for all $t \geq 0$ and to raise the corresponding identity from Theorem 5.2 (iii) into power $t$.

Remark 5.6 (Lévy measures in the quasi-self-dual case). In order to construct a Lévy measure $\nu$ satisfying (5.4), note that

$$
e^{\frac{\alpha}{2} x_{i}} d \nu(x)=e^{\frac{\alpha}{2}\left(K_{i} x\right)_{i}} d \nu\left(K_{i} x\right),
$$

meaning that the measure $\nu_{0}$ with density $\frac{d \nu_{0}}{d \nu}(x)=e^{\frac{\alpha}{2} x_{i}}$ is $K_{i}$-invariant. Therefore, in the background one always needs to have a $K_{i}$-invariant Lévy measure.

Since the Lebesgue measure on $\mathbb{R}^{n}$ is $K_{i}$-invariant, a simple example of $\nu_{0}$ is provided by the Lebesgue measure restricted onto $B_{R}$, where $B_{R}=\{x:\|x\| \leq R\}$ is the ball of radius $R$ in the $\|\cdot\| / \|$-norm. A further implication of the $K_{i}$-invariance property of the Lebesgue measure on $\mathbb{R}^{n}$ is that the Lebesgue density $p_{\nu_{0}}$ of an absolutely continuous $K_{i}$-invariant measure $\nu_{0}$ is also $K_{i}$-invariant; i.e., $p_{\nu_{0}}(x)=p_{\nu_{0}}\left(K_{i} x\right)$ for almost every $x \in \mathbb{R}^{n}$. Then (5.4) can be equivalently written as $p_{\nu}(x)=e^{-\alpha x_{i}} p_{\nu}\left(K_{i} x\right)$. Clearly, condition (4.3) is always satisfied for a finite $\nu$ without an atom at the origin, which then yields the compound Poisson part of $\xi$ from $\eta=e^{\xi} \in \operatorname{QSD}_{i}(\lambda, \alpha)$. The integrability condition on $\eta^{\alpha}$ additionally requires that

$$
\int_{\|x\|>1} e^{\alpha x_{j}} e^{-\frac{\alpha}{2} x_{i}} d \nu_{0}(x)<\infty, \quad j=1, \ldots, n
$$

see [34, Thm. 25.17].
Remark 5.7 (determining $\alpha$ from the carrying costs in the risk-neutral case). Assume that $\mathbf{E} \eta_{j}=1$ for all $j=1, \ldots, n$ and $\eta \in \operatorname{QSD}_{i}(\lambda, \alpha)$ with given $\lambda$. Since $\varphi_{\xi}^{\mathbf{Q}}\left(-\boldsymbol{\imath} e_{j}\right)=\mathbf{E} \eta_{j}=1$, we see that

$$
\begin{equation*}
\gamma_{j}=-\int_{\mathbb{R}^{n}}\left(e^{x_{j}}-1-x_{j} \mathbb{\mathbb { 1 }}\|x\| \leq 1\right) d \nu(x)-\frac{1}{2} a_{j j}, \quad j=1, \ldots, n \tag{5.6}
\end{equation*}
$$

If $\alpha=1$, then the above condition for $j=i$ yields (4.5) (or (5.5) for $\alpha=1$ and $\lambda=0$ ). Indeed, it suffices to check that

$$
\begin{aligned}
& \int_{\mathbb{R}^{n}}\left(1-e^{x_{i}}+x_{i} e^{\frac{1}{2} x_{i} \mathbb{1}\|x\| \leq 1}\right) d \nu(x) \\
& \quad=\int_{\left\{x_{i}<0\right\}}\left(1-e^{x_{i}}+x_{i} e^{\frac{1}{2} x_{i} \mathbb{1}\|x\| \leq 1}\right) e^{-x_{i}} d \nu\left(K_{i} x\right)+\int_{\left\{x_{i}>0\right\}}\left(1-e^{x_{i}}+x_{i} e^{\frac{1}{2} x_{i} \mathbb{1}\|x\| \leq 1}\right) d \nu(x) \\
& \quad=\int_{\left\{y_{i}>0\right\}}\left(e^{y_{i}}-1-y_{i} e^{\frac{1}{2} y_{i}} \mathbb{\mathbb { H }}\|y\| \leq 1\right) d \nu(y)+\int_{\left\{x_{i}>0\right\}}\left(1-e^{x_{i}}+x_{i} e^{\frac{1}{2} x_{i}} \mathbb{1}_{\|x\| \leq 1}\right) d \nu(x)=0 .
\end{aligned}
$$

However, for nonvanishing $\lambda$ we need to combine (5.6) with (5.5) to see that $\alpha$ must satisfy

$$
-\int_{\mathbb{R}^{n}}\left(e^{x_{i}}-1-x_{i} \mathbb{1}\|x\| \leq 1\right) d \nu(x)-\frac{1}{2} a_{i i}=\int_{\|x\| \leq 1} x_{i}\left(1-e^{\frac{\alpha}{2} x_{i}}\right) d \nu(x)-\frac{\alpha}{2} a_{i i}-\lambda_{i}
$$

or, equivalently,

$$
\begin{equation*}
a_{i i} \alpha=a_{i i}-2 \lambda_{i}+2 \int_{\mathbb{R}^{n}}\left(e^{x_{i}}-1-x_{i} e^{\frac{\alpha}{2} x_{i}} \mathbb{1}_{\|x\| \leq 1}\right) d \nu(x) . \tag{5.7}
\end{equation*}
$$

It should be noted that in the Lévy process setting from Corollary 5.5 the values of $\alpha$ calculated for all $t \geq 0$ coincide.

Remark 5.8 (finite mean case). Assume that $\eta=e^{\xi} \in \operatorname{QSD}_{i}(\lambda, \alpha)$. If, as in Remark 4.4, $\xi$ has a finite mean, then (5.5) is replaced by

$$
\begin{equation*}
\mu_{i}=\int_{\mathbb{R}^{n}} x_{i}\left(1-e^{\frac{\alpha}{2} x_{i}}\right) d \nu(x)-\frac{\alpha}{2} a_{i i}-\lambda_{i}, \tag{5.8}
\end{equation*}
$$

where $\mu$ is the expectation of $\xi$. If $\mathbf{Q}$ is a martingale measure for $\eta_{i}$, then $\varphi_{\xi}^{\mathbf{Q}}\left(-\boldsymbol{\imath} e_{i}\right)=\mathbf{E} e^{\xi_{i}}=1$ yields that

$$
\begin{equation*}
\mu_{i}=-\int_{\mathbb{R}^{n}}\left(e^{x_{i}}-1-x_{i}\right) d \nu(x)-\frac{1}{2} a_{i i} . \tag{5.9}
\end{equation*}
$$

Combining (5.8) with (5.9) yields

$$
\begin{align*}
a_{i i} \alpha & =a_{i i}-2 \lambda_{i}+2 \int_{\mathbb{R}^{n}}\left(e^{x_{i}}-1-x_{i} e^{\frac{\alpha}{2} x_{i}}\right) d \nu(x) \\
& =a_{i i}-2 \lambda_{i}+2 \int_{\mathbb{R}}\left(e^{x_{i}}-1-x_{i} e^{\frac{\alpha}{2} x_{i}}\right) d \nu_{i}\left(x_{i}\right), \tag{5.10}
\end{align*}
$$

where $\nu_{i}$ is the marginal Lévy measure defined by $\nu_{i}(B)=\nu\left(\left\{x \in \mathbb{R}^{n}: x_{i} \in B\right\}\right)$ for Borel $B \subset \mathbb{R}, 0 \notin B$; see [34, Prop. 11.10].

Compared to (5.7), (5.10) yields a considerable simplification in calculating $\alpha$. Since $\nu_{i}$ is the Lévy measure corresponding to $\eta_{i}$, it is possible to calculate $\alpha$ from only the distribution of the $i$ th component of $\eta$ and the corresponding carrying costs $\lambda_{i}$.

In the purely non-Gaussian case (i.e., if $A$ vanishes) it is useful to write the integral in (5.10) as its principal value. Recall that the principal value of an integral is defined as

$$
f_{\mathbb{R}} g(x) d \nu(x)=\lim _{R \rightarrow+\infty} \int_{-R}^{R} g(x) d \nu(x) .
$$

Then the principal value of the integral of $x_{i} e^{\frac{\alpha}{2} x_{i}}$ vanishes, since $d \nu_{i}\left(x_{i}\right)=e^{-\frac{\alpha}{2} x_{i}} d \nu_{0 i}\left(x_{i}\right)$ for a symmetric measure $\nu_{0 i}$, and

$$
\begin{aligned}
\lambda_{i} & =f_{\mathbb{R}}\left(e^{x_{i}}-1\right) d \nu_{i}\left(x_{i}\right)=f_{\mathbb{R}}\left(e^{x_{i}}-1\right) e^{-\frac{\alpha}{2} x_{i}} d \nu_{0 i}\left(x_{i}\right) \\
& =f_{\mathbb{R}}\left(e^{\left(1-\frac{\alpha}{2}\right) x_{i}}-e^{-\frac{\alpha}{2} x_{i}}\right) d \nu_{0 i}\left(x_{i}\right)
\end{aligned}
$$

If $\nu_{0 i}$ has a finite Laplace transform $\psi$ on the real line, then $\alpha$ solves

$$
\lambda_{i}=\psi\left(1-\frac{\alpha}{2}\right)-\psi\left(-\frac{\alpha}{2}\right)
$$

Example 5.9 (log-normal model with carrying costs). By Corollary 5.4, among all of the loginfinitely divisible distributions only the log-normal one can be quasi-self-dual of two orders with respect to the same numeraire. Applying (5.7) for the univariate log-normal case with $a_{i i}=\sigma^{2}>0($ and vanishing $\nu)$ yields that

$$
\alpha=1-\frac{2 \lambda}{\sigma^{2}}
$$

as stated in [9, 12]. Hence, the univariate log-normal distribution in the Black-Scholes setting is self-dual and quasi-self-dual of order $1-\frac{2 \lambda}{\sigma^{2}}$ at the same time. By (5.7), this is also true for multivariate log-normal models from Example 4.5 being self-dual with respect to the $i$ th numeraire; i.e., this distribution is at the same time quasi-self-dual of order $\alpha=1-2 \lambda_{i} / a_{i i}$ with respect to the same numeraire.

Example 5.10 (determining $\boldsymbol{\alpha}$ for nontrivial Lévy measures). Start with the univariate case (i.e., $n=1$ ) and choose $\nu_{0}$ from Remark 5.6 to be the centered Gaussian measure with variance $\beta^{2}>0$. If normalized to have the total mass one, $\nu$ becomes the density of the normal law with mean $-\frac{\alpha \beta^{2}}{2}$ and variance $\beta^{2}$. Solving (5.7) or, equivalently, (5.10) for this particular measure $\nu$ and $a_{i i}=\sigma^{2}>0$ yields that

$$
\alpha=\frac{1}{\beta^{2} \sigma^{2}}\left(2 \text { LambertW }\left(\frac{\beta^{2}}{\sigma^{2}} \exp \left\{\frac{\beta^{2}(\lambda+1)}{\sigma^{2}}\right\}\right) \sigma^{2}+\beta^{2} \sigma^{2}-2 \beta^{2} \lambda-2 \beta^{2}\right)
$$

where LambertW $(x)=g(x)$ is the principal branch of the LambertW function that satisfies $g(x) e^{g(x)}=x$ for all $x$. In the purely non-Gaussian case the required power is given by

$$
\alpha=1-\frac{2}{\beta^{2}} \log (1+\lambda)
$$

In the multivariate case, we start with $\nu_{0}$ being the centered Gaussian law having positive definite covariance matrix $B$ that satisfies Theorem $5.3(1)$ for some fixed $i$ and define measure $\nu$ with density

$$
\frac{d \nu}{d \nu_{0}}(x)=e^{-\frac{\alpha}{2} x_{i}}
$$

Then (5.4) holds and the $i$ th marginal $\nu_{i}$ of $\nu$ has the density $e^{-\frac{\alpha}{2} x_{i}}$ with respect to the $i$ th marginal of $\nu_{0}$, the latter being the centered Gaussian law with variance $\beta^{2}=b_{i i}>0$. Since the $i$ th marginal for the normalized $\nu$ coincides with the Lévy measure constructed above in the univariate case, we obtain the same $\alpha$ as in the univariate case with $\beta=\sqrt{b_{i i}}$.

## 6. Distributions of self-dual random variables.

6.1. Characterization and examples. In this section, we specialize the results from section 2 for self-dual random variables. Denote by $\bar{F}(x)=\mathbf{P}(\eta>x)$ the tail of the cumulative distribution function of a positive random variable $\eta$ and by

$$
\bar{F}_{I}(z)=\int_{0}^{z} \bar{F}(t) d t, \quad z \geq 0,
$$

the integrated tail. Note that $\bar{F}_{I}(0)=0$ and $\bar{F}_{I}(\infty)=1$ in the case $\mathbf{E} \eta=1$.
Theorem 6.1. An integrable positive random variable $\eta$ is self-dual if and only if $\bar{F}_{I}(\infty)=1$ and

$$
\begin{equation*}
z \bar{F}_{I}\left(z^{-1}\right)=\bar{F}_{I}(z) \quad \text { for all } z>0 \tag{6.1}
\end{equation*}
$$

Proof. It is easy to check that

$$
\bar{F}_{I}(z)=\mathbf{E} \min (\eta, z), \quad z \geq 0
$$

Now apply Theorem 1.2 or Theorem $2.4($ iii ) to the payoff function $f(\eta)=\min (\eta, z)$ to see that

$$
\bar{F}_{I}(z)=\mathbf{E} \min (\eta, z)=\mathbf{E}\left[\min \left(\eta^{-1}, z\right) \eta\right]=\mathbf{E} \min (1, z \eta)=z \bar{F}_{I}\left(z^{-1}\right) .
$$

In the opposite direction, (6.1) yields that

$$
\mathbf{E} \max (\eta, z)=\mathbf{E}[\eta+z-\min (\eta, z)]=\mathbf{E}[1+z \eta-\min (1, z \eta)]=\mathbf{E} \max (1, z \eta) ;
$$

i.e., by rescaling we arrive at the univariate case of the self-duality property given in Theorem 2.4(ii).

For $\eta$ being a positive, integrable, and absolutely continuous random variable with probability density $p_{\eta}$, Theorem 2.8 yields that $\eta$ is self-dual if and only if

$$
\begin{equation*}
p_{\eta}(x)=x^{-3} p_{\eta}\left(x^{-1}\right) \quad \text { for almost all } x>0 . \tag{6.2}
\end{equation*}
$$

If $\xi=\log \eta$, the self-duality of $\eta$ (i.e., the exponential self-duality of $\xi$ ) is equivalent to $p_{\xi}(x)=e^{-x} p_{\xi}(-x)$ for almost all $x \in \mathbb{R}$; i.e., $e^{\frac{1}{2} x} p_{\xi}(x)$ is an even function. For instance, the probability density of the log-normal distribution of mean one satisfies (6.2). It is also satisfied by mixtures of log-normal densities that appear in the (uncorrelated) Hull-White stochastic volatility model; see [21, Thm. 3.1]. The self-duality property of stochastic volatility models is explored in [12, Thm. 3.1].

Example 6.2 (log-normal model). If $S_{T}=F \eta$ has the log-normal distribution, the BlackScholes formula yields that

$$
\begin{equation*}
\mathbf{E} \max (F \eta, k)=F \Phi\left(d^{\prime}\right)+k \Phi\left(d^{\prime \prime}\right), \tag{6.3}
\end{equation*}
$$

where $k, F>0$,

$$
d^{\prime}=\lambda+\frac{1}{2 \lambda} \log \frac{F}{k}, \quad d^{\prime \prime}=\lambda+\frac{1}{2 \lambda} \log \frac{k}{F}, \quad \lambda=\frac{1}{2} \sigma \sqrt{T},
$$

and $\Phi$ is the cumulative distribution function for the standard normal variable. Note that the conventional Black-Scholes formula is obtained by subtracting $k$ from (6.3) and then discounting. By looking at the right-hand side of (6.3), it is easy to see that it is symmetric with respect to $F$ and $k$; i.e., $\eta$ is a self-dual random variable.

The right-hand side of (6.3) defines a (symmetric) norm on $\mathbb{R}_{+}^{2}$ called the Hüsler-Reiss norm of $x=(k, F)$; see [28]. Thus, the derivative given by the maximum of the asset price and the strike has the price given by the discounted norm of the vector composed of the forward and the strike. Notably, expression (6.3) appears in the literature on extreme values (see [23]) as the limit distribution of coordinatewise maxima for triangular arrays of bivariate Gaussian vectors with correlation $\varrho(n)$ that approaches one with rate $(1-\varrho(n)) \log n \rightarrow \lambda^{2} \in[0, \infty]$ as $n \rightarrow \infty$.

In order to construct further examples of probability density functions $p_{\eta}$ that satisfy (6.2), it suffices to define $p_{\eta}(x)$ for $x \geq 1$ and then extend it for $x \in(0,1)$ using (6.2) with a subsequent normalization to ensure that the total mass is one. Clearly, one has to bear in mind that $\mathbf{E} \eta=1$ presumes the integrability of $x p_{\eta}(x)$ (alongside with $p_{\eta}(x)$ itself) at zero and infinity.

Example 6.3 (self-dual random variables with heavy tails). The log-normal distribution has a light tail at infinity. It is possible to construct a self-dual heavy-tailed distribution by setting

$$
p(x)= \begin{cases}c_{\gamma} x^{\gamma} & \text { if } x \in(0,1]  \tag{6.4}\\ c_{\gamma} x^{-(3+\gamma)} & \text { if } x>1\end{cases}
$$

for $\gamma>-1$, where $c_{\gamma}=(1+\gamma)(2+\gamma) /(3+2 \gamma)$ normalizes the probability density.
Example 6.4 (discrete self-dual random variable). If $\eta$ takes values $\frac{1}{2}, 1,2$ with probabilities $\frac{1}{3}, \frac{1}{2}, \frac{1}{6}$, then Theorem 2.8(b) implies that $\eta$ is self-dual.

Example 6.5. Consider the $\ell_{p}$-norm on $\mathbb{R}^{2}$, which is clearly symmetric. Evaluating the $\ell_{p}$-norm of $(t, 1)$, we arrive at

$$
\left(t^{p}+1\right)^{1 / p}=\mathbf{E} \max (t, \eta)=t \mathbf{P}(\eta \leq t)+\int_{t}^{\infty} x p_{\eta}(x) d x, \quad t>0
$$

assuming that $\eta$ is absolutely continuous with density $p_{\eta}$. Differentiating with respect to $t$ yields that

$$
\mathbf{P}(\eta \leq t)=t^{p-1}\left(t^{p}+1\right)^{1 / p-1}
$$

The density of $\eta$ satisfies (6.2).
6.2. Moments of self-dual random variables. It is immediate that all self-dual random variables have expectation one. Carr and Lee [12, Cor. 2.11] show that

$$
\begin{equation*}
\mathbf{E} \eta^{n}=\mathbf{E} \eta^{-n+1}, \quad n \geq 1 \tag{6.5}
\end{equation*}
$$

In particular, if $\mathbf{E} \eta^{2}<\infty$, then

$$
\operatorname{Cov}\left(\eta^{-1}, \eta\right)=1-\mathbf{E} \eta^{-1}=(\mathbf{E} \eta)^{2}-\mathbf{E} \eta^{2}=-\operatorname{Var}(\eta) .
$$

Theorem 6.6. Each nontrivial self-dual variable $\eta$ with a finite third moment has a positive skewness $\mathbf{E}(\eta-\mathbf{E} \eta)^{3} /(\operatorname{Var}(\eta))^{3 / 2}$.

Proof. In view of (6.5),

$$
\begin{aligned}
\mathbf{E}(\eta-\mathbf{E} \eta)^{3} & =\mathbf{E}\left(\eta^{3}-3 \mathbf{E} \eta^{2}+2\right) \\
& =\mathbf{E}\left(\eta^{3}-3 \mathbf{E} \eta^{2}+2+6(\eta-1)+\eta^{-1}-\eta^{2}\right) \\
& =\mathbf{E}\left[(\eta-1)^{2}\left(\eta+\eta^{-1}-2\right)\right] \geq 0 .
\end{aligned}
$$

This also shows that the skewness vanishes if and only if $\eta=1$ almost surely.
Remark 6.7 (product of self-dual variables). If $\eta_{1}$ and $\eta_{2}$ are two independent self-dual random variables, then

$$
\begin{aligned}
\mathbf{E} \max \left(k, \eta_{1} \eta_{2}\right) & =\mathbf{E}\left[\mathbf{E}\left(\max \left(k, \eta_{1} \eta_{2}\right) \mid \eta_{1}\right)\right] \\
& =\mathbf{E}\left[\mathbf{E}\left(\max \left(k \eta_{2}, \eta_{1}\right) \mid \eta_{2}\right)\right]=\mathbf{E} \max \left(k \eta_{1} \eta_{2}, 1\right)
\end{aligned}
$$

i.e., $\eta_{1} \eta_{2}$ is self-dual. By taking successive products, it is possible to construct a sequence of self-dual random variables whose logarithms build a random walk. Note, however, that the values of this random walk at different time points are not jointly self-dual; cf. Remark 3.3.

Recall that the exponentially self-dual and infinitely divisible random variable $\xi$ can be characterized in terms of its generating triplet $\left(\sigma^{2}, \nu, \gamma\right)$ by applying Theorem 4.1. The obtained univariate result corresponds to [19]. If $\mathbf{E}|\xi|^{3}<\infty$, then [14, Prop. 3.13] yields that

$$
\mathbf{E}(\xi-\mathbf{E} \xi)^{3}=\int_{\mathbb{R}} x^{3} d \nu(x)=\int_{0}^{\infty} x^{3}\left(1-e^{x}\right) d \nu(x)
$$

Thus, the skewness of exponentially self-dual $\xi$ is negative except in the log-normal case, where it is zero.
6.3. Quasi-self-dual variables and asymmetry corrections. Let $S_{T}=S_{0} a \eta$ for $S_{0}, a>0$ with $\eta$ being a general positive random variable, so that the forward price is given by $F=S_{0} a$. Assume that $\eta$ is absolutely continuous with nonvanishing density $p_{\eta}$ and $\mathbf{E} \eta=1$. Then it is possible to find a function $q_{a \eta}$ such that

$$
\begin{align*}
\mathbf{E} f\left(S_{T}\right)=\mathbf{E}\left[f\left(S_{0} /(a \eta)\right) q_{a \eta}(a \eta)\right] & =\mathbf{E}\left[f\left(F /\left(a^{2} \eta\right)\right) q_{a \eta}(a \eta)\right] \\
& =\mathbf{E}\left[f\left(\left(S_{0}\right)^{2} / S_{T}\right) q_{a \eta}(a \eta)\right] \tag{6.6}
\end{align*}
$$

for each function $f: \mathbb{R}_{+} \mapsto \mathbb{R}_{+}$. Indeed, it suffices to choose

$$
q_{a \eta}(x)=\frac{p_{a \eta}\left(x^{-1}\right)}{x^{2} p_{a \eta}(x)}=\frac{p_{S_{T}}\left(x^{-1} S_{0}\right)}{x^{2} p_{S_{T}}\left(x S_{0}\right)} .
$$

By choosing $x=a \eta=S_{T} / S_{0}$, we arrive at

$$
\mathbf{E} f\left(S_{T}\right)=\mathbf{E}\left[f\left(\frac{\left(S_{0}\right)^{2}}{S_{T}}\right)\left(\frac{S_{T}}{S_{0}}\right)^{-2} \frac{p_{S_{T}}\left(\left(S_{0}\right)^{2} / S_{T}\right)}{p_{S_{T}}\left(S_{T}\right)}\right] .
$$

Apart from trivial cases, the density $p_{S_{T}}$ of $S_{T}$ depends on $T$. In view of applications to semistatic hedging described in [12], it is beneficial if the correcting expression

$$
q_{S_{t}}(x)=\frac{p_{S_{t}}\left(\left(S_{0}\right)^{2} / x\right)}{p_{S_{t}}(x)}
$$

at any time $t \in[0, T]$ depends only on $x$ and $S_{0}$ but not on $t$. This is the case if $\eta$ is self-dual with no carrying costs (then $q_{S_{t}}(x)=\left(x / S_{0}\right)^{3}, x>0$, by Theorem 1.2) or quasi-self-dual with parameters $a=e^{\lambda}$ and some $\alpha \neq 0$, which is the case if and only if $q_{S_{t}}(x)=\left(x / S_{0}\right)^{2+\alpha}, x>0$. In the latter case (6.6) turns into

$$
\mathbf{E} f(F \eta)=\mathbf{E}\left[f\left(\frac{F}{a^{2} \eta}\right) a^{\alpha} \eta^{\alpha}\right] .
$$

By letting $f(x)=(x-k)_{+}^{\alpha}$ and noticing that $\mathbf{E}\left[(a \eta)^{\alpha}\right]=1$ in the quasi-self-dual case, this implies the following property:

$$
\mathbf{E}\left[(F \eta-k)_{+}^{\alpha}\right]=a^{-\alpha} \mathbf{E}\left[\left(F-k a^{2} \eta\right)_{+}^{\alpha}\right]=\mathbf{E}\left[\eta^{\alpha}\right] \mathbf{E}\left[\left(F-k \eta\left(\mathbf{E}\left[\eta^{\alpha}\right]\right)^{-2 / \alpha}\right)_{+}^{\alpha}\right],
$$

which can be termed as the power put-call symmetry, since for $a=1$ one has $\mathbf{E}(F \eta-k)_{+}^{\alpha}=$ $\mathbf{E}(F-k \eta)_{+}^{\alpha}$.

## 7. Barrier options and semistatic hedging.

7.1. Time-dependent framework. Consider a finite horizon model with the asset prices given by

$$
S_{t}=S_{0} \circ e^{t \lambda} \circ \eta_{t}=S_{0} \circ e^{t \lambda+\xi_{t}}=\left(S_{01} e^{t \lambda_{1}+\xi_{t 1}}, \ldots, S_{0 n} e^{t \lambda_{n}+\xi_{t n}}\right), \quad t \in[0, T],
$$

where $\lambda \in \mathbb{R}^{n}$ represent deterministic carrying costs and all components of $\left(\eta_{t}\right)_{t \in[0, T]}=$ $\left(e^{\xi_{t}}\right)_{t \in[0, T]}$ are martingales with $\left(\xi_{t}\right)_{t \in[0, T]}$ being a Lévy process. Fix $i \in\{1, \ldots, n\}$ and assume that $\eta_{t} \in \operatorname{QSD}_{i}(t \lambda, \alpha)$ for every $t \in[0, T]$. The latter condition is satisfied (with $\alpha=1$ and $\lambda=0$ ) for all exponentially self-dual Lévy models with no carrying costs analyzed in section 4 and for quasi-self-dual Lévy models from section 5 for nonvanishing $\lambda$; see Corollary 5.5.

Let $\tau$ be a stopping time with values in $[0, T]$ and let $\mathfrak{F}_{\tau}$ be the corresponding stopping $\sigma$-algebra. Since $\left(\xi_{t}\right)_{t \in[0, T]}$ is a Lévy process, $\left(\xi_{\tau}, \xi_{T}\right)$ and $\left(\xi_{\tau}, \xi_{\tau}+\xi_{T-\tau}^{\prime}\right)$ share the same distribution, where $\left(\xi_{t}^{\prime}\right)_{t \in[0, T]}$ is an independent copy of the process $\left(\xi_{t}\right)_{t \in[0, T]}$. Hence, $\left(S_{\tau}, S_{T}\right)$ and $\left(S_{\tau}, S_{\tau} \circ e^{\lambda(T-\tau)+\xi_{T-\tau}^{\prime}}\right)$ also coincide in distribution. Then

$$
\begin{aligned}
\mathbf{E}\left[f\left(S_{T}\right) \mid \mathfrak{F}_{\tau}\right]=\mathbf{E}\left[f\left(S_{T}\right) \mid S_{\tau}\right] & =\mathbf{E}\left[f \left(S_{\tau} \circ e^{\left.\left.\lambda(T-\tau)+\xi_{T-\tau}^{\prime}\right) \mid S_{\tau}\right]}\right.\right. \\
& =\mathbf{E}\left[f\left(S_{\tau} \circ e^{\left.\left.\lambda(T-\tau)+\xi_{T-\tau}^{\prime}\right) \mid \mathfrak{F}_{\tau}\right]}\right]\right.
\end{aligned}
$$

where $f$ is any integrable payoff function. The quasi-self-duality of $\eta_{T-\tau}^{\prime}=e^{\xi_{T-\tau}^{\prime}}$ with respect to the $i$ th numeraire adjusted for conditional expectations (see Remark 2.7) yields that

$$
\mathbf{E}\left[f\left(S_{T}\right) \mid \mathfrak{F}_{\tau}\right]=\mathbf{E}\left[f\left(S_{\tau} \circ e^{K_{i}\left(\lambda(T-\tau)+\xi_{T-\tau}^{\prime}\right)}\right) e^{\alpha\left(\lambda_{i}(T-\tau)+\xi_{(T-\tau) i}^{\prime}\right)} \mid \mathfrak{F}_{\tau}\right],
$$

whence

$$
\begin{align*}
& \mathbf{E}\left[f\left(S_{T}\right) \mid \mathfrak{F}_{\tau}\right]  \tag{7.1}\\
& =\mathbf{E}\left[\left.f\left(\frac{S_{T 1} S_{\tau i}}{S_{T i}}, \ldots, \frac{S_{T(i-1)} S_{\tau i}}{S_{T i}}, \frac{S_{\tau i}^{2}}{S_{T i}}, \frac{S_{T(i+1)} S_{\tau i}}{S_{T i}}, \ldots, \frac{S_{T n} S_{\tau i}}{S_{T i}}\right)\left(\frac{S_{T i}}{S_{\tau i}}\right)^{\alpha} \right\rvert\, \mathfrak{F}_{\tau}\right] ;
\end{align*}
$$

cf. Remark 2.6. If $S_{\tau i}=H$ almost surely for a constant $H$, then

$$
\begin{align*}
& \mathbf{E}\left[f\left(S_{T}\right) \mid \mathfrak{F}_{\tau}\right]  \tag{7.2}\\
& =\mathbf{E}\left[\left.f\left(\frac{S_{T 1} H}{S_{T i}}, \ldots, \frac{S_{T(i-1)} H}{S_{T i}}, \frac{H^{2}}{S_{T i}}, \frac{S_{T(i+1)} H}{S_{T i}}, \ldots, \frac{S_{T n} H}{S_{T i}}\right)\left(\frac{S_{T i}}{H}\right)^{\alpha} \right\rvert\, \mathfrak{F}_{\tau}\right] .
\end{align*}
$$

Identity (7.1) for $\alpha=1, \lambda=0$ yields the self-dual case, and in the univariate case $n=1$ corresponds to [12, eq. (5.3)]. Classical examples with trivial carrying costs (i.e., $\lambda=0$ ) are options on futures or options on shares with dividend yield being equal to the risk-free interest rate. For the univariate quasi-self-dual case, see [12, Cor. 6.10].

Remark 7.1. Instead of the quasi-self-duality and Lévy properties, it is possible to impose (7.1) for stopping times $\tau \in[0, T]$ that might appear in relation to hedging of particular barrier options.
7.2. Multivariate hedging with a single univariate barrier. Assume a risk-neutral setting for a price process $\left(S_{t}\right)_{t \in[0, T]}$ and fix a barrier level at $H>0$ such that $S_{0 i} \neq H$ with given $i \in\{1, \ldots, n\}$. For simplicity of notation, define function $\hat{\varkappa}_{i}: \mathbb{E}^{n} \mapsto \mathbb{E}^{n}$ acting as

$$
\hat{\varkappa}_{i}\left(S_{T}, H\right)=\frac{H}{S_{T i}}\left(S_{T 1}, \ldots, S_{T(i-1)}, H, S_{T(i+1)}, \ldots, S_{T n}\right)
$$

Define $\Xi_{t i}$ to be the (closed) line segment with endpoints $S_{0 i}$ and $S_{t i}$ and let

$$
\tau_{H}=\inf \left\{t \geq 0: H \in \Xi_{t i}\right\} \quad \text { and } \quad \chi=\mathbb{1}_{\tau_{H} \leq T}
$$

cf. [12, sec. 5.2$]$, where a slightly different way to handle the two cases when the initial price is, respectively, lower and higher than the barrier is used. Furthermore, assume that the asset price dynamics satisfies (7.1) for the stopping time $\tau=\tau_{H}$ and that $S_{\tau_{H} i}=H$ almost surely in the event that $\left\{\tau_{H} \leq T\right\}$. Condition (7.1) is ensured by assuming that the asset price dynamics follows a Lévy-driven process satisfying the conditions of Theorem 5.3 (for $\alpha=1, \lambda=0$, corresponding to the conditions stated in Theorem 4.1 for self-dual cases). The equality $S_{\tau_{H} i}=H$ almost surely in the event that $\left\{\tau_{H} \leq T\right\}$ is guaranteed by the sample path continuity of the $i$ th component of $\left(S_{t}\right)_{t \in[0, T]}$. If the $i$ th component of (the Lévy process) $\left(\xi_{t}\right)_{t \in[0, T]}$ is not continuous, the symmetry condition (5.4) on the Lévy measure implies the presence of jumps of both signs, so it is much more difficult to ensure that $S_{\tau_{H}}=H$ almost surely.

Remark 7.2 (applicability and accuracy). In a risk-neutral setting, by applying an independent common stochastic clock that is continuous with respect to calendar time, the suggested hedging strategies are applicable and accurate for independently time-changed self-dual exponential Lévy models without jumps in the $i$ th component. This is seen by means of conditioning arguments described in [12, Thms. 4.2 and 5.8, Rem. 4.3]. They are also applicable and
accurate for quasi-self-dual exponential Lévy models without jumps in the $i$ th component. For super- or subreplication in the presence of jumps in the $i$ th component, see Example 7.3 and Remark 7.4.

Take any integrable payoff function $f$ and consider an option with payoff $\chi f\left(S_{T}\right)$, i.e., the knock-in option with barrier $H$ for the $i$ th asset. In order to replicate this option using only options that depend on the terminal value $S_{T}$, consider a European claim on

$$
\begin{equation*}
G\left(S_{T}\right)=f\left(S_{T}\right) \mathbb{1}_{H \in \Xi_{T i}}+\left(\frac{S_{T i}}{H}\right)^{\alpha} f\left(\hat{\varkappa}_{i}\left(S_{T}, H\right)\right)\left(\mathbb{1}_{H \in \Xi_{T i}}-\mathbb{1}_{S_{T i}=H}\right) . \tag{7.3}
\end{equation*}
$$

Here one has to bear in mind that this is feasible only if the considered claims are liquid or can be replicated by liquid instruments. In relation to this, Carr and Laurence [11] mention that "all major banks stand ready to provide over-the-counter quotes on customized baskets," which obviously helps to construct the desired hedges and even to replicate general payoff functions. Furthermore, there is a fast growing literature about sub- and superreplication of multiasset instruments; see, e.g., [25] and the references therein.

In the event that $\left\{\tau_{H}>T\right\}$, the claim in (7.3) expires worthless as desired. If the barrier knocks in, we can exchange (7.3) for a claim on $f\left(S_{T}\right)$ at zero cost. To confirm this, define $\hat{\Xi}_{t i}$ to be the (closed) line segment with endpoints $S_{0 i}$ and $H^{2} S_{t i}^{-1}$. Note that $H \notin \hat{\Xi}_{t i}$ if and only if $H \in \Xi_{t i} \backslash\left\{S_{t i}\right\}$. Hence, in the event that $\left\{\tau_{H} \leq T\right\}$, by (7.2), we have

$$
\begin{aligned}
\mathbf{E}\left[f\left(S_{T}\right) \mid \mathfrak{F}_{\tau}\right] & =\mathbf{E}\left[f\left(S_{T}\right) \mathbb{1}_{H \in \Xi_{T i} \mid} \mid \mathfrak{F}_{\tau}\right]+\mathbf{E}\left[f\left(S_{T}\right) \mathbb{1}_{H \notin \Xi_{T i}} \mid \mathfrak{F}_{\tau}\right] \\
& =\mathbf{E}\left[f\left(S_{T}\right) \mathbb{1}_{H \in \Xi_{T i}} \mid \mathfrak{F}_{\tau}\right]+\mathbf{E}\left[\left.\left(\frac{S_{T i}}{H}\right)^{\alpha} f\left(\hat{\varkappa}_{i}\left(S_{T}, H\right)\right) \mathbb{1}_{H \notin \hat{\Xi}_{T i} \mid} \right\rvert\, \mathfrak{F}_{\tau}\right] \\
& =\mathbf{E}\left[f\left(S_{T}\right) \mathbb{1}_{H \in \Xi_{T i}} \mid \mathfrak{F}_{\tau}\right]+\mathbf{E}\left[\left.\left(\frac{S_{T i}}{H}\right)^{\alpha} f\left(\hat{\varkappa}_{i}\left(S_{T}, H\right)\right)\left(\mathbb{1}_{H \in \Xi_{T i}}-\mathbb{1}_{S_{T i}=H}\right) \right\rvert\, \mathfrak{F}_{\tau}\right] .
\end{aligned}
$$

For simplicity, we assume from now on that $S_{T i}$ has a nonatomic distribution, so that (7.3) becomes

$$
\begin{equation*}
G\left(S_{T}\right)=\left(f\left(S_{T}\right)+\left(\frac{S_{T i}}{H}\right)^{\alpha} f\left(\hat{\varkappa}_{i}\left(S_{T}, H\right)\right)\right) \mathbb{1}_{H \in \Xi_{T i}} \tag{7.4}
\end{equation*}
$$

Consider a general basket call $f\left(S_{T}\right)=\left(\sum_{j=1}^{n} u_{j} S_{T j}-k\right)_{+}$. By (7.4), the hedge for the knock-in basket call with payoff function $\chi f\left(S_{T}\right)$ is given by the derivative with payoff function

$$
\left\{\left(\sum_{j=1}^{n} u_{j} S_{T j}-k\right)_{+}+\left(\frac{S_{T i}}{H}\right)^{\alpha-1}\left(u_{i} H-\left(\frac{k}{H} S_{T i}-\sum_{j=1, j \neq i}^{n} u_{j} S_{T j}\right)\right)_{+}\right\}_{H \in \Xi_{T i}},
$$

which depends only on $S_{T}$.
If (7.1) holds with $\alpha=1$, this hedge becomes

$$
\begin{equation*}
\left\{\left(\sum_{j=1}^{n} u_{j} S_{T j}-k\right)_{+}+\left(u_{i} H-\left(\frac{k}{H} S_{T i}-\sum_{j=1, j \neq i}^{n} u_{j} S_{T j}\right)\right)_{+}\right\} \mathbb{1}_{H \in \Xi_{T i}}, \tag{7.5}
\end{equation*}
$$

which is the sum of a basket call and a spread put with the knocking condition depending only on the $i$ th component $S_{T i}$ at maturity.

In some cases the knocking condition at maturity can be incorporated into the payoff function. For this, note that we can write all integrable payoff functions in the form

$$
f\left(S_{T}\right)=f_{0}\left(S_{T}\right) \mathbb{1}_{S_{T} \in \Theta}, \quad \text { where } \quad \Theta=\{x: f(x) \neq 0\},
$$

with $\Theta$ possibly being $\mathbb{R}_{+}^{n}$. For example, the basket call $\left(\sum_{j=1}^{n} u_{j} S_{T j}-k\right)_{+}$can be written as $\left(\sum_{j=1}^{n} u_{j} S_{T j}-k\right) \mathbb{1}_{\sum_{j=1}^{n} u_{j} S_{T j}>k}$. Of course if $H \notin \Xi_{T i}$ would imply that $\hat{\varkappa}_{i}\left(S_{T}, H\right) \notin \Theta$ and $S_{T} \notin \Theta$ at the same time, then it would be possible to omit $\mathbb{1}_{H \in \Xi_{T i}}$ in (7.4), but this is not the case for standard payoff functions. If $H \in \Xi_{T i}$ implies that $S_{T} \notin \Theta$ (resp., $\hat{\varkappa}_{i}\left(S_{T}, H\right) \notin \Theta$ ), then the first (second) summand in (7.4) is always zero. If, furthermore, $H \notin \Xi_{T i}$ implies that $\hat{\varkappa}_{i}\left(S_{T}, H\right) \notin \Theta\left(S_{T} \notin \Theta\right)$, then we can omit the first (second) summand in (7.4) and hedge with the second (first) summand without the knocking condition $\mathbb{1}_{H \in \Xi_{T i}}$; i.e., in (7.5) we can hedge with a conventional basket option.

Example 7.3. Consider a bivariate price process $\left(S_{t 1}, S_{t 2}\right)_{t \in[0, T]}$ in a risk-neutral setting satisfying (7.1) with $\alpha=1$ and $i=1$ for the stopping time $\tau=\tau_{H}=\inf \left\{t: S_{t 1} \leq H\right\}$ with barrier $H$ such that $0<H<S_{01}$. First assume again that $\left(S_{t 1}\right)_{t \in[0, T]}$ cannot jump over $H$. For the spread option

$$
f\left(S_{T 1}, S_{T 2}\right)=\left(a S_{T 1}-b S_{T 2}-k\right)_{+}, \quad a, b>0,
$$

assume additionally that $a H \leq k$ and define $\chi=\mathbb{1}_{\tau_{H} \leq T}$. By using the hedging strategy described in (7.4) and $a H \leq k$, we obtain a hedge for $\chi f\left(S_{T 1}, S_{T 2}\right)$ by

$$
\begin{aligned}
\left(a S_{T 1}\right. & \left.-b S_{T 2}-k\right)_{+} \mathbb{1}_{H \in\left[S_{T 1}, S_{01}\right]}+\left(a H-\frac{k}{H} S_{T 1}-b S_{T 2}\right)_{+} \mathbb{1}_{H \in\left[S_{T 1}, S_{01}\right]} \\
& =\left(a H-\frac{k}{H} S_{T 1}-b S_{T 2}\right)_{+} \mathbb{1}_{H \in\left[S_{T 1}, S_{01}\right]}=\left(a H-\frac{k}{H} S_{T 1}-b S_{T 2}\right)_{+}
\end{aligned}
$$

i.e., it is possible to hedge with a basket put. Therefore, the related knock-out option can be hedged with a long position in the spread call with payoff function $f\left(S_{T 1}, S_{T 2}\right)=\left(a S_{T 1}-\right.$ $\left.b S_{T 2}-k\right)_{+}$and a short position in the above hedge. Note that we assumed only that $b>0$ so that the knock-in level can, but need not, be deep out of the money. If $\left(S_{t 1}\right)_{t \in[0, T]}$ is driven by a Lévy process such that $\left(S_{t 1}\right)_{t \in[0, T]}$ can jump over the barrier $H$, we get a superreplication in the case of the knock-in option and a more problematic subreplication in the case of the knock-out option.

Assuming (7.1) for the stopping time $\tau_{H}$ with $\alpha \neq 1$, where $\left(S_{t 1}\right)_{t \in[0, T]}$ does not jump over $H$, the hedge for the knock-in option has to be modified as

$$
\left(\frac{S_{T 1}}{H}\right)^{\alpha-1}\left(a H-\frac{k}{H} S_{T 1}-b S_{T 2}\right)_{+}
$$

while the modification for the knock-out is now obvious. This example can easily be extended in a higher-dimensional setting as long as all risky assets without a knocking barrier enter the payoff function with a minus sign.

Remark 7.4 (jumps and superreplication). Example 7.3 illustrates that in the presence of jumps in the $i$ th component it is possible that the suggested hedge results in super- or subreplications, it is even possible that hedges provide superreplications for small and subreplications for big jumps, etc. However, similarly to the univariate case (cf. [12, Remark 5.17]), we can give the following criterion for superreplication. Recall that $G$ is a hedge portfolio given by (7.3) or (7.4). Suppose that

$$
\tilde{G}\left(S_{\tau_{H}}\right)=\mathbf{E}\left[G\left(S_{T}\right)-f\left(S_{T}\right) \mid \mathfrak{F}_{\tau_{H}}\right] \quad \text { almost surely, }
$$

where for up-and-in (down-and-in) options $\tilde{G}$ is some increasing (decreasing) function in the $i$ th component. Then the hedge superreplicates the knock-in contracts, since $\tilde{G}\left(S_{\tau_{H}}\right) \geq$ $\tilde{G}\left(\left(S_{\tau_{H}}, \ldots, H, \ldots, S_{\tau_{H} n}\right)\right)=0$. A sufficient condition for Lévy driven models is that the function $G\left(S_{T}\right)-f\left(S_{T}\right)$ itself increases (decreases) in the $i$ th component.

Example 7.5. Consider a bivariate price process in a risk-neutral setting. Assume that $S_{01}>k>0$ and that (7.1) holds for $\tau=\tau_{k}=\inf \left\{t: S_{t 1} \leq k\right\}, \alpha=1$, and $i=1$, while $\left(S_{t 1}\right)_{t \in[0, T]}$ cannot jump over $k$. Introduce the (possibly negative integrable) payoff function

$$
g\left(S_{T 1}, S_{T 2}\right)=\left(S_{T 1}-k\right)+\left(S_{T 2} \wedge\left(S_{T 1}-k\right)\right),
$$

where $a \wedge b=\min (a, b)$.
By (7.4), with $\alpha=1$ we obtain a hedge for $\mathbb{1}_{\tau_{k} \leq T} g\left(S_{T 1}, S_{T 2}\right)$ as

$$
\left\{\left(\left(S_{T 1}-k\right)+\left(S_{T 2} \wedge\left(S_{T 1}-k\right)\right)\right)+\left(\left(k-S_{T 1}\right)+\left(S_{T 2} \wedge\left(k-S_{T 1}\right)\right)\right)\right\}_{\mathbb{1}_{k \in\left[S_{T 1}, S_{01}\right]} .}
$$

Here we can get rid of the indicator function by noticing that the above payoff function can be written as

$$
\left(\left(S_{T 1}-k\right)+\left(S_{T 2} \wedge\left(S_{T 1}-k\right)\right)\right)-\left(S_{T 1}+S_{T 2}-k\right)_{+}+\left(k+S_{T 2}-S_{T 1}\right)_{+} .
$$

Furthermore, for the related knock-out option we get a hedge given by a long position in the basket call with payoff function $\left(S_{T 1}+S_{T 2}-k\right)_{+}$and a short position in the put spread with payoff function $\left(k-\left(S_{T 1}-S_{T 2}\right)\right)_{+}$.
7.3. Examples of hedging in jointly self-dual cases. In this section we create hedges for more complex instruments and bivariate models satisfying (7.1) for two different numeraires.

Example 7.6 (options with knocking conditions depending on two assets). Assume that there is a risk-neutral setting for a price process $\left(S_{t}\right)_{t \in[0, T]}=\left(S_{t 1}, S_{t 2}\right)_{t \in[0, T]}$, where (7.1) holds for both assets with $\alpha=1$ and the subsequently defined stopping times. Furthermore, let $k_{x}, k_{y}>0$ be constants such that $k_{x}<S_{01}, S_{02}<k_{y}$ and define the stopping times $\tau_{i x}=\inf \left\{t>0: S_{t i} \leq k_{x}\right\}, \tau_{i y}=\inf \left\{t>0: S_{t i} \geq k_{y}\right\}$ and the corresponding stopping $\sigma$ algebras $\mathfrak{F}_{\tau_{i x}}, \mathfrak{F}_{\tau_{i y}}, i=1,2$, as well as the stopping time $\tau=\tau_{1 x} \wedge \tau_{2 x}$ with the corresponding stopping $\sigma$-algebra $\mathfrak{F}_{\tau}$. In order to ensure exact hedges, assume also that the price processes cannot jump over the barriers $k_{x}$ and $k_{y}$, respectively. These conditions are satisfied for models obtained by independent time changes (by a continuous common stochastic clock) applied to the jointly self-dual models based on Example 4.5.

Consider the claims

$$
\begin{aligned}
& X=\left(S_{T 1}-S_{T 2}-k_{x}\right)_{+} \mathbb{1}_{\tau=\tau_{1 x} \leq T}+\left(S_{T 2}-S_{T 1}-k_{x}\right)_{+} \mathbb{1}_{\tau_{1 x} \neq \tau=\tau_{2 x} \leq T}, \\
& Y=\left(k_{y}-S_{T 1}-S_{T 2}\right)_{+}\left(\mathbb{1}_{\tau_{1 y} \leq T<\tau_{2 y}}-\mathbb{1}_{\tau_{2 y} \leq T<\tau_{1 y}}\right) ;
\end{aligned}
$$

i.e., $X$ is a knock-in spread option on the difference between the share which first hits $k_{x}$ and the other one being knocked in only if at least one share hits $k_{x}$ before $T$. At maturity, $Y$ is a long position in a basket put if and only if the first, but not the second, asset price hits the price level $k_{y}$ and a short position in the same basket put if and only the second, but not the first, asset hits $k_{y}$ before $T$.

The claim $X$ can be hedged with a long position in the European basket put with payoff $\left(k_{x}-S_{T 1}-S_{T 2}\right)_{+}$. The claim $Y$ can be hedged by entering a long position in the spread option with payoff $\left(S_{T 1}-S_{T 2}-k_{y}\right)_{+}$along with a short position in the spread option with payoff $\left(S_{T 2}-S_{T 1}-k_{y}\right)_{+}$. To see that, apply identity (7.2) for $\alpha=1$, so that

$$
\begin{align*}
& \mathbf{E}\left[\left(k_{z}-S_{T 1}-S_{T 2}\right)_{+} \mid \mathfrak{F}_{\tau_{1 z}}\right]=\mathbf{E}\left[\left(S_{T 1}-S_{T 2}-k_{z}\right)_{+} \mid \mathfrak{F}_{\tau_{1 z}}\right],  \tag{7.6}\\
& \mathbf{E}\left[\left(k_{z}-S_{T 1}-S_{T 2}\right)_{+} \mid \mathfrak{F}_{\tau_{2 z}}\right]=\mathbf{E}\left[\left(S_{T 2}-S_{T 1}-k_{z}\right)_{+} \mid \mathfrak{F}_{\tau_{2 z}}\right],
\end{align*}
$$

$z=x, y$, while the value of the European spread option with payoff $\left(S_{T 1}-S_{T 2}-k\right)_{+}$(resp., $\left.\left(S_{T 2}-S_{T 1}-k\right)_{+}\right)$remains unchanged by applying (7.2) at $\tau_{2}$ (resp., $\tau_{1}$ ).

As far as $X$ is concerned we have that in the case where $\{\tau>T\}$ neither $X$ is knocked in nor is the basket put in the hedge portfolio in the money, since $S_{T 1}, S_{T 2}>k_{x}$. If $\left\{\tau=\tau_{1 x} \leq T\right\}$, by (7.6) we can exchange the long position in the hedge portfolio for the needed spread; if $\tau_{1 x} \neq \tau=\tau_{2 x} \leq T$, the same is true due to (7.7).

As far as $Y$ is concerned we have that if $\left\{\tau_{i y}>T\right\}, i=1,2$, both instruments in the hedging portfolio are out of the money since $S_{T 1}, S_{T 2} \leq k_{y}$, i.e., have payoff zero, as does $Y$. In the event that we first have $\left\{\tau_{1 y} \leq T\right\}$, we can change the long position in the spread option with payoff $\left(S_{T 1}-S_{T 2}-k_{y}\right)_{+}$in a long position in the basket put with payoff $\left(k_{y}-S_{T 1}-S_{T 2}\right)_{+}$while leaving the short position unchanged. Provided that additionally $\left\{\tau_{2 y} \in\left[\tau_{1 y}, T\right]\right\}$, by (7.7), we can also exchange the short position for the same basket put, so that we can close our positions as required; otherwise, i.e., $\left\{\tau_{2 y}>T\right\}$, the long position in the hedge portfolio yields a potentially needed payoff of $Y$, while the short position still matures worthless. In the event that we first have $\left\{\tau_{2 y} \leq T\right\}$, by (7.7), we can change the short position in the needed basket put while leaving the long position unchanged. If, furthermore, $\left\{\tau_{1 y} \in\left[\tau_{2 y}, T\right]\right\}$, we can again close our position due to (7.6). Otherwise, unlike the long position, the short position in the hedge portfolio may be in the money at maturity, but in that case $Y$ at maturity would be a long position for the hedger with the same payoff.

In the case of nonvanishing carrying costs $\lambda=\left(\lambda_{1}, \lambda_{2}\right)$, Theorem 5.3 implies that the bivariate jointly self-dual Black-Scholes model (see Example 4.5) satisfies (7.1) for two powers $\alpha_{1}$ and $\alpha_{2}$ calculated as in Example 5.9 for the corresponding components. Also in this case asset prices cannot jump over the barriers; i.e., we can apply (7.2) in the same way to see that the hedge should theoretically be modified to a long position in the European derivative with payoff $\left(S_{T 1}-S_{T 2}-k_{y}\right)_{+}\left(k_{y}^{-1} S_{T 1}\right)^{\alpha_{1}-1}$ along with a short position in the European derivative with payoff $\left(S_{T 2}-S_{T 1}-k_{y}\right)_{+}\left(k_{y}^{-1} S_{T 2}\right)^{\alpha_{2}-1}$.

For creating semistatic hedges of barrier spread options with certain knocking conditions, e.g., claims of the form

$$
Z=\left(S_{T 1}-S_{T 2}-k\right)_{+} \mathbb{1}_{S_{t 1}>S_{t 2}, \forall t \in[0, T]}, \quad k>0,
$$

in equal carrying cost cases the full strength of the joint self-duality is not needed. It suffices to assume exchangeability being implied by the joint self-duality; see Corollary 3.1 and [29] for details including model characterizations, weakening of the exchangeability assumption, and hedges for several related derivatives.
7.4. Semistatic superhedges of basket options. The following superhedges may be quite expensive for replication purposes. Thus, they seem to be more useful if one would like to speculate with a basket option and get some additional money by writing a different knock-in basket option, where the maximum loss should be limited to the initially invested capital.

In what follows we work in the same setting as in section 7.2 with the additional assumptions that $\alpha=1$ and $S_{0 i}>H>0$. Define again the stopping time $\tau_{H}=\inf \left\{t: S_{t i} \leq H\right\}$ and let $\mathfrak{F}_{\tau_{H}}$ be the corresponding $\sigma$-algebra. Consider the knock-in basket option with the following payoff function:

$$
\chi\left(\sum_{j=1}^{n} u_{j} S_{T j}-k\right)_{+}, \quad k, u_{i}>0, u_{j} \in \mathbb{R} \text { for } j=1, \ldots, n, j \neq i
$$

where $\chi=\mathbb{1}_{\tau_{H} \leq T}$. By (7.2), for $\alpha=1$ we have

$$
\mathbf{E}\left[\left(\sum_{j=1}^{n} u_{j} S_{T j}-k\right)_{+} \mid \mathfrak{F}_{\tau_{H}}\right]=\mathbf{E}\left[\left.\left(\sum_{j=1, j \neq i}^{n} u_{j} S_{T j}-\frac{k}{H} S_{T i}+u_{i} H\right)_{+} \right\rvert\, \mathfrak{F}_{\tau_{H}}\right] .
$$

Hence, the maximum loss of buying the basket option in the right-hand side of the above equation and short selling the initial knock-in basket call does not exceed the initial costs of this strategy.

Note that, in this setting, jumps over the barrier $H$ in Lévy driven models add a further aspect of superhedging.

Acknowledgments. The authors are grateful to Daniel Neuenschwander and Pierre Patie for helpful discussions, to Rolf Burgermeister, Peter Carr, and Markus Liechti for useful hints, and to the referees for stimulating remarks.

## REFERENCES

[1] L. B. G. Andersen, J. Andreasen, and D. Eliezer, Static replication of barrier options: Some general results, J. Comput. Finance, 5 (2002), pp. 1-25.
[2] J. Andreasen, Behind the mirror, Risk, 14 (2001), pp. 108-110.
[3] D. S. Bates, The crash of ' 87 -was it expected? The evidence from options markets, J. Finance, 46 (1991), pp. 1009-1044.
[4] D. S. Bates, The skewness premium: Option pricing under asymmetric processes, Adv. Futures Options Res., 9 (1997), pp. 51-82.
[5] J. Bowie and P. Carr, Static simplicity, Risk, 7 (1994), pp. 45-49.
[6] D. T. Breeden and R. H. Litzenberger, Prices of state-contingent claims implicit in options prices, J. Business, 51 (1978), pp. 621-651.
[7] P. Carr, European Put Call Symmetry, Tech. report, Cornell University, Ithaca, NY, 1994.
[8] P. Carr and M. Chesney, American Put Call Symmetry, preprint, 1996.
[9] P. Carr and A. Chou, Breaking barriers, Risk, 10 (1997), pp. 139-145.
[10] P. Carr, K. Ellis, and V. Gupta, Static hedging of exotic options, J. Finance, 53 (1998), pp. 1165-1190.
[11] P. Carr and P. Laurence, Multi-asset stochastic local variance contracts, Math. Finance, to appear.
[12] P. Carr and R. Lee, Put-call symmetry: Extensions and applications, Math. Finance, 19 (2009), pp. 523-560.
[13] P. Carr and D. B. Madan, Towards a theory of volatility trading, in Volatility, R. Jarrow, ed., Risk Publications, London, 1998, pp. 417-427.
[14] R. Cont and P. Tankov, Financial Modelling with Jump Processes, Chapman \& Hall/CRC, London, 2004.
[15] J. Detemple, American options: Symmetry properties, in Option Pricing, Interest Rates and Risk Management, J. Cvitanić, E. Jouini, and M. Musiela, eds., Cambridge University Press, Cambridge, UK, 2001, pp. 67-104.
[16] E. Eberlein, A. Papapantoleon, and A. N. Shiryaev, Esscher transform and the duality principle for multidimensional semimartingales, Ann. Appl. Probab., 19 (2009), pp. 1944-1971.
[17] E. Eberlein, A. Papapantoleon, and A. N. Shiryaev, On the duality principle in option pricing: Semimartingale setting, Finance Stoch., 12 (2008), pp. 265-292.
[18] J. Fajardo and E. Mordecki, Skewness Premium with Lévy Processes, Working paper, IBMEC, Rio de Janeiro, Brazil, 2006.
[19] J. FAJARDO AND E. Mordecki, Symmetry and duality in Lévy markets, Quant. Finance, 6 (2006), pp. 219-227.
[20] O. Grabbe, The pricing of call and put options on foreign exchange, J. Int. Money Finance, 2 (1983), pp. 239-253.
[21] A. Gulishashvili and E.-M. Stein, Asymptotic behavior of the distribution of the stock price in models with stochastic volatility, C. R. Math. Acad. Sci. Paris, 343 (2006), pp. 519-523.
[22] G. M. Henkin and A. A. Shananin, Bernstein theorems and Radon transform. Application to the theory of production functions, in Mathematical Problems of Tomography, I. M. Gel'fand and S. G. Gindikin, eds., AMS, Providence, RI, 1990, pp. 189-223.
[23] J. HüsLer and R.-D. Reiss, Maxima of normal random vectors: Between independence and complete dependence, Statist. Probab. Lett., 7 (1989), pp. 283-286.
[24] J. Jacod and A. N. Shiryaev, Limit Theorems for Stochastic Processes, 2nd ed., Springer, Berlin, 2002.
[25] P. LaURENCE AND T.-H. WANG, Sharp distribution free lower bounds for spread options and the corresponding optimal subreplicating portfolios, Insurance Math. Econom., 44 (2009), pp. 35-47.
[26] R. McDonald and M. Schroder, A parity result for American options, J. Comput. Finance, 1 (1998).
[27] R. C. Merton, The theory of rational option pricing, Bell J. Econ. Manag. Sci., 4 (1973), pp. $141-183$.
[28] I. Molchanov, Convex geometry of max-stable distributions, Extremes, 11 (2008), pp. 235-259.
[29] I. Molchanov and M. Schmutz, Semi-static Hedging under Exchangeability Type Conditions, Tech. report, University of Bern, Bern, Switzerland, 2009; also available online from http://arxiv.org/abs/ 0901.4914.
[30] K. Mosler, Multivariate Dispersion, Central Regions and Depth. The Lift Zonoid Approach, Lecture Notes in Statist. 165, Springer, Berlin, 2002.
[31] R. Poulsen, Barrier options and their static hedges: Simple derivations and extensions, Quant. Finance, 6 (2006), pp. 327-335.
[32] S. A. Ross, Options and efficiency, Quart. J. Econ., 90 (1976), pp. 75-89.
[33] K. Sato, Subordination depending on a parameter, in Probability Theory and Mathematical Statistics, B. Grigelionis, J. Kubilius, V. Palauskas, H. Pragarauskas, R. Rudzkis, and V. Statulevičius, eds., Vol. 2, Mokslas, Vilnius, Lithuania, 1990, pp. 372-382.
[34] K. Sato, Lévy Processes and Infinitely Divisible Distributions, Cambridge University Press, Cambridge, UK, 1999.
[35] K. Sato, Density Transformation in Lévy Processes, Lecture Notes, http://www.maphysto.dk/ publications/MPS-LN/2000/7.pdf (2000).
[36] M. Schmutz, Semi-static Hedging for Certain Margrabe Type Options with Barriers, Tech. report, University of Bern, Bern, Switzerland, 2008; also available online from http://arxiv.org/abs/0810.5146.
[37] M. Schroder, Changes of numeraire for pricing futures, forwards and options, Rev. Financ. Stud., 12 (1999), pp. 1143-1163.

# On the Microstructural Hedging Error* 

Christian Y. Robert ${ }^{\dagger}$ and Mathieu Rosenbaum ${ }^{\ddagger}$


#### Abstract

We consider the issue of hedging a European derivative security in the presence of microstructure noise. In a market where the efficient price of the asset is driven by a stochastic volatility process, we assume an agent wants to use a (possibly misspecified) local volatility-type replication strategy. Focusing on microstructure noise effects, our goal is to evaluate the error between the theoretical, but practically unfeasible, strategy and its market adapted versions. The microstructural hedging error is in particular due to transaction price discreteness and endogenous trading times. Thus, we consider a transaction price model that accommodates such inherent properties of ultrahigh frequency data with the assumption of a continuous semimartingale efficient price. In this framework, we study two hedging strategies derived from the local volatility-type hedging strategy: (i) the hedging portfolio is rebalanced every time that the transaction price moves; (ii) the hedging portfolio is rebalanced only once the transaction price has varied by more than a selected value. To assess these strategies, we use an asymptotic approach where the number of rebalancing transactions goes to infinity. For the first strategy, we show that, because of microstructure noise effects, the hedging error does not vanish. However, an optimal strategy of the second type enables us to reduce it significantly.


Key words. continuous-time processes, limit theorems for martingales, stable convergence in law, microstructure noise, discrete hedging, endogenous trading times

AMS subject classifications. 60F05, 62P05, 91B70, 91B24
DOI. 10.1137/090764578

1. Introduction. We consider a market with a risky asset whose efficient price is driven by a continuous Itô semimartingale. In this context, pricing and hedging strategies are usually built under the assumption of a "frictionless" market. The fundamental conditions for such a market can be summarized as follows:

- It is possible to borrow and lend cash at a risk-free interest rate.
- The transaction price is equal to the efficient price, irrespective of the volume of the transaction and of its sign (buy or sell).
- One can buy or sell instantaneously and continuously.
- There are no transaction costs.
- The asset is perfectly divisible (it is possible to buy or sell any fraction of a share). Moreover, short selling is authorized.
The failure of one of the preceding conditions makes the problem of hedging a derivative security more complex. The case of restrictions on short selling is considered in [15], and

[^35]the presence of liquidity costs is studied in [3]. The consequences of transaction costs in conjunction with discrete-time hedging operations has also been extensively studied; see, among others, [16], [17], and [19]. In this paper, we are interested in a model where transaction prices are bound to lie on a tick grid. Thus, we consider a framework where the second of the preceding assumptions is no more in force and where the third one becomes irrelevant.

It is well admitted that, in term of prices modeling, continuous Itô semimartingales are natural processes for building low frequency financial strategies. This is no longer true if one is interested in high frequency strategies. For example, one has to consider specific characteristics of transaction prices such as irregular temporal spacing and discreteness. The usual way to model this gap is to consider that the efficient price is contaminated in the high frequencies by the so-called microstructure noise. This microstructure noise occurs for several reasons, one of the most obvious being the fact that market prices have to lie on the tick grid, which in particular means that the observed price process is not continuous but only stepwise continuous.

Microstructure noise effects have been studied mainly in the perspective of volatility estimation; see, among others, [1], [13], [18], [23], [25], [27], and [28]. In this paper, we explore the issue of hedging a European derivative security when taking into account the discreteness of the transaction prices. This has two important consequences. The first one is the impossibility to buy or sell a share at the efficient price: the microstructure noise leads to a cost (possibly negative) that cannot be avoided. The second one comes from the fact that the transaction price changes a finite number of times in a given time period. Therefore, it is reasonable to assume that one waits for a price change before rebalancing the hedging portfolio.

To treat the problem of measuring the microstructural hedging error, we assume an agent considers a hedging procedure which is derived from a usual strategy in a frictionless market. ${ }^{1}$ This usual strategy is given by a (possibly misspecified) local volatility-type replicating portfolio; see, in particular, [5]. The market model we use is the model with uncertainty zones introduced and discussed in [22]. Indeed, it accommodates the stylized facts of prices and durations together with a semimartingale efficient price; see [22]. In this model, the microstructure noise and the durations between price changes are endogenous. ${ }^{2}$ This enables us to obtain realistic dynamics for both prices and durations.

We study two hedging strategies, where the usual continuous-time hedging portfolio is in fact rebalanced at some random trading times: (i) the hedging portfolio is rebalanced every time that the transaction price moves; (ii) the hedging portfolio is rebalanced only once the transaction price has varied by more than a selected value. Strategy (i) is, of course, unrealistic in practice because of the various transaction costs. However, it is important to evaluate the hedging error in this context in order to assess the maximal cost due to microstructure. In particular, it will provide us a reference result for comparing with the results of strategy (ii). Note also that, in some sense, strategy (i) can be seen as an original way to evaluate the impact on the hedging error of implicit transaction costs linked to bid-ask bounce; see section 3.2. Strategy (ii) is quite intuitive from a practitioner's point of view. Indeed, rebalancing

[^36]the portfolio when the price moves "significantly" is much more natural than a timely based rebalancing strategy.

We develop an asymptotic approach for studying the hedging error in the spirit of [2] and [12]. This asymptotic point of view is also used in [6], [9], and [26], whereas $L^{2}$ measures are considered in [7], [8], and [11]. We show that, as the number of transactions goes to infinity, because of the presence of microstructure noise, the hedging error does not vanish for the first strategy. However, it can be reduced significantly by using a strategy of the second type in an optimal way.

This paper is organized as follows. In section 2, we recall the construction of the model with uncertainty zones. The hedging strategies are described in section 3 . We state in section 4 the associated theorems on the microstructural hedging error. A numerical study can be found in section 5 , and the proofs are relegated to section 6 .
2. Model with uncertainty zones. Here we describe our model for the last traded price, namely the model with uncertainty zones introduced in [22]. This model accommodates the stylized facts of prices and durations together with a semimartingale efficient price; see [22]. In order to make the paper self-contained, we repeat the precise description of the model that can be found in [23] (up to slight modifications relative to our specific problem).
2.1. Description of the model. In an idealistic framework, where the efficient price would be observed, market participants would trade when the efficient price would cross the tick grid. In practice, there is some uncertainty about the efficient price value so that market participants are reluctant to price changes. Hence, there is a modification of the transaction price only if some buyers and sellers are truly convinced that the efficient price is sufficiently far from the last traded price. We introduce a parameter $\eta$ that quantifies the aversion to price changes (with respect to the tick size) of the market participants and propose a model that takes into account this aversion.

Let $\left(X_{t}\right)_{t \geq 0}$ denote the theoretical, efficient price of the asset. On a rich enough filtered probability space $\left(\Omega,\left(\mathcal{F}_{t}\right)_{t \geq 0}, \mathbb{P}\right)$, we assume that the logarithm of the efficient price is an $\mathcal{F}_{t}$-adapted continuous semimartingale of the form

$$
\begin{equation*}
\mathrm{d} \log X_{t}=a_{t} \mathrm{~d} t+\sigma_{t-} \mathrm{d} W_{t} \tag{2.1}
\end{equation*}
$$

where $\left(W_{t}\right)_{t \geq 0}$ is a standard $\mathcal{F}$-Brownian motion, $\left(a_{t}\right)_{t \geq 0}$ is a progressively measurable process with locally bounded sample paths, and $\left(\sigma_{t}\right)_{t \geq 0}$ is a positive $\mathcal{F}_{t}$-adapted process with càdlàg sample paths.

The tick grid on which transaction prices are bound to lie is defined as $\{k \alpha ; k \in \mathbb{N}\}$, with $\alpha$ the tick size. For $k \in \mathbb{N}$ and $0<\eta<1$, we define the zone $U_{k}$ by $U_{k}=[0, \infty) \times\left(d_{k}, u_{k}\right)$, with

$$
d_{k}=(k+1 / 2-\eta) \alpha \quad \text { and } \quad u_{k}=(k+1 / 2+\eta) \alpha .
$$

Thus, $U_{k}$ is a band around the midtick grid value $(k+1 / 2) \alpha$; see Figure 1. Note that, when $\eta$ is smaller than $1 / 2$, there is no overlap between the zones.

We assume that the transaction price may jump from price $k^{\prime} \alpha$ to price $k \alpha$, with $k^{\prime} \neq k$ only once the efficient price has exited down the zone $U_{k}$ or exited up the zone $U_{k-1}$ and provided that market conditions are favorable for a transaction to occur. In a way, the
transaction price changes only when the efficient price is close to a new multiple value of $\alpha$ and market participants want to trade. The zones $\left(U_{k}\right)_{k \in \mathbb{N}}$ represent bands inside of which the efficient price cannot trigger a change of the transaction price. Consequently, they will be referred to as the uncertainty zones.

More specifically, let us explain the construction of the sequence $\left(\tau_{i}\right)_{i \geq 0}$ of the exit times from the uncertainty zones, which will lead to a change in the transaction price. Let $\tau_{0}=0$, and assume without loss of generality that $\tau_{1}$ is the exit time of $\left(X_{t}\right)_{t \geq 0}$ from the set $\left(d_{k_{0}-1}, u_{k_{0}}\right)$, where $k_{0}=X_{0}^{(\alpha)} / \alpha$, with $X_{0}^{(\alpha)}$ the value of $X_{0}$ rounded to the nearest multiple of $\alpha$. We introduce a sequence $\left(L_{i}\right)_{i \geq 1}$ of $\mathcal{F}_{\tau_{i}}$-measurable discrete random variables which represent the absolute value in the number of ticks of the price jump between the $i$ th and the $(i+1)$ st transaction leading to a price change. As explained later, the distribution of this variable will depend on the value of some market quantities at time $\tau_{i}$. Then define $\tau_{i+1}$ recursively as the exit time of $\left(X_{t}\right)_{t>\tau_{i}}$ from the set $\left(d_{k_{i}-L_{i}}, u_{k_{i}+L_{i}-1}\right)$, where $k_{i}=X_{\tau_{i}}^{(\alpha)} / \alpha$, that is,

$$
\begin{equation*}
\tau_{i+1}=\inf \left\{t: t>\tau_{i}, X_{t}=X_{\tau_{i}}^{(\alpha)}-\alpha\left(L_{i}-\frac{1}{2}+\eta\right) \text { or } X_{t}=X_{\tau_{i}}^{(\alpha)}+\alpha\left(L_{i}-\frac{1}{2}+\eta\right)\right\} \tag{2.2}
\end{equation*}
$$

In particular, if $X_{\tau_{i}}=d_{j}$ for some $j \in \mathbb{N}, \tau_{i+1}$ is the exit time of $\left(X_{t}\right)_{t>\tau_{i}}$ from the set $\left(d_{j-L_{i}}, u_{j+L_{i}-1}\right)$, and if $X_{\tau_{i}}=u_{j}$ for some $j \in \mathbb{N}, \tau_{i+1}$ is the exit time of $\left(X_{t}\right)_{t>\tau_{i}}$ from the set $\left(d_{j-L_{i}+1}, u_{j+L_{i}}\right)$.

Let $P_{0}$ be the opening price. We define the last traded price $\left(P_{t}\right)_{t \geq 0}$ as the càdlàg piecewise constant process built from the $\left(\tau_{i}, P_{\tau_{i}}\right)_{i \geq 0}$, where $P_{\tau_{i}}=X_{\tau_{i}}^{(\alpha)}$. It is, of course, quite unrealistic to assume that the transaction prices change exactly at the exit times $\tau_{i}$. However, it enables us to simplify some computations. Moreover, under additional assumptions, the results given in section 4 still hold in the initial model of [23], where delays caused by the reaction times of market participants and/or by the trading process are allowed.

Note that $\alpha L_{i}$ is the absolute value of the price jump between the $i$ th and the $(i+1)$ st transactions with price change and that

$$
\begin{equation*}
P_{\tau_{i}}=X_{\tau_{i}}+\operatorname{sign}\left(X_{\tau_{i}}-X_{\tau_{i-1}}\right)(1 / 2-\eta) \alpha=X_{\tau_{i}}+\operatorname{sign}\left(P_{\tau_{i}}-P_{\tau_{i-1}}\right)(1 / 2-\eta) \alpha \tag{2.3}
\end{equation*}
$$

We eventually explain the conditional distribution of the jump sizes in ticks between consecutive transaction prices. We assume that the jump sizes are bounded (which is empirically not restrictive) and denote by $m$ their maximal value. For $k=1, \ldots, m$ and $t>0$, let

$$
N_{\alpha, t, k}^{(a)}=\sum_{\tau_{i} \leq t} \mathbb{I}_{\left\{\left|X_{\tau_{i}}-X_{\tau_{i-1}}\right|=\alpha(k-1+2 \eta)\right\}} \quad \text { and } \quad N_{\alpha, t, k}^{(c)}=\sum_{\tau_{i} \leq t} \mathbb{I}_{\left\{\left|X_{\tau_{i}}-X_{\tau_{i-1}}\right|=\alpha k\right\}}
$$

be, respectively, the number of alternations and continuations of $k$ ticks. An alternation (continuation) of $k$ ticks is a jump of $k$ ticks whose direction is opposite to (the same as) that of the preceding jump; see Figure 1. Remark that, for small (large) values of $\eta$, one will mainly observe alternations (continuations). The number of price changes in $[0, t]$ is given by

$$
N_{\alpha, t}=\sum_{k=1}^{m}\left(N_{\alpha, t, k}^{(a)}+N_{\alpha, t, k}^{(c)}\right)
$$



Figure 1. Example of trajectories of the efficient price and of the observed price.

Let $\left(\chi_{t}\right)_{t \geq 0}$ be an $M$-dimensional $\mathcal{F}_{t}$-adapted process. For simplicity, we assume that the process $\chi$ is a continuous Itô semimartingale with progressively measurable drift with locally bounded sample paths and a positive $\mathcal{F}_{t}$-adapted volatility matrix whose elements have càdlàg sample paths. We define the filtration $\mathcal{E}$ as the complete right-continuous filtration generated by $\left(X_{t}, \chi_{t}, N_{\alpha, t, k}^{(a)}, N_{\alpha, t, k}^{(c)}, k=1, \ldots, m\right)$. We assume that, conditional on $\mathcal{E}_{\tau_{i}}, L_{i}$ is a discrete random variable on $\llbracket 1, m \rrbracket$ satisfying

$$
\begin{equation*}
\mathbb{P}_{\mathcal{E}_{\tau_{i}}}\left[L_{i}=k\right]=p_{k}\left(\chi_{\tau_{i}}\right), \quad 1 \leq k \leq m, \tag{2.4}
\end{equation*}
$$

for some positive differentiable with bounded derivative functions $p_{k}$. In practice, $\chi_{t}$ may represent quantities related, for example, to the traded volume, the bid-ask spread, or the bid and ask depths. For the applications, specific forms for the $p_{k}$ are given in [22].

### 2.2. Discussion.

- The model with uncertainty zones accommodates the inherent properties of prices, durations, and microstructure noise together with a semimartingale efficient price. In particular, this model allows for discrete prices, a bid-ask bounce, and an inverse relation between durations and volatility. Moreover, the usual behaviors of the autocorrelograms and crosscorrelograms of returns and microstructure noise, both in calendar and tick time, are reproduced. Eventually, this model leads to jumps in the price of several ticks, the size of the jumps
being determined by explanatory variables involving, for example, the order book. Mostly, the model with uncertainty zones is clearly validated on real data. These results are studied in detail in [22].
- As explained in the previous section, $\eta$ quantifies the aversion to price changes (with respect to the tick size) of the market participants. Indeed, $\eta$ controls the width of the uncertainty zones. In the tick unit, the larger the $\eta$, the farther from the last traded price the efficient price has to be so that a price change occurs. In some sense, a small $\eta(<1 / 2)$ means that the tick size appears too large to the market participants and a large $\eta$ means that the tick size appears too small.
- There are several other ways to interpret the parameter $\eta$, notably from a practitioner's perspective. For example, one can think that, in very high frequencies, the order book cannot "follow" the efficient price and is reluctant to price changes. This reluctancy could be characterized by $\eta$. Another possibility is to view $\eta$ as a measure of the usual price depth explored by the transaction volumes.
- It is quite easy to estimate the value of $\eta$; see [23]. Moreover, it is shown in [22] that this parameter remains remarkably stable in time for a large number of assets. Finally, all types of configuration between the tick size, the order of magnitude of the volatility, and the value of $\eta$ can be found in the market.

3. Hedging strategies. We consider a European derivative security with expiration date $T$ and payoff $F\left(X_{T}\right)$, where $F$ is a regular enough payoff function. ${ }^{3}$ We present in this section our benchmark frictionless hedging strategy and two strategies adapted to our uncertainty zone market. We assume in the remainder of this paper that all assets are perfectly divisible and, without loss of generality, that the riskless borrowing and lending rate is zero.
3.1. Benchmark frictionless hedging strategy. The benchmark frictionless hedging strategy is that of an agent deciding (possibly wrongly) that the volatility of the efficient price at time $t$ is equal to $\sigma\left(t, X_{t}\right)$ for a regular enough function $\sigma(t, x)$. If true, such an assumption on the volatility enables one to build a self-financing replicating portfolio of stocks and riskless bonds whose marked-to-model price at time $t$ is of the form $C\left(t, X_{t}\right)$. The function $C$ satisfies

$$
\dot{C}_{t}(t, x)+\frac{1}{2} \sigma^{2}(t, x) x^{2} \ddot{C}_{x x}(t, x)=0, \quad C(T, x)=F(x),
$$

with $\dot{C}_{t}(t, x)=\partial C(t, x) / \partial t, \ddot{C}_{x x}(t, x)=\partial^{2} C(t, x) / \partial x^{2}$; see [4].
Suppose the agent implements this strategy in a frictionless market. It leads to a benchmark frictionless hedging portfolio whose value $\Pi_{t}$ satisfies

$$
\Pi_{t}=C\left(0, X_{0}\right)+\int_{0}^{t} \dot{C}_{x}\left(u, X_{u}\right) \mathrm{d} X_{u} .
$$

Note that, if the model is misspecified, $\Pi_{t}$ is different from $C\left(t, X_{t}\right)$; see [5].

[^37]Finally, we assume that for some $M>0$ there exists a sequence of closed sets $\mathcal{S}_{n} \subset[0, M]$ such that the function $C(t, x)$ satisfies

$$
\left|\frac{\partial^{\gamma+\beta} C(t, x)}{\partial t^{\gamma} \partial x^{\beta}}\right|<+\infty
$$

for all $(t, x) \in\{[0, T-1 / n) \times[1 / n,+\infty)\} \cup\left\{[T-1 / n, T] \times\left\{[1 / n,+\infty) / \mathcal{S}_{n}\right\}\right\}$, and

$$
\rho_{n}=\inf \left\{T-2 / n \leq t \leq T, X_{t} \in \mathcal{S}_{n}\right\}
$$

is such that $\mathbb{P}\left[\rho_{n}>T\right] \rightarrow 1$, with the convention $\inf \{\emptyset\}=+\infty$. Remark that, for example, if the benchmark frictionless hedging portfolio is built thanks to the Black-Scholes model, the preceding assumption holds for a European call with strike $K$ taking $\mathcal{S}_{n}=[K-1 / n, K+1 / n]$.
3.2. Hedging strategies in an uncertainty zone market. We assume the real market is the uncertainty zone market, where one can buy and sell at the last traded price, irrespective of the volume of the transaction and of its sign. So, we naturally impose that the times when the hedging portfolio may be rebalanced are the times where the transaction price moves. ${ }^{4}$ Thus, we assume that the hedging portfolio can be rebalanced only at the transaction times $\tau_{i}$. Therefore, here the trading strategies are piecewise constant. In this setting, we consider strategies such that, if $\tau_{i}$ is a rebalancing time, the number of shares in the risky asset at time $\tau_{i}$ is $\dot{C}_{x}\left(\tau_{i}, X_{\tau_{i}}\right)$. Such a strategy is feasible under the assumption that $\eta$ is known so that $X_{\tau_{i}}$ can be computed. This is not a restrictive assumption since $\eta$ can be easily and accurately estimated; see [23]. Finally, recall that our model treats discreteness of prices and reproduces bid-ask bounce. This can be intrinsically connected to implicit transaction costs in the spirit of [24]. Indeed, assume that in the benchmark hedging strategy the gamma of the derivatives is positive; then, when the investor faces a series of oscillations of the price, the buy (resp., sell) price is systematically higher (resp., lower) than the efficient price; see section 4.2.

In the next section, we consider two hedging strategies: (i) the hedging portfolio is rebalanced every time that the transaction price moves; (ii) the hedging portfolio is rebalanced only once the transaction price has varied by more than a selected value.
4. Asymptotic distributions of the microstructural hedging error. In our setting, the microstructural hedging error is due to the following:

- discrete trading: the hedging portfolio is rebalanced a finite number of times;
- microstructure noise on the price: between two rebalancing times, the variation of the market price (multiple of the tick size) differs from the variation of the efficient price.
We analyze this microstructure hedging error in two steps. First, we assume that there is no microstructure noise on the price but that the trading times are endogenous (for all $i, P_{\tau_{i}}=X_{\tau_{i}}$ ): we expect results more or less similar to those in [2], [12], and [26], where exogenous trading times are considered. Second, we assume the presence of the endogenous microstructure noise and discuss the two hedging strategies.

Our asymptotic results use the notion of stable convergence in law that we recall here. Let $Z_{\alpha}$ be a family of random variables (taking their values in $\mathbb{D}[0, T]$, the space of càdlàg functions

[^38]endowed with the Skorokhod topology $J_{1}$ ). Let $\alpha_{n}$ be a deterministic sequence tending to zero as $n$ tends to infinity and $\mathcal{I}$ be a sub- $\sigma$-field of $\mathcal{F}$. We say that $Z_{\alpha_{n}}$ converges $\mathcal{I}$-stably to $Z$ as $\alpha_{n}$ tends to zero $\left(Z_{\alpha_{n}} \xrightarrow{\mathcal{I}-\mathcal{L}_{s}} Z\right)$ if, for every $\mathcal{I}$-measurable bounded real random variable $V$, ( $V, Z_{\alpha_{n}}$ ) converges in law to ( $V, Z$ ) as $n$ tends to infinity. This is a slightly stronger mode of convergence than the weak convergence; see [14] for details and equivalent definitions. Finally, we say that $Z_{\alpha}$ converges $\mathcal{I}$-stably to $Z$ as $\alpha$ tends to zero if, for any sequence $\alpha_{n}$ tending to zero, $Z_{\alpha_{n}} \xrightarrow{\text { I- }}{ }^{\mathcal{L} s} Z$.

We are now able to state our main results. Our asymptotics is to consider that the tick size is going to zero. Even if the tick size is fixed on the markets, in our model it is a natural way to make the number of rebalancing transactions go to infinity.
4.1. Hedging error without microstructure noise on the price. Let $\phi(t)=\sup \left\{\tau_{i}: \tau_{i}<\right.$ $t\}$. In the absence of microstructure noise on the price, the hedging error is given by

$$
L_{\alpha, t}^{(1)}=\int_{0}^{t}\left[\dot{C}_{x}\left(u, X_{u}\right)-\dot{C}_{x}\left(\phi(u), X_{\phi(u)}\right)\right] \mathrm{d} X_{u} .
$$

Let $f_{t}$ be the limit in probability of $\alpha^{2} N_{\alpha, t}$ as $\alpha \rightarrow 0$. We have (see [23])

$$
f_{t}=\int_{0}^{t} \varphi\left(\chi_{u}\right) \sigma_{u}^{2} X_{u}^{2} \mathrm{~d} u
$$

where

$$
\varphi\left(\chi_{u}\right)=\left(\sum_{j=1}^{m} p_{j}\left(\chi_{u}\right) j(j-1+2 \eta)\right)^{-1} .
$$

The inverse of the function $f$, denoted by $\theta$, is a time change for which the observation times are asymptotically uniformly distributed on any finite interval. Let $\Delta X_{\tau_{i}}=X_{\tau_{i+1}}-X_{\tau_{i}}$, and let $\tau_{i_{t}}=\inf \left\{\tau_{i}, \tau_{i} \geq t\right\}$ and $\tau_{i_{t}+1}$ (resp., $\tau_{i_{t}-1}$ ) be the first $\tau_{i}$ after (resp., before) $\tau_{i_{t}}$. We introduce the asymptotic conditional fourth moment of the normalized price change at time $t$ :

$$
\begin{aligned}
\mu_{4}\left(\chi_{t}\right) & =\lim _{\alpha \rightarrow 0} \mathbb{E}_{\mathcal{E}_{\tau_{t}}}\left[\alpha^{-4}\left(\Delta X_{\tau_{i_{t}}}\right)^{4}\right] \\
& =\sum_{k=1}^{m} k(k-1+2 \eta)\left(k^{2}-k+2 \eta k+4 \eta^{2}-4 \eta+1\right) p_{k}\left(\chi_{t}\right) .
\end{aligned}
$$

Let $\mathcal{I}$ be the filtration generated by the processes $X$ and $\chi$. We have the following result.
Theorem 4.1. As a tends to 0,

$$
\begin{equation*}
N_{\alpha, t}^{1 / 2} L_{\alpha, t}^{(1)} \xrightarrow{\mathcal{I}-\mathcal{\mathcal { L }} s} L_{t}^{(1)}:=f_{t}^{1 / 2} \int_{0}^{t} c_{f_{s}}^{(1)} \mathrm{d} \mathcal{W}_{f_{s}}^{(1)} \tag{4.1}
\end{equation*}
$$

in $\mathbb{D}[0, T]$, where $\mathcal{W}^{(1)}$ is a Brownian motion defined on an extension of the filtered probability space $\left(\Omega,\left(\mathcal{F}_{t}\right)_{t \geq 0}, \mathbb{P}\right)$ and independent of all the preceding quantities, and $c_{s}^{(1)}$ is such that

$$
\left(c_{s}^{(1)}\right)^{2}=\frac{1}{6} \ddot{C}_{x x}^{2}\left(\theta_{s}, X_{\theta_{s}}\right) \mu_{4}\left(\chi_{\theta_{s}}\right) .
$$

We see that the variance of the hedging error is proportional to the reciprocal of the number of rebalancing transactions and depends on the local volatility gamma of the derivative security.

Note that if the hedging portfolio were rebalanced at the exogenous times $\tau_{i}=i \alpha^{2} T$, for $i=1, \ldots, n_{\alpha}=\alpha^{-2}$, then we would have $f_{t}=t, \mu_{4}\left(X_{t}\right)=3 \sigma_{t}^{4} X_{t}^{4}$ and our asymptotic hedging error would be equal to those in [2]. We conclude that, although endogenous transaction times are considered, our result is in agreement with [2] and [12].
4.2. Hedging error with microstructure noise. In the presence of microstructure noise on the price, the transaction prices differ from the efficient prices. The hedging error is now given by

$$
L_{\alpha, t}^{(2)}=\int_{0}^{t} \dot{C}_{x}\left(u, X_{u}\right) \mathrm{d} X_{u}-\int_{0}^{t} \dot{C}_{x}\left(\phi(u), X_{\phi(u)}\right) \mathrm{d} P_{u} .
$$

Let

$$
\pi_{a}\left(\chi_{t}\right)=\lim _{\alpha \rightarrow 0} \mathbb{E}_{\mathcal{E}_{\tau_{i_{t}}}}\left[\mathbb{I}_{\left\{\Delta X_{\tau_{i_{t}}} \Delta X_{\tau_{i_{t}-1}}<0\right\}}\right]=\sum_{k=1}^{m} \frac{k}{2 k-1+2 \eta} p_{k}\left(\chi_{t}\right)
$$

be the asymptotic conditional probability that the next price change at time $t$ is due to an alternation, and let

$$
\mu_{1, a}^{*}\left(\chi_{t}\right)=\lim _{\alpha \rightarrow 0} \mathbb{E}_{\mathcal{E}_{\tau_{i_{t}}}}\left[\alpha^{-1}\left|\Delta X_{\tau_{i_{t}}}\right| \mathbb{I}_{\left\{\Delta X_{\tau_{i_{t}}} \Delta X_{\tau_{i_{t}-1}}<0\right\}}\right]=\sum_{k=1}^{m} \frac{k(k-1+2 \eta)}{2 k-1+2 \eta} p_{k}\left(\chi_{t}\right)
$$

be the asymptotic conditional expectation of the absolute value of the normalized price change at time $t$ when the price change is due to an alternation. We have the following result.

Theorem 4.2. As $\alpha$ tends to 0 ,

$$
L_{\alpha, t}^{(2)} \xrightarrow{\mathcal{I}-\mathcal{L}_{s}} L_{t}^{(2)}:=\int_{0}^{t} a_{f_{s}}^{(2)} \mathrm{d} s+\int_{0}^{t} b_{f_{s}}^{(2)} \mathrm{d} X_{s}+\int_{0}^{t} c_{f_{s}}^{(2)} \mathrm{d} \mathcal{W}_{f_{s}}^{(2)}
$$

in $\mathbb{D}[0, T]$, where $\mathcal{W}^{(2)}$ is a Brownian motion defined on an extension of the filtered probability space $\left(\Omega,\left(\mathcal{F}_{t}\right)_{t \geq 0}, \mathbb{P}\right)$ and independent of all the preceding quantities, and $a_{s}^{(2)}, b_{s}^{(2)}$, and $c_{s}^{(2)}$ are such that

$$
\begin{aligned}
a_{s}^{(2)} & =-(1-2 \eta) \ddot{C}_{x x}\left(\theta_{s}, X_{\theta_{s}}\right) \mu_{1, a}^{*}\left(\chi_{\theta_{s}}\right) \varphi\left(\chi_{\theta_{s}}\right), \\
b_{s}^{(2)} & =(1-2 \eta) \dot{C}_{x}\left(\theta_{s}, X_{\theta_{s}}\right) \mu_{1, a}^{*}\left(\chi_{\theta_{s}}\right) \varphi\left(\chi_{\theta_{s}}\right), \\
\left(c_{s}^{(2)}\right)^{2} & =(1-2 \eta)^{2} \dot{C}_{x}^{2}\left(\theta_{s}, X_{\theta_{s}}\right) \varphi\left(\chi_{\theta_{s}}\right)\left(\pi_{a}\left(\chi_{\theta_{s}}\right) \varphi^{-1}\left(\chi_{\theta_{s}}\right)-\left(\mu_{1, a}^{*}\left(\chi_{\theta_{s}}\right)\right)^{2}\right) .
\end{aligned}
$$

It is worth noticing that the microstructural hedging error process is not renormalized as in the previous case. It means that the hedging error does not vanish even if the number of rebalancing transactions goes to infinity and thus is of the same order of magnitude as the usual tracking error. However, if $\eta=1 / 2$, the changes of the transaction prices coincide with the changes of the efficient prices at the exit times of the uncertainty zones $\left(\Delta P_{\tau_{i}}=\Delta X_{\tau_{i}}\right)$ and the error due to the microstructure noise on the price vanishes.

The first two components of the asymptotic hedging error are quite unusual in this kind of asymptotic distribution theory. The first term can be interpreted as a bias due to implicit transaction costs linked to bid-ask bounce. The second one is due to the asymmetry between alternations and continuations. Indeed, when an alternation occurs, $\Delta P_{\tau_{i}}-\Delta X_{\tau_{i}}=(1-$ $2 \eta) \operatorname{sign}\left(\Delta X_{\tau_{i}}\right)$, while, when a continuation occurs, $\Delta P_{\tau_{i}}-\Delta X_{\tau_{i}}=0$. For further developments on the presence of limiting bias terms, see [20]. Moreover, remark that the quadratic variation of the asymptotic hedging error is

$$
(1-2 \eta)^{2} \int_{0}^{t} \dot{C}_{x}^{2}\left(s, X_{s}\right) \pi_{a}\left(\chi_{s}\right) \mathrm{d} f_{s}
$$

It now depends on the local volatility delta of the derivative security and on the proportion of alternations. Consequently, the variance of the microstructural hedging error increases with the position delta and the proportion of alternation in the price. Indeed, compared with the previous case, one now faces an additional microstructural hedging error of first order.

Finally, recall that the strategy presented in this section is unrealistic in practice. However, it provides us a reference result for comparing with the results of more realistic strategies, which we explain in the next section.
4.3. Optimal rebalancing level in the presence of microstructure noise. We now build strategies where the portfolio is rebalanced when the price moves "significantly." ${ }^{5}$ It is probably more natural than a timely based rebalancing strategy and should reduce the impact of the microstructure noise. So, we now assume that the hedging portfolio is rebalanced only once the price has changed by $l_{\alpha}$ ticks. For simplicity, here we assume that $m=1$. We will choose $l_{\alpha}$ such that $l_{\alpha} \rightarrow \infty$ and $\alpha l_{\alpha} \rightarrow 0$ as $\alpha \rightarrow 0$. In this way, hedging errors due, respectively, to microstructure noise on the price and discrete-time rebalancing will disappear as the tick goes to zero. Note that it is more or less equivalent to considering that the transaction price lies on a subgrid with new tick size $\alpha l_{\alpha}$ and that the hedging portfolio is rebalanced as soon as the transaction price takes a new value on this subgrid.

Let $N_{\alpha, t}^{(l)}$ be the number of rebalancing transactions between 0 and $t$. We define the sequence $\left(\tau_{i}^{(l)}\right)_{i \geq 1}$ recursively by

$$
\tau_{i+1}^{(l)}=\inf \left\{t: t>\tau_{i}^{(l)}, X_{t}=X_{\tau_{i}}^{(\alpha)}-\alpha\left(l_{\alpha}-\frac{1}{2}+\eta\right) \text { or } X_{t}=X_{\tau_{i}}^{(\alpha)}+\alpha\left(l_{\alpha}-\frac{1}{2}+\eta\right)\right\} .
$$

Let $f_{t}^{(l)}$ be the limit in probability of $\left(\alpha l_{\alpha}\right)^{2} N_{\alpha, t}^{(l)}$ as $\alpha \rightarrow 0$. We have

$$
f_{t}^{(l)}=\int_{0}^{t} \sigma_{u}^{2} X_{u}^{2} \mathrm{~d} u
$$

Let $\phi^{(l)}(t)=\sup \left\{\tau_{i}^{(l)}: \tau_{i}^{(l)}<t\right\}$. The hedging error is given by

$$
L_{\alpha, t}^{(3)}=\int_{0}^{t} \dot{C}_{x}\left(u, X_{u}\right) \mathrm{d} X_{u}-\int_{0}^{t} \dot{C}_{x}\left(\phi^{(l)}(u), X_{\phi^{(l)}(u)}\right) \mathrm{d} P_{u}
$$

[^39]Let $\Delta X_{\tau_{i-1}^{(l)}}=X_{\tau_{i}^{(l)}}-X_{\tau_{i-1}^{(l)}}$. This error can be decomposed in the following way:

$$
\begin{aligned}
L_{\alpha, t}^{(3)}= & L_{\alpha, t}^{(3,1)}+L_{\alpha, t}^{(3,2)}+L_{\alpha, t}^{(3,3)} \\
:= & \sum_{i=1}^{N_{\alpha, t}^{(l)}} \int_{\tau_{i-1}^{(l)}}^{\tau_{i}^{(l)}}\left(\dot{C}_{x}\left(u, X_{u}\right)-\dot{C}_{x}\left(\tau_{i-1}^{(l)}, X_{\tau_{i-1}}^{(l)}\right)\right) \mathrm{d} X_{u} \\
& +\sum_{i=1}^{N_{\alpha, t}^{(l)}} \dot{C}_{x}\left(\tau_{i-1}^{(l)}, X_{\tau_{i-1}(l)}^{(l)}\right)\left(\Delta X_{\tau_{i-1}}^{(l)}-\Delta P_{\tau_{i-1}^{(l)}}^{(l)}\right) \\
& +\int_{\substack{\tau_{\alpha, t}^{(l)} \\
N_{\alpha, t}^{(l)}}} \dot{C}_{x}\left(u, X_{u}\right) \mathrm{d} X_{u}-\dot{C}_{x}\left(\tau_{N_{\alpha, t}^{(l)}(l)}^{(l)} X_{\tau_{N_{\alpha, t}}^{(l)}}^{(l)}\right)\left(P_{t}-P_{\substack{\tau_{N, t}^{(l)} \\
N_{\alpha, t}^{(l)}}}\right) .
\end{aligned}
$$

We have the following result.
Theorem 4.3. As a tends to 0,

$$
\begin{aligned}
& \left(\alpha l_{\alpha}\right)^{-1} L_{\alpha, t}^{(3,1)} \xrightarrow[\rightarrow]{\mathcal{I}-\mathcal{L}_{s}} L_{t}^{(3,1)}:=\int_{0}^{t} a_{f_{s}^{(l)}}^{(3,1)} \mathrm{d} s+\int_{0}^{t} b_{f_{s}^{(l)}}^{(3,1)} \mathrm{d} X_{s}+\int_{0}^{t} c_{f_{s}^{(l)}}^{(3,1)} \mathrm{d} \mathcal{W}_{f_{s}^{(l)}}^{(3,1)}, \\
& l_{\alpha} L_{\alpha, t}^{(3,2)} \xrightarrow{\mathcal{I}-\mathcal{L} s} L_{t}^{(3,2)}:=\int_{0}^{t} a_{f_{s}^{(l)}}^{(3,2)} \mathrm{d} s+\int_{0}^{t} b_{f_{s}^{(l)}}^{(3,2)} \mathrm{d} X_{s}+\int_{0}^{t} c_{f_{s}^{(l)}}^{(3,2)} \mathrm{d} \mathcal{W}_{f_{s}^{(l)}}^{(3,2)}, \\
& \left(l_{\alpha} \vee\left(\alpha l_{\alpha}\right)^{-1}\right) L_{\alpha, t}^{(3,3)} \xrightarrow{\mathcal{I}-\mathcal{E}_{s} s} 0
\end{aligned}
$$

in $\mathbb{D}[0, T]$, where $\mathcal{W}^{(3,1)}$ and $\mathcal{W}^{(3,2)}$ are Brownian motions defined on an extension of the filtered probability space $\left(\Omega,\left(\mathcal{F}_{t}\right)_{t \geq 0}, \mathbb{P}\right)$ and independent of all the preceding quantities, $a_{s}^{(3,1)}$, $b_{s}^{(3,1)}$, and $c_{s}^{(3,1)}$ are such that

$$
a_{s}^{(3,1)}=0, \quad b_{s}^{(3,1)}=0, \quad\left(c_{s}^{(3,1)}\right)^{2}=\frac{1}{6} \ddot{C}_{x x}^{2}\left(\theta_{s}, X_{\theta_{s}}\right),
$$

and $a_{s}^{(3,2)}, b_{s}^{(3,2)}$, and $c_{s}^{(3,2)}$ are such that

$$
\begin{aligned}
a_{s}^{(3,2)} & =-(1-2 \eta) \ddot{C}_{x x}\left(\theta_{s}, X_{\theta_{s}}\right), \\
b_{s}^{(3,2)} & =\frac{(1-2 \eta)}{2} \dot{C}_{x}\left(\theta_{s}, X_{\theta_{s}}\right), \\
\left(c_{s}^{(3,2)}\right)^{2} & =\frac{(1-2 \eta)^{2}}{4} \dot{C}_{x}^{2}\left(\theta_{s}, X_{\theta_{s}}\right)
\end{aligned}
$$

The optimal asymptotic rate for $l_{\alpha}$ results from a trade-off between the variances of the renormalized asymptotic errors.

Theorem 4.4. Let $l_{\alpha}=\alpha^{-1 / 2}$. As $\alpha$ tends to 0 ,

$$
\left(N_{\alpha, t}\right)^{1 / 4} L_{\alpha, t}^{(3)} \xrightarrow{\mathcal{I}-\mathcal{L} s} L_{t}^{(3)}:=\left(f_{t}^{(l)}\right)^{1 / 4}\left(\int_{0}^{t} a_{f_{s}^{(l)}}^{(3)} \mathrm{d} s+\int_{0}^{t} b_{f_{s}^{(l)}}^{(3)} \mathrm{d} X_{s}+\int_{0}^{t} c_{f_{s}^{(l)}}^{(3)} \mathrm{d} \mathcal{W}_{f_{s}^{(l)}}^{(3)}\right)
$$

in $\mathbb{D}[0, T]$, where $\mathcal{W}^{(3)}$ is a Brownian motion defined on an extension of the filtered probability space $\left(\Omega,\left(\mathcal{F}_{t}\right)_{t \geq 0}, \mathbb{P}\right)$ and independent of all the preceding quantities, and $b_{s}^{(3)}$ and $c_{s}^{(3)}$ are such that

$$
\begin{aligned}
a_{s}^{(3)} & =-(1-2 \eta) \ddot{C}_{x x}\left(\theta_{s}, X_{\theta_{s}}\right), \\
b_{s}^{(3)} & =\frac{(1-2 \eta)}{2} \dot{C}_{x}\left(\theta_{s}, X_{\theta_{s}}\right), \\
\left(c_{s}^{(3)}\right)^{2} & =\frac{(1-2 \eta)^{2}}{4} \dot{C}_{x}^{2}\left(\theta_{s}, X_{\theta_{s}}\right)+\frac{1}{6} \ddot{C}_{x x}^{2}\left(\theta_{s}, X_{\theta_{s}}\right) .
\end{aligned}
$$

This optimal strategy allows one to reduce the hedging error significantly in the presence of microstructure noise. Interestingly, the quadratic variation of the asymptotic hedging error now depends both on the delta and on the gamma of the derivative security.

Note that the optimal $l_{\alpha}$ is of the same order of magnitude as the square root of the number of times where the hedging portfolio is rebalanced and of the same order of magnitude as the reciprocal of the asymptotic variance of the hedging error. Asymptotically, to decrease the variance by a factor of 2 , one must rebalance four times more often with a target level $\left(\alpha l_{\alpha}\right)$ divided by 2 .
5. A numerical study. In this section, we give numerical results about the hedging errors of a European call with strike $K=100$ and maturity $T=1$. We consider the following model for the underlying asset for the trading period. The efficient price is given by the Black-Scholes dynamics

$$
\mathrm{d} X_{t}=\sigma X_{t} \mathrm{~d} W_{t}, \quad x_{0}=100, \quad t \in[0, T]
$$

where $\sigma=0.01$, and we take $\alpha=0.05, \eta=0.05$, and, for $i \geq 1, L_{i}=1$. These parameters are chosen to be in agreement with real data; see [22]. Finally, we assume the benchmark strategy of the agent is the Black-Scholes strategy with $\sigma=0.01$. Here we give statistics for the number of rebalancings (NR) and the histogram of the hedging error for the different strategies over 1000 Monte Carlo simulations; see Figure 2. Note that $L_{\alpha, T}^{(3)}$ is computed for price moves of 5 ticks.

|  | Average of NR | Standard deviation of NR |
| :---: | :---: | :---: |
| Every price move rebalancing | 3487 | 108 |
| 5 ticks rebalancing | 20.22 | 3.55 |

We clearly see that rebalancing the portfolio each time the price changes induces a strong negative bias in the hedging error. Rebalancing less frequently enables us to significantly improve the hedging error, which is in agreement with the theoretical results.
6. Proofs. The proofs of our four theorems are based on the methodology for limit theorems in the presence of endogenous observation times developed in [23]. In particular, Lemmas $1,2,3,4,6,7$, and 11 in [23] hold since they share the same assumptions as in this paper. The key idea of this methodology is to work in a modified time in which the observation times are equidistant and to use stability properties of the convergence in law in $\mathbb{D}[0, T]$. In a first step, discrete-time processes taking new values at the transaction times are built. This is done considering that the transaction times are equispaced. Then it is established that the


Figure 2. Histograms of $L_{\alpha, T}^{(i)}$ for $i=1$ (in green, error due to discrete trading times), 2 (in red, error due to discrete trading times and microstructure noise), and 3 (in grey, error in the case of "optimal" rebalancing).
piecewise constant processes stably converge in law by using Theorem IX.7.3 in [14]. In [23], the required assumptions for this theorem are shown to hold through Lemma 8, 9, 10, 12, and 13. Note that, for our specific results, only Lemmas 10, 12, and 13 have to be modified, the other ones still being in force (up to obvious modifications). In a second step, a time change that enables us to come back in calendar time is used (Lemma 15 in [23]).
6.1. Preliminary remarks. In all the proofs, $c$ denotes a positive constant that may vary from line to line and $\left(\alpha_{n}\right)_{n \geq 0}$ is a sequence tending to zero. So, we write $\tau_{i, n}$ for $\tau_{i}$ and $L_{i, n}$ for $L_{i}$. We define $\mathcal{E}^{n}$ as the complete right-continuous filtration generated by $\left(X_{t}, \chi_{t}, N_{\alpha_{n}, t, k}^{(a)}, N_{\alpha_{n}, t, k}^{(c)}, k=1, \ldots, m\right)$ and the filtration $\mathcal{H}^{n}$ by

$$
\mathcal{H}_{u}^{n}=\mathcal{E}_{\tau_{\left\lfloor\alpha_{n}^{-2} u\right\rfloor, n}^{n}} .
$$

Moreover, without loss of generality, we consider that the semimartingale $\chi$ is a one-dimensional process of the form

$$
\chi_{t}=\chi_{0}+\int_{0}^{t} a_{u}^{\chi} \mathrm{d} u+\int_{0}^{t} \sigma_{u-}^{\chi} \mathrm{d} \hat{W}_{u} \quad \text { for } t \leq T,
$$

with $\hat{W}$ a Brownian motion on $(\Omega, \mathcal{F}, \mathbb{P})$, and we set $\chi_{t}=\chi_{T}$ for $t>T$.
We write $\mathcal{F}=\mathcal{F}_{1} \otimes \mathcal{F}_{2}$. The processes $X$ and $\chi$ are measurable with respect to $\mathcal{F}_{1}$, and $\mathcal{F}_{2}$ is the filtration generated by a Brownian motion $W^{\prime}$, independent of $\mathcal{F}_{1}$. Let $\Phi$ denote the
cumulative distribution function of a standard Gaussian random variable. We define

$$
\begin{gathered}
g_{t, n}=\sup \left\{\tau_{j, n}: \tau_{j, n}<t\right\}, \\
L_{t}^{\prime}=\sum_{k=1}^{m} k \mathbb{I}\left\{\Phi\left(\frac{W_{t}^{\prime}-W_{g_{t, n}}^{\prime}}{\sqrt{t-g_{t, n}}}\right) \in\left[\sum_{j=1}^{k-1} p_{j}\left(\chi_{t}\right), \sum_{j=1}^{k} p_{j}\left(\chi_{t}\right)\right)\right\},
\end{gathered}
$$

and $L_{i, n}=L_{\tau_{i, n}}^{\prime}$, with the usual convention $\sum_{j=1}^{0} p_{j}\left(\chi_{t}\right)=0$. So defined, for $k=1, \ldots, m$, we have $\mathbb{P}_{\mathcal{E}_{\tau_{i, n}}^{n}}\left[L_{\tau_{i, n}}=k\right]=p_{k}\left(\chi_{\tau_{i, n}}\right)$.

Here we modify the localization procedure and change of probability in [23] slightly. More precisely, we follow the same lines as in [23], but we consider before the change of probability the sequence of processes $X^{n}$ defined on $[0, T]$ by

$$
X_{t}^{n}=X_{t \wedge \rho_{n}}+2 M n \int_{0}^{t} \mathbb{I}_{\{T-2 / n \leq s \leq T\} \cap\left\{\rho_{n}=T-2 / n\right\}} \mathrm{d} s
$$

This process is a continuous Itô semimartingale which coincides with $X$ on $\left[0, \rho_{n}\right]$ and does not take values in $\dot{\mathcal{S}}_{n}$ for $t \in[T-1 / n, T]$. Hence, without restriction, we can prove the results under the following supplementary assumption.

Assumption 1. For all $t \in[0, T], a_{t}=-\sigma_{t-}^{2} / 2$ and there exists a constant $M>0$ such that, for all $\gamma \in \mathbb{N}, \beta \in \mathbb{N}$, and $0 \leq \gamma+\beta \leq 3$,

$$
\left|\log \left(X_{t}\right)\right|+\left|\chi_{t}\right|+\left|\sigma_{t}\right|+\left|\sigma_{t}^{\chi}\right|+\left|a_{t}^{\chi}\right|+\left|\frac{\partial^{\gamma+\beta} C\left(t, X_{t}\right)}{\partial t^{\gamma} \partial x^{\beta}}\right| \leq M
$$

Finally, it is useful to view our price process as a time-changed Brownian motion. For that purpose, we introduce the process $\left(Z_{t}\right)_{t \geq 0}$ with an infinite bracket at infinity defined by

$$
Z_{t}=X_{t \wedge T}+\int_{0}^{t} \mathbb{I}_{s>T} \mathrm{~d} W_{s}
$$

Let

$$
\mathcal{T}(s)=\inf \left\{t \geq 0:\langle Z\rangle_{t}>s\right\}
$$

By the Dubins-Schwarz theorem for continuous local martingales (see, for example, [21, Theorem V.1.6]), there exists an $\mathcal{F}_{\mathcal{T}(s)}$-adapted Brownian motion $\left(B_{s}\right)_{s \geq 0}$ such that, for $t \geq 0$,

$$
B_{\langle Z\rangle_{t}}+x_{0}=Z_{t} .
$$

Hence, for $t \in[0, T], B_{\langle X\rangle_{t}}+x_{0}=X_{t}$. From now on, we redefine $\tau_{i, n}$ and $L_{i, n}$ in the same way as in (2.2) and (2.4) replacing $X_{t}$ by $Z_{t}$. For simplicity, from now on we keep the notation $X$ for $Z$. We set $\nu_{i, n}=\langle X\rangle_{\tau_{i, n}}$ and $\Delta \nu_{i, n}=\nu_{i+1, n}-\nu_{i, n}$.
6.2. Proofs of Theorem 4.1. The hedging error is given by

$$
\begin{aligned}
L_{\alpha_{n}, t}^{(1)}= & \sum_{i=1}^{N_{\alpha_{n}, t}} \int_{\tau_{i, n-1}}^{\tau_{i, n}}\left(\dot{C}_{x}\left(u, X_{u}\right)-\dot{C}_{x}\left(\tau_{i, n-1}, X_{\tau_{i, n-1}}\right)\right) \mathrm{d} X_{u} \\
& +\int_{\tau_{N_{\alpha_{n}, t}}}^{t}\left(\dot{C}_{x}\left(u, X_{u}\right)-\dot{C}_{x}\left(\tau_{\tau_{N_{\alpha_{n}, t}}}, X_{\tau_{N_{\alpha_{n}, t}}}\right)\right) \mathrm{d} X_{u} .
\end{aligned}
$$

Let us define

$$
\begin{aligned}
K_{i}^{(1)}(n) & =\alpha_{n}^{-1} \int_{\tau_{i-1, n}}^{\tau_{i, n}}\left(\dot{C}_{x}\left(u, X_{u}\right)-\dot{C}_{x}\left(\tau_{i-1, n}, X_{\tau_{i-1, n}}\right)\right) \mathrm{d} X_{u} \\
& =\alpha_{n}^{-1} \int_{\nu_{i-1, n}}^{\nu_{i, n}}\left(\dot{C}_{x}\left(\mathcal{T}(v), B_{v}\right)-\dot{C}_{x}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right)\right) \mathrm{d} B_{v}
\end{aligned}
$$

and

$$
K^{(1)}(n)_{u}=\sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} K_{i}^{(1)}(n)
$$

The following lemma is intended to replace Lemma 10 in [23].
Lemma 6.1. For $\varepsilon>0$, as $n \rightarrow \infty$, we have

$$
\sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[K_{i}^{(1)}(n)^{2} \mathbb{I}_{\left\{\left|K_{i}^{(1)}(n)\right|>\varepsilon\right\}}\right] \rightarrow \mathbb{P} 0
$$

Proof. For $\nu_{i-1, n} \leq v \leq \nu_{i, n}$, we have

$$
\begin{aligned}
& \dot{C}_{x}\left(\mathcal{T}(v), B_{v}\right)-\dot{C}_{x}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right) \\
= & \ddot{C}_{x t}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right)\left(\mathcal{T}(v)-\mathcal{T}\left(\nu_{i-1, n}\right)\right) \\
& +\ddot{C}_{x x}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right)\left(B_{v}-B_{\nu_{i-1, n}}\right)+R_{i, n}^{(1)}(v) \\
= & \frac{\ddot{C}_{x t}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right)}{\sigma^{2}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right) B_{\nu_{i-1, n}}^{2}}\left(v-\nu_{i-1, n}\right) \\
& +\ddot{C}_{x x}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right)\left(B_{v}-B_{\nu_{i-1, n}}\right)+R_{i, n}^{(2)}(v),
\end{aligned}
$$

where $\left|R_{i, n}^{(j)}(v)\right| \leq c\left(\alpha_{n}^{2}+\left(\Delta \nu_{i-1, n}\right)^{2}\right)$ for $j=1,2$. Itô's formula gives

$$
\int_{\nu_{i-1, n}}^{\nu_{i, n}}\left(B_{v}-B_{\nu_{i-1, n}}\right) \mathrm{d} B_{v}=\frac{1}{2}\left[\left(\Delta B_{\nu_{i-1, n}}\right)^{2}-\Delta \nu_{i-1, n}\right]
$$

and

$$
\int_{\nu_{i-1, n}}^{\nu_{i, n}}\left(v-\nu_{i-1, n}\right) \mathrm{d} B_{v}=\Delta B_{\nu_{i-1, n}} \Delta \nu_{i-1, n}-\int_{\nu_{i-1, n}}^{\nu_{i, n}}\left(B_{v}-B_{\nu_{i-1, n}}\right) \mathrm{d} v .
$$

It follows that

$$
\left|K_{i}^{(1)}(n)\right| \leq c \alpha_{n}^{-1}\left(\Delta \nu_{i-1, n}+\left(\Delta B_{\nu_{i-1, n}}\right)^{2}\right)+\alpha_{n}^{-1}\left|\int_{\nu_{i-1, n}}^{\nu_{i, n}} R_{i, n}^{(2)}(v) \mathrm{d} B_{v}\right|
$$

By the Bürkholder-Davis-Gundy (BDG) inequality, we get

$$
\begin{aligned}
\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left|\int_{\nu_{i-1, n}}^{\nu_{i, n}} R_{i, n}^{(2)}(v) \mathrm{d} B_{v}\right|^{p}\right] & \leq c_{p} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(\int_{\nu_{i-1, n}}^{\nu_{i, n}}\left(R_{i, n}^{(2)}(v)\right)^{2} \mathrm{~d} v\right)^{p / 2}\right] \\
& \leq c \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(\left(\alpha_{n}^{2}+\left(\Delta \nu_{i-1, n}\right)^{2}\right)^{2}\left(\Delta \nu_{i-1, n}\right)\right)^{p / 2}\right] \\
& \leq c \alpha_{n}^{3 p}
\end{aligned}
$$

And then, for $p \geq 1$

$$
\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left|K_{i}^{(1)}(n)\right|^{p}\right] \leq c \alpha_{n}^{p}
$$

Using together the Cauchy-Schwarz and Markov inequalities, we deduce that

$$
\sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[K_{i}^{(1)}(n)^{2} \mathbb{I}_{\left\{\left|K_{i}(n)\right|>\varepsilon\right\}}\right] \leq c \varepsilon^{-1} \alpha_{n}^{5 / 2}
$$

and the result follows.
We now define the processes $Z$ and $Z(n)$ by $Z_{u}=X_{\theta_{u}}$ and $Z(n)_{u}=Z_{f_{\tau_{\left\lfloor\alpha_{n}^{-2} u\right\rfloor, n}}}$. The following lemma is intended to replace Lemma 12 in [23].

Lemma 6.2. As $n \rightarrow \infty$, we have

$$
\left\langle K^{(1)}(n), Z(n)\right\rangle_{u} \xrightarrow{\mathbb{P}} 0
$$

Proof. First note that $\left\langle K_{1}(n), Z(n)\right\rangle_{u}$ is equal to

$$
\sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[K_{i}^{(1)}(n) \Delta B_{\nu_{i-1, n}}\right]
$$

By Itô's formula, we have

$$
\begin{aligned}
& \alpha_{n} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[K_{i}^{(1)}(n) \Delta B_{\nu_{i-1, n}}\right] \\
= & \frac{\ddot{C}_{x t}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right)}{\sigma^{2}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right) B_{\nu_{i-1, n}}^{2}} \\
& \times \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\Delta \nu_{i-1, n}\left(\Delta B_{\nu_{i-1, n}}\right)^{2}-\Delta B_{\nu_{i-1, n}} \int_{\nu_{i-1, n}}^{\nu_{i, n}}\left(B_{v}-B_{\nu_{i-1, n}}\right) \mathrm{d} v\right] \\
& +\frac{1}{2} \ddot{C}_{x x}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right) \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(\Delta B_{\nu_{i-1, n}}\right)^{3}-\Delta B_{\nu_{i-1, n}} \Delta \nu_{i-1, n}\right] \\
& +\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\Delta B_{\nu_{i-1, n}} \int_{\nu_{i-1, n}}^{\nu_{i, n}} R_{i, n}^{(2)}(v) \mathrm{d} B_{v}\right] .
\end{aligned}
$$

Using the same arguments as in the proof of Lemma 12 in [23], we deduce that

$$
\alpha_{n}^{-1} \sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(\Delta B_{\nu_{i-1, n}}\right)^{3}-\Delta B_{\nu_{i-1, n}} \Delta \nu_{i-1, n}\right] \xrightarrow{\mathbb{P}} 0 .
$$

Moreover, it is easy to see that

$$
\left|\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\Delta \nu_{i-1, n}\left(\Delta B_{\nu_{i-1, n}}\right)^{2}-\Delta B_{\nu_{i-1, n}} \int_{\nu_{i-1, n}}^{\nu_{i, n}}\left(B_{v}-B_{\nu_{i-1, n}}\right) \mathrm{d} v\right]\right| \leq c \alpha_{n}^{4}
$$

and that by the BDG inequality

$$
\left|\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\Delta B_{\nu_{i-1, n}} \int_{\nu_{i-1, n}}^{\nu_{i, n}} R_{i, n}^{(2)}(v) \mathrm{d} B_{v}\right]\right| \leq c \alpha_{n}^{4} .
$$

The result follows.
The following lemma is intended to replace Lemma 13 in [23].
Lemma 6.3. Let

$$
m_{k}=k(k-1+2 \eta)\left(k^{2}-k+2 \eta k+4 \eta^{2}-4 \eta+1\right) .
$$

As $n \rightarrow \infty$, we have

$$
\sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(K_{i}^{(1)}(n)\right)^{2}\right] \xrightarrow{\mathbb{P}} \frac{1}{6} \int_{0}^{\langle X\rangle_{\theta_{u}}} \ddot{C}_{x x}^{2}\left(\mathcal{T}(v), B_{v}\right) \sum_{k=1}^{m} m_{k} p_{k}\left(\chi_{\mathcal{T}(v)}\right) \varphi\left(\chi_{\mathcal{T}(v)}\right) d v
$$

Proof. First, note that
$\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(K_{i}^{(1)}(n)\right)^{2}\right]=\alpha_{n}^{-2} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\int_{\nu_{i-1, n}}^{\nu_{i, n}}\left(\dot{C}_{x}\left(\mathcal{T}(v), B_{v}\right)-\dot{C}_{x}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right)\right)^{2} \mathrm{~d} v\right]$.
Then, by Itô's formula, we have that

$$
\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\int_{\nu_{i-1, n}}^{\nu_{i, n}}\left(B_{v}-B_{\nu_{i-1, n}}\right)^{2} \mathrm{~d} v\right]
$$

is equal to

$$
\frac{1}{6} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(B_{\nu_{i, n}}-B_{\nu_{i-1, n}}\right)^{4}\right]-\frac{2}{3} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\int_{\nu_{i-1, n}}^{\nu_{i, n}}\left(B_{v}-B_{\nu_{i-1, n}}\right)^{3} \mathrm{~d} B_{v}\right] .
$$

Now consider for $t \geq \nu_{i-1, n}$ the martingale

$$
M_{t}=\int_{\nu_{i-1, n}}^{t \wedge \nu_{i, n}}\left(B_{v}-B_{\nu_{i-1, n}}\right)^{3} \mathrm{~d} B_{v} .
$$

We have

$$
\left|M_{t}\right| \leq c\left(\alpha_{n}^{4}+\alpha_{n}^{2} \Delta \nu_{i-1, n}\right)
$$

and

$$
\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\Delta \nu_{i-1, n}\right] \leq c \alpha_{n}^{2} .
$$

By the optional sampling theorem and the dominated convergence theorem, we deduce that

$$
\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{\nu_{i, n}}\right]=0 .
$$

So,

$$
\begin{aligned}
\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\int_{\nu_{i-1, n}}^{\nu_{i, n}}\left(B_{v}-B_{\nu_{i-1, n}}\right)^{2} \mathrm{~d} v\right] & =\frac{1}{6} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(B_{\nu_{i, n}}-B_{\nu_{i-1, n}}\right)^{4}\right] \\
& =\frac{1}{6} \alpha_{n}^{4} \sum_{k=1}^{m} m_{k} p_{k}\left(\chi_{\mathcal{T}\left(\nu_{i-1, n}\right)}\right) .
\end{aligned}
$$

Then

$$
\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\int_{\nu_{i-1, n}}^{\nu_{i, n}}\left(v-\nu_{i-1, n}\right)^{2} \mathrm{~d} v\right]=\frac{1}{3} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(\nu_{i, n}-\nu_{i-1, n}\right)^{3}\right] \leq c \alpha_{n}^{6} .
$$

Moreover,

$$
\left|\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\int_{\nu_{i-1, n}}^{\nu_{i, n}}\left(v-\nu_{i-1, n}\right)\left(B_{v}-B_{\nu_{i-1, n}}\right) \mathrm{d} v\right]\right| \leq c \alpha_{n}^{5} .
$$

It follows that

$$
\sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(K_{i}^{(1)}(n)\right)^{2}\right]=\frac{1}{6} \alpha_{n}^{2} \sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \sum_{k=1}^{m} \ddot{C}_{x x}^{2}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right) m_{k} p_{k}\left(\chi_{\mathcal{T}\left(\nu_{i-1, n}\right)}\right)+R_{n}
$$

where $R_{n}$ goes to zero in probability. Using Lemma 3 in [23], we get

$$
\sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{n}(i-1)}^{n}}\left[\left(K_{i}^{(1)}(n)\right)^{2}\right] \xrightarrow{\mathbb{P}} \frac{1}{6} \int_{0}^{\langle X\rangle_{\theta_{u}}} \ddot{C}_{x x}^{2}\left(\mathcal{T}(v), B_{v}\right) \sum_{k=1}^{m} m_{k} p_{k}\left(\chi_{\mathcal{T}(v)}\right) \varphi\left(\chi_{\mathcal{T}(v)}\right) \mathrm{d} v
$$

An obvious application of the BDG inequality leads to the following lemma.
Lemma 6.4. As $n \rightarrow \infty$, we have

$$
\int_{\tau_{N_{\alpha_{n}}, t}}^{t}\left(\dot{C}_{x}\left(u, X_{u}\right)-\dot{C}_{x}\left(\tau_{N_{\alpha_{n}, t}}, X_{\tau_{N_{\alpha_{n}}, t}}\right)\right) d X_{u} \xrightarrow{\mathbb{P}} 0
$$

Eventually, the proof of Theorem 4.1 follows from Theorem IX.7.3 in [14] by using the same arguments as in Lemma 15 in [23].
6.3. Proof of Theorem 4.2. The hedging error is given by

$$
\begin{aligned}
L_{\alpha_{n}, t}^{(2)}= & \sum_{i=1}^{N_{\alpha_{n}, t}} \int_{\tau_{i-1, n}}^{\tau_{i, n}}\left(\dot{C}_{x}\left(u, X_{u}\right)-\dot{C}_{x}\left(\tau_{i-1, n}, X_{\tau_{i-1, n}}\right)\right) \mathrm{d} X_{u} \\
& +\sum_{i=1}^{N_{\alpha_{n}, t}} \dot{C}_{x}\left(\tau_{i-1, n}, X_{\tau_{i-1, n}}\right)\left(\Delta X_{\tau_{i-1, n}}-\Delta P_{\tau_{i-1, n}}\right) \\
& +\int_{\tau_{N_{\alpha_{n}, t}}}^{t} \dot{C}_{x}\left(u, X_{u}\right) \mathrm{d} X_{u} .
\end{aligned}
$$

It is important to note that

$$
\frac{\Delta P_{\tau_{i-1, n}}}{\Delta X_{\tau_{i-1, n}}}=\sum_{k=1}^{m}\left(\mathbb{I}_{\left\{\left|\Delta X_{\tau_{i-1, n}}\right|=\alpha_{n} k\right\}}+\frac{k}{k-1+2 \eta} \mathbb{I}_{\left\{\left|\Delta X_{\tau_{i-1, n}}\right|=\alpha_{n}(k-1+2 \eta)\right\}}\right) .
$$

Let us define

$$
K_{i}^{(2)}(n)=\alpha_{n} K_{i}^{(1)}(n)+M_{i}^{(1)}(n),
$$

where

$$
M_{i}^{(1)}(n)=D_{i, n} \dot{C}_{x}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right) \Delta B_{\nu_{i-1, n}}
$$

and

$$
D_{i, n}=\frac{1-2 \eta}{k-1+2 \eta} \mathbb{I}_{\left\{\left|\Delta B_{\nu_{i-1, n}}\right|=\alpha_{n}(k-1+2 \eta)\right\} .} .
$$

Note that

$$
L_{\alpha_{n}, t}^{(2)}=\sum_{i=1}^{N_{\alpha_{n}, t}} K_{i}^{(2)}(n)+\int_{\tau_{N_{\alpha_{n}}, t}}^{t} \dot{C}_{x}\left(u, X_{u}\right) \mathrm{d} X_{u}
$$

We have the following lemma.
Lemma 6.5. As $n \rightarrow \infty$, we have

$$
\sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor}\left[\alpha_{n} K_{i}^{(1)}(n)\right] \xrightarrow{\mathbb{P}} 0
$$

Proof. The result follows from Theorem 4.1.
An obvious application of Lemma 3 in [23] leads to the following lemma.
Lemma 6.6. As $n \rightarrow \infty$, we have

$$
\int_{\tau_{N_{\alpha_{n}, t}}}^{t} \dot{C}_{x}\left(u, X_{u}\right) d X_{u} \xrightarrow{\mathbb{P}} 0
$$

Therefore, $L_{\alpha_{n}, t}^{(2)}$ has the same limit as $\sum_{i=1}^{N_{\alpha_{n}, t}} M_{i}^{(1)}(n)$.
We define the sets of the upward and downward barriers by $\mathcal{D}=\cup d_{k}$ and $\mathcal{U}=\cup u_{k}$. We have that $\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(1)}(n)\right]$ is equal to

$$
-\alpha_{n}(1-2 \eta) \dot{C}_{x}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}\left(\nu_{i-1, n}\right)}\right) b_{k}\left(\mathbb{I}_{B_{\nu_{i-1, n}} \in \mathcal{U}}-\mathbb{I}_{B_{\nu_{i-1, n}} \in \mathcal{D}}\right),
$$

with $b_{k}=k(2 k-1+2 \eta)^{-1}$. We see that $M_{i}^{(1)}(n)$ is not conditionally centered. So, a direct use of the conditional expectation does not enable us to derive the limit. Consequently, we approximate $M_{i}^{(1)}(n)$ by the sum of two centered terms plus a bias term. Let

$$
M_{i}^{(2)}(n)=-\alpha_{n} \frac{(1-2 \eta)}{2} \dot{C}_{x}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right)\left(\mathbb{I}_{B_{\nu_{i, n}} \in \mathcal{U}}-\mathbb{I}_{B_{\nu_{i, n}} \in \mathcal{D}}\right) .
$$

We have

$$
\begin{aligned}
& \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(2)}(n)\right] \\
= & -\alpha_{n} \frac{(1-2 \eta)}{2} \dot{C}_{x}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right)\left(1-2 \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}\left(\nu_{i-1, n}\right)}\right) b_{k}\right)\left(\mathbb{I}_{B_{\nu_{i-1}, n} \in \mathcal{U}}-\mathbb{I}_{B_{\nu_{i-1, n}} \in \mathcal{D}}\right) .
\end{aligned}
$$

It follows that

$$
\begin{aligned}
& \sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor}\left(M_{i}^{(2)}(n)-\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(2)}(n)\right]\right) \\
= & o_{P}(1)+\sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(1)}(n)\right] \\
& +\frac{(1-2 \eta)}{2} \alpha_{n} \sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor-1}\left(\dot{C}_{x}\left(\mathcal{T}\left(\nu_{i, n}\right), B_{\nu_{i, n}}\right)-\dot{C}_{x}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right)\right)\left(\mathbb{I}_{B_{\nu_{i, n}} \in \mathcal{U}}-\mathbb{I}_{B_{\nu_{i, n}} \in \mathcal{D}}\right) .
\end{aligned}
$$

Using the fact that $\dot{C}_{x}\left(\mathcal{T}\left(\nu_{i, n}\right), B_{\nu_{i, n}}\right)$ is equal to

$$
\begin{aligned}
& \dot{C}_{x}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right) \\
+ & \frac{\ddot{C}_{x t}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right)}{\sigma^{2}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right) B_{\nu_{i-1, n}}^{2}} \Delta \nu_{i-1, n}+\ddot{C}_{x x}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right) \Delta B_{\nu_{i-1, n}}+R_{i, n}^{(2)}\left(\nu_{i-1, n}\right)
\end{aligned}
$$

and that

$$
\begin{aligned}
\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\Delta B_{\nu_{i-1, n}} \mathbb{I}_{\mathcal{D}_{\nu_{i, n}} \in \mathcal{U}}\right] & =-\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\Delta B_{\nu_{i-1, n}} \mathbb{I}_{B_{\nu_{i, n}} \in \mathcal{D}}\right] \\
& =\alpha_{n} \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}\left(\nu_{i-1, n}\right)}\right) b_{k}(k-1+2 \eta)
\end{aligned}
$$

together with Lemma 9 in [10], we deduce that

$$
\frac{(1-2 \eta)}{2} \alpha_{n} \sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor-1}\left(\dot{C}_{x}\left(\mathcal{T}\left(\nu_{i, n}\right), B_{\nu_{i, n}}\right)-\dot{C}_{x}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right)\right)\left(\mathbb{I}_{B_{\nu_{i, n}} \in \mathcal{U}}-\mathbb{I}_{B_{\nu_{i, n}} \in \mathcal{D}}\right)
$$

tends in probability to $L_{u}$ defined by

$$
L_{u}=(1-2 \eta) \int_{0}^{\langle X\rangle_{\theta_{u}}} \ddot{C}_{x x}\left(\mathcal{T}(v), B_{v}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) b_{k}(k-1+2 \eta) \varphi\left(\chi_{\mathcal{T}(v)}\right) d v
$$

Therefore, $\sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} M_{i}^{(1)}(n)+L_{u}$ is equal to

$$
\sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor}\left(\left[M_{i}^{(1)}(n)-\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(1)}(n)\right]\right]+\left[M_{i}^{(2)}(n)-\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(2)}(n)\right]\right]\right)+o_{P}(1)
$$

So, we define

$$
\tilde{K}_{i}^{(2)}(n)=\left[M_{i}^{(1)}(n)-\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(1)}(n)\right]\right]+\left[M_{i}^{(2)}(n)-\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(2)}(n)\right]\right]
$$

Finally $L_{\alpha_{n}, t}^{(2)}+L_{f_{t}}$ has the same limit as $\sum_{i=1}^{N_{\alpha_{n}, t}} \tilde{K}_{i}^{(2)}(n)$.
The following lemma is intended to replace Lemma 10 in [23].
Lemma 6.7. For $\varepsilon>0$, as $n \rightarrow \infty$, we have

$$
\sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}}\left[\left(\tilde{K}_{i}^{(2)}(n)\right)^{2} \mathbb{I}_{\left\{\left|\tilde{K}_{i}^{(2)}(n)\right|>\varepsilon\right\}}\right] \xrightarrow{\mathbb{P}} 0
$$

Proof. It is easy to see that

$$
\left|\tilde{K}_{i}^{(2)}(n)\right| \leq c \alpha_{n} .
$$

Using together the Cauchy-Schwarz and Markov inequalities, we get

$$
\sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(\tilde{K}_{i}^{(2)}(n)\right)^{2} \mathbb{I}_{\left\{\left|\tilde{K}_{i}^{(2)}(n)\right|>\varepsilon\right\}}\right] \leq c \varepsilon^{-1} \alpha_{n}^{5 / 2}
$$

Let

$$
\tilde{K}^{(2)}(n)_{u}=\sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \tilde{K}_{i}^{(2)}(n) .
$$

The following lemma is intended to replace Lemma 12 in [23].
Lemma 6.8. As $n \rightarrow \infty$, we have

$$
\left\langle\tilde{K}^{(2)}(n), Z(n)\right\rangle_{u} \xrightarrow{\mathbb{P}}(1-2 \eta) \int_{0}^{\langle X\rangle_{\theta_{u}}} \dot{C}_{x}\left(\mathcal{T}(v), B_{v}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) a_{k} \varphi\left(\chi_{\mathcal{T}(v)}\right) \mathrm{d} v
$$

where $a_{k}=b_{k}(k-1+2 \eta)$.
Proof. Note that $\left\langle\tilde{K}^{(2)}(n), Z(n)\right\rangle_{u}$ is equal to

$$
\sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{n}(i-1)}^{n}}\left[\tilde{K}_{i}^{(2)}(n) \Delta B_{\nu_{i-1, n}}\right]
$$

We have
$\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\tilde{K}_{i}^{(2)}(n) \Delta B_{\nu_{i-1, n}}\right]=\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(1)}(n) \Delta B_{\nu_{i-1, n}}\right]+\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(2)}(n) \Delta B_{\nu_{i-1, n}}\right]$.

First,

$$
\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(1)}(n) \Delta B_{\nu_{i-1, n}}\right]=\dot{C}_{x}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right) \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(\Delta B_{\nu_{i-1, n}}\right)^{2} D_{i, n}\right]
$$

and

$$
\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(\Delta B_{\nu_{i-1, n}}\right)^{2} D_{i, n}\right]=\alpha_{n}^{2}(1-2 \eta) \sum_{k=1}^{m} a_{k} p_{k}\left(\chi_{\mathcal{T}\left(\nu_{i-1, n}\right)}\right) .
$$

It follows that

$$
\begin{aligned}
& \sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(1)}(n) \Delta B_{\nu_{i-1, n}}\right] \\
& \xrightarrow{\mathbb{P}}(1-2 \eta) \int_{0}^{\langle X\rangle_{\theta_{u}}} \dot{C}_{x}\left(\mathcal{T}(v), B_{v}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) a_{k} \varphi\left(\chi_{\mathcal{T}(v)}\right) \mathrm{d} v .
\end{aligned}
$$

Second,

$$
\begin{aligned}
& \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(2)}(n) \Delta B_{\nu_{i-1, n}}\right] \\
= & -\alpha_{n} \frac{(1-2 \eta)}{2} \dot{C}_{x}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right) \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\Delta B_{\nu_{i-1, n}}\left(\mathbb{I}_{B_{\nu_{i, n}} \in \mathcal{U}}-\mathbb{I}_{B_{\nu_{i, n}} \in \mathcal{D}}\right)\right] \\
= & -\alpha_{n}^{2}(1-2 \eta) \dot{C}_{x}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}\left(\nu_{i-1, n}\right)}\right) a_{k}\left(\mathbb{I}_{B_{\nu_{i-1, n}} \in \mathcal{U}}-\mathbb{I}_{B_{\nu_{i-1, n}} \in \mathcal{D}}\right),
\end{aligned}
$$

and so, by using the same arguments as in Lemma 12 in [23],

$$
\sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(2)}(n) \Delta B_{\nu_{i-1, n}}\right] \xrightarrow{\mathbb{P}} 0 .
$$

The following lemma is intended to replace Lemma 12 in [23].
Lemma 6.9. As $n \rightarrow \infty$, we have

$$
\sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(\tilde{K}_{i}^{(2)}(n)\right)^{2}\right] \xrightarrow{\mathbb{P}}(1-2 \eta)^{2} \int_{0}^{\langle X\rangle_{\theta_{u}}} \dot{C}_{x}^{2}\left(\mathcal{T}(v), B_{v}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) b_{k} \varphi\left(\chi_{\mathcal{T}(v)}\right) d v
$$

We have

$$
\begin{aligned}
\left(\tilde{K}_{i}^{(2)}(n)\right)^{2}= & {\left[M_{i}^{(1)}(n)-\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(1)}(n)\right]\right]^{2}+\left[M_{i}^{(2)}(n)-\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(2)}(n)\right]\right]^{2} } \\
& +2\left(M_{i}^{(1)}(n)-\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(1)}(n)\right]\right)\left(M_{i}^{(2)}(n)-\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(2)}(n)\right]\right) .
\end{aligned}
$$

(i) We have

$$
\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(M_{i}^{(1)}(n)\right)^{2}\right]=\alpha_{n}^{2}(1-2 \eta)^{2} \dot{C}_{x}^{2}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}\left(\nu_{i-1, n}\right)}\right) b_{k}
$$

and

$$
\left(\mathbb{E}_{\underset{\mathcal{H}_{n}^{2}(i-1)}{n}}\left[M_{i}^{(1)}(n)\right]\right)^{2}=\alpha_{n}^{2}(1-2 \eta)^{2} \dot{C}_{x}^{2}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right)\left(\sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}\left(\nu_{i-1, n}\right)}\right) b_{k}\right)^{2}
$$

It follows that

$$
\begin{aligned}
& \sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(M_{i}^{(1)}(n)-\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(1)}(n)\right]\right)^{2}\right] \\
& \xrightarrow{\mathbb{P}}(1-2 \eta)^{2} \int_{0}^{\langle X\rangle_{\theta_{u}}} \dot{C}_{x}^{2}\left(\mathcal{T}(v), B_{v}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) b_{k} \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) c_{k} \varphi\left(\chi_{\mathcal{T}(v)}\right) \mathrm{d} v,
\end{aligned}
$$

with $c_{k}=(k-1+2 \eta)(2 k-1+2 \eta)^{-1}$.
(ii) Note that

$$
\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(M_{i}^{(2)}(n)\right)^{2}\right]=\alpha_{n}^{2} \frac{(1-2 \eta)^{2}}{4} \dot{C}_{x}^{2}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right)
$$

and

$$
\left(\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(2)}(n)\right]\right)^{2}=\alpha_{n}^{2} \frac{(1-2 \eta)^{2}}{4} \dot{C}_{x}^{2}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right)\left(1-2 \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}\left(\nu_{i-1, n}\right)}\right) b_{k}\right)^{2}
$$

Consequently,

$$
\begin{aligned}
& \sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(M_{i}^{(2)}(n)-\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(2)}(n)\right]\right)^{2}\right] \\
& \xrightarrow{\mathbb{P}}(1-2 \eta)^{2} \int_{0}^{\langle X\rangle_{\theta_{u}}} \dot{C}_{x}^{2}\left(\mathcal{T}(v), B_{v}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) b_{k} \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) c_{k} \varphi\left(\chi_{\mathcal{T}(v)}\right) \mathrm{d} v .
\end{aligned}
$$

(iii) We have

$$
\begin{aligned}
& \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(1)}(n) M_{i}^{(2)}(n)\right] \\
= & -\alpha_{n} \frac{(1-2 \eta)}{2} \dot{C}_{x}^{2}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right) \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[D_{i, n} \Delta B_{\nu_{i-1, n}}\left(\mathbb{I}_{B_{\nu_{i, n}} \in \mathcal{U}}-\mathbb{I}_{B_{\nu_{i, n}} \in \mathcal{D}}\right)\right] \\
= & \alpha_{n}^{2} \frac{(1-2 \eta)^{2}}{2} \dot{C}_{x}^{2}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right) \\
& \times \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}\left(\nu_{i-1, n}\right)}\right) b_{k}(k-1+2 \eta)\left(\mathbb{I}_{B_{\nu_{i-1, n}} \in \mathcal{U}}-\mathbb{I}_{B_{\nu_{i-1, n}} \in \mathcal{D}}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
& \mathbb{E}_{\substack{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}}\left[M_{i}^{(1)}(n)\right] \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(2)}(n)\right] \\
= & \alpha_{n}^{2} \frac{(1-2 \eta)^{2}}{2} \dot{C}_{x}^{2}\left(\mathcal{T}\left(\nu_{i-1, n}\right), B_{\nu_{i-1, n}}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}\left(\nu_{i-1, n}\right)}\right) b_{k}\left(1-2 \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}\left(\nu_{i-1, n}\right)}\right) b_{k}\right) .
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& \sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(M_{i}^{(1)}(n)-\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(1)}(n)\right]\right)\left(M_{i}^{(2)}(n)-\mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[M_{i}^{(2)}(n)\right]\right)\right] \\
& \stackrel{\mathbb{P}}{\rightarrow}-\frac{(1-2 \eta)^{2}}{2} \int_{0}^{\langle X\rangle_{\theta_{u}}} \dot{C}_{x}^{2}\left(\mathcal{T}(v), B_{v}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) b_{k}\left(1-2 \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) b_{k}\right) \varphi\left(\chi_{\mathcal{T}(v)}\right) \mathrm{d} v .
\end{aligned}
$$

(iv) Finally,

$$
\begin{aligned}
& \sum_{i=1}^{\left\lfloor\alpha_{n}^{-2} u\right\rfloor} \mathbb{E}_{\mathcal{H}_{\alpha_{n}^{2}(i-1)}^{n}}\left[\left(\tilde{K}_{i}^{(2)}(n)\right)^{2}\right] \\
& \xrightarrow{\mathbb{P}}(1-2 \eta)^{2} \int_{0}^{\langle X\rangle_{\theta_{u}}} \dot{C}_{x}^{2}\left(\mathcal{T}(v), B_{v}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) b_{k} \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) c_{k} \varphi\left(\chi_{\mathcal{T}(v)}\right) \mathrm{d} v \\
& +(1-2 \eta)^{2} \int_{0}^{\langle X\rangle_{\theta_{u}}} \dot{C}_{x}^{2}\left(\mathcal{T}(v), B_{v}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) b_{k} \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) c_{k} \varphi\left(\chi_{\mathcal{T}(v)}\right) \mathrm{d} v \\
& -(1-2 \eta)^{2} \int_{0}^{\langle X\rangle_{\theta_{u}}} \dot{C}_{x}^{2}\left(\mathcal{T}(v), B_{v}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) b_{k}\left(1-2 \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) b_{k}\right) \varphi\left(\chi_{\mathcal{T}(v)}\right) \mathrm{d} v \\
& =(1-2 \eta)^{2} \int_{0}^{\langle X\rangle_{\theta_{u}}} \dot{C}_{x}^{2}\left(\mathcal{T}(v), B_{v}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) b_{k} \varphi\left(\chi_{\mathcal{T}(v)}\right) \mathrm{d} v .
\end{aligned}
$$

We now use the decompositions ${ }^{6}$ introduced in Theorem IX.7.3 in [14]. By Lemma 8 in [23],

$$
\begin{aligned}
\langle Z(n)\rangle_{u} \xrightarrow{\mathbb{P}}\langle Z\rangle_{u}=\langle X\rangle_{\theta_{u}} & =\int_{0}^{\theta_{u}} X_{v}^{2} \sigma_{v}^{2} \mathrm{~d} v=\int_{0}^{u} \varphi^{-1}\left(\chi_{\theta_{s}}\right) \mathrm{d} s \\
& =\int_{0}^{u} \sum_{k=1}^{m} p_{k}\left(\chi_{\theta_{s}}\right) k(k-1+2 \eta) \mathrm{d} s \\
& =: \int_{0}^{u} v_{s}^{2} \mathrm{~d} s .
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
\left\langle\tilde{K}^{(2)}(n), Z(n)\right\rangle_{u} & \xrightarrow{\mathbb{P}}(1-2 \eta) \int_{0}^{\langle X\rangle_{\theta_{u}}} \dot{C}_{x}\left(\mathcal{T}(v), B_{v}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(\nu)}\right) \frac{k(k-1+2 \eta)}{2 k-1+2 \eta} \varphi\left(\chi_{\mathcal{T}(v)}\right) \mathrm{d} v \\
& =\int_{0}^{u}(1-2 \eta) \dot{C}_{x}\left(\theta_{s}, X_{\theta_{s}}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\theta_{s}}\right) \frac{k(k-1+2 \eta)}{2 k-1+2 \eta} \mathrm{~d} s=: \int_{0}^{u} u_{s} v_{s}^{2} \mathrm{~d} s
\end{aligned}
$$

[^40]and
\[

$$
\begin{aligned}
\left\langle\tilde{K}^{(2)}(n)\right\rangle_{u} & \xrightarrow{\mathbb{P}}(1-2 \eta)^{2} \int_{0}^{\langle X\rangle_{\theta_{u}}} \dot{C}_{x}^{2}\left(\mathcal{T}(v), B_{v}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\mathcal{T}(v)}\right) \frac{k}{(2 k-1+2 \eta)} \varphi\left(\chi_{\mathcal{T}(v)}\right) \mathrm{d} v \\
& =\int_{0}^{u}(1-2 \eta)^{2} \dot{C}_{x}^{2}\left(\theta_{s}, X_{\theta_{s}}\right) \sum_{k=1}^{m} p_{k}\left(\chi_{\theta_{s}}\right) b_{k} \mathrm{~d} s=: \int_{0}^{u}\left(u_{s}^{2} v_{s}^{2}+w_{s}^{2}\right) \mathrm{d} s
\end{aligned}
$$
\]

It follows that

$$
\begin{aligned}
& v_{s}^{2}= \varphi^{-1}\left(\chi_{\theta_{s}}\right) \\
& u_{s}^{2}=(1-2 \eta)^{2} \dot{C}_{x}^{2}\left(\theta_{s}, X_{\theta_{s}}\right)\left(\sum_{k=1}^{m} p_{k}\left(\chi_{\theta_{s}}\right) \frac{k(k-1+2 \eta)}{2 k-1+2 \eta}\right)^{2} \varphi^{2}\left(\chi_{\theta_{s}}\right) \\
& w_{s}^{2}=(1-2 \eta)^{2} \dot{C}_{x}^{2}\left(\theta_{s}, X_{\theta_{s}}\right) \varphi\left(\chi_{\theta_{s}}\right)\left(\sum_{k=1}^{m} p_{k}\left(\chi_{\theta_{s}}\right) \frac{k}{(2 k-1+2 \eta)} \varphi^{-1}\left(\chi_{\theta_{s}}\right)\right. \\
&\left.-\left(\sum_{k=1}^{m} p_{k}\left(\chi_{\theta_{s}}\right) \frac{k(k-1+2 \eta)}{2 k-1+2 \eta}\right)^{2}\right)
\end{aligned}
$$

The quantity $w_{s}^{2}$ is positive since

$$
\frac{k-1+2 \eta}{2 k-1+2 \eta}<\sqrt{\frac{k-1+2 \eta}{2 k-1+2 \eta}}<1
$$

and, by the Cauchy-Schwarz inequality,

$$
\left(\sum_{k=1}^{m} p_{k}\left(\chi_{\theta_{s}}\right) \frac{k \sqrt{k-1+2 \eta}}{\sqrt{2 k-1+2 \eta}}\right)^{2} \leq\left(\sum_{k=1}^{m} p_{k}\left(\chi_{\theta_{s}}\right) \frac{k}{(2 k-1+2 \eta)}\right)\left(\sum_{k=1}^{m} p_{k}\left(\chi_{\theta_{s}}\right) k(k-1+2 \eta)\right)
$$

We set $b_{s}^{(2)}=u_{s}$ and $\left(c_{s}^{(2)}\right)^{2}=w_{s}^{2}$. The proof of Theorem 4.2 follows from Theorem IX.7.3 in [14] and by using the same arguments as in Lemma 15 in [23].

### 6.4. Proof of Theorem 4.3.

First component: $L_{\alpha, t}^{(3,1)}$. Let us recall that

$$
L_{\alpha_{n}, t}^{(3,1)}=\sum_{i=1}^{N_{\alpha n, t}^{(l)}} \int_{\tau_{i-1, n}^{(l)}}^{\tau_{i, n}^{(l)}}\left(\dot{C}_{x}\left(u, X_{u}\right)-\dot{C}_{x}\left(\tau_{i-1, n}^{(l)}, X_{\tau_{i-1, n}^{(l)}}\right)\right) \mathrm{d} X_{u}
$$

Lemma 6.10. As $n \rightarrow \infty$, we have

$$
\left(\alpha_{n} l_{\alpha_{n}}\right)^{2} N_{\alpha_{n}, t}^{(l)} \xrightarrow{\mathbb{P}}\langle X\rangle_{t}
$$

Proof. We have

$$
\left(\alpha_{n} l_{\alpha_{n}}\right)^{2} N_{\alpha_{n}, t}^{(l)}=\left(\alpha_{n} l_{\alpha_{n}}\right)^{2} \sum_{\nu_{i, n}^{(l)} \leq\langle X\rangle_{t}} 1 .
$$

The result follows by using the same arguments as those in Lemmas 1, 3, and 4 in [23].
The proof of the stable convergence in law of $\left(\alpha_{n} l_{\alpha_{n}}\right)^{-1} L_{\alpha_{n}, t}^{(3,1)}$ in Theorem 4.3 is proved in the same way as in Theorem 4.1 by using the previous lemma and by replacing $\alpha_{n}$ by $\alpha_{n} l_{\alpha_{n}}$ (when needed).

Second component: $L_{\alpha, t}^{(3,2)}$. Recall that

$$
L_{\alpha_{n}, t}^{(3,2)}=\sum_{i=1}^{N_{\alpha, n}^{(l)}} \dot{C}_{x}\left(\tau_{i-1, n}^{(l)}, X_{\tau_{i-1, n}^{(l)}}\right)\left(\Delta X_{\tau_{i-1, n}^{(l)}}-\Delta P_{\tau_{i-1, n}^{(l)}}\right)
$$

The proof of the stable convergence in law of $l_{\alpha_{n}} L_{\alpha_{n}, t}^{(3,2)}$ in Theorem 4.3 is proved by using the same arguments as in Theorem 4.2 by choosing $m=l_{\alpha_{n}}$ and $p_{m}\left(\chi_{t}\right)=1$ for all $t$.

Third component: $L_{\alpha, t}^{(3,3)}$. Recall that

$$
L_{\alpha_{n}, t}^{(3,3)}=\int_{\substack{\tau^{(l)} \\ N_{\alpha, t}^{(l)}}}^{t} \dot{C}_{x}\left(u, X_{u}\right) \mathrm{d} X_{u}-\dot{C}_{x}\left(\tau_{N_{\alpha, t}^{(l)}, X_{\tau_{N_{\alpha, t}}^{(l)}}^{(l)}}^{(l)}\right)\left(P_{t}-P_{\tau_{N_{\alpha, t}^{(l)}}^{(l)}}\right)
$$

The proof of the convergence in probability of $\left(l_{\alpha_{n}} \vee\left(\alpha_{n} l_{\alpha_{n}}\right)^{-1}\right) L_{\alpha_{n}, t}^{(3,3)}$ results from the fact that

$$
\left|\left(X_{t}-X_{\substack{\tau_{N_{\alpha n}, t}^{(l)}}}\right)-\left(P_{t}-P_{\substack{\tau_{N_{\alpha, t}}^{(l)}}}\right)\right| \leq c \alpha_{n} .
$$

6.5. Proof of Theorem 4.4. When $l_{\alpha_{n}}=\alpha_{n}^{-1 / 2}$, we have $l_{\alpha_{n}}=\left(\alpha_{n} l_{\alpha_{n}}\right)^{-1}$. So one has to consider the stable convergence in law of $l_{\alpha_{n}}\left(L_{\alpha_{n}, t}^{(3,1)}+L_{\alpha_{n}, t}^{(3,2)}\right)$. The proof of Theorem 4.4 follows by using the same type of arguments as previously. Remark that the quadratic covariation between $L_{t}^{(3,1)}$ and $L_{t}^{(3,2)}$ is null.

Acknowledgment. We are very grateful to two anonymous referees whose relevant comments have significantly improved a former version of the paper.

## REFERENCES

[1] O. E. Barndorff-Nielsen, P. R Hansen, A. Lunde, and N. Shephard, Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise, Econometrica, 76 (2009), pp. 1481-1536.
[2] D. Bertsimas, L. Kogan, and A. W. Lo, When is time continuous?, J. Financial Econ., 55 (2000), pp. 173-204.
[3] U. Cetin, M. Soner, and N. Touzi, Option hedging for small investors under liquidity costs, Finance Stoch., to appear.
[4] D. Duffie, Dynamic Asset Pricing Theory, 3rd ed., Princeton Ser. Finance, Princeton University Press, Princeton, NJ, 2002.
[5] N. El Karoui, M. Jeanblanc-Picqué, and S. Shreve, Robustness of the Black and Scholes formula, Math. Finance, 8 (1998), pp. 93-126.
[6] M. Fukasawa, Discretization Error of Stochastic Integrals, Working paper, 2009.
[7] C. Geiss and S. Geiss, On an approximation problem for stochastic integrals where random time nets do not help, Stochastic Process. Appl., 116 (2006), pp. 407-422.
[8] S. Geiss, Quantitative approximation of certain stochastic integrals, Stoch. Stoch. Rep., 73 (2002), pp. 241-270.
[9] S. Geiss and A. Toivola, Weak convergence of error processes in discretizations of stochastic integrals and Besov spaces, Bernoulli, 15 (2009), pp. 925-954.
[10] V. Genon-Catalot and J. Jacod, On estimating the diffusion coefficient for multidimensional processes, Ann. Inst. H. Poincaré Probab. Statist., 29 (1993), pp. 119-151.
[11] E. Gobet and E. Temam, Discrete time hedging errors for options with irregular pay-offs, Finance Stoch., 5 (2001), pp. 357-367.
[12] T. Hayashi and P. A. Mykland, Hedging errors: An asymptotic approach, Math. Finance, 15 (2005), pp. 309-343.
[13] J. Jacod, Y. Li, P. A. Mykland, M. Podolskij, and M. Vetter, Microstructure noise in the continuous case: The pre-averaging approach, Stochastic Process. Appl., 119 (2009), pp. 2249-2276.
[14] J. Jacod and A. N. Shiryaev, Limit Theorems for Stochastic Processes, 2nd ed., Springer-Verlag, Berlin, 2003.
[15] E. Jouini and H. Kallal, Arbitrage in securities markets with short sales constraints, Math. Finance, 5 (1995), pp. 197-232.
[16] Y. M. Kabanov and G. Last, Hedging under transaction costs in currency markets: A continuous-time model, Math. Finance, 12 (2002), pp. 63-70.
[17] Y. M. Kabanov and M. M. Safarian, On Leland's strategy of option pricing with transactions costs, Finance Stoch., 1 (1997), pp. 239-250.
[18] I. Kalnina and O. B. Linton, Estimating quadratic variation consistently in the presence of correlated measurement error, J. Econometrics, 147 (2008), pp. 47-59.
[19] P. F. Koehl, H. Pham, and N. Touzi, Hedging in discrete time under transaction costs and continuoustime limit, J. Appl. Probab., 36 (1999), pp. 163-178.
[20] Y. Li, P. A. Mykland, E. Renault, L. Zhang, and X. Zheng, Realized Volatility When Sampling Times Are Possibly Endogenous, Working paper, 2009.
[21] D. Revuz and M. Yor, Continuous Martingales and Brownian Motion, 3rd ed., Springer-Verlag, Berlin, 1999.
[22] C. Y. Robert and M. Rosenbaum, A new approach for the dynamics of ultra high frequency data: The model with uncertainty zones, J. Financial Econometrics, to appear.
[23] C. Y. Robert and M. Rosenbaum, Volatility and covariation estimation when microstructure noise and trading times are endogenous, Math. Finance, to appear.
[24] R. Roll, A simple implicit measure of the effective bid-ask spread in an efficient market, J. Finance, 39 (2009), pp. 1127-1139.
[25] M. Rosenbaum, Integrated volatility and round-off error, Bernoulli, 15 (2009), pp. 687-720.
[26] P. Tankov and E. Voltchkova, Asymptotic analysis of hedging errors in models with jumps, Stochastic Process. Appl., 119 (2009), pp. 2004-2027.
[27] L. Zhang, Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach, Bernoulli, 12 (2006), pp. 1019-1043.
[28] L. Zhang, P. A. Mykland, and Y. Aït-Sahalia, A tale of two time scales: Determining integrated volatility with noisy high-frequency data, J. Amer. Statist. Assoc., 100 (2005), pp. 1394-1411.

# Option Pricing in Hilbert Space-Valued Jump-Diffusion Models Using Partial Integro-Differential Equations* 

Peter Hepperger ${ }^{\dagger}$


#### Abstract

Hilbert space-valued jump-diffusion models are employed for various markets and derivatives. Examples include swaptions, which depend on continuous forward curves, and basket options on stocks. Usually, no analytical pricing formulas are available for such products. Numerical methods, on the other hand, suffer from exponentially increasing computational effort with increasing dimension of the problem, the "curse of dimension." In this paper, we present an efficient approach using partial integro-differential equations. The key to this method is a dimension reduction technique based on a Karhunen-Loève expansion, which is also known as proper orthogonal decomposition. Using the eigenvectors of a covariance operator, the differential equation is projected to a low-dimensional problem. Convergence results for the projection are given, and the numerical aspects of the implementation are discussed. An approximate solution is computed using a sparse grid combination technique and discontinuous Galerkin discretization. The main goal of this article is to combine the different analytical and numerical techniques needed, presenting a computationally feasible method for pricing European options. Numerical experiments show the effectiveness of the algorithm.


Key words. Hilbert space-valued stochastic analysis, infinite-dimensional jump diffusion, partial integrodifferential equation, proper orthogonal decomposition, basket options, electricity swaps

AMS subject classifications. 91B25, 60H35, 35R15, 35R09
DOI. 10.1137/09077271X

## 1. Introduction.

Pricing with partial differential equations. The pricing of options written on a single asset by means of partial differential equations (PDEs) is a well-studied problem. Efficient numerical solutions have been proposed not only for the classical Black-Scholes setting but also for Lévy-driven processes (cf., e.g., $[24,12,13]$ and the references therein). Due to the jump part, an additional nonlocal term occurs when using Lévy models. Thus, partial integro-differential equations (PIDEs) have to be solved instead of plain PDEs.

The pricing problem becomes much more involved when the derivative (e.g., a basket option) depends on more than one asset. PIDE methods have recently been extended to such multidimensional settings $[28,36]$. The main problem here is the exponentially increasing computational effort for high-dimensional problems. While this curse of dimension can be reduced effectively by applying sparse grid discretizations, numerically feasible dimensions $n$ are still moderate, usually $n \leq 10$. On the other hand, real world products often imply much higher dimensional equations. In fact, derivatives (e.g., swaptions) may even depend on

[^41]a continuum of "assets" and thus be infinite dimensional. This makes dimension reduction techniques an interesting tool. To motivate the subsequently presented techniques, consider the following example, which naturally arises in energy markets.

Motivating example: Electricity swaptions. Since the liberalization of many European electricity markets during the 1990s, producers, consumers, and speculators have traded in electricity on energy exchanges. The Scandinavian Nordpool, the European Energy Exchange (EEX) in Germany, and the Amsterdam Power Exchange (APX) are the largest European trading centers for electricity. Traded products include spot, forward, and futures contracts and options on these. The most liquidly traded underlyings are contracts of futures type. These are agreements traded at time $t$ for a constant delivery of 1 MW of electricity over a certain future period of time $\left[T_{1}, T_{2}\right]$, while in return a fixed rate $F\left(t ; T_{1}, T_{2}\right)$ is paid during this delivery period. These products are also called electricity swaps. The relation of spot and forward prices is not clearly defined for electricity due to its nonstorability [3, 4, 6]. This difficulty can be avoided by directly modeling the forward curve under a risk neutral measure [ 1,20 ], similar to the Heath-Jarrow-Morton approach for bond markets. For every maturity $u \in\left[T_{1}, T_{2}\right]$, let

$$
S(t, u):=\lim _{v \rightarrow u} F(t ; u, v)
$$

be the corresponding value of the forward curve at time $t \leq u$. In practice, the forward curve is constructed from a discrete set of available market prices [5]. Several authors propose exponential additive forward curve models of diffusion or jump-diffusion type [7, 20], i.e.,

$$
\begin{align*}
X(t, u) & =\int_{0}^{t} \gamma(s, u) d s+\sum_{k=1}^{n_{W}} \int_{0}^{t} \sigma_{k}(s, u) d W_{k}(s)+\sum_{k=1}^{n_{J}} \int_{0}^{t} \int_{\mathbb{R}} \eta_{k}(s, u) y \widetilde{M}_{k}(d y, d s),  \tag{1.1}\\
S(t, u) & =S(0, u) \exp (X(t, u))
\end{align*}
$$

where $W_{k}$ are scalar Brownian motions and $\widetilde{M}_{k}$ are compensated random jump measures, all of them independent. Assuming a constant interest rate $r \geq 0$, a European call option on the swap $F\left(T ; T_{1}, T_{2}\right)$ with strike rate $K$ and maturity $T \leq T_{1}$ has the discounted price

$$
\begin{aligned}
V(t) & =e^{-r T} E\left[\left(T_{2}-T_{1}\right)\left(F\left(T ; T_{1}, T_{2}\right)-K\right)^{+} \mid \mathcal{F}_{t}\right] \\
& =e^{-r T}\left(T_{2}-T_{1}\right) E\left[\left(\int_{T_{1}}^{T_{2}} w\left(u ; T_{1}, T_{2}\right) S(T, u) d u-K\right)^{+} \mid \mathcal{F}_{t}\right]
\end{aligned}
$$

at time $t \leq T$, where

$$
w\left(u ; T_{1}, T_{2}\right)=e^{-r u} / \int_{T_{1}}^{T_{2}} e^{-r s} d s
$$

is a deterministic, nonnegative discounting factor. In general, there is no explicit representation for the probability density of the integral term inside the conditional expectation. This makes pricing more difficult than in the one-dimensional standard setting. One possible valuation method is to use log-normal approximation formulas [9]. These are fast but sometimes give poor results in the presence of jumps. Numerical Monte Carlo simulations, on the other hand, are precise but rather time consuming since they converge slowly. In the present work,
we will discuss an efficient numerical approximation method which is based on solving Hilbert space-valued PIDEs and dimension reduction techniques.

Outline of the article. Both basket options and options depending on forward curves can be viewed as special cases of Hilbert space-valued problems. The main goal of this article is to present an efficient method for pricing European options on such derivatives. In section 2, we state the general model and hypotheses for our approach. The corresponding Hilbert space-valued PIDE for pricing European options is derived in section 2.3. This PIDE can be approximated with finite-dimensional equations. To this end, we apply Karhunen-Loève approximation, which is closely related to proper orthogonal decomposition and factor analysis. In section 3, the dimension reduction method is described. We show how to construct an optimal set of approximating basis functions by solving an eigenvector problem. This basis set is then used to transform the PIDE, and the main results of this work are presented. After proving existence and uniqueness, convergence results and error bounds are given for the solutions of the approximating problems. In section 4, we discuss how to perform the dimension reduction numerically, using a Galerkin approach and the sparse grid combination technique. Finally, we apply the pricing method to test problems (a basket option and an electricity swaption) and demonstrate the efficiency of the method with numerical experiments in section 5.
2. Hilbert space-valued jump diffusion. We now state the Hilbert space-valued model used throughout this article. For a definition of stochastic processes and integration in Hilbert spaces with respect to Brownian motion, see, e.g., [15, 22]. An overview of Poisson random measures in Hilbert spaces can be found in [17], and the Lévy case is treated in [25]. Infinitedimensional stochastic analysis and its applications to interest-rate theory and Heath-JarrowMorton models are presented in [10].
2.1. Exponential additive model. Let $D \subset \mathbb{R}^{m}$ and let $\mu_{D}$ be a measure on $D$. Then

$$
H:=L^{2}\left(D, \mu_{D}\right)
$$

is a separable Hilbert space. For every $h \in H$, we denote the corresponding norm by

$$
\|h\|_{H}:=\sqrt{\int_{D}[h(u)]^{2} \mu_{D}(u)} .
$$

For electricity swaptions, we choose $D=\left[T_{1}, T_{2}\right]$, and $\mu_{D}=\lambda_{D}$ is the Lebesgue measure. Then $H$ can be interpreted as the space of forward curves on the delivery period $D$. For basket options, we choose $D$ to be a finite (though possibly large) index set with $n$ entries, each index corresponding to one asset. The measure $\mu_{D}$ is then simply the counting measure, and the norm $\|\cdot\|_{H}$ is the Euclidean norm on $\mathbb{R}^{n}$.

Consider the $H$-valued exponential-additive stochastic model

$$
\begin{align*}
X(t) & =\int_{0}^{t} \gamma(s) d s+\int_{0}^{t} \sigma(s) d W(s)+\int_{0}^{t} \int_{H}[\eta(s)](\xi) \widetilde{M}(d \xi, d s),  \tag{2.1}\\
S(t) & =S(0) \exp (X(t))
\end{align*}
$$

for $t \in[0, T]$, with initial value $S(0) \in H$. Subsequently, we will write $f(t, u):=[f(t)](u)$ for every $f:[0, T] \rightarrow H, u \in D$, and similarly $g(t, h):=[g(t)](h)$ for every $g:[0, T] \rightarrow$ $\mathrm{L}(H, H), h \in H$. The exponential function in the model is defined pointwise, i.e., $S(t, u)=$ $S(0, u) \exp (X(t, u))$ for a.e. $u \in D$. The diffusion part is driven by an $H$-valued Wiener process $W$ whose covariance is described by a symmetric, nonnegative definite trace class operator $Q$. The driving process $\widetilde{M}$ for the jump part is the compensated random measure of an $H$-valued compound Poisson process

$$
J(t)=\sum_{i=1}^{N(t)} Y_{i},
$$

which is independent of $W$. Here, $N$ denotes a Poisson process with intensity $\lambda$, and $Y_{i} \sim P^{Y}$ $(i=1,2, \ldots)$ are independent and identically distributed (i.i.d.) on $H$. The corresponding Lévy measure is denoted by $\nu=\lambda P^{Y}$. Further, denote by $\mathrm{L}(H, H)$ the space of all bounded linear operators on $H$ and let $\gamma:[0, T] \rightarrow H, \sigma:[0, T] \rightarrow \mathrm{L}(H, H)$, and $\eta:[0, T] \rightarrow \mathrm{L}(H, H)$ be deterministic functions. In particular, this model generalizes the electricity swaption model (1.1) driven by a sum of scalar processes.

The following hypothesis is assumed to hold throughout this article. For an introduction to time-dependent Bochner spaces, such as $\mathrm{L}^{2}(0, T ; H)$, see [16, Chap. 5.9].

Assumption 2.1. Suppose that the second exponential moment of the jump distribution $Y$ exists:

$$
E\left[e^{2\|Y\|_{H}}\right]=\int_{H} e^{2\|\xi\|_{H}} P^{Y}(d \xi)<\infty .
$$

Assume further that $\|\eta(t)\|_{\mathrm{L}(H, H)} \leq 1$ for a.e. $t \in[0, T], \gamma \in \mathrm{L}^{2}(0, T ; H)$, and $\sigma \in \mathrm{L}^{2}(0, T$; $\mathrm{L}(H, H))$.

Under Assumption 2.1, $(X(t))_{t \geq 0}$ is an additive process with finite activity and finite expectation. This simplifies notation, since no truncation of large jumps is needed in the characteristic function.

Theorem 2.2. The process $(X(t))_{t \geq 0}$ is square-integrable:

$$
\begin{equation*}
\sup _{0 \leq t \leq T} E\left[\|X(t)\|_{H}^{2}\right]<\infty . \tag{2.2}
\end{equation*}
$$

Proof. The definition of the process $(X(t))_{t \geq 0}$ yields

$$
E\left[\|X(t)\|_{H}^{2}\right] \leq 3 E\left[\left\|\int_{0}^{t} \gamma(s) d s\right\|_{H}^{2}+\left\|\int_{0}^{t} \sigma(s) d W(s)\right\|_{H}^{2}+\left\|\int_{0}^{t} \int_{H} \eta(s, \xi) \widetilde{M}(d \xi, d s)\right\|_{H}^{2}\right] .
$$

We now apply three different results to the three integrals on the right-hand side. For the first one, we use the basic properties of Bochner integrals and Jensen's inequality to obtain

$$
\left\|\int_{0}^{t} \gamma(s) d s\right\|_{H}^{2} \leq \int_{0}^{t}\|\gamma(s)\|_{H}^{2} d s \leq\|\gamma\|_{\mathrm{L}^{2}(0, T ; H)}^{2} .
$$

By the definition of the integral with respect to $H$-valued Gaussian processes and the results from [15, eqs. (4.8) and (4.10)], we have

$$
E\left\|\int_{0}^{t} \sigma(s) d W(s)\right\|_{H}^{2} \leq(\operatorname{tr} Q) E \int_{0}^{t}\|\sigma(s)\|_{\mathrm{L}(H, H)}^{2} d s=(\operatorname{tr} Q)\|\sigma\|_{\mathrm{L}^{2}(0, T ; \mathrm{L}(H, H))}^{2}
$$

where $\operatorname{tr} Q$ denotes the trace of the covariance operator of $W$. Finally, from Young's inequality and [18, Prop. 3.3] we get

$$
\begin{aligned}
E\left[\left\|\int_{0}^{t} \int_{H} \eta(s, \xi) \widetilde{M}(d \xi, d s)\right\|_{H}^{2}\right] & \leq C \int_{0}^{t} \int_{H}\|\eta(s, \xi)\|_{H}^{2} \nu(d \xi) d s \\
& \leq C \lambda T \int_{H}\|y\|_{H}^{2} P^{Y}(d y)
\end{aligned}
$$

Combining the above estimates and employing Assumption 2.1 yields (2.2), since the righthand side in each estimate is independent of $t$.

Theorem 2.3. The characteristic function of $X(t)$ is given by

$$
\begin{align*}
& E\left[e^{i\langle X(t), h\rangle_{H}}\right]=\exp \left[i\left\langle\int_{0}^{t} \gamma(s) d s, h\right\rangle_{H}-\frac{1}{2}\left\langle\left[\int_{0}^{t} \sigma(s) Q \sigma^{*}(s) d s\right](h), h\right\rangle_{H}\right.  \tag{2.3}\\
&\left.+\int_{0}^{t} \int_{H}\left(e^{i\langle\eta(s, \xi), h\rangle_{H}}-1-i\langle\eta(s, \xi), h\rangle_{H}\right) \nu(d \xi) d s\right]
\end{align*}
$$

for every $h \in H$, where $\sigma^{*}(s)$ is the adjoint operator of $\sigma(s)$.
Proof. The drift $\gamma$ is deterministic, and so the first term on the right-hand side of (2.3) is trivial. Since we have finite second moments by Theorem 2.2, we may apply [22, Thm. 4] to obtain the characteristic function of the diffusion and jump parts. It remains to verify the expression for the covariance operator of the diffusion. Applying [15, Prop. 4.13] yields

$$
E\left[\left\langle\int_{0}^{t} \sigma(s) d W(s), h_{1}\right\rangle_{H}\left\langle\int_{0}^{t} \sigma(s) d W(s), h_{2}\right\rangle_{H}\right]=\left\langle\left[\int_{0}^{t} \sigma(s) Q \sigma^{*}(s) d s\right]\left(h_{1}\right), h_{2}\right\rangle_{H}
$$

for every $h_{1}, h_{2} \in H$. The integral on the right-hand side is a Bochner integral with values in $\mathrm{L}(H, H)$.
2.2. Equivalent exponential Lévy model. As already stated, the main goal of this article is pricing European options depending on $S(T)=S_{0} \exp (X(T))$. Since by definition the payoff of such an option depends only on the value of $X$ at time $t=T$, the characteristic function of $X(T)$ completely determines the price of the product. Therefore, it is possible to construct a (time homogeneous) Lévy process $\left(X_{L}(t)\right)_{t \geq 0}$ which produces the same terminal distribution and thus the same price (cf. [31, Chap. 11]). We denote the Borel sets on $H$ by $\mathcal{B}(H)$ and the indicator function of a set $B \in \mathcal{B}(H)$ by $\chi_{B}$. The Lévy-Khinchin triplet of the

Lévy process is given by

$$
\begin{align*}
A_{L}\left(h_{1}, h_{2}\right) & =\frac{1}{T}\left\langle\left[\int_{0}^{T} \sigma(s) Q \sigma^{*}(s) d s\right]\left(h_{1}\right), h_{2}\right\rangle_{H}  \tag{2.4}\\
\gamma_{L} & =\frac{1}{T} \int_{0}^{T} \gamma(t) d t \\
\nu_{L}(B) & =\frac{1}{T} \int_{0}^{T} \int_{H} \chi_{B}(\eta(t, \xi)) \nu(d \xi) d t \quad \text { for } B \in \mathcal{B}(H),
\end{align*}
$$

where $A_{L}: H \times H \rightarrow \mathbb{R}$ is a bilinear covariance operator, $\gamma_{L} \in H$, and $\nu_{L}$ is a finite activity Lévy measure on $H$. Note that the resulting characteristic function of $X_{L}$ at time $t=T$ is identical to (2.3).

Let $I=\{1,2, \ldots, \operatorname{dim} H\}$, if $H$ has finite dimension, or $I=\mathbb{N}$ otherwise. In order to obtain an explicit Lévy representation for $X(T)$, define

$$
Q_{L}: \begin{cases}H & \rightarrow H, \\ h & \mapsto \frac{1}{T}\left[\int_{0}^{T} \sigma(s) Q \sigma^{*}(s) d s\right](h) .\end{cases}
$$

This is a symmetric nonnegative definite operator with finite trace, since, by construction and by the proof of Theorem 2.2, we have

$$
\begin{aligned}
\sum_{l \in I}\left\langle Q_{L} p_{l}, p_{l}\right\rangle_{H} & =\sum_{l \in I} \frac{1}{T} E\left[\left\langle\int_{0}^{t} \sigma(s) d W(s), p_{l}\right\rangle_{H}\left\langle\int_{0}^{t} \sigma(s) d W(s), p_{l}\right\rangle_{H}\right] \\
& =\frac{1}{T} E\left\|\int_{0}^{t} \sigma(s) d W(s)\right\|_{H}^{2} \leq(\operatorname{tr} Q)\|\sigma\|_{\mathrm{L}^{2}(0, T ; \mathrm{L}(H, H))}^{2}
\end{aligned}
$$

for every orthonormal basis $\left(p_{l}\right)_{l \in I}$ of $H$. Consequently, the operator $Q_{L}$ is compact by [15, Prop. C.3] and, in particular, a trace class operator. By [14, Thm. 1.2.1], there is a unique Gaussian probability measure with mean 0 and covariance operator $Q_{L}$. Moreover, there is a corresponding $Q_{L}$-Wiener process by [15, Prop. 4.2], which we denote by $W_{L}$.

In addition, we introduce the $H$-valued compound Poisson process

$$
J_{L}(t)=\sum_{i=1}^{N(t)} Y_{L, i}
$$

where $Y_{L, i} \sim P^{Y_{L}}$ are i.i.d. random variables with values in $H$, and

$$
P^{Y_{L}}\left(Y_{L, i} \in B\right)=\frac{1}{T} \int_{0}^{T} \int_{H} \chi_{B}\left(\eta_{k}(t, \xi)\right) \nu(d \xi) d t
$$

for every Borel set $B \in \mathcal{B}(H)$. The random measure corresponding to $J_{L}$ is denoted by $M_{L}$ and its compensated version by $\widetilde{M}_{L}$. Combined, we obtain the Lévy process

$$
X_{L}(t)=\Gamma t+W_{L}(t)+\int_{0}^{t} \int_{H} \zeta \widetilde{M}_{L}(d \zeta, d s)
$$

with the same distribution as $X$ at $t=T$. The last two summands are $H$-martingales. We will use this model subsequently in place of (2.1), since it is completely equivalent with respect to European option pricing.

Due to the existence of second moments (Theorem 2.2), the bounded linear covariance operator

$$
\mathcal{C}_{X}: \begin{cases}H & \rightarrow H^{\prime} \cong H  \tag{2.5}\\ h & \mapsto E\left[\left\langle X_{L}(T)-E\left[X_{L}(T)\right], h\right\rangle_{H}\left\langle X_{L}(T)-E\left[X_{L}(T)\right], \cdot\right\rangle_{H}\right]\end{cases}
$$

is well defined, and $E\left[X_{L}(T)\right]=\Gamma T$. For notational convenience, define the centered process

$$
Z_{L}(t):=X_{L}(t)-E\left[X_{L}(t)\right]=X_{L}(t)-\Gamma t, \quad t \in[0, T] .
$$

The following theorem gives a summary of the properties of $\mathcal{C}_{X}$.
Theorem 2.4. The operator $\mathcal{C}_{X}$ defined in (2.5) is a symmetric and nonnegative definite trace class operator (and thus compact).

Proof. Since $\mathcal{C}_{X}$ is a covariance operator, it is symmetric and nonnegative definite by definition. It remains to prove that the trace of $\mathcal{C}_{X}$ is finite. To this end, let $\left(p_{l}\right)_{l \in I}$ be an arbitrary orthonormal base of $H$. Then, by dominated convergence,

$$
\begin{aligned}
\operatorname{tr} \mathcal{C}_{X} & =\sum_{l \in I}\left\langle\mathcal{C}_{X} p_{l}, p_{l}\right\rangle_{H}=\sum_{l \in I} E\left[\left\langle Z_{L}(T), p_{l}\right\rangle_{H}^{2}\right]=E\left[\sum_{l \in I}\left\langle Z_{L}(T), p_{l}\right\rangle_{H}^{2}\right] \\
& =E\left[\left\|Z_{L}(T)\right\|_{H}^{2}\right]<\infty
\end{aligned}
$$

holds. We use [15, Prop. C.3] again to conclude that $\mathcal{C}_{X}$ is compact and thus a trace class operator.

We denote the eigenspace of $\mathcal{C}_{X}$ corresponding to eigenvalue 0 by $E_{0}\left(\mathcal{C}_{X}\right)$. Its orthogonal complement $E_{0}\left(\mathcal{C}_{X}\right)^{\perp}$ is the linear span of all eigenvectors corresponding to positive eigenvalues. This is the subspace of $H$ to which the centered process $Z_{L}$ is restricted almost surely, since

$$
E\left[\left\langle Z_{L}(t), h\right\rangle_{H}^{2}\right]=0 \quad \text { for every } h \in E_{0}\left(\mathcal{C}_{X}\right) \text { and a.e. } t \geq 0
$$

Finally, in analogy to the requirement of a nonvanishing volatility in one-dimensional jump-diffusion models, an assumption on the $Q_{L}$-Wiener process $W_{L}$ is needed.

Assumption 2.5. Assume that the restriction of $Q_{L}$ to the subspace $E_{0}\left(\mathcal{C}_{X}\right)^{\perp} \subset H$ is positive definite, i.e.,

$$
\left\langle Q_{L} h, h\right\rangle_{H}>0 \quad \text { for every } h \in E_{0}\left(\mathcal{C}_{X}\right)^{\perp} \backslash\{0\} .
$$

This means that $Z_{L}$ has a nonvanishing Brownian component for all directions in $H$, which are not almost surely orthogonal to the trajectory of the process. This is necessary for the coercivity property (Gårding's inequality) of the PDE which we are going to derive in the next section.
2.3. Hilbert space-valued PIDE. We consider a European option depending on the value of $S$ at time $t=T$. Since $S(T)$ is a deterministic function of $X(T)$, and $X(T)$ is distributed like $X_{L}(T)=\Gamma T+Z_{L}(T)$, we may equivalently consider a European option whose payoff $G$ is a function of $Z_{L}$. Hence, if $z \in H$ is the value of $Z_{L}$ at time $t$, the value $V$ of the option at time $t \leq T$, discounted to time 0 , is described by

$$
\begin{equation*}
V(t, z):=e^{-r T} E\left[G\left(z+Z_{L}(T-t)\right)\right] . \tag{2.6}
\end{equation*}
$$

Note that $V$ is a martingale under the risk neutral measure. In this section, an Itô formula for Hilbert space-valued random variables and the martingale property of $V$ are employed to derive a PIDE for $V$.

First, let us recall the definition of derivatives on Hilbert spaces. We denote by $D_{z} V(t, z) \in$ $L(H, \mathbb{R})$ and $D_{z}^{2} V(t, z) \in L(H, H)$ continuous linear operators such that

$$
V(t, z+\zeta)=V(t, z)+\left[D_{z} V(t, z)\right](\zeta)+\frac{1}{2}\left\langle\left[D_{z}^{2} V(t, z)\right](\zeta), \zeta\right\rangle_{H}+o\left(\|\zeta\|_{H}^{2}\right)
$$

for every $\zeta \in H$. It is often convenient to identify $D_{z}^{2} V(t, z)$ with a bilinear form on $H \times H$, setting

$$
\left[D_{z}^{2} V(t, z)\right]\left(\zeta_{1}, \zeta_{2}\right):=\left\langle\left[D_{z}^{2} V(t, z)\right]\left(\zeta_{1}\right), \zeta_{2}\right\rangle_{H} .
$$

The partial derivative with respect to time is denoted with $\partial_{t} V(t, z)$. We assume the following regularity condition for $V$, which is, in particular, a prerequisite for Itô's formula. However, note that this hypothesis is not necessary for the convergence results in section 3.4.

Assumption 2.6. Suppose that $V \in C^{1,2}((0, T) \times H, \mathbb{R}) \cap C([0, T] \times H, \mathbb{R})$; i.e., $V$ is continuously differentiable with respect to $t$ and twice continuously differentiable with respect to z. Moreover, assume that the operator norms $\left\|D_{z}^{2} V(t, z)\right\|,\left\|D_{z} V(t, z)\right\|$, and $\left\|\partial_{t} V(t, z)\right\|$ are bounded.

As a direct consequence of this assumption, $V$ satisfies the Lipschitz condition

$$
|V(t, z)-V(t, z+\zeta)| \leq K_{V}\|\zeta\|_{H} \quad \text { for every } \zeta \in H
$$

with constant $K_{V}:=\sup _{(s, y) \in[0, T] \times H}\left\|D_{z} V(s, y)\right\|$. We are now able to calculate the stochastic dynamics of $V$ using Itô's formula.

Theorem 2.7. The discounted price $V$ given by (2.6) satisfies

$$
\begin{align*}
d V & \left(t, Z_{L}(t)\right) \\
= & \partial_{t} V\left(t, Z_{L}(t-)\right) d t+\frac{1}{2} \operatorname{tr}\left(\left[D_{z}^{2} V\left(t, Z_{L}(t-)\right)\right] Q_{L}\right) d t \\
& +\int_{H}\left\{V\left(t, Z_{L}(t-)+\zeta\right)-V\left(t, Z_{L}(t-)\right)-\left[D_{z} V\left(t, Z_{L}(t-)\right)\right](\zeta)\right\} \nu_{L}(d \zeta) d t  \tag{2.7}\\
& +D_{z} V\left(t, Z_{L}(t-)\right) d W_{L}(t)+\int_{H}\left[V\left(t, Z_{L}(t-)+\zeta\right)-V\left(t, Z_{L}(t-)\right)\right] \widetilde{M}_{L}(d \zeta, d t) .
\end{align*}
$$

Proof. By Itô's formula [22, Thm. 3], we obtain

$$
\begin{align*}
& V\left(t, Z_{L}(t)\right)  \tag{2.8}\\
&= V\left(0, Z_{L}(0)\right)+\int_{0}^{t} \partial_{t} V\left(s, Z_{L}(s-)\right) d s \\
&+\frac{1}{2}\left\langle\left\langle\int_{0} D_{z}^{2} V\left(s, Z_{L}(s-)\right) d W_{L}(s) ; W_{L}(\cdot)\right\rangle\right\rangle_{t} \\
&+\int_{0}^{t} D_{z} V\left(s, Z_{L}(s-)\right) d Z_{L}(s) \\
&+\sum_{0 \leq s \leq t}\left[V\left(s, Z_{L}(s-)+\Delta Z_{L}(s)\right)-V\left(s, Z_{L}(s-)\right)-\left[D_{z} V\left(s, Z_{L}(s-)\right)\right]\left(\Delta Z_{L}(s)\right)\right]
\end{align*}
$$

where $\left\langle\langle X ; Y\rangle\right.$ denotes the predictable quadratic covariation of $\langle X, Y\rangle_{H}$.
We first calculate the covariation. From [15, Cor. 4.14], we know that

$$
\begin{aligned}
& E\left\langle\int_{t_{1}}^{t_{2}} D_{z}^{2} V\left(s, Z_{L}(s-)\right) d W_{L}(s), W_{L}\left(t_{2}\right)-W_{L}\left(t_{1}\right)\right\rangle_{H} \\
& \quad=\int_{t_{1}}^{t_{2}} \operatorname{tr}\left(\left[D_{z}^{2} V\left(s, Z_{L}(s-)\right)\right] Q_{L}\right) d s
\end{aligned}
$$

for every $0 \leq t_{1} \leq t_{2}$. Consequently, by independence of the increments of $W_{L}$, we get

$$
\begin{equation*}
\left\langle\left\langle\int_{0}^{\cdot} D_{z}^{2} V\left(s, Z_{L}(s-)\right) d W_{L}(s) ; W_{L}(\cdot)\right\rangle\right\rangle_{t}=\int_{0}^{t} \operatorname{tr}\left(\left[D_{z}^{2} V\left(s, Z_{L}(s-)\right)\right] Q_{L}\right) d s \tag{2.9}
\end{equation*}
$$

For the next term in (2.8), we use the dynamics of $Z_{L}$ to obtain

$$
\begin{align*}
& \int_{0}^{t} D_{z} V\left(s, Z_{L}(s-)\right) d Z_{L}(s) \\
&= \int_{0}^{t} D_{z} V\left(s, Z_{L}(s-)\right) d\left[W_{L}(s)+\int_{0}^{s} \int_{H} \zeta \widetilde{M}_{L}\left(d \zeta, d s_{2}\right)\right] \\
&= \int_{0}^{t} D_{z} V\left(s, Z_{L}(s-)\right) d W_{L}(s)  \tag{2.10}\\
&+\int_{0}^{t} D_{z} V\left(s, Z_{L}(s-)\right) d\left[\sum_{0 \leq s_{2} \leq s} \Delta Z_{L}\left(s_{2}\right)-\int_{0}^{s} \int_{H} \zeta \nu_{L}(d \zeta) d s_{2}\right] .
\end{align*}
$$

A theorem for interchanging linear operators and Bochner integrals [16, App. E, Thm. 8] yields

$$
\begin{equation*}
\int_{0}^{t} D_{z} V\left(s, Z_{L}(s-)\right) d\left[\int_{0}^{s} \int_{H} \zeta \nu_{L}(d \zeta) d s_{2}\right]=\int_{0}^{t} \int_{H}\left[D_{z} V\left(s, Z_{L}(s-)\right)\right](\zeta) \nu_{L}(d \zeta) d s \tag{2.11}
\end{equation*}
$$

Plugging (2.9), (2.10), and (2.11) into the Itô dynamics (2.8) finishes the proof.

In order to get a slightly more explicit form of the trace expression in (2.7), let $\left(p_{l}\right)_{l \in I}$ be an arbitrary orthonormal basis of $H$. By definition of the trace, this yields

$$
\operatorname{tr}\left(\left[D_{z}^{2} V\left(t, Z_{L}(t-)\right)\right] Q_{L}\right)=\sum_{l \in I}\left[D_{z}^{2} V\left(t, Z_{L}(t-)\right)\right]\left(Q_{L} p_{l}, p_{l}\right) .
$$

Theorem 2.8. The discounted price $V$ of a European option with payoff $G\left(Z_{L}(T)\right)$ at maturity $T$ satisfies the PIDE

$$
\begin{align*}
-\partial_{t} V(t, z)= & \frac{1}{2} \sum_{l \in I}\left[D_{z}^{2} V(t, z)\right]\left(Q_{L} p_{l}, p_{l}\right)  \tag{2.12}\\
& +\int_{H}\left\{V(t, z+\zeta)-V(t, z)-\left[D_{z} V(t, z)\right](\zeta)\right\} \nu_{L}(d \zeta),
\end{align*}
$$

with terminal condition

$$
V(T, z)=e^{-r T} G(z),
$$

for a.e. $t \in(0, T), z \in E_{0}\left(\mathcal{C}_{X}\right)^{\perp}$.
Proof. We employ Theorem 2.7. The penultimate term in (2.7),

$$
\int_{0}^{t} D_{z} V\left(s, Z_{L}(s-)\right) d W_{L}(s)
$$

is a martingale by [15, Thm. 4.12]. In order to show that the integral with respect to the compensated Poisson measure is a martingale too, we apply [29, Thm. 3.11]. The prerequisite for this theorem is a strong integrability condition, which is satisfied due to [29, Thm. 3.12], since

$$
\int_{0}^{t} \int_{H} E\left|V\left(s, Z_{L}(s-)+\zeta\right)-V\left(s, Z_{L}(s-)\right)\right| \nu_{L}(d \zeta) d s \leq t \int_{H} K_{V}\|\zeta\|_{H} \nu_{L}(d \zeta)<\infty
$$

The remaining integral terms in (2.7) are continuous in $t$ and of finite variation. Consequently, the martingale property of $V$, together with the fact that continuous martingales of finite variation are almost surely constant [27, Thm. 27], yields the PIDE.
3. Dimension reduction for the PIDE. The main goal of this section is to introduce a low-dimensional approximation of the $H$-valued process $X_{L}$. To this end, the KarhunenLoève expansion of $X_{L}(T)$ is used, which is in fact identical to proper orthogonal decomposition (POD). It is also closely related to principal component analysis (which is commonly used for data analysis) and factor analysis (which uses additional error terms in the decomposition). All of these methods are based on the construction of a small set of orthogonal basis elements which can be used to approximate $X_{L}$ in some $L^{2}$-norm. For an overview of POD methods in the context of deterministic differential equations, see [21]. An introduction to KarhunenLoève expansions of stochastic processes can be found in [23, Chap. 37]. Numerical aspects of the method and most of the theory needed here are presented in [32]. After deriving the low-dimensional approximating PIDE, we will show existence and uniqueness of a solution. Finally, we study convergence of the calculated option prices and give error estimates.
3.1. Karhunen-Loève approximation for jump diffusion. While principal component analysis and factor analysis are usually applied to analyze empirical data, we apply the Karhunen-Loève method directly to our model. The dimension reduction takes place in the state space of the previously $H$-valued process. Let us now give a mathematically precise formulation of what is meant by "approximating $X_{L}(T)$."

Definition 3.1. A sequence of orthonormal elements $p_{l} \in H=L^{2}(D), l \in I$, is called $a$ POD basis for $X_{L}(T)$ if it solves the minimization problem

$$
\min _{\left\langle p_{i}, p_{j}\right\rangle_{H}=\delta_{i j}} E\left[\left\|X_{L}(T)-\left(E\left[X_{L}(T)\right]+\sum_{l=1}^{d} p_{l}\left\langle Z_{L}(T), p_{l}\right\rangle_{H}\right)\right\|_{H}^{2}\right]
$$

for every $d \in I$.
In other words, a POD basis is a set of deterministic orthonormal functions such that we expect the projection of the random vector $Z_{L}(T)=X_{L}(T)-E\left[X_{L}(T)\right] \in H$ onto the first $d$ elements of this basis to be a good approximation.

Remark 3.2. Since we are interested in pricing European options, we include only the value of $X_{L}$ at time $t=T$ in Definition 3.1. However, we could approximate the whole trajectory of $X$ as well by using the difference to its projection in the space $\mathrm{L}^{2}(0, T ; H)$. This may be useful for pricing path-dependent derivatives, which we will study separately in future work.

Approximation with a POD basis is equivalent to using the partial sum of the first $d$ elements of a Karhunen-Loève expansion, which itself is closely connected to the eigenvector problem of the covariance operator $\mathcal{C}_{X}$ defined in (2.5). The following theorem shows that the eigenvectors of $\mathcal{C}_{X}$ are indeed the POD basis we are looking for.

Theorem 3.3. A sequence of orthonormal eigenvectors $\left(p_{l}\right)_{l \in I}$ of the operator $\mathcal{C}_{X}$, ordered by the size of the corresponding eigenvalues $\mu_{1} \geq \mu_{2} \geq \cdots \geq 0$, solves the maximization problem

$$
\max _{\left\langle p_{i}, p_{j}\right\rangle_{H}=\delta_{i j}} \sum_{l=1}^{d}\left\langle\mathcal{C}_{X} p_{l}, p_{l}\right\rangle_{H}
$$

for every $d \in I$. The maximum value is

$$
\sum_{l=1}^{d}\left\langle\mathcal{C}_{X} p_{l}, p_{l}\right\rangle_{H}=\sum_{l=1}^{d} \mu_{l} .
$$

Moreover, the eigenvectors are a POD basis in the sense of Definition 3.1, and the expectation of the projection error is

$$
E\left[\left\|Z_{L}(T)-\sum_{l=1}^{d} p_{l}\left\langle Z_{L}(T), p_{l}\right\rangle_{H}\right\|_{H}^{2}\right]=\sum_{l=d+1}^{\operatorname{dim} H} \mu_{l} .
$$

Proof. This is an application of [32, Thm. 2.7 and Prop. 2.8].
3.2. Projection on $d$-dimensional subspace. Subsequently, let $\left(p_{l}\right)_{l \in I}$ and $\left(\mu_{l}\right)_{l \in I}$ denote the orthonormal basis and eigenvalues from Theorem 3.3. Further, let

$$
U_{d}:=\operatorname{span}\left\{p_{1}, p_{2}, \ldots, p_{d}\right\} \subset H
$$

be the $d$-dimensional subspace spanned by the eigenvectors corresponding to the largest eigenvalues. We will assume that $d \leq \operatorname{dim}\left(E_{0}\left(\mathcal{C}_{X}\right)^{\perp}\right)$, i.e., $\mu_{1} \geq \cdots \geq \mu_{d}>0$, as there is no need to include eigenvectors of the covariance operator corresponding to eigenvalue 0 . Indeed, one may project the PIDE to $E_{0}\left(\mathcal{C}_{X}\right)^{\perp}$ without any error, since the projection of $Z_{L}$ on $E_{0}\left(\mathcal{C}_{X}\right)$ is almost surely 0 . Define the projection operator

$$
\mathcal{P}_{d}: \begin{cases}H & \rightarrow U_{d} \cong \mathbb{R}^{d}, \\ z & \mapsto x:=\sum_{l=1}^{d}\left\langle z, p_{l}\right\rangle_{H} p_{l} .\end{cases}
$$

We identify $U_{d}$ with $\mathbb{R}^{d}$ via the isometry

$$
\iota:\left\{\begin{align*}
\left(U_{d},\|\cdot\|_{H}\right) & \rightarrow\left(\mathbb{R}^{d},\|\cdot\|\right),  \tag{3.1}\\
x & \mapsto\left(\left\langle x, p_{l}\right\rangle_{H}\right)_{l=1}^{d}
\end{align*}\right.
$$

In particular, we identify $\mathcal{P}_{d} z$ with the sequence $\left(\left\langle z, p_{l}\right\rangle_{H}\right)_{l=1}^{d}$.
We introduce the finite-dimensional approximation

$$
\begin{equation*}
V_{d}(t, x):=e^{-r T} E\left[G\left(x+\mathcal{P}_{d} Z_{L}(T-t)\right)\right] \tag{3.2}
\end{equation*}
$$

for $x \in U_{d} \cong \mathbb{R}^{d}$. We do not assume a finite-dimensional analogue of the regularity assumption (Assumption 2.6), which is hard to verify in practice. Instead, we impose a simple condition on the payoff function.

Assumption 3.4. Suppose that the payoff function $G$ is Lipschitz continuous on $H$ with Lipschitz constant $K_{G}$.

Remark 3.5. Assumption 3.4 is not necessarily satisfied for payoffs depending on the exponential of $Z_{L}(T)$, e.g., a plain call option depending on $S(T)$. However, this can be remedied easily. In the specific case of a call, we can apply a put-call parity. More generally, every payoff can be truncated to a bounded domain (e.g., by multiplying with a smooth cutoff function). A payoff function has finite expectation; hence the error introduced by truncation is arbitrarily small. Since we have to localize the computational domain for any numerical calculation anyway (compare section 4.2), Assumption 3.4 is no substantial restriction.

The following theorem shows that Assumption 3.4 is actually enough to recover the regularity of $V_{d}$.

Theorem 3.6. The finite-dimensional approximation $V_{d}:[0, T] \times U_{d} \rightarrow \mathbb{R}$ defined in (3.2) satisfies $V_{d} \in C^{1,2}\left((0, T) \times U_{d}, \mathbb{R}\right) \cap C\left([0, T] \times U_{d}, \mathbb{R}\right)$. Moreover, the partial derivatives $\partial_{x_{i}} V_{d}(t, x), \partial_{x_{i}} \partial_{x_{j}} V_{d}(t, x)$, and $\partial_{t} V_{d}(t, x)$ are functions of at most linear growth in $\|x\|$ for $i, j=1, \ldots, d$.

Proof. The first step of the proof is to show the existence of a smooth density for the random variable $\mathcal{P}_{d} Z_{L}(t)$ for a.e. $t \geq 0$. To achieve this, a fast decay condition for its characteristic function

$$
\hat{\mu}_{t}(x):=E\left(e^{i\left\langle x, \mathcal{P}_{d} Z_{L}(t)\right\rangle}\right) \in \mathbb{C}
$$

is needed. We have

$$
\begin{aligned}
E\left[\left\langle\mathcal{P}_{d} W_{L}(t), x_{1}\right\rangle_{U_{d}}\right. & \left.\left\langle\mathcal{P}_{d} W_{L}(t), x_{2}\right\rangle_{U_{d}}\right] \\
& =\sum_{k, l=1}^{d}\left\langle x_{1}, p_{k}\right\rangle_{U_{d}}\left\langle x_{2}, p_{l}\right\rangle_{U_{d}} E\left[\left\langle W_{L}(t), p_{k}\right\rangle_{H}\left\langle W_{L}(t), p_{l}\right\rangle_{H}\right]=\left\langle Q_{L} x_{1}, x_{2}\right\rangle_{H}
\end{aligned}
$$

for every $x_{1}, x_{2} \in U_{d}$. Thus, the covariance operator for the diffusion part of $\mathcal{P}_{d} Z_{L}$ is given by $\mathcal{P}_{d} Q_{L} \mathcal{P}_{d}$. The same arguments as in the proof of Theorem 2.3 yield

$$
\hat{\mu}_{t}(x)=\exp \left(-\frac{1}{2} t\left\langle\mathcal{P}_{d} Q_{L} x, x\right\rangle_{U_{d}}+\int_{0}^{t} \int_{H}\left(e^{i\left\langle\mathcal{P}_{d} \eta(s, \zeta), x\right\rangle_{U_{d}}}-1-i\left\langle\mathcal{P}_{d} \eta(s, \zeta), x\right\rangle_{U_{d}}\right) \nu_{L}(d \zeta) d s\right)
$$

for every $x \in U_{d}$. Using Assumptions 2.1 and 2.5, this implies

$$
\begin{aligned}
\left|\hat{\mu}_{t}(x)\right| \leq & \exp \left(-\frac{1}{2} t\left\langle\mathcal{P}_{d} Q_{L} x, x\right\rangle_{H}\right. \\
& \left.+\int_{0}^{t} \int_{H}\left|e^{i\left\langle\mathcal{P}_{d} \eta(s, \zeta), x\right\rangle_{H}}-1-i\left\langle\mathcal{P}_{d} \eta(s, \zeta), x\right\rangle_{H}\right| \nu_{L}(d \zeta) d s\right) \\
\leq & \exp \left[t\left(-\frac{1}{2} C_{1}\|x\|^{2}+C_{2}\|x\|+C_{3}\right)\right]
\end{aligned}
$$

with positive constants $C_{1}, C_{2}$, and $C_{3}$ depending on $d$. In particular, we have

$$
\lim _{\|x\| \rightarrow \infty}\|x\|^{n} \hat{\mu}_{t}(x)=0 \quad \text { for every } n \in \mathbb{N} .
$$

Similarly, we obtain

$$
\left|\partial_{x}^{\alpha} \hat{\mu}_{t}(x)\right| \leq p_{\alpha}(t,\|x\|)\left|\hat{\mu}_{t}(x)\right| \quad \text { and } \quad\left|\partial_{x}^{\alpha} \partial_{t} \hat{\mu}_{t}(x)\right| \leq q_{\alpha}(t,\|x\|)\left|\hat{\mu}_{t}(x)\right|
$$

for every multiindex $\alpha \in \mathbb{N}_{0}^{d}$, where $p_{\alpha}$ and $q_{\alpha}$ are polynomials. Consequently, for every $t \in(0, T), \hat{\mu}_{t}$ and $\partial_{t} \hat{\mu}_{t}$ are elements of the Schwartz space

$$
\mathcal{S}\left(\mathbb{R}^{d}\right)=\left\{f \in C^{\infty}\left(\mathbb{R}^{d}\right): \lim _{\|x\| \rightarrow \infty}\|x\|^{n} \partial^{\alpha} f(x)=0 \text { for every } \alpha \in \mathbb{N}_{0}^{d}, n \in \mathbb{N}_{0}\right\} .
$$

From [31, Prop. 28.1], we know that $\mathcal{P}_{d} Z_{L}(t)$ has a density $g_{t} \in C^{\infty}\left(\mathbb{R}^{d}\right)$, given by

$$
g_{t}(y):=(2 \pi)^{-d} \int_{\mathbb{R}^{d}} e^{-i\langle y, x\rangle} \hat{\mu}_{t}(x) d x
$$

Moreover, by the properties of $\hat{\mu}_{t}$ and [35, Thm. V.2.8], we obtain $g_{t} \in \mathcal{S}\left(\mathbb{R}^{d}\right)$ and $\partial_{t} g_{t} \in \mathcal{S}\left(\mathbb{R}^{d}\right)$. Finally, note that $V_{d}$ can be written as a convolution of the payoff and the density:

$$
V_{d}(t, x)=e^{-r T} \int_{\mathbb{R}^{d}} G(x+y) g_{T-t}(y) d y=e^{-r T} \int_{\mathbb{R}^{d}} G(y) g_{T-t}(y-x) d y .
$$

Due to Assumption 3.4, we have $|G(y)| \leq|G(0)|+K_{G}\|y\|$. Hence, for $x \in U_{d}, t \in(0, T)$, we may compute

$$
\partial_{x}^{\alpha} V_{d}(t, x)=-e^{-r T} \int_{\mathbb{R}^{d}} G(x+y) \partial_{x}^{\alpha} g_{T-t}(y) d y
$$

and

$$
\partial_{t} V_{d}(t, x)=-e^{-r T} \int_{\mathbb{R}^{d}} G(x+y) \partial_{t} g_{T-t}(y) d y
$$

for every $\alpha \in \mathbb{N}_{0}^{d}$. This proves continuity of the derivatives. In addition, we obtain

$$
\begin{aligned}
\left|\partial_{x}^{\alpha} V_{d}(t, x)\right| & \leq e^{-r T} \int_{\mathbb{R}^{d}}|G(y+x)|\left|\partial_{x}^{\alpha} g_{T-t}(y)\right| d y \\
& \left.\leq \int_{\mathbb{R}^{d}}| | G(0) \mid+K_{G}\|x\|+K_{G}\|y\|\right)\left|\partial_{x}^{\alpha} g_{T-t}(y)\right| d y
\end{aligned}
$$

for every $\alpha \in \mathbb{N}_{0}^{d}$. Similarly,

$$
\begin{aligned}
\left|\partial_{t} V_{d}(t, x)\right| & =e^{-r T}\left|\int_{\mathbb{R}^{d}} G(y) \partial_{t} g_{T-t}(y-x) d y\right| \\
& \left.\leq \int_{\mathbb{R}^{d}}| | G(0) \mid+K_{G}\|x\|+K_{G}\|y\|\right)\left|\partial_{t} g_{T-t}(y)\right| d y .
\end{aligned}
$$

Thus, the growth condition is shown.
It remains to prove that $V_{d}$ is also continuous for $t \rightarrow T$. This is, however, a direct consequence of the fact that

$$
\lim _{t \rightarrow 0} E\left[\left\|Z_{L}(t)\right\|_{H}\right]=0
$$

and thus

$$
\begin{aligned}
\left|V_{d}(t, x)-V_{d}(T, x)\right| & \leq e^{-r T} E\left[\left|G\left(x+\mathcal{P}_{d} Z_{L}(T-t)\right)-G(x)\right|\right] \\
& \leq C E\left[\left\|Z_{L}(T-t)\right\|_{H}\right] \rightarrow 0 \quad \text { for } t \rightarrow T .
\end{aligned}
$$

Theorem 3.7. The function $V_{d}$ defined in (3.2) is a classical solution of the finite-dimensional PIDE

$$
\begin{align*}
-\partial_{t} V_{d}(t, x)= & \frac{1}{2} \sum_{l=1}^{d}\left[D_{x}^{2} V_{d}(t, x)\right]\left(\mathcal{P}_{d} Q_{L} p_{l}, \mathcal{P}_{d} p_{l}\right)  \tag{3.3}\\
& +\int_{H}\left\{V_{d}\left(t, x+\mathcal{P}_{d} \zeta\right)-V_{d}(t, x)-\left[D_{x} V_{d}(t, x)\right]\left(\mathcal{P}_{d} \zeta\right)\right\} \nu_{L}(d \zeta)
\end{align*}
$$

with terminal condition

$$
V_{d}(T, x)=V(T, x)=e^{-r T} G(x),
$$

for $t \in(0, T), x \in U_{d}$.
Proof. The stochastic dynamics of $\mathcal{P}_{d} Z_{L}(t)$ are given by

$$
d\left(\mathcal{P}_{d} Z_{L}(t)\right)=d\left(\mathcal{P}_{d} W_{L}(t)\right)+\int_{H} \mathcal{P}_{d} \zeta \widetilde{M}_{L}(d \zeta, d s)
$$

The process $\mathcal{P}_{d} W_{L}(t)=\sum_{i=1}^{d}\left\langle W_{L}(t), p_{l}\right\rangle_{H} p_{l}$ is a $d$-dimensional Wiener process with correlation operator $\mathcal{P}_{d} Q_{L} \mathcal{P}_{d}$. The integral with respect to $\widetilde{M}_{L}$ can easily be rewritten as an integral over $U_{d}$, since the integrand depends only on the projection $\mathcal{P}_{d} \zeta \in U_{d}$.

We apply the finite-dimensional version of Itô's formula (cf., e.g., [11, Thm. 8.18]) to $V_{d}\left(t, \mathcal{P}_{d} Z_{L}(t)\right)$. In contrast to the Hilbert space-valued case, bounded derivatives are not needed here. The properties of $V_{d}$ shown in Theorem 3.6 are sufficient. By the same arguments as in the proof of Theorem 2.7, we obtain the following:

$$
\begin{aligned}
& d V_{d}\left(t, \mathcal{P}_{d} Z_{L}(t)\right) \\
& =\partial_{t} V_{d} d t+\frac{1}{2} \operatorname{tr}\left(\left[D_{x}^{2} V_{d}\right] \mathcal{P}_{d} Q_{L} \mathcal{P}_{d}\right) d t \\
& \quad+\int_{H}\left\{V_{d}\left(t, \mathcal{P}_{d} Z_{L}(t-)+\mathcal{P}_{d} \zeta\right)-V_{d}\left(t, \mathcal{P}_{d} Z_{L}(t-)\right)-\left[D_{x} V_{d}\right]\left(\mathcal{P}_{d} \zeta\right)\right\} \nu_{L}(d \zeta) d t \\
& \quad+D_{x} V_{d} d\left(\mathcal{P}_{d} W_{L}(t)\right)+\int_{H}\left[V_{d}\left(t, \mathcal{P}_{d} Z_{L}(t-)+\mathcal{P}_{d} \zeta\right)-V_{d}\left(t, \mathcal{P}_{d} Z_{L}(t-)\right)\right] \widetilde{M}_{L}(d \zeta, d t) .
\end{aligned}
$$

Proceeding exactly as in the proof of Theorem 2.8, we obtain (3.3).
The PIDE in Theorem 3.7 is of course nothing more than a projected version of the $\operatorname{PIDE}(2.12)$ for $V$. The derivatives $D_{x} V_{d}(t, x) \in \mathrm{L}\left(\mathbb{R}^{d}, \mathbb{R}\right)$ and $D_{x}^{2} V_{d}(t, x) \in \mathrm{L}\left(\mathbb{R}^{d}, \mathbb{R}^{d}\right)$ can be interpreted as a vector and a matrix, respectively. In particular, we have

$$
\begin{aligned}
\sum_{l=1}^{d}\left[D_{x}^{2} V_{d}(t, x)\right]\left(\mathcal{P}_{d} Q_{L} p_{l}, \mathcal{P}_{d} p_{l}\right) & =\sum_{l=1}^{d} \sum_{i, j=1}^{d} \partial_{x_{i}} \partial_{x_{j}} V_{d}(t, x)\left\langle\mathcal{P}_{d} Q_{L} p_{l}, p_{j}\right\rangle_{H}\left\langle\mathcal{P}_{d} p_{l}, p_{i}\right\rangle_{H} \\
& =\sum_{i, j=1}^{d}\left\langle Q_{L} p_{i}, p_{j}\right\rangle_{H} \partial_{x_{i}} \partial_{x_{j}} V_{d}(t, x)
\end{aligned}
$$

To simplify notation, we define coefficients

$$
a_{i j}:=\frac{1}{2}\left\langle Q_{L} p_{i}, p_{j}\right\rangle_{H} .
$$

The PIDE can thus be written as

$$
\begin{align*}
-\partial_{t} V_{d}(t, x)= & \sum_{i, j=1}^{d} a_{i j} \partial_{x_{i}} \partial_{x_{j}} V_{d}(t, x)  \tag{3.4}\\
& +\int_{H}\left\{V_{d}\left(t, x+\mathcal{P}_{d} \zeta\right)-V_{d}(t, x)-\sum_{i=1}^{d}\left\langle\zeta, p_{i}\right\rangle_{H} \partial_{x_{i}} V_{d}(t, x)\right\} \nu_{L}(d \zeta) .
\end{align*}
$$

Moreover, we have the following ellipticity property.
Theorem 3.8. The matrix $\left(a_{i j}\right)_{i, j=1}^{d}$ is symmetric positive definite.
Proof. This is a direct consequence of Assumption 2.5, since

$$
\sum_{i, j=1}^{d} y_{i} a_{i j} y_{j}=\frac{1}{2}\left\langle Q_{L} \sum_{i=1}^{d} y_{i} p_{i}, \sum_{i=1}^{d} y_{i} p_{i}\right\rangle_{H}
$$

for every $y \in \mathbb{R}^{d}$, and $\left\|\sum_{i=1}^{d} y_{i} p_{i}\right\|_{H}=\|y\|$ due to the isometry (3.1).
3.3. Variational formulation and uniqueness. We have already shown that the approximation $V_{d}(t, x)$ is a classical solution of the finite-dimensional PIDE (3.4). In this section, we introduce the corresponding variational formulation in appropriate Hilbert spaces and show uniqueness of the weak solution. Since the payoff is not necessarily bounded, and thus not an element of $\mathrm{L}^{2}\left(\mathbb{R}^{d}\right)$, we use weighted Sobolev spaces instead. Let $\rho_{\theta}$ be the weight function with exponential decay defined by

$$
\rho_{\theta}: \begin{cases}\mathbb{R}^{d} & \rightarrow \mathbb{R}, \\ x & \mapsto e^{-\theta \sqrt{1+\|x\|^{2}}}\end{cases}
$$

with a parameter $\theta>0$. We define the scalar products

$$
\langle\psi, \varphi\rangle_{\mathrm{L}^{2}, \theta}:=\int_{\mathbb{R}^{d}} \psi(x) \varphi(x) \rho_{\theta}(x) d x
$$

and

$$
\langle\psi, \varphi\rangle_{\mathrm{H}^{k, \theta}}:=\sum_{\alpha \in \mathbb{N}_{0}^{d},|\alpha| \leq k}\left\langle\partial^{\alpha} \psi, \partial^{\alpha} \varphi\right\rangle_{\mathrm{L}^{2}, \theta}
$$

for functions $\psi, \varphi: \mathbb{R}^{d} \rightarrow \mathbb{R}$. The corresponding Hilbert spaces are denoted by $\mathrm{L}^{2, \theta}\left(\mathbb{R}^{d}\right)$ and $\mathrm{H}^{k, \theta}\left(\mathbb{R}^{d}\right)$ (cf., e.g., [2, Chap. 3.1]). In particular, we consider the Gelfand triplet

$$
\mathrm{H}^{1, \theta}\left(\mathbb{R}^{d}\right) \hookrightarrow \mathrm{L}^{2, \theta}\left(\mathbb{R}^{d}\right) \hookrightarrow\left(\mathrm{H}^{1, \theta}\left(\mathbb{R}^{d}\right)\right)^{\prime}
$$

Finally, we define the bilinear form

$$
\begin{align*}
a(\psi, \varphi):= & \int_{\mathbb{R}^{d}} \sum_{i, j=1}^{d} a_{i j} \partial_{x_{i}} \psi(x) \partial_{x_{j}} \varphi(x) \rho_{\theta}(x) d x+\int_{\mathbb{R}^{d}} \sum_{i=1}^{d} b_{i}(x) \partial_{x_{i}} \psi(x) \varphi(x) \rho_{\theta}(x) d x  \tag{3.5}\\
& -\int_{H} \int_{\mathbb{R}^{d}}\left[\psi\left(x+\mathcal{P}_{d} \zeta\right)-\psi(x)-\sum_{i=1}^{d}\left\langle\zeta, p_{i}\right\rangle_{H} \partial_{x_{i}} \psi(x)\right] \varphi(x) \rho_{\theta}(x) d x \nu_{L}(d \zeta)
\end{align*}
$$

for $\psi, \varphi \in \mathrm{H}^{1, \theta}\left(\mathbb{R}^{d}\right)$, where

$$
b_{i}(x):=-\frac{\theta \sum_{j=1}^{d} a_{i j} x_{j}}{\sqrt{1+\|x\|^{2}}} \text { for } i=1, \ldots, d
$$

The coefficients $b_{i}(x)$ satisfy

$$
\left|b_{i}(x)\right| \leq \theta d \max _{j=1, \ldots, d}\left|a_{i j}\right|
$$

and are therefore bounded on $\mathbb{R}^{d}$.
We can now state a variational form of (3.4).
Theorem 3.9. The function $V_{d}$ defined in (3.2) satisfies

$$
\begin{equation*}
-\left\langle\partial_{t} V_{d}(t, \cdot), \varphi\right\rangle_{\mathrm{L}^{2, \theta}}+a\left(V_{d}(t, \cdot), \varphi\right)=0 \tag{3.6}
\end{equation*}
$$

for every $\varphi \in \mathrm{H}^{1, \theta}\left(\mathbb{R}^{d}\right)$ and a.e. $t \in(0, T)$, with terminal condition

$$
\begin{equation*}
V_{d}(T, x)=G(x) \tag{3.7}
\end{equation*}
$$

Proof. We first note that $V_{d}(t, \cdot) \in \mathrm{H}^{2, \theta}\left(\mathbb{R}^{d}\right)$ and $\partial_{t} V_{d}(t, \cdot) \in \mathrm{L}^{2, \theta}\left(\mathbb{R}^{d}\right)$ hold for every $t \in(0, T)$ due to Theorem 3.6. Starting with (3.4), partial integration yields

$$
\begin{aligned}
- & \left\langle\partial_{t} V_{d}(t, x), \varphi\right\rangle_{\mathrm{L}^{2}, \theta} \\
= & -\sum_{i, j=1}^{d} a_{i j} \int_{\mathbb{R}^{d}} \partial_{x_{i}} V_{d}(t, x) \partial_{x_{j}}\left(\varphi \rho_{\theta}\right)(x) d x \\
& +\int_{H} \int_{\mathbb{R}^{d}}\left[V_{d}\left(t, x+\mathcal{P}_{d} \zeta\right)-V_{d}(t, x)-\sum_{i=1}^{d}\left\langle\zeta, p_{i}\right\rangle_{H} \partial_{x_{i}} V_{d}(t, x)\right] \varphi(x) \rho_{\theta}(x) d x \nu_{L}(d \zeta) .
\end{aligned}
$$

Using the product rule, we obtain

$$
\sum_{j=1}^{d} a_{i j} \partial_{x_{j}}\left(\varphi \rho_{\theta}\right)(x)=\sum_{j=1}^{d} a_{i j} \partial_{x_{j}} \varphi(x) \rho_{\theta}(x)+\varphi(x) \frac{-\theta \sum_{j=1}^{d} a_{i j} x_{j}}{\sqrt{1+\|x\|^{2}}} \rho_{\theta}(x)
$$

Theorem 3.10. The bilinear form a defined in (3.5) is continuous and satisfies Gårding's inequality. More precisely, there are constants $C>0, c_{1} \geq 0$, and $c_{2}>0$ (possibly depending on d) such that

$$
\begin{equation*}
|a(\psi, \varphi)| \leq C\|\psi\|_{\mathrm{H}^{1}, \theta}\|\varphi\|_{\mathrm{H}^{1, \theta}} \tag{3.8}
\end{equation*}
$$

and

$$
\begin{equation*}
a(\psi, \psi)+c_{1}\|\psi\|_{\mathrm{L}^{2, \theta}}^{2} \geq c_{2}\|\psi\|_{\mathrm{H}^{1, \theta}}^{2} \tag{3.9}
\end{equation*}
$$

hold for every $\psi, \varphi \in \mathrm{H}^{1, \theta}\left(\mathbb{R}^{d}\right)$.
Proof. First, we show continuity. From the definition of $a$, we obtain

$$
\begin{aligned}
&|a(\psi, \varphi)| \\
& \leq \sum_{i, j=1}^{d}\left|a_{i j}\right| \int_{\mathbb{R}^{d}}\left|\partial_{x_{i}} \psi(x)\right|\left|\partial_{x_{j}} \varphi(x)\right| \rho_{\theta}(x) d x+\sum_{i=1}^{d} \int_{\mathbb{R}^{d}}\left|b_{i}(x)\right|\left|\partial_{x_{i}} \psi(x)\right||\varphi(x)| \rho_{\theta}(x) d x \\
&+\int_{H} \int_{\mathbb{R}^{d}}\left|\psi\left(x+\mathcal{P}_{d} \zeta\right)\right||\varphi(x)| \rho_{\theta}(x) d x \nu_{L}(d \zeta)+\int_{H} \int_{\mathbb{R}^{d}}|\psi(x)||\varphi(x)| \rho_{\theta}(x) d x \nu_{L}(d \zeta) \\
&+\sum_{i=1}^{d} \int_{H} \int_{\mathbb{R}^{d}}\left|\left\langle\zeta, p_{i}\right\rangle_{H}\right|\left|\partial_{x_{i}} \psi(x) \| \varphi(x)\right| \rho_{\theta}(x) d x \nu_{L}(d \zeta) .
\end{aligned}
$$

The Cauchy-Schwarz inequality yields

$$
\begin{aligned}
|a(\psi, \varphi)| \leq & \max _{i, j=1, \ldots, d}\left|a_{i j}\right| \sum_{i, j=1}^{d}\left\|\partial_{x_{i}} \psi\right\|_{\mathrm{L}^{2}, \theta}\left\|\partial_{x_{j}} \varphi\right\|_{\mathrm{L}^{2}, \theta} \\
& +\max _{i=1, \ldots, d}\left\|b_{i}\right\|_{\mathrm{L}^{\infty}} \sum_{i=1}^{d}\left\|\partial_{x_{i}} \psi\right\|_{\mathrm{L}^{2}, \theta}\|\varphi\|_{\mathrm{L}^{2}, \theta}+2 \int_{H}\|\psi\|_{\mathrm{L}^{2}, \theta}\|\varphi\|_{\mathrm{L}^{2}, \theta} \nu_{L}(d \zeta) \\
& +\int_{H} \sum_{i=1}^{d}\left|\left\langle\zeta, p_{i}\right\rangle_{H}\right|\left\|\partial_{x_{i}} \psi\right\|_{\mathrm{L}^{2}, \theta}\|\varphi\|_{\mathrm{L}^{2}, \theta} \nu_{L}(d \zeta) .
\end{aligned}
$$

Due to

$$
\int_{H} \nu_{L}(d \zeta)<\infty \quad \text { and } \quad \int_{H}\|\zeta\|_{H} \nu_{L}(d \zeta)<\infty
$$

this proves (3.8).
For the proof of Gårding's inequality, we start with the ellipticity property from Theorem 3.8. For every $\zeta \in \mathbb{R}^{d}$,

$$
\sum_{i, j=1}^{d} a_{i j} \zeta_{i} \zeta_{j} \geq c \sum_{i=1}^{d} \zeta_{i}^{2}
$$

holds, with a constant $c>0$. Hence

$$
\begin{aligned}
& c \int_{\mathbb{R}^{d}} \sum_{i=1}^{d}\left|\partial_{x_{i}} \psi(x)\right|^{2} \rho_{\theta}(x) d x \leq \int_{\mathbb{R}^{d}} \sum_{i, j=1}^{d} a_{i j} \partial_{x_{i}} \psi(x) \partial_{x_{j}} \psi(x) \rho_{\theta}(x) d x \\
&= a(\psi, \psi)-\int_{\mathbb{R}^{d}} \sum_{i=1}^{d} b_{i}(x) \partial_{x_{i}} \psi(x) \psi(x) \rho_{\theta}(x) d x \\
& \quad+\int_{H} \int_{\mathbb{R}^{d}}\left[\psi\left(x+\mathcal{P}_{d} \zeta\right)-\psi(x)-\sum_{i=1}^{d}\left\langle\zeta, p_{i}\right\rangle_{H} \partial_{x_{i}} \psi(x)\right] \psi(x) \rho_{\theta}(x) d x \nu_{L}(d \zeta) .
\end{aligned}
$$

The same calculations as in the proof of continuity above yield

$$
\begin{aligned}
c \sum_{i=1}^{d}\left\|\partial_{x_{i}} \psi\right\|_{\mathrm{L}^{2, \theta}}^{2} & \leq a(\psi, \psi)+C_{1} \sum_{i=1}^{d}\left\|\partial_{x_{i}} \psi\right\|_{\mathrm{L}^{2, \theta}}\|\psi\|_{\mathrm{L}^{2, \theta}}+C_{2}\|\psi\|_{\mathrm{L}^{2, \theta}}^{2} \\
& \leq a(\psi, \psi)+C_{1}\left(\frac{\varepsilon}{2} \sum_{i=1}^{d}\left\|\partial_{x_{i}} \psi\right\|_{\mathrm{L}^{2, \theta}}^{2}+\frac{d}{2 \varepsilon}\|\psi\|_{\mathrm{L}^{2, \theta}}^{2}\right)+C_{2}\|\psi\|_{\mathrm{L}^{2, \theta}}^{2},
\end{aligned}
$$

where we have used Young's inequality in the last estimate. Choosing $\varepsilon$ so small that

$$
C_{1} \frac{\varepsilon}{2} \leq \frac{1}{2} c
$$

and setting

$$
c_{1}=\frac{1}{2} c+\frac{C_{1} d}{2 \varepsilon}+C_{2} \quad \text { and } \quad c_{2}=\frac{1}{2} c
$$

yields (3.9).

With the following two theorems, we show that $V_{d}$ is indeed the unique solution of the PIDE. We start with a lemma requiring stronger regularity hypotheses for $G$. Afterwards, we give a result for arbitrary Lipschitz continuous payoffs, including, in particular, call and put options.

Lemma 3.11. Suppose that $G \in \mathrm{H}^{2, \theta}\left(\mathbb{R}^{d}\right)$ has bounded first and second derivatives. Then $V_{d}$ is the unique solution of (3.6), with terminal condition (3.7), in the space $\mathcal{W}(0, T)$ defined by

$$
\mathcal{W}(0, T):=\left\{f: \mathbb{R}^{d} \rightarrow \mathbb{R}: f \in \mathrm{~L}^{2}\left(0, T ; \mathrm{H}^{1, \theta}\right), \partial_{t} f \in \mathrm{~L}^{2}\left(0, T ;\left(\mathrm{H}^{1, \theta}\left(\mathbb{R}^{d}\right)\right)^{\prime}\right)\right\} .
$$

Proof. The bilinear form $a$ is continuous and satisfies Gårding's inequality by Theorem 3.10. Therefore, the PIDE has a unique solution in the space $\mathcal{W}(0, T)$ by [37, Thm. 26.1]. On the other hand, we know from Theorem 3.7 that $V_{d}$ satisfies (3.6). It remains to prove that $V_{d} \in \mathcal{W}(0, T)$.

Using the same notation as in the proof of Theorem 3.6, we have

$$
V_{d}(t, x)=e^{-r T} \int_{\mathbb{R}^{d}} G(x+y) g_{T-t}(y) d y
$$

where $g_{T-t} \in \mathcal{S}\left(\mathbb{R}^{d}\right)$ is the density of $\mathcal{P}_{d} Z_{L}(T-t)$. Consequently,

$$
\int_{\mathbb{R}^{d}} V_{d}^{2}(t, x) \rho_{\theta}(x) d x \leq \int_{\mathbb{R}^{d}}\left(\int_{\mathbb{R}^{d}}\left(C_{1}+C_{2}\|x\|+C_{3}\|y\|\right) g_{T-t}(y) d y\right)^{2} \rho_{\theta}(x) d x
$$

Since

$$
\int_{\mathbb{R}^{d}}\|x\|^{i} \rho_{\theta}(x) d x<\infty(i \in\{1,2\}), \quad \int_{\mathbb{R}^{d}} g_{T-t}(y) d y=1
$$

and

$$
\int_{\mathbb{R}^{d}}\|y\| g_{T-t}(y) d y=E\left[\left\|\mathcal{P}_{d} Z_{L}(T-t)\right\|\right]<\infty
$$

hold for every $t \in(0, T)$, this implies $V_{d} \in \mathrm{~L}^{2}\left(0, T ; \mathrm{L}^{2, \theta}\left(\mathbb{R}^{d}\right)\right)$. Moreover, we have

$$
\partial_{x}^{\alpha} V_{d}(t, x)=e^{-r T} \int_{\mathbb{R}^{d}} \partial^{\alpha} G(x+y) g_{T-t}(y) d y
$$

for every $\alpha \in \mathbb{N}_{0}^{d},|\alpha| \in\{1,2\}$. Due to the boundedness of the derivatives of $G$, the following holds:

$$
\int_{\mathbb{R}^{d}}\left(\partial_{x}^{\alpha} V_{d}(t, x)\right)^{2} \rho_{\theta}(x) d x \leq C \int_{\mathbb{R}^{d}}\left(\int_{\mathbb{R}^{d}} g_{T-t}(y) d y\right)^{2} \rho_{\theta}(x) d x=C \int_{\mathbb{R}^{d}} \rho_{\theta}(x) d x
$$

Consequently, $V_{d} \in L^{2}\left(0, T ; \mathrm{H}^{2, \theta}\left(\mathbb{R}^{d}\right)\right)$. Since $V_{d}$ satisfies the PIDE (3.4), its time derivative $\partial_{t} V_{d}$ can be expressed in terms of its first and second spatial derivatives. Thus, we also have $\partial_{t} V_{d} \in L^{2}\left(0, T ; \mathrm{L}^{2, \theta}\left(\mathbb{R}^{d}\right)\right)$, and the proof is finished.

Theorem 3.12. Let $G$ satisfy Assumption 3.4. Then $V_{d}$ is an element of the space $\mathcal{W}(0, T)$ defined in Lemma 3.11 and the unique solution of (3.6) with terminal condition (3.7).

Proof. The key of the proof is to approximate $G$ with a sequence of smooth functions to which Lemma 3.11 can be applied. To this end, let $\left(\psi_{n}\right)_{n \in \mathbb{N}} \subset \mathrm{C}_{0}^{\infty}\left(\mathbb{R}^{d}\right)$ be a sequence of standard mollifiers with compact support. Define the convolution

$$
G_{n}(x):=\int_{\mathbb{R}^{d}} G(x-y) \psi_{n}(y) d y \in \mathrm{C}^{\infty}\left(\mathbb{R}^{d}\right), \quad x \in \mathbb{R}^{d}
$$

Since $G$ is by assumption Lipschitz, and thus uniformly continuous, the following uniform approximation property holds by [16, Thm. C.6]:

$$
\forall \varepsilon>0 \exists N \in \mathbb{N}:\left|G_{n}(x)-G(x)\right| \leq \varepsilon \quad \text { for every } n \geq N, x \in \mathbb{R}^{d} .
$$

Moreover, for every $\xi \in \mathbb{R}^{d}$ and every multiindex $\alpha \in \mathbb{N}_{0}^{d}$, we have

$$
\begin{aligned}
\left|\partial^{\alpha} G_{n}(x+\xi)-\partial^{\alpha} G_{n}(x)\right| & \leq \int_{\mathbb{R}^{d}}|G(x+\xi-y)-G(x-y)|\left|\partial^{\alpha} \psi_{n}(y)\right| d y \\
& \leq K_{G}\|\xi\| \int_{\mathbb{R}^{d}}\left|\partial^{\alpha} \psi_{n}(y)\right| d y
\end{aligned}
$$

Thus, in particular, the first and second derivatives of $G_{n}$ are bounded for every $n \in \mathbb{N}$.
Now let

$$
V_{d}^{n}(t, x):=e^{-r T} E\left[G_{n}\left(x+\mathcal{P}_{d} Z_{L}(T-t)\right)\right]
$$

be the price function associated with payoff $G_{n}$. By Lemma 3.11, $V_{d} \in \mathcal{W}(0, T)$ is the unique solution of (3.4) with terminal condition $V_{d}^{n}(T, x)=G_{n}(x)$. Moreover, the PIDE with terminal value $G(x)$ also has a unique solution, which we denote by $\widetilde{V}_{d} \in \mathcal{W}(0, T)$. From [37, Thm. 26.1], we obtain

$$
\begin{equation*}
\left\|V_{d}^{n}-\widetilde{V}_{d}\right\|_{\mathrm{L}^{2}\left(0, T ; \mathrm{L}^{2, \theta}\right)} \leq C\left\|G_{n}-G\right\|_{\mathrm{L}^{2, \theta}} \rightarrow 0 \quad \text { for } n \rightarrow \infty \tag{3.10}
\end{equation*}
$$

On the other hand, by the proof of Lemma 3.11, we have $V_{d} \in \mathrm{~L}^{2}\left(0, T ; \mathrm{L}^{2, \theta}\right)$ for every payoff $G$ satisfying Assumption 3.4. Thus, using the notation from the proof of Theorem 3.6, we get

$$
\begin{align*}
\| V_{d}^{n} & \left.-V_{d} \|_{\mathrm{L}^{2}\left(0, T ; \mathrm{L}^{2}, \theta\right.}^{2}\right) \\
& =e^{-r T} \int_{0}^{T} \int_{\mathbb{R}^{d}}\left(\int_{\mathbb{R}^{d}}\left(G(x+y)-G_{n}(x+y)\right) g_{T-t}(y) d y\right)^{2} \rho_{\theta}(x) d x d t  \tag{3.11}\\
& \leq e^{-r T} \int_{0}^{T} \int_{\mathbb{R}^{d}}\left(\int_{\mathbb{R}^{d}}\left|G(x+y)-G_{n}(x+y)\right| g_{T-t}(y) d y\right)^{2} \rho_{\theta}(x) d x d t \\
& \rightarrow 0 \text { for } n \rightarrow \infty .
\end{align*}
$$

Combining (3.10) and (3.11), we obtain $V_{d}=\widetilde{V}_{d}$.
3.4. Convergence of finite-dimensional approximation. We are interested in the convergence of the finite-dimensional approximation $V_{d}$ to the true price function $V$. The fair price of the option is given by $V(0,0)$, since we assume the Lévy process $Z_{L}$ starts in 0 . In particular, we are looking for a pointwise convergence result.

Theorem 3.13. Let $\mu_{1} \geq \mu_{2} \geq \cdots \geq 0$ be the eigenvalues of the covariance operator $\mathcal{C}_{X}$ defined in (2.5). Then there exists a constant $C>0$ such that

$$
\left|V_{d}(0,0)-V(0,0)\right| \leq C \sqrt{\sum_{l=d+1}^{\operatorname{dim} H} \mu_{l}} .
$$

Proof. By definition of $V$ and $V_{d}$, we have

$$
\begin{aligned}
\left|V_{d}(0,0)-V(0,0)\right| & =e^{-r T}\left|E\left[G\left(Z_{L}(T)\right)-G\left(\mathcal{P}_{d} Z_{L}(T)\right)\right]\right| \\
& \leq e^{-r T} E\left[K_{G}\left\|Z_{L}(T)-\mathcal{P}_{d} Z_{L}(T)\right\|_{H}\right]
\end{aligned}
$$

Since $\|\cdot\|_{\mathrm{L}^{1}} \leq C\|\cdot\|_{\mathrm{L}^{2}}$ for finite measure spaces, we may apply Theorem 3.3 to obtain

$$
\left|V_{d}(0,0)-V(0,0)\right| \leq C \sqrt{E\left[\left\|Z_{L}(T)-\mathcal{P}_{d} Z_{L}(T)\right\|_{H}^{2}\right]}=C \sqrt{\sum_{l=d+1}^{\operatorname{dim} H} \mu_{l}}
$$

Depending on the properties of the covariance operator $\mathcal{C}_{X}$, bounds for the decay of the eigenvalues $\mu_{l}$ can be found. To this end, we define the kernel

$$
K:\left\{\begin{array}{l}
D \times D \rightarrow \mathbb{R}  \tag{3.12}\\
(u, v) \mapsto E\left[Z_{L}(T, u) Z_{L}(T, v)\right]
\end{array}\right.
$$

where as before $D \subset \mathbb{R}^{n}$ and $H=\mathrm{L}^{2}\left(D, \mu_{D}\right)$. Then $K$ is indeed the kernel of the covariance operator $\mathcal{C}_{X}$, since by Fubini's theorem

$$
\begin{aligned}
\int_{D} \int_{D} K(u, v) h_{1}(u) \mu_{D}(d u) h_{2}(v) \mu_{D}(d v) & =E\left[\left\langle Z_{L}(T), h_{1}\right\rangle_{H}\left\langle Z_{L}(T), h_{2}\right\rangle_{H}\right] \\
& =\left\langle\mathcal{C}_{X} h_{1}, h_{2}\right\rangle_{H}
\end{aligned}
$$

for every $h_{1}, h_{2} \in H$. Consequently,

$$
\mathcal{C}_{X} h_{1}(\cdot)=\int_{D} K(\cdot, v) h_{1}(v) \mu_{D}(d v)
$$

Moreover, $K$ is an element of $\mathrm{L}^{2}(D \times D)$, since

$$
\begin{aligned}
\int_{D} \int_{D} K^{2}(u, v) \mu_{D}(d u) \mu_{D}(d u) & \leq \int_{D} \int_{D} E\left[Z_{L}(T, u)^{2}\right] E\left[Z_{L}(T, v)^{2}\right] \mu_{D}(d u) \mu_{D}(d u) \\
& =E\left[\left\|X_{L}(T)-\Gamma T\right\|_{H}^{2}\right]^{2}<\infty
\end{aligned}
$$

by Theorem 2.2. We now give a result for the eigenvalue decay, depending on the properties of $K$.

Theorem 3.14. Let $D \subset \mathbb{R}^{m}$ be a bounded Borel set and $\mu_{D}$ be the Lebesgue measure. If the kernel $K$ is piecewise $\mathrm{H}^{k} \otimes \mathrm{~L}^{2}$ on $D \times D$ for a $k \in \mathbb{N}$, then there exists a constant $C$ such that

$$
\mu_{l} \leq C l^{-\frac{k}{d}} \quad \text { for every } l \in I .
$$

Moreover, if the kernel $K$ is piecewise analytic, then there are constants $C_{1}, C_{2}$ such that

$$
\mu_{l} \leq C_{1} e^{-C_{2} l^{\frac{1}{d}}} \quad \text { for every } l \in I .
$$

Proof. These results are shown in [32, Props. 2.18 and 2.21].
Remark 3.15. A precise definition of "piecewise $\mathrm{H}^{k} \otimes \mathrm{~L}^{2}$ " as used in the hypothesis of Theorem 3.14 is given in [34, Def. 3.1]. The hypothesis of Theorem 3.14 is fulfilled for the jump-diffusion model if the drift $\gamma$, the volatility $\sigma$, the Brownian covariance operator $Q$, and the jump-dampening factors $\eta$ satisfy corresponding piecewise smoothness criteria.
4. Numerical methods. In this section, methods for the numerical solution of the European option pricing problem are described. In particular, the computation of a suitable basis for the dimension reduction described in the previous section is addressed. Moreover, we give a short overview of the techniques needed to solve the resulting finite-dimensional PIDE.
4.1. Computation of the POD basis. As we have seen in section 3.1, the construction of a POD basis for $X(T)$ can be reduced to the eigenvalue problem

$$
\mathcal{C}_{X} p_{l}=\mu_{l} p_{l}, \quad l \in I
$$

with the covariance operator $\mathcal{C}_{X}: H \rightarrow H^{\prime} \cong H=L^{2}\left(D, \mu_{D}\right)$ defined in (2.5). In general, the eigenvectors $p_{l}$ are not known analytically. However, it is possible to compute good approximations numerically. For finite index sets $D$ (basket options), the solution of the symmetric eigenvalue problem can be performed by standard methods (e.g., QR-algorithm). For all subsequent results, we will therefore assume the more complicated setting that $D \subset \mathbb{R}^{m}$ is a bounded Borel set and $\mu_{D}$ is the Lebesgue measure. This holds, e.g., for electricity swaptions (with $m=1$ ). In this case, some further approximation is needed to obtain a finitedimensional problem eventually. To this end, we employ a finite element discretization and solve a projected eigenvalue problem. Convergence results for this technique can be shown under certain regularity conditions on the covariance kernel $K$ defined in (3.12). The general theory of Galerkin approximations of Karhunen-Loève expansions is discussed in [32].

Let $U_{\Delta x} \subset H=\mathrm{L}^{2}\left(D, \mu_{D}\right)$ be a finite element subspace with discretization parameter $\Delta x$. For $l=1, \ldots, \operatorname{dim} U_{\Delta x}$, the Galerkin approximations $\left(\mu_{l}^{\Delta x}, p_{l}^{\Delta x}\right) \subset \mathbb{R} \times U_{h}$ of the eigenpairs $\left(\mu_{l}, p_{l}\right) \subset \mathbb{R} \times H$ are, by definition, solutions of the following problem:

$$
\begin{aligned}
\forall \varphi \in U_{\Delta x}:\left\langle\mathcal{C}_{X} p_{l}^{\Delta x}, \varphi\right\rangle_{H} & =\int_{D}\left(\int_{D} K(u, v) p_{l}^{\Delta x}(v) \mu_{D}(d v)\right) \varphi(u) \mu_{D}(d u) \\
& \stackrel{!}{=} \mu_{l}^{\Delta x}\left\langle p_{l}^{\Delta x}, \varphi\right\rangle_{H} .
\end{aligned}
$$

This is equivalent to the eigenvalue problem

$$
\begin{equation*}
\mathcal{P}_{\Delta x} \mathcal{C}_{X} \mathcal{P}_{\Delta x} p_{l}^{\Delta x}=\mu_{l}^{\Delta x} p_{l}^{\Delta x} \tag{4.1}
\end{equation*}
$$

where $\mathcal{P}_{\Delta x}: H \rightarrow U_{\Delta x}$ is the projection operator onto the finite element subspace. The operator $\mathcal{P}_{\Delta x} \mathcal{C}_{X} \mathcal{P}_{\Delta x}$ is self-adjoint and compact due to the properties of the projection and Theorem 2.4. The following theorem gives an error bound for the approximation of $Z_{L}(T)$ obtained with this Galerkin method.

Theorem 4.1. Let the covariance kernel $K$ defined in (3.12) be a piecewise smooth function. Further, let $U_{\Delta x} \subset H$ be a finite element space of piecewise polynomials of degree $q \in \mathbb{N}$, where $\Delta x$ denotes the mesh width of the regular triangulation. Finally, let $\mu_{l}$ be the true eigenvalues of the covariance operator $\mathcal{C}_{X}$, and let $p_{i}^{\Delta x}$ be orthonormal solutions of the projected eigenvalue problem (4.1). Then there exists a constant $C$ such that

$$
E\left[\left\|Z_{L}(T)-\sum_{l=1}^{d}\left\langle Z_{L}(T), p_{l}^{\Delta x}\right\rangle_{H} p_{l}^{\Delta x}\right\|_{H}^{2}\right] \leq C \Delta x^{2 q+1}+\sum_{l=d+1}^{\operatorname{dim} H} \mu_{l}
$$

holds for every $d \leq \operatorname{dim} U_{\Delta x}$.
Proof. The estimate is taken from [32, Prop. 3.3]. The necessary assumption [32, Ass. 3.1] is satisfied due to the piecewise smoothness of the kernel and [34, Thm. 1.5].

Note that all the results from sections 3.2 and 3.3 are still valid when we use $\left(p_{l}^{\Delta x}\right)_{l=1}^{d}$ instead of $\left(p_{l}\right)_{l=1}^{d}$ if we replace Assumption 2.5 with the following hypothesis:

$$
\begin{equation*}
\left\langle Q_{L} h, h\right\rangle_{H}>0 \quad \text { for every } h \in \operatorname{span}\left\{p_{l}^{\Delta x}, l=1, \ldots, d\right\} \backslash\{0\} . \tag{4.2}
\end{equation*}
$$

In this case, we denote the unique solution of the corresponding finite-dimensional PIDE (projected to $\operatorname{span}\left\{p_{l}^{\Delta x}, l=1, \ldots, d\right\}$ ) by $V_{d}^{\Delta x}$.

Corollary 4.2. Let the hypotheses of Theorem 4.1 and (4.2) hold. Then there exists a constant $C>0$ such that

$$
\left|V_{d}^{\Delta x}(0,0)-V(0,0)\right| \leq C \sqrt{(\Delta x)^{2 q+1}+\sum_{l=d+1}^{\operatorname{dim} H} \mu_{l}}
$$

Proof. The estimate follows from Theorem 4.1 by exactly the same arguments as in the proof of Theorem 3.13.
4.2. Numerical PIDE solution. In this subsection, we give an overview of the numerical methods for solving the finite-dimensional PIDE (3.4). We will not go into details about convergence results for PIDE solvers but refer to recent literature instead. Finite difference methods for integro-differential equations are analyzed, e.g., in [12, 30], and finite elements and wavelet compression techniques are described in [24, 36, 28].

Localization. We consider the PIDE (3.4) whose spatial domain is the whole of $\mathbb{R}^{d}$. The first step towards a numerical solution is therefore the localization to a bounded domain. To this end, we restrict the spatial part of the equation to a $d$-dimensional cuboid

$$
\Omega_{R}:=\left[-R_{1}, R_{1}\right] \times\left[-R_{2}, R_{2}\right] \times \cdots \times\left[-R_{d}, R_{d}\right] .
$$

This simple domain can be described by a single vector $R=\left(R_{1}, \ldots, R_{d}\right) \in \mathbb{R}^{d}$. The probabilistic interpretation of this truncation is that we price a barrier option whose value is set to 0
if the process $Z_{L}$ leaves $\Omega_{R}$ at any time before maturity. Under polynomial growth conditions for the payoff function $G$, one can show that the truncation error decreases exponentially with $\|R\|$ (cf., e.g., [36, Thm. 3.3.2]). The error is of course higher if the solution is evaluated closer to the boundary of $\Omega_{R}$.

Analysis of the dependence of the localization error on the truncation parameter in onedimensional problems shows that good results are achieved for $R_{1}$ greater than a certain multiple of the standard deviation of the jump-diffusion process at time $T$ (cf., e.g., [11, Fig. 12.1]). By construction of the POD basis $\left(p_{l}\right)_{l=1}^{d}$, we have

$$
\left\langle\mathcal{C}_{X} p_{l}, p_{l}\right\rangle_{H}=\mu_{l} \quad \text { for } l=1, \ldots, d .
$$

Hence, the eigenvalues $\mu_{l}$ describe the variance in direction of the POD vectors. This suggests an adaptive truncation strategy, setting

$$
\begin{equation*}
R_{l}=C \sqrt{\mu_{l}}, \quad l=1, \ldots, d \tag{4.3}
\end{equation*}
$$

with a constant $C>0$. This heuristic choice accounts for the decreasing variance in the coordinate directions (compare Theorem 3.14). It results in smaller domains $\Omega_{R}$ (compared to cubic domains) and allows for more accurate discretization results using the same number of grid points. We introduce artificial homogeneous Dirichlet boundary conditions and set

$$
V_{d}^{h}(t, x)=0 \quad \text { for every } t \in[0, T], x \in \partial \Omega_{R},
$$

where $\partial \Omega_{R}$ is the boundary of the domain.
Discretization. We will solve the PIDE using a vertical method of lines; i.e., we first discretize the spatial operators and apply some time stepping for ordinary differential equations afterwards. Since we have already derived the variational formulation of the PIDE in Theorem 3.9, we can directly apply a finite element method to approximate the spatial derivatives in (3.6). Usually, finite elements have several advantages compared to finite differences. In particular, they allow for an easy discretization of geometrically complex domains, adaptive refinement, and higher-order approximations. Moreover, the theory of weak solutions allows for lower regularity assumptions than the classical differential operator.

However, in the specific setting of option pricing, these arguments are only partly valid. First, we may choose the shape of our computational domain arbitrarily due to localization. As described in the previous paragraph, we truncate the domain to a $d$-dimensional cuboid. Second, we have already proven that $V_{d}$ is a smooth classical solution to (3.4). Moreover, despite the simple shape of the domain, a significant overhead is needed to compute and store the geometric information about finite elements.

Hence, we apply finite differences for the numerical experiments in section 5. We define a regular but anisotropic grid $G_{\alpha}$ on $\Omega_{R}$. The grid is described by a multiindex $\alpha=$ $\left(\alpha_{1}, \ldots, \alpha_{d}\right) \in \mathbb{N}^{d}$, and the mesh size in each coordinate is given by $h_{i}:=2 R_{i} / 2^{\alpha_{i}}, i=1, \ldots, d$. The grid contains the points

$$
x(\beta):=\left(-R_{i}+\beta_{i} h_{i}\right)_{i=1}^{d} \in \mathbb{R}^{d}, \quad \beta \in\left\{0,1, \ldots, 2^{\alpha_{1}}\right\} \times \cdots \times\left\{0,1, \ldots, 2^{\alpha_{d}}\right\} .
$$

The corresponding discretized subspace is denoted by $U_{d}^{h} \subset U_{d} \subset H$. The partial derivatives are approximated by central differences as follows:

$$
\partial_{x_{i}} V_{d}(t, x(\beta)) \approx \frac{1}{2 h_{i}}\left[V_{d}\left(t, x\left(\beta+e_{i}\right)\right)-V_{d}\left(t, x\left(\beta-e_{i}\right)\right)\right],
$$

where $e_{i}$ is the $i$ th canonical unit vector. For the second derivatives, we have

$$
\partial_{x_{j}} \partial_{x_{i}} V_{d}(t, x(\beta)) \approx\left\{\begin{aligned}
& \frac{1}{4 h_{i} h_{j}}\left[V_{d}\left(t, x\left(\beta+e_{i}+e_{j}\right)\right)-V_{d}\left(t, x\left(\beta+e_{i}-e_{j}\right)\right)\right. \\
&\left.-V_{d}\left(t, x\left(\beta-e_{i}+e_{j}\right)\right)+V_{d}\left(t, x\left(\beta-e_{i}-e_{j}\right)\right)\right], i \neq j, \\
& \frac{1}{h_{i}^{2}}\left[V_{d}\left(t, x\left(\beta+e_{i}\right)\right)-2 V_{d}(t, x(\beta))+V_{d}\left(t, x\left(\beta-e_{i}\right)\right)\right], i=j .
\end{aligned}\right.
$$

The discretization of the nonlocal integral term will be addressed separately in the next paragraph.

After applying the finite difference scheme for a sparse grid approximation in space, the resulting system of ordinary differential equations is solved with an appropriate time stepping scheme. Since the differential part of the PIDE is of parabolic type, we choose a discontinuous Galerkin method of order 1 for this purpose. Details on the topic (in particular, error estimates) can be found in [33, Chap. 12]. Defining a partition $0=t_{0}<\cdots<t_{n_{T}}=T$ of $[0, T]$, we calculate a solution in the space

$$
S_{d}^{h}:=\left\{v \in \mathrm{~L}^{2}\left(0, T ; U_{d}^{h}\right) ; v \chi_{\left(t_{m-1}, t_{m}\right]} \in \Pi_{1}\left(t_{m-1}, t_{m} ; U_{d}^{h}\right), m=1, \ldots, n_{T}\right\}
$$

where $\Pi_{1}\left(t_{m-1}, t_{m} ; U_{d}^{h}\right)$ is the space of polynomials of degree at most 1 having values in $U_{d}^{h}$. This method yields a stable algorithm, allowing for large time steps even in the presence of convection terms.

Nonlocal integral terms. One of the main difficulties when solving the PIDE is the nonlocal nature of the integral term. In contrast to differential operators, this term involves the solution on the whole of $\mathbb{R}^{d}$. Since we have already introduced artificial zero boundary conditions, the solution can easily be continued with 0 outside the truncated domain $\Omega_{R}$. However, numerical quadrature formulas will in general lead to full matrices. This is contradictory to the essential use of sparse matrices even for problems on relatively low-dimensional spaces. An efficient way to reduce the number of nonzero matrix entries is by using wavelet compression schemes. These make use of the fast decline of entries corresponding to wavelet basis functions with larger distance of their supports. Entries close to 0 are then discarded (cf. the references given at the beginning of this section).

Wavelets are the method of choice if the jump distributions in the additive model (2.1) cover large parts of the domain. However, the examples we are going to examine in section 5 are of type (1.1), i.e., Hilbert space-valued jump diffusions with scalar driving processes. Here, at every time $t \in[0, T]$, jumps are restricted to one-dimensional subspaces, spanned by $\eta_{k}(t) \in H$. If the number of driving jump processes is low, a different approach is feasible, too. Instead of wavelet compression, direct numerical quadrature (Gauß or Newton-Cotes methods) can be applied. We will now have a closer look at this specific case.

Let $0=t_{0}<t_{1}<\cdots<t_{N_{T}}=T, N_{T} \in \mathbb{N}$, be the time discretization grid. It is not necessary to compute the Lévy measure $\nu_{L}$ defined in (2.4) explicitly. Instead, we define measures

$$
\begin{aligned}
\nu_{L}\left(t_{k}, t_{l} ; B\right) & :=\frac{1}{t_{l}-t_{k}} \int_{t_{k}}^{t_{l}} \int_{H} \chi_{B}(\eta(s, \xi)) \nu(d \xi) d s, \\
\nu_{L}(t ; B) & :=\int_{H} \chi_{B}(\eta(t, \xi)) \nu(d \xi)
\end{aligned}
$$

for any $B \subset \mathcal{B}(H), 0 \leq k<l \leq n_{T}$, and $t \geq 0$. For each time step $\left[t_{k}, t_{k+1}\right]$ in the method of lines, we use integrals with respect to $\nu_{L}\left(t_{k}, t_{k+1} ; \cdot\right)$, i.e., the equivalent Lévy measure for the current time interval. We are looking for an approximation of the integral

$$
\begin{gathered}
\int_{H}\left[V_{d}\left(t, x+\mathcal{P}_{d} \zeta\right)-V_{d}(t, x)-\sum_{i=1}^{d}\left\langle\zeta, p_{i}\right\rangle_{H} \partial_{x_{i}} V_{d}(t, x)\right] \nu_{L}\left(t_{k}, t_{k+1} ; d \zeta\right) \\
=\int_{H} V_{d}\left(t, x+\mathcal{P}_{d} \zeta\right) \nu_{L}\left(t_{k}, t_{k+1} ; d \zeta\right)-\lambda V_{d}(t, x) \\
\quad-\sum_{i=1}^{d} \partial_{x_{i}} V_{d}(t, x) \int_{H}\left\langle\zeta, p_{i}\right\rangle_{H} \nu_{L}\left(t_{k}, t_{k+1} ; d \zeta\right) .
\end{gathered}
$$

To reduce the computational effort for the method of lines, the time steps $\Delta t$ are chosen rather large. In order to nevertheless achieve sufficient accuracy for the approximation of the above integrals, additional substeps of each interval $\left[t_{k}, t_{k+1}\right]$ are introduced. Let $n_{s} \in \mathbb{N}$ be the number of substeps for each major time step. Then we have

$$
\begin{aligned}
\int_{H} V_{d}(t, x & \left.+\mathcal{P}_{d} \zeta\right) \nu_{L}\left(t_{k}, t_{k+1} ; d \zeta\right) \\
& \approx \frac{1}{n_{s}} \sum_{j=1}^{n_{s}} \int_{H} V_{d}\left(t, x+\mathcal{P}_{d} \zeta\right) \nu_{L}\left(t_{k}+j \frac{t_{k+1}-t_{k}}{n_{s}} ; d \zeta\right) \\
& =\frac{\lambda}{n_{s}} \sum_{j=1}^{n_{s}} \int_{H} V_{d}\left(t, x+\mathcal{P}_{d}\left[\eta\left(t_{k}+j \frac{t_{k+1}-t_{k}}{n_{s}}, y\right)\right]\right) P^{Y}(d y)
\end{aligned}
$$

for $0 \leq k<n_{T}$. Similarly, we obtain

$$
\int_{H}\left\langle\zeta, p_{i}\right\rangle_{H} \nu_{L}\left(t_{k}, t_{k+1} ; d \zeta\right) \approx \lambda \frac{1}{n_{s}} \sum_{j=1}^{n_{s}} \int_{H}\left\langle\eta\left(t_{k}+j \frac{t_{k+1}-t_{k}}{n_{s}}, y\right), p_{i}\right\rangle_{H} P^{Y}(d y)
$$

The integrals with respect to $P^{Y}$ can then be approximated with quadrature formulas. This requires interpolation of the function at quadrature points, which are not necessarily identical to grid points. Since we have assumed that jumps occur only along a relatively small subspace of $\Omega_{R}$, only a small number of grid points is involved, and the corresponding matrices remain sparse. The idea is illustrated in Figure 1 for a single jump process, where $P^{Y}$ is defined on $\mathbb{R}$ and the jumps are given by $\eta(t) y \in H$.


Figure 1. Grid points (on full grid) involved in the numerical quadrature of jump term integrals in the time interval $\left[t_{1}, t_{2}\right]$.

Sparse grids. Besides the nonlocality of the integral term, the exponentially increasing computational complexity with increasing dimension $d$ is the major numerical problem when solving the PIDE. This curse of dimension can be broken by using sparse grids. A comprehensive overview of this topic can be found in [8]. Figure 2 shows a sparse grid in two-dimensional space.

In particular, we make use of the so-called combination technique [26]. Thus, we use a standard PIDE solver on a series of full, regular, but anisotropic grids. Instead of using all grids $G_{\alpha}$ with $\|\alpha\|_{\infty} \leq M \in \mathbb{N}$, only grids satisfying

$$
\begin{equation*}
M \leq\|\alpha\|_{1} \leq M+d-1 \tag{4.4}
\end{equation*}
$$

for some $M \in \mathbb{N}$ are employed. Denote the approximation of $V_{d}$ on the grid $G_{\alpha}$ by $\widetilde{V}_{d}^{\alpha}$. An approximation $\widetilde{V}_{d}^{M}$, corresponding to a sparse grid solution, can then be obtained by linear combination as follows:

$$
\begin{equation*}
\widetilde{V}_{d}^{M}(x):=\sum_{k=1}^{d}(-1)^{k+1}\binom{d-1}{k-1} \sum_{\|\alpha\|_{1}=M+d-k} \widetilde{V}_{d}^{\alpha} . \tag{4.5}
\end{equation*}
$$

Since artificial zero boundary conditions are applied, grid points on the boundary $\partial \Omega_{R}$ do not have to be included.


Figure 2. Construction of sparse tensor product (left) and sparse grid without boundary points (right) in $\mathbb{R}^{2}$.

Because of the anisotropic truncation of the domain introduced in (4.3), an equal number of refinements in every coordinate would result in finer mesh widths for coordinates which are in fact the least important ones. This mismatch can be remedied by introducing additional constraints on the multiindex of grids used. In order to achieve similar mesh widths, we demand

$$
\alpha_{i} \leq M+\ln \left(\frac{R_{i}}{R_{1}}\right) / \ln (2), \quad i=1, \ldots, d,
$$

which is equivalent to $2^{\alpha_{i}} \leq 2^{M} \frac{R_{i}}{R_{1}}$. Since this yields a set of grids different from the one described by condition (4.4) alone, the corresponding coefficients in (4.5) have to be modified. For a detailed presentation of how to choose coefficients in anisotropic sparse grids, see [19].
5. Numerical experiments. In this section, we will examine the performance of the presented option pricing method. To this end, both the dimension reduction and the PIDE solver have been implemented in $\mathrm{C}++$. The program was applied to test problems which are described in detail below.

### 5.1. Test problems.

Basket option. The first test problem we consider is a basket option on six assets. We use a jump-diffusion model similar to that of [38]. The corresponding Hilbert space is $H=$ $\left(\mathbb{R}^{6},\langle\cdot, \cdot\rangle_{2}\right)$, where $\langle\cdot, \cdot\rangle_{2}$ denotes the Euclidean scalar product. The six-dimensional price process $S=\left(S_{1}, \ldots, S_{6}\right)$ satisfies the dynamics

$$
\frac{d S_{i}(t)}{S_{i}(t)}=r d t+\sigma_{i} d W_{i}(t)+\eta_{i}^{0} d\left[N_{0}(t)-\lambda_{0} t\right]+\eta_{i}^{1} d\left[N_{i}(t)-\lambda_{i} t\right], \quad i=1, \ldots, 6,
$$

where $r=0.05$ is the constant risk-free interest rate. The scalar-valued Brownian motions $W_{i}$ are supposed to be correlated with correlation matrix

$$
\left(\rho_{i j}\right)_{i, j=1}^{6}=\left(\begin{array}{cccccc}
1 & 0.8 & 0.6 & 0.4 & 0.2 & 0 \\
0.8 & 1 & 0.8 & 0.6 & 0.4 & 0.2 \\
0.6 & 0.8 & 1 & 0.8 & 0.6 & 0.4 \\
0.4 & 0.6 & 0.8 & 1 & 0.8 & 0.6 \\
0.2 & 0.4 & 0.6 & 0.8 & 1 & 0.8 \\
0 & 0.2 & 0.4 & 0.6 & 0.8 & 1
\end{array}\right) .
$$

We set the volatilities to $\sigma_{i}=0.2, i=1, \ldots, 6$. The processes $N_{0}$ and $N_{i}$ are independent Poisson processes with intensities $\lambda_{0}=\lambda_{i}=1$, describing jumps common for all assets and independent jumps for individual assets, respectively. The discounted value of every asset is thus a martingale under the risk neutral measure. The relative jump heights are set to $\eta_{i}^{0}=0.2$ and $\eta_{i}^{1}=0.05$. We can write the value of each asset as the exponential of a Lévy process as follows:

$$
\begin{aligned}
S_{i}(t)=S_{i}(0) \exp \{ & \left(r-\frac{1}{2} \sigma_{i}^{2}-\eta_{i}^{0} \lambda_{0}-\eta_{i}^{1} \lambda_{1}\right) t+\sigma_{i} W(t) \\
& \left.+\ln \left(1+\eta_{i}^{0}(t)\right) N_{0}(t)+\ln \left(1+\eta_{i}^{1}(t)\right) N_{i}(t)\right\}, \quad i=1, \ldots, 6
\end{aligned}
$$

We choose the initial value $S_{i}(0)=\frac{100}{6}, i=1, \ldots, 6$, and strike

$$
K=100 \cdot e^{r T}=E\left[\sum_{i=1}^{6} S_{i}(T)\right]
$$

The discounted price of the basket option with maturity $T=1.0$ at time $t \leq T$ is

$$
V(t)=e^{-r T} E\left[\left(\sum_{i=1}^{6} S_{i}(T)-K\right)^{+} \mid \mathcal{F}_{t}\right]
$$

Electricity swaption. The second problem is an instance of the example described in section 1, an option on an electricity swap. We use the exponential additive model given in (1.1). The corresponding Hilbert space is $H=\mathrm{L}^{2}\left(\left[T_{1}, T_{2}\right], \lambda\right)$, where $D=\left[T_{1}, T_{2}\right]$ is the delivery period of the swap and $\lambda$ is the Lebesgue measure. For the diffusive part of the model, we use two factors, similar to as in [20]. The volatilities are given by

$$
\sigma_{1}(s, u) \equiv 0.15, \quad \sigma_{2}(s, u)=0.3 e^{-1.4(u-s)}
$$

Moreover, we assume additional normally distributed jumps, which yield a Merton model. For the jumps, we use a compound Poisson process with intensity $\lambda_{1}=12$ and jump distribution $Y \sim \mathcal{N}(0.1,0.1)$. The additional factor for dampening the jumps is

$$
\eta_{1}(s, u)=0.5-0.5 \frac{u-T_{1}}{T_{2}-T_{1}}
$$

Since an electricity swap requires no payment before the delivery period starts, it is a martingale under the risk neutral measure. Thus, we need the following drift term (compare [11, Chap. 8.4.1]):

$$
\gamma(s, u)=-\frac{1}{2} \sum_{i=1}^{2} \sigma_{i}^{2}(s, u)-\lambda_{1} \int_{\mathbb{R}}\left(e^{\eta_{1}(s, u) y}-1-\eta_{1}(s, u) y\right) P^{Y}(d y)
$$

The risk-free interest rate is assumed to be constant at $r=0.05$. We consider a monthly swap maturing in one year, i.e., $T=T_{1}=1, T_{2}=T_{1}+30 / 365$. The initial forward curve at time $t=0$ is $S_{0}(u) \equiv 50, u \in\left[T_{1}, T_{2}\right]$, and the strike is $K=50$. It remains to specify the discretization of the delivery period $\left[T_{1}, T_{2}\right]$. As we have a continuous forward curve model, we may use an arbitrary number of discretization points. On energy markets, monthly swaps on electricity are usually based on daily prices. Thus we will use exactly $n=30$ components. We will further assume that there are 8 delivery hours per day. Setting $u_{i}=T_{1}+(i-1) \frac{T_{2}-T_{1}}{30}$ for $i=1, \ldots, 30$, we obtain

$$
V(t)=e^{-r T} \cdot 8 \cdot 30 \cdot E\left[\left(\sum_{i=1}^{30} w\left(u_{i} ; T_{1}, T_{2}\right) S\left(T, u_{i}\right)-K\right)^{+} \mid \mathcal{F}_{t}\right]
$$

for the discounted price of the swaption, without making any discretization error.
5.2. Results. Since no analytical solutions are available, a large number of Monte Carlo (MC) simulations were performed for each test problem to obtain a precise solution. All errors were computed using these MC reference values. In order to make use of the easily parallelized methods (MC simulation as well as the sparse grid combination technique), the experiments were run on a Linux workstation with six Opteron processors at 2.7 GHz .

Results for basket option. We first examine the number of POD components needed to obtain a sufficiently good approximation. The eigenvalues $\mu_{i}$ of the covariance operator $\mathcal{C}_{X}$ defined in (2.5) are given in Table 1. They exhibit a strong decay, which is not uncommon (also for real market data). The corresponding explained variability, defined by $\sum_{j=1}^{i} \lambda_{j} / \sum_{j=1}^{6} \lambda_{j}$, is also shown in the table.

Table 1
Basket option-Eigenvalues and explained variability.

| $i$ | $\mu_{i}$ | $\mu_{i} / \mu_{1}$ | Expl. var. |
| :---: | :---: | :---: | :---: |
| 1 | 0.4490 | 1.0000 | 0.8096 |
| 2 | 0.0623 | 0.1389 | 0.9221 |
| 3 | 0.0177 | 0.0395 | 0.9541 |
| 4 | 0.0106 | 0.0237 | 0.9732 |
| 5 | 0.0079 | 0.0177 | 0.9875 |
| 6 | 0.0069 | 0.0154 | 1.0000 |

The computed POD components are displayed in Figure 3. They resemble those known from fixed income markets. In particular, the first three basis vectors represent the typical shift, tilt, and bend. Further components feature higher frequencies.


Figure 3. Basket-The first four POD basis components.

We now examine the errors of the PIDE method. The "exact" reference solution used here is the result from $10^{7} \mathrm{MC}$ simulations with a standard deviation of 0.0063 (estimated from 100 MC series). It turns out that, due to the stable discontinuous Galerkin method, the mesh size for the time variable has little influence on the PIDE results. Thus we do all computations with fixed, equidistant time steps of size $\Delta t=\frac{1}{10}$. The spatial grid, on the other hand, has considerable impact on the accuracy. Figure 4 displays the (signed) relative error for different dimensions $d$ of the projected problem. The number $N:=2^{M}$ is the maximal number of discretization points in one coordinate and is always taken to be a power of 2 . Two effects can be observed here. For each fixed value of $N$, the method converges to a certain limit when the dimension of the problem is increased. These limits, in turn, converge to the exact solution with increasingly fine meshes. Thus, we might accidentally get a very precise result for a low-dimensional computation on a coarse grid if the two errors (from dimension and mesh) happen to cancel each other. In practice, both the dimension and the number of grid points have to be chosen large enough in order to guarantee a precise result. In our case, $d=4$ and $N=2^{8}$ are sufficient to obtain relative errors below $1 \%$. Approximating the basket value with a log-normal distribution (Lévy approximation), on the other hand, yields a relative error of $3.1 \%$.

Finally, we have a look at the computational time needed for the PIDE method. To this end, we fix $N=2^{9}$ and plot the error for various dimensions. Figure 5 displays the results; both $y$-axes (for time and error) have a logarithmic scale. The error decreases approximately exponentially with increasing dimension. The computational time, on the other hand, in-


Figure 4. Basket—Relative error of PIDE method with different meshes.
creases exponentially. Note that the solution of the problem without dimension reduction takes approximately 2 minutes, even though we use the sparse grid combination technique. However, the increase of the projected problem dimension $d$ by one increases the computational effort by only a factor of 4 to 5 , despite the fact that we use up to $2^{9}=512$ grid points in each coordinate. A reasonably precise solution (in practice within the bounds of the model error) can be computed within a few seconds. While this is slightly faster than the MC method, it is not an extreme gain. The computational effort might be greatly reduced by using more sophisticated grids, featuring more grid points around the origin and fewer close to the boundary of the domain, which we will not consider here.

Results for electricity swaption. In the test case for electricity swaptions, two POD basis components are already sufficient to explain almost $100 \%$ of the volatility, and the sum of the remaining eigenvalues satisfies $\sum_{i=3}^{\infty} \mu_{i} \approx 0$. This is of course due to the strong correlation between the price changes for different maturities $u \in\left[T_{1}, T_{2}\right]$, which makes the dimension reduction technique a particularly well suited method for this type of derivative. The forward curve defined on the delivery interval does not change its shape arbitrarily. The two POD components accurately describe the possible shape changes in our (rather simple) test setting. We are thus able to compute accurate prices by solving a two-dimensional PIDE.

Figure 6 displays the relative error and time of the PIDE method for different mesh widths. As was to be expected, the computational effort increases exponentially in the number of grid refinements (and thus linear in the total number of grid points). The relative error decreases exponentially. However, in contrast to the basket option, a very accurate solution can be


Figure 5. Basket—Relative error and computational time of PIDE method with at most $N=2^{9}$ grid points per coordinate.


Figure 6. Swaption-Relative error and computational time of PIDE method for dimension $d=2$.
computed for the electricity swaption within a fraction of a second. This is made possible by the very efficient dimension reduction from $n=30$ to $d=2$.

A comparison of the PIDE method, MC simulation, and the log-normal approximation is displayed in Figure 7. The log-normal approximation yields a relative error of $3.5 \%$, which is by far the largest of all the methods. MC simulation yields very good results for $10^{6}$ runs and above. However, with $n=30$ it takes considerably longer than in the case of the sixdimensional basket. The standard deviation after $10^{6}$ simulations is 0.4667 (estimated from 100 MC series). The PIDE solver is indeed the fastest and most accurate method for this second test problem.


Figure 7. Swaption-PIDE solutions $(d=2), M C$ simulations, and log-normal approximation.
6. Conclusion. In this article, numerical pricing of European options in Hilbert spacevalued jump-diffusion models is discussed. We have presented a feasible approach employing PIDEs and a dimension reduction method based on Karhunen-Loève expansion. The PIDE can be projected to an approximating low-dimensional equation by solving an eigenvalue problem. Existence, uniqueness, and convergence of the approximating solutions have been shown based on the variational formulation of the PIDE. The numerical solution of the problem is based on a sparse grid combination method for spatial discretization and a discontinuous Galerkin time stepping method.

Numerical experiments have been performed for two applications: an electricity swaption and a basket option. The swaption can be priced very efficiently with the presented algorithm, which gives more accurate results in less time than MC simulation. For the basket option, the PIDE method yields comparable performance to MC simulation, depending on the correlation
of the assets. Using PIDEs, however, should have considerable advantages when pricing pathdependent options or options driven by infinite activity Lévy models, both topics for future research. In addition, future work will be concerned with adaptive and graded grids, further improving the performance of the method.

Acknowledgments. Many thanks go to Claudia Klüppelberg and Fred Espen Benth for helpful discussions and critical reading of the manuscript. Their advice and useful suggestions are highly appreciated. I would also like to thank the referees for their helpful remarks.

## REFERENCES

[1] N. Audet, P. Heiskanen, J. Keppo, and I. Vehviläinen, Modeling electricity forward curve dynamics in the Nordic market, in Modelling Prices in Competitive Electricity Markets, D. W. Bunn, ed., Wiley Series in Financial Economics, John Wiley and Sons, Chichester, 2004, pp. 251-265.
[2] A. Bensoussan and J.-L. Lions, Impulse Control and Quasi-variational Inequalities, Gauthier-Villars, Montrouge, France, 1984.
[3] F. E. Benth, Á. Cartea, and R. Kiesel, Pricing forward contracts in power markets by the certainty equivalence principle: Explaining the sign of the market risk premium, J. Bank. Finance, 32 (2008), pp. 2006-2021.
[4] F. E. Benth and S. Koekebakker, Stochastic modeling of financial electricity contracts, Energ. Econ., 30 (2008), pp. 1116-1157.
[5] F. E. Benth, S. Koekebakker, and F. Ollmar, Extracting and applying smooth forward curves from average-based commodity contracts with seasonal variation, J. Derivatives, 52 (2007), pp. 62-66.
[6] F. E. Benth and T. Meyer-Brandis, The information premium for non-storable commodities, J. Energy Markets, 2 (2009), pp. 111-140.
[7] F. E. Benth, J. Săltyté Benth, and S. Koekebakker, Stochastic Modeling of Electricity and Related Markets, World Scientific, Hackensack, NJ, 2008.
[8] H.-J. Bungartz and M. Griebel, Sparse grids, Acta Numer., 13 (2004), pp. 147-269.
[9] R. Carmona and V. Durrleman, Generalizing the Black-Scholes formula to multivariate contingent claims, J. Comput. Finance, 9 (2006), pp. 43-67.
[10] R. Carmona and M. Tehranchi, Interest Rate Models: An Infinite-Dimensional Stochastic Analysis Perspective, Springer, Berlin, 2006.
[11] R. Cont and P. Tankov, Financial Modelling with Jump Processes, Chapman \& Hall/CRC Financ. Math. Ser., Chapman \& Hall/CRC, Boca Raton, FL, 2004.
[12] R. Cont and E. Voltchkova, A finite difference scheme for option pricing in jump diffusion and exponential Lévy models, SIAM J. Numer. Anal., 43 (2005), pp. 1596-1626.
[13] R. Cont and E. Voltchkova, Integro-differential equations for option prices in exponential Lévy models, Finance Stoch., 9 (2005), pp. 299-325.
[14] G. Da Prato and J. Zabczyk, Second Order Partial Differential Equations in Hilbert Spaces, Cambridge University Press, Cambridge, UK, 2002.
[15] G. Da Prato and J. Zabczyk, Stochastic Equations in Infinite Dimensions, Cambridge University Press, Cambridge, UK, 1992.
[16] L. C. Evans, Partial Differential Equations, American Mathematical Society, Providence, RI, 1998.
[17] E. Hausenblas, A note on the Itô formula of stochastic integrals in Banach spaces, Random Oper. Stochastic Equations, 14 (2006), pp. 45-58.
[18] E. Hausenblas, Existence, uniqueness and regularity of parabolic SPDEs driven by Poisson random measure, Electron. J. Probab., 10 (2005), pp. 1496-1546.
[19] M. Hegland, J. Garcke, and V. Challis, The combination technique and some generalisations, Linear Algebra Appl., 420 (2007), pp. 249-275.
[20] R. Kiesel, G. Schindlmayr, and R. Börger, A two-factor model for the electricity forward market, Quant. Finance, 9 (2009), pp. 279-287.
[21] K. Kunisch and S. Volkwein, Galerkin proper orthogonal decomposition methods for parabolic problems, Numer. Math., 90 (2001), pp. 117-148.
[22] H. Kunita, Stochastic integrals based on martingales taking values in Hilbert space, Nagoya Math. J., 38 (1970), pp. 41-52.
[23] M. Loève, Probability Theory II, 4th ed., Springer, New York, 1978.
[24] A.-M. Matache, T. von Petersdorff, and C. Schwab, Fast deterministic pricing of options on Lévy-driven assets, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 37-71.
[25] S. Peszat and J. Zabczyk, Stochastic Partial Differential Equations with Lévy Noise: An Evolution Equation Approach, Cambridge University Press, Cambridge, UK, 2007.
[26] C. Pflaum and A. Zhou, Error analysis of the combination technique, Numer. Math., 84 (1999), pp. 327-350.
[27] P. E. Protter, Stochastic Integration and Differential Equations, 2nd ed., Springer, Berlin, Heidelberg, New York, 2005.
[28] N. Reich, Wavelet Compression of Anisotropic Integrodifferential Operators on Sparse Tensor Product Spaces, Ph.D. thesis, ETH Zürich, Zürich, Switzerland, 2008; available online from http://e-collection. ethbib.ethz.ch/eserv/eth:30174/eth-30174-02.pdf.
[29] B. Rüdiger and G. Ziglio, Itô formula for stochastic integrals w.r.t. compensated Poisson random measures on separable Banach spaces, Stochastics, 78 (2006), pp. 377-410.
[30] E. W. Sachs and A. K. Strauss, Efficient solution of a partial integro-differential equation in finance, Appl. Numer. Math., 58 (2008), pp. 1687-1703.
[31] K. Sato, Lévy Processes and Infinitely Divisible Distributions, Cambridge University Press, Cambridge, UK, 1999.
[32] C. Schwab and R. A. Todor, Karhunen-Loève approximation of random fields by generalized fast multipole methods, J. Comput. Phys., 217 (2006), pp. 100-122.
[33] V. Thomée, Galerkin Finite Element Methods for Parabolic Problems, 2nd ed., Springer, Berlin, Heidelberg, New York, 2006.
[34] R. A. Todor, Robust eigenvalue computation for smoothing operators, SIAM J. Numer. Anal., 44 (2006), pp. 865-878.
[35] D. Werner, Funktionalanalysis, 5th ed., Springer, Berlin, Heidelberg, New York, 2005.
[36] C. Winter, Wavelet Galerkin Schemes for Option Pricing in Multidimensional Lévy Models, Ph.D. thesis, ETH Zürich, Zürich, Switzerland, 2009; available online from http://e-collection.ethbib.ethz. ch/eserv/eth:41555/eth-41555-02.pdf.
[37] J. Wloka, Partial Differential Equations, Cambridge University Press, Cambridge, UK, 1987.
[38] G. Xu and H. Zheng, Approximate basket options valuation for a jump-diffusion model, Insurance Math. Econom., 45 (2009), pp. 188-194.

# Optimal Trade Execution and Absence of Price Manipulations in Limit Order Book Models* 

Aurélien Alfonsi ${ }^{\dagger}$ and Alexander Schied ${ }^{\ddagger}$

Abstract. We analyze the existence of price manipulation and optimal trade execution strategies in a model for an electronic limit order book with nonlinear price impact and exponential resilience. Our main results show that, under general conditions on the shape function of the limit order book, placing deterministic trade sizes at trading dates that are homogeneously spaced is optimal within a large class of adaptive strategies with arbitrary trading dates. This extends results from our earlier work with A. Fruth. Perhaps even more importantly, our analysis yields as a corollary that our model does not admit price manipulation strategies. This latter result contrasts the recent findings of Gatheral [Quant. Finance, to appear], where, in a related but different model, exponential resilience was found to give rise to price manipulation strategies when price impact is nonlinear.

Key words. liquidity risk, optimal portfolio liquidation, block trade execution, limit order book, market impact model, price manipulation strategies

AMS subject classifications. 91B26, 91G99, 49K99, 93E20
DOI. 10.1137/090762786

1. Introduction. The problem of optimal trade execution is concerned with the optimal acquisition or liquidation of large asset positions. In doing so, it is usually beneficial to split up the large order into a sequence of partial orders, which are then spread over a certain time horizon, so as to reduce the overall price impact and the trade execution costs. The optimization problem at hand is thus to find a trading strategy that minimizes a cost criterion under the constraint of overall order trade execution within a given time frame. There are several reasons why studying this problem is interesting.

First, liquidity risk is one of the least understood sources of financial risk, and one of its various aspects is the risk resulting from price impact created by trading large positions. Due to the nonlinear feedback effects on dynamic trading strategies, market impact risk is probably also among the most fascinating aspects of liquidity risk for mathematicians. The optimal trade execution problem allows studying market impact risk in its purest form. Moreover, the results obtained for this problem can serve as building blocks in a realistic analysis of more complex problems such as the hedging of derivatives in illiquid markets.

[^42]Second, the mathematical analysis of optimal trade execution strategies can help in the ongoing search for viable market impact models. As argued by Huberman and Stanzl [14] and Gatheral [12], any reasonable market impact model should not admit price manipulation strategies in the sense that there are no "round trips" (i.e., trading strategies with zero balance in shares) whose expected trading costs are negative. Since every round trip can be regarded as the execution of a zero-size order, a solution of the optimal trade execution problem also includes an analysis of price manipulation strategies in the model (at least as a limiting case when the order size tends to zero).

In recent years, the problem of optimal trade execution was considered for various market impact models and cost functions by authors such as Bertsimas and Lo [9], Almgren and Chriss [6, 5], Almgren [4], Obizhaeva and Wang [16], Almgren and Lorenz [7], Schied and Schöneborn [18], and Schied, Schöneborn, and Tehranchi [19], and by our joint papers with A. Fruth $[1,2]$, to mention only a few.

Here we continue our analysis from [1, 2]. Instead of focussing on the two-sided limit order book model in [2], our emphasis is now on a zero-spread market impact model that is obtained from the one in [2] by collapsing the bid-ask spread. There are several advantages to introducing this model. First, it is easier to analyze than the two-sided model, while it still allows one to transfer results to the two-sided framework; ${ }^{1}$ see section 2.6. Second, most other market impact models in the literature, such as those suggested by Almgren and Chriss $[6,5]$ or Gatheral [12], do not include a bid-ask spread. Therefore, these models and their features can be compared much better to our zero-spread model than to the two-sided model. Third, it is easier to detect model irregularities, such as price manipulation strategies, in the zero-spread model. In the two-sided model, such irregularities may not emerge on an explicit level.

In our model, the limit order book consists of a certain distribution of limit ask orders at prices higher than the current price, while for lower prices there is a continuous distribution of limit buy orders. We consider a large trader who is placing market orders in this order book and thereby shifts prices according to the volume of available limit orders. Since the distribution of limit orders is allowed to be nonuniform, the price impact created by a market order is typically a nonlinear function of the order size. In reaction to price shocks created by market orders, there is a subsequent recovery of the price within a certain time span. That is, the price evolution will exhibit a certain resilience. Thus, the price impact of a market order will neither be completely instantaneous nor entirely permanent but will decay exponentially with a time-dependent resilience rate. As in [2], we consider the following two distinct possibilities for modeling the resilience of the limit order book after a large market order: the exponential recovery of volume impact, or the exponential recovery of price impact.

This model is quite close to descriptions of price impact on limit order books found in empirical studies such as Biais, Hillion, and Spatt [10], Potters and Bouchaud [17], Bouchaud et al. [11], and Weber and Rosenow [20]. In particular, the existence of a strong resilience effect, which stems from the placement of new limit orders close to the bid-ask spread, seems to be a well-established fact, although its quantitative features seem to be the subject of an ongoing discussion.

[^43]While in [2], we considered only strategies whose trades are placed at equidistant times, we now allow the trading times to be stopping times. This problem description is more natural than prescribing a priori the dates at which trading may take place. It is also more realistic than the idealization of trading in continuous time. In addition, the time-inhomogeneous description allows us to account for time-varying liquidity and thus in particular for the wellknown U-shaped patterns in intraday market parameters; see, e.g., [15].

Optimal trade execution in this extended framework leads to the problem of optimizing simultaneously over both trading times and sizes. This problem is more complex than the one considered in [2] and requires new arguments. Nevertheless, our main results show that the unique optimum is attained by placing the deterministic trade sizes identified in [2] at trading dates that are homogeneously spaced with respect to the average resilience rate in between trades.

As a corollary, we show that neither of the two variants of our model admits price manipulation strategies in the sense of Huberman and Stanzl [14] and Gatheral [12], provided that the shape function of the limit order book belongs to a certain class of functions (which differs slightly for each variant). This corollary is surprising in view of recent results by Gatheral [12]. There it was shown that, in a closely related but different market impact model, exponential resilience leads to the existence of price manipulation strategies as soon as price impact is nonlinear.

This paper is organized as follows. In section 2.1, we introduce our market impact model with its two variants. The cost optimization problem is explained in section 2.2. In section 2.3, we state our main results for the case of a block-shaped limit order book, which corresponds to linear price impact. This special case is much simpler than the case with nonlinear price impact. We therefore give a self-contained description and proof for this case, so that the reader can gain a quick intuition on why our results are true. The proofs for the blockshaped case rely on the results from our earlier paper [1] with A. Fruth and are provided in section 3.1. The main results for the model variant with reversion of volume impact are stated in section 2.4. The corresponding proofs are given in section 3.2. The results for the model variant with reversion of price impact are stated in section 2.5 , while proofs are given in section 3.3. In section 2.6, we explain how our results can be transferred to the case of a two-sided limit order book model.
2. Setup and main results. In this section, we first introduce the two variants of our market impact model and formulate the optimization problem. We then state our results for the particularly simple case of a block-shaped limit order book. Subsequently, we formulate our theorems for each model variant individually. Finally, we explain how the results for the models described in section 2.1 can be transferred to the case of a nonvanishing bid-ask spread.
2.1. Description of the market impact models. The model variants that we consider here are time-inhomogeneous versions of the zero-spread models introduced in [2, Appendix A]. The general aim is to model the dynamics of a limit order book that is exposed to repeated market orders by a large trader whose goal is to liquidate a portfolio of $X_{0}$ shares within a certain time period $[0, T]$. The case $X_{0}>0$ corresponds to a long position and hence to a sell program, and the case $X_{0}<0$ corresponds to a buy program. Here we neglect the bid-ask
spread of the limit order book, but in section 2.6 we will explain how our results can be carried over to limit order book models with nonvanishing bid-ask spread. In these two-sided models, buy orders impact only the ask side of the limit order book and sell orders affect only the bid side. Nevertheless, we will see that the optimal strategies are the same as in the zero-spread models.

When the large trader is inactive, the dynamics of the limit order book are determined by the actions of noise traders only. We assume that the corresponding unaffected price process $S^{0}$ is a right-continuous martingale on a given filtered probability space $\left(\Omega,\left(\mathcal{F}_{t}\right), \mathcal{F}, \mathbb{P}\right)$ such that $S_{0}^{0}$ is $\mathbb{P}$-a.s. equal to some constant $S_{0}$. The actual price process $S$ is driven by the dynamics of $S^{0}$ and by the response of the limit order book to the market orders of the large traders. The key to modeling this response is to start by first describing the volume impact process $E$. If at time $t$ the trader places a market order of size $\xi_{t}$, where $\xi_{t}>0$ stands for a buy order and $\xi_{t}<0$ for a sell order, the volume impact process jumps from $E_{t}$ to

$$
\begin{equation*}
E_{t+}:=E_{t}+\xi_{t} . \tag{1}
\end{equation*}
$$

When the large trader is inactive in between market orders, $E$ reverts back at a given rate. In our first model variant, the Model with volume impact reversion, we assume that the volume impact process reverts on an exponential scale with a deterministic, time-dependent rate $t \mapsto \rho_{t}$, called resilience speed. More precisely, we assume that

$$
\begin{equation*}
d E_{t}=-\rho_{t} E_{t} d t \tag{2}
\end{equation*}
$$

while the large investor is not placing buy orders. Equations (1) and (2) completely determine the dynamics of the volume impact process $E$ in our first model variant.

In the next step, we describe the relation between volume impact and price impact. To this end, we assume a continuous distribution of bid and ask orders away from the unaffected price $S_{t}^{0}$. This distribution is described by a continuous function $f: \mathbb{R} \rightarrow[0, \infty)$ that satisfies $f(x)>0$ for almost every $x$, the shape function. Its intuitive meaning is that the number of shares offered at price $S_{t}^{0}+x$ is given by $f(x) d x$. Thus, a volume impact of $E_{t}$ shares corresponds to a price impact of $D_{t}$, which is given implicitly via

$$
\int_{0}^{D_{t}} f(x) d x=E_{t} .
$$

By introducing the antiderivative of $f$,

$$
F(y):=\int_{0}^{y} f(x) d x, \quad y \in \mathbb{R},
$$

the relation between the volume impact process $E$ and the price impact process $D$ can be expressed as follows:

$$
\begin{equation*}
E_{t}=F\left(D_{t}\right) \quad \text { and } \quad D_{t}=F^{-1}\left(E_{t}\right) \tag{3}
\end{equation*}
$$

Here we have used our assumption that $f>0$ a.e. so that $F$ is indeed invertible. Given the price impact process $D$, the actual price process $S$ is defined as

$$
\begin{equation*}
S_{t}=S_{t}^{0}+D_{t} \tag{4}
\end{equation*}
$$

Thus, if at time $t$ the trader places a market order of size $\xi_{t}$, then the price process jumps from $S_{t}$ to

$$
S_{t+}=S_{t}^{0}+D_{t+}=S_{t}^{0}+F^{-1}\left(E_{t}+\xi_{t}\right)
$$

see Figure 1. Hence, the price impact $D_{t+}-D_{t}$ will be a nonlinear function of the order size $\xi_{t}$ unless $f$ is constant between $D_{t}$ and $D_{t+}$. The choice of a shape function that is constant throughout $\mathbb{R}$ corresponds to linear market impact and to the zero-spread version of the block-shaped limit order book model of Obizhaeva and Wang [16]. This zero-spread version was introduced in [1].


Figure 1. The price impact of a buy market order of size $\xi_{t}>0$ is defined by the equation $\xi_{t}=$ $\int_{D_{t}}^{D_{t+}} f(x) d x=F\left(D_{t+}\right)-F\left(D_{t}\right)$.

Instead of an exponential resilience of the volume impact as described in (2), one can also assume an exponential reversion of the price impact $D$. This means that one has to replace (2) by

$$
\begin{equation*}
d D_{t}=-\rho_{t} D_{t} d t \tag{5}
\end{equation*}
$$

while the large investor is not placing buy orders. The resulting model variant will be called the Model with price impact reversion. We now summarize the definitions of our two model variants.

Definition 2.1. The dynamics of the Model with volume impact reversion are described by (1), (2), (3), and (4). The Model with price impact reversion is defined via (1), (5), (3), and (4). ${ }^{2}$

With the reversion of the processes $D$ and $E$ as described in (2) and (5), we model the well-established empirical fact that order books exhibit a certain resilience as to the price impact of a large buy market order. That is, after the initial impact the best ask price reverts back to its previous position; cf. Biais, Hillion, and Spatt [10], Potters and Bouchaud [17], Bouchaud et al. [11], and Weber and Rosenow [20] for empirical studies.

[^44]Note that we assume that the shape function of the limit order book is neither timedependent nor subject to additional randomness. This assumption is similar to the assumption of fixed, time-independent impact functions in the models by Bertsimas and Lo [9], Almgren [4], or Gatheral [12]. It can also be justified at least partially by the observation that the shapes of empirical limit order books for certain liquid stocks can be relatively stable over time. More importantly, it was noted in [2] that our optimal strategies are remarkably robust with respect to changes in the shape function $f$. One can therefore expect that a moderate randomization of $f$ will have only relatively small effects on the optimal strategy. A corresponding analysis will be the subject of future research.

We now introduce three example classes for shape functions that satisfy the assumptions of our main results.

Example 2.2 (block shape). The simplest example corresponds to a block-shaped limit order book:

$$
\begin{equation*}
f(x) \equiv q \quad \text { for some constant } q>0 . \tag{6}
\end{equation*}
$$

It corresponds to the zero-spread version of the block-shaped limit order book model of Obizhaeva and Wang [16], which was introduced in [1]. In this case, the price impact function is linear: $F^{-1}(x)=x / q$. It follows that the processes $D$ and $E$ are related via $E_{t}=q D_{t}$, and so the model variants with volume and price impact reversion coincide. Results and proofs are particularly easy in this case. We therefore discuss it separately in section 2.3.

Example 2.3 (positive power-law shape). Consider the class of shape functions

$$
\begin{equation*}
f(x)=\lambda|x|^{\alpha}, \quad \text { where } \lambda, \alpha>0 . \tag{7}
\end{equation*}
$$

In this case, $F(x)=\frac{\lambda}{1+\alpha}|x|^{1+\alpha} \operatorname{sign} x$, and so $F^{-1}(x)=\left(\frac{1+\alpha}{\lambda}|x|\right)^{1 /(1+\alpha)} \operatorname{sign} x$. Hence, price impact follows a power law. The choice $\alpha=1$ corresponds to square-root impact, which is a particularly popular choice and admits certain justifications; see [12]. See also [8] for empirical results on power-law impact. We will see later that the shape functions from the class (7) satisfy the assumptions of our main results. Moreover, as was kindly pointed out to us by Jim Gatheral, our Model with volume impact reversion is equivalent to the Model with price impact reversion when we replace $\rho_{t}$ by $\widetilde{\rho}_{t}:=\frac{\rho_{t}}{1+\alpha}$. Indeed, when the large trader is not active during the interval $[t, t+s)$, volume impact reversion implies that $E_{t+s}=e^{-\int_{t}^{t+s} \rho_{u} d u} E_{t}$. Hence,

$$
D_{t+s}=F^{-1}\left(E_{t+s}\right)=e^{-\int_{t}^{t+s} \tilde{\rho}_{u} d u} F^{-1}\left(E_{t}\right)=e^{-\int_{t}^{t+s} \tilde{\rho}_{u} d u} D_{t},
$$

and so $D$ satisfies $d D_{r}=-\widetilde{\rho}_{r} D_{r} d r$ in $[t, t+s)$.
Example 2.4 (negative power-law shape). Consider the shape functions of the form

$$
f(x)= \begin{cases}\frac{q}{\left(1+\lambda_{+} x\right)^{\alpha_{+}}} & \text {for } x>0 \\ \frac{q}{\left(1-\lambda_{-} x\right)^{\alpha_{-}}} & \text {for } x<0\end{cases}
$$

where $q$ and $\lambda_{ \pm}$are positive constants and $\alpha_{ \pm} \in(0,1]$. We will see later that these shape functions satisfy the assumptions of our main results.
2.2. The cost optimization problem. We assume that the large trader needs to liquidate a portfolio of $X_{0}$ shares until time $T$ and that trading can occur at $N+1$ trades within the time interval $[0, T]$. An admissible sequence of trading times will be a sequence $\mathcal{T}=\left(\tau_{0}, \ldots, \tau_{N}\right)$ of stopping times such that $0=\tau_{0} \leq \tau_{1} \leq \cdots \leq \tau_{N}=T$. For such an admissible sequence of trading times, $\mathcal{T}$, we define a $\mathcal{T}$-admissible trading strategy as a sequence $\boldsymbol{\xi}=\left(\xi_{0}, \xi_{1}, \ldots, \xi_{N}\right)$ of random variables such that

- $X_{0}+\sum_{n=0}^{N} \xi_{n}=0$ (i.e., the strategy liquidates the given portfolio $X_{0}$ ),
- each $\xi_{n}$ is measurable with respect to $\mathcal{F}_{\tau_{n}}$, and
- each $\xi_{n}$ is bounded from below.

The quantity $\xi_{n}$ corresponds to the size of the market order placed at time $\tau_{n}$. Note that we do not a priori require that all $\xi_{n}$ have the same sign; i.e., we also allow for an alternation of buy and sell orders. But we assume that there is some bound on the size of sell orders. Finally, an admissible strategy is a pair $(\mathcal{T}, \boldsymbol{\xi})$ consisting of an admissible sequence of trading times $\mathcal{T}$ and a $\mathcal{T}$-admissible trading strategy $\boldsymbol{\xi}$.

Let us now define the costs incurred by an admissible strategy $(\mathcal{T}, \boldsymbol{\xi})$. When at time $\tau_{n}$ a buy market order of size $\xi_{n}>0$ is placed, the trader will purchase $f(x) d x$ shares at price $S_{\tau_{n}}^{0}+x$, with $x$ ranging from $D_{\tau_{n}}$ to $D_{\tau_{n}+}$. Hence, the total cost of the buy market order amounts to

$$
\begin{equation*}
\pi_{\tau_{n}}(\boldsymbol{\xi}):=\int_{D_{\tau_{n}}}^{D_{\tau_{n}+}}\left(S_{\tau_{n}}^{0}+x\right) f(x) d x=S_{\tau_{n}}^{0} \xi_{n}+\int_{D_{\tau_{n}}}^{D_{\tau_{n}+}} x f(x) d x \tag{8}
\end{equation*}
$$

Similarly, for a sell market order $\xi_{n}<0$, the trader will sell $f(x) d x$ shares at price $S_{\tau_{n}}^{0}+x$, with $x$ ranging from $D_{\tau_{n}+}$ to $D_{\tau_{n}}$. Since the costs of sales should be negative, formula (8) is also valid in the case of a sell order.

The average cost $\mathcal{C}(\boldsymbol{\xi}, \mathcal{T})$ of an admissible strategy $(\boldsymbol{\xi}, \mathcal{T})$ is defined as the expected value of the total costs incurred by the consecutive market orders:

$$
\begin{equation*}
\mathcal{C}(\boldsymbol{\xi}, \mathcal{T})=\mathbb{E}\left[\sum_{n=0}^{N} \pi_{\tau_{n}}(\boldsymbol{\xi})\right] . \tag{9}
\end{equation*}
$$

The problem at hand is thus to minimize the average $\operatorname{cost} \mathcal{C}(\boldsymbol{\xi}, \mathcal{T})$ over all admissible strategies $(\boldsymbol{\xi}, \mathcal{T})$. In doing this, we will assume for simplicity throughout this paper that the function $F$ is unbounded in the sense that

$$
\begin{equation*}
\lim _{x \uparrow \infty} F(x)=\infty \quad \text { and } \quad \lim _{x \downarrow-\infty} F(x)=-\infty \tag{10}
\end{equation*}
$$

That is, we assume that the limit order book has infinite depth. Relaxing this assumption is possible but would require additional constraints on the order sizes in admissible strategies and thus complicate the problem description.

In our earlier paper with A. Fruth, [2], we considered the case of a constant resilience $\rho$ and a fixed, equidistant time spacing $\mathcal{T}_{e q}=\{i T / N \mid i=0, \ldots, N\}$. In this setting, we determined trading strategies that minimize the $\operatorname{cost} \mathcal{C}\left(\boldsymbol{\xi}, \mathcal{T}_{\text {eq }}\right)$ among all $\mathcal{T}_{\text {eq }}$-admissible trading strategies $\boldsymbol{\xi}$. Our goal in this paper consists in simultaneously minimizing over trade times
and sizes. Also, in our present setting of an inhomogeneous resilience function $\rho_{t}$ it is natural to replace the equidistant time spacing by the homogeneous time spacing

$$
\mathcal{T}^{*}=\left(t_{0}^{*}, \ldots, t_{N}^{*}\right)
$$

defined via

$$
\int_{t_{i-1}^{*}}^{t_{i}^{*}} \rho_{s} d s=\frac{1}{N} \int_{0}^{T} \rho_{s} d s, \quad i=1, \ldots, N .
$$

We also define

$$
\begin{equation*}
a^{*}:=e^{-\frac{1}{N} \int_{0}^{T} \rho_{u} d u} . \tag{11}
\end{equation*}
$$

Our main result states that, under certain technical assumptions, $\mathcal{T}^{*}$ is in fact the unique optimal time grid for portfolio liquidation with $N+1$ trades in $[0, T]$. In addition, the unique optimal $\mathcal{T}^{*}$-admissible strategies for both model variants, i.e., for the reversion of price or volume impact, are given by the corresponding trading strategies in [2].

As a corollary to our main results, we are able to show that our models do not admit price manipulation strategies in the following sense, introduced by Huberman and Stanzl [14] (see also Gatheral [12]).

Definition 2.5. $A$ round trip is an admissible strategy $(\overline{\boldsymbol{\xi}}, \overline{\mathcal{T}})$ for $X_{0}=0$. A price manipulation strategy is a round trip $(\overline{\boldsymbol{\xi}}, \overline{\mathcal{T}})$ such that $\mathcal{C}(\overline{\boldsymbol{\xi}}, \overline{\mathcal{T}})<0$.

Our main results also imply that our models do not possess the following model irregularity, which was introduced and discussed in our joint paper [3] with A. Slynko.

Definition 2.6. A model admits transaction-triggered price manipulation if the expected execution costs of a sell (buy) program can be decreased by intermediate buy (sell) trades. More precisely, there is transaction-triggered price manipulation if there exist $X_{0} \in \mathbb{R}$ and a corresponding admissible strategy $(\overline{\boldsymbol{\xi}}, \overline{\mathcal{T}})$ such that

$$
\begin{equation*}
\mathcal{C}(\overline{\boldsymbol{\xi}}, \overline{\mathcal{T}})<\inf \{\mathcal{C}(\boldsymbol{\xi}, \mathcal{T}) \mid(\boldsymbol{\xi}, \mathcal{T}) \text { is admissible and all trades in } \boldsymbol{\xi} \text { have the same sign }\} . \tag{12}
\end{equation*}
$$

By taking $X_{0}=0$ in Definition 2.6, one sees that standard price manipulation in the sense of Definition 2.5 can be regarded as a special case of transaction-triggered price manipulation. It follows that the absence of transaction-triggered price manipulation implies the absence of standard price manipulation. It is possible, however, to construct models that admit transaction-triggered price manipulation but not standard price manipulation. This happens, for instance, if we take a constant shape function $f \equiv q$ and replace exponential resilience by Gaussian decay of price impact; see [3].
2.3. Main results for the block-shaped limit order book. We first discuss our problem in the particularly easy case of a block-shaped limit order book in which $f(x)=q$. In that case, our two model variants with the respective reversion of price and volume impact coincide. It follows from the results in [1] that for every admissible sequence of trading times $\mathcal{T}=\left(\tau_{0}, \ldots, \tau_{N}\right)$ there is a $\mathcal{T}$-admissible trading strategy that minimizes the cost $\mathcal{C}(\cdot, \mathcal{T})$ among all $\mathcal{T}$-admissible trading strategies. This strategy can even be computed explicitly; see [1, Theorem 3.1]. In the following theorem, we consider the problem of optimizing jointly over trading times and sizes.


Figure 2. Relative gain between the extra liquidity cost of the optimal strategy on the optimal grid $\mathcal{T}^{*}$ and the optimal strategy on the equidistant grid $\mathcal{T}_{\text {eq }}$ as a function of $N$ when $T=1,2$, and 5 with the resilience function $\rho(t)=10+8 \cos (t / 2 \pi)$.

Theorem 2.7. In a block-shaped limit order book, there is a $\mathbb{P}$-a.s. unique optimal strategy $\left(\boldsymbol{\xi}^{*}, \mathcal{T}^{*}\right)$ consisting of homogeneous time spacing $\mathcal{T}^{*}$ and the deterministic trading strategy $\boldsymbol{\xi}^{*}$ defined by

$$
\begin{equation*}
\xi_{0}^{*}=\xi_{N}^{*}=\frac{-X_{0}}{2+(N-1)\left(1-a^{*}\right)} \quad \text { and } \quad \xi_{1}^{*}=\cdots=\xi_{N-1}^{*}=\xi_{0}^{*}\left(1-a^{*}\right), \tag{13}
\end{equation*}
$$

where $a^{*}$ is as in (11).
While the preceding theorem is a special case of our main results, Theorems 2.11 and 2.17, it admits a particularly easy proof based on the results in [1]. This proof is given in section 3.1.

Corollary 2.8. In a block-shaped limit order book, there is neither standard nor transactiontriggered price manipulation.

An obvious extension of our model is to allow the resilience rate $\rho_{t}$ to be a progressively measurable stochastic process. In this case, optimal strategies will look different. However, the absence of price manipulation remains valid even for this case; see Remark 3.2 at the end of section 3.1.

Figure 2 gives an illustration of the situation when $\rho(t)=a+b \cos (t /(2 \pi)), 0 \leq t \leq T$. For $a>b>0$, the resilience is greater near the opening and the closure of the stock exchange, and $T$ represents the trade duration in days. We plot here the relative gain, i.e., the quotient of the respective expected costs, for the optimal strategies corresponding to the optimal time grid $\mathcal{T}^{*}$
and the equidistant time grid $\mathcal{T}_{\text {eq }}$. More precisely, we plot the quotient of the respective cost functions defined in (32).
2.4. Main results for volume impact reversion. In this section, we state our main results for the Model with volume impact reversion. They hold under the following assumption. Its first part covers Examples 2.2 and 2.4. Its second part covers the important case of power-law price impact as introduced in Example 2.3.

Assumption 2.9. In the Model with volume impact reversion, we assume in addition to (10) that the shape function $f$ satisfies one of the following conditions (a) and (b):
(a) $f$ is nondecreasing on $\mathbb{R}_{-}$and nonincreasing on $\mathbb{R}_{+}$.
(b) $f(x)=\lambda|x|^{\alpha}$ for constants $\lambda, \alpha>0$.

We start by looking at optimal trading strategies when an admissible sequence of trading times $\mathcal{T}=\left(\tau_{0}, \ldots, \tau_{N}\right)$ is fixed. If $\boldsymbol{\xi}$ is a $\mathcal{T}$-admissible trading strategy and it happens that $\tau_{i}=$ $\tau_{i+1}$, then the corresponding trades, $\xi_{i}$ and $\xi_{i+1}$, are executed simultaneously. We therefore say that two $\mathcal{T}$-admissible trading strategies $\boldsymbol{\xi}$ and $\overline{\boldsymbol{\xi}}$ are equivalent if $\xi_{i}+\xi_{i+1}=\bar{\xi}_{i}+\bar{\xi}_{i+1}$ $\mathbb{P}$-a.s. on $\left\{\tau_{i}=\tau_{i+1}\right\}$.

Proposition 2.10. Suppose that an admissible sequence of trading times $\mathcal{T}$ is given and that Assumption 2.9 holds. Then there exists a $\mathcal{T}$-admissible trading strategy $\boldsymbol{\xi}^{V, \mathcal{T}}, \mathbb{P}$-a.s. unique up to equivalence, that minimizes the cost $\mathcal{C}(\cdot, \mathcal{T})$ among all $\mathcal{T}$-admissible trading strategies. Moreover, $\boldsymbol{\xi}^{V, \mathcal{T}} \neq 0$ for $X_{0} \neq 0$ up to equivalence, and all components of $\boldsymbol{\xi}^{V, \mathcal{T}}$ have the same sign. For $X_{0}=0$, all components of $\boldsymbol{\xi}^{V, \mathcal{T}}$ are zero.

As we will see in the proof of Proposition 2.10, the optimal trading strategy $\boldsymbol{\xi}^{V, \mathcal{T}}$ can be implicitly characterized via a certain nonlinear equation. Our main result for the case of volume impact reversion states, however, that things become much easier when optimizing simultaneously over trading times and sizes.

Theorem 2.11. Under Assumption 2.9, for every $X_{0} \neq 0$ there is a $\mathbb{P}$-a.s. unique optimal strategy $\left(\boldsymbol{\xi}^{V}, \mathcal{T}^{*}\right)$ consisting of homogeneous time spacing $\mathcal{T}^{*}$ and deterministic trading strategy $\boldsymbol{\xi}^{V}$ that is defined as follows. The initial market order $\xi_{0}^{V}$ is the unique solution of the equation

$$
\begin{equation*}
F^{-1}\left(-X_{0}-N \xi_{0}^{V}\left(1-a^{*}\right)\right)=\frac{F^{-1}\left(\xi_{0}^{V}\right)-a^{*} F^{-1}\left(a^{*} \xi_{0}^{V}\right)}{1-a^{*}} \tag{14}
\end{equation*}
$$

the intermediate orders are given by

$$
\begin{equation*}
\xi_{1}^{V}=\cdots=\xi_{N-1}^{V}=\xi_{0}^{V}\left(1-a^{*}\right), \tag{15}
\end{equation*}
$$

and the final order is determined by

$$
\xi_{N}^{V}=-X_{0}-\xi_{0}^{V}-(N-1) \xi_{0}^{V}\left(1-a^{*}\right) .
$$

Moreover, $\xi_{0}^{V} \neq 0$ and all components of $\boldsymbol{\xi}^{V}$ have the same sign. That is, $\boldsymbol{\xi}^{V}$ consists only of nontrivial sell orders for $X_{0}>0$ and only of nontrivial buy orders for $X_{0}<0$.

The preceding results imply the following corollary relating to Definitions 2.5 and 2.6.
Corollary 2.12. Under Assumption 2.9, the Model with volume impact reversion admits neither standard nor transaction-triggered price manipulation.

The absence of price manipulation stated in the preceding corollary remains valid even for the case in which resilience $\rho_{t}$ is stochastic; see Remark 3.2 at the end of section 3.1.

Corollary 2.12 shows that, in our Model with volume reversion, exponential resilience of price impact is well compatible with nonlinear impact governed by a shape function that satisfies Assumption 2.9. We thus deduce that, at least from a theoretical perspective, exponential resilience of a limit order book is a viable possibility for describing the decay of market impact. This contrasts Gatheral's observation [12] that, in a related but different continuoustime model, exponential decay of price impact gives rise to price manipulation in the sense of Definition 2.5 as soon as price impact is nonlinear. Given the strong contrasts between the results in [12] and Corollary 2.12, it is interesting to discuss the relations between the model in [12] and our model.

Remark 2.13 (relation to Gatheral's model). In [12], a continuous-time model is introduced, which is closely related to our model. Formulating a discrete-time variant within our setting leads to the following definition for the actual price process:

$$
\begin{equation*}
S_{t}^{J G}=S_{t}^{0}+\sum_{\tau_{n}<t} h\left(\xi_{n}\right) \psi\left(t-\tau_{n}\right) . \tag{16}
\end{equation*}
$$

Here $h: \mathbb{R} \rightarrow \mathbb{R}$ is the price impact function, and $\psi: \mathbb{R}_{+} \rightarrow \mathbb{R}_{+}$is the decay kernel. The decay kernel describes the time decay of price impact. If we take $\psi(t)=e^{-\rho t}$ for a constant $\rho>0$, (16) takes the form

$$
\begin{equation*}
S_{t}^{J G}=S_{t}^{0}+\sum_{\tau_{n}<t} h\left(\xi_{n}\right) e^{-\rho\left(t-\tau_{n}\right)} . \tag{17}
\end{equation*}
$$

It was shown in section 4 of [12] that the continuous-time version of (17) admits price manipulation in the sense of Definition 2.5 as soon as the function $h$ is not linear. By approximating a continuous-time price manipulation strategy with discrete-time strategies, one sees that this result carries over to the discrete-time framework (17).

When taking a constant resilience speed $\rho$ in our Model with volume impact reversion, the volume impact process is of the form $E_{t}=\sum_{\tau_{n}<t} \xi_{n} e^{-\rho\left(t-\tau_{n}\right)}$, and so our price process is given by

$$
\begin{equation*}
S_{t}=S_{t}^{0}+F^{-1}\left(\sum_{\tau_{n}<t} \xi_{n} e^{-\rho\left(t-\tau_{n}\right)}\right) \tag{18}
\end{equation*}
$$

Thus, the difference between the two models is that in (17) the nonlinear price impact function is applied to each individual trade, while in (18) the price impact is obtained as a nonlinear function of the volume impact. Therefore, both models are different, and there is no contradiction between the results in [12] and Corollary 2.12.

Remark 2.14 (continuous-time limit). Let us briefly discuss the asymptotic behavior of the optimal strategy when the number $N$ of trades tends to infinity. Since for any $N$ we have that $\xi_{0}^{V}$ lies strictly between 0 and $-X_{0}$, we can extract a subsequence that converges to some $\xi_{0}^{V, \infty}$. One therefore checks that the right-hand side of (14) tends to

$$
h_{V}^{\infty}\left(\xi_{0}^{V, \infty}\right):=F^{-1}\left(\xi_{0}^{V, \infty}\right)+\frac{\xi_{0}^{V, \infty}}{f\left(F^{-1}\left(\xi_{0}^{V, \infty}\right)\right)} .
$$

Since $N\left(1-a^{*}\right) \rightarrow \int_{0}^{T} \rho_{s} d s$, the left-hand side of (14) converges as well, and so $\xi_{0}^{V, \infty}$ must be a solution $y$ of the equation

$$
F^{-1}\left(-X_{0}-y \int_{0}^{T} \rho_{s} d s\right)=h_{V}^{\infty}(y)
$$

Note that, under our assumptions, $h_{V}^{\infty}$ is strictly increasing. Hence, the preceding equation has a unique solution, which consequently must be the limit of $\xi_{0}^{V}$ as $N \uparrow \infty$. It follows, moreover, that $N \xi_{1}^{V} \rightarrow \xi_{0}^{V, \infty} \int_{0}^{T} \rho_{s} d s$ and that

$$
\xi_{N}^{V} \longrightarrow-X_{0}-\xi_{0}^{V, \infty}-\xi_{0}^{V} \int_{0}^{T} \rho_{s} d s=: \xi_{T}^{V, \infty}
$$

Thus, the optimal strategy, described in "resilience time" $r(t):=\int_{0}^{t} \rho_{s} d s$, consists of an initial block trade of size $\xi_{0}^{V, \infty}$, continuous buying at constant rate $\xi_{0}^{V, \infty}$ during $(0, T)$, and a final block trade of size $\xi_{T}^{V, \infty}$. Transforming back to standard time leaves the initial and final block trades unaffected, and continuous buying in $(0, T)$ now occurs at the time-dependent rate $\rho_{t} \xi_{0}^{V, \infty}$.

One can expect that this limiting strategy could be optimal in the following continuoustime variant of our model, which is similar to the setup in [13]. A strategy is a predictable process $t \mapsto X_{t}$ of bounded total variation, which describes the number of shares in the portfolio of the trader at time $t$. Given such a strategy, the process $E$ is defined via $E_{0}=0$ and

$$
\begin{equation*}
d E_{t}=d X_{t}-\rho_{t} E_{t} d t \tag{19}
\end{equation*}
$$

and $D$ is given by $D_{t}=F\left(E_{t}\right)$. The optimal strategy obtained above then corresponds to

$$
d X_{t}^{*}=\xi_{0}^{V, \infty} \delta_{0}(d t)+\xi_{0}^{V, \infty} \rho_{t} d t+\xi_{T}^{V, \infty} \delta_{T}(d t)
$$

2.5. Main results for reversion of price impact. In this section, we state our main results for the Model with reversion of price impact. This case is analytically more complicated than the Model with volume impact, because the quantity that decays exponentially is no longer a linear function of the order size. We therefore need a stronger assumption.

Assumption 2.15. In addition to (10), we assume that the shape function $f$ satisfies one of the following conditions (a) and (b):
(a) $f$ is twice differentiable on $\mathbb{R} \backslash\{0\}$, nondecreasing on $\mathbb{R}_{-}$, and nonincreasing on $\mathbb{R}_{+}$and satisfies

$$
\begin{gather*}
x \mapsto x f^{\prime}(x) / f(x) \text { is nondecreasing on } \mathbb{R}_{-}, \text {nonincreasing on } \mathbb{R}_{+}, \\
\text {and }(-1,0] \text {-valued, } \tag{20}
\end{gather*}
$$

$$
\begin{equation*}
1+x \frac{f^{\prime}(x)}{f(x)}+2 x^{2}\left(\frac{f^{\prime}(x)}{f(x)}\right)^{2}-x^{2} \frac{f^{\prime \prime}(x)}{f(x)} \geq 0 \quad \text { for all } x \geq 0 \tag{21}
\end{equation*}
$$

(b) $f(x)=\lambda|x|^{\alpha}$ for constants $\lambda, \alpha>0$.

We will see in Example 2.19 that Assumption 2.15(a) is satisfied for the power-law shape functions from Example 2.4. Also, we know already from Example 2.3 that under Assumption 2.15(b) the Model with reversion of price impact is equivalent to a model with reversion of volume impact.

We start by looking at optimal trading strategies when an admissible sequence of trading times $\mathcal{T}=\left(\tau_{0}, \ldots, \tau_{N}\right)$ is fixed. As in section 2.4 , we say that two $\mathcal{T}$-admissible trading strategies $\boldsymbol{\xi}$ and $\overline{\boldsymbol{\xi}}$ are equivalent if $\xi_{i}+\xi_{i+1}=\bar{\xi}_{i}+\bar{\xi}_{i+1} \mathbb{P}$-a.s. on $\left\{\tau_{i}=\tau_{i+1}\right\}$.

Proposition 2.16. Suppose that an admissible sequence of trading times $\mathcal{T}$ is given and that Assumption 2.15 holds. Then there exists a $\mathcal{T}$-admissible trading strategy $\boldsymbol{\xi}^{P, \mathcal{T}}, \mathbb{P}$-a.s. unique up to equivalence, that minimizes the cost $\mathcal{C}(\cdot, \mathcal{T})$. Moreover, $\boldsymbol{\xi}^{P, \mathcal{T}} \neq 0$ for $X_{0} \neq 0$ up to equivalence, and all components of $\boldsymbol{\xi}^{P, \mathcal{T}}$ have the same sign. For $X_{0}=0$, all components of $\boldsymbol{\xi}^{P, \mathcal{T}}$ are zero.

As in Proposition 2.10, computing the optimal trading strategy $\boldsymbol{\xi}^{P, \mathcal{T}}$ for an arbitrary sequence $\mathcal{T}$ can be quite complicated. But again the structure becomes much easier when also optimizing over the sequence of trading times $\mathcal{T}$. To state the corresponding result, let us recall from (11) the definition of $a^{*}$ and let us introduce the function

$$
h_{P, a^{*}}(x):=x \frac{f\left(x / a^{*}\right) / a^{*}-a^{*} f(x)}{f\left(x / a^{*}\right)-a^{*} f(x)} .
$$

We will see in Lemma 3.8(a) that $h_{P, a^{*}}(x)$ is indeed well defined for all $x \in \mathbb{R}$ as soon as Assumption 2.15 is satisfied.

Theorem 2.17. Suppose that that the shape function $f$ satisfies Assumption 2.15. Then for $X_{0} \neq 0$ there is a $\mathbb{P}$-a.s. unique optimal strategy $\left(\boldsymbol{\xi}^{P}, \mathcal{T}^{*}\right)$, consisting of homogeneous time spacing $\mathcal{T}^{*}$ and deterministic trading strategy $\boldsymbol{\xi}^{P}$, that is defined as follows. The initial market order $\xi_{0}^{P}$ is the unique solution of the equation

$$
\begin{equation*}
F^{-1}\left(-X_{0}-N\left[\xi_{0}^{P}-F\left(a^{*} F^{-1}\left(\xi_{0}^{P}\right)\right)\right]\right)=h_{P, a^{*}}\left(F^{-1}\left(\xi_{0}^{P}\right)\right) \tag{22}
\end{equation*}
$$

the intermediate orders are given by

$$
\begin{equation*}
\xi_{1}^{P}=\cdots=\xi_{N-1}^{P}=\xi_{0}^{P}-F\left(a^{*} F^{-1}\left(\xi_{0}^{P}\right)\right) \tag{23}
\end{equation*}
$$

and the final order is determined by

$$
\xi_{N}^{P}=-X_{0}-N \xi_{0}^{P}+(N-1) F\left(a^{*} F^{-1}\left(\xi_{0}^{P}\right)\right)
$$

Moreover, $\xi_{0}^{P} \neq 0$ and all components of $\boldsymbol{\xi}^{P}$ have the same sign. That is, $\boldsymbol{\xi}^{P}$ consists only of nontrivial sell orders for $X_{0}>0$ and only of nontrivial buy orders for $X_{0}<0$.

Again, the preceding results lead to the following corollary. Its conclusion regarding the absence of price manipulation remains valid even for the case in which resilience $\rho_{t}$ is stochastic; see Remark 3.2 at the end of section 3.1.

Corollary 2.18. Under Assumption 2.15, the Model with price impact reversion admits neither standard nor transaction-triggered price manipulation.

We continue this section by showing that the power-law shape functions from Example 2.4 satisfy Assumption 2.15(a).

Example 2.19 (negative power-law shape). Let us show that the power-law shape functions from Example 2.4 satisfy Assumption 2.15(a). For checking (20) and (21), we concentrate on the branch of $f$ on the positive part of the real line. So let us suppose that

$$
f(x)=\frac{q}{(1+\lambda x)^{\alpha}} \quad \text { for } x>0,
$$

with $\alpha \in[0,1], q, \lambda>0$. We have $x f^{\prime}(x) / f(x)=-\frac{\alpha \lambda x}{1+\lambda x} \in(-1,0]$, which is nonincreasing on $\mathbb{R}_{+}$. Moreover, for $x \geq 0$, we have

$$
1+x \frac{f^{\prime}(x)}{f(x)}+2 x^{2}\left(\frac{f^{\prime}(x)}{f(x)}\right)^{2}-x^{2} \frac{f^{\prime \prime}(x)}{f(x)}=\frac{1+(2-\alpha) \lambda x+\left(1-2 \alpha+\alpha^{2}\right)(\lambda x)^{2}}{(1+\lambda x)^{2}} \geq 0 .
$$

Remark 2.20. With a positive power-law shape, we know from Example 2.3 that the Model with volume impact reversion $\rho_{t}$ is equivalent to the Model with price impact reversion $\tilde{\rho}_{t}=$ $\rho_{t} /(1+\alpha)$. Thus, the strategy defined via (14) and (15) with $a^{*}$ is the same as the one defined by (22) and (23) with $\tilde{a}^{*}=\left(a^{*}\right)^{\frac{1}{1+\alpha}}$. This can be checked by a straightforward calculation.

Remark 2.21. As in Remark 2.14, we can study the asymptotic behavior of the optimal strategy as the number $N$ of trades tends to infinity. First, one checks that $h_{P, a^{*}}$ converges to

$$
h_{P}^{\infty}(x):=x\left(1+\frac{f(x)}{f(x)+x f^{\prime}(x)}\right)
$$

and that $N\left(y-F\left(a^{*} F^{-1}(y)\right)\right)$ tends to $F^{-1}(y) f\left(F^{-1}(y)\right) \int_{0}^{T} \rho_{s} d s$. Now suppose that the equation

$$
F^{-1}\left(-X_{0}-F^{-1}(y) f\left(F^{-1}(y)\right) \int_{0}^{T} \rho_{s} d s\right)=h_{P}^{\infty}\left(F^{-1}(y)\right)
$$

has a unique solution that lies strictly between 0 and $-X_{0}$ and which we will call $\xi_{0}^{P, \infty}$. We then see as in Remark 2.14 that $\xi_{0}^{P, \infty}$ is the limit of $\xi_{0}^{P}$ when $N \uparrow \infty$. Next, $N \xi_{1}^{P}$ converges to $F^{-1}\left(\xi_{0}^{P, \infty}\right) f\left(F^{-1}\left(\xi_{0}^{P, \infty}\right)\right) \int_{0}^{T} \rho_{s} d s$ and $\xi_{N}^{P}$ to

$$
\xi_{T}^{P, \infty}:=-X_{0}-\xi_{0}^{P, \infty}-F^{-1}\left(\xi_{0}^{P, \infty}\right) f\left(F^{-1}\left(\xi_{0}^{P, \infty}\right)\right) \int_{0}^{T} \rho_{s} d s
$$

This yields a description of the continuous-time limit in "resilience time" $r(t):=\int_{0}^{t} \rho_{s} d s$. Using a time change as in Remark 2.14, we obtain that the optimal strategy consists of an initial block order of $\xi_{0}^{P, \infty}$ shares at time 0 , continuous buying at rate $\rho_{t} F^{-1}\left(\xi_{0}^{P, \infty}\right) f\left(F^{-1}\left(\xi_{0}^{P, \infty}\right)\right)$ during $(0, T)$, and a final block order of $\xi_{T}^{P, \infty}$ shares at time $T$. One might guess that this strategy should be optimal in the continuous-time model in which strategies are predictable processes $t \mapsto X_{t}$ of total bounded variation and the volume impact process satisfies

$$
d E_{t}=d X_{t}-\rho_{t} g\left(E_{t}\right) d t
$$

for $g(x)=f\left(F^{-1}(x)\right) F^{-1}(x)$; see also Remark 2.14.
2.6. Two-sided limit order book models. We now explain how our results can be used to analyze models for an electronic limit order book with a nonvanishing and dynamic bid-ask spread. To this end, we focus on a buy program with $X_{0}<0$ (the case of a sell program is analogous). Therefore, emphasis is on buy orders, and we concentrate first on the upper part of the limit order book, which consists of shares offered at various ask prices. The lowest ask price at which shares are offered is called the best ask price. When the large trader is inactive, the dynamics of the limit order book are determined by the actions of noise traders only. We assume that the corresponding unaffected best ask price $A^{0}$ is a right-continuous martingale on a given filtered probability space $\left(\Omega,\left(\mathcal{F}_{t}\right), \mathcal{F}, \mathbb{P}\right)$ and satisfies $A_{0}^{0}=A_{0} \mathbb{P}$-a.s. for some constant $A_{0}$. Above the unaffected best ask price $A_{t}^{0}$, we assume a continuous distribution for ask limit orders: the number of shares offered at price $A_{t}^{0}+x$ with $x \geq 0$ is given by $f(x) d x$, where $f$ is a given shape function.

The actual best ask price at time $t$, i.e., the best ask price after taking the price impact of previous buy orders of the large trader into account, is denoted by $A_{t}$. It lies above the unaffected best ask price, and the price impact on ask prices caused by the actions of the large trader is denoted by

$$
D_{t}^{A}:=A_{t}-A_{t}^{0} .
$$

A buy market order of $\xi_{t}>0$ shares placed by the large trader at time $t$ will consume all the ask limit orders offered at prices between $A_{t}$ and $A_{t+}:=A_{t}^{0}+D_{t+}^{A}$, where $D_{t+}^{A}$ is determined by the condition $\int_{D_{t}^{A}}^{D_{t}^{A}} f(x) d x=\xi_{t}$. Thus, the process $D^{A}$ captures only the impact of market buy orders on the current best ask price. We also define the volume impact on ask orders by $E_{t}^{A}:=F\left(D_{t}^{A}\right)$, where again $F(x)=\int_{0}^{x} f(y) d y$.

On the bid side of the limit order book, we have an unaffected best bid process, $B_{t}^{0}$. All we assume on its dynamics is $B_{t}^{0} \leq A_{t}^{0}$ at all times $t$. The distribution of bids below $B_{t}^{0}$ is modeled by the restriction of the shape function $f$ to the domain $(-\infty, 0)$. In analogy to the ask part, we introduce the the price impact on bid prices by $D_{t}^{B}:=B_{t}-B_{t}^{0}$. The process $D^{B}$ will be nonpositive. A sell market order of $\xi_{t}<0$ shares placed at time $t$ will consume all the shares offered at prices between $B_{t}$ and $B_{t+}:=B_{t}^{0}+D_{t+}^{B}$, where $D_{t+}^{B}$ is determined by the condition

$$
\xi_{t}=F\left(D_{t+}^{B}\right)-F\left(D_{t}^{B}\right)=: E_{t+}^{B}-E_{t}^{B}
$$

for $E_{s}^{B}:=F\left(D_{s}^{B}\right)$. As before, there are now two distinct variants for modeling the reversion of the volume and price impact processes while the trader is not active. More precisely, we assume that

$$
\begin{array}{rlll}
d E_{t}^{A}=-\rho_{t} E_{t}^{A} d t & \text { and } & d E_{t}^{B}=-\rho_{t} E_{t}^{B} d t & \text { for reversion of volume impact, } \\
d D_{t}^{A}=-\rho_{t} D_{t}^{A} d t & \text { and } & d D_{t}^{B}=-\rho_{t} D_{t}^{B} d t & \text { for reversion of price impact. }
\end{array}
$$

The respective model variants will be called the two-sided limit order book models with reversion of volume or price impact.

Finally, we define the costs of an admissible strategy $(\mathcal{T}, \boldsymbol{\xi})$. We can argue as in section 2.2
that the costs incurred at time $\tau_{n}$ should be defined as

$$
\bar{\pi}_{\tau_{n}}(\boldsymbol{\xi})= \begin{cases}\int_{D_{\tau_{n}}^{A}}^{D_{\tau_{n}+}^{A}} f(x) d x & \text { for } \xi_{n}>0  \tag{25}\\ 0 & \text { for } \xi_{n}=0 \\ \int_{D_{\tau_{n}}^{B}}^{D_{\tau_{n}+}^{B}} f(x) d x & \text { for } \xi_{n}<0\end{cases}
$$

The following result compares the costs in the two-sided model to the costs $\pi_{\tau_{n}}(\boldsymbol{\xi})$ defined in (8).

Proposition 2.22. Suppose that $A^{0}=S^{0}$. Then, for any strategy $\boldsymbol{\xi}$, we have $\bar{\pi}_{\tau_{n}}(\boldsymbol{\xi}) \geq \pi_{\tau_{n}}(\boldsymbol{\xi})$ for all $n$, with equality if all trades in $\boldsymbol{\xi}$ are nonnegative.

The preceding result can be proved by arguments given in [2, section A]. Together with the observation that there is no transaction-triggered price manipulation in the models introduced in section 2.1, it provides the key to transferring results on the basic model to the order book model. We thus have the following corollary.

Corollary 2.23. Suppose that $A^{0}=S^{0}$.
(a) Under Assumption 2.9, the strategy $\left(\boldsymbol{\xi}^{V}, \mathcal{T}^{*}\right)$ defined in Theorem 2.11 is the unique optimal strategy in the two-sided limit order book Model with volume impact reversion.
(b) Under Assumption 2.15, the strategy $\left(\boldsymbol{\xi}^{P}, \mathcal{T}^{*}\right)$ defined in Theorem 2.17 is the unique optimal strategy in the two-sided limit order book Model with price impact reversion.
3. Proofs. In a first step, note that the average costs introduced in (9) are of the form

$$
\mathcal{C}(\boldsymbol{\xi}, \mathcal{T})=\mathbb{E}\left[\sum_{n=0}^{N} \pi_{\tau_{n}}(\boldsymbol{\xi})\right]=\mathbb{E}\left[\sum_{n=0}^{N} \xi_{n} S_{\tau_{n}}^{0}\right]+\mathbb{E}\left[\sum_{n=0}^{N} \int_{D_{\tau_{n}}}^{D_{\tau_{n}+}} x f(x) d x\right] .
$$

Due to the martingale property of $S^{0}$, optional stopping, and the fact that $\sum_{n=0}^{N} \xi_{n}=-X_{0}$, the first expectation on the right-hand side is equal to $-X_{0} S_{0}$. Next, note that the process $D$ evolves deterministically once the values of $\tau_{0}(\omega), \ldots, \tau_{N}(\omega)$ and $\xi_{0}(\omega), \ldots, \xi_{N}(\omega)$ are given. Thus, when the functional $\sum_{n=0}^{N} \int_{D_{\tau_{n}}}^{D_{\tau_{n}+}} x f(x) d x$ admits a unique deterministic minimizer, this minimizer must be equal to the unique optimal strategy.

To formulate the resulting deterministic optimization problem, it will be convenient to work with the quantities

$$
\begin{equation*}
\alpha_{k}:=\int_{\tau_{k-1}}^{\tau_{k}} \rho_{s} d s, \quad k=1, \ldots, N, \tag{26}
\end{equation*}
$$

instead of the $\tau_{k}$ themselves. The condition $0=\tau_{0} \leq \tau_{1} \leq \cdots \leq \tau_{N}=T$ is clearly equivalent to $\boldsymbol{\alpha}:=\left(\alpha_{1}, \ldots, \alpha_{N}\right)$ belonging to

$$
\mathcal{A}:=\left\{\boldsymbol{\alpha}:=\left(\alpha_{1}, \ldots, \alpha_{N}\right) \in \mathbb{R}_{+}^{N} \mid \sum_{k=1}^{N} \alpha_{k}=\int_{0}^{T} \rho_{s} d s\right\} .
$$

By abuse of notation, we will write

$$
\begin{equation*}
E_{n} \text { and } D_{n} \text { instead of } E_{\tau_{n}} \text { and } D_{\tau_{n}} \tag{27}
\end{equation*}
$$

as long as there is no possible confusion. We will also write

$$
\begin{equation*}
E_{n+}=E_{n}+\xi_{n} \text { and } D_{n+}=D_{n}+\xi_{n} \text { instead of } E_{\tau_{n}+} \text { and } D_{\tau_{n}+} \tag{28}
\end{equation*}
$$

Then

$$
\begin{array}{ll}
E_{k+1}=e^{-\alpha_{k+1}} E_{k+}=e^{-\alpha_{k+1}}\left(E_{k}+\xi_{k}\right) & \text { for volume impact reversion } \\
D_{k+1}=e^{-\alpha_{k+1}} D_{k+}=e^{-\alpha_{k+1}} F^{-1}\left(\xi_{k}+F\left(D_{k}\right)\right) & \text { for price impact reversion } \tag{29}
\end{array}
$$

With this notation, it follows that there exist two deterministic functions $C^{V}, C^{P}: \mathbb{R}^{N+1} \times$ $\mathcal{A} \rightarrow \mathbb{R}$ such that

$$
\sum_{n=0}^{N} \int_{D_{n}}^{D_{n+}} x f(x) d x= \begin{cases}C^{V}(\boldsymbol{\xi}, \boldsymbol{\alpha}) & \text { in the Model with volume impact reversion }  \tag{30}\\ C^{P}(\boldsymbol{\xi}, \boldsymbol{\alpha}) & \text { in the Model with price impact reversion }\end{cases}
$$

We will show in the respective sections 3.2 and 3.3 that, under our assumptions, the functions $C^{V}$ and $C^{P}$ have unique minima within the set $\Xi \times \mathcal{A}$, where

$$
\Xi:=\left\{\boldsymbol{x}=\left(x_{0}, \ldots, x_{N}\right) \in \mathbb{R}^{N+1} \mid X_{0}+\sum_{n=0}^{N} x_{n}=0\right\}
$$

When working with deterministic trading strategies in $\Xi$ rather than with random variables, we will mainly use Roman letters such as $\boldsymbol{x}$ instead of Greek letters such as $\boldsymbol{\xi}$. Last, we introduce the functions

$$
\begin{equation*}
\tilde{F}(x):=\int_{0}^{x} z f(z) d z \quad \text { and } \quad G=\tilde{F} \circ F^{-1} \tag{31}
\end{equation*}
$$

We conclude this section with the following easy lemma.
Lemma 3.1. For $X_{0}<0$, there is no $\boldsymbol{x} \in \Xi$ such that $E_{n+}=E_{n}+x_{n} \leq 0$ (or, equivalently, $D_{n+} \leq 0$ ) for all $n=0, \ldots, N$.

Proof. Since the effect of resilience is to drive the extra spread back to zero, we have $E_{n+} \geq x_{0}+\cdots+x_{n}$ up to and including the first $n$ at which $x_{0}+\cdots+x_{n}>0$. Since $x_{0}+\cdots+x_{N}=-X_{0}>0$, the result follows.
3.1. Proofs for a block-shaped limit order book. In this section, we give quick and direct proofs for our results in the case of a block-shaped limit order book with $f(x)=q$. In this setting, our two model variants coincide; see Example 2.2. As explained in [1], the cost function in (30) is an increasing affine function of

$$
\begin{equation*}
C(\boldsymbol{x}, \boldsymbol{\alpha})=\frac{1}{2}\langle\boldsymbol{x}, M(\boldsymbol{\alpha}) \boldsymbol{x}\rangle, \quad \boldsymbol{x} \in \Xi, \quad \alpha \in \mathcal{A} \tag{32}
\end{equation*}
$$

where $\langle\cdot, \cdot\rangle$ is the usual Euclidean inner product and $M(\boldsymbol{\alpha})$ is the positive definite symmetric matrix with entries

$$
M(\boldsymbol{\alpha})_{n, m}=\exp \left(-\int_{\tau_{n \wedge m}}^{\tau_{n \vee m}} \rho_{u} d u\right)=\exp \left(-\left|\sum_{i=1}^{n} \alpha_{i}-\sum_{j=1}^{m} \alpha_{j}\right|\right), \quad 0 \leq n, m \leq N .
$$

Proof of Theorem 2.7. For $\boldsymbol{\alpha}$ belonging to

$$
\mathcal{A}^{*}:=\left\{\boldsymbol{\alpha} \in \mathcal{A} \mid \alpha_{i}>0, i=1, \ldots, N\right\},
$$

the inverse $M(\boldsymbol{\alpha})^{-1}$ of the matrix $M(\boldsymbol{\alpha})$ can be computed explicitly, and the unique optimal trading strategy for fixed $\boldsymbol{\alpha}$ is

$$
\boldsymbol{x}^{*}(\boldsymbol{\alpha})=\frac{-X_{0}}{\left\langle\mathbf{1}, M(\boldsymbol{\alpha})^{-1} \mathbf{1}\right\rangle} M(\boldsymbol{\alpha})^{-1} \mathbf{1} .
$$

By [1, Theorem 3.1], the vector $M(\boldsymbol{\alpha})^{-1} \mathbf{1}$ has only strictly positive components for $\boldsymbol{\alpha} \in \mathcal{A}^{*}$. It follows that

$$
\begin{align*}
\min _{x \in \Xi} C(\boldsymbol{x}, \boldsymbol{\alpha}) & =C\left(\boldsymbol{x}^{*}(\boldsymbol{\alpha}), \boldsymbol{\alpha}\right)=\frac{X_{0}^{2}}{2\left\langle\mathbf{1}, M(\boldsymbol{\alpha})^{-1} \mathbf{1}\right\rangle} \\
& =\frac{X_{0}^{2}}{2}\left(\frac{2}{1+e^{-\alpha_{1}}}+\sum_{n=2}^{N} \frac{1-e^{-\alpha_{n}}}{1+e^{-\alpha_{n}}}\right)^{-1}  \tag{33}\\
& =\frac{X_{0}^{2}}{2}\left(\sum_{n=1}^{N} \frac{2}{1+e^{-\alpha_{n}}}-(N-1)\right)^{-1} .
\end{align*}
$$

Minimizing $\min _{\boldsymbol{x} \in \Xi} C(\boldsymbol{x}, \boldsymbol{\alpha})$ over $\boldsymbol{\alpha} \in \mathcal{A}^{*}$ is thus equivalent to maximizing $\sum_{n=1}^{N} \frac{2}{1+e^{-\alpha_{n}}}$. The function $a \mapsto \frac{2}{1+e^{-a}}$ is strictly concave in $a>0$. Hence,

$$
\sum_{n=1}^{N} \frac{2}{1+e^{-\alpha_{n}}} \leq \frac{2 N}{1+e^{-\frac{1}{N} \sum_{n=1}^{N} \alpha_{n}}}=\frac{2 N}{1+e^{-\frac{1}{N} \int_{0}^{T} \rho_{u} d u}}
$$

with equality if and only if $\boldsymbol{\alpha}=\boldsymbol{\alpha}^{*}$, where $\boldsymbol{\alpha}^{*}$ corresponds to homogeneous time spacing $\mathcal{T}^{*}$, i.e.,

$$
\begin{equation*}
\alpha_{i}^{*}=\frac{1}{N} \int_{0}^{T} \rho_{s} d s, \quad i=1, \ldots, N . \tag{34}
\end{equation*}
$$

Next, $C(\boldsymbol{x}, \boldsymbol{\alpha})$ is clearly jointly continuous in $\boldsymbol{x} \in \Xi$ and $\boldsymbol{\alpha} \in \mathcal{A}$, so $\inf _{\boldsymbol{x} \in \Xi} C(\boldsymbol{x}, \boldsymbol{\alpha})$ is upper semicontinuous in $\boldsymbol{\alpha}$. One thus sees that the minimum cannot be attained at the boundary of $\mathcal{A}$. Finally, the formula (13) for the optimal trading strategy with homogeneous time spacing can be found in [1, Remark 3.2] or in [2, Corollary 6.1].

Proof of Corollary 2.8. The result follows immediately from Theorem 2.7.
Remark 3.2 (stochastic resilience). Suppose that the resilience rate $\rho_{t}$ is not necessarily deterministic but can also be a progressively measurable stochastic process. We assume,
moreover, that $\rho_{t}$ is strictly positive and integrable. Then the expected costs of any admissible strategy $(\mathcal{T}, \boldsymbol{\xi})$ will still be of the form

$$
\mathcal{C}(\boldsymbol{\xi}, \mathcal{T})=\mathbb{E}[C(\boldsymbol{\xi}, \boldsymbol{\alpha})] .
$$

Since $C(\boldsymbol{\xi}, \boldsymbol{\alpha}) \geq 0$ for every round trip by Corollary 2.8, the same is true for $\mathcal{C}(\boldsymbol{\xi}, \mathcal{T})$, and hence there cannot be price manipulation for any round trip, even under stochastic resilience. The same argument also applies in the contexts of sections 2.4 and 2.5 .
3.2. Proofs for reversion of volume impact. We need a few lemmas before we can prove our main results for reversion of volume impact. For simplicity, we will first prove our results for $X_{0}<0$. The case for $X_{0}>0$ will then follow by the symmetry of the problem formulation. The case of round trips with $X_{0}=0$ will be analyzed by a limiting procedure in the proof of Proposition 2.10.

Lemma 3.3 formulates analytic properties implied by Assumption 2.9. In fact, only these properties will be needed in the remainder of the proof, and so our results remain valid for all shape functions $f$ that satisfy the conclusions of Lemma 3.3. The next step in our proof is to prove existence and uniqueness of optimal strategies. This is done in Lemma 3.6 by exploiting the fact that, in the Model with volume impact reversion, the cost functional is coercive and strictly convex. Note, however, that convexity will be lost in the Model with price impact reversion. The final and most delicate step of the proof is to show that the first-order condition for optimality yields a nonlinear equation that uniquely determines the optimal strategy. The recursive formula of Lemma 3.4 is a preliminary step into that direction.

Lemma 3.3. Under Assumption 2.9, the following conclusions hold:
(a) For each $a \in(0,1)$, the function $h_{V, a}(y)=F^{-1}(y)-a F^{-1}(a y)$ is strictly increasing on $\mathbb{R}$.
(b) For all $a, b \in(0,1)$ and $\nu>0$, we have the inequalities

$$
\begin{gather*}
h_{V, a}^{-1}(\nu(1-a))>b \cdot h_{V, b}^{-1}(\nu(1-b)),  \tag{35}\\
b \cdot h_{V, b}^{-1}(\nu(1-b))<F(\nu) . \tag{36}
\end{gather*}
$$

(c) The function $H_{V}:(y, a) \in(0, \infty) \times(0,1) \mapsto\left(\frac{F^{-1}(y)-a F^{-1}(a y)}{1-a}, a y \frac{F^{-1}(y)-F^{-1}(a y)}{1-a}\right) \in \mathbb{R}^{2}$ is one-to-one.
Proof of Lemma 3.3 under Assumption 2.9(a). Assumption 2.9(a) states that $f$ is increasing on $\mathbb{R}_{-}$and decreasing on $\mathbb{R}_{+}$. Part (a) of the assertion thus follows from [2, Remark 2].

For the proof of part (b), let $y:=h_{V, a}^{-1}(\nu(1-a))$. Then $y>0$, since $h_{V, a}(0)=0$. Note also that $F^{-1}$ is convex on $\mathbb{R}_{+}$. Let $\widehat{f}$ be its derivative. Then

$$
\begin{aligned}
\nu & =\frac{F^{-1}(y)-a F^{-1}(a y)}{1-a}=F^{-1}(a y)+\frac{F^{-1}(y)-F^{-1}(a y)}{1-a} \\
& =F^{-1}(a y)+\frac{1}{1-a} \int_{a y}^{y} \widehat{f}(x) d x<F^{-1}(y)+y \widehat{f}(y)=: g(y)
\end{aligned}
$$

Clearly, $g$ is a strictly increasing function on $\mathbb{R}_{+}$, and so we have $y>g^{-1}(\nu)$.

Next, let $z:=b \cdot h_{V, b}^{-1}(\nu(1-b))$. Then,

$$
\begin{aligned}
\nu & =\frac{F^{-1}(z / b)-b F^{-1}(z)}{1-b}=F^{-1}(z)+\frac{F^{-1}(z / b)-F^{-1}(z)}{1-b} \\
& =F^{-1}(z)+\frac{1}{1-b} \int_{z}^{z / b} \widehat{f}(x) d x \geq F^{-1}(z)+z \widehat{f}(z)=g(z),
\end{aligned}
$$

since $\widehat{f}(z)=1 / f\left(F^{-1}(z)\right)$ is nondecreasing for $z>0$. Thus, $z \leq g^{-1}(\nu)<h_{V, a}^{-1}(\nu(1-a))$, and (35) follows. For (36), it now suffices to note that $g(z)>F^{-1}(z)$.

To prove part (c), let $a_{1}, a_{2} \in(0,1)$ and $y_{1}, y_{2}>0$ and assume that $H_{V}\left(a_{1}, y_{1}\right)=$ $H_{V}\left(a_{2}, y_{2}\right)$. Since

$$
\frac{F^{-1}(y)-a F^{-1}(a y)}{1-a}=F^{-1}(y)+a \frac{F^{-1}(y)-F^{-1}(a y)}{1-a}
$$

we get

$$
\begin{align*}
F^{-1}\left(y_{1}\right)+a_{1} \frac{F^{-1}\left(y_{1}\right)-F^{-1}\left(a_{1} y_{1}\right)}{1-a_{1}} & =F^{-1}\left(y_{2}\right)+a_{2} \frac{F^{-1}\left(y_{2}\right)-F^{-1}\left(a_{2} y_{2}\right)}{1-a_{2}}, \\
a_{1} y_{1} \frac{F^{-1}\left(y_{1}\right)-F^{-1}\left(a_{1} y_{1}\right)}{1-a_{1}} & =a_{2} y_{2} \frac{F^{-1}\left(y_{2}\right)-F^{-1}\left(a_{2} y_{2}\right)}{1-a_{2}} . \tag{37}
\end{align*}
$$

Assume that $y_{1} \neq y_{2}$, say, $y_{1}>y_{2}>0$. Multiplying the first identity by $y_{1}$ and subtracting the second identity yields

$$
\begin{equation*}
y_{1} \frac{F^{-1}\left(y_{1}\right)-F^{-1}\left(y_{2}\right)}{y_{1}-y_{2}}=\frac{a_{2}}{1-a_{2}}\left[F^{-1}\left(y_{2}\right)-F^{-1}\left(a_{2} y_{2}\right)\right] . \tag{3}
\end{equation*}
$$

Since $\left(F^{-1}\right)^{\prime}(y)=\widehat{f}(y)$ is nondecreasing for $y>0$, we obtain that

$$
y_{1} \frac{F^{-1}\left(y_{1}\right)-F^{-1}\left(y_{2}\right)}{y_{1}-y_{2}} \geq y_{1} \widehat{f}\left(y_{2}\right) \quad \text { and } \quad \frac{a_{2}}{1-a_{2}}\left[F^{-1}\left(y_{2}\right)-F^{-1}\left(a_{2} y_{2}\right)\right] \leq a_{2} y_{2} \widehat{f}\left(y_{2}\right),
$$

which contradicts the previous equation, since $y_{1}>y_{2} \geq a_{2} y_{2}$. Therefore, we must have $y_{1}=y_{2}$.

It is therefore sufficient to show that

$$
\left.\widetilde{h}(a):=\frac{a}{1-a}\left[F^{-1}(y)-F^{-1}(a y)\right], \quad a \in\right] 0,1[,
$$

is one-to-one for any $y>0$. Its derivative is equal to

$$
\begin{equation*}
\widetilde{h}^{\prime}(a)=\frac{1}{(1-a)^{2}}\left[F^{-1}(y)-F^{-1}(a y)\right]-\frac{a y}{1-a} \widehat{f}(a y) . \tag{39}
\end{equation*}
$$

Using again that $\widehat{f}(y)$ is nondecreasing for $y>0$, we get

$$
F^{-1}(y)-F^{-1}(a y)>(1-a) y \widehat{f}(a y)
$$

and in turn $\widetilde{h}^{\prime}(a)>0$.

Proof of Lemma 3.3 under Assumption 2.9(b). Assumption 2.9(b) states that $f(x)=\lambda|x|^{\alpha}$ for constants $\lambda, \alpha>0$. Thus,

$$
\begin{equation*}
h_{V, a}(y)=\left(\frac{1+\alpha}{\lambda}\right)^{\frac{1}{1+\alpha}}\left(1-a^{\frac{2+\alpha}{1+\alpha}}\right)|y|^{\frac{1}{1+\alpha}} \operatorname{sign} y, \tag{40}
\end{equation*}
$$

and so part (a) of the assertion follows.
As for part (b) of the assertion, note that

$$
h_{V, a}^{-1}(\nu(1-a))=\frac{\lambda}{1+\alpha} \nu^{1+\alpha}\left(\frac{1-a}{1-a^{\frac{2+\alpha}{1+\alpha}}}\right)^{1+\alpha} .
$$

Hence, when $a$ goes from 0 to 1 , the value of $h_{V, a}^{-1}(\nu(1-a))$ decreases from $\frac{\lambda}{1+\alpha} \nu^{1+\alpha}$ to $\frac{\lambda}{1+\alpha} \nu^{1+\alpha}\left(\frac{1+\alpha}{2+\alpha}\right)^{1+\alpha}$. Using the shorthand notation $\gamma:=1+\alpha$, the inequality (35) will thus follow if we can show that

$$
\begin{equation*}
b^{\frac{1}{\gamma}} \frac{1-b}{1-b^{\frac{1+\gamma}{\gamma}}}<\frac{\gamma}{1+\gamma} \tag{41}
\end{equation*}
$$

for $0<b<1$ and $\gamma>1$. The preceding inequality is equivalent to

$$
1-b<\frac{\gamma}{1+\gamma}\left(b^{-1 / \gamma}-b\right) .
$$

But the function on the right-hand side is a strictly convex decreasing function of $b$ whose derivative at $b=1$ is -1 . This proves the asserted inequalities and in turn (35). Inequality (36) is obvious, given our formulas for $h_{V, b}^{-1}(\nu(1-b))$ and $F$.

To prove (c), we use the same argument as under Assumption 2.9(a). First, let us observe that $\Psi: x \in \mathbb{R}_{+} \mapsto x\left(x^{1 /(1+\alpha)}-1\right) /(x-1)$ is increasing, which can be easily checked by derivation. If $y_{1} \neq y_{2}$, say, $z=y_{1} / y_{2}>1$, we get $\Psi(z)=\Psi\left(a_{2}\right)$ from (38), which is not possible, since $z>1>a_{2}$. Thus $y_{1}=y_{2}$, and we get from (37) that $\Psi\left(a_{1}\right)=\Psi\left(a_{2}\right)$, which gives $a_{1}=a_{2}$.

Let us turn to the calculation of the cost derivatives. With (31), the cost function (30) in the Model with volume impact reversion can be represented as

$$
\begin{equation*}
C^{V}(\boldsymbol{x}, \boldsymbol{\alpha})=\sum_{n=0}^{N}\left[G\left(E_{n}+x_{n}\right)-G\left(E_{n}\right)\right], \quad \boldsymbol{x} \in \Xi, \quad \boldsymbol{\alpha} \in \mathcal{A}, \tag{42}
\end{equation*}
$$

where

$$
E_{0}=0 \quad \text { and } \quad E_{n}=\sum_{i=0}^{n-1} x_{i} e^{-\sum_{k=i+1}^{n} \alpha_{k}}, \quad 1 \leq n \leq N .
$$

Lemma 3.4. For $i=0, \ldots, N-1$, we have the following recursive formula:

$$
\begin{equation*}
\frac{\partial C^{V}}{\partial x_{i}}=F^{-1}\left(E_{i}+x_{i}\right)-e^{-\alpha_{i+1}} F^{-1}\left(E_{i+1}\right)+e^{-\alpha_{i+1}} \frac{\partial C^{V}}{\partial x_{i+1}} . \tag{43}
\end{equation*}
$$

Moreover, for $i=1, \ldots, N$,

$$
\begin{equation*}
\frac{\partial C^{V}}{\partial \alpha_{i}}=E_{i} \sum_{n=i}^{N}\left[F^{-1}\left(E_{n}+x_{n}\right)-F^{-1}\left(E_{n}\right)\right] e^{-\sum_{k=i+1}^{n} \alpha_{k}} \tag{44}
\end{equation*}
$$

Proof. To prove (43), we first need to calculate $\partial E_{n} / \partial x_{i}$. We obtain

$$
\frac{\partial E_{n}}{\partial x_{i}}=0 \quad \text { if } i \geq n \quad \text { and } \quad \frac{\partial E_{n}}{\partial x_{i}}=e^{-\sum_{k=i+1}^{n} \alpha_{k}} \quad \text { if } i<n .
$$

Using the fact that $G^{\prime}=F^{-1}$, we therefore get

$$
\begin{aligned}
\frac{\partial C^{V}}{\partial x_{i}}= & F^{-1}\left(E_{i}+x_{i}\right)+\sum_{n=i+1}^{N}\left[F^{-1}\left(E_{n}+x_{n}\right)-F^{-1}\left(E_{n}\right)\right] e^{-\sum_{k=i+1}^{n} \alpha_{k}} \\
= & F^{-1}\left(E_{i}+x_{i}\right)-e^{-\alpha_{i+1}} F^{-1}\left(E_{i+1}\right) \\
& +e^{-\alpha_{i+1}}\left(F^{-1}\left(E_{i+1}+x_{i+1}\right)+\sum_{n=i+2}^{N}\left[F^{-1}\left(E_{n}+x_{n}\right)-F^{-1}\left(E_{n}\right)\right] e^{-\sum_{k=i+2}^{n} \alpha_{k}}\right),
\end{aligned}
$$

which yields (43).
For the proof of (44), we first have to compute $\partial E_{n} / \partial \alpha_{i}$. We obtain

$$
\frac{\partial E_{n}}{\partial \alpha_{i}}=0 \quad \text { if } i>n \quad \text { and } \quad \frac{\partial E_{n}}{\partial \alpha_{i}}=-\sum_{m=0}^{i-1} x_{m} e^{-\sum_{k=m+1}^{n} \alpha_{k}} \quad \text { for } i \leq n
$$

From here, we get

$$
\begin{aligned}
\frac{\partial C^{V}}{\partial \alpha_{i}} & =-\sum_{n=i}^{N}\left[F^{-1}\left(E_{n}+x_{n}\right)-F^{-1}\left(E_{n}\right)\right] \sum_{m=0}^{i-1} x_{m} e^{-\sum_{k=m+1}^{n} \alpha_{k}} \\
& =E_{i} \sum_{n=i}^{N}\left[F^{-1}\left(E_{n}+x_{n}\right)-F^{-1}\left(E_{n}\right)\right] e^{-\sum_{k=i+1}^{n} \alpha_{k}},
\end{aligned}
$$

which is (44).
Remark 3.5. A consequence of this lemma is that homogeneous time spacing $\boldsymbol{\alpha}^{*}$ and optimal strategy $\boldsymbol{\xi}^{V}$ given in [2] yield a critical point for the minimization in ( $\boldsymbol{x}, \boldsymbol{\alpha}$ ). Indeed, we then have $E_{i}=a^{*} \xi_{0}^{V}$ for any $i$, and therefore $\frac{\partial C^{V}}{\partial \alpha_{i}}$ does not depend on $i$.

Lemma 3.6. For each $\boldsymbol{\alpha} \in \mathcal{A}$, the function $C^{V}(\cdot, \boldsymbol{\alpha})$ has a minimizer $\boldsymbol{x}^{*}(\boldsymbol{\alpha}) \in \Xi$, which is unique up to equivalence.

Proof. First, note that we may assume without loss of generality that $\boldsymbol{\alpha} \in \mathcal{A}^{*}=\{\boldsymbol{\alpha} \in \mathcal{A} \mid$ $\left.\alpha_{i}>0, i=1, \ldots, N\right\}$. Indeed, if $\alpha_{i}=0$, we can merge the trades $x_{i-1}$ and $x_{i}$ into a single one and reduce $N$ to $N-1$.

We next extend the arguments in [2, Lemma B.1] to prove the existence of a unique minimizer of $C^{V}(\cdot, \boldsymbol{\alpha})$ in $\boldsymbol{\Xi}$.

Using the convention $\sum_{k=n+1}^{n} \alpha_{k}:=0$, we obtain by rearranging the sum in (42) that

$$
\begin{aligned}
C^{V}(\boldsymbol{x}, \boldsymbol{\alpha})= & G\left(\sum_{i=0}^{N} x_{i} e^{-\sum_{k=i+1}^{N} \alpha_{k}}\right)-G(0) \\
& +\sum_{n=0}^{N-1}\left[G\left(\sum_{i=0}^{n} x_{i} e^{-\sum_{k=i+1}^{n} \alpha_{k}}\right)-G\left(e^{-\alpha_{n+1}} \sum_{i=0}^{n} x_{i} e^{-\sum_{k=i+1}^{n} \alpha_{k}}\right)\right]
\end{aligned}
$$

Let us define the linear map $T: \mathbb{R}^{N+1} \rightarrow \mathbb{R}^{N+1}$ via

$$
(T \boldsymbol{x})_{n}=\sum_{i=0}^{n} x_{i} e^{-\sum_{k=i+1}^{n} \alpha_{k}}, \quad n=0, \ldots, N .
$$

We can thus write

$$
\begin{equation*}
C^{V}(\boldsymbol{x}, \boldsymbol{\alpha})=G\left((T \boldsymbol{x})_{N}\right)-G(0)+\sum_{n=0}^{N-1}\left[G\left((T \boldsymbol{x})_{n}\right)-G\left(e^{-\alpha_{n+1}}(T \boldsymbol{x})_{n}\right)\right] . \tag{45}
\end{equation*}
$$

First, note that $G$ is strictly convex, since $G^{\prime}=F^{-1}$ is strictly increasing. Second, for $a \in(0,1)$, the function $x \rightarrow G(x)-G(a x)$ is also strictly convex, because its derivative is equal to the strictly increasing function $h_{V, a}$ in Lemma 3.3. And third, $T$ is one-to-one. Hence, $C^{V}(\cdot, \boldsymbol{\alpha})$ is strictly convex in is first argument, and there can be at most one minimizer.

To show the existence of a minimizer, note that $G^{\prime}=F^{-1}$ is increasing with $F^{-1}(0)=0$, and hence $G(y)-G(a y) \geq(1-a)|y| \cdot\left|F^{-1}(a y)\right|$. Therefore, (45) yields

$$
\begin{aligned}
C^{V}(\boldsymbol{x}, \boldsymbol{\alpha}) \geq & G\left((T \boldsymbol{x})_{N}\right)-G(0) \\
& +\sum_{n=0}^{N-1}\left(1-e^{-\alpha_{n+1}}\right) \cdot\left|F^{-1}\left(e^{-\alpha_{n+1}}(T \boldsymbol{x})_{n}\right)\right| \cdot\left|(T \boldsymbol{x})_{n}\right| .
\end{aligned}
$$

Hence,

$$
C^{V}(\boldsymbol{x}, \boldsymbol{\alpha}) \geq \Lambda\left(|T \boldsymbol{x}|_{\infty}\right)-G(0)
$$

where $|\cdot|_{\infty}$ is the $\ell^{\infty}$-norm on $\mathbb{R}^{N+1}$ and $\Lambda$ is the function

$$
\Lambda(y):=G(y) \wedge G(-y) \wedge \min _{n=0, \ldots, N-1}\left\{|y| \cdot\left(1-a_{n+1}\right)\left(\left|F^{-1}\left(a_{n+1} \cdot y\right)\right| \wedge\left|F^{-1}\left(-a_{n+1} \cdot y\right)\right|\right)\right\}
$$

where $a_{n+1}:=e^{-\alpha_{n+1}}$. Since $F$ is unbounded, both $G(y)$ and $\left|F^{-1}(y)\right|$ tend to $+\infty$ for $|y| \rightarrow \infty$, and the fact that $T$ is one-to-one implies that $\Lambda\left(|T \boldsymbol{x}|_{\infty}\right) \rightarrow+\infty$ for $|\boldsymbol{x}| \rightarrow \infty$. Note also that by assumption $\alpha_{n}>0$ for each $n$. Hence, $C^{V}(\cdot, \boldsymbol{\alpha})$ must attain its minimum on $\Xi$.

We are now in a position to prove the main results for the Model with reversion of volume impact.

Proof of Proposition 2.10. The result for $X_{0}<0$ will follow if we can show that the minimizer in Lemma 3.6 consists only of strictly positive components. Here we may assume
without loss of generality that the admissible sequence of trading times is strictly increasing, or equivalently that $\boldsymbol{\alpha} \in \mathcal{A}^{*}$, for otherwise we can simply merge two trades occurring at the same time into a single trade.

If $\boldsymbol{x}$ is the minimizer of $C^{V}(\cdot, \boldsymbol{\alpha})$ on $\Xi$, then there must be a Lagrange multiplier $\nu$ such that $\boldsymbol{x}$ is a critical point of $\boldsymbol{y} \mapsto C^{V}(\boldsymbol{y}, \boldsymbol{\alpha})-\nu \sum_{i=0}^{N} y_{i}$. Hence, (43) yields that

$$
\begin{equation*}
\nu\left(1-a_{i+1}\right)=F^{-1}\left(E_{i}+x_{i}\right)-a_{i+1} F^{-1}\left(E_{i+1}\right)=h_{V, a_{i+1}}\left(E_{i}+x_{i}\right), \quad i=0, \ldots, N-1 \tag{46}
\end{equation*}
$$

where $a_{i+1}=e^{-\alpha_{i+1}}$ and $h_{V, a}$ is as in Lemma 3.3. For the final trade, we have

$$
\begin{equation*}
\nu=F^{-1}\left(E_{N}+x_{N}\right) \tag{47}
\end{equation*}
$$

Since $h_{V, a}(0)=0=F^{-1}(0)$ and both $h_{V, a}$ and $F^{-1}$ are strictly increasing, we conclude that $E_{0}+x_{0}, \ldots, E_{N}+x_{N}$ all have the same sign as $\nu$. Thus, $\nu>0$ by Lemma 3.1. Next, (46) implies that $E_{i}+x_{i}=h_{V, a_{i+1}}^{-1}\left(\nu\left(1-a_{i+1}\right)\right)$ and hence $E_{i+1}=a_{i+1} h_{V, a_{i+1}}^{-1}\left(\nu\left(1-a_{i+1}\right)\right)$. Using (46) once again yields
$x_{0}=h_{V, a_{1}}^{-1}\left(\nu\left(1-a_{1}\right)\right) \quad$ and $\quad x_{i}=h_{V, a_{i+1}}^{-1}\left(\nu\left(1-a_{i+1}\right)\right)-a_{i} h_{V, a_{i}}^{-1}\left(\nu\left(1-a_{i}\right)\right), \quad i=1, \ldots, N-1$.
The inequality (35) thus gives $x_{i}>0$ for $i=0, \ldots, N-1$. As for the final trade, (47) gives $x_{N}=F(\nu)-a_{N} h_{V, a_{N}}^{-1}\left(\nu\left(1-a_{N}\right)\right)$, which is strictly positive by (36).

Now we consider the case $X_{0}=0$. Suppose that $(\boldsymbol{x}, \boldsymbol{\alpha})$ is a round trip such that $C^{V}(\boldsymbol{x}, \boldsymbol{\alpha}) \leq 0$. Again we can assume without loss of generality that $\boldsymbol{\alpha} \in \mathcal{A}^{*}$. Then

$$
C^{V}(\boldsymbol{x}, \boldsymbol{\alpha})=\lim _{\varepsilon \downarrow 0} C^{V}(\boldsymbol{x}+\varepsilon \mathbf{1}, \boldsymbol{\alpha})
$$

But $C^{V}(\boldsymbol{x}+\varepsilon \mathbf{1}, \boldsymbol{\alpha})>0$ for each $\varepsilon>0$, due to our previous results. Hence, we must have $C^{V}(\boldsymbol{x}, \boldsymbol{\alpha})=0$. The strict convexity of $C^{V}(\cdot, \boldsymbol{\alpha})$, established in the proof of Lemma 3.6, implies that there can be at most one minimizer of $C^{V}(\cdot, \boldsymbol{\alpha})$ in the class of round trips. Since we clearly have $C^{V}(\mathbf{0}, \boldsymbol{\alpha})=0$, we must conclude that $\boldsymbol{x}=\mathbf{0}$.

Proof of Theorem 2.11. We will show that for $X_{0}<0$ the admissible strategy $\left(\boldsymbol{\xi}^{V}, \boldsymbol{\alpha}^{*}\right)$, defined via (14), (15), and (34), is the unique minimizer of $C^{V}$ on $\Xi \times \mathcal{A}$. The first step is to show the existence of a minimizer. To this end, note that Proposition 2.10 allows us to restrict the minimization of $C^{V}$ to $\Xi_{+} \times \mathcal{A}$, where $\Xi_{+}=\left\{\boldsymbol{x} \in \Xi \mid x_{i} \geq 0, i=0, \ldots, N\right\}$. The set $\Xi_{+} \times \mathcal{A}$ is in fact the product of two compact simplices, and so the continuity of $C^{V}$ yields the existence of a global minimizer, which lies in $\Xi_{+} \times \mathcal{A}$.

We next argue that any minimizer must belong to the relative interior of $\Xi_{+} \times \mathcal{A}$. To this end, suppose that $\boldsymbol{x} \in \Xi_{+}$and $\boldsymbol{\alpha} \in \mathcal{A}$ are given such that $\alpha_{i}=0$ for some $i$. We then define $\overline{\boldsymbol{\alpha}}:=\left(\alpha_{0}, \ldots, \alpha_{i-1}, \alpha_{i+1} / 2, \alpha_{i+1} / 2, \alpha_{i+2}, \ldots, \alpha_{N}\right)$ and $\overline{\boldsymbol{x}}:=\left(x_{0}, \ldots, x_{i-2}, x_{i-1}+x_{i}, 0\right.$, $\left.x_{i+1}, \ldots, x_{N}\right)$ and observe that $C^{V}(\overline{\boldsymbol{x}}, \overline{\boldsymbol{\alpha}})=C^{V}(\boldsymbol{x}, \boldsymbol{\alpha})$. But Proposition 2.10 implies that $\overline{\boldsymbol{x}}$ cannot be optimal for $\overline{\boldsymbol{\alpha}}$, since $\bar{x}_{i}=0$. In particular, $(\boldsymbol{x}, \boldsymbol{\alpha})$ cannot be optimal. Thus, the $\boldsymbol{\alpha}$-component of any minimizer must lie in the relative interior of $\mathcal{A}$. Finally, for $\boldsymbol{\alpha}$ in the relative interior of $\mathcal{A}$, Proposition 2.10 states that $\overline{\boldsymbol{x}}^{*}(\boldsymbol{\alpha})$ belongs to the relative interior of $\Xi_{+}$.

Now suppose that $(\boldsymbol{x}, \boldsymbol{\alpha})$ is a minimizer of $C^{V}$. Due to the preceding step, there must be Lagrange multipliers $\nu$ and $\lambda$ such that $(\boldsymbol{x}, \boldsymbol{\alpha})$ is a critical point of $(\boldsymbol{y}, \boldsymbol{\beta}) \mapsto C^{V}(\boldsymbol{y}, \boldsymbol{\beta})-$ $\nu \sum_{i=0}^{N} y_{i}-\lambda \sum_{j=1}^{N} \beta_{j}$. The identity (43) thus again yields

$$
\begin{equation*}
\nu\left(1-e^{-\alpha_{i+1}}\right)=F^{-1}\left(E_{i}+x_{i}\right)-e^{-\alpha_{i+1}} F^{-1}\left(E_{i+1}\right), \quad i=0, \ldots, N-1, \tag{48}
\end{equation*}
$$

and

$$
\begin{equation*}
\nu=F^{-1}\left(E_{N}+x_{N}\right) \tag{49}
\end{equation*}
$$

Using the same argument as in the proof of Proposition 2.10, we have $\nu>0$. Note that this can also be obtained by writing

$$
-X_{0}=\sum_{i=0}^{N} x_{i}=F(\nu)+\sum_{i=1}^{N}\left(1-a_{i}\right) h_{1, a_{i}}^{-1}\left(\nu\left(1-a_{i}\right)\right)
$$

Indeed, the right-hand side is strictly increasing in $\nu\left(F\right.$ and the functions $h_{1, a_{i}}^{-1}$ are strictly increasing) and vanishes for $\nu=0$, so $\nu>0$.

Next, (44) gives

$$
\begin{equation*}
\lambda=E_{j} \sum_{n=j}^{N}\left[F^{-1}\left(E_{n}+x_{n}\right)-F^{-1}\left(E_{n}\right)\right] e^{-\sum_{k=j+1}^{n} \alpha_{k}}, \quad j=1, \ldots, N \tag{50}
\end{equation*}
$$

We now rewrite the sum in (50) as follows:

$$
\begin{aligned}
& \sum_{n=j}^{N}\left[F^{-1}\left(E_{n}+x_{n}\right)-F^{-1}\left(E_{n}\right)\right] e^{-\sum_{k=j+1}^{n} \alpha_{k}} \\
& =-F^{-1}\left(E_{j}\right) \\
& \quad+\left[F^{-1}\left(E_{j}+x_{j}\right)-F^{-1}\left(E_{j+1}\right) e^{-\alpha_{j+1}}\right]+\cdots \\
& \quad+\left[F^{-1}\left(E_{N-1}+x_{N-1}\right)-F^{-1}\left(E_{N}\right) e^{-\alpha_{N}}\right] e^{-\sum_{k=j+1}^{N-1} \alpha_{k}} \\
& \quad+F^{-1}\left(E_{N}+x_{N}\right) e^{-\sum_{k=j+1}^{N} \alpha_{k}}
\end{aligned}
$$

Plugging in (48) and (49), simplifications occur and we get

$$
\sum_{n=j}^{N}\left[F^{-1}\left(E_{n}+x_{n}\right)-F^{-1}\left(E_{n}\right)\right] e^{-\sum_{k=j+1}^{n} \alpha_{k}}=\nu-F^{-1}\left(E_{j}\right)
$$

Plugging this back into (50) yields $\lambda=\left(\nu-F^{-1}\left(E_{j}\right)\right) E_{j}$ for $j=1, \ldots, N$. Solving this equation together with (48) for $\nu$ and $\lambda$ implies that necessarily

$$
\begin{aligned}
& \nu=\frac{F^{-1}\left(E_{i-1}+x_{i-1}\right)-e^{-\alpha_{i}} F^{-1}\left(e^{-\alpha_{i}}\left(E_{i-1}+x_{i-1}\right)\right)}{1-e^{-\alpha_{i}}} \\
& \lambda=e^{-\alpha_{i}}\left(E_{i-1}+x_{i-1}\right) \frac{F^{-1}\left(E_{i-1}+x_{i-1}\right)-F^{-1}\left(e^{-\alpha_{i}}\left(E_{i-1}+x_{i-1}\right)\right)}{1-e^{-\alpha_{i}}}
\end{aligned}
$$

for $i=1, \ldots, N$. Lemma 3.3(c) thus implies that

$$
\alpha_{1}=\cdots=\alpha_{N} \quad \text { and } \quad x_{0}=E_{1}+x_{1}=\cdots=E_{N-1}+x_{N-1}
$$

This gives $\boldsymbol{\alpha}=\boldsymbol{\alpha}^{*}$. Moreover, (15) holds, since $x_{i}=\left(1-a^{*}\right) x_{0}$ for $i=1, \ldots, N-1$. We also get $E_{i}=a^{*} x_{0}$ for $i=1, \ldots, N$. Note next that $x_{N}=-X_{0}-x_{0}-(N-1)\left(1-a^{*}\right) x_{0}$, and therefore $E_{N}+x_{N}=-X_{0}-N\left(1-a^{*}\right) x_{0}$. Equation (14) now follows from the fact that

$$
F^{-1}\left(-X_{0}-N\left(1-a^{*}\right)\right)=F^{-1}\left(E_{N}+x_{N}\right)=\frac{\partial C^{V}}{\partial x_{N}}(\boldsymbol{x}, \boldsymbol{\alpha})=\nu=\frac{F^{-1}\left(x_{0}\right)-a^{*} F^{-1}\left(a^{*} x_{0}\right)}{1-a^{*}} .
$$

This concludes the proof of the theorem.
Proof of Corollary 2.12. The result follows immediately from Theorem 2.11.
3.3. Proofs for reversion of price impact. The general strategy of the proof is similar to the one in section 3.2, although there are also some differences. We start with two lemmas on properties of the functions satisfying Assumption 2.15. Their conclusions are more important than Assumption 2.15 itself, as the validity of the conclusions of Lemmas 3.7 and 3.8 will imply the validity of Theorem 2.17.

Lemma 3.9 provides recursive identities for the gradient of our cost functional. These identities are needed to derive equations for critical points of the constraint optimization problem. The existence of such critical points is guaranteed by Lemma 3.11. Uniqueness, however, must be proved by another method as in section 3.2, because the cost functional is no longer convex for price impact reversion.

Again, it will be enough to prove our results for $X_{0}<0$. The case for $X_{0}>0$ will then follow by the symmetry of the problem formulation. The case of round trips with $X_{0}=0$ will be analyzed by a limiting procedure.

Lemma 3.7. Under Assumption 2.15, the following conclusions hold:
(a) $x \mapsto x f(x)$ is increasing on $\mathbb{R}$ (or, equivalently, $\tilde{F}$ is convex).
(b) For all $a \in(0,1), x \mapsto a f(a x) / f(x)$ is nondecreasing on $\mathbb{R}_{+}$and nonincreasing on $\mathbb{R}_{-}$ and takes values in $(0,1)$.
(c) For all $x>0,(0,1) \ni a \longmapsto \frac{1-a^{2} f(a x) / f(x)}{1-a f(a x) / f(x)}$ is increasing.
(d) For all $x>0,(0,1) \ni a \longmapsto a^{-1} \frac{1-a^{2} f(x) / f(x / a)}{1-a f(x) / f(x / a)}$ is decreasing.

Proof of Lemma 3.7 under Assumption 2.15(a). (a) The derivative is positive, since $x f^{\prime}(x) / f(x)>-1$ by Assumption 2.15(a).
(b) Since $x \mapsto x f(x)$ is increasing, $a f(a x) / f(x)=[a x f(a x)] /[x f(x)] \in(0,1)$. The derivative of $x \mapsto a f(a x) / f(x)$ is equal to $\left[a^{2} f^{\prime}(a x) f(x)-a f(a x) f^{\prime}(x)\right] / f(x)^{2}$. It is nonnegative on $\mathbb{R}_{+}$and nonpositive on $\mathbb{R}_{-}$if and only if

$$
\frac{a f^{\prime}(a x)}{f(a x)} \geq \frac{f^{\prime}(x)}{f(x)} \quad \text { for } x \geq 0 \quad \text { and } \quad \frac{a f^{\prime}(a x)}{f(a x)} \leq \frac{f^{\prime}(x)}{f(x)} \quad \text { for } x \leq 0
$$

These conditions hold as a direct consequence of (20).
(c) For a fixed $x \geq 0$, we set $\psi(a)=a f(a x) / f(x)$, which takes values in $(0,1)$. We need to show that

$$
\frac{d}{d a} \frac{1-a \psi(a)}{1-\psi(a)}=\frac{(1-a) \psi^{\prime}(a)-\psi(a)(1-\psi(a))}{(1-\psi(a))^{2}}>0
$$

This condition holds if and only if

$$
\frac{\psi^{\prime}(a)}{\psi(a)}>\frac{1-\psi(a)}{1-a}
$$

It is thus sufficient to show that $\psi^{\prime} / \psi$ is nonincreasing, since then we would have

$$
1-\psi(a)<\int_{a}^{1} \frac{\psi^{\prime}(u)}{\psi(u)} d u \leq(1-a) \frac{\psi^{\prime}(a)}{\psi(a)}
$$

This leads to requiring $\psi \psi^{\prime \prime}-\left(\psi^{\prime}\right)^{2} \leq 0$, which in turn leads to the following condition:

$$
1+(a x)^{2}\left(\frac{f^{\prime}(a x)}{f(a x)}\right)^{2}-(a x)^{2} \frac{f^{\prime \prime}(a x)}{f(a x)} \geq 0 \quad \text { for } a \in(0,1)
$$

The latter condition is ensured by assumption (21), since $x f^{\prime}(x) / f(x) \in(-1,0]$ and thus

$$
\left(\frac{x f^{\prime}(x)}{f(x)}\right)^{2}+\frac{x f^{\prime}(x)}{f(x)}<0
$$

(d) We fix $x>0$ and let $\tilde{\psi}(a):=a f(x) / f(x / a)$. We need to show that

$$
\frac{d}{d a} a^{-1} \frac{1-a \tilde{\psi}(a)}{1-\tilde{\psi}(a)}=\frac{\tilde{\psi}(a)-1+a \tilde{\psi}^{\prime}(a)(1-a)}{a^{2}(1-\tilde{\psi}(a))^{2}}<0
$$

This condition holds if and only if

$$
a \tilde{\psi}^{\prime}(a)<\frac{1-\tilde{\psi}(a)}{1-a}
$$

Hence, it is enough to show that $a \mapsto a \tilde{\psi}^{\prime}(a)$ is nondecreasing, because then we would have

$$
1-\tilde{\psi}(a)>\int_{a}^{1} u \tilde{\psi}^{\prime}(u) d u \geq(1-a) a \tilde{\psi}^{\prime}(a)
$$

Some calculations lead to

$$
\frac{d}{d a} a \tilde{\psi}^{\prime}(a)=\frac{1}{f(x / a)}\left(1+\frac{x}{a} \frac{f^{\prime}(x / a)}{f(x / a)}+2\left(\frac{x}{a} \frac{f^{\prime}(x / a)}{f(x / a)}\right)^{2}-\left(\frac{x}{a}\right)^{2} \frac{f^{\prime \prime}(x / a)}{f(x / a)}\right)
$$

which is nonnegative by assumption (21).
Proof of Lemma 3.7 under Assumption 2.15(b). Points (a) and (b) are trivial. To check (c) and (d), we have to show that

$$
a \in(0,1), \quad a \mapsto \frac{1-a^{2+\alpha}}{1-a^{1+\alpha}}, \quad \text { and } \quad a \mapsto \frac{a-a^{2+\alpha}}{1-a^{2+\alpha}}=1-\frac{1-a}{1-a^{2+\alpha}}
$$

are increasing. It is, however, easy to check by derivation that $a \mapsto \frac{1-a^{\gamma}}{1-a^{\beta}}$ is increasing on $(0,1)$ when $0<\beta<\gamma$, which gives the result.

Lemma 3.8. Under Assumption 2.15, the following conclusions hold:
(a) For each $a \in(0,1)$, the function $h_{P, a}(x)=x \frac{f(x / a) / a-a f(x)}{f(x / a)-a f(x)}$ is well defined for $x \in \mathbb{R}$ and is strictly increasing.
(b) For all $a, b \in(0,1)$ and $\nu>0$, we have the inequalities

$$
h_{P, a}^{-1}(\nu) / a>h_{P, b}^{-1}(\nu) \quad \text { and } \quad h_{P, b}^{-1}(\nu)<\nu .
$$

(c) The function $H_{P}:(x, a) \in(0, \infty) \times(0,1) \mapsto\left(x \frac{f(x / a) / a-a f(x)}{f(x / a)-a f(x)},-x^{2} f(x) \frac{f(x / a)(1 / a-1)}{f(x / a)-a f(x)}\right)$ is one-to-one.
Proof. (a) First, let us observe that the denominator of $h_{P, a}$ is positive, since $x \mapsto x f(x)$ is increasing by Lemma 3.7(a). We have

$$
\begin{equation*}
h_{P, a}(x)=x\left(1+\frac{a^{-1}-1}{1-a f(x) / f(x / a)}\right) . \tag{51}
\end{equation*}
$$

Again by Lemma 3.7, the fraction is positive and, as a function of $x$, nondecreasing on $\mathbb{R}_{+}$ and nonincreasing $\mathbb{R}_{-}$, which gives the result.
(b) It is clear from (51) that $h_{P, a}(x)>x$ for $x>0$, and therefore $h_{P, a}^{-1}(x)<x$. Let us now consider $a, b \in(0,1), \nu>0$ and set $x^{\prime}=h_{P, a}^{-1}(\nu) / a, x=h_{P, b}^{-1}(\nu)$. Then both $x$ and $x^{\prime}$ are positive, and we need to show that $x^{\prime}>x$. It follows that

$$
\nu=x^{\prime} \frac{f\left(x^{\prime}\right)-a^{2} f\left(a x^{\prime}\right)}{f\left(x^{\prime}\right)-a f\left(a x^{\prime}\right)}=x \frac{f(x / b) / b-b f(x)}{f(x / b)-b f(x)} .
$$

Let us suppose by way of contradiction that $x^{\prime} \leq x$. Then, using Lemma 3.7 (b) and the fact that $u \in[0,1) \mapsto(1-a u) /(1-u)$ is increasing, we get

$$
\frac{1-a^{2} f(a x) / f(x)}{1-a f(a x) / f(x)} \geq \frac{1-a^{2} f\left(a x^{\prime}\right) / f\left(x^{\prime}\right)}{1-a f\left(a x^{\prime}\right) / f\left(x^{\prime}\right)} \geq b^{-1} \frac{1-b^{2} f(x) / f(x / b)}{1-b f(x) / f(x / b)} .
$$

Again by Lemma 3.7, the left-hand side is increasing with respect to $a$ and the right-hand side is decreasing with respect to $b$. Moreover, both have the same limit,

$$
\frac{2+x f^{\prime}(x) / f(x)}{1+x f^{\prime}(x) / f(x)},
$$

when $a \uparrow 1$ and $b \uparrow 1$, which leads to a contradiction.
(c) Let $\left(a_{1}, y_{1}\right),\left(a_{2}, y_{2}\right) \in(0,1) \times(0, \infty)$ be such that $H_{P}\left(a_{1}, y_{1}\right)=H_{P}\left(a_{2}, y_{2}\right)$. By (51), we then have

$$
\left\{\begin{array}{l}
y_{1}\left(1+\gamma_{1}\right)=y_{2}\left(1+\gamma_{2}\right),  \tag{52}\\
y_{1}^{2} f\left(y_{1}\right) \gamma_{1}=y_{2}^{2} f\left(y_{2}\right) \gamma_{2},
\end{array} \quad \text { where } \gamma_{i}:=\frac{\left(a_{i}^{-1}-1\right) f\left(y_{i} / a_{i}\right)}{f\left(y_{i} / a_{i}\right)-a_{i} f\left(y_{i}\right)} \text { for } i=1,2\right.
$$

Let us assume, for example, that $\gamma_{2} \leq \gamma_{1}$ and set $\eta=\gamma_{2} / \gamma_{1} \in(0,1]$. Eliminating $y_{1}$ in (52) yields

$$
\phi(\eta):=\left(\frac{1+\eta \gamma_{1}}{1+\gamma_{1}}\right)^{2} f\left(y_{2} \frac{1+\eta \gamma_{1}}{1+\gamma_{1}}\right)-\eta f\left(y_{2}\right)=0 .
$$

Since $x \mapsto x f(x)$ is increasing by Lemma 3.7(a), we have

$$
\eta \in(0,1), \phi(\eta)<\frac{1-\eta}{1+\gamma_{1}} f\left(y_{2}\right)<0 .
$$

Thus, $\eta=1$ is the only zero of $\phi(\eta)$. We may thus conclude that $\gamma_{1}=\gamma_{2}$ and in turn that $y_{1}=y_{2}$. Finally, the equality $\gamma_{1}=\gamma_{2}$ leads to $a_{1}=a_{2}$ due to Lemma 3.7(d), since

$$
1+\gamma_{i}=a_{i}^{-1} \frac{1-a_{i}^{2} f\left(y_{i}\right) / f\left(y_{i} / a_{i}\right)}{1-a_{i} f\left(y_{i}\right) / f\left(y_{i} / a_{i}\right)}
$$

In the Model with price impact reversion, we need to minimize the following cost functional:

$$
\begin{equation*}
C^{P}\left(x_{0}, \ldots, x_{n}, \boldsymbol{\alpha}\right)=\sum_{n=0}^{N} G\left(F\left(D_{n}\right)+x_{n}\right)-G\left(F\left(D_{n}\right)\right), \tag{53}
\end{equation*}
$$

where $D_{0}=0$ and $D_{n}=e^{-\alpha_{n}} F^{-1}\left(x_{n-1}+F\left(D_{n-1}\right)\right)$ for $1 \leq n \leq N$. By $\widehat{f}(x)=1 / f\left(F^{-1}(x)\right)$, we again denote the derivative of $F^{-1}$.

Lemma 3.9. We have the following recursive formula for $i=0, \ldots, N-1$ :

$$
\begin{equation*}
\frac{\partial C^{P}}{\partial x_{i}}=F^{-1}\left(F\left(D_{i}\right)+x_{i}\right)+e^{-\alpha_{i+1}} f\left(D_{i+1}\right) \widehat{f}\left(x_{i}+F\left(D_{i}\right)\right)\left[\frac{\partial C^{P}}{\partial x_{i+1}}-D_{i+1}\right] \tag{54}
\end{equation*}
$$

Moreover, for $j=1, \ldots, N$,

$$
\begin{equation*}
\frac{\partial C^{P}}{\partial \alpha_{j}}=-D_{j} f\left(D_{j}\right)\left(\frac{\partial C^{P}}{\partial x_{j}}-D_{j}\right) \tag{55}
\end{equation*}
$$

Proof. We have $D_{1}=e^{-\alpha_{1}} F^{-1}\left(x_{0}\right)$ and $D_{n}=e^{-\alpha_{n}} F^{-1}\left(x_{n-1}+F\left(D_{n-1}\right)\right)$ for $1 \leq n \leq N$. Thus, we obtain the following recursive relations between the derivatives of $D_{n}$ with respect to $x_{i}$ :

$$
\begin{aligned}
& \frac{\partial D_{n}}{\partial x_{i}}=0 \quad \text { for } i \geq n, \quad \frac{\partial D_{n}}{\partial x_{n-1}}=e^{-\alpha_{n}} \widehat{f}\left(x_{n-1}+F\left(D_{n-1}\right)\right), \\
& \frac{\partial D_{n}}{\partial x_{i}}=e^{-\alpha_{i+1}} f\left(D_{i+1}\right) \widehat{f}\left(x_{i}+F\left(D_{i}\right)\right) \frac{\partial D_{n}}{\partial x_{i+1}} \quad \text { for } 1 \leq i \leq n-1 .
\end{aligned}
$$

Thus, by (53),

$$
\begin{align*}
\frac{\partial C^{P}}{\partial x_{i}}= & F^{-1}\left(x_{i}+F\left(D_{i}\right)\right)+\sum_{n=i+1}^{N} f\left(D_{n}\right)\left[F^{-1}\left(x_{n}+F\left(D_{n}\right)\right)-D_{n}\right] \frac{\partial D_{n}}{\partial x_{i}}  \tag{56}\\
= & F^{-1}\left(x_{i}+F\left(D_{i}\right)\right)+e^{-\alpha_{i+1}} f\left(D_{i+1}\right) \widehat{f}\left(x_{i}+F\left(D_{i}\right)\right)\left[F^{-1}\left(x_{i+1}+F\left(D_{i+1}\right)\right)-D_{i+1}\right] \\
& +e^{-\alpha_{i+1}} f\left(D_{i+1}\right) \widehat{f}\left(x_{i}+F\left(D_{i}\right)\right) \sum_{n=i+2}^{N} f\left(D_{n}\right)\left[F^{-1}\left(x_{n}+F\left(D_{n}\right)\right)-D_{n}\right] \frac{\partial D_{n}}{\partial x_{i+1}} .
\end{align*}
$$

By (56), the sum in the preceding line satisfies

$$
\sum_{n=i+2}^{N} f\left(D_{n}\right)\left[F^{-1}\left(x_{n}+F\left(D_{n}\right)\right)-D_{n}\right] \frac{\partial D_{n}}{\partial x_{i+1}}=\frac{\partial C^{P}}{\partial x_{i+1}}-F^{-1}\left(x_{i+1}+F\left(D_{i+1}\right)\right)
$$

Hence,

$$
\frac{\partial C^{P}}{\partial x_{i}}=F^{-1}\left(x_{i}+F\left(D_{i}\right)\right)+e^{-\alpha_{i+1}} f\left(D_{i+1}\right) \widehat{f}\left(x_{i}+F\left(D_{i}\right)\right)\left[\frac{\partial C^{P}}{\partial x_{i+1}}-D_{i+1}\right],
$$

which is formula (54).
As to (55), we again use the recursive scheme at the beginning of this proof to obtain formulas for the derivatives of $D_{n}$ with respect to $\alpha_{j}$ :

$$
\begin{array}{ll}
\frac{\partial D_{n}}{\partial \alpha_{j}}=0 \quad \text { for } j>n, & \frac{\partial D_{n}}{\partial \alpha_{n}}=-D_{n}, \\
\frac{\partial D_{n}}{\partial \alpha_{j}}=-D_{j} f\left(D_{j}\right) \frac{\partial D_{n}}{\partial x_{j}} & \text { for } 1 \leq j \leq n-1 .
\end{array}
$$

We therefore obtain from (53)

$$
\begin{aligned}
\frac{\partial C^{P}}{\partial \alpha_{i}}= & \sum_{n=i}^{N} f\left(D_{n}\right)\left[F^{-1}\left(x_{n}+F\left(D_{n}\right)\right)-D_{n}\right] \frac{\partial D_{n}}{\partial \alpha_{i}} \\
= & -D_{i} f\left(D_{i}\right)\left[F^{-1}\left(x_{i}+F\left(D_{i}\right)\right)-D_{i}\right] \\
& -\sum_{n=i+1}^{N} f\left(D_{n}\right) D_{i} f\left(D_{i}\right)\left[F^{-1}\left(x_{n}+F\left(D_{n}\right)\right)-D_{n}\right] \frac{\partial D_{n}}{\partial x_{i}} \\
= & -D_{i} f\left(D_{i}\right)\left(F^{-1}\left(x_{i}+F\left(D_{i}\right)\right)-D_{i}+\frac{\partial C^{P}}{\partial x_{i}}-F^{-1}\left(x_{i}+F\left(D_{i}\right)\right)\right) \\
= & -D_{i} f\left(D_{i}\right)\left(\frac{\partial C^{P}}{\partial x_{i}}-D_{i}\right) .
\end{aligned}
$$

Remark 3.10. A consequence of this lemma is that the optimal strategy given by [2] on the homogeneous time spacing grid $\mathcal{T}^{*}$ is a critical point for the minimization in $(\boldsymbol{x}, \boldsymbol{\alpha})$. Indeed, we then have $D_{i}=a^{*} F^{-1}\left(\xi_{0}^{P}\right)$ for any $i$, and therefore $\frac{\partial C^{P}}{\partial \alpha_{i}}$ does not depend on $i$.

Lemma 3.11. Assume that $\boldsymbol{\alpha} \in \mathcal{A}^{*}$. Then $C^{P}(\boldsymbol{x}, \boldsymbol{\alpha}) \rightarrow \infty$ as $|\boldsymbol{x}| \rightarrow \infty$ under Assumption 2.15.

Proof. Equation (53) yields

$$
\begin{aligned}
& C^{P}(\boldsymbol{x}, \boldsymbol{\alpha}) \\
& =\sum_{n=0}^{N} \tilde{F}\left(F^{-1}\left(F\left(D_{n}\right)+x_{n}\right)\right)-\tilde{F}\left(D_{n}\right) \\
& =\sum_{n=0}^{N-1}\left[\tilde{F}\left(F^{-1}\left(F\left(D_{n}\right)+x_{n}\right)\right)-\tilde{F}\left(e^{-\alpha_{n+1}} F^{-1}\left(F\left(D_{n}\right)+x_{n}\right)\right)\right]+\tilde{F}\left(F^{-1}\left(F\left(D_{N}\right)+x_{N}\right)\right) .
\end{aligned}
$$

Let $\bar{a}=\max _{i=1, \ldots, N} e^{-\alpha_{i}}<1$. Since $x \mapsto x f(x)$ is increasing on $\mathbb{R}$, we have for $x \in \mathbb{R}, a \in[0, \bar{a}]$

$$
\tilde{F}(x)-\tilde{F}(a x)=\int_{a x}^{x} y f(y) d y \geq \int_{\bar{a} x}^{x} y f(y) d y \geq \bar{a}(1-\bar{a}) x^{2} f(\bar{a} x)=: H(x) .
$$

Defining $T_{2}(\boldsymbol{x})=\left(x_{0}, x_{1}+F^{-1}\left(D_{1}\right), \ldots, x_{N}+F^{-1}\left(D_{N}\right)\right)$, we thus get

$$
C^{P}(\boldsymbol{x}, \boldsymbol{\alpha}) \geq H\left(\left|T_{2}(\boldsymbol{x})\right|_{\infty}\right)
$$

From (20), $x \mapsto x f(x)$ is increasing, and therefore $H(x) \rightarrow+\infty$ as $|x| \rightarrow+\infty$. It is therefore sufficient to have $T_{2}(\boldsymbol{x}) \rightarrow+\infty$ for $|\boldsymbol{x}| \rightarrow+\infty$. To this end, let $\left(\boldsymbol{x}^{k}\right)$ be a sequence such that the sequence $\left(T_{2}\left(\boldsymbol{x}^{k}\right)\right)$ is bounded. We will show that $\left(\boldsymbol{x}^{k}\right)$ then must also be bounded. It is clear that the first coordinate $x_{0}^{k}$ is bounded. Therefore, $F^{-1}\left(D_{1}^{k}\right)$ is also bounded, which in turn implies that the second coordinate of $\left(T_{2}\left(\boldsymbol{x}^{k}\right)\right)$ is bounded. We then get that $\left(x_{1}^{k}\right)$ is bounded. An easy induction on coordinates thus gives the desired result.

We are now in position to prove the main results for the Model with price impact reversion.
Proof of Proposition 2.16. Let us first assume $X_{0}<0$. We can assume without loss of generality that $\boldsymbol{\alpha} \in \mathcal{A}^{*}$, for otherwise we can simply merge two trades occurring at the same time into a single trade. If $\boldsymbol{x}$ is the minimizer of $C^{P}(\cdot, \boldsymbol{\alpha})$ on $\Xi$, then there must be a Lagrange multiplier $\nu$ such that $\boldsymbol{x}$ is a critical point of $\boldsymbol{y} \mapsto C^{P}(\boldsymbol{y}, \boldsymbol{\alpha})-\nu \sum_{i=0}^{N} y_{i}$. Hence, (54) yields that

$$
\nu=h_{P, a_{i+1}}\left(D_{i+1}\right), \quad i=0, \ldots, N-1,
$$

where $a_{i+1}=e^{-\alpha_{i+1}}$ and $h_{P, a}$ is defined as in Lemma 3.8. Since $D_{i+1}=a_{i+1} F^{-1}\left(x_{i}+F\left(D_{i}\right)\right)$, we get with Lemma 3.8 that

$$
x_{0}=F\left(h_{P, a_{1}}^{-1}(\nu) / a_{1}\right), x_{i}=F\left(h_{P, a_{i+1}}^{-1}(\nu) / a_{i+1}\right)-F\left(h_{P, a_{i}}^{-1}(\nu)\right), \quad i=1, \ldots, N-1 .
$$

For the last trade, we also get that $\nu=F^{-1}\left(x_{N}+F\left(D_{N}\right)\right)$ and $x_{N}=F(\nu)-F\left(h_{P, a_{N}}^{-1}(\nu)\right)$. Therefore, summing all the trades, we get

$$
\begin{equation*}
-X_{0}=F(\nu)+\sum_{i=1}^{N}\left[F\left(h_{P, a_{i}}^{-1}(\nu) / a_{i}\right)-F\left(h_{P, a_{i}}^{-1}(\nu)\right)\right] . \tag{57}
\end{equation*}
$$

Now let us observe that $F$ is increasing on $\mathbb{R}$, and, for any $a \in(0,1), y \mapsto F(y / a)-F(y)$ is increasing (its derivative is positive by Lemma 3.7(a)). Besides, $F$ and $h_{P, a}^{-1}$ are increasing for any $a \in(0,1)$, and therefore $\nu$ is uniquely determined by the above equation. We have, moreover, $\nu>0$ because the left-hand side vanishes when $\nu$ is equal to 0 . This proves that there is a unique critical point, which then is necessarily the global minimum of $C^{P}$ by Lemma 3.11.

Next, $x_{i}>0$ for $i=0, \ldots, N$, due to Lemma 3.7 and the fact that $F$ is increasing.
Finally, we consider the case $X_{0}=0$. As in the proof of Proposition 2.10, we can show that a round trip such that $C^{P}(\boldsymbol{x}, \boldsymbol{\alpha}) \leq 0$ necessarily satisfies $C^{P}(\boldsymbol{x}, \boldsymbol{\alpha})=0$. Moreover, for $\boldsymbol{\alpha} \in \mathcal{A}^{*}$, we see looking at the proof of Proposition 2.16 that $(0, \ldots, 0)$ is the only critical point when $X_{0}=0$ since we necessarily have $\nu=0$ by (57). Therefore, it is also the unique minimum of $C^{P}$ by Lemma 3.11.

Proof of Theorem 2.17. The existence of a minimizer $\left(\boldsymbol{\xi}^{P}, \boldsymbol{\alpha}^{*}\right)$ and the fact that it belongs to $\Xi_{+} \times \mathcal{A}^{*}$ follow exactly as in the proof of Theorem 2.11.

Now suppose that $(\boldsymbol{x}, \boldsymbol{\alpha})$ is a minimizer of $C^{P}$ for $X_{0}<0$. Due to the preceding step, there must be Lagrange multipliers $\nu, \lambda \in \mathbb{R}$ such that $(\boldsymbol{x}, \boldsymbol{\alpha})$ is a critical point of $(\boldsymbol{y}, \boldsymbol{\beta}) \mapsto$ $C^{P}(\boldsymbol{y}, \boldsymbol{\beta})-\nu \sum_{i=0}^{N} y_{i}-\lambda \sum_{j=1}^{N} \beta_{j}$.

From (54), we easily obtain that for $i=1, \ldots, N$

$$
\nu=\frac{e^{-\alpha_{i}} f\left(D_{i}\right)}{f\left(e^{\alpha_{i}} D_{i}\right)}\left[\nu-D_{i}\right]+e^{\alpha_{i}} D_{i}
$$

and $\nu=F^{-1}\left(x_{N}+F\left(D_{N}\right)\right)$ for the last trade. We then deduce from (55) that

$$
\begin{aligned}
& \nu=D_{i} \frac{e^{\alpha_{i}} f\left(e^{\alpha_{i}} D_{i}\right)-e^{-\alpha_{i}} f\left(D_{i}\right)}{f\left(e^{\alpha_{i}} D_{i}\right)-e^{-\alpha_{i}} f\left(D_{i}\right)} \\
& \lambda=-D_{i}^{2} f\left(D_{i}\right) \frac{\left(e^{\alpha_{i}}-1\right) f\left(e^{\alpha_{i}} D_{i}\right)}{f\left(e^{\alpha_{i}} D_{i}\right)-e^{-\alpha_{i}} f\left(D_{i}\right)},
\end{aligned}
$$

i.e., $(\nu, \lambda)=H_{P}\left(D_{i}, a_{i}\right)$, with $a_{i}=e^{-\alpha_{i}}$. As in the proof of Proposition 2.16, we get (57), which (by our standing assumption $X_{0}<0$ ) ensures $\nu>0$ and in turn $D_{i}>0$ for $i=1, \ldots, N$. Due to Lemma 3.8, $H_{P}$ is one-to-one on $(0, \infty) \times(0,1)$, and therefore $\alpha_{1}=\cdots=\alpha_{N}$ and $D_{1}=\cdots=D_{N}$. Then $D_{1}=a^{*} F^{-1}\left(x_{0}\right)$. Since $D_{i+1}=a^{*} F^{-1}\left(x_{i}+F\left(D_{i}\right)\right)$, we get $x_{i}=$ $x_{0}-F\left(D_{i}\right)=x_{0}-F\left(a^{*} F^{-1}\left(x_{0}\right)\right)$, and therefore $x_{N}=-X_{0}-N x_{0}+(N-1) F\left(a^{*} F^{-1}\left(x_{0}\right)\right)$. Combining this with $\nu=F^{-1}\left(x_{N}+F\left(D_{N}\right)\right)$, we get

$$
F^{-1}\left(-X_{0}-N\left[x_{0}-F\left(a^{*} F^{-1}\left(x_{0}\right)\right)\right]\right)=h_{P, a^{*}}\left(F^{-1}\left(x_{0}\right)\right) .
$$

We refer the reader to [2, Lemma C.3] for the existence, uniqueness, and positivity of the solution $x_{0}$ of this equation. It follows that there is a unique critical point of $C^{P}$ on $\Xi_{+} \times \mathcal{A}^{*}$, which is necessarily the global minimum.

Proof of Corollary 2.18. The result follows immediately from Proposition 2.16 and Theorem 2.17.
4. Conclusion. We have introduced two variants of a market impact model in which price impact is a nonlinear function of volume impact and in which either volume or price impact reverts on an exponential scale. In both model variants, there are unique optimal strategies for the liquidation or acquisition of asset positions when optimality is defined in terms of the minimization of the expected liquidation costs. Existence and structure of these strategies allow us to conclude that our market impact model admits neither price manipulation in the sense of Huberman and Stanzl [14] nor transaction-triggered price manipulation in the sense of Alfonsi, Schied, and Slynko [3].

Our optimal execution strategies turn out to be deterministic, because we are minimizing the expected execution costs. As argued by Almgren and Chriss [6, 5], trade execution strategies used in practice should also take volatility risk into account, which may lead to adaptive strategies. We refer the reader to $[19,18]$. For future research, it would also be interesting to allow certain model parameters to be random.

Acknowledgment. The authors thank Jim Gatheral for discussions.

## REFERENCES

[1] A. Alfonsi, A. Fruth, and A. Schied, Constrained portfolio liquidation in a limit order book model, in Advances in Mathematics of Finance, Banach Center Publ. 83, Polish Acad. Sci. Inst. Math., Warsaw, 2008, pp. 9-25.
[2] A. Alfonsi, A. Fruth, and A. Schied, Optimal execution strategies in limit order books with general shape functions, Quant. Finance, 10 (2010), pp. 143-157.
[3] A. Alfonsi, A. Schied, and A. Slynko, Order Books Resilience, Price Manipulation, and the Positive Portfolio Problem, http://ssrn.com/abstract=1498514 (2009).
[4] R. Almgren, Optimal execution with nonlinear impact functions and trading-enhanced risk, Appl. Math. Finance, 10 (2003), pp. 1-18.
[5] R. Almgren and N. Chriss, Optimal execution of portfolio transactions, J. Risk, 3 (2000), pp. 5-39.
[6] R. Almgren and N. Chriss, Value under liquidation, Risk, 12 (1999), pp. 61-63.
[7] R. Almgren and J. Lorenz, Adaptive arrival price, in Algorithmic Trading III: Precision, Control, Execution, B. R. Bruce, ed., Institutional Investor, New York, 2007, pp. 59-66.
[8] R. Almgren, C. Thum, E. Hauptmann, and E. Li, Equity market impact, Risk, 18 (2005), pp. 57-62.
[9] D. Bertsimas and A. Lo, Optimal control of execution costs, J. Financial Markets, 1 (1998), pp. 1-50.
[10] B. Biais, P. Hillion, and C. Spatt, An empirical analysis of the limit order book and order flow in Paris Bourse, J. Finance, 50 (1995), pp. 1655-1689.
[11] J. P. Bouchaud, Y. Gefen, M. Potters, and M. Wyart, Fluctuations and response in financial markets: The subtle nature of "random" price changes, Quant. Finance, 4 (2004), pp. 176-190.
[12] J. Gatheral, No-dynamic-arbitrage and market impact, Quant. Finance, to appear.
[13] J. Gatheral, A. Schied, and A. Slynko, Transient Linear Price Impact and Fredholm Integral Equations, http://ssrn.com/abstract=1531466 (2010).
[14] G. Huberman and W. Stanzl, Price manipulation and quasi-arbitrage, Econometrica, 72 (2004), pp. 1247-1275.
[15] B. Mönch, Strategic Trading in Illiquid Markets, Springer, Berlin, 2005.
[16] A. Obizhaeva and J. Wang, Optimal Trading Strategy and Supply/Demand Dynamics, preprint, 2005.
[17] M. Potters and J.-P. Bouchaud, More statistical properties of order books and price impact, Phys. A, 324 (2003), pp. 133-140.
[18] A. Schied and T. Schöneborn, Risk aversion and the dynamics of optimal liquidation strategies in illiquid markets, Finance Stoch., 13 (2009), pp. 181-204.
[19] A. Schied, T. Schöneborn, and M. Tehranchi, Optimal basket liquidation for CARA investors is deterministic, Appl. Math. Finance, to appear.
[20] P. Weber and B. Rosenow, Order book approach to price impact, Quant. Finance, 5 (2005), pp. 357364.

# Term Structure Models Driven by Wiener Processes and Poisson Measures: Existence and Positivity* 

Damir Filipović ${ }^{\dagger}$, Stefan Tappe ${ }^{\ddagger}$, and Josef Teichmann ${ }^{\ddagger}$

Abstract. In the spirit of [T. Björk et al., Finance Stoch., 1 (1997), pp. 141-174], we investigate term structure models driven by Wiener processes and Poisson measures with forward curve dependent volatilities. This includes a full existence and uniqueness proof for the corresponding Heath-Jarrow-Mortontype term structure equation. Furthermore, we characterize positivity preserving models by means of the characteristic coefficients. A key role in our investigation is played by the method of the moving frame, which allows us to transform term structure equations to time-dependent SDEs.

Key words. term structure models driven by Wiener processes and Poisson measures, Heath-Jarrow-MortonMusiela equation, forward curve spaces, positivity preserving models

AMS subject classifications. 91B28, 60 H 15
DOI. 10.1137/090758593

1. Introduction. Interest rate theory deals with zero-coupon bonds, which are subject to a stochastic evolution due to daily trading of related products like coupon bearing bonds, swaps, caps, floors, swaptions, etc. Zero-coupon bonds, which are financial assets paying the holder one unit of cash at maturity time $T$, are conceptually important products, since one can easily write all other products as derivatives on them. We always assume default-free bonds; i.e., there are no counterparty risks in the considered markets. The Heath-Jarrow-Morton (HJM) methodology takes the bond market on the whole as today's aggregation of information on interest rates, and one tries to model future flows of information by a stochastic evolution equation on the set of possible scenarios of bond prices. For the set of possible scenarios of bond prices, the forward rate proved to be a flexible and useful parametrization, since it maps possible states of the bond market to open subsets of (Hilbert) spaces of forward rate curves. Under some regularity assumptions, the price of a zero-coupon bond at $t \leq T$ can be written as

$$
P(t, T)=\exp \left(-\int_{t}^{T} f(t, u) d u\right)
$$

where $f(t, T)$ is the forward rate for date $T$. We usually assume the forward rate to be continuous in maturity time $T$. The classical continuous framework for the evolution of the

[^45]forward rates goes back to Heath, Jarrow, and Morton [23]. They assume that, for every date $T$, the forward rates $f(t, T)$ follow an Itô process of the form
\[

$$
\begin{equation*}
d f(t, T)=\alpha(t, T) d t+\sum_{j=1}^{d} \sigma^{j}(t, T) d W_{t}^{j}, \quad t \in[0, T] \tag{1.1}
\end{equation*}
$$

\]

where $W=\left(W^{1}, \ldots, W^{d}\right)$ is a standard Brownian motion in $\mathbb{R}^{d}$.
There are several reasons for generalizing the HJM framework (1.1) by introducing jumps. Namely, this allows us to model the impact of unexpected news about the economy, such as interventions by central banks, credit events, or (natural) disasters. Indeed, there is strong statistical evidence in the finance literature that empirical features of the data cannot be captured by continuous models. We also mention that for the particular case of deterministic integrands $\alpha$ and $\sigma$ in (1.1), as it is for the Vasiček model, the log returns for discounted zerocoupon bonds are normally distributed, which, however, is not true for empirically observed log returns; see the discussion in [37, Chap. 5].

Björk et al. (see $[3,4]$ ), Eberlein et al. (see $[15,14,10,12,13,11]$ ), and others (see [38, $26,24]$ ) thus proposed replacing the classical Brownian motion $W$ in (1.1) by a more general driving noise, also taking into account the occurrence of jumps. Carmona and Tehranchi [6] proposed models based on infinite dimensional Wiener processes; see also [16]. In the spirit of Björk et al. [3] and Carmona and Tehranchi [6], we focus on term structure models of the type

$$
\begin{equation*}
d f(t, T)=\alpha(t, T) d t+\sigma(t, T) d W_{t}+\int_{E} \gamma(t, x, T)(\mu(d t, d x)-F(d x) d t) \tag{1.2}
\end{equation*}
$$

where $W$ denotes a (possibly infinite dimensional) Wiener process and, in addition, $\mu$ is a homogeneous Poisson random measure on $\mathbb{R}_{+} \times E$ with compensator $d t \otimes F(d x)$, where $E$ denotes the mark space.

In what follows, it will be convenient to switch to the alternative parametrization

$$
r_{t}(\xi):=f(t, t+\xi), \quad \xi \geq 0
$$

which is due to Musiela [32]. Then, we may regard $\left(r_{t}\right)_{t \geq 0}$ as one stochastic process with values in $H$, that is,

$$
r: \Omega \times \mathbb{R}_{+} \rightarrow H
$$

where $H$ denotes a Hilbert space of forward curves $h: \mathbb{R}_{+} \rightarrow \mathbb{R}$ to be specified later. Recall that we always assume that forward rate curves are continuous. Denoting by $\left(S_{t}\right)_{t \geq 0}$ the shift semigroup on $H$, that is, $S_{t} h=h(t+\cdot)$, (1.2) becomes in integrated form

$$
\begin{align*}
r_{t}(\xi)= & S_{t} h_{0}(\xi)+\int_{0}^{t} S_{t-s} \alpha(s, s+\xi) d s+\int_{0}^{t} S_{t-s} \sigma(s, s+\xi) d W_{s} \\
& +\int_{0}^{t} \int_{E} S_{t-s} \gamma(s, x, s+\xi)(\mu(d s, d x)-F(d x) d s), \quad t \geq 0 \tag{1.3}
\end{align*}
$$

where $h_{0} \in H$ denotes the initial forward curve and $S_{t-s}$ operates on the functions $\xi \mapsto$ $\alpha(s, s+\xi), \xi \mapsto \sigma(s, s+\xi)$, and $\xi \mapsto \gamma(s, x, s+\xi)$.

From a financial modeling point of view, one would rather consider drift and volatilities to be functions of the prevailing forward curve, that is,

$$
\begin{aligned}
\alpha & : H \\
\sigma^{j} & : H \\
\gamma & \rightarrow H \quad \text { for all } j, \\
\gamma & \times E \rightarrow H .
\end{aligned}
$$

For example, the volatilities could be of the form $\sigma^{j}(h)=\phi_{j}\left(\ell_{1}(h), \ldots, \ell_{p}(h)\right)$ for some $p \in \mathbb{N}$ with $\phi_{j}: \mathbb{R}^{p} \rightarrow H$ and $\ell_{i}: H \rightarrow \mathbb{R}$. We may think of $\ell_{i}(h)=\frac{1}{\xi_{i}} \int_{0}^{\xi_{i}} h(\eta) d \eta$ (benchmark yields) or $\ell_{i}(h)=h\left(\xi_{i}\right)$ (benchmark forward rates).

The implied bond market

$$
\begin{equation*}
P(t, T)=\exp \left(-\int_{0}^{T-t} r_{t}(\xi) d \xi\right) \tag{1.4}
\end{equation*}
$$

is free of arbitrage if we can find an equivalent (local) martingale measure $\mathbb{Q} \sim \mathbb{P}$ such that the discounted bond prices

$$
\begin{equation*}
\exp \left(-\int_{0}^{t} r_{s}(0) d s\right) P(t, T), \quad t \in[0, T] \tag{1.5}
\end{equation*}
$$

are local $\mathbb{Q}$-martingales for all maturities $T$. In what follows, we will directly specify the HJM equation under a martingale measure. More precisely, we will assume that the drift $\alpha=\alpha_{\mathrm{HJM}}: H \rightarrow H$ is given by

$$
\begin{equation*}
\alpha_{\mathrm{HJM}}(h):=\sum_{j} \sigma^{j}(h) \Sigma^{j}(h)-\int_{E} \gamma(h, x)\left(e^{\Gamma(h, x)}-1\right) F(d x) \tag{1.6}
\end{equation*}
$$

for all $h \in H$, where we have set

$$
\begin{align*}
\Sigma^{j}(h)(\xi) & :=\int_{0}^{\xi} \sigma^{j}(h)(\eta) d \eta \quad \text { for all } j,  \tag{1.7}\\
\Gamma(h, x)(\xi) & :=-\int_{0}^{\xi} \gamma(h, x)(\eta) d \eta . \tag{1.8}
\end{align*}
$$

According to [3] (if the Brownian motion is infinite dimensional; see also [16]), condition (1.6) guarantees that the discounted zero-coupon bond prices (1.5) are local martingales for all maturities $T$, whence the bond market (1.4) is free of arbitrage. In the classical situation, where the model is driven by a finite dimensional standard Brownian motion, (1.6) is the well-known HJM drift condition derived in [23].

Our requirements lead to the forward rates $\left(r_{t}\right)_{t \geq 0}$ in (1.3) being a solution of the stochastic equation

$$
\begin{align*}
r_{t}= & S_{t} h_{0}+\int_{0}^{t} S_{t-s} \alpha_{\mathrm{HJM}}\left(r_{s}\right) d s+\int_{0}^{t} S_{t-s} \sigma\left(r_{s}\right) d W_{s} \\
& +\int_{0}^{t} \int_{E} S_{t-s} \gamma\left(r_{s-}, x\right)(\mu(d s, d x)-F(d x) d s), \quad t \geq 0 \tag{1.9}
\end{align*}
$$

and it raises the question of whether this equation possesses a solution. To our knowledge, there has not yet been an explicit proof for the existence of a solution to the Poisson measure driven equation (1.9). We thus provide such a proof in our paper; see Theorem 3.2. For term structure models driven by a Brownian motion, the existence proof has been provided in [16] and for the Lévy case in [19]. We also refer the reader to the related papers [35] and [28].

In the spirit of [8] and [36], an $H$-valued stochastic process $\left(r_{t}\right)_{t \geq 0}$ satisfying (1.9) is a so-called mild solution for the (semilinear) SPDE

$$
\left\{\begin{align*}
d r_{t} & =\left(\frac{d}{d \xi} r_{t}+\alpha_{\mathrm{HJM}}\left(r_{t}\right)\right) d t+\sigma\left(r_{t}\right) d W_{t}+\int_{E} \gamma\left(r_{t-}, x\right)(\mu(d t, d x)-F(d x) d t)  \tag{1.10}\\
r_{0} & =h_{0}
\end{align*}\right.
$$

where $\frac{d}{d \xi}$ becomes the infinitesimal generator of the strongly continuous semigroup of shifts $\left(S_{t}\right)_{t \geq 0}$.

As in [20], we understand SPDEs as time-dependent transformations of time-dependent SDEs with infinite dimensional state space. More precisely, on an enlarged space $\mathcal{H}$ of forward curves $h: \mathbb{R} \rightarrow \mathbb{R}$, which are indexed by the whole real line, equipped with the strongly continuous group $\left(U_{t}\right)_{t \in \mathbb{R}}$ of shifts, we solve the SDE

$$
\left\{\begin{align*}
d f_{t}= & U_{-t} \ell \alpha_{\mathrm{HJM}}\left(\pi U_{t} f_{t}\right) d t+U_{-t} \ell \sigma\left(\pi U_{t} f_{t}\right) d W_{t}  \tag{1.11}\\
& +\int_{E} U_{-t} \ell \gamma\left(\pi U_{t} f_{t-}, x\right)(\mu(d t, d x)-F(d x) d t) \\
f_{0}= & \ell h_{0}
\end{align*}\right.
$$

where $\ell: H \rightarrow \mathcal{H}$ is an isometric embedding and $\pi: \mathcal{H} \rightarrow H$ is the orthogonal projection on $H$, and afterwards we transform the solution process $\left(f_{t}\right)_{t \geq 0}$ by $r_{t}:=\pi U_{t} f_{t}$ in order to obtain a mild solution for (1.10). Notice that (1.11) corresponds just to the original HJM dynamics in (1.2), where, of course, the forward rate $f_{t}(T)$ has no economic interpretation for $T<t$. Thus, we will henceforth refer to (1.11) as the HJM equation.

We emphasize that knowledge about the HJM equation (1.11) is not necessary in order to deduce existence and uniqueness for the forward curve evolution (1.10). The only thing we require in order to apply the existence result from [20] is that we can embed the space $H$ of forward curves into a larger Hilbert space $\mathcal{H}$ on which the shift semigroup extends to a group. The point of the "method of the moving frame" from [20] is to transform an SPDE into a time-dependent SDE on $\mathcal{H}$, but we do not need the particular structure of this extension.

Our existence result of this paper generalizes that for pure diffusion models from [16] and for term structure models driven by Lévy processes from [19] (we also mention the papers [35] and [28] concerning Lévy term structure models); see Corollary 3.3. As described above, in order to establish the proof we apply a result from [20], which contains existence and uniqueness results for general SPDEs on Hilbert spaces. Other references for SPDEs on Hilbert spaces are [1] and [29]. Since the HJM drift term $\alpha_{H J M}$ is of the particular form (1.6), we have to work out sufficient conditions which permit an application of this existence result.

In practice, we are interested in term structure models producing positive forward curves, since negative forward rates are very rarely observed. After establishing the existence issue, we shall therefore focus on positivity preserving term structure models and give a characterization of such models. The HJM equation (1.11) on the enlarged function space will be the key for
analyzing positivity of forward curves. Indeed, the "method of the moving frame" (see [20]) allows us to use standard stochastic analysis (see [25]) for our investigations. It will turn out that the conditions

$$
\begin{aligned}
& \sigma^{j}(h)(\xi)=0 \quad \text { for all } \xi \in(0, \infty), h \in \partial P_{\xi} \text { and all } j, \\
& h+\gamma(h, x) \in P \quad \text { for all } h \in P \text { and } F \text {-almost all } x \in E, \\
& \gamma(h, x)(\xi)=0 \quad \text { for all } \xi \in(0, \infty), h \in \partial P_{\xi} \text { and } F \text {-almost all } x \in E,
\end{aligned}
$$

where $P$ denotes the convex cone of all nonnegative forward curves and $\partial P_{\xi}$ is the "edge" consisting of all nonnegative forward curves $h$ with $h(\xi)=0$, are necessary and sufficient for the positivity preserving property; see Theorem 4.13. For this purpose, we provide a general positivity preserving result (see Theorem 4.11), which is of independent interest and can also be applied to other function spaces. Positivity results for the diffusion case have been worked out in [27] and [30]. In particular, we would like to mention the important and beautiful work [34], where, through an application of a general support theorem, positivity is proved. We shall also apply this general argument for our purposes.

The remainder of this text is organized as follows. In section 2 we introduce the space $H_{\beta}$ of forward curves. Using this space, we prove in section 3, under appropriate regularity assumptions, the existence of a unique solution for the Heath-Jarrow-Morton-Musiela (HJMM) equation (1.10). The positivity issue of term structure models is treated in section 4. There, we first show the necessary conditions with a general semimartingale argument. The sufficient conditions are proved to hold true by switching on the jumps "slowly". This allows for a reduction to results from [34]. For the sake of lucidity, we postpone the proofs of some auxiliary results to the appendix.
2. The space of forward curves. In this section, we introduce the space of forward curves on which we will solve the HJMM equation (1.10) in section 3.

We fix an arbitrary constant $\beta>0$. Let $H_{\beta}$ be the space of all absolutely continuous functions $h: \mathbb{R}_{+} \rightarrow \mathbb{R}$ such that

$$
\|h\|_{\beta}:=\left(|h(0)|^{2}+\int_{\mathbb{R}_{+}}\left|h^{\prime}(\xi)\right|^{2} e^{\beta \xi} d \xi\right)^{1 / 2}<\infty
$$

Let $\left(S_{t}\right)_{t \geq 0}$ be the shift semigroup on $H_{\beta}$ defined by $S_{t} h:=h(t+\cdot)$ for $t \in \mathbb{R}_{+}$.
Since forward curves should flatten for large times to maturity $\xi$, the choice of $H_{\beta}$ is reasonable from an economic point of view.

Moreover, let $\mathcal{H}_{\beta}$ be the space of all absolutely continuous functions $h: \mathbb{R} \rightarrow \mathbb{R}$ such that

$$
\|h\|_{\beta}:=\left(|h(0)|^{2}+\int_{\mathbb{R}}\left|h^{\prime}(\xi)\right|^{2} e^{\beta|\xi|} d \xi\right)^{1 / 2}<\infty
$$

Let $\left(U_{t}\right)_{t \in \mathbb{R}}$ be the shift group on $\mathcal{H}_{\beta}$ defined by $U_{t} h:=h(t+\cdot)$ for $t \in \mathbb{R}$.
The linear operator $\ell: H_{\beta} \rightarrow \mathcal{H}_{\beta}$ defined by

$$
\ell(h)(\xi):=\left\{\begin{array}{ll}
h(0), & \xi<0, \\
h(\xi), & \xi \geq 0,
\end{array} \quad h \in H_{\beta}\right.
$$

is an isometric embedding with adjoint operator $\pi:=\ell^{*}: \mathcal{H}_{\beta} \rightarrow H_{\beta}$ given by $\pi(h)=\left.h\right|_{\mathbb{R}_{+}}$, $h \in \mathcal{H}_{\beta}$.

Theorem 2.1. Let $\beta>0$ be arbitrary.

1. The space $\left(H_{\beta},\|\cdot\|_{\beta}\right)$ is a separable Hilbert space.
2. For each $\xi \in \mathbb{R}_{+}$, the point evaluation $h \mapsto h(\xi): H_{\beta} \rightarrow \mathbb{R}$ is a continuous linear functional.
3. $\left(S_{t}\right)_{t \geq 0}$ is a $C_{0}$-semigroup on $H_{\beta}$ with infinitesimal generator $\frac{d}{d \xi}: \mathcal{D}\left(\frac{d}{d \xi}\right) \subset H_{\beta} \rightarrow H_{\beta}$, $\frac{d}{d \xi} h=h^{\prime}$, and domain

$$
\mathcal{D}\left(\frac{d}{d \xi}\right)=\left\{h \in H_{\beta} \mid h^{\prime} \in H_{\beta}\right\}
$$

4. Each $h \in H_{\beta}$ is continuous and bounded, and the limit $h(\infty):=\lim _{\xi \rightarrow \infty} h(\xi)$ exists.
5. $H_{\beta}^{0}:=\left\{h \in H_{\beta} \mid h(\infty)=0\right\}$ is a closed subspace of $H_{\beta}$.
6. There are universal constants $C_{1}, C_{2}, C_{3}, C_{4}>0$, depending only on $\beta$, such that for all $h \in H_{\beta}$ we have the estimates

$$
\begin{align*}
\left\|h^{\prime}\right\|_{L^{1}\left(\mathbb{R}_{+}\right)} & \leq C_{1}\|h\|_{\beta},  \tag{2.1}\\
\|h\|_{L^{\infty}\left(\mathbb{R}_{+}\right)} & \leq C_{2}\|h\|_{\beta},  \tag{2.2}\\
\|h-h(\infty)\|_{L^{1}\left(\mathbb{R}_{+}\right)} & \leq C_{3}\|h\|_{\beta},  \tag{2.3}\\
\left\|(h-h(\infty))^{4} e^{\beta \bullet}\right\|_{L^{1}\left(\mathbb{R}_{+}\right)} & \leq C_{4}\|h\|_{\beta}^{4} \tag{2.4}
\end{align*}
$$

7. For each $\beta^{\prime}>\beta$, we have $H_{\beta^{\prime}} \subset H_{\beta}$ and the relation

$$
\begin{equation*}
\|h\|_{\beta} \leq\|h\|_{\beta^{\prime}}, \quad h \in H_{\beta^{\prime}} \tag{2.5}
\end{equation*}
$$

and there is a universal constant $C_{5}>0$, depending only on $\beta$ and $\beta^{\prime}$, such that for all $h \in H_{\beta^{\prime}}$ we have the estimate

$$
\begin{equation*}
\left\|(h-h(\infty))^{2} e^{\beta \bullet}\right\|_{L^{1}\left(\mathbb{R}_{+}\right)} \leq C_{5}\|h\|_{\beta^{\prime}}^{2} \tag{2.6}
\end{equation*}
$$

8. The space $\left(\mathcal{H}_{\beta},\|\cdot\|_{\beta}\right)$ is a separable Hilbert space, $\left(U_{t}\right)_{t \in \mathbb{R}}$ is a $C_{0}$-group on $\mathcal{H}_{\beta}$, and, for each $\xi \in \mathbb{R}$, the point evaluation $h \mapsto h(\xi), \mathcal{H}_{\beta} \rightarrow \mathbb{R}$ is a continuous linear functional.
9. The diagram

commutes for every $t \in \mathbb{R}_{+}$, that is,

$$
\begin{equation*}
\pi U_{t} \ell=S_{t} \quad \text { for all } t \in \mathbb{R}_{+} \tag{2.7}
\end{equation*}
$$

Proof. See the appendix.
3. Existence of term structure models driven by Wiener processes and Poisson measures. In this section, we establish existence and uniqueness of the HJMM equation (1.10) with diffusive and jump components on the Hilbert spaces introduced in the previous section.

Let $0<\beta<\beta^{\prime}$ be arbitrary real numbers. We denote by $H_{\beta}$ and $H_{\beta^{\prime}}$ the Hilbert spaces of the previous section, equipped with the strongly continuous semigroup $\left(S_{t}\right)_{t \geq 0}$ of shifts, which has the infinitesimal generator $\frac{d}{d \xi}$.

In what follows, $\left(\Omega, \mathcal{F},\left(\mathcal{F}_{t}\right)_{t \geq 0}, \mathbb{P}\right)$ denotes a filtered probability space satisfying the usual conditions.

Let $U$ be another separable Hilbert space, and let $Q \in L(U)$ be a compact, self-adjoint, strictly positive linear operator. Then there exist an orthonormal basis $\left\{e_{j}\right\}$ of $U$ and a bounded sequence $\lambda_{j}$ of strictly positive real numbers such that

$$
Q u=\sum_{j} \lambda_{j}\left\langle u, e_{j}\right\rangle e_{j}, \quad u \in U
$$

namely, the $\lambda_{j}$ are the eigenvalues of $Q$, and each $e_{j}$ is an eigenvector corresponding to $\lambda_{j}$; see, e.g., [41, Thm. VI.3.2].

The space $U_{0}:=Q^{1 / 2}(U)$, equipped with the inner product

$$
\langle u, v\rangle_{U_{0}}:=\left\langle Q^{-1 / 2} u, Q^{-1 / 2} v\right\rangle_{U},
$$

is another separable Hilbert space and $\left\{\sqrt{\lambda_{j}} e_{j}\right\}$ is an orthonormal basis.
Let $W$ be a $Q$-Wiener process [8, pp. 86-87]. We assume that $\operatorname{tr}(Q)=\sum_{j} \lambda_{j}<\infty$. Otherwise, which is the case if $W$ is a cylindrical Wiener process, there always exists a separable Hilbert space $U_{1} \supset U$ on which $W$ has a realization as a finite trace class Wiener process; see [8, Chap. 4.3].

We denote by $L_{2}^{0}\left(H_{\beta}\right):=L_{2}\left(U_{0}, H_{\beta}\right)$ the space of Hilbert-Schmidt operators from $U_{0}$ into $H_{\beta}$, which, endowed with the Hilbert-Schmidt norm

$$
\|\Phi\|_{L_{2}^{0}\left(H_{\beta}\right)}:=\sqrt{\sum_{j} \lambda_{j}\left\|\Phi e_{j}\right\|^{2}}, \quad \Phi \in L_{2}^{0}\left(H_{\beta}\right),
$$

itself is a separable Hilbert space.
According to [8, Prop. 4.1], the sequence of stochastic processes $\left\{\beta^{j}\right\}$ defined as $\beta^{j}:=$ $\frac{1}{\sqrt{\lambda_{j}}}\left\langle W, e_{j}\right\rangle$ is a sequence of real-valued independent $\left(\mathcal{F}_{t}\right)$-Brownian motions and we have the expansion

$$
\begin{equation*}
W=\sum_{j} \sqrt{\lambda_{j}} \beta^{j} e_{j}, \tag{3.1}
\end{equation*}
$$

where the series is convergent in the space $M^{2}(U)$ of $U$-valued square-integrable martingales. Let $\Phi: \Omega \times \mathbb{R}_{+} \rightarrow L_{2}^{0}\left(H_{\beta}\right)$ be an integrable process; i.e., $\Phi$ is predictable and satisfies

$$
\mathbb{P}\left(\int_{0}^{T}\left\|\Phi_{t}\right\|_{L_{2}^{0}\left(H_{\beta}\right)}^{2} d t<\infty\right)=1 \quad \text { for all } T \in \mathbb{R}_{+}
$$

Setting $\Phi^{j}:=\sqrt{\lambda_{j}} \Phi e_{j}$ for each $j$, we have

$$
\begin{equation*}
\int_{0}^{t} \Phi_{s} d W_{s}=\sum_{j} \int_{0}^{t} \Phi_{s}^{j} d \beta_{s}^{j}, \quad t \in \mathbb{R}_{+} \tag{3.2}
\end{equation*}
$$

where the convergence is uniform on compact time intervals in probability; see [8, Thm. 4.3].
Let $(E, \varepsilon)$ be a measurable space which we assume to be a Blackwell space (see $[9,22]$ ). We remark that every Polish space with its Borel $\sigma$-field is a Blackwell space.

Furthermore, let $\mu$ be a homogeneous Poisson random measure on $\mathbb{R}_{+} \times E$; see [25, Def. II.1.20]. Then its compensator is of the form $d t \otimes F(d x)$, where $F$ is a $\sigma$-finite measure on $(E, \varepsilon)$.

Let measurable vector fields $\sigma: H_{\beta} \rightarrow L_{2}^{0}\left(H_{\beta}^{0}\right)$ and $\gamma: H_{\beta} \times E \rightarrow H_{\beta^{\prime}}^{0}$ be given, where the subspace $H_{\beta}^{0}$ is as defined in Theorem 2.1. For each $j$, we define $\sigma^{j}: H_{\beta} \rightarrow H_{\beta}^{0}$ as $\sigma^{j}(h):=\sqrt{\lambda_{j}} \sigma(h) e_{j}$. We shall now focus on the HJMM equation (1.10).

Assumption 3.1. We assume there exists a measurable function $\Phi: E \rightarrow \mathbb{R}_{+}$satisfying

$$
\begin{equation*}
|\Gamma(h, x)(\xi)| \leq \Phi(x), \quad h \in H_{\beta}, x \in E, \text { and } \xi \in \mathbb{R}_{+} \tag{3.3}
\end{equation*}
$$

a constant $L>0$ such that

$$
\begin{align*}
\left\|\sigma\left(h_{1}\right)-\sigma\left(h_{2}\right)\right\|_{L_{2}^{0}\left(H_{\beta}\right)} & \leq L\left\|h_{1}-h_{2}\right\|_{\beta},  \tag{3.4}\\
\left(\int_{E} e^{\Phi(x)}\left\|\gamma\left(h_{1}, x\right)-\gamma\left(h_{2}, x\right)\right\|_{\beta^{\prime}}^{2} F(d x)\right)^{1 / 2} & \leq L\left\|h_{1}-h_{2}\right\|_{\beta} \tag{3.5}
\end{align*}
$$

for all $h_{1}, h_{2} \in H_{\beta}$, and a constant $M>0$ such that

$$
\begin{equation*}
\int_{E} e^{\Phi(x)}\left(\|\gamma(h, x)\|_{\beta^{\prime}}^{2} \vee\|\gamma(h, x)\|_{\beta^{\prime}}^{4}\right) F(d x) \leq M \tag{3.6}
\end{equation*}
$$

for all $h \in H_{\beta}$. Furthermore, we assume that for each $h \in H_{\beta}$ the map

$$
\begin{equation*}
\alpha_{2}(h):=-\int_{E} \gamma(h, x)\left(e^{\Gamma(h, x)}-1\right) F(d x) \tag{3.8}
\end{equation*}
$$

is absolutely continuous with weak derivative

$$
\begin{equation*}
\frac{d}{d \xi} \alpha_{2}(h)=\int_{E} \gamma(h, x)^{2} e^{\Gamma(h, x)} F(d x)-\int_{E} \frac{d}{d \xi} \gamma(h, x)\left(e^{\Gamma(h, x)}-1\right) F(d x) . \tag{3.9}
\end{equation*}
$$

Remark 3.2. The proof of Proposition 3.1 gives rise to the following remarks concerning conditions (3.8), (3.9) from Assumption 3.1.

- For each $h \in H_{\beta}$, the map $\alpha_{2}(h)$ is well defined (that is, for every $\xi \in \mathbb{R}_{+}$the integral in (3.8) exists) and continuous (but not necessarily absolutely continuous) with $\alpha_{2}(h)(0)=0$ and $\lim _{\xi \rightarrow \infty} \alpha_{2}(h)(\xi)=0$.
- If $\alpha_{2}(h)$ is absolutely continuous for some $h \in H_{\beta}$, then we have $\alpha_{2}(h) \in H_{\beta}^{0}$.
- For Lévy-driven term structure models, we can verify conditions (3.8) and (3.9) directly; see Corollary 3.3.
- Let $h \in H_{\beta}$ be such that $\gamma(h, x) \in C^{1}\left(\mathbb{R}_{+}\right)$for all $x \in E$. Suppose for each $\xi \in \mathbb{R}_{+}$ there exist $\delta=\delta(h, \xi)>0$ and $C=C(h, \xi)>0$ such that

$$
\left|\frac{d}{d \eta} \gamma(h, x)(\eta)\right| \leq C\|\gamma(h, x)\|_{\beta^{\prime}} \quad \text { for all } x \in E \text { and } \eta \in(\xi-\delta, \xi+\delta) \cap \mathbb{R}_{+}
$$

Then, we even have $\alpha_{2}(h) \in C^{1}\left(\mathbb{R}_{+}\right)$with derivative (3.9).
Proposition 3.1. Suppose Assumption 3.1 is fulfilled. Then we have $\alpha_{\mathrm{HJM}}\left(H_{\beta}\right) \subset H_{\beta}^{0}$ and there is a constant $K>0$ such that

$$
\begin{equation*}
\left\|\alpha_{\mathrm{HJM}}\left(h_{1}\right)-\alpha_{\mathrm{HJM}}\left(h_{2}\right)\right\|_{\beta} \leq K\left\|h_{1}-h_{2}\right\|_{\beta} \tag{3.10}
\end{equation*}
$$

for all $h_{1}, h_{2} \in H_{\beta}$.
Proof. Note that $\alpha_{\mathrm{HJM}}=\alpha_{1}+\alpha_{2}$, where

$$
\alpha_{1}(h):=\sum_{j} \sigma^{j}(h) \Sigma^{j}(h), \quad h \in H_{\beta},
$$

and $\alpha_{2}$ is given by (3.8). By [16, Cor. 5.1.2], we have $\sigma^{j}(h) \Sigma^{j}(h) \in H_{\beta}^{0}, h \in H_{\beta}$ for all $j$. For an arbitrary $h \in H_{\beta}$, we obtain, by using [16, Cor. 5.1.2] again,

$$
\sum_{j}\left\|\sigma^{j}(h) \Sigma^{j}(h)\right\|_{\beta} \leq \sqrt{3\left(C_{3}^{2}+2 C_{4}\right)} \sum_{j}\left\|\sigma^{j}(h)\right\|_{\beta}^{2}=\sqrt{3\left(C_{3}^{2}+2 C_{4}\right)}\|\sigma(h)\|_{L_{2}^{0}\left(H_{\beta}\right)}^{2},
$$

and hence we deduce that $\alpha_{1}\left(H_{\beta}\right) \subset H_{\beta}^{0}$.
Let $h \in H_{\beta}$ be arbitrary. For all $x \in E$ and $\xi \in \mathbb{R}_{+}$, we have by (2.2) and (2.5)

$$
\begin{equation*}
|\gamma(h, x)(\xi)| \leq C_{2}\|\gamma(h, x)\|_{\beta} \leq C_{2}\|\gamma(h, x)\|_{\beta^{\prime}} \tag{3.11}
\end{equation*}
$$

and for all $x \in E$ and $\xi \in \mathbb{R}_{+}$we have by (3.3), (2.3), and (2.5)

$$
\begin{equation*}
\left|e^{\Gamma(h, x)(\xi)}-1\right| \leq e^{\Phi(x)}|\Gamma(h, x)(\xi)| \leq e^{\Phi(x)}\|\gamma(h, x)\|_{L^{1}\left(\mathbb{R}_{+}\right)} \leq C_{3} e^{\Phi(x)}\|\gamma(h, x)\|_{\beta^{\prime}} . \tag{3.12}
\end{equation*}
$$

Estimates (3.11), (3.12), and (3.7) show that $\lim _{\xi \rightarrow \infty} \alpha_{2}(h)(\xi)=0$. From (3.3), (3.11), (3.7), and (2.6), it follows that

$$
\begin{aligned}
& \int_{\mathbb{R}_{+}}\left(\int_{E} \gamma(h, x)(\xi)^{2} e^{\Gamma(h, x)(\xi)} F(d x)\right)^{2} e^{\beta \xi} d \xi \\
& \leq C_{2}^{2} M \int_{\mathbb{R}_{+}}\left(\int_{E} \gamma(h, x)(\xi)^{2} e^{\Gamma(h, x)(\xi)} F(d x)\right) e^{\beta \xi} d \xi \\
& \leq C_{2}^{2} M \int_{E} e^{\Phi(x)} \int_{\mathbb{R}_{+}} \gamma(h, x)(\xi)^{2} e^{\beta \xi} d \xi F(d x) \\
& \leq C_{2}^{2} M C_{5} \int_{E} e^{\Phi(x)}\|\gamma(h, x)\|_{\beta^{\prime}}^{2} F(d x) \leq C_{2}^{2} M^{2} C_{5} .
\end{aligned}
$$

We obtain by (3.12), Hölder's inequality, (3.7), and (2.5)

$$
\begin{aligned}
& \int_{\mathbb{R}_{+}}\left(\int_{E} \frac{d}{d \xi} \gamma(h, x)(\xi)\left(e^{\Gamma(h, x)(\xi)}-1\right) F(d x)\right)^{2} e^{\beta \xi} d \xi \\
& \leq C_{3}^{2} \int_{\mathbb{R}_{+}}\left(\int_{E}\left|\frac{d}{d \xi} \gamma(h, x)(\xi)\right| e^{\frac{1}{2} \Phi(x)} e^{\frac{1}{2} \Phi(x)}\|\gamma(h, x)\|_{\beta^{\prime}} F(d x)\right)^{2} e^{\beta \xi} d \xi \\
& \leq C_{3}^{2} M \int_{E} e^{\Phi(x)} \int_{\mathbb{R}_{+}}\left|\frac{d}{d \xi} \gamma(h, x)(\xi)\right|^{2} e^{\beta \xi} d \xi F(d x) \\
& \leq C_{3}^{2} M \int_{E} e^{\Phi(x)}\|\gamma(h, x)\|_{\beta^{\prime}}^{2} F(d x) \leq C_{3}^{2} M^{2}
\end{aligned}
$$

In view of (3.9), we conclude that $\alpha_{2}\left(H_{\beta}\right) \subset H_{\beta}^{0}$, and hence $\alpha_{\mathrm{HJM}}\left(H_{\beta}\right) \subset H_{\beta}^{0}$.
Let $h_{1}, h_{2} \in H_{\beta}$ be arbitrary. By [16, Cor. 5.1.2], Hölder's inequality, (3.4), and (3.6), we have

$$
\begin{aligned}
& \left\|\alpha_{1}\left(h_{1}\right)-\alpha_{1}\left(h_{2}\right)\right\|_{\beta} \\
& \leq \sqrt{3\left(C_{3}^{2}+2 C_{4}\right)} \sum_{j}\left(\left\|\sigma^{j}\left(h_{1}\right)\right\|_{\beta}+\left\|\sigma^{j}\left(h_{2}\right)\right\|_{\beta}\right)\left\|\sigma^{j}\left(h_{1}\right)-\sigma^{j}\left(h_{2}\right)\right\|_{\beta} \\
& \leq \sqrt{3\left(C_{3}^{2}+2 C_{4}\right)} \sqrt{\sum_{j}\left(\left\|\sigma^{j}\left(h_{1}\right)\right\|_{\beta}+\left\|\sigma^{j}\left(h_{2}\right)\right\|_{\beta}\right)^{2}} \sqrt{\sum_{j}\left\|\sigma^{j}\left(h_{1}\right)-\sigma^{j}\left(h_{2}\right)\right\|_{\beta}^{2}} \\
& \leq \sqrt{6\left(C_{3}^{2}+2 C_{4}\right)\left(\left\|\sigma\left(h_{1}\right)\right\|_{L_{2}^{0}\left(H_{\beta}\right)}+\left\|\sigma\left(h_{2}\right)\right\|_{L_{2}^{0}\left(H_{\beta}\right)}\right)\left\|\sigma\left(h_{1}\right)-\sigma\left(h_{2}\right)\right\|_{L_{2}^{0}\left(H_{\beta}\right)}} \\
& \leq 2 M L \sqrt{6\left(C_{3}^{2}+2 C_{4}\right)\left\|h_{1}-h_{2}\right\|_{\beta} .}
\end{aligned}
$$

Furthermore, by (3.9),

$$
\left\|\alpha_{2}\left(h_{1}\right)-\alpha_{2}\left(h_{2}\right)\right\|_{\beta}^{2} \leq 4\left(I_{1}+I_{2}+I_{3}+I_{4}\right)
$$

where we have put

$$
\begin{aligned}
I_{1} & :=\int_{\mathbb{R}_{+}}\left(\int_{E} \gamma\left(h_{1}, x\right)(\xi)^{2}\left(e^{\Gamma\left(h_{1}, x\right)(\xi)}-e^{\Gamma\left(h_{2}, x\right)(\xi)}\right) F(d x)\right)^{2} e^{\beta \xi} d \xi \\
I_{2} & :=\int_{\mathbb{R}_{+}}\left(\int_{E} e^{\Gamma\left(h_{2}, x\right)(\xi)}\left(\gamma\left(h_{1}, x\right)(\xi)^{2}-\gamma\left(h_{2}, x\right)(\xi)^{2}\right) F(d x)\right)^{2} e^{\beta \xi} d \xi \\
I_{3} & :=\int_{\mathbb{R}_{+}}\left(\int_{E} \frac{d}{d \xi} \gamma\left(h_{1}, x\right)(\xi)\left(e^{\Gamma\left(h_{1}, x\right)(\xi)}-e^{\Gamma\left(h_{2}, x\right)(\xi)}\right) F(d x)\right)^{2} e^{\beta \xi} d \xi \\
I_{4} & :=\int_{\mathbb{R}_{+}}\left(\int_{E}\left(e^{\Gamma\left(h_{2}, x\right)(\xi)}-1\right)\left(\frac{d}{d \xi} \gamma\left(h_{1}, x\right)(\xi)-\frac{d}{d \xi} \gamma\left(h_{2}, x\right)(\xi)\right) F(d x)\right)^{2} e^{\beta \xi} d \xi .
\end{aligned}
$$

We get for all $x \in E$ and $\xi \in \mathbb{R}_{+}$by (3.3), (2.3), and (2.5)

$$
\begin{align*}
& \left|e^{\Gamma\left(h_{1}, x\right)(\xi)}-e^{\Gamma\left(h_{2}, x\right)(\xi)}\right| \leq e^{\Phi(x)}\left|\Gamma\left(h_{1}, x\right)(\xi)-\Gamma\left(h_{2}, x\right)(\xi)\right|  \tag{3.13}\\
& \leq e^{\Phi(x)}\left\|\gamma\left(h_{1}, x\right)-\gamma\left(h_{2}, x\right)\right\|_{L^{1}\left(\mathbb{R}_{+}\right)} \leq C_{3} e^{\Phi(x)}\left\|\gamma\left(h_{1}, x\right)-\gamma\left(h_{2}, x\right)\right\|_{\beta^{\prime}}
\end{align*}
$$

Relations (3.13), Hölder's inequality, (3.5), (2.4), (2.5), and (3.7) give us

$$
\begin{aligned}
I_{1} & \leq C_{3}^{2} \int_{\mathbb{R}_{+}}\left(\int_{E} \gamma\left(h_{1}, x\right)(\xi)^{2} e^{\frac{1}{2} \Phi(x)} e^{\frac{1}{2} \Phi(x)}\left\|\gamma\left(h_{1}, x\right)-\gamma\left(h_{2}, x\right)\right\|_{\beta^{\prime}} F(d x)\right)^{2} e^{\beta \xi} d \xi \\
& \leq C_{3}^{2} L^{2}\left\|h_{1}-h_{2}\right\|_{\beta}^{2} \int_{E} e^{\Phi(x)} \int_{\mathbb{R}_{+}} \gamma\left(h_{1}, x\right)(\xi)^{4} e^{\beta \xi} d \xi F(d x) \\
& \leq C_{3}^{2} L^{2} C_{4}\left\|h_{1}-h_{2}\right\|_{\beta}^{2} \int_{E} e^{\Phi(x)}\left\|\gamma\left(h_{1}, x\right)\right\|_{\beta^{\prime}}^{4} F(d x) \leq C_{3}^{2} L^{2} C_{4} M\left\|h_{1}-h_{2}\right\|_{\beta}^{2} .
\end{aligned}
$$

For every $\xi \in \mathbb{R}_{+}$, we obtain by (3.11) and (3.7)

$$
\begin{align*}
& \int_{E} e^{\Phi(x)}\left(\gamma\left(h_{1}, x\right)(\xi)+\gamma\left(h_{2}, x\right)(\xi)\right)^{2} F(d x) \\
& \leq 2 \int_{E} e^{\Phi(x)}\left(\gamma\left(h_{1}, x\right)(\xi)^{2}+\gamma\left(h_{2}, x\right)(\xi)^{2}\right) F(d x)  \tag{3.14}\\
& \leq 2 C_{2}^{2}\left(\int_{E} e^{\Phi(x)}\left\|\gamma\left(h_{1}, x\right)\right\|_{\beta^{\prime}}^{2} F(d x)+\int_{E} e^{\Phi(x)}\left\|\gamma\left(h_{2}, x\right)\right\|_{\beta^{\prime}}^{2} F(d x)\right) \leq 4 C_{2}^{2} M .
\end{align*}
$$

Using (3.3), Hölder's inequality, (3.14), (2.6), and (3.5), we get

$$
\begin{aligned}
& I_{2} \leq \int_{\mathbb{R}_{+}}\left(\int_{E}\left(\gamma\left(h_{1}, x\right)(\xi)+\gamma\left(h_{2}, x\right)(\xi)\right) e^{\frac{1}{2} \Phi(x)}\right. \\
&\left.\quad \times e^{\frac{1}{2} \Phi(x)}\left(\gamma\left(h_{1}, x\right)(\xi)-\gamma\left(h_{2}, x\right)(\xi)\right) F(d x)\right)^{2} e^{\beta \xi} d \xi \\
& \leq 4 C_{2}^{2} M \int_{E} e^{\Phi(x)} \int_{\mathbb{R}_{+}}\left(\gamma\left(h_{1}, x\right)(\xi)-\gamma\left(h_{2}, x\right)(\xi)\right)^{2} e^{\beta \xi} d \xi F(d x) \\
& \leq 4 C_{2}^{2} M C_{5} \int_{E} e^{\Phi(x)}\left\|\gamma\left(h_{1}, x\right)(\xi)-\gamma\left(h_{2}, x\right)(\xi)\right\|_{\beta^{\prime}}^{2} F(d x) \\
& \leq 4 C_{2}^{2} M C_{5} L^{2}\left\|h_{1}-h_{2}\right\|_{\beta}^{2} .
\end{aligned}
$$

Using (3.13), Hölder's inequality, (3.5), (2.5), and (3.7) gives us

$$
\begin{aligned}
I_{3} \leq & C_{3}^{2} \int_{\mathbb{R}_{+}} \\
& \left(\int_{E}\left|\frac{d}{d \xi} \gamma\left(h_{1}, x\right)(\xi)\right| e^{\frac{1}{2} \Phi(x)} e^{\frac{1}{2} \Phi(x)}\left\|\gamma\left(h_{1}, x\right)-\gamma\left(h_{2}, x\right)\right\|_{\beta^{\prime}} F(d x)\right)^{2} \\
\leq & C_{3}^{2} L^{2}\left\|h_{1}-h_{2}\right\|_{\beta}^{2} \int_{E} e^{\Phi(x)} \int_{\mathbb{R}_{+}}\left|\frac{d}{d \xi} \gamma\left(h_{1}, x\right)(\xi)\right|^{2} e^{\beta \xi} d \xi F(d x) \\
\leq & C_{3}^{2} L^{2}\left\|h_{1}-h_{2}\right\|_{\beta}^{2} \int_{E} e^{\Phi(x)}\left\|\gamma\left(h_{1}, x\right)\right\|_{\beta^{\prime}}^{2} F(d x) \leq C_{3}^{2} L^{2} M\left\|h_{1}-h_{2}\right\|_{\beta}^{2} .
\end{aligned}
$$

We obtain by (3.12), Hölder's inequality, (3.7), (2.5), and (3.5)

$$
\begin{aligned}
I_{4} \leq & C_{3}^{2} \int_{\mathbb{R}_{+}}\left(\int_{E}\left\|\gamma\left(h_{2}, x\right)\right\|_{\beta^{\prime}} e^{\frac{1}{2} \Phi(x)} e^{\frac{1}{2} \Phi(x)}\left|\frac{d}{d \xi} \gamma\left(h_{1}, x\right)(\xi)-\frac{d}{d \xi} \gamma\left(h_{2}, x\right)(\xi)\right| F(d x)\right)^{2} \\
& \times e^{\beta \xi} d \xi \\
\leq & C_{3}^{2} M \int_{E} e^{\Phi(x)} \int_{\mathbb{R}_{+}}\left|\frac{d}{d \xi} \gamma\left(h_{1}, x\right)(\xi)-\frac{d}{d \xi} \gamma\left(h_{2}, x\right)(\xi)\right|^{2} e^{\beta \xi} d \xi F(d x) \\
\leq & C_{3}^{2} M \int_{E} e^{\Phi(x)}\left\|\gamma\left(h_{1}, x\right)-\gamma\left(h_{2}, x\right)\right\|_{\beta^{\prime}}^{2} F(d x) \leq C_{3}^{2} M L^{2}\left\|h_{1}-h_{2}\right\|_{\beta}^{2} .
\end{aligned}
$$

Summing up, we deduce that there is a constant $K>0$ such that (3.10) is satisfied for all $h_{1}, h_{2} \in H_{\beta}$.

Theorem 3.2. Suppose Assumption 3.1 is fulfilled. Then, for each initial curve $h_{0} \in$ $L^{2}\left(\Omega, \mathcal{F}_{0}, \mathbb{P} ; H_{\beta}\right)$ there exists a unique adapted, càdlàg, mean-square continuous $\mathcal{H}_{\beta}$-valued solution $\left(f_{t}\right)_{t \geq 0}$ for the HJM equation (1.11) with $f_{0}=\ell h_{0}$ satisfying

$$
\begin{equation*}
\mathbb{E}\left[\sup _{t \in[0, T]}\left\|f_{t}\right\|_{\beta}^{2}\right]<\infty \quad \text { for all } T \in \mathbb{R}_{+} \tag{3.15}
\end{equation*}
$$

and there exists a unique adapted, càdlàg, mean-square continuous mild and weak $H_{\beta}$-valued solution $\left(r_{t}\right)_{t \geq 0}$ for the HJMM equation (1.10) with $r_{0}=h_{0}$ satisfying

$$
\begin{equation*}
\mathbb{E}\left[\sup _{t \in[0, T]}\left\|r_{t}\right\|_{\beta}^{2}\right]<\infty \quad \text { for all } T \in \mathbb{R}_{+} \tag{3.16}
\end{equation*}
$$

which is given by $r_{t}:=\pi U_{t} f_{t}, t \geq 0$. Moreover, the implied bond market (1.4) is free of arbitrage.

Proof. By virtue of Theorem 2.1, Proposition 3.1, (3.4), (3.5), (3.7), and (2.5), all assumptions from [20, Cor. 10.9] are fulfilled, which therefore applies and establishes the claimed existence and uniqueness result.

For all $h \in H_{\beta}, x \in E$, and $\xi \in \mathbb{R}_{+}$, we have by (3.3), (2.3), and (2.5)

$$
\begin{align*}
& \left|e^{\Gamma(h, x)(\xi)}-1-\Gamma(h, x)(\xi)\right| \leq \frac{1}{2} e^{\Phi(x)} \Gamma(h, x)(\xi)^{2} \\
& \leq \frac{1}{2} e^{\Phi(x)}\|\gamma(h, x)\|_{L^{1}\left(\mathbb{R}_{+}\right)}^{2} \leq \frac{C_{3}^{2}}{2} e^{\Phi(x)}\|\gamma(h, x)\|_{\beta^{\prime}}^{2} \tag{3.17}
\end{align*}
$$

Integrating (1.6), we obtain, by using [16, Lemma 4.3.2], (3.17), and (3.7),

$$
\int_{0}^{\bullet} \alpha_{\mathrm{HJM}}(h)(\eta) d \eta=\frac{1}{2} \sum_{j} \Sigma^{j}(h)^{2}+\int_{E}\left(e^{\Gamma(h, x)}-1-\Gamma(h, x)\right) F(d x)
$$

for all $h \in H_{\beta}$. Combining [3, Prop. 5.3] and [16, Lemma 4.3.3] (the latter result is required only if $W$ is infinite dimensional), the probability measure $\mathbb{P}$ is a local martingale measure, and hence the bond market (1.4) is free of arbitrage.

The case of Lévy-driven HJMM models is now a special case. We assume that the mark space is $E=\mathbb{R}^{e}$ for some positive integer $e \in \mathbb{N}$, equipped with its Borel $\sigma$-algebra $\mathcal{E}=\mathcal{B}\left(\mathbb{R}^{e}\right)$. The measure $F$ is given by

$$
\begin{equation*}
F(B):=\sum_{k=1}^{e} \int_{\mathbb{R}} \mathbb{1}_{B}\left(x e_{k}\right) F_{k}(d x), \quad B \in \mathcal{B}\left(\mathbb{R}^{e}\right), \tag{3.18}
\end{equation*}
$$

where $F_{1}, \ldots, F_{e}$ are Lévy measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ satisfying

$$
\begin{align*}
& F_{k}(\{0\})=0, \quad k=1, \ldots, e  \tag{3.19}\\
& \int_{\mathbb{R}}\left(|x|^{2} \wedge 1\right) F_{k}(d x)<\infty, \quad k=1, \ldots, e \tag{3.20}
\end{align*}
$$

and where the $\left(e_{k}\right)_{k=1, \ldots, e}$ denote the unit vectors in $\mathbb{R}^{e}$. Note that definition (3.18) implies

$$
\begin{equation*}
\int_{\mathbb{R}^{e}} g(x) F(d x)=\sum_{k=1}^{e} \int_{\mathbb{R}} g\left(x e_{k}\right) F_{k}(d x) \tag{3.21}
\end{equation*}
$$

for any nonnegative measurable function $g: \mathbb{R}^{e} \rightarrow \mathbb{R}$. In particular, the support of $F$ is contained in $\bigcup_{k=1}^{e} \operatorname{span}\left\{e_{k}\right\}$, the union of the coordinate axes in $\mathbb{R}^{e}$. For each $k=1, \ldots, e$, let $\delta^{k}: H_{\beta} \rightarrow H_{\beta^{\prime}}^{0}$ be a vector field. We define $\gamma: H_{\beta} \times \mathbb{R}^{e} \rightarrow H_{\beta^{\prime}}^{0}$ as

$$
\begin{equation*}
\gamma(h, x):=\sum_{k=1}^{e} \delta^{k}(h) x_{k} . \tag{3.22}
\end{equation*}
$$

Then, (1.10) corresponds to the situation where the term structure model is driven by several real-valued, independent Lévy processes with Lévy measures $F_{k}$. For all $h \in H_{\beta}$ and $\xi \in \mathbb{R}_{+}$, we set

$$
\Delta^{k}(h)(\xi):=-\int_{0}^{\xi} \delta^{k}(h)(\eta) d \eta, \quad k=1, \ldots, e .
$$

Assumption 3.3. We assume there exist constants $N, \epsilon>0$ such that for all $k=1, \ldots, e$ we have

$$
\begin{align*}
& \int_{\{|x|>1\}} e^{z x} F_{k}(d x)<\infty, \quad z \in[-(1+\epsilon) N,(1+\epsilon) N],  \tag{3.23}\\
& \left|\Delta^{k}(h)(\xi)\right| \leq N, \quad h \in H_{\beta}, \xi \in \mathbb{R}_{+}, \tag{3.24}
\end{align*}
$$

a constant $L>0$ such that (3.4) and

$$
\begin{equation*}
\left\|\delta^{k}\left(h_{1}\right)-\delta^{k}\left(h_{2}\right)\right\|_{\beta^{\prime}} \leq L\left\|h_{1}-h_{2}\right\|_{\beta}, \quad k=1, \ldots, e, \tag{3.25}
\end{equation*}
$$

are satisfied for all $h_{1}, h_{2} \in H_{\beta}$, and a constant $M>0$ such that (3.6) and

$$
\begin{equation*}
\left\|\delta^{k}(h)\right\|_{\beta^{\prime}} \leq M, \quad k=1, \ldots, e \tag{3.26}
\end{equation*}
$$

are satisfied for all $h \in H_{\beta}$.
Now, we obtain the statement of [19, Thm. 4.6] as a corollary.
Corollary 3.3. Suppose Assumption 3.3 is fulfilled. Then, for each initial curve $h_{0} \in$ $L^{2}\left(\Omega, \mathcal{F}_{0}, \mathbb{P} ; H_{\beta}\right)$ there exists a unique adapted, càdlàg, mean-square continuous $\mathcal{H}_{\beta}$-valued solution $\left(f_{t}\right)_{t \geq 0}$ for the HJM equation (1.11) with $f_{0}=\ell h_{0}$ satisfying (3.15), and there exists a unique adapted, càdlàg, mean-square continuous mild and weak $H_{\beta}$-valued solution $\left(r_{t}\right)_{t \geq 0}$ for the HJMM equation (1.10) with $r_{0}=h_{0}$ satisfying (3.16), which is given by $r_{t}:=\pi U_{t} \bar{f}_{t}$, $t \geq 0$. Moreover, the implied bond market (1.4) is free of arbitrage.

Proof. Using (3.24), the measurable function $\Phi: \mathbb{R}^{e} \rightarrow \mathbb{R}_{+}$defined as

$$
\Phi(x):=N \sum_{k=1}^{e}\left|x_{k}\right|, \quad x \in \mathbb{R}^{e},
$$

satisfies (3.3). For each $k=1, \ldots, e$ and every $m \in \mathbb{N}$ with $m \geq 2$, we have

$$
\begin{equation*}
\int_{\mathbb{R}}|x|^{m} e^{|z x|} F_{k}(d x)<\infty, \quad z \in(-(1+\epsilon) N,(1+\epsilon) N) \tag{3.27}
\end{equation*}
$$

Indeed, let $z \in(-(1+\epsilon) N,(1+\epsilon) N)$ be arbitrary. There exists $\eta \in(0, \epsilon)$ such that $|z| \leq$ $(1+\eta) N$. By (3.20), (3.23), and the basic inequality $x^{m} \leq m!e^{x}$ for $x \geq 0$, we obtain

$$
\begin{aligned}
& \int_{\mathbb{R}}|x|^{m} e^{|z x|} F_{k}(d x) \leq \int_{\mathbb{R}}|x|^{m} e^{(1+\eta) N|x|} F_{k}(d x) \\
& \leq 2 \int_{\left\{|x| \leq \frac{\ln 2}{(1+\eta) N}\right\}}|x|^{m} F_{k}(d x)+\frac{m!}{((\epsilon-\eta) N)^{m}} \int_{\left\{|x|>\frac{\ln 2}{(1+\eta) N}\right\}} e^{(1+\epsilon) N|x|} F_{k}(d x)<\infty
\end{aligned}
$$

proving (3.27). Taking into account (3.21), (3.27), relations (3.25), (3.26) imply (3.5), (3.7). Furthermore, (3.27), the elementary inequalities

$$
\begin{aligned}
\left|e^{x}-1-x\right| & \leq \frac{1}{2} x^{2} e^{|x|}, \quad x \in \mathbb{R} \\
\left|e^{x}-1\right| & \leq|x| e^{|x|}, \quad x \in \mathbb{R}
\end{aligned}
$$

and Lebesgue's theorem show that the cumulant generating functions

$$
\Psi_{k}(z)=\int_{\mathbb{R}}\left(e^{z x}-1-z x\right) F_{k}(d x), \quad k=1, \ldots, e,
$$

belong to class $C^{\infty}$ on the open interval $(-(1+\epsilon) N,(1+\epsilon) N)$ with derivatives

$$
\begin{aligned}
\Psi_{k}^{\prime}(z) & =\int_{\mathbb{R}} x\left(e^{z x}-1\right) F_{k}(d x), \\
\Psi_{k}^{(m)}(z) & =\int_{\mathbb{R}} x^{m} e^{z x} F_{k}(d x), \quad m \geq 2 .
\end{aligned}
$$

Therefore, and because of (3.21), we can, for an arbitrary $h \in H_{\beta}$, write $\alpha_{2}(h)$, which is defined in (3.8), as

$$
\alpha_{2}(h)=-\sum_{k=1}^{e} \delta^{k}(h) \Psi_{k}^{\prime}\left(-\int_{0}^{\bullet} \delta^{k}(\eta) d \eta\right) .
$$

Hence, $\alpha_{2}(h)$ is absolutely continuous with weak derivative (3.9). Consequently, Assumption 3.1 is fulfilled and Theorem 3.2 applies.

Note that the boundedness assumptions (3.6), (3.7) for Theorem 3.2 (resp., (3.6), (3.26) for Corollary 3.3) cannot be weakened substantially. For example, for arbitrage-free term structure models driven by a single Brownian motion, it was shown in [31, sect. 4.7] that for the simple case of proportional volatility, that is, $\sigma(h)=\sigma_{0} h$ for some constant $\sigma_{0}>0$, solutions necessarily explode. We mention, however, that [35, sect. 6] contains some existence results for Lévy term structure models with linear volatility.

## 4. Positivity preserving term structure models driven by Wiener processes and Poisson

 measures. In applications, we are often interested in term structure models producing positive forward curves. In this section, we characterize HJMM forward curve evolutions of the type (1.10), which preserve positivity, by means of the characteristics of the SPDE. In the case of short rate models, this can be characterized by the positivity of the short rate, a onedimensional Markov process. In case of an infinite-factor evolution, as described by a generic HJMM equation (see, for instance, [2]), this problem is much more delicate. Indeed, one has to find conditions such that a Markov process defined by the HJMM equation (on a Hilbert space of forward rate curves) stays in a "small" set of curves, namely the convex cone of positive curves bounded by a nonsmooth set. Our strategy to solve this problem is the following: First, we show by general semimartingale methods necessary conditions for positivity. These necessary conditions are basically described by the fact that the Itô drift is inward pointing and that the volatilities are parallel at the boundary of the set of nonnegative functions. Taking those conditions, we can also prove that the Stratonovich drift is inward pointing, since parallel volatilities produce parallel Stratonovich corrections (a fact which is not true for general closed convex sets but holds true for the set of nonnegative functions $P$ ). Then, we reduce the sufficiency proof to two steps: First, we essentially apply results from [34] in order to solve the pure diffusion case, and then we "slowly" switch on the jumps to see the general result.Let $H_{\beta}$ be the space of forward curves introduced in section 2 for some fixed $\beta>0$. We introduce the half-spaces

$$
H_{\xi}^{+}:=\left\{h \in H_{\beta} \mid h(\xi) \geq 0\right\}, \quad \xi \in \mathbb{R}_{+},
$$

and define the closed, convex cone

$$
P:=\bigcap_{\xi \in \mathbb{R}_{+}} H_{\xi}^{+}
$$

consisting of all nonnegative forward curves from $H_{\beta}$. In what follows, we shall use that, by the continuity of the functions from $H_{\beta}$, we can write $P$ as

$$
P=\bigcap_{\xi \in(0, \infty)} H_{\xi}^{+}
$$

Furthermore, we define the edges

$$
\partial P_{\xi}:=\{h \in P \mid h(\xi)=0\}, \quad \xi \in(0, \infty) .
$$

First, we consider the positivity problem for general forward curve evolutions, where the HJM drift condition (1.6) is not necessarily satisfied, and afterwards we apply our results to the arbitrage-free situation.

We emphasize that, in what follows, we assume the existence of solutions. Sufficient conditions for existence and uniqueness are provided in [20] (we also mention the related articles [1] and [29]) for general SPDEs and in the previous section 3 for the HJMM term structure equation (1.10).

As in the previous section, we work on the space $H_{\beta}$ of forward curves from section 2 for some $\beta>0$. At first glance, it looks reasonable to treat the positivity problem by working with weak solutions on $H_{\beta}$. However, this is unfeasible, because the point evaluations at $\xi \in(0, \infty)$, i.e., a linear functional $\zeta \in H_{\beta}$ such that $h(\xi)=\langle\zeta, h\rangle$ for all $h \in H_{\beta}$, never belong to the domain $\mathcal{D}\left(\left(\frac{d}{d \xi}\right)^{*}\right)$ of the adjoint operator. Indeed, a well-known mollifying technique shows that for each $\xi \in(0, \infty)$ the linear functional $h \mapsto h^{\prime}(\xi): \mathcal{D}\left(\frac{d}{d \xi}\right) \rightarrow \mathbb{R}$ is unbounded.

Therefore, treating the positivity problem with weak solutions does not bring an immediate advantage; hence we shall work with mild solutions on $H_{\beta}$.

Let measurable vector fields $\alpha: H_{\beta} \rightarrow H_{\beta}, \sigma: H_{\beta} \rightarrow L_{2}^{0}\left(H_{\beta}\right)$, and $\gamma: H_{\beta} \times E \rightarrow H_{\beta}$ be given. Currently, we do not assume that the drift term $\alpha$ is given by the HJM drift condition (1.6). For each $j$, we define $\sigma^{j}: H_{\beta} \rightarrow H_{\beta}$ as $\sigma^{j}(h):=\sqrt{\lambda_{j}} \sigma(h) e_{j}$. We assume that for each $h_{0} \in P$ the HJM equation

$$
\left\{\begin{align*}
d f_{t}= & U_{-t} \ell \alpha\left(\pi U_{t} f_{t}\right) d t+U_{-t} \ell \sigma\left(\pi U_{t} f_{t}\right) d W_{t}  \tag{4.1}\\
& +\int_{E} U_{-t} \ell \gamma\left(\pi U_{t} f_{t-}, x\right)(\mu(d t, d x)-F(d x) d t) \\
f_{0}= & \ell h_{0}
\end{align*}\right.
$$

has at least one $\mathcal{H}_{\beta}$-valued solution $\left(f_{t}\right)_{t \geq 0}$. Then, because of (2.7), the transformation $r_{t}:=$ $\pi U_{t} f_{t}, t \geq 0$, is a mild $H_{\beta}$-valued solution of the HJMM equation

$$
\left\{\begin{align*}
d r_{t} & =\left(\frac{d}{d \xi} r_{t}+\alpha\left(r_{t}\right)\right) d t+\sigma\left(r_{t}\right) d W_{t}+\int_{E} \gamma\left(r_{t-}, x\right)(\mu(d t, d x)-F(d x) d t)  \tag{4.2}\\
r_{0} & =h_{0}
\end{align*}\right.
$$

Definition 4.1. The HJMM equation (4.2) is said to be positivity preserving if for all $h_{0} \in$ $L^{2}\left(\Omega, \mathcal{F}_{0}, \mathbb{P} ; H_{\beta}\right)$ with $\mathbb{P}\left(h_{0} \in P\right)=1$ there exists a solution $\left(f_{t}\right)_{t \geq 0}$ of (4.1) with $f_{0}=\ell h_{0}$ such that $\mathbb{P}\left(\bigcap_{t \in \mathbb{R}_{+}}\left\{r_{t} \in P\right\}\right)=1$, where $r_{t}:=\pi U_{t} f_{t}, t \geq 0$.

Remark 4.1. Note that the seemingly weaker condition $\mathbb{P}\left(r_{t} \in P\right)=1$ for all $t \in \mathbb{R}_{+}$ is equivalent to the condition of the previous definition due to the càdlàg property of the trajectories.

Definition 4.2. The HJMM equation (4.2) is said to be locally positivity preserving if for all $h_{0} \in L^{2}\left(\Omega, \mathcal{F}_{0}, \mathbb{P} ; H_{\beta}\right)$ with $\mathbb{P}\left(h_{0} \in P\right)=1$ there exists a solution $\left(f_{t}\right)_{t \geq 0}$ of (4.1) with $f_{0}=\ell h_{0}$ and a strictly positive stopping time $\tau$ such that $\mathbb{P}\left(\bigcap_{t \in \mathbb{R}_{+}}\left\{r_{t \wedge \tau} \in P\right\}\right)=1$, where $r_{t}:=\pi U_{t} f_{t}$, $t \geq 0$.

Lemma 4.3. Let $h_{0} \in P$ be arbitrary, and let $\left(f_{t}\right)_{t \geq 0}$ be a solution for (4.1) with $f_{0}=\ell h_{0}$. Set $r_{t}:=\pi U_{t} f_{t}, t \geq 0$. The following two statements are equivalent:

1. We have $\mathbb{P}\left(\bigcap_{t \in \mathbb{R}_{+}}\left\{r_{t} \in P\right\}\right)=1$.
2. We have $\mathbb{P}\left(\bigcap_{t \in[0, T]}\left\{f_{t}(T) \geq 0\right\}\right)=1$ for all $T \in(0, \infty)$.

Proof. The claim follows, because the processes $\left(r_{t}\right)_{t \geq 0}$ and $\left(f_{t}(T)\right)_{t \in[0, T]}$ for an arbitrary $T \in(0, \infty)$ are càdlàg and because the functions from $H_{\beta}$ are continuous.

Assumption 4.2. We assume that the vector fields $\alpha: H_{\beta} \rightarrow H_{\beta}$ and $\sigma: H_{\beta} \rightarrow L_{2}^{0}\left(H_{\beta}\right)$ are continuous and that $h \mapsto \int_{B} \gamma(h, x) F(d x)$ is continuous on $H_{\beta}$ for all $B \in \mathcal{E}$ with $F(B)<\infty$.

Remark 4.3. Notice that, by Hölder's inequality, Assumption 4.2 is implied by Assumptions 4.6 and 4.7 and therefore in particular by Assumption 3.1.

Proposition 4.4. Suppose Assumption 4.2 is fulfilled. If (4.2) is positivity preserving, then we have

$$
\begin{align*}
& \int_{E} \gamma(h, x)(\xi) F(d x)<\infty \quad \text { for all } \xi \in(0, \infty), h \in \partial P_{\xi},  \tag{4.3}\\
& \alpha(h)(\xi)-\int_{E} \gamma(h, x)(\xi) F(d x) \geq 0 \quad \text { for all } \xi \in(0, \infty), h \in \partial P_{\xi},  \tag{4.4}\\
& \sigma^{j}(h)(\xi)=0 \quad \text { for all } \xi \in(0, \infty), h \in \partial P_{\xi} \text { and all } j,  \tag{4.5}\\
& h+\gamma(h, x) \in P \quad \text { for all } h \in P \text { and } F \text {-almost all } x \in E . \tag{4.6}
\end{align*}
$$

Remark 4.4. Observe that condition (4.6) implies

$$
\begin{equation*}
\gamma(h, x)(\xi) \geq 0 \quad \text { for all } \xi \in(0, \infty), h \in \partial P_{\xi} \text { and } F \text {-almost all } x \in E \text {. } \tag{4.7}
\end{equation*}
$$

Therefore, condition (4.3) is equivalent to

$$
\int_{E}|\gamma(h, x)(\xi)| F(d x)<\infty \quad \text { for all } \xi \in(0, \infty), h \in \partial P_{\xi}
$$

Consequently, conditions (4.3) and (4.4) can be unified to

$$
\int_{E}|\gamma(h, x)(\xi)| F(d x) \leq \alpha(h)(\xi)
$$

for all $\xi \in(0, \infty)$ and $h \in \partial P_{\xi}$.
Proof. Let $h_{0} \in P$ be arbitrary, and let $\left(f_{t}\right)_{t \geq 0}$ be a solution for (4.1) with $f_{0}=\ell h_{0}$ such that $\mathbb{P}\left(\bigcap_{t \in \mathbb{R}_{+}}\left\{r_{t} \in P\right\}\right)=1$, where $r_{t}:=\pi U_{t} f_{t}, t \geq 0$. By Lemma 4.3, for each $T \in(0, \infty)$ and every stopping time $\tau \leq T$ we have

$$
\begin{equation*}
\mathbb{P}\left(f_{\tau}(T) \geq 0\right)=1 \tag{4.8}
\end{equation*}
$$

Let $\phi \in U_{0}^{\prime}$ be a linear functional such that $\phi^{j}:=\phi e_{j} \neq 0$ for only finitely many $j$, and let $\psi: E \rightarrow \mathbb{R}$ be a measurable function of the form $\psi=c \mathbb{1}_{B}$ with $c>-1$ and $B \in \mathcal{E}$ satisfying $F(B)<\infty$. Let $Z$ be the Doléans-Dade exponential

$$
Z_{t}=\varepsilon\left(\sum_{j} \phi^{j} \beta^{j}+\int_{0}^{\bullet} \int_{E} \psi(x)(\mu(d s, d x)-F(d x) d s)\right)_{t}, \quad t \geq 0
$$

By [25, Thm. I.4.61], the process $Z$ is a solution of

$$
Z_{t}=1+\sum_{j} \phi^{j} \int_{0}^{t} Z_{s} d \beta_{s}^{j}+\int_{0}^{t} \int_{E} Z_{s-} \psi(x)(\mu(d s, d x)-F(d x) d s), \quad t \geq 0
$$

and, since $\psi>-1$, the process $Z$ is a strictly positive local martingale. There exists a strictly positive stopping time $\tau_{1}$ such that $Z^{\tau_{1}}$ is a martingale. Due to the method of the moving frame (see [20]), we can use standard stochastic analysis to proceed further. For an arbitrary $T \in(0, \infty)$, integration by parts yields (see [25, Thm. I.4.52])

$$
\begin{align*}
f_{t}(T) Z_{t}= & \int_{0}^{t} f_{s-}(T) d Z_{s}+\int_{0}^{t} Z_{s-} d f_{s}(T)+\left\langle f(T)^{c}, Z^{c}\right\rangle_{t}  \tag{4.9}\\
& +\sum_{s \leq t} \Delta f_{s}(T) \Delta Z_{s}, \quad t \geq 0
\end{align*}
$$

Taking into account the dynamics

$$
\begin{align*}
f_{t}(T)= & \ell h_{0}(T)+\int_{0}^{t} U_{-s} \ell \alpha\left(\pi U_{s} f_{s}\right)(T) d s+\sum_{j} \int_{0}^{t} U_{-s} \ell \sigma^{j}\left(\pi U_{s} f_{s}\right)(T) d \beta_{s}^{j}  \tag{4.10}\\
& +\int_{0}^{t} \int_{E} U_{-s} \ell \gamma\left(\pi U_{s} f_{s-}, x\right)(T)(\mu(d s, d x)-F(d x) d s), \quad t \geq 0
\end{align*}
$$

we have

$$
\begin{align*}
\left\langle f(T)^{c}, Z^{c}\right\rangle_{t} & =\sum_{j} \phi^{j} \int_{0}^{t} Z_{s} U_{-s} \ell \sigma^{j}\left(\pi U_{s} f_{s}\right)(T) d s, \quad t \geq 0,  \tag{4.11}\\
\sum_{s \leq t} \Delta f_{s}(T) \Delta Z_{s} & =\int_{0}^{t} \int_{E} Z_{s-} \psi(x) U_{-s} \ell \gamma\left(\pi U_{s} f_{s-}, x\right)(T) \mu(d s, d x), \quad t \geq 0 . \tag{4.12}
\end{align*}
$$

Incorporating (4.10), (4.11), and (4.12) into (4.9), we obtain

$$
\begin{align*}
f_{t}(T) Z_{t}=M_{t}+\int_{0}^{t} Z_{s-} & \left(U_{-s} \ell \alpha\left(\pi U_{s} f_{s-}\right)(T)+\sum_{j} \phi^{j} U_{-s} \ell \sigma^{j}\left(\pi U_{s} f_{s-}\right)(T)\right.  \tag{4.13}\\
& \left.+\int_{E} \psi(x) U_{-s} \ell \gamma\left(\pi U_{s} f_{s-}, x\right)(T) F(d x)\right) d s, \quad t \geq 0,
\end{align*}
$$

where $M$ is a local martingale with $M_{0}=0$. There exists a strictly positive stopping time $\tau_{2}$ such that $M^{\tau_{2}}$ is a martingale.

By Assumption 4.2, there exist a strictly positive stopping time $\tau_{3}$ and a constant $\tilde{\alpha}>0$ such that

$$
\left|U_{-\left(t \wedge \tau_{3}\right)} \ell \alpha\left(\pi U_{t \wedge \tau_{3}} f_{\left(t \wedge \tau_{3}\right)-}\right)(T)\right| \leq \tilde{\alpha}, \quad t \geq 0
$$

Let $B:=\left\{x \in E: h_{0}+\gamma\left(h_{0}, x\right) \notin P\right\}$. In order to prove (4.6), it suffices, since $F$ is $\sigma$-finite, to show that $F(B \cap C)=0$ for all $C \in \mathcal{E}$ with $F(C)<\infty$. Suppose, on the contrary, there exists $C \in \mathcal{E}$ with $F(C)<\infty$ such that $F(B \cap C)>0$. By the continuity of the functions from $H_{\beta}$, there exists $T \in(0, \infty)$ such that $F\left(B_{T} \cap C\right)>0$, where $B_{T}:=\left\{x \in E: h_{0}(T)+\gamma\left(h_{0}, x\right)(T)<\right.$ $0\}$. We obtain

$$
\int_{B_{T} \cap C} \gamma\left(h_{0}, x\right)(T) F(d x) \leq \int_{B_{T} \cap C}\left(h_{0}(T)+\gamma\left(h_{0}, x\right)(T)\right) F(d x)<0 .
$$

By Assumption 4.2 and the left continuity of the process $f_{-}$, there exist $\eta>0$ and a strictly positive stopping time $\tau_{4} \leq T$ such that

$$
\int_{B_{T} \cap C} U_{-\left(t \wedge \tau_{4}\right)} \ell \gamma\left(\pi U_{\left(t \wedge \tau_{4}\right)} f_{\left(t \wedge \tau_{4}\right)-}, x\right)(T) F(d x) \leq-\eta, \quad t \geq 0
$$

Let $\phi:=0, \psi:=\frac{\tilde{\alpha}+1}{\eta} \mathbb{1}_{B_{T} \cap C}$, and $\tau:=\bigwedge_{i=1}^{4} \tau_{i}$. Taking expectation in (4.13), we obtain $\mathbb{E}\left[f_{\tau}(T) Z_{\tau}\right]<0$, implying $\mathbb{P}\left(f_{\tau}(T)<0\right)>0$, which contradicts (4.8). This yields (4.6).

From now on, we assume that $h_{0} \in \partial P_{T}$ for an arbitrary $T \in(0, \infty)$.
Suppose that $\sigma^{j}\left(h_{0}\right)(T) \neq 0$ for some $j$. By the continuity of $\sigma$ (see Assumption 4.2), there exist $\eta>0$ and a strictly positive stopping time $\tau_{4} \leq T$ such that

$$
\left|U_{-\left(t \wedge \tau_{4}\right)} \ell \sigma^{j}\left(\pi U_{t \wedge \tau_{4}} f_{\left(t \wedge \tau_{4}\right)-}\right)(T)\right| \geq \eta, \quad t \geq 0
$$

Let $\phi \in U_{0}^{\prime}$ be the linear functional given by $\phi^{j}=-\operatorname{sign}\left(\sigma^{j}\left(h_{0}\right)(T)\right) \frac{\tilde{\alpha}+1}{\eta}$, and $\phi^{k}=0$ for $k \neq j$. Furthermore, let $\psi:=0$ and $\tau:=\bigwedge_{i=1}^{4} \tau_{i}$. Taking expectation in (4.13) yields $\mathbb{E}\left[f_{\tau}(T) Z_{\tau}\right]<0$, implying $\mathbb{P}\left(f_{\tau}(T)<0\right)>0$, which contradicts (4.8). This proves (4.5).

Now suppose $\int_{E} \gamma\left(h_{0}, x\right)(T) F(d x)=\infty$. Using Assumption 4.2, relation (4.7), and the $\sigma$-finiteness of $F$, there exist $B \in \mathcal{E}$ with $F(B)<\infty$ and a strictly positive stopping time $\tau_{4} \leq T$ such that

$$
-\frac{1}{2} \int_{B} U_{-\left(t \wedge \tau_{4}\right)} \ell \gamma\left(\pi U_{t \wedge \tau_{4}} f_{\left(t \wedge \tau_{4}\right)-}, x\right)(T) F(d x) \leq-(\tilde{\alpha}+1), \quad t \geq 0
$$

Let $\phi:=0, \psi:=-\frac{1}{2} \mathbb{1}_{B}$, and $\tau:=\bigwedge_{i=1}^{4} \tau_{i}$. Taking expectation in (4.13), we obtain $\mathbb{E}\left[f_{\tau}(T) Z_{\tau}\right]<$ 0 , implying $\mathbb{P}\left(f_{\tau}(T)<0\right)>0$, which contradicts (4.8). This yields (4.3).

Since $F$ is $\sigma$-finite, there exists a sequence $\left(B_{n}\right)_{n \in \mathbb{N}} \subset \mathcal{E}$ with $B_{n} \uparrow E$ and $F\left(B_{n}\right)<\infty$, $n \in \mathbb{N}$. Next, we show for all $n \in \mathbb{N}$ the relation

$$
\begin{equation*}
\alpha\left(h_{0}\right)(T)+\int_{E} \psi_{n}(x) \gamma\left(h_{0}, x\right)(T) F(d x) \geq 0 \tag{4.14}
\end{equation*}
$$

where $\psi_{n}:=-\left(1-\frac{1}{n}\right) \mathbb{1}_{B_{n}}$. Suppose, on the contrary, that (4.14) is not satisfied for some $n \in \mathbb{N}$. Using Assumption 4.2, there exist $\eta>0$ and a strictly positive stopping time $\tau_{4} \leq T$ such that

$$
\begin{aligned}
& U_{-\left(t \wedge \tau_{4}\right)} \ell \alpha\left(\pi U_{t} f_{\left(t \wedge \tau_{4}\right)-}\right)(T) \\
& +\int_{E} \psi_{n}(x) U_{-\left(t \wedge \tau_{4}\right)} \ell \gamma\left(\pi U_{t \wedge \tau_{4}} f_{\left(t \wedge \tau_{4}\right)-}, x\right)(T) F(d x) \leq-\eta, \quad t \geq 0 .
\end{aligned}
$$

Let $\phi:=0$ and $\tau:=\bigwedge_{i=1}^{4} \tau_{i}$. Taking expectation in (4.13), we obtain $\mathbb{E}\left[f_{\tau}(T) Z_{\tau}\right]<0$, implying $\mathbb{P}\left(f_{\tau}(T)<0\right)>0$, which contradicts (4.8). This yields (4.14). By (4.14), (4.3), and Lebesgue's theorem, we get (4.4).

Remark 4.5. The Cox-Ingersoll-Ross (CIR) model [7] is celebrated for its feature of producing nonnegative interest rates. At this point, it is worth pointing out that the HullWhite extension of the CIR model (HWCIR) is not positivity preserving. Indeed, in Musiela
parametrization its term structure dynamics are given by

$$
\left\{\begin{aligned}
d r_{t} & =\left(\frac{d}{d \xi} r_{t}+\alpha_{\mathrm{HJM}}\left(r_{t}\right)\right) d t+\sigma\left(r_{t}\right) d W_{t} \\
r_{0} & =h_{0}
\end{aligned}\right.
$$

where $W$ is a one-dimensional Wiener process and the vector fields $\alpha_{\mathrm{HJM}}, \sigma: H_{\beta} \rightarrow H_{\beta}^{0}$ are defined as

$$
\begin{aligned}
\alpha_{\mathrm{HJM}}(h) & :=\rho^{2}|h(0)| \lambda \Lambda, & h \in H_{\beta}, \\
\sigma(h) & :=\rho \sqrt{|h(0)|} \lambda, & h \in H_{\beta},
\end{aligned}
$$

where $\rho>0$ is a constant, $h \mapsto h(0): H_{\beta} \rightarrow \mathbb{R}$ denotes the evaluation of the short rate, and $\lambda \in H_{\beta}^{0}$ is a function with $\lambda(x)>0$ for all $x \geq 0$ such that $\Lambda=\int_{0}^{\bullet} \lambda(\eta) d \eta$ satisfies a certain Riccati equation; see [21, sect. 6.2] for more details. Note that Assumption 4.2 is satisfied, because the vector fields $\alpha_{\mathrm{HJM}}, \sigma: H_{\beta} \rightarrow H_{\beta}^{0}$ are continuous, but condition (4.5) from Proposition 4.4 does not hold, because $\sigma$ depends only on the current state of the short rate. Hence, the HWCIR model cannot be positivity preserving.

Indeed, we can also verify this directly as follows. According to [21, sect. 6.2], for any initial curve $h_{0} \in H_{\beta}$ the short rate $R_{t}=r_{t}(0), t \geq 0$, has the dynamics

$$
\left\{\begin{align*}
d R_{t} & =\left(b(t)-c\left|R_{t}\right|\right) d t+\rho \sqrt{\left|R_{t}\right|} d W_{t}  \tag{4.15}\\
R_{0} & =h_{0}(0)
\end{align*}\right.
$$

for some constant $c \in \mathbb{R}$ and a time-dependent function $b=b\left(h_{0}\right): \mathbb{R}_{+} \rightarrow \mathbb{R}$. Due to [40], for each starting point $h_{0}(0) \in \mathbb{R}$ the $\operatorname{SDE}$ (4.15) has a unique strong solution, and, according to [18], this solution is nonnegative if and only if $h_{0}(0) \geq 0$ and $b(t) \geq 0$ for all $t \in \mathbb{R}_{+}$. By [17, Prop. 5.2], the forward rates are given by

$$
f(t, T)=\int_{t}^{T} b(s) \lambda(T-s) d s+\lambda(T-t) R_{t}
$$

which implies for the initial forward curve

$$
h_{0}(T)=\int_{0}^{T} b(s) \lambda(T-s) d s+\lambda(T) h_{0}(0), \quad T \geq 0
$$

Having in mind that $\lambda(t)>0$ for all $t \geq 0$, we see that for certain nonnegative initial curves $h_{0} \in P$ the function $b$ can also reach negative values, which yields negative short rates. For example, take an initial curve $h_{0} \in P$ with $h_{0}(0)>0$ and $h_{0}(T)=0$ for some $T>0$.

We shall now present sufficient conditions for positivity preserving term structure models. In what follows, we will require the following linear growth and Lipschitz conditions.

Assumption 4.6. We assume that $\int_{E}\|\gamma(0, x)\|_{\beta}^{2} F(d x)<\infty$ and that there is a constant $K>0$ such that

$$
\left(\int_{E}\left\|\gamma\left(h_{1}, x\right)-\gamma\left(h_{2}, x\right)\right\|_{\beta}^{2} F(d x)\right)^{1 / 2} \leq K\left\|h_{1}-h_{2}\right\|_{\beta}
$$

for all $h_{1}, h_{2} \in H_{\beta}$.
Assumption 4.7. We assume there is a constant $L>0$ such that

$$
\begin{aligned}
\left\|\alpha\left(h_{1}\right)-\alpha\left(h_{2}\right)\right\|_{\beta} & \leq L\left\|h_{1}-h_{2}\right\|_{\beta}, \\
\left\|\sigma\left(h_{1}\right)-\sigma\left(h_{2}\right)\right\|_{L_{2}^{0}\left(H_{\beta}\right)} & \leq L\left\|h_{1}-h_{2}\right\|_{\beta}
\end{aligned}
$$

for all $h_{1}, h_{2} \in H_{\beta}$.
This ensures existence and uniqueness of solutions by [20, Cor. 10.9].
Lemma 4.5. Suppose Assumptions 4.6 and 4.7 are fulfilled, and suppose for each $h_{0} \in P$ we have $\mathbb{P}\left(\bigcap_{t \in \mathbb{R}_{+}}\left\{r_{t} \in P\right\}\right)=1$, where $\left(r_{t}\right)_{t \geq 0}$ denotes the mild solution for (4.2) with $r_{0}=h_{0}$. Then, the HJMM equation (4.2) is positivity preserving.

Proof. Let $h_{0} \in L^{2}\left(\Omega, \mathcal{F}_{0}, \mathbb{P} ; H_{\beta}\right)$ with $\mathbb{P}\left(h_{0} \in P\right)=1$ be arbitrary. There exists a sequence $\left(h_{n}\right)_{n \in \mathbb{N}} \subset L^{2}\left(\Omega, \mathcal{F}_{0}, \mathbb{P} ; H_{\beta}\right)$ such that $h_{n} \rightarrow h_{0}$ in $L^{2}\left(\Omega, \mathcal{F}_{0}, \mathbb{P} ; H_{\beta}\right)$ and such that for each $n \in \mathbb{N}$ we have $\mathbb{P}\left(h_{n} \in P\right)=1$ and $h_{n}$ has only a finite number of values. By assumption, we have $\mathbb{P}\left(\bigcap_{t \in \mathbb{R}_{+}}\left\{r_{t}^{n} \in P\right\}\right)=1$ for all $n \in \mathbb{N}$. Applying [20, Prop. 9.1] yields

$$
\mathbb{E}\left[\sup _{t \in[0, T]}\left\|r_{t}-r_{t}^{n}\right\|_{\beta}^{2}\right] \rightarrow 0 \quad \text { for all } T \in \mathbb{R}_{+}
$$

showing that $\mathbb{P}\left(\bigcap_{t \in \mathbb{R}_{+}}\left\{r_{t} \in P\right\}\right)=1$.
In view of Lemma 4.7, we prepare an auxiliary result, which is proved in the appendix.
Lemma 4.6. Let $\tau$ be a bounded stopping time. We define the new filtration $\left(\tilde{\mathcal{F}}_{t}\right)_{t \geq 0}$ by $\tilde{\mathcal{F}}_{t}:=\mathcal{F}_{\tau+t}$, the new $U$-valued process $\tilde{W}$ by $\tilde{W}_{t}:=W_{\tau+t}-W_{\tau}$, and the new random measure $\tilde{\mu}$ on $\mathbb{R}_{+} \times E$ by $\tilde{\mu}(\omega ; B):=\mu\left(\omega ; B_{\tau(\omega)}\right), B \in \mathcal{B}\left(\mathbb{R}_{+}\right) \otimes \mathcal{E}$, where

$$
B_{\tau}:=\left\{(t+\tau, x) \in \mathbb{R}_{+} \times E:(t, x) \in B\right\}
$$

Then, $\tilde{W}$ is a $Q$-Wiener process with respect to $\left(\tilde{\mathcal{F}}_{t}\right)_{t \geq 0}$ and $\tilde{\mu}$ is a homogeneous Poisson random measure on $\mathbb{R}_{+} \times E$ with respect to $\left(\tilde{\mathcal{F}}_{t}\right)_{t \geq 0}$ having the compensator $d t \otimes F(d x)$. Moreover, we have the expansion

$$
\begin{equation*}
\tilde{W}=\sum_{j} \sqrt{\lambda_{j}} \tilde{\beta}^{j} e_{j} \tag{4.16}
\end{equation*}
$$

where $\tilde{\beta}^{j}$ defined as $\tilde{\beta}_{t}^{j}:=\beta_{\tau+t}^{j}-\beta_{\tau}^{j}$ is a sequence of real-valued independent $\left(\tilde{\mathcal{F}}_{t}\right)$-Brownian motions. Furthermore, if $\left(r_{t}\right)_{t \geq 0}$ is a weak solution for (4.2), then the $\left(\tilde{\mathcal{F}}_{t}\right)$-adapted process $\left(\tilde{r}_{t}\right)_{t \geq 0}$ defined by $\tilde{r}_{t}:=r_{\tau+t}$ is a weak solution for

$$
\left\{\begin{align*}
d \tilde{r}_{t} & =\left(\frac{d}{d \xi} \tilde{r}_{t}+\alpha\left(\tilde{r}_{t}\right)\right) d t+\sigma\left(\tilde{r}_{t}\right) d \tilde{W}_{t}+\int_{E} \gamma\left(\tilde{r}_{t-}, x\right)(\tilde{\mu}(d t, d x)-F(d x) d t)  \tag{4.17}\\
\tilde{r}_{0} & =r_{\tau}
\end{align*}\right.
$$

Proof. See the appendix.
Lemma 4.7. Suppose Assumptions 4.6 and 4.7 are fulfilled. If (4.2) is locally positivity preserving and we have (4.6), then (4.2) is positivity preserving.

Proof. Let $h_{0} \in L^{2}\left(\Omega, \mathcal{F}_{0}, \mathbb{P} ; H_{\beta}\right)$ with $\mathbb{P}\left(h_{0} \in P\right)=1$ be arbitrary. Moreover, let $\left(r_{t}\right)_{t \geq 0}$ be the mild solution for (4.2) with $r_{0}=h_{0}$. We define the stopping time

$$
\begin{equation*}
\tau_{0}:=\inf \left\{t \geq 0: r_{t} \notin P\right\} . \tag{4.18}
\end{equation*}
$$

By the closedness of $P$ and (4.6), we have $r_{\tau_{0}} \in P$ almost surely on $\left\{\tau_{0}<\infty\right\}$. We claim that $\mathbb{P}\left(\tau_{0}=\infty\right)=1$. Assume, on the contrary, that

$$
\begin{equation*}
\mathbb{P}\left(\tau_{0}<N\right)>0 \tag{4.19}
\end{equation*}
$$

for some $N \in \mathbb{N}$. Let $\tau$ be the bounded stopping time $\tau:=\tau_{0} \wedge N$. We define the new filtration $\left(\tilde{\mathcal{F}}_{t}\right)_{t \geq 0}$, the new $Q$-Wiener process $\tilde{W}$, and the new Poisson random measure $\tilde{\mu}$ as in Lemma 4.6. Note that $r_{\tau} \in L^{2}\left(\Omega, \tilde{F}_{0}, \mathbb{P} ; H_{\beta}\right)$, because, by (3.16), we have

$$
\mathbb{E}\left[\left\|r_{\tau}\right\|_{\beta}^{2}\right] \leq \mathbb{E}\left[\sup _{t \in[0, N]}\left\|r_{t}\right\|_{\beta}^{2}\right]<\infty .
$$

By Lemma 4.6, the $\left(\tilde{\mathcal{F}}_{t}\right)$-adapted process $\tilde{r}_{t}:=r_{\tau+t}$ is the unique mild solution for (4.17). Since (4.2) is locally positivity preserving and $\mathbb{P}\left(r_{\tau} \in P\right)=1$, there exists a strictly positive stopping time $\tau_{1}$ such that $\mathbb{P}\left(\bigcap_{t \in \mathbb{R}_{+}} \tilde{r}_{t \wedge \tau_{1}} \in P\right)=1$. Since $\left\{\tau_{0}<N\right\} \subset\left\{\tau_{0}=\tau\right\}$, we obtain

$$
r_{\tau_{0}+t} \in P \quad \text { almost surely on }\left[0, \tau_{1}\right] \cap\left\{\tau_{0}<N\right\}
$$

which is a contradiction because of (4.19) and definition (4.18) of $\tau_{0}$. Consequently, we have $\mathbb{P}\left(\tau_{0}=\infty\right)=1$, whence (4.2) is positivity preserving.

Assumption 4.8. We assume that $\sigma \in C^{2}\left(H_{\beta} ; L_{2}^{0}\left(H_{\beta}\right)\right)$, and that the vector field

$$
\begin{equation*}
h \mapsto \sum_{j} D \sigma^{j}(h) \sigma^{j}(h) \tag{4.20}
\end{equation*}
$$

is globally Lipschitz on $H_{\beta}$.
Remark 4.9. Note that Assumption 4.8 is satisfied if $\sigma \in C_{b}^{2}\left(H_{\beta} ; L_{2}^{0}\left(H_{\beta}\right)\right)$ and the series (4.20) converges for every $h \in H_{\beta}$.

Lemma 4.8. Suppose Assumption 4.8 and relation (4.5) are fulfilled. Then we have

$$
\left(\sum_{j} D \sigma^{j}(h) \sigma^{j}(h)\right)(\xi)=0 \quad \text { for all } \xi \in(0, \infty), h \in \partial P_{\xi}
$$

Proof. Let $\xi \in(0, \infty)$ be arbitrary. It suffices to show $\left(D \sigma^{j}(h) \sigma^{j}(h)\right)(\xi)=0$ for all $h \in \partial P_{\xi}$ and all $j$. Therefore, let $j$ be fixed and denote $\sigma=\sigma^{j}$. By assumption for all $h \geq 0$ with $h(\xi)=0$, we have that $\sigma(h)(\xi)=0$. In other words, the volatility vector field $\sigma$ is parallel to the boundary at boundary elements of $P$. We denote the local flow of the Lipschitz vector field $\sigma$ by Fl , which is defined on a small time interval $]-\epsilon, \epsilon[$ around time 0 and a small neighborhood of each element $h \in P$. We first state that the flow Fl leaves the set $P$ invariant, i.e., $\mathrm{Fl}_{t}(h) \geq 0$ if $h \geq 0$, by convexity and closedness of the cone of positive functions due to [39]. Indeed, $P$ is a closed and convex cone, whose supporting hyperplanes $l$
(a linear functional $l$ is called a supporting hyperplane of $P$ at $h$ if $l(P) \geq 0$ and $l(h)=0)$ are given by appropriate positive measures $\mu$ on $\mathbb{R}_{+}$via

$$
l(h)=\int_{\mathbb{R}_{+}} h(\xi) \mu(d \xi),
$$

whence condition (4) from [39] is fulfilled due to (4.5). Next we show that even more holds: the solution $\mathrm{Fl}_{t}(h)$ evaluated at $\xi$ vanishes if $h(\xi)=0$, which we show directly. Indeed, let us additionally fix $h \in \partial P_{\xi}$, i.e., $h \geq 0$ and $h(\xi)=0$. Looking now at the Picard-Lindelöf approximation scheme

$$
c^{(n+1)}(t)=h+\int_{0}^{t} \sigma\left(c^{(n)}(s)\right) d s
$$

with $c^{(n)}(0)=h$ and $c^{(0)}(s)=h$ for $\left.s, t \in\right]-\epsilon, \epsilon[$ and $n \geq 0$, we see by induction that under our assumptions

$$
c^{(n)}(t)(\xi)=0
$$

for all $n \geq 0$ and $t \in]-\epsilon, \epsilon[$ for the given fixed element $h$. Consequently-as $n \rightarrow \infty$-we obtain that $\mathrm{Fl}_{t}(h)(\xi)=0$, which is the limit of $c^{(n)}(t)$. Therefore,

$$
(D \sigma(h) \sigma(h))(\xi)=\left.\frac{d}{d s}\right|_{s=0} \sigma\left(\mathrm{Fl}_{s}(h)\right)(\xi)=0,
$$

since $\mathrm{Fl}_{t}(h) \geq 0$ by invariance and $\mathrm{Fl}_{t}(h)(\xi)=0$ by the previous consideration lead to $\sigma\left(\mathrm{Fl}_{t}(h)\right)(\xi)=0$ for $\left.t \in\right]-\epsilon, \epsilon[$. Notice that we did not need the global Lipschitz property of the Stratonovich correction for the proof of this lemma.

Before we show sufficiency for the HJMM equation (4.2) with jumps, we consider the pure diffusion case. Notice that, due to Lemma 4.8, the condition (4.4) is in fact equivalent to the very same condition formulated with the Stratonovich drift $\sigma^{0}$, defined in (4.21), instead of $\alpha$, since the Stratonovich correction vanishes at the boundary of $P$.

In order to treat the pure diffusion case, we apply [34], which, by using the support theorem provided in [33], offers a general characterization of stochastic invariance of closed sets for SPDEs.

Other results for positivity preserving SPDEs, where, in contrast to our framework, the state space is an $L^{2}$-space, can be found in [27] and [30]. The results from [30] have been used in [35] in order to derive some positivity results for Lévy term structure models on $L^{2}$-spaces.

Proposition 4.9. Suppose Assumptions 4.7 and 4.8 are fulfilled and $\gamma \equiv 0$. If conditions (4.4) and (4.5) are satisfied, then (4.2) is positivity preserving.

Proof. In view of Lemma 4.5, it suffices to show that for all $h_{0} \in P$ we have $\mathbb{P}\left(\bigcap_{t \in \mathbb{R}_{+}}\left\{r_{t} \in\right.\right.$ $P\})=1$, where $\left(r_{t}\right)_{t \geq 0}$ denotes the mild solution for (4.2) with $r_{0}=h_{0}$. Moreover, we may assume that the vector fields $\alpha, \sigma$ and

$$
\begin{equation*}
h \mapsto \sigma^{0}(h):=\alpha(h)-\frac{1}{2} \sum_{j} D \sigma^{j}(h) \sigma^{j}(h) \tag{4.21}
\end{equation*}
$$

are bounded in order to apply Nakayama's beautiful support theorem from [34]. Indeed, for $n \in \mathbb{N}$ we choose a bump function $\psi_{n} \in C^{\infty}\left(H_{\beta} ;[0,1]\right)$ such that $\psi_{n} \equiv 1$ on $\overline{B_{n}(0)}$ and
$\operatorname{supp}\left(\psi_{n}\right) \subset B_{n+1}(0)$ and define the vector fields

$$
\begin{aligned}
\alpha_{n}(h):=\psi_{n}(h) \alpha(h), & h \in H_{\beta}, \\
\sigma_{n}(h):=\psi_{n}(h) \sigma(h), & h \in H_{\beta} .
\end{aligned}
$$

These vector fields and

$$
h \mapsto \sigma_{n}^{0}(h):=\alpha_{n}(h)-\frac{1}{2} \sum_{j} D \sigma_{n}^{j}(h) \sigma_{n}^{j}(h)
$$

are bounded by the Lipschitz continuity of $\alpha, \sigma$, and $\sigma^{0}$, and Assumptions 4.7 and 4.8 as well as conditions (4.4) and (4.5) are again satisfied.

Now, we show that the semigroup Nagumo condition (3) from [33, Prop. 1.1] is fulfilled due to conditions (4.4) and (4.5). Introducing the distance $d_{P}(h)$ from $P$ as the minimal distance of $h \in H_{\beta}$ from $P$, we can formulate Nagumo's condition as

$$
\begin{equation*}
\liminf _{t \downarrow 0} \frac{1}{t} d_{P}\left(S_{t} h+t \sigma^{0}(h)+t \sigma(h) u\right)=0 \tag{4.22}
\end{equation*}
$$

for all $u \in U_{0}$ and $h \in P$. Now fix $h \in P$ and $u \in U_{0}$ and introduce the abbreviation $\tilde{\sigma}=\sigma^{0}+\sigma(\cdot) u$; then we obviously have

$$
\left\|S_{t} h+t \sigma^{0}(h)+t \sigma(h) u-S_{t} \mathrm{Fl}_{t}^{\tilde{\sigma}}(h)\right\|_{\beta}=t\left\|\tilde{\sigma}(h)-S_{t} \frac{\mathrm{Fl}_{t}^{\tilde{\sigma}}(h)-h}{t}\right\|_{\beta}, \quad t>0,
$$

which means that

$$
\lim _{t \downarrow 0} \frac{1}{t}\left\|S_{t} h+t \sigma^{0}(h)+t \sigma(h) u-S_{t} \mathrm{Fl}_{t}^{\tilde{\sigma}}(h)\right\|_{\beta}=0 .
$$

Hence, Nagumo's condition (4.22) can equivalently be formulated as

$$
\begin{equation*}
\liminf _{t \downarrow 0} \frac{1}{t} d_{P}\left(S_{t} \mathrm{Fl}_{t}^{\sigma}(h)\right)=0 \tag{4.23}
\end{equation*}
$$

for the particular choice of $u \in U_{0}$ and $h \in P$, since the shortest distance projector onto $P$ is a Lipschitz continuous map. Due to conditions (4.4), (4.5) and Lemma 4.8, the semiflow Fl ${ }^{\tilde{\sigma}}$ leaves $P$ invariant by [39], the semigroup $\left(S_{t}\right)_{t \geq 0}$ certainly, too, and therefore we have $d_{P}\left(S_{t} \mathrm{Fl}_{t}^{\tilde{\sigma}}(h)\right)=0, t \geq 0$, whence Nagumo's condition (4.23) is more than satisfied.

Finally, the next result states the sufficient conditions under which we can conclude that (4.2) is positivity preserving.

Proposition 4.10. Suppose Assumptions 4.6, 4.7, and 4.8 and conditions (4.3)-(4.6) are fulfilled. Then, (4.2) is positivity preserving.

Proof. Since the measure $F$ is $\sigma$-finite, there exists a sequence $\left(B_{n}\right)_{n \in \mathbb{N}} \subset \mathcal{E}$ with $B_{n} \uparrow E$ and $F\left(B_{n}\right)<\infty$ for all $n \in \mathbb{N}$. Let $h_{0} \in L^{2}\left(\Omega, \mathcal{F}_{0}, \mathbb{P} ; H_{\beta}\right)$ be arbitrary. Relations (4.4), (4.7), (4.5), Proposition 4.9, and (4.6) together with the closedness of $P$ yield that, for each $n \in \mathbb{N}$, the mild solution $\left(r_{t}^{n}\right)_{t \geq 0}$ of the SPDE

$$
\left\{\begin{align*}
d r_{t}^{n}= & \left(\frac{d}{d \xi} r_{t}^{n}+\alpha\left(r_{t}^{n}\right)-\int_{B_{n}} \gamma\left(r_{t}^{n}, x\right) F(d x)\right) d t+\sigma\left(r_{t}^{n}\right) d W_{t}  \tag{4.24}\\
& +\int_{B_{n}} \gamma\left(r_{t-}^{n}, x\right) \mu(d t, d x) \\
r_{0}^{n}= & h_{0}
\end{align*}\right.
$$

satisfies $\mathbb{P}\left(\bigcap_{t \in \mathbb{R}_{+}} r_{t \wedge \tau}^{n} \in P\right)=1$, where $\tau$ denotes the strictly positive stopping time

$$
\tau:=\inf \left\{t \geq 0: \mu\left([0, t] \times B_{n}\right)=1\right\} .
$$

By virtue of Lemma 4.7, for each $n \in \mathbb{N}(4.24)$ is positivity preserving. According to [20, Prop. 9.1], we have

$$
\mathbb{E}\left[\sup _{t \in[0, T]}\left\|r_{t}-r_{t}^{n}\right\|_{\beta}^{2}\right] \rightarrow 0 \quad \text { for all } T \in \mathbb{R}_{+},
$$

proving that (4.2) is positivity preserving.
Theorem 4.11. Suppose Assumptions 4.6, 4.7, and 4.8 are fulfilled. Then, for each initial curve $h_{0} \in L^{2}\left(\Omega, \mathcal{F}_{0}, \mathbb{P} ; H_{\beta}\right)$ there exists a unique adapted, càdlàg, mean-square continuous $\mathcal{H}_{\beta}$ valued solution $\left(f_{t}\right)_{t \geq 0}$ for the HJM equation (4.1) with $f_{0}=\ell h_{0}$ satisfying (3.15), and there exists a unique adapted, càdlàg, mean-square continuous mild and weak $H_{\beta}$-valued solution $\left(r_{t}\right)_{t \geq 0}$ for the HJMM equation (4.2) with $r_{0}=h_{0}$ satisfying (3.16), which is given by $r_{t}:=$ $\pi U_{t} f_{t}, t \geq 0$. Moreover, (4.2) is positivity preserving if and only if we have (4.3)-(4.6).

Proof. The statement follows from [20, Cor. 10.9], Proposition 4.4 (see also Remark 4.3), and Proposition 4.10.

Remark 4.10. Note that Theorem 4.11 is also valid on other state spaces. The only requirements are that the Hilbert space $H$ consists of real-valued, continuous functions, on which the point evaluations are continuous linear functionals, and that the shift semigroup extends to a strongly continuous group on a larger Hilbert space $\mathcal{H}$.

Remark 4.11. For the particular situation where (4.2) has no jumps, Theorem 4.11 corresponds to the statement of [30, Thm. 3], where positivity on weighted $L^{2}$-spaces is investigated. Since point evaluations are discontinuous functionals on $L^{2}$-spaces, the conditions in [30] are formulated by taking other appropriate linear functionals.

We shall now consider the arbitrage-free situation. Let $\alpha=\alpha_{\mathrm{HJM}}: H_{\beta} \rightarrow H_{\beta}$ in (4.2) be defined according to the HJM drift condition (1.6).

Proposition 4.12. Conditions (4.3)-(4.6) are satisfied if and only if we have (4.5), (4.6), and

$$
\begin{equation*}
\gamma(h, x)(\xi)=0, \quad \xi \in(0, \infty), h \in \partial P_{\xi}, \text { and } F \text {-almost all } x \in E . \tag{4.25}
\end{equation*}
$$

Proof. Provided (4.5), (4.6) are fulfilled, conditions (4.3), (4.4) are satisfied if and only if we have (4.3) and

$$
\begin{equation*}
-\int_{E} \gamma(h, x)(\xi) e^{\Gamma(h, x)(\xi)} F(d x) \geq 0, \quad \xi \in(0, \infty), h \in \partial P_{\xi} \tag{4.26}
\end{equation*}
$$

because the drift $\alpha$ is given by (1.6). By (4.7), relations (4.3), (4.26) are fulfilled if and only if we have (4.25).

Now let, as in section 3, measurable vector fields $\sigma: H_{\beta} \rightarrow L_{2}^{0}\left(H_{\beta}^{0}\right)$ and $\gamma: H_{\beta} \times E \rightarrow H_{\beta^{\prime}}^{0}$ be given, where $\beta^{\prime}>\beta$ is a real number.

Theorem 4.13. Suppose Assumptions 3.1 and 4.8 are fulfilled. Then, the statement of Theorem 3.2 is valid, and, in addition, the HJMM equation (1.10) is positivity preserving if and only if we have (4.5), (4.6), and (4.25).

Proof. The statement follows from Theorem 3.2, Theorem 4.11, and Proposition 4.12.
Finally, let us consider the Lévy case treated at the end of section 3. In this framework, the following statement is valid.

Proposition 4.14. Conditions (4.5), (4.6), and (4.25) are satisfied if and only if we have (4.5) and

$$
\begin{align*}
& h+\delta^{k}(h) x \in P, \quad h \in P, k=1, \ldots, e, \text { and } F_{k} \text {-almost all } x \in \mathbb{R},  \tag{4.27}\\
& \delta^{k}(h)(\xi)=0, \quad \xi \in(0, \infty), h \in \partial P_{\xi}, \text { and all } k=1, \ldots, e \text { with } F_{k}(\mathbb{R})>0 . \tag{4.28}
\end{align*}
$$

Proof. The claim follows from definition (3.18) of $F$ and definition (3.22) of $\gamma$.
Corollary 4.15. Suppose Assumption 3.3 and 4.8 are fulfilled. Then, the statement of Corollary 3.3 is valid, and, in addition, the HJMM equation (1.10) is positivity preserving if and only if we have (4.5), (4.27), and (4.28).

Proof. The assertion follows from Theorem 4.13 and Proposition 4.14.
Our above results on arbitrage-free, positivity preserving term structure models apply in particular for local state dependent volatilities. The following two results are obvious.

Proposition 4.16. Suppose for each $j$ there exists $\tilde{\sigma}^{j}: \mathbb{R}_{+} \times \mathbb{R} \rightarrow \mathbb{R}$, and suppose there are $\tilde{\gamma}: \mathbb{R}_{+} \times \mathbb{R} \times E \rightarrow \mathbb{R}$ and $\phi: \mathbb{R}_{+} \rightarrow(0, \infty)$ such that

$$
\begin{aligned}
\sigma^{j}(h)(\xi) & =\phi\left(\|h\|_{\beta}\right) \tilde{\sigma}^{j}(\xi, h(\xi)), \quad(h, \xi) \in H_{\beta} \times \mathbb{R}_{+}, \quad \text { for all } j, \\
\gamma(h, x)(\xi) & =\phi\left(\|h\|_{\beta}\right) \tilde{\gamma}(\xi, h(\xi), x), \quad(h, x, \xi) \in H_{\beta} \times E \times \mathbb{R}_{+} .
\end{aligned}
$$

Then, conditions (4.5), (4.6), and (4.25) are fulfilled if and only if

$$
\begin{align*}
& \tilde{\sigma}^{j}(\xi, 0)=0, \quad \xi \in(0, \infty), \quad \text { for all } j,  \tag{4.29}\\
& y+z \tilde{\gamma}(\xi, y, x) \geq 0, \quad \xi \in(0, \infty), y \in \mathbb{R}_{+}, z \in \phi\left(\mathbb{R}_{+}\right),  \tag{4.30}\\
& \quad \text { and } F \text {-almost all } x \in E, \\
& \tilde{\gamma}(\xi, 0, x)=0, \quad \xi \in(0, \infty) \text { and } F \text {-almost all } x \in E \tag{4.31}
\end{align*}
$$

Lévy term structure models with local state dependent volatilities have been studied in [35] and [28]. In the framework of Proposition 4.14, we obtain the following result.

Proposition 4.17. Suppose for each $j$ there is $\tilde{\sigma}^{j}: \mathbb{R}_{+} \times \mathbb{R} \rightarrow \mathbb{R}$, for all $k=1, \ldots$, e there is $\tilde{\delta}^{k}: \mathbb{R}_{+} \times \mathbb{R} \rightarrow \mathbb{R}$, and there exists $\phi: \mathbb{R}_{+} \rightarrow(0, \infty)$ such that

$$
\begin{aligned}
\sigma^{j}(h)(\xi) & =\phi\left(\|h\|_{\beta}\right) \tilde{\sigma}^{j}(\xi, h(\xi)), & (h, \xi) \in H_{\beta} \times \mathbb{R}_{+}, & \text {for all } j, \\
\delta^{k}(h)(\xi) & =\phi\left(\|h\|_{\beta}\right) \tilde{\delta}^{k}(\xi, h(\xi)), & (h, \xi) \in H_{\beta} \times \mathbb{R}_{+}, & k=1, \ldots, e .
\end{aligned}
$$

Then, conditions (4.5), (4.27), and (4.28) are fulfilled if and only if we have (4.29) and

$$
\begin{align*}
& y+z x \tilde{\delta}^{k}(\xi, y) \geq 0, \quad \xi \in(0, \infty), y \in \mathbb{R}_{+}, z \in \phi\left(\mathbb{R}_{+}\right), k=1, \ldots, e,  \tag{4.32}\\
& \quad \text { and } F_{k} \text {-almost all } x \in \mathbb{R}, \\
& \tilde{\delta}^{k}(\xi, 0)=0, \quad \xi \in(0, \infty) \text { and all } k=1, \ldots, e \text { with } F_{k}(\mathbb{R})>0 . \tag{4.33}
\end{align*}
$$

Section 5 in [35] contains some positivity results for Lévy-driven term structure models on weighted $L^{2}$-spaces. Using Proposition 4.17, we can derive the analogous statements of [35, Thm. 4] on our $H_{\beta}$-spaces.

Remark 4.12. For local state dependent volatilities, we can establish sufficient conditions on the mappings $\tilde{\sigma}^{j}, \tilde{\gamma}, \phi$ (resp., $\tilde{\sigma}^{j}, \tilde{\delta}^{k}, \phi$ ) such that Assumptions 3.1 and 4.8 (resp., Assumptions 3.3 and 4.8) are fulfilled, which allows us to combine Theorem 4.13 and Proposition 4.16 (resp., Corollary 4.15 and Proposition 4.17). We obtain such sufficient conditions by modifying the conditions from [16, Prop. 5.4.1] in an appropriate manner.

Appendix. Attached proofs. In this appendix, we gather the proofs of results which we have postponed for the sake of lucidity.

Proof of Theorem 2.1. Note that $H_{\beta}$ is the space $H_{w}$ from [16, sect. 5.1] with weight function $w(\xi)=e^{\beta \xi}, \xi \in \mathbb{R}_{+}$. Hence, the first six statements follow from [16, Thm. 5.1.1, Cor. 5.1.1].

For each $\beta^{\prime}>\beta$, the observation

$$
\int_{\mathbb{R}_{+}}\left|h^{\prime}(\xi)\right|^{2} e^{\beta \xi} d \xi \leq \int_{\mathbb{R}_{+}}\left|h^{\prime}(\xi)\right|^{2} e^{\beta^{\prime} \xi} d \xi, \quad h \in H_{\beta^{\prime}}
$$

shows that $H_{\beta^{\prime}} \subset H_{\beta}$ and (2.5). For an arbitrary $h \in H_{\beta^{\prime}}$, we have, by Hölder's inequality,

$$
\begin{aligned}
& \int_{\mathbb{R}_{+}}|h(\xi)-h(\infty)|^{2} e^{\beta \xi} d \xi=\int_{\mathbb{R}_{+}}\left(\int_{\xi}^{\infty} h^{\prime}(\eta) e^{\frac{1}{2} \beta^{\prime} \eta} e^{-\frac{1}{2} \beta^{\prime} \eta} d \eta\right)^{2} e^{\beta \xi} d \xi \\
& \leq \int_{\mathbb{R}_{+}}\left(\int_{\mathbb{R}_{+}}\left|h^{\prime}(\eta)\right|^{2} e^{\beta^{\prime} \eta} d \eta\right)\left(\int_{\xi}^{\infty} e^{-\beta^{\prime} \eta} d \eta\right) e^{\beta \xi} d \xi \leq \frac{1}{\beta^{\prime}\left(\beta^{\prime}-\beta\right)}\|h\|_{\beta^{\prime}}^{2} .
\end{aligned}
$$

Choosing $C_{5}:=\frac{1}{\beta^{\prime}\left(\beta^{\prime}-\beta\right)}$ proves (2.6).
It is clear that $\|\cdot\|_{\beta}$ is a norm on $\mathcal{H}_{\beta}$. First, we prove that there is a constant $K_{1}>0$ such that

$$
\begin{equation*}
\left\|h^{\prime}\right\|_{L^{1}(\mathbb{R})} \leq K_{1}\|h\|_{\beta}, \quad h \in \mathcal{H}_{\beta} \tag{A.1}
\end{equation*}
$$

Setting $K_{1}:=\sqrt{\frac{2}{\beta}}$, this is established by Hölder's inequality:

$$
\begin{align*}
& \int_{\mathbb{R}}\left|h^{\prime}(\xi)\right| d \xi=\int_{\mathbb{R}}\left|h^{\prime}(\xi)\right| e^{\frac{1}{2} \beta|\xi|} e^{-\frac{1}{2} \beta|\xi|} d \xi \\
& \leq\left(\int_{\mathbb{R}}\left|h^{\prime}(\xi)\right|^{2} e^{\beta|\xi|} d \xi\right)^{1 / 2}\left(\int_{\mathbb{R}} e^{-\beta|\xi|} d \xi\right)^{1 / 2}=\sqrt{\frac{2}{\beta}}\left(\int_{\mathbb{R}}\left|h^{\prime}(\xi)\right|^{2} e^{\beta|\xi|} d \xi\right)^{1 / 2} \tag{A.2}
\end{align*}
$$

As a consequence of (A.1), for each $h \in \mathcal{H}_{\beta}$ the limits $h(\infty):=\lim _{\xi \rightarrow \infty} h(\xi)$ and $h(-\infty):=$ $\lim _{\xi \rightarrow-\infty} h(\xi)$ exist. This allows us to define the new norm

$$
|h|_{\beta}:=\left(|h(-\infty)|^{2}+\int_{\mathbb{R}}\left|h^{\prime}(\xi)\right|^{2} e^{\beta|\xi|} d \xi\right)^{1 / 2}, \quad h \in \mathcal{H}_{\beta}
$$

From (A.2), we also deduce that

$$
\begin{equation*}
\left\|h^{\prime}\right\|_{L^{1}(\mathbb{R})} \leq K_{1}|h|_{\beta}, \quad h \in \mathcal{H}_{\beta} . \tag{A.3}
\end{equation*}
$$

Setting $K_{2}:=1+K_{1}$, from (A.1) and (A.3) it follows that

$$
\begin{align*}
\|h\|_{L^{\infty}(\mathbb{R})} & \leq K_{2}\|h\|_{\beta}  \tag{A.4}\\
\|h\|_{L^{\infty}(\mathbb{R})} & \leq K_{2}|h|_{\beta} \tag{A.5}
\end{align*}
$$

for all $h \in \mathcal{H}_{\beta}$. Estimate (A.4) shows that, for each $\xi \in \mathbb{R}$, the point evaluation $h \mapsto h(\xi)$, $\mathcal{H}_{\beta} \rightarrow \mathbb{R}$ is a continuous linear functional.

Using (A.4) and (A.5), we conclude that

$$
\frac{1}{\left(1+K_{2}^{2}\right)^{\frac{1}{2}}}\|h\|_{\beta} \leq|h|_{\beta} \leq\left(1+K_{2}^{2}\right)^{\frac{1}{2}}\|h\|_{\beta}, \quad h \in \mathcal{H}_{\beta},
$$

which shows that $\|\cdot\|_{\beta}$ and $|\cdot|_{\beta}$ are equivalent norms on $\mathcal{H}_{\beta}$.
Consider the separable Hilbert space $\mathbb{R} \times L^{2}(\mathbb{R})$ equipped with the norm $\left(|\cdot|^{2}+\|\cdot\|_{L^{2}(\mathbb{R})}^{2}\right)^{\frac{1}{2}}$. Then, the linear operator $T:\left(\mathcal{H}_{\beta},|\cdot|_{\beta}\right) \rightarrow \mathbb{R} \times L^{2}(\mathbb{R})$ given by

$$
T h:=\left(h(-\infty), h^{\prime} e^{\frac{1}{2} \beta|\cdot|}\right), \quad h \in \mathcal{H}_{\beta},
$$

is an isometric isomorphism with inverse

$$
\left(T^{-1}(u, g)\right)(x)=u+\int_{-\infty}^{x} g(\eta) e^{-\frac{1}{2} \beta|\eta|} d \eta, \quad(u, g) \in \mathbb{R} \times L^{2}(\mathbb{R})
$$

Since $\|\cdot\|_{\beta}$ and $|\cdot|_{\beta}$ are equivalent, $\left(\mathcal{H}_{\beta},\|\cdot\|_{\beta}\right)$ is a separable Hilbert space.
Next, we claim that

$$
\mathcal{D}_{0}:=\left\{g \in \mathcal{H}_{\beta} \mid g^{\prime} \in \mathcal{H}_{\beta}\right\}
$$

is dense in $\mathcal{H}_{\beta}$. Indeed, $C_{c}^{\infty}(\mathbb{R})$ is dense in $L^{2}(\mathbb{R})$; see [5, Cor. IV.23]. Fix $h \in \mathcal{H}_{\beta}$ and let $\left(g_{n}\right)_{n \in \mathbb{N}} \subset C_{c}^{\infty}(\mathbb{R})$ be an approximating sequence of $h^{\prime} e^{\frac{1}{2} \beta|\cdot|}$ in $L^{2}(\mathbb{R})$. Then, we have $h_{n}:=T^{-1}\left(h(-\infty), g_{n}\right) \in \mathcal{D}_{0}$ for all $n \in \mathbb{N}$ and $h_{n} \rightarrow h$ in $\mathcal{H}_{\beta}$.

For each $t \in \mathbb{R}$ and $h \in \mathcal{H}_{\beta}$, the function $U_{t} h$ is again absolutely continuous. We claim that there exists a constant $K_{3}>0$ such that

$$
\begin{equation*}
\left\|U_{t} h\right\|_{\beta}^{2} \leq\left(K_{3}+e^{\beta|t|}\right)\|h\|_{\beta}^{2}, \quad(t, h) \in \mathbb{R} \times \mathcal{H}_{\beta} . \tag{A.6}
\end{equation*}
$$

Indeed, using (A.4) we obtain

$$
\begin{aligned}
\left\|U_{t} h\right\|_{\beta}^{2} & =|h(t)|^{2}+\int_{0}^{\infty}\left|h^{\prime}(\xi+t)\right|^{2} e^{\beta \xi} d \xi+\int_{-\infty}^{0}\left|h^{\prime}(\xi+t)\right|^{2} e^{-\beta \xi} d \xi \\
& =|h(t)|^{2}+e^{-\beta t} \int_{t}^{\infty}\left|h^{\prime}(\xi)\right|^{2} e^{\beta \xi} d \xi+e^{\beta t} \int_{-\infty}^{t}\left|h^{\prime}(\xi)\right|^{2} e^{-\beta \xi} d \xi \\
& \leq\left(K_{2}^{2}+1+e^{\beta|t|}\right)\|h\|_{\beta}^{2}, \quad h \in \mathcal{H}_{\beta} .
\end{aligned}
$$

Setting $K_{3}:=1+K_{2}^{2}$, this establishes (A.6). Hence, we have $U_{t} h \in \mathcal{H}_{\beta}$ for all $t \in \mathbb{R}$ and $h \in \mathcal{H}_{\beta}$ and $U_{t} \in L\left(\mathcal{H}_{\beta}\right), t \in \mathbb{R}$.

It remains to show strong continuity of the group $\left(U_{t}\right)_{t \in \mathbb{R}}$. Using the observation

$$
h(\xi+t)-h(\xi)=t \int_{0}^{1} h^{\prime}(\xi+s t) d s, \quad(\xi, t, h) \in \mathbb{R} \times \mathbb{R} \times \mathcal{H}_{\beta}
$$

and (A.6), we obtain for each $g \in \mathcal{D}_{0}$ the convergence

$$
\begin{aligned}
& \left\|U_{t} g-g\right\|_{\beta}^{2}=|g(t)-g(0)|^{2}+\int_{\mathbb{R}}\left|g^{\prime}(\xi+t)-g^{\prime}(\xi)\right|^{2} e^{\beta|\xi|} d \xi \\
& \leq|g(t)-g(0)|^{2}+t^{2} \int_{0}^{1} \int_{\mathbb{R}}\left|g^{\prime \prime}(\xi+s t)\right|^{2} e^{\beta|\xi|} d \xi d s \\
& \leq|g(t)-g(0)|^{2}+t^{2} \int_{0}^{1}\left\|U_{s t} g^{\prime}\right\|_{\beta}^{2} d s \\
& \leq|g(t)-g(0)|^{2}+t^{2}\left\|g^{\prime}\right\|_{\beta}^{2} \int_{0}^{1}\left(K_{3}+e^{\beta s|t|}\right) d s \\
& =|g(t)-g(0)|^{2}+\left(K_{3} t^{2}+\frac{|t|}{\beta}\left(e^{\beta|t|}-1\right)\right)\left\|g^{\prime}\right\|_{\beta}^{2} \rightarrow 0 \quad \text { as } t \rightarrow 0 .
\end{aligned}
$$

Hence, $\left(U_{t}\right)_{t \in \mathbb{R}}$ is strongly continuous on $\mathcal{D}_{0}$. But for any $h \in \mathcal{H}_{\beta}$ and $\epsilon>0$ there exists $g \in \mathcal{D}_{0}$ with $\|h-g\|_{\beta}<\frac{\epsilon}{4 \sqrt{K_{3}+e^{\beta}}}$. Combining this with (A.6) yields

$$
\begin{aligned}
\left\|U_{t} h-h\right\|_{\beta} & \leq\left\|U_{t}(h-g)\right\|_{\beta}+\left\|U_{t} g-g\right\|_{\beta}+\|g-h\|_{\beta} \\
& <\sqrt{K_{3}+e^{\beta|t|}} \frac{\epsilon}{4 \sqrt{K_{3}+e^{\beta}}}+\left\|U_{t} g-g\right\|_{\beta}+\frac{\epsilon}{4 \sqrt{K_{3}+e^{\beta}}}<\epsilon
\end{aligned}
$$

for $t \in \mathbb{R}$ small enough. We conclude that $\left(U_{t}\right)_{t \in \mathbb{R}}$ is a $C_{0}$-group on $\mathcal{H}_{\beta}$.
Finally, relation (2.7) follows from the definitions of $\ell$ and $\pi$.
Proof of Lemma 4.6. Note that $\tilde{W}$ is a continuous $\left(\tilde{\mathcal{F}}_{t}\right)$-adapted process with $\tilde{W}_{0}=0$, and $\tilde{\mu}$ is an integer-valued random measure on $\mathbb{R}_{+} \times E$.

We fix an arbitrary $u \in U$. The process

$$
M_{t}:=\frac{\exp \left(i\left\langle u, W_{t}\right\rangle\right)}{\mathbb{E}\left[\exp \left(i\left\langle u, W_{t}\right\rangle\right)\right]}, \quad t \geq 0
$$

is a complex-valued martingale, because for all $s, t \in \mathbb{R}_{+}$with $s<t$ the random variable $W_{t}-$ $W_{s}$ and the $\sigma$-algebra $\mathcal{F}_{s}$ are independent. The martingale $\left(M_{t}\right)_{t \geq 0}$ admits the representation

$$
M_{t}=\exp \left(i\left\langle u, W_{t}\right\rangle+\frac{t}{2}\langle Q u, u\rangle\right), \quad t \geq 0
$$

According to the optional stopping theorem, the process $\left(M_{t+\tau}\right)_{t \geq 0}$ is a nowhere vanishing complex $\left(\tilde{\mathcal{F}}_{t}\right)$-martingale. Thus, for $s, t \in \mathbb{R}_{+}$with $s<t$ we obtain

$$
\mathbb{E}\left[\left.\frac{M_{t+\tau}}{M_{s+\tau}} \right\rvert\, \tilde{\mathcal{F}}_{s}\right]=1 .
$$

For each $C \in \tilde{\mathcal{F}}_{s}$, we get

$$
\mathbb{E}\left[\mathbb{1}_{C} \exp \left(i\left\langle u, \tilde{W}_{t}-\tilde{W}_{s}\right\rangle\right)\right]=\mathbb{P}(C) \exp \left(-\frac{t-s}{2}\langle Q u, u\rangle\right) .
$$

Hence, the random variable $\tilde{W}_{t}-\tilde{W}_{s}$ and the $\sigma$-algebra $\tilde{\mathcal{F}}_{s}$ are independent, and $\tilde{W}_{t}-\tilde{W}_{s}$ has a Gaussian distribution with covariance operator $(t-s) Q$. The expansion (4.16) follows from (3.1).

Now we fix $v \in \mathbb{R}$ and $B \in \mathcal{E}$ with $F(B)<\infty$. The process

$$
N_{t}:=\frac{\exp (i v \mu([0, t] \times B))}{\mathbb{E}[\exp (i v \mu([0, t] \times B))]}, \quad t \geq 0
$$

is a complex-valued martingale, because for all $s, t \in \mathbb{R}_{+}$with $s<t$ the random variable $\mu((s, t] \times B)$ and the $\sigma$-algebra $\mathcal{F}_{s}$ are independent. By [25, Thm. II.4.8], the martingale $\left(N_{t}\right)_{t \geq 0}$ admits the representation

$$
N_{t}=\exp \left(i v \mu([0, t] \times B)-\left(e^{i v}-1\right) F(B) t\right), \quad t \geq 0
$$

According to the optional stopping theorem, the process $\left(N_{t+\tau}\right)_{t \geq 0}$ is a nowhere vanishing complex $\left(\tilde{\mathcal{F}}_{t}\right)$-martingale. Thus, for $s, t \in \mathbb{R}_{+}$with $s<t$ we obtain

$$
\mathbb{E}\left[\left.\frac{N_{t+\tau}}{N_{s+\tau}} \right\rvert\, \tilde{\mathscr{F}}_{s}\right]=1 .
$$

For each $C \in \tilde{\mathcal{F}}_{s}$, we get

$$
\mathbb{E}\left[\mathbb{1}_{C} \exp (i v \tilde{\mu}((s, t] \times B))\right]=\mathbb{P}(C) \exp \left(\left(e^{i v}-1\right) F(B)(t-s)\right)
$$

Hence, the random variable $\tilde{\mu}((s, t] \times B)$ and the $\sigma$-algebra $\tilde{\mathcal{F}}_{s}$ are independent, and $\tilde{\mu}((s, t] \times B)$ has a Poisson distribution with mean $(t-s) F(B)$.

Next, we claim that

$$
\begin{equation*}
\int_{\tau}^{\tau+t} \Phi_{s} d W_{s}=\int_{0}^{t} \Phi_{\tau+s} d \tilde{W}_{s} \tag{A.7}
\end{equation*}
$$

for every predictable process $\Phi: \Omega \times \mathbb{R}_{+} \rightarrow L_{2}^{0}\left(H_{\beta}\right)$ satisfying

$$
\mathbb{P}\left(\int_{0}^{t}\left\|\Phi_{s}\right\|_{L_{2}^{0}\left(H_{\beta}\right)}^{2} d s<\infty\right)=1
$$

for all $t \in \mathbb{R}_{+}$, and

$$
\begin{equation*}
\int_{\tau}^{\tau+t} \Psi(s, x)(\mu(d s, d x)-F(d x) d s)=\int_{0}^{t} \Psi(\tau+s, x)(\tilde{\mu}(d s, d x)-F(d x) d s) \tag{A.8}
\end{equation*}
$$

for every predictable process $\Psi: \Omega \times \mathbb{R}_{+} \times E \rightarrow H_{\beta}$ satisfying

$$
\mathbb{P}\left(\int_{0}^{t} \int_{E}\|\Psi(s, x)\|_{\beta}^{2} F(d x) d s<\infty\right)=1
$$

for all $t \in \mathbb{R}_{+}$. If $\Phi, \Psi$ are elementary and $\tau$ is a simple stopping time, then (A.7), (A.8) hold by inspection. The general case follows by localization.

If $\left(r_{t}\right)_{t \geq 0}$ is a weak solution to (4.2), for every $\zeta \in \mathcal{D}\left(\left(\frac{d}{d \xi}\right)^{*}\right)$ relations (A.7), (A.8) yield

$$
\begin{aligned}
\left\langle\zeta, r_{\tau+t}\right\rangle= & \left\langle\zeta, r_{\tau}\right\rangle+\int_{\tau}^{\tau+t}\left(\left\langle\left(\frac{d}{d \xi}\right)^{*} \zeta, r_{s}\right\rangle+\left\langle\zeta, \alpha\left(r_{s}\right)\right\rangle\right) d s+\int_{\tau}^{\tau+t}\left\langle\zeta, \sigma\left(r_{s}\right)\right\rangle d W_{s} \\
& +\int_{\tau}^{\tau+t} \int_{E}\left\langle\zeta, \gamma\left(r_{s-}, x\right)\right\rangle(\mu(d s, d x)-F(d x) d s) \\
= & \left\langle\zeta, r_{\tau}\right\rangle+\int_{0}^{t}\left(\left\langle\left(\frac{d}{d \xi}\right)^{*} \zeta, r_{\tau+s}\right\rangle+\left\langle\zeta, \alpha\left(r_{\tau+s}\right)\right\rangle\right) d s+\int_{0}^{t}\left\langle\zeta, \sigma\left(r_{\tau+s}\right)\right\rangle d \tilde{W}_{s} \\
& +\int_{0}^{t} \int_{E}\left\langle\zeta, \gamma\left(r_{(\tau+s)-}, x\right)\right\rangle(\tilde{\mu}(d s, d x)-F(d x) d s) .
\end{aligned}
$$

Hence, $\left(\tilde{r}_{t}\right)_{t \geq 0}$ is a weak solution for (4.17).
Acknowledgment. The authors thank two anonymous referees for their helpful remarks and suggestions.

## REFERENCES

[1] S. Albeverio, V. Mandrekar, and B. Rüdiger, Existence of mild solutions for stochastic differential equations and semilinear equations with non Gaussian Lévy noise, Stochastic Process. Appl., 119 (2009), pp. 835-863.
[2] F. Baudoin and J. Teichmann, Hypoellipticity in infinite dimensions and an application in interest rate theory, Ann. Appl. Probab., 15 (2005), pp. 1765-1777.
[3] T. Björk, G. Di Masi, Y. Kabanov, and W. Runggaldier, Towards a general theory of bond markets, Finance Stoch., 1 (1997), pp. 141-174.
[4] T. Björk, Y. Kabanov, and W. Runggaldier, Bond market structure in the presence of marked point processes, Math. Finance, 7 (1997), pp. 211-239.
[5] H. Brézis, Analyse fonctionnelle: Théorie et applications, 4th ed., Masson, Paris, 1993.
[6] R. Carmona and M. Tehranchi, Interest Rate Models: An Infinite Dimensional Stochastic Analysis Perspective, Springer-Verlag, Berlin, 2006.
[7] J. Cox, J. Ingersoll, and S. Ross, A theory of the term structure of interest rates, Econometrica, 53 (1985), pp. 385-408.
[8] G. Da Prato and J. Zabczyk, Stochastic Equations in Infinite Dimensions, Cambridge University Press, New York, 1992.
[9] C. Dellacherie and P. A. Meyer, Probabilités et potentiel, Hermann, Paris, 1982.
[10] E. Eberlein, J. Jacod, and S. Raible, Lévy term structure models: No-arbitrage and completeness, Finance Stoch., 9 (2005), pp. 67-88.
[11] E. Eberlein and W. Kluge, Calibration of Lévy term structure models, in Advances in Mathematical Finance: In Honor of D. B. Madan, M. Fu, R. A. Jarrow, J.-Y. Yen, and R. J. Elliott, eds., Birkhäuser, Boston, MA, 2007, pp. 147-172.
[12] E. Eberlein and W. Kluge, Exact pricing formulae for caps and swaptions in a Lévy term structure model, J. Comput. Finance, 9 (2006), pp. 99-125.
[13] E. Eberlein and W. Kluge, Valuation of floating range notes in Lévy term structure models, Math. Finance, 16 (2006), pp. 237-254.
[14] E. Eberlein and F. Özkan, The defaultable Lévy term structure: Ratings and restructuring, Math. Finance, 13 (2003), pp. 277-300.
[15] E. Eberlein and S. Raible, Term structure models driven by general Lévy processes, Math. Finance, 9 (1999), pp. 31-53.
[16] D. Filipović, Consistency Problems for Heath-Jarrow-Morton Interest Rate Models, Springer-Verlag, Berlin, 2001.
[17] D. Filipović, Term-Structure Models: A Graduate Course, Springer-Verlag, Berlin, 2009.
[18] D. Filipović, Time-inhomogeneous affine processes, Stochastic Process. Appl., 115 (2005), pp. 639-659.
[19] D. Filipović and S. Tappe, Existence of Lévy term structure models, Finance Stoch., 12 (2008), pp. 83115.
[20] D. Filipović, S. Tappe, and J. Teichmann, Jump-diffusions in Hilbert spaces: Existence, stability and numerics, Stochastics, to appear.
[21] D. Filipović and J. Teichmann, On the geometry of the term structure of interest rates, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 460 (2004), pp. 129-167.
[22] R. K. Getoor, On the construction of kernels, in Séminaire de Probabilités IX, Springer-Verlag, Berlin, 1975, pp. 443-463.
[23] D. Heath, R. Jarrow, and A. Morton, Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation, Econometrica, 60 (1992), pp. 77-105.
[24] M. Hyll, Affine term structures and short-rate realizations of forward rate models driven by jumpdiffusion processes, in Essays on the Term Structure of Interest Rates, Ph.D. thesis, Stockholm School of Economics, Stockholm, Sweden, 2000, pp. 1-23.
[25] J. Jacod and A. N. Shiryaev, Limit Theorems for Stochastic Processes, Springer-Verlag, Berlin, 2003.
[26] R. Jarrow and D. B. Madan, Option pricing using the term structure of interest rates to hedge systematic discontinuities in asset returns, Math. Finance, 5 (1995), pp. 311-336.
[27] P. Kotelenez, Comparison methods for a class of function valued stochastic partial differential equations, Probab. Theory Related Fields, 93 (1992), pp. 1-19.
[28] C. Marinelli, Local well-posedness of Musiela's SPDE with Lévy noise, Math. Finance, 20 (2010), pp. 341-363.
[29] C. Marinelli, C. Prévôt, and M. Röckner, Regular dependence on initial data for stochastic evolution equations with multiplicative Poisson noise, J. Funct. Anal., 258 (2010), pp. 616-649.
[30] A. Milian, Comparison theorems for stochastic evolution equations, Stoch. Stoch. Rep., 72 (2002), pp. 79108.
[31] A. Morton, Arbitrage and Martingales, Ph.D. thesis, Cornell University, Ithaca, NY, 1989.
[32] M. Musiela, Stochastic PDEs and term structure models, in Journées Internationales de Finance, IGRAFFI, La Baule, France, 1993.
[33] T. Nakayama, Support theorem for mild solutions of SDE's in Hilbert spaces, J. Math. Sci. Univ. Tokyo, 11 (2004), pp. 245-311.
[34] T. Nakayama, Viability theorem for SPDE's including HJM framework, J. Math. Sci. Univ. Tokyo, 11 (2004), pp. 313-324.
[35] S. Peszat and J. Zabczyk, Heath-Jarrow-Morton-Musiela Equation of Bond Market, Preprint IMPAN 677, Warsaw, Poland; available online from http://www.impan.pl/Preprints/p677.pdf (2007).
[36] S. Peszat and J. Zabczyk, Stochastic Partial Differential Equations with Lévy Noise, Cambridge University Press, Cambridge, UK, 2007.
[37] S. Raible, Lévy Processes in Finance: Theory, Numerics, and Empirical Facts, Ph.D. thesis, University of Freiburg, Freiburg, Germany, 2000.
[38] H. Shirakawa, Interest rate option pricing with Poisson-Gaussian forward rate curve processes, Math. Finance, 1 (1991), pp. 77-94.
[39] P. Volkmann, Über die Invarianz konvexer Mengen und Differentialgleichungen in einem normierten Raume, Math. Ann., 203 (1973), pp. 201-210.
[40] S. Watanabe and T. Yamada, On the uniqueness of solutions of stochastic differential equations II, J. Math. Kyoto Univ., 11 (1971), pp. 553-563.
[41] D. Werner, Funktionalanalysis, Springer-Verlag, Berlin, 2002.

# Default Intensities Implied by CDO Spreads: Inversion Formula and Model Calibration* 

Rama Cont $^{\dagger}$, Romain Deguest ${ }^{\ddagger}$, and Yu Hang Kan ${ }^{\ddagger}$


#### Abstract

We propose a simple computational method for constructing an arbitrage-free collateralized debt obligation (CDO) pricing model which matches a prespecified set of CDO tranche spreads. The key ingredient of the method is an inversion formula for computing the aggregate default rate in a portfolio, as a function of the number of defaults, from its expected tranche notionals. This formula can be seen as an analogue of the Dupire formula for portfolio credit derivatives. Together with a quadratic programming method for recovering expected tranche notionals from CDO spreads, our inversion formula leads to an efficient nonparametric method for calibrating CDO pricing models. Contrarily to the base correlation method, our method yields an arbitrage-free model. Comparing this approach to other calibration methods, we find that model-dependent quantities such as the forward starting tranche spreads and jump-to-default ratios are quite sensitive to the calibration method used, even within the same model class. On the other hand, comparing the local intensity functions implied by different credit portfolio models reveals that apparently different models, such as the static Student-t copula models and the reduced-form affine jump-diffusion models, lead to similar marginal loss distributions and tranche spreads.


Key words. portfolio credit derivatives, collateralized debt obligation, inverse problem, effective intensity, default intensity, expected tranche notionals, Dupire formula, quadratic programming, calibration, tranche, Duffie-Gârleanu model, Student-t copula, Herbertsson model

AMS subject classifications. 60G55, 90C $20,62 \mathrm{M} 05$
DOI. 10.1137/09076800X

The inadequacy of widely used static factor models, such as the Gaussian copula model and its various extensions, for pricing and hedging portfolio credit derivatives, as emphasized by the recent turmoil in credit derivatives markets, has led to the development of various dynamic models for portfolio credit risk.

One of the main obstacles in implementing and using these dynamic models has been the availability of efficient calibration algorithms. Once models are calibrated to market data, they can be compared in terms of pricing and hedging performance. Previous studies on dynamic models have mostly been based on black-box optimization procedures applied to nonconvex least squares minimization problems. The lack of convexity entails that the convergence and stability of these methods are not guaranteed, casting doubts on the reproducibility of calibration results and their stability.

Recovering implied default rates from market data is by nature an ill-posed problem. Although the actual default rate (intensity) may depend on the past market history, it has

[^46]been argued $[6,7,21]$ that the information contained in collateralized debt obligation (CDO) tranche spreads can be used at best to recover the local intensity function, defined as the conditional expectation of the portfolio default intensity given the loss level. The local intensity function is analogous to the local volatility function introduced by Dupire [10] for equity derivatives. It summarizes all information available from CDO tranche spreads on the marginal loss distributions of the portfolio and provides a common basis to compare different models.

Herbertsson [13] and Lopatin and Misirpashaev [18] have used parametric methods to recover local intensity functions from CDO data. Laurent, Cousin, and Fermanian [15] propose an implied tree method for reconstructing the local intensity function. Reformulating the calibration of default intensity as a stochastic control problem, Cont and Minca [6] proposed a stable nonparametric approach based on relative entropy minimization for recovering the local intensity function.

We propose in this paper an alternative, and simpler, approach based on an analytical inversion formula for the local intensity function, which is analogous to the Dupire formula in diffusion models [10]. This formula allows us to compute the local intensity function of a portfolio from its expected tranche notionals. This yields a simple computational method for constructing an arbitrage-free CDO pricing model which matches a prespecified set of tranche spreads.

Together with a quadratic programming method for recovering expected tranche notionals from CDO spreads, our inversion formula leads to an efficient nonparametric method for calibrating CDO pricing models. In a first step, we extract the expected tranche notionals from the CDO spreads by solving a quadratic minimization problem under linear constraints. Next, the default intensity is computed from the expected tranche notionals using the inversion formula. Unlike the calibration methods introduced in [6, 13], our method requires only relatively simple mathematical techniques.

Comparing this approach to other calibration methods using iTraxx Europe index CDO spreads, we find that model-dependent quantities such as the forward starting tranche spreads and jump-to-default ratios are quite sensitive to the calibration method used, even within the same model class. On the other hand, comparing the local intensity functions implied by different credit portfolio models reveals that apparently different models, such as static Student-t copula models and reduced-form affine jump-diffusion models, lead to similar marginal loss distributions and tranche spreads.

Figure 1 gives an overview of this paper, and the details are structured as follows. Section 1 derives our main results concerning the existence and expression of the local intensity function given the expected tranche notionals. Section 2 proposes a nonparametric method to recover the local intensity function from the CDO market data. Section 3 compares this calibration method with the parametric approach introduced by Herbertsson [13] and the entropy minimization algorithm proposed by Cont and Minca [6]. Section 4 compares the local intensities implied by various credit portfolio loss models. Section 5 summarizes our main findings and discusses some implications. Proofs are presented in Appendix A.

1. An inversion formula for the local intensity function. In this section, we first introduce the notions of effective intensity, local intensity function, and expected tranche notionals. Then, we present two theorems that are related to the inversion formula of the local intensity

## Credit portfolio loss models

Simulation

## CDO market Data

## Quadratic

programming

Expected tranche notionals

Inversion formula


Figure 1. Application of the inversion formula to recover the local intensity function.
function. Those theorems are the key results for computing the local intensity function implied by a credit portfolio loss model or the expected tranche notionals obtained from the market data.
1.1. Local intensity function. We model credit events using as filtered probability space $\left(\Omega, \mathcal{F},\left(\mathcal{F}_{t}\right)_{t \in\left[0, T^{*}\right]}, \mathbb{Q}\right)$, where $\Omega$ is the set of market scenarios, the filtration $\left(\mathcal{F}_{t}\right)_{t \in\left[0, T^{*}\right]}$ represents the flow of information up to a terminal date $T^{*}$, and $\mathbb{Q}$ is a probability measure representing the market pricing rule (pricing measure).

Consider an equally weighted credit portfolio (index) consisting of $n$ names. Our main interest is the aggregate portfolio loss due to defaults, modeled as

$$
\begin{equation*}
L_{t}=\delta N_{t} \tag{1}
\end{equation*}
$$

where $\left(N_{t}\right)$ is a point process representing the number of defaults and $\delta$ is the loss at each default, assumed to be constant. Without loss of generality, we set $N_{0}=0$ and assume that the timing of default events is independent of the interest rates. We assume the existence of a (risk-neutral) default intensity.

Assumption 1. The point process $\left(N_{t}\right)$ admits an intensity: there exists a nonnegative $\mathcal{F}_{t}$-predictable process $\left(\lambda_{t}\right)$ such that, for all $t \in\left[0, T^{*}\right]$,

$$
\int_{0}^{t} \lambda_{s} d s<\infty \quad \mathbb{Q} \text {-a.s. }
$$

The portfolio default intensity can be seen as the conditional probability per unit time of the next default:

$$
\lambda_{t}=\lim _{\Delta t \rightarrow 0} \frac{1}{\Delta t} \mathbb{Q}\left(N_{t^{-}+\Delta t}-N_{t^{-}}=1 \mid \mathcal{F}_{t^{-}}\right)
$$

It is characterized by the fact that, for any nonnegative $\mathcal{F}_{t}$-predictable process $\left(C_{t}\right)$,

$$
E^{\mathbb{Q}}\left[\int_{0}^{T^{*}} C_{s} d N_{s}\right]=E^{\mathbb{Q}}\left[\int_{0}^{T^{*}} C_{s} \lambda_{s} d s\right]
$$

In general, $\left(\lambda_{t}\right)$ can be path dependent; i.e., it may depend on the entire market history. However, as shown in $[6,7]$, one can construct a Markovian pricing model whose marginal distributions mimic those of $\left(L_{t}\right)$. The intensity $\left(\lambda_{t}^{\text {eff }}\right)$ of this Markovian projection is called the effective default intensity [6].

Definition 1 (local intensity function). Consider a loss process satisfying Assumption 1 with

$$
\forall t \in\left(0, T^{*}\right], \quad E^{\mathbb{Q}}\left[\lambda_{t} \mid \mathcal{F}_{0}\right]<\infty
$$

The local intensity function $a:\left(0, T^{*}\right] \times\{0,1, \ldots, n\} \mapsto \mathbb{R}_{+}$at time 0 is defined as

$$
\begin{equation*}
a(t, i):=E^{\mathbb{Q}}\left[\lambda_{t} \mid N_{t^{-}}=i, \mathcal{F}_{0}\right] . \tag{2}
\end{equation*}
$$

If $\mathbb{Q}\left(N_{t^{-}}=i \mid \mathcal{F}_{0}\right)=0$, we set $a(t, i)=0$ by convention.
We call $\lambda_{t}^{\text {eff }}:=a\left(t, N_{t^{-}}\right)$the effective intensity of the loss process: it is the best estimate for the default intensity given the portfolio loss level. $a(t, i)$ may also be viewed as a forward default rate for the portfolio given that $i$ defaults have occurred in the portfolio [21]. Similarly to the local volatility function in diffusion models [10], the local intensity function summarizes all the necessary information to price non-path-dependent portfolio credit derivatives: for any point process $\left(N_{t}\right)$ that satisfies Assumption 1, there exists a Markovian point process $\left(\widetilde{N}_{t}\right)$ with transition rate $a\left(t, \widetilde{N}_{t^{-}}\right)$such that $\left(N_{t}\right)$ and $\left(\widetilde{N}_{t}\right)$ have the same marginal distributions at all dates $t \in\left[0, T^{*}\right][7]$. Moreover, given the local intensity function, the marginal distribution can be computed by solving the forward Kolmogorov equations [21].

Given that the information content of market prices can be summarized in the local intensity function, various calibration methods have been proposed in the recent literature for recovering the local intensity function from CDO tranche spreads. Examples include parametric methods introduced by Herbertsson [13] and Lopatin and Misirpashaev [18] and a nonparametric entropy minimization algorithm proposed by Cont and Minca [6]. In section 2, we will introduce a novel nonparametric calibration method which makes use of the inversion formula that will be shown later in this section.

As the local intensity function can be defined for a wide range of credit portfolio loss processes, it provides a common basis to compare models defined in different manners. We will further study this aspect in section 4.
1.2. Expected tranche notionals. Consider the equity tranche of a synthetic CDO with detachment point $K$. The expected remaining notional value of this equity tranche at time $T>t$ is equal to

$$
P_{t}(T, K):=E^{\mathbb{Q}}\left[\left(K-L_{T}\right)^{+} \mid \mathcal{F}_{t}\right] .
$$

We follow the notation in [7] and call this quantity the expected tranche notional with maturity $T$ and strike $K$. For simplicity, we will fix the observation time at 0 in the remainder of the paper and drop the subscript $t$ in the notation of the expected tranche notionals. Expected tranche notionals verify the following static arbitrage constraints (a proof is given in Appendix A).

Property 1 (static arbitrage constraints).
(a) $P(T, K) \geq 0$,
(b) $P(T, 0)=0$,
(c) $P(0, K)=K$,
(d) $K \mapsto P(T, K)$ is convex,
(e) $P\left(T_{2}, K_{1}\right)-P\left(T_{1}, K_{1}\right) \geq P\left(T_{2}, K_{2}\right)-P\left(T_{1}, K_{2}\right)$ for any $T_{1} \leq T_{2}, K_{1} \leq K_{2}$,
(f) $K \mapsto P(T, K)$ is continuous and piecewise linear on $[(i-1) \delta, i \delta], i=1, \ldots, n$.

Cont and Savescu [7] show that the expected tranche notionals can be computed directly from the local intensity function by solving a system of forward differential equations: for $T \in\left(0, T^{*}\right], i=1, \ldots, n$,

$$
\begin{align*}
& \partial_{T} P(T, i \delta)=-a(T, 0) P(T, \delta)-\sum_{k=1}^{i-1} a(T, k) \nabla_{K}^{2} P(T,(k-1) \delta),  \tag{3}\\
& \text { with initial condition } \quad P(0, i \delta)=i \delta,
\end{align*}
$$

where $\nabla_{K}$ is the forward difference operator in strike:

$$
\nabla_{K} F(T, i \delta):=F(T,(i+1) \delta)-F(T, i \delta)
$$

for any function $F:\left[0, T^{*}\right] \times\{i \delta: i=0, \ldots, n-1\} \mapsto \mathbb{R}$. In fact, the forward equations (3) are a result of the forward Kolmogorov equations with the identities

$$
\mathbb{Q}\left(L_{T}=i \delta \mid \mathcal{F}_{0}\right)= \begin{cases}\frac{P(T, \delta)}{\delta}, & i=0,  \tag{4}\\ \frac{\nabla_{K}^{2} P(T,(i-1) \delta)}{\delta}, & i=1, \ldots, n-1\end{cases}
$$

1.3. Inversion formula and Markovian projection. To compute the local intensity function implied by a credit portfolio loss model, Theorem 1 shows that we can first compute the expected tranche notionals under the model assumption and then convert them into a local intensity function using an inversion formula. This approach avoids computing the conditional expectation of the default intensity $\left(\lambda_{t}\right)$, which can be a difficult task.

Theorem 1 (inversion formula). Consider a portfolio loss process $L_{t}=\delta N_{t}$, where the point process $\left(N_{t}\right)$ verifies Assumption 1 and

$$
\forall T \in\left(0, T^{*}\right], \quad E^{\mathbb{Q}}\left[\lambda_{T} \mid \mathcal{F}_{0}\right]<\infty
$$

The local intensity function (2) is given by

$$
a(T, i)=E^{\mathbb{Q}}\left[\lambda_{T} \mid N_{T^{-}}=i, \mathcal{F}_{0}\right]= \begin{cases}\frac{-\partial_{T} P(T, \delta)}{P(T, \delta)}, & i=0,  \tag{5}\\ \frac{-\nabla_{K} \partial_{T} P(T, i \delta)}{\nabla_{K}^{2} P(T,(i-1) \delta)}, & i=1, \ldots, n-1, \\ 0, & i=n,\end{cases}
$$

for all $T \in\left(0, T^{*}\right]$, where $P(T, i \delta)=E^{\mathbb{Q}}\left[\left(i \delta-L_{T}\right)^{+} \mid \mathcal{F}_{0}\right]$ is the expected tranche notional.

The proof is given in Appendix A.
In a practical situation, the default intensity $\left(\lambda_{t}\right)$ is unobservable, but, if sufficiently many tranche spreads are quoted, expected tranche notionals can be recovered from market data. We will present a nonparametric method to recover the expected tranche notionals from the tranche spreads in section 2. Given such a set of values of expected tranche notionals, Theorem 2 shows that we can construct a Markovian loss process with an intensity in the form of (5) consistent with these values. This result, which is analogous to the Dupire formula for local volatility [10], is particularly useful when we want to recover a local intensity function from either the market data or a model that does not satisfy Assumption 1, such as the static factor models.

Theorem 2 (the local intensity function that is implied by expected tranche notionals). We let $\{P(T, i \delta)\}_{T \in\left[0, T^{*}\right], i=0, \ldots, n}$ be a (complete) set of expected tranche notionals verifying Property 1 and define the function $a:\left(0, T^{*}\right] \times\{0,1, \ldots, n\}$ by

$$
a(T, i)= \begin{cases}\frac{-\partial_{T} P(T, \delta)}{P(T, \delta)}, & i=0  \tag{6}\\ \frac{-\nabla_{K} \partial_{T} P(T, i \delta)}{\nabla_{K}^{2} P(T,(i-1) \delta)}, & i=1, \ldots, n-1 \\ 0, & i=n\end{cases}
$$

for all $T \in\left(0, T^{*}\right]$. If $a(.,$.$) is bounded, there exists a Markovian point process \left(M_{t}\right)$ with intensity $\gamma_{t}=a\left(t, M_{t-}\right)$ defined on some probability space $\left(\Omega_{0}, \mathcal{G},\left(\mathcal{G}_{t}\right), \mathbb{Q}_{0}\right)$ such that

$$
\forall T \in\left[0, T^{*}\right], \quad \forall i \in\{0, \ldots, n\}, \quad P(T, i \delta)=E^{\mathbb{Q}_{0}}\left[\left(i \delta-\delta M_{T}\right)^{+} \mid \mathcal{G}_{0}\right]
$$

If in addition

$$
\begin{equation*}
\nabla_{K} \partial_{T} P(T, i \delta)<0 \tag{7}
\end{equation*}
$$

the intensity function $a(T, i)$ is strictly positive for all $i<n$.
Proof. Property 1 entails that the function $a(.,$.$) defined in the theorem is nonnegative.$ Consider a standard Poisson process $\left(N_{t}\right)$ constructed on a probability space $\left(\Omega_{0}, \mathcal{G},\left(\mathcal{G}_{t}\right), \mathbb{P}\right)$. Denote by $\tau_{1}<\tau_{2}<\cdots$ the jump times of $\left(N_{t}\right)$ and set $M_{t}=N_{t \wedge \tau_{n}} .\left(\mathcal{G}_{t}\right)$ is the filtration generated by $\left(M_{t}\right)$. Now define the nonnegative predictable process

$$
\gamma_{t}:=a\left(t, M_{t-}\right)
$$

Since the function $a(.,$.$) is assumed to be bounded, we know that for all t \in\left[0, T^{*}\right]$

$$
\int_{0}^{t} \gamma_{s} d s<\infty \quad \mathbb{P} \text {-a.s. }
$$

We now apply the change of measure theorem for point processes [3, Chap. VI, sect. 2, Thm. T3] and define a new probability measure $\mathbb{Q}_{0}$ by

$$
\left.\frac{d \mathbb{Q}_{0}}{d \mathbb{P}}\right|_{\mathcal{G}_{t}}=\exp \left(\int_{0}^{t}\left(1-\gamma_{s}\right) d s\right) \prod_{\tau_{j}<t} \gamma_{\tau_{j}}
$$

Under $\mathbb{Q}_{0},\left(M_{t}\right)$ is a Markovian point process with intensity $\left(\gamma_{t}\right)$.
As shown in [7], the function $u(T, i)=E^{\mathbb{Q}_{0}}\left[\left(i \delta-\delta M_{T}\right)^{+} \mid \mathcal{G}_{0}\right]$ is a solution of (3). On the other hand, substituting $\{a(T, i)\}_{T \in\left[0, T^{*}\right], i=0, \ldots, n}$ into the forward equations (3) shows that the expected tranche notionals $\{P(T, i \delta)\}_{T \in\left[0, T^{*}\right], i=0, \ldots, n}$ solve (3). The boundedness of $a(.,$. entails that the linear system of ODEs (3) has a unique solution, so

$$
P(T, i \delta)=E^{\mathbb{Q}_{0}}\left[\left(i \delta-\delta M_{T}\right)^{+} \mid \mathcal{G}_{0}\right]
$$

for all $T \in\left[0, T^{*}\right], i=0, \ldots, n$.
Finally, under (7), the positivity of $a(T, i)$ is immediate from (6).
Formula (5) is analogous to the Dupire formula [10], which expresses the local volatility as a function of the call prices:

$$
\sigma^{2}(T, K)=\frac{2}{K^{2}} \frac{\partial_{T} C(T, K)}{\partial_{K}^{2} C(T, K)}, \quad T \geq 0, K \geq 0
$$

where $C(T, K)$ is the call price with maturity $T$ and strike $K$. In the diffusion framework, asset prices take values in $[0, \infty)$, which leads to a marginal probability density defined on $[0, \infty)$. This explains why a continuum of call prices in both strike and maturity is required to recover the local volatility function. On the other hand, since we model the portfolio loss as a point process with finite state space $(i \delta)_{i=0, \ldots, n}$, we require only a set of expected tranche notionals with $n$ strikes equal to the possible loss levels to recover the local intensity function.

Schönbucher [21] shows a similar formula expressed in terms of the marginal distribution:

$$
\begin{equation*}
a(T, i)=\frac{-\sum_{k=0}^{i} \partial_{T} \mathbb{Q}\left(L_{T}=i \delta \mid \mathcal{F}_{0}\right)}{\mathbb{Q}\left(L_{T}=i \delta \mid \mathcal{F}_{0}\right)}, \quad i=0, \ldots, n-1, \quad T \in\left(0, T^{*}\right] . \tag{8}
\end{equation*}
$$

However, formula (5) appears to have an important advantage over (8). As we will discuss in section 2, the value of a CDO tranche can be expressed as a linear combination of a small set of expected tranche notionals. In this case, recovering the expected tranche notionals from the CDO market data can be achieved efficiently and the results can be used to compute the local intensity function using formula (5). However, this procedure will become more difficult if we consider the marginal distribution because we need the marginal distribution at all loss levels to express the CDO mark-to-market values.
2. Nonparametric estimation of the local intensity function. We now present a nonparametric method for recovering the local intensity function from CDO tranche spreads. The idea is to first extract the expected tranche notionals from the CDO market data using a nonparametric approach and then compute the local intensity function by using formula (5), based on the results in Theorem 2.

We briefly recall the structure of index CDO tranches and their relationship with the expected tranche notionals [5] and introduce a quadratic programming method to recover expected tranche notionals from the CDO market data. Subsequently, we will outline the calibration algorithm for the local intensity function. We show that, unlike our proposed method, which yields an arbitrage-free pricing model, base-correlation interpolation does not guarantee the absence of arbitrage.
2.1. CDOs and expected tranche notionals. Consider a CDO tranche defined by an interval $[a, b], 0 \leq a<b \leq 1 . a$ and $b$ are called, respectively, the attachment and detachment points of the tranche and expressed in percentage of the total notional value. A synthetic CDO tranche swap is a bilateral contract in which the protection seller agrees to pay all portfolio loss within the interval $[a, b]$ in exchange for a periodic spread $s^{[a, b]}$ on the remaining notional value and an upfront payment $U^{[a, b]}$ on the initial notional value $b-a$.

Assume that the tranche is incepted at time 0 and that the spread $s^{[a, b]}$ and the upfront payment $U^{[a, b]}$ are given such that the mark-to-market value of the tranche is equal to zero. Then, we have ${ }^{1}$

$$
\begin{align*}
U^{[a, b]}(b-a)= & \sum_{j=1}^{m} D\left(0, t_{j}\right)\left[P\left(t_{j}, a\right)-P\left(t_{j}, b\right)-P\left(t_{j-1}, a\right)+P\left(t_{j-1}, b\right)\right] \\
& -s^{[a, b]} \sum_{t_{j}>0} D\left(0, t_{j}\right)\left(t_{j}-t_{j-1}\right)\left[P\left(t_{j}, b\right)-P\left(t_{j}, a\right)\right] \tag{9}
\end{align*}
$$

where $D\left(0, t_{j}\right)$ is the discount factor from time $t_{j}$ to 0 , and $0=t_{0}<t_{1}<\cdots<t_{m}$ are the payment times, where the last payment time $t_{m}$ corresponds to the expiration time. Notice that equality (9) is linear in the expected tranche notionals with strikes equal to the attachment and detachment points and maturities equal to the payment times.
2.2. Recovering expected tranche notionals via quadratic programming. Assume that at time 0 we observe the spreads and upfront payments of CDO tranches $\left[\kappa_{i-1}, \kappa_{i}\right]$ for $i=1, \ldots, I$. Without loss of generality, we assume $I \leq n^{2}$ and denote the payment times $0=t_{0}<t_{1}<\cdots<t_{m}$ with expiration time $t_{m}{ }^{3}$

Our goal is to recover a local intensity function from the CDO market data using the results in Theorem 2. For computational purposes, we approximate the derivatives in formula (5) by finite differences and therefore consider expected tranche notionals with maturities on a discrete time grid. In particular, we focus on the set of expected tranche notionals $\left\{P\left(t_{j}, i \delta\right)\right\}_{j=0, \ldots, m ; i=1, \ldots, n}$, which is represented in vector form:

$$
\mathbf{p}=\left[P\left(t_{0}, \delta\right), \ldots, P\left(t_{0}, n \delta\right), \ldots, P\left(t_{m}, \delta\right), \ldots, P\left(t_{m}, n \delta\right)\right]^{\mathrm{T}} \in \mathbb{R}^{n(m+1)}
$$

By definition, $\left\{P\left(t_{j}, i \delta\right)\right\}_{j=0, \ldots, m ; i=1, \ldots, n}$ has to satisfy Property 1. In order to apply Theorem 2, the function defined by (5) has to be bounded, which means that the denominator of formula (5) has to be strictly positive. Moreover, we impose additional condition (7) to ensure that the local intensity function is strictly positive. Taking all constraints into account, $\left\{P\left(t_{j}, i \delta\right)\right\}_{j=0, \ldots, m ; i=1, \ldots, n}$ must verify the following conditions.

Condition 1.
(a) $P\left(t_{j}, \delta\right)>0, j=1, \ldots, m$;
(b) $\nabla_{K}^{2} P\left(t_{j},(i-1) \delta\right)>0, j=1, \ldots, m, i=1, \ldots, n-1$;

[^47](c) $\nabla_{K} P\left(t_{j-1}, i \delta\right)>\nabla_{K} P\left(t_{j}, i \delta\right), j=1, \ldots, m, i=0, \ldots, n-1$.

Since the relations in Condition 1 are linear in the expected tranche notionals, they can be written in matrix form (see Appendix B):

$$
\begin{equation*}
\mathbf{B p}<\mathbf{0}, \tag{10}
\end{equation*}
$$

where $\mathbf{B}$ is a $2 n m \times n(m+1)$ matrix.
In order to be consistent with the CDO market data, the expected tranche notionals must satisfy (9) for each tranche. Although (9) may involve expected tranche notionals with strikes not equal to the multiples of $\delta$ (because attachment and detachment points may not be multiples of $\delta$ ), Property 1(f) shows that we can always compute an expected tranche notional by linearly interpolating its neighbors which have strikes equal to the multiples of $\delta$. Therefore, we can express (9) in terms of $\left\{P\left(t_{j}, i \delta\right)\right\}_{j=0, \ldots, m ; i=1, \ldots, n}$. Writing this in matrix form (see Appendix B), we have

$$
\begin{equation*}
\mathbf{A p}=\mathbf{b} \tag{11}
\end{equation*}
$$

where $\mathbf{b} \in \mathbb{R}^{I}, \mathbf{A}$ is an $I \times n(m+1)$ matrix, and both depend on the CDO market data and the discount factors. Thus, calibrating a set of expected tranche notionals that satisfies Condition 1 and is consistent with the CDO data is equivalent to finding a solution of the linear system (10)-(11). However, this system has either no or infinitely many solutions.

Proposition 1. Given the CDO market data, there are either no or infinitely many sets of expected tranche notionals with maturities $t_{0}<\cdots<t_{m}$ and strikes $\delta<\cdots<n \delta$ which satisfy (10)-(11).

In order to pinpoint a unique set of expected tranche notionals, we consider a convex optimization problem under constraints (10)-(11). Considering a convex function $f$ : $\mathbb{R}_{+}^{n(m+1)} \mapsto \mathbb{R}$, we calibrate the expected tranche notionals by solving

$$
\begin{equation*}
\min _{\mathbf{p}} f(\mathbf{p}) \quad \text { subject to } \quad \mathbf{A} \mathbf{p}=\mathbf{b}, \quad \mathbf{B} \mathbf{p} \leq-\mathbf{e} . \tag{12}
\end{equation*}
$$

Here, we replace the strict inequalities in (10) by inequalities with an error vector $\mathbf{e}>0$ that can be chosen arbitrarily. In fact, if (10)-(11) has a solution, then we can always pick $\mathbf{e}$ such that (12) is feasible. If there exists a set of expected tranche notionals $\mathbf{p}$ that satisfies Condition 1 and is consistent with the CDO market data, solving (12) gives us a unique solution. In particular, we propose as selection criterion

$$
\begin{equation*}
f(\mathbf{p})=\sum_{j=0}^{m} \sum_{i=1}^{n} w_{i j}\left(P\left(t_{j}, i \delta\right)-\widetilde{P}\left(t_{j}, i \delta\right)\right)^{2}, \tag{13}
\end{equation*}
$$

where $\left(w_{i j}\right)_{j=0, \ldots, m-1 ; i=1, \ldots, n-1}$ are weights, and $\left\{\widetilde{P}\left(t_{j}, i \delta\right)\right\}_{j=0, \ldots, m ; i=1, \ldots, n-1}$ is a reference set of expected tranche notionals. Using this objective function, we can impose a "prior" view by choosing the reference expected tranche notionals; e.g., the reference expected tranche notionals can be computed from a particular credit model. Furthermore, (12) reduces to a quadratic programming problem which can be solved efficiently [2, 22].
2.3. Numerical issues. Notice that CDO tranche payments are typically made every quarter. In this case, the expected tranche notionals $\left\{P\left(t_{j}, i \delta\right)\right\}_{j=0, \ldots, m ; i=1, \ldots, n}$ obtained by solving (12) are sparsely spaced in maturity. In order to obtain a finer set of expected tranche notionals, a simple method is to linearly interpolate $\left\{P\left(t_{j}, i \delta\right)\right\}_{j=0, \ldots, m ; i=1, \ldots, n}$ across maturities. This guarantees that the finer set will also satisfy Condition 1. However, this method will give extremely large values to the local intensity function for short maturities and may lead to computational instability. The reason is the following.

Assume we want to compute the expected tranche notionals $\left\{P\left(T_{j}, i \delta\right)\right\}_{j=0, \ldots, q ; i=1, \ldots, n}$ on the finer time grid $\left(T_{j}\right)_{j=0, \ldots, q}$, which includes the payment times $\left(t_{j}\right)_{j=0, \ldots, m}$. For $i>0$, we compute $P\left(T_{1}, i \delta\right)$ and $P\left(T_{2}, i \delta\right)$, where $T_{1}<T_{2}<t_{1}$ by linearly interpolating the values $P\left(t_{0}, i \delta\right)$ and $P\left(t_{1}, i \delta\right)$. Then, the local intensity function computed from (5) is equal to

$$
a\left(T_{1}, i\right)=\frac{\left(-\nabla_{K} P\left(T_{2}, i \delta\right)+\nabla_{K} P\left(T_{1}, i \delta\right)\right) /\left(T_{2}-T_{1}\right)}{\nabla_{K}^{2} P\left(T_{1},(i-1) \delta\right)},
$$

where we approximate the partial derivatives by finite differences. Since we compute $P\left(T_{2}, i \delta\right)$ and $P\left(T_{1}, i \delta\right)$ by linear interpolation, the numerator of $a\left(T_{1}, i\right)$ is strictly positive and has the same value for any $T_{1}<T_{2}<t_{1}$. On the other hand, if $T_{1}$ is close to 0 , the denominator $\nabla_{K}^{2} P\left(T_{1},(i-1) \delta\right)$ is also close to 0 because $K \mapsto P(0, K)$ is linear. Therefore, $a\left(T_{1}, i\right)$ becomes extremely large when $T_{1}$ is small.

To overcome this problem, we propose the following method.

## AlGORITHM 1. Maturity interpolation of expected tranche notionals.

1. Construct an arbitrage-free set $\left\{P\left(t_{j}, i \delta\right)\right\}_{j=0, \ldots, m ; i=0, \ldots, n}$ of expected tranche notionals by solving (12).
2. Fix an integer $r$ such that $n \leq r<q$ and $0=T_{0}<T_{1}<\cdots<T_{r}<t_{1}$. Arbitrarily set a positive but sufficiently small value for the local intensity function at times $\left(T_{j}\right)_{j=0, \ldots, r-1}$ such that $\nabla_{K} P\left(T_{r}, i \delta\right)>\nabla_{K} P\left(t_{1}, i \delta\right)$ for all $i$.
3. Compute the expected tranche notionals for maturities $\left(T_{j}\right)_{j=0, \ldots, r}$ using forward equations (3). Note that the set of expected tranche notionals at the maturity $T_{r}$, $\left\{P\left(T_{r}, i \delta\right)\right\}_{i=0, \ldots, n}$, automatically satisfies the strict convexity constraint in Condition 1(b).
4. Linearly interpolate expected tranche notionals in maturity starting from maturity $T_{r}$.

Since the linear interpolation starts from maturity $T_{r}$ and $\left\{P\left(T_{r}, i \delta\right)\right\}_{i=0, \ldots, n}$ satisfies the strict convexity constraint in Condition 1(b), the denominator in the local intensity function formula (5) is strictly positive at maturity $T_{r}$. Therefore, having extremely large values for the local intensity function at short maturities is avoided.

Remark 1. The purpose of Algorithm 1 is to compute the expected tranche notionals on a finer time grid but not to build a complete set of expected tranche notionals. In particular, if $T \mapsto P(T, i \delta)$ is piecewise linear, it will no longer be differentiable at certain points. If one is interested in recovering a complete set of expected tranche notionals and applying Theorem 2, all necessary conditions must be verified carefully.
2.4. Calibration algorithm for the local intensity function. The above considerations lead to the following algorithm for computing a local intensity function from a discrete set of

CDO tranche spreads:
AlGORITHM 2. Quadratic programming calibration for the local intensity function.

1. Compute matrices $\mathbf{A}$ and $\mathbf{b}$ in (11) according to the CDO market data and matrix $\mathbf{B}$ in (10) according to Condition 1 (see Appendix B).
2. Solve quadratic programming problem (12) with objective function (13) and obtain a set of expected tranche notionals which is consistent with the CDO market data.
3. Apply Algorithm 1 to obtain expected tranche notionals on a finer time grid if desired.
4. Convert the calibrated expected tranche notionals into a local intensity function using formula (5).
2.5. Arbitrage opportunities when using base-correlation interpolation. To price nonstandard CDO tranches, or equivalently expected tranche notionals, it is common to use the base-correlation interpolation method under the Gaussian copula framework. For example, if we want to price tranches $[5 \%, 6 \%]$ and $[6 \%, 7 \%]$ of the iTraxx investment grade (IG) portfolio, we first calibrate the one-factor Gaussian copula model [17] to the standard tranches and obtain the base correlations [16] at the standard strikes $3 \%, 6 \%, 9 \%, 12 \%$, and $22 \%$. Then, we interpolate the base correlations for other strikes, say $5 \%$ and $7 \%$ in this example. After that, we compute the expected tranche notionals at strikes $5 \%$ and $7 \%$ using the two different base correlations obtained by interpolation and price the corresponding CDO tranches. However, we show that this method does not guarantee absence of arbitrage.

Figure 2 shows the base correlations of the one-factor Gaussian copula model calibrated to the iTraxx data in Table 2. By linearly interpolating the base correlations, we compute the upfront payments of nonstandard tranches [ $5 \%, 6 \%$ ] and $[6 \%, 7 \%]$ with a fixed periodic spread 100 bps in Table 1. At first glance, we can see that the upfront payment of the more senior tranche $[6 \%, 7 \%]$ is larger than the upfront payment of tranche [ $5 \%, 6 \%$ ]. To show that it leads to an arbitrage opportunity, we take the following positions:

- we buy protection on tranche $[5 \%, 6 \%]$;
- we sell protection on tranche [ $6 \%, 7 \%$ ].

At inception time $t_{0}$, the cash flow of our positions is equal to the difference of the upfront payments:

$$
(879.3 \mathrm{bps}-819.5 \mathrm{bps})(1 \%)=0.598 \mathrm{bps},
$$

which is strictly positive.
At payment time $t_{j}$, the net premium received from our positions, $\operatorname{Prem}\left(t_{j}\right)$, is equal to

$$
\begin{aligned}
\operatorname{Prem}\left(t_{j}\right)= & 100 \operatorname{bps}\left(t_{j}-t_{j-1}\right)\left[\left(7 \%-L_{t_{j}}\right)^{+}-\left(6 \%-L_{t_{j}}\right)^{+}\right] \\
& -100 \operatorname{bps}\left(t_{j}-t_{j-1}\right)\left[\left(6 \%-L_{t_{j}}\right)^{+}-\left(5 \%-L_{t_{j}}\right)^{+}\right] \\
= & 100 \operatorname{bps}\left(t_{j}-t_{j-1}\right)\left[\left(5 \%-L_{t_{j}}\right)^{+}-2\left(6 \%-L_{t_{j}}\right)^{+}+\left(7 \%-L_{t_{j}}\right)^{+}\right]
\end{aligned}
$$

which is positive, due to the fact that $K \mapsto(K-L)^{+}$is a convex function. Therefore, the value of the premium leg at expiration time $t_{m}$ is positive and equal to

$$
\text { value of premium leg at expiration }=\sum_{j=1}^{m} D\left(t_{j}, t_{m}\right)^{-1} \operatorname{Prem}\left(t_{j}\right) \geq 0,
$$



Figure 2. Base correlations of one-factor Gaussian copula model. Data: 5 Y iTraxx Europe $I G S 9$ on 25 March 2008.

## Table 1

Upfront payments of 5-year iTraxx IG CDO tranches with 100 bps periodic spread. Pricing method: onefactor Gaussian copula model with linearly interpolated base correlations.

| Tranche | $5 \%-6 \%$ | $6 \%-7 \%$ |
| :---: | :---: | :---: |
| Upfront payment | 819.5 bps | 879.3 bps |

where $D\left(t_{j}, t_{m}\right)$ is the risk-free zero coupon bond price at time $t_{j}$ with maturity $t_{m}$.
On the other hand, at payment time $t_{j}$, the net default payment received from the positions, $\operatorname{Def}\left(t_{j}\right)$, is equal to

$$
\begin{aligned}
\operatorname{Def}\left(t_{j}\right)= & {\left[\left(5 \%-L_{t_{j}}\right)^{+}-2\left(6 \%-L_{t_{j}}\right)^{+}+\left(7 \%-L_{t_{j}}\right)^{+}\right] } \\
& -\left[\left(5 \%-L_{t_{j-1}}\right)^{+}-2\left(6 \%-L_{t_{j-1}}\right)^{+}+\left(7 \%-L_{t_{j-1}}\right)^{+}\right]
\end{aligned}
$$

and the value of the default leg at expiration time $t_{m}$ is equal to

$$
\text { value of default leg at expiration }=\sum_{j=1}^{m} D\left(t_{j}, t_{m}\right)^{-1} \operatorname{Def}\left(t_{j}\right)
$$

In section A.4, we show that the value of the default leg is also positive. Therefore, our total payoff (sum of premium leg and default leg) at expiration is positive. Recall that at the inception of the tranches at time $t_{0}$ we received a strictly positive cash flow. As a result, our trading strategy shows an arbitrage opportunity.

In contrast to our method in Algorithm 2, this example illustrates that interpolation of base correlation can result in arbitrage opportunities.
3. Application to iTraxx tranches. In this section, we illustrate the calibration method introduced in section 2 by using the iTraxx data. We refer to this calibration method as the quadratic programming method (QP). The weights in the objective function (13) are chosen to be proportional to the reference expected tranche notionals which are computed from a flat local intensity function equal to 1 . In addition, we compare our results to two alternative calibration methods:

- A parametric method.

Herbertsson [13] specifies the local intensity function in the parametric form

$$
\begin{equation*}
a(t, i)=(n-i) \sum_{k=0}^{i} b_{k}, \tag{11}
\end{equation*}
$$

where the parameters $b_{0}, \ldots, b_{n}$ are constants. This model states that the local intensity function is constant except when defaults occur. There is no sign restriction on $b_{k}$ as long as the local intensity function remains nonnegative. To recover the parameters from the CDO spreads, we minimize the sum across observed tranches of squared differences between model and market CDO spreads using a gradient-based algorithm. Note that this objective function is not convex; hence the gradient-based algorithm may not necessarily yield a global minimum, and the minimum need not be unique.

- Entropy minimization algorithm.

Cont and Minca [6] introduced a nonparametric method for recovering the local intensity function as the solution of a relative entropy minimization problem

$$
\inf _{\mathbb{Q} \in \Lambda} \mathbb{E}^{\mathbb{Q}_{0}}\left[\frac{d \mathbb{Q}}{d \mathbb{Q}_{0}} \ln \left(\frac{d \mathbb{Q}}{d \mathbb{Q}_{0}}\right)\right]
$$

subject to calibration constraints (9). $\mathbb{Q}_{0}$ denotes the law of a prior Markovian point process, and $\Lambda$ is the set of laws of Markovian point processes, equivalent to $\mathbb{Q}_{0}$. To implement this algorithm, we choose as prior measure the law of a standard Poisson process stopped at $n$.
These three methods are applied to the 5 -year iTraxx Europe IG Index CDO tranche spreads on 20 September 2006 and 25 March 2008, a portfolio consisting of 125 names. The recovery rate is assumed to be $R=40 \%$. Table 2 shows the result of the calibration.

Table 2
Calibrated CDO tranche spreads of 5 Y iTraxx Europe IG Series 6 on 20 September 2006 and Series 9 on 25 March 2008. Quotes are given in bps, except for equity tranches, which are quoted as upfront in percentage with 500 bps periodic coupons.

|  | $0 \%-3 \%$ | $3 \%-6 \%$ | $6 \%-9 \%$ | $9 \%-12 \%$ | $12 \%-22 \%$ | $22 \%-100 \%$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 20-Sep-06 |  |  |  |  |  |  |
| Market bid | $11.8 \%$ | 53.8 | 14.0 | 5.8 | 2.1 | 0.8 |
| Market ask | $12.0 \%$ | 55.3 | 15.5 | 6.8 | 2.9 | 1.3 |
| QP | $11.9 \%$ | 54.6 | 14.8 | 6.3 | 2.5 | 1.0 |
| Parametric | $11.9 \%$ | 54.5 | 14.8 | 6.3 | 2.5 | 1.1 |
| Entropy min | $11.9 \%$ | 54.5 | 14.8 | 6.3 | 2.5 | 1.1 |
| 25-Mar-08 |  |  |  |  |  |  |
| Market bid | $37.7 \%$ | 441.6 | 270.2 | 174.4 | 97.4 | 42.8 |
| Market ask | $39.7 \%$ | 466.6 | 290.2 | 189.4 | 110.7 | 46.9 |
| QP | $38.4 \%$ | 451.9 | 279.0 | 181.1 | 103.2 | 44.3 |
| Parametric | $38.7 \%$ | 454.1 | 280.2 | 181.9 | 104.1 | 44.8 |
| Entropy min | $38.6 \%$ | 453.3 | 279.5 | 181.2 | 103.4 | 44.6 |



Figure 3. Local intensity functions implied from 5 Y iTraxx Europe $I G$ tranche spreads using the quadratic programming method (Algorithm 2) Left: S6 on 20 September 2006. Right: S9 on 25 March 2008.


Figure 4. Local intensity functions implied from 5 Y iTraxx Europe IG tranche spreads using the parametric model [13]. Left: S6 on 20 September 2006. Right: S9 on 25 March 2008.


Figure 5. Local intensity functions implied from 5 Y iTraxx Europe IG tranche spreads using the nonparametric entropy minimization method [6]. Left: S6 on 20 September 2006. Right: S9 on 25 March 2008.
3.1. Local intensity functions. Figures 3 to 5 show local intensity functions implied from iTraxx CDO spreads using the three different approaches presented above on two different dates: 20 September 2006 and 25 March 2008. These local intensity functions exhibit qualita-
tively similar features, but the value they imply for the default intensity can be quite different. Also, for each calibration method, we observe that the general shape of the local intensity function calibrated to the 2006 dataset is similar to the one calibrated to the 2008 dataset.

For the quadratic programming method (Figure 3), we observe that, at any fixed time, the local intensity functions stay at a low level when the number of defaults is small but sharply increases around 5 defaults: this sharp increase signals the onset of contagion. After that, the local intensity functions stay almost flat at a high level when the number of defaults is larger than 5 . The term structure is relatively flat in all examples.

The parametric approach (Figure 4) yields, by construction, smoother local intensity functions. When the loss is large, unlike the local intensity functions obtained via quadratic programming, the parametric local intensity functions decrease gradually towards zero as the number of defaults increases. Interestingly, this feature is also observed in the local intensity functions obtained with the nonparametric entropy minimization algorithm (Figure 5), but the decrease is much faster for short times than in the parametric model. Furthermore, the maximum attained value of the local intensity function obtained via the entropy minimization algorithm is substantially lower than by both the parametric and quadratic programming methods.
3.2. Stability analysis. A crucial property of a calibration method is its stability with respect to the inputs. To examine the stability of the various calibration methods considered above, we apply a $1 \%$ proportional shift to all CDO market spreads, recalibrate the local intensity function to the shifted CDO spreads, and measure the magnitude of the changes using the Frobenius norm:

$$
\left(\sum_{i=0}^{n} \sum_{j=0}^{q}\left|a\left(T_{j}, i\right)-\widehat{a}\left(T_{j}, i\right)\right|^{2}\right)^{1 / 2}
$$

where $\left\{a\left(T_{j}, i\right)\right\}$ and $\left\{\widehat{a}\left(T_{j}, i\right)\right\}$ are, respectively, the local intensity functions calibrated to the original and perturbed CDO tranche spreads. The smaller the value of this norm, the more stable is the method. From Table 3, we observe that the entropy minimization algorithm is substantially more stable than the other two methods with respect to a change in the inputs, while the parametric approach is the most unstable one among the three. This result is in line with findings in similar studies using equity derivatives [8].

Table 3
Frobenius norm of the changes in the local intensity function with respect to $1 \%$ proportional increase in the CDO spreads. Data: 5 Y iTraxx Europe IG S6 on 20 September 2006 and S9 on 25 March 2008.

|  | QP | Parametric | Entropy min |
| :---: | :---: | :---: | :---: |
| 20-Sep-06 | 56.2 | 32116.2 | $2.0 \times 10^{-2}$ |
| 25-Mar-08 | 673.2 | 728.3 | $2.0 \times 10^{-1}$ |

3.3. Marginal distributions and expected losses. Figure 6 shows the marginal distribution of the default process on 25 March 2008. We see that the marginal distributions are similar across the three calibration methods at year 1 but have more significant differences


Figure 6. Marginal distribution at year 1 and year 4, truncated up to 20 defaults. Data: 5 Y iTraxx Europe IG S9 on 25 March 2008.


Figure 7. Term structure of the differences in the expected loss $E\left(L_{T}\right)-E\left(L_{T-0.25}\right)$. Data: 5 Y iTraxx Europe IG S9 on 25 March 2008.
for longer times at year 4. This suggests that pricing non-path-dependent credit derivatives is more sensitive to the choice of the calibration method used to recover the local intensity function for longer maturities.

Another important quantity that we study is the expected portfolio loss. Figure 7 shows the differences of the expected losses on a quarterly basis, i.e., $E\left(L_{T}\right)-E\left(L_{T-0.25}\right)$. Observe that the differences in the expected losses are almost flat for all calibration methods when the time is less than 2 years. When time increases, the difference of the expected loss computed from the parametric method increases gradually. On the other hand, the other two calibration methods give similar differences of the expected loss along time, except a sharp increase at year 5 for the quadratic programming method.
3.4. Forward starting tranche spreads. Forward starting tranches provide protection against tranche losses in a prespecified future period $[t, T]$. The distinguishing feature is that defaults occurring prior to the starting date do not affect the subordination of the tranche. In particular, a forward tranche with attachment-detachment interval $[a, b]$ can be valued as the forward value of a tranche with adjusted interval $\left[a^{\prime}, b^{\prime}\right]$, where $a^{\prime}=\min \left(1, a+L_{t}\right)$ and $b^{\prime}=\min \left(1, b+L_{t}\right)$. This dependence of the payoff on the loss makes the forward tranche path dependent.

Table 4 shows the spreads of forward starting tranches which start in year 1 and mature 3 years later. As we can see, the forward tranche spreads are significantly different across the

Table 4
Spreads of forward starting tranches which start in year 1 and mature 3 years later. Data: 5 Y iTraxx Europe IG S6 on 20 September 2006 and S9 on 25 March 2008.

|  | 20 September 2006 |  |  | 25 March 2008 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | QP | Parametric | Entropy min | QP | Parametric | Entropy min |
| $0 \%-3 \%$ | 12.05 | 12.25 | 14.26 | 53.46 | 36.92 | 65.92 |
| $3 \%-6 \%$ | 2.72 | 17.89 | 33.62 | 93.79 | 290.65 | 482.23 |
| $6 \%-9 \%$ | 2.46 | 3.18 | 7.46 | 92.46 | 142.25 | 236.22 |
| $9 \%-12 \%$ | 2.21 | 0.79 | 4.14 | 91.45 | 63.45 | 170.80 |
| $12 \%-22 \%$ | 1.59 | 0.36 | 4.03 | 89.36 | 34.49 | 165.59 |
| $22 \%-100 \%$ | 0.03 | 0.15 | 0.69 | 37.99 | 13.38 | 27.60 |

Table 5
Model uncertainty ratio of the forward starting tranche spreads.

|  | $0 \%-3 \%$ | $3 \%-6 \%$ | $6 \%-9 \%$ | $9 \%-12 \%$ | $12 \%-22 \%$ | $22 \%-100 \%$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 20-Sep-06 | $17 \%$ | $171 \%$ | $115 \%$ | $141 \%$ | $184 \%$ | $226 \%$ |
| 25-Mar-08 | $56 \%$ | $134 \%$ | $92 \%$ | $99 \%$ | $136 \%$ | $93 \%$ |

calibration methods. Table 5 shows the model uncertainty ratio [4], which is defined as

$$
\text { model uncertainty ratio }=\frac{s^{\max }-s^{\min }}{s^{\text {ave }}}
$$

where $s^{\max }, s^{\text {min }}$, and $s^{a v e}$ are the maximum, minimum, and average forward spreads, respectively, across the calibration methods. Notice that the ratio is larger than $90 \%$ for all tranches except for the equity tranche. This reveals a serious problem because, even if we price exotic credit derivatives in the same modeling framework (local intensity framework in this case), there is substantial uncertainty in pricing exotic credit derivatives due to the choice of calibration method.
3.5. Jump-to-default ratios. In the local intensity framework, the market is complete and the self-financing strategy to replicate the payoff of a CDO tranche involves trading the underlying index default swap. The corresponding hedge ratio, which is known as the jump-to-default ratio, is defined by

$$
\frac{v^{[a, b]}\left(t, N_{t}+1\right)-v^{[a, b]}\left(t, N_{t}\right)}{v^{\text {index }}\left(t, N_{t}+1\right)-v^{\text {index }}\left(t, N_{t}\right)},
$$

where $v^{[a, b]}(t, m)$ and $v^{\text {index }}(t, m)$ denote the mark-to-market values per unit notional of tranche $[a, b]$ and the index default swap, respectively, conditional on $m$ defaults occurring by time $t$. More details on this subject can be found in [5, 15].

Table 6 shows the jump-to-default ratios computed from the local intensities in section 3.1. Interestingly, we observe that the jump-to-default ratios generated by the quadratic programming and the entropy minimization methods are quite similar. However, substantial differences are observed when comparing to the parametric method. This implies that uncertainty due to the choice of the calibration method not only affects the pricing of credit derivatives, as we have shown in section 3.4, but may also have a large impact on hedging strategies for portfolio credit derivatives [5].

Table 6
Jump-to-default ratios computed from the calibrated local intensity functions. Data: 5 Y iTraxx Europe $I G$ S6 on 20 September 2006 and S9 on 25 March 2008.

|  | 20 September 2006 |  |  | 25 March 2008 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | QP | Parametric | Entropy min | QP | Parametric | Entropy min |
| $0 \%-3 \%$ | 6.29 | 20.97 | 6.32 | 1.03 | 3.62 | 1.60 |
| $3 \%-6 \%$ | 2.12 | 5.16 | 3.51 | 1.69 | 3.31 | 2.33 |
| $6 \%-9 \%$ | 1.63 | 2.00 | 2.23 | 1.68 | 2.65 | 2.15 |
| $9 \%-12 \%$ | 1.52 | 1.02 | 1.72 | 1.68 | 2.08 | 1.97 |
| $12 \%-22 \%$ | 1.47 | 0.48 | 1.39 | 1.68 | 1.48 | 1.76 |
| $22 \%-100 \%$ | 0.67 | 0.22 | 0.61 | 0.81 | 0.66 | 0.75 |

4. Local intensity function implied by credit portfolio loss models. Market practice has been to calibrate credit portfolio models to market observations of index spreads and index tranche spreads and use the resulting parameters to price nonstandard or illiquid products. As observed in section 3, even in a local intensity model, the marginal distributions for the portfolio loss generated by the model can vary substantially depending on the calibration methods. This raises the question of whether there is also a substantial difference of loss distributions across different models when these models are calibrated to the same market data.

We compare the local intensity functions implied by six different models: Herbertsson model [13], bivariate spread-loss model [1], shot-noise model [12], one-factor Gaussian copula model [17], one-factor Student-t copula model [14], and bottom-up affine jump-diffusion model $[9,19,11]$. The first three are top-down models, which means that the portfolio default intensity is directly specified. The one-factor Gaussian and Student-t copula models are bottom-up static factor models, and the affine jump-diffusion model is a dynamic bottom-up model.

All models are calibrated to the iTraxx Europe IG Series 9 CDO data on 25 March 2008. Table 7 shows the calibration results. The Gaussian and Student-t copula models are calibrated using the base-correlation method [16]. Except for the shot-noise model and the affine jump-diffusion model, all models yield tranche spreads well within the bid-ask intervals.

Table 7
Calibration of different models to 5 Y iTraxx Europe IG Series 9 tranche spreads on 25 March 2008. Quotes are given in bps, except for equity tranches, which are quoted as upfront in percent with 500 bps periodic coupons.

|  | $0 \%-3 \%$ | $3 \%-6 \%$ | $6 \%-9 \%$ | $9 \%-12 \%$ | $12 \%-22 \%$ | $22 \%-100 \%$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 25-Mar-08 |  |  |  |  |  |  |
| Market bid | $37.7 \%$ | 441.6 | 270.2 | 174.4 | 97.4 | 42.8 |
| Market ask | $39.7 \%$ | 466.6 | 290.2 | 189.4 | 110.7 | 46.9 |
| Bivariate spread-loss model | $38.7 \%$ | 454.1 | 280.2 | 181.9 | 104.1 | 44.8 |
| Shot-noise | $43.8 \%$ | 463.5 | 219.5 | 159.9 | 128.0 | 40.7 |
| Gaussian copula | $38.7 \%$ | 454.1 | 280.2 | 181.9 | 104.1 | 43.3 |
| Student-t copula | $38.7 \%$ | 454.1 | 280.2 | 181.9 | 104.1 | 44.9 |
| Affine jump-diffusion | $48.2 \%$ | 493.0 | 244.8 | 186.5 | 154.4 | 37.2 |

In order to compute the local intensity functions, we first compute the expected tranche notionals for each model. For the dynamic models, we apply Theorem 1 and convert the


Figure 8. Local intensity functions implied by credit portfolio loss models. Data: 5 Y iTraxx Europe IG S9 on 25 March 2008.
expected tranche notionals into local intensity functions based on formula (5). For the copula models, we compute the implied portfolio default intensity using Theorem 2: the local intensity function is computed using (6). The local intensity functions are shown in Figure 8.
4.1. Herbertsson model. In section 3, we have presented the parametric model introduced by Herbertsson [13], in which the portfolio default intensity has the functional form (14). Since the portfolio default intensity depends only on the credit portfolio loss level, this is one of the "simplest" models in the sense that the effective default intensity $\left(\lambda_{t}^{\text {eff }}\right)$ is the same as the portfolio default intensity $\left(\lambda_{t}\right)$. It can serve as a benchmark to compare with other models.
4.2. Bivariate spread-loss model. Arnsdorff and Halperin [1] introduce the bivariate spread-loss model in which the portfolio default intensity depends not only on the loss process but also on a mean-reverting diffusion process $\left(Y_{t}\right)$ which generates spread volatility. The portfolio default intensity is given by

$$
\lambda_{t}=Y_{t} F\left(N_{t}\right),
$$



Figure 9. Change of the local intensity function implied by the bivariate spread-loss model with respect to a small increase of the risk factor value $Y_{0}$ at time 0 .
where $F$ is called the contagion function. The factor $\left(Y_{t}\right)$ generates spread volatility between default dates and follows

$$
d \ln Y_{t}=\kappa\left(b-\ln Y_{t}\right) d t+\sigma d W_{t}
$$

where $\left(W_{t}\right)$ is a standard Brownian motion.
From Figure 8, we observe that the local intensity functions obtained from the bivariate spread-loss model and Herbertsson's parametric model are similar. The reasons is that Arnsdorff and Halperin parameterize the contagion function $F$ in the same way as Herbertsson specifies the local intensity function.

Another interesting feature to investigate is how the local intensity function changes with respect to the initial value of the risk factor $Y_{0}$. In Figure 9, we show the time evolution of the difference of the local intensity function $a(t, i)$ calibrated with two different risk factors $Y_{0}$. We represent only this difference for $i=0$ and $i=5$ and observe that a change in the risk factor will mostly affect the local intensity function at short times. This observation is consistent with the fact that, since the risk factor is specified as mean reverting, it will revert back to its average in the long run and give similar local intensity functions for longer times.
4.3. Shot-noise model. In order to better capture the possibility of extreme events, Gaspar and Schmidt [12] propose the shot-noise model in which the portfolio default intensity has the form

$$
\lambda_{t}=\eta_{t}+J_{t}
$$

where $\left(\eta_{t}\right)$ is a continuous affine process

$$
d \eta_{t}=\kappa\left(b-\eta_{t}\right) d t+\sigma \sqrt{\eta_{t}} d W_{t}
$$

with $\left(W_{t}\right)$ a standard Brownian motion, and $\left(J_{t}\right)$ is a non-Gaussian Ornstein-Uhlenbeck process which represents the shot-noise:

$$
J_{t}=\sum_{\tau_{i} \leq t} Y_{i} e^{-\alpha\left(t-\tau_{i}\right)}
$$

where $\tau_{i}, i=1,2, \ldots$, are the jump times of a Poisson process and $Y_{i}, i=1,2, \ldots$, are independent and identically distributed variables independent of the $\tau_{i}$. We assume here
that $Y_{i}$ has an exponential distribution. Under this setting, they provide a semianalytical expression for the local intensity function:

$$
\begin{equation*}
a(T, k)=\frac{\left.\frac{\partial^{k}}{\partial \theta^{k}}\right|_{\theta=-1} \frac{\partial}{\partial T} \frac{1}{\theta} S(\theta, T)}{\left.\frac{\partial^{k}}{\partial \theta^{k}}\right|_{\theta=-1} S(\theta, T)}, \tag{15}
\end{equation*}
$$

where $S(\theta, T)$ is the Laplace transform of the cumulative portfolio default intensity

$$
S(\theta, T)=E^{\mathbb{Q}}\left[e^{\theta \int_{0}^{T} \lambda_{s} d s} \mid \mathcal{F}_{0}\right], \quad \theta<0
$$

which has a closed-form expression (see Appendix C). However, formula (15) is of little use for computing the local intensity function for a typical credit portfolio which consists of more than 100 names because we have to (numerically) differentiate $S(\theta, T)$ with respect to $\theta$ more than 100 times, which yields an unstable result. This example illustrates the difficulty of computing the local intensity function, even with a semianalytical formula. On the other hand, we can easily overcome this problem by using the inversion formula (5).

In Figure 8, we see that the local intensity function implied by the shot-noise model first increases sharply when the number of defaults is small and then slowly increases when the number of defaults gets larger. This is consistent with the argument given by Gaspar and Schmidt [12]: since the default intensity $\left(\lambda_{t}\right)$ is not observed, the loss level $\left(L_{t}\right)$ is used as a statistic to estimate the default intensity. If the loss increases, it is more likely that the default intensity is high. Therefore, it leads to an increasing local intensity function in the loss level.
4.4. Gaussian and Student-t copula models. Because of its tractability, the one-factor Gaussian copula model has been the financial industry benchmark despite some well-known drawbacks. The Student-t copula model, which embeds the Gaussian copula model as a limit when the degree of freedom goes to infinity, is widely used as well. More precisely, given a family of marginal default time distributions $\left(F_{i}, i=1, \ldots, n\right)$, the joint distribution of the default times $\tau_{i}$ is modeled by first defining latent factors

$$
X_{i}=S\left(\rho Z_{0}+\sqrt{1-\rho^{2}} Z_{i}\right), \quad \text { with } S=\left\{\begin{array}{cl}
1 & \text { for Gaussian copula } \\
\sqrt{\frac{\nu}{V}} & \text { for Student-t copula }
\end{array}\right.
$$

where $Z_{0}, Z_{i} \sim \mathcal{N}(0,1), V \sim \chi_{\nu}^{2}$ are independent variables, and then defining the default times by

$$
\tau_{i}=F_{i}^{-1}\left(F_{X_{i}}\left(X_{i}\right)\right),
$$

where $F_{X_{i}}($.$) denotes the distribution of X_{i}$. We refer readers to [14] for details.
To study the local intensity function implied by these two bottom-up models, we first calibrate them using the base-correlation method [16]. Then, we study the local intensity function corresponding to different base correlations. Notice that these two models are static, which means that there is no default intensity defined in this framework: we are in effect representing the expected tranche notionals in these models in terms of an equivalent local
intensity function, which then enables us to compare these static models with the dynamic models presented above.

Figure 8 shows the local intensity functions implied by the two copula models with base correlations corresponding to tranche [ $6 \%, 9 \%$ ]. Observe that the main difference between the two local intensity functions is that for the Gaussian copula model there is a sharp increase for short times when the number of defaults is larger than 100. Other than that, both local intensity functions appear to have a smooth dome shape. This relatively restricted form of local intensity function can explain why a single correlation cannot fit the full set of CDO tranche spreads and why we usually observe a base-correlation skew. Also, since the Student-t copula embeds the Gaussian copula as a limit, it is not surprising that the local intensity functions implied by the two models are similar in general shape. This also suggests that the additional degree of freedom in the Student-t copula is still not able to generate a flexible enough local intensity function to match the full set of CDO market data.
4.5. Affine jump-diffusion model. Many "bottom-up" reduced form models [9, 11, 19, 20] based on diffusive or jump-diffusion dynamics for default intensities have been proposed for pricing portfolio credit derivatives. Most of these models are built in the "doubly stochastic" framework by specifying the default intensities for each name in the portfolio. A prominent example, which lends itself to implementation, is the model proposed by Duffie and Gârleanu [9], where the default intensities follow correlated affine jump-diffusion processes. We consider the extension of this model considered in Mortensen [19] here. The default intensity for name $i$ is represented as

$$
\lambda_{i, t}=X_{t}^{i}+a_{i} X_{t}^{0}
$$

where $a_{1}, \ldots, a_{n}$ are parameters, and $\left(X_{t}^{i}\right), i=0, \ldots, n$, are independent affine jump diffusions with

$$
d X_{t}^{i}=\kappa_{i}\left(b_{i}-X_{t}^{i}\right) d t+\sigma_{i} \sqrt{X_{t}^{i}} d W_{t}^{i}+d J_{t}^{i}
$$

where $\left(W_{t}^{i}\right), i=0, \ldots, n$, are independent Brownian motions and $\left(J_{t}^{i}\right), i=0, \ldots, n$, are independent compound Poisson processes with exponentially distributed jumps. This general specification is theoretically appealing, but the calibration to 125 individual credit default swap spreads and six tranche spreads of a CDO involves the solution to a nonlinear optimization problem in dimension 881: 125 factor loadings, 126 initial risk factor values, and 630 parameters for the risk factor dynamics. Eckner [11] proposes a parsimonious version of this model, which we will adopt here.

Interestingly, Figure 8 shows that the local intensity function implied by the affine jumpdiffusion model [11] is similar to the one implied by the one-factor Student-t copula model. This is a surprising result because the two modeling frameworks are fundamentally different: one is dynamic, while the other one is static.
5. Conclusion. We have proposed a simple and efficient calibration method for recovering the default intensity of a portfolio from CDO spreads. Our method is based on two ingredients: a nonparametric method based on quadratic programming for recovering expected tranche notionals from CDO spreads, and an inversion formula for computing the local intensity function from the expected tranche notionals. This method is shown to be much more stable, with respect to changes in inputs, than the commonly used nonlinear least squares method
based on parametric models (see, e.g., [13]). Contrarily to the base-correlation method, our method yields an arbitrage-free model.

Comparing our calibration algorithm to a parametric calibration method [13] and to a nonparametric entropy minimization method [6] using iTraxx Europe index CDO spreads, we observe that these different calibration methods lead to quite different values of default intensity while maintaining a good match to the observations: this illustrates clearly the illposedness of the calibration problem. We also find that model-dependent quantities such as forward starting tranche spreads and jump-to-default ratios are quite sensitive to the calibration method used, even within the same model class.

On the other hand, comparing the local intensity functions implied by different credit portfolio models reveals that apparently different models, such as static Student-t copula models and reduced-form affine jump-diffusion models, lead to similar marginal loss distributions and tranche spreads. Thus, market prices alone are insufficient to discriminate between these model classes.

These results emphasize the importance of model uncertainty when addressing the pricing and hedging of portfolio credit derivatives and call for more research in this direction.

## Appendix A. Proofs.

A.1. Proof of Property 1. Properties 1(a)-(c) are immediate results from the definition. Since the payoff function $\left(K-L_{T}\right)^{+}$is convex in $K$ and taking expectation of the payoff function preserves convexity, Property 1(d) holds. Since $\left(L_{t}\right)$ is an increasing process, we know that for $T_{1} \leq T_{2}, K_{1} \leq K_{2}$

$$
\left(K_{1}-L_{T_{2}}\right)^{+}-\left(K_{1}-L_{T_{1}}\right)^{+} \geq\left(K_{2}-L_{T_{2}}\right)^{+}-\left(K_{2}-L_{T_{1}}\right)^{+} .
$$

Taking the expectation on both sides, we obtain Property 1(e).
For Property $1(\mathrm{f})$, consider $K \in[(i-1) \delta, i \delta]$ for any $i \in\{1, \ldots, n\}$ and $T \in\left[0, T^{*}\right]$. From the definition of $P(T, K)$, we have

$$
\begin{equation*}
P(T, K)=E^{\mathbb{Q}}\left[\left(K-L_{T}\right)^{+} \mid \mathcal{F}_{0}\right]=\sum_{k=0}^{i-1}(K-k \delta) \mathbb{Q}\left(L_{T}=k \delta \mid \mathcal{F}_{0}\right), \tag{16}
\end{equation*}
$$

which shows immediately that $K \mapsto P(T, K)$ is linear on $[(i-1) \delta, i \delta]$ and that $K \mapsto P(T, K)$ is a continuous function.
A.2. Proof of Theorem 1. Since $E^{\mathbb{Q}}\left[\lambda_{T} \mid \mathcal{F}_{0}\right]<\infty$ for all $T \in\left(0, T^{*}\right]$, the local intensity function defined by (2) satisfies $a(T, i)<\infty$ for all $T$ and $i$. Therefore the forward equations (3) hold and have the solution $P(T, i \delta)=E^{\mathbb{Q}}\left[\left(i \delta-L_{T}\right)^{+} \mid \mathcal{F}_{0}\right]$. By rewriting (3) in matrix form for $T \in\left(0, T^{*}\right]$, we have

$$
\begin{equation*}
\mathbf{P}^{\prime}=-\mathbf{M a}, \tag{17}
\end{equation*}
$$

where $\mathbf{P}^{\prime}=\left[\partial_{T} P(T, \delta), \ldots, \partial_{T} P(T, n \delta)\right]^{\mathrm{T}}, \mathbf{a}=[a(T, 0), \ldots, a(T, n-1)]^{\mathrm{T}}$, and $\mathbf{M}$ is an $n$-by- $n$ lower triangular matrix with entries

$$
\begin{aligned}
& \mathbf{M}(i, 1)=P(T, \delta), \\
& \mathbf{M}(i, j)=\nabla_{K}^{2} P(T,(j-2) \delta), \quad j=2, \ldots, i,
\end{aligned}
$$

for $i=1, \ldots, n$ and $\mathbf{M}(i, j)=0$ otherwise. From (4), we know that

$$
\begin{aligned}
& \mathbf{M}(i, 1)=\delta \mathbb{Q}\left(N_{T}=0 \mid \mathcal{F}_{0}\right), \\
& \mathbf{M}(i, j)=\delta \mathbb{Q}\left(N_{T}=j-1 \mid \mathcal{F}_{0}\right), \quad j=2, \ldots, i,
\end{aligned}
$$

for $i=1, \ldots, n$. Let $\mathcal{K} \subseteq\{0, \ldots, n-1\}$ be the set of integers $k$ such that $\mathbb{Q}\left(N_{T}=k \mid \mathcal{F}_{0}\right)=0$. We know that the $(k+1)$ st column of $\mathbf{M}$ is equal to zero for all $k \in \mathcal{K}$, and this implies that $\partial_{T} P(T, k \delta)=\partial_{T} P(T,(k-1) \delta)$. Moreover, in order to respect the convention given in the definition of the local intensity function (2), for all $k \in \mathcal{K}$, we have $a(T, k)=0$.

Now, let $\widetilde{\mathbf{M}}$ be the $(n-|\mathcal{K}|) \times(n-|\mathcal{K}|)$ matrix resulting from the elimination of the $(k+1)$ st row and column in $\mathbf{M}$ for all $k \in \mathcal{K} . \widetilde{\mathbf{M}}$ is a lower triangular matrix with $\widetilde{\mathbf{M}}(i, j)>0$ for all $i$ and $j \leq i$, and therefore it is invertible. Since, for all $k \in \mathcal{K}$, we have set $a(T, k)=0$ and noticed that $\partial_{T} P(T, k \delta)=\partial_{T} P(T,(k-1) \delta)$, we can rewrite the linear system (17) as

$$
\begin{equation*}
\widetilde{\mathbf{P}^{\prime}}=-\widetilde{\mathbf{M}} \widetilde{\mathbf{a}}, \tag{18}
\end{equation*}
$$

where $\widetilde{\mathbf{P}^{\prime}}$ and $\widetilde{\mathbf{a}}$ are $(n-|\mathcal{K}|)$-vectors which result from the elimination of the $(k+1)$ st entries of $\mathbf{P}^{\prime}$ and a, respectively, for all $k \in \mathcal{K}$. Multiplying both sides of (18) by $\widetilde{\mathbf{M}}^{-1}$, we have

$$
\begin{equation*}
\widetilde{\mathbf{a}}=-\widetilde{\mathbf{M}}^{-1} \widetilde{\mathbf{P}^{\prime}} \tag{19}
\end{equation*}
$$

Therefore, the local intensity function $a(.,$.$) is uniquely determined by the set of expected$ tranche notionals via (19). By the Gaussian elimination method, it is easy to check that (19) is equivalent to expression (5). Therefore, the local intensity function must have the form given in (5).
A.3. Proof of Proposition 1. Assume there exists a vector $\mathbf{p} \in \mathbb{R}^{n(m+1)}$ that satisfies (10)-(11). Using the fact that $\operatorname{rank}(\mathbf{A}) \leq I<n(m+1)$, we know from the rank-nullity theorem that there exists a nonzero $\mathbf{y} \in \mathbb{R}^{n(m+1)}$ such that

$$
\mathrm{Ay}=\mathbf{0} .
$$

Then, we choose the smallest $\mu_{0}<0$ and largest $\mu_{1}>0$ such that

$$
\mathbf{B}\left(\mathbf{p}+\mu_{0} \mathbf{y}\right) \leq \mathbf{0}, \quad \mathbf{B}\left(\mathbf{p}+\mu_{1} \mathbf{y}\right) \leq \mathbf{0} .
$$

Then, for any $\mu \in\left(\mu_{0}, \mu_{1}\right)$, we have

$$
\mathbf{B}(\mathbf{p}+\mu \mathbf{y})<\mathbf{0} .
$$

Therefore, the vector $\mathbf{p}+\mu \mathbf{y}$ also satisfies constraints (10)-(11) for all $\mu \in\left(\mu_{0}, \mu_{1}\right)$; i.e., there are infinitely many solutions for the system (10)-(11).
A.4. Existence of arbitrage in the base-correlation model. To show that the total default payments received from our trading strategy in section 2.5 are positive, we consider different scenarios at the last payment time, or equivalently the expiration time, $t_{m}$.

- For $L_{t_{m-1}} \leq 5 \%$, we would have received only a default payment at the expiration time $t_{m}$. Then, we have

$$
\begin{aligned}
& \text { value of default leg at expiration } \\
&=\left(5 \%-L_{t_{m}}\right)^{+}-2\left(6 \%-L_{t_{m}}\right)^{+}+\left(7 \%-L_{t_{m}}\right)^{+} \\
&-\left(5 \%-L_{t_{m-1}}\right)^{+}+2\left(6 \%-L_{t_{m-1}}\right)^{+}-\left(7 \%-L_{t_{m-1}}\right)^{+} \\
&= \begin{cases}0 & \text { if } L_{t_{m}} \leq 5 \% \\
L_{t_{m}}-5 \% & \text { if } 5 \%<L_{t_{m}} \leq 6 \%, \\
7 \%-L_{t_{m}} & \text { if } 6 \%<L_{t_{m}} \leq 7 \%, \\
0 & \text { if } 7 \%<L_{t_{m}}\end{cases} \\
& \geq 0
\end{aligned}
$$

- For $5 \%<L_{t_{m-1}} \leq 6 \%$, we would have received a certain number of default payments before time $t_{m}$ from tranche $[5 \%, 6 \%]$. Taking the reinvestment of those received payments at a risk-free rate into account, we know that the total value of the received payments at time $t_{m}$ is greater than $L_{t_{m-1}}-5 \%$. Then, we have
value of default leg at expiration

$$
\begin{aligned}
& \geq\left(L_{t_{m-1}}-5 \%\right) \\
&-2\left(6 \%-L_{t_{m}}\right)^{+}+\left(7 \%-L_{t_{m}}\right)^{+} \\
&+2\left(6 \%-L_{t_{m-1}}\right)^{+}-\left(7 \%-L_{t_{m-1}}\right)^{+} \\
&= \begin{cases}L_{t_{m}}-5 \% & \text { if } L_{t_{m}} \leq 6 \% \\
7 \%-L_{t_{m}} & \text { if } 6 \%<L_{t_{m}} \leq 7 \%, \\
0 & \text { if } 7 \%<L_{t_{m}}\end{cases} \\
& \geq 0
\end{aligned}
$$

- For $6 \%<L_{t_{m-1}} \leq 7 \%$, we would have
- received default payments from tranche $[5 \%, 6 \%]$ before time $t_{m}$, with value greater than $1 \%$ at time $t_{m}$,
- paid default payments for tranche $[6 \%, 7 \%]$ before time $t_{m}$, with value greater than $L_{t_{m-1}}-6 \%$ at time $t_{m}$.
Since we would have received all default payments from tranche [ $5 \%, 6 \%$ ] before paying any default payments for tranche $[6 \%, 7 \%]$, the difference in timing of the reinvestment tells us that the total net value of default payments received before time $t_{m}$ is greater than $7 \%-L_{t_{m-1}}$. Then, we have
value of default leg at expiration

$$
\begin{aligned}
& \geq\left(7 \%-L_{t_{m-1}}\right)+\left(7 \%-L_{t_{m}}\right)^{+}-\left(7 \%-L_{t_{m-1}}\right)^{+} \\
& =\left(7 \%-L_{t_{m}}\right)^{+} \geq 0
\end{aligned}
$$

- For $7 \%<L_{t_{m-1}}$, all possible default payments would have been made before time $t_{m}$, and we would have
- received default payments from tranche [5\%,6\%], with value greater than $1 \%$ at time $t_{m}$,
- paid default payments for tranche [6\%,7\%], with value greater than $1 \%$ at time $t_{m}$. Again, since we would have received all default payments from tranche $[5 \%, 6 \%]$ before paying any default payments for tranche [ $6 \%, 7 \%$ ], the total net value of default payments received before time $t_{m}$ is greater than 0 .
Therefore, the value of the default leg at expiration is positive.
Appendix B. Matrix representation of constraints. Recall that we write the expected tranche notionals in vector form:

$$
\mathbf{p}=\left[P\left(t_{0}, \delta\right), \ldots, P\left(t_{0}, n \delta\right), \ldots, P\left(t_{m}, \delta\right), \ldots, P\left(t_{m}, n \delta\right)\right]^{\mathrm{T}} \in \mathbb{R}^{n(m+1)}
$$

Since $t_{0}=0$ and $N_{0}=0$, we must have $P\left(t_{0}, i \delta\right)=i \delta$ for all $i$.
B.1. Pricing constraints. Assume that there are $I$ CDO tranches $\left[\kappa_{i-1}, \kappa_{i}\right], i=1, \ldots, I$, written on the reference portfolio with $\kappa_{0}=0$ and $\kappa_{I}=1$. Denoting $0=t_{0}<t_{1}<\cdots<t_{m}$ the payment dates of the CDO tranches, we can rewrite (9) for each tranche and obtain the following linear pricing constraints:

- For each of the mezzanine tranches $\left[\kappa_{i-1}, \kappa_{i}\right], i=2, \ldots, I-1$, without upfront payment, we have

$$
\begin{aligned}
0= & \sum_{j=1}^{m} D\left(0, t_{j}\right)\left[\left(1+s^{\left[\kappa_{i-1}, \kappa_{i}\right]} \Delta t_{j}\right) P\left(t_{j}, \kappa_{i}\right)-\left(1+s^{\left[\kappa_{i-1}, \kappa_{i}\right]} \Delta t_{j}\right) P\left(t_{j}, \kappa_{i-1}\right)\right. \\
= & \left.\quad-P\left(t_{j-1}, \kappa_{i}\right)+P\left(t_{j-1}, \kappa_{i-1}\right)\right] \\
& -\sum_{j=1}^{m-1}\left[D\left(t_{0}, \kappa_{i-1}\right)-D\left(0, t_{1}\right) P\left(t_{0}, \kappa_{i}\right)\right. \\
& \left.+\sum_{j=1}^{m-1}\left[1+s^{\left[\kappa_{i-1}, \kappa_{i}\right]} \Delta t_{j}\right)-D\left(0, t_{j+1}\right)\right] P\left(t_{j}, \kappa_{i-1}\right) \\
& \left.-D\left(0, t_{j}\right)\left(1+s^{\left[\kappa_{i-1}, \kappa_{i}\right]} \Delta t_{j}\right)-D\left(0, t_{j+1}\right)\right] P\left(t_{j}, \kappa_{i}\right) \\
& \left.-s^{\left[\kappa_{i-1}, \kappa_{i}\right]} \Delta t_{m}\right) P\left(t_{m}, \kappa_{i-1}\right)+D\left(0, t_{m}\right)\left(1+s^{\left[\kappa_{i-1}, \kappa_{i}\right]} \Delta t_{m}\right) P\left(t_{m}, \kappa_{i}\right)
\end{aligned}
$$

where $\Delta t_{j}=t_{j}-t_{j-1}$.

- For the most senior tranche $\left[\kappa_{I-1}, \kappa_{I}\right]$ with $\kappa_{I-1}<n \delta$, each default in the portfolio will reduce the notional value by $\frac{1}{n}-\delta$. In this case, the expected remaining notional value of tranche $\left[\kappa_{I-1}, \kappa_{I}\right]$ at payment time $t_{j}$ is equal to

$$
\begin{aligned}
& E^{\mathbb{Q}}\left[\left(1-L_{t_{j}}\right)-\left(\kappa_{I-1}-L_{t_{j}}\right)^{+}-\left(\frac{1}{n}-\delta\right) N_{t_{j}}\right] \\
= & E^{\mathbb{Q}}\left[\left(1-\frac{N_{t_{j}}}{n}\right)-\left(\kappa_{I-1}-L_{t_{j}}\right)^{+}\right] \\
= & \frac{1}{n \delta} P\left(t_{j}, n \delta\right)-P\left(t_{j}, \kappa_{I-1}\right) .
\end{aligned}
$$

Taking this into account, we have a slightly different equality for the most senior tranche:

$$
\begin{aligned}
0= & D\left(0, t_{1}\right) P\left(t_{0}, \kappa_{I-1}\right)-D\left(0, t_{1}\right) P\left(t_{0}, n \delta\right) \\
& -\sum_{j=1}^{m-1}\left[D\left(0, t_{j}\right)\left(1+s^{\left[\kappa_{I-1}, \kappa_{I}\right]} \Delta t_{j}\right)-D\left(0, t_{j+1}\right)\right] P\left(t_{j}, \kappa_{I-1}\right) \\
& +\sum_{j=1}^{m-1}\left[D\left(0, t_{j}\right)\left(1+\frac{s^{\left[\kappa_{I-1}, \kappa_{I}\right]} \Delta t_{j}}{n \delta}\right)-D\left(0, t_{j+1}\right)\right] P\left(t_{j}, n \delta\right) \\
& -D\left(0, t_{m}\right)\left(1+s^{\left[\kappa_{I-1}, \kappa_{I}\right]} \Delta t_{m}\right) P\left(t_{m}, \kappa_{I-1}\right)+D\left(0, t_{m}\right)\left(1+\frac{s^{\left[\kappa_{I-1}, \kappa_{I}\right]} \Delta t_{m}}{n \delta}\right) P\left(t_{m}, n \delta\right) .
\end{aligned}
$$

- For the equity tranche $\left[\kappa_{0}, \kappa_{1}\right]$ with an upfront payment $U^{\left[\kappa_{0}, \kappa_{1}\right]}$, we have

$$
\begin{aligned}
-U^{\left[\kappa_{0}, \kappa_{1}\right]} \kappa_{1}= & -D\left(0, t_{1}\right) P\left(t_{0}, \kappa_{1}\right) \\
& +\sum_{j=1}^{m-1}\left[D\left(0, t_{j}\right)\left(1+s^{\left[\kappa_{0}, \kappa_{1}\right]} \Delta t_{j}\right)-D\left(0, t_{j+1}\right)\right] P\left(t_{j}, \kappa_{1}\right) \\
& +D\left(0, t_{m}\right)\left(1+s^{\left[\kappa_{0}, \kappa_{1}\right]} \Delta t_{m}\right) P\left(t_{m}, \kappa_{1}\right) .
\end{aligned}
$$

One may notice that the attachment/detachment points $\left(\kappa_{i}\right)_{i=1, \ldots, I}$ of the CDO tranches are not necessary multiples of $\delta$; i.e., the existence of $k$ such that $\kappa_{i}=k \delta$ is not guaranteed. Nonetheless, Property 1 gives us the piecewise linearity of $P(T,$.$) on each (i \delta,(i+1) \delta)$, and therefore we can write

$$
\mathbf{p}_{\kappa}=\mathbf{A}_{1} \mathbf{p}
$$

where
$\mathbf{p}_{\kappa}=\left[P\left(t_{0}, \kappa_{1}\right), \ldots, P\left(t_{0}, \kappa_{I-1}\right), P\left(t_{0}, n \delta\right), \ldots, P\left(t_{m}, \kappa_{1}\right), \ldots, P\left(t_{m}, \kappa_{I-1}\right), P\left(t_{m}, n \delta\right)\right]^{\mathrm{T}} \in \mathbb{R}^{I(m+1)}$, and matrix $\mathbf{A}_{1}$ is a linear interpolation operator equal to

$$
\mathbf{A}_{1}=\left[\begin{array}{lll}
\tilde{\mathbf{A}}_{1} & & \\
& \ddots & \\
& & \tilde{\mathbf{A}}_{1}
\end{array}\right]_{I(m+1) \times n(m+1)}
$$

where the nonzero entries of $\tilde{\mathbf{A}}_{1} \in \mathbb{R}^{I \times n}$ are

$$
\begin{array}{ll}
\tilde{\mathbf{A}}_{1}\left(i,\left\lfloor\frac{\kappa_{i}}{\delta}\right\rfloor\right)=-\frac{\kappa_{i}}{\delta}+\left(\left\lfloor\frac{\kappa_{i}}{\delta}\right\rfloor+1\right), & i=1, \ldots, I-1, \\
\tilde{\mathbf{A}}_{1}\left(i,\left\lfloor\frac{\kappa_{i}}{\delta}\right\rfloor+1\right)=\frac{\kappa_{i}}{\delta}-\left\lfloor\frac{\kappa_{i}}{\delta}\right\rfloor, & i=1, \ldots, I-1, \\
\tilde{\mathbf{A}}_{1}(I, n)=1 . &
\end{array}
$$

Then, since each pricing constraint is linear in $\mathbf{p}_{\kappa}$ and the transformation from $\mathbf{p}$ to $\mathbf{p}_{\kappa}$ is linear as well, it is easy to see that

$$
\mathbf{A}_{2} \mathbf{A}_{1} \mathbf{p}=\mathbf{b}
$$

where

$$
\mathbf{b}=\left[\begin{array}{llll}
-U^{\left[\kappa_{0}, \kappa_{1}\right]} \kappa_{1} & 0 & \cdots & 0
\end{array}\right]^{\mathrm{T}} \in \mathbb{R}^{I}
$$

and the nonzero entries of $\mathbf{A}_{2} \in \mathbb{R}^{I \times I(m+1)}$ are

$$
\begin{aligned}
& \mathbf{A}_{2}(1,1)=-D\left(0, t_{1}\right) \\
& \mathbf{A}_{2}(1, I j+1)=D\left(0, t_{j}\right)\left(1+s^{\left[\kappa_{0}, \kappa_{1}\right]} \Delta t_{j}\right)-D\left(0, t_{j+1}\right), \quad j=1, \ldots, m-1 \\
& \mathbf{A}_{2}(1, I m+1)=D\left(0, t_{m}\right)\left(1+s^{\left[\kappa_{0}, \kappa_{1}\right]} \Delta t_{m}\right)
\end{aligned}
$$

for $i=2, \ldots, I-1$,

$$
\begin{aligned}
& \mathbf{A}_{2}(i, i-1)=D\left(0, t_{1}\right) \\
& \mathbf{A}_{2}(i, i)=-D\left(0, t_{1}\right) \\
& \mathbf{A}_{2}(i, I j+i-1)=-D\left(0, t_{j}\right)\left(1+s^{\left[\kappa_{i-1}, \kappa_{i}\right]} \Delta t_{j}\right)+D\left(0, t_{j+1}\right), \\
& \mathbf{A}_{2}(i, I j+i)=D\left(0, t_{j}\right)\left(1+s^{\left[\kappa_{i-1}, \kappa_{i}\right]} \Delta t_{j}\right)-D\left(0, t_{j+1}\right), \\
& \mathbf{A}_{2}(i, I m+i-1)=-D\left(0, t_{m}\right)\left(1+s^{\left[\kappa_{i-1}, \kappa_{i}\right]} \Delta t_{m}\right), \\
& \mathbf{A}_{2}(i, I m+i)=D\left(0, t_{m}\right)\left(1+s^{\left[\kappa_{i-1}, \kappa_{i}\right]} \Delta t_{m}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
& \mathbf{A}_{2}(I, I-1)=D\left(0, t_{1}\right) \\
& \mathbf{A}_{2}(I, I)=-D\left(0, t_{1}\right) \\
& \mathbf{A}_{2}(I, I j+I-1)=-D\left(0, t_{j}\right)\left(1+s^{\left[\kappa_{I-1}, \kappa_{I}\right]} \Delta t_{j}\right)+D\left(0, t_{j+1}\right), \quad j=1, \ldots, m-1, \\
& \mathbf{A}_{2}(I, I j+I)=D\left(0, t_{j}\right)\left(1+\frac{s^{\left[\kappa_{I-1}, \kappa_{I}\right]} \Delta t_{j}}{n \delta}\right)-D\left(0, t_{j+1}\right), \\
& \mathbf{A}_{2}(I, I(m+1)-1)=-D\left(0, t_{m}\right)\left(1+s^{\left[\kappa_{I-1}, \kappa_{I}\right]} \Delta t_{m}\right) \\
& \mathbf{A}_{2}(I, I(m+1))=D\left(0, t_{m}\right)\left(1+\frac{s^{\left[\kappa_{I-1}, \kappa_{I}\right]} \Delta t_{m}}{n \delta}\right)
\end{aligned}
$$

B.2. Relations in Condition 1. The strict positivity constraints in Condition 1(a) can be written as

$$
\mathbf{B}_{0} \mathbf{p}<\mathbf{0}
$$

where

And the strict convexity constraints in Condition 1(b) can be written as
$\mathbf{B}_{1} \mathbf{p}<\mathbf{0}$,
where

$$
\begin{gathered}
\mathbf{B}_{1}=\left[\begin{array}{ccc|cccc}
0 & \cdots & 0 & \tilde{\mathbf{B}}_{1} & & & \\
\vdots & & \vdots & & \ddots & \\
0 & \cdots & 0 & & & \tilde{\mathbf{B}}_{1}
\end{array}\right]_{(n-1) m \times n(m+1)} \\
\tilde{\mathbf{B}}_{1}=\left[\begin{array}{ccccccc}
2 & -1 & 0 & \cdots & & & \\
-1 & 2 & -1 & 0 & \cdots & & \\
& & \ddots & \ddots & \ddots & \ddots & \\
& & & 0 & -1 & 2 & -1
\end{array}\right]_{(n-1) \times n}
\end{gathered}
$$

Finally, Condition 1(c) can be written as

$$
\mathbf{B}_{2} \mathbf{B}_{3} \mathbf{p}<\mathbf{0}
$$

where
and
$\mathbf{B}_{3}=\left[\begin{array}{cccc|cccc|c|cccc|cccc}-1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & & 0 & 0 & & \cdots & 0 \\ 0 & -1 & \ddots & 0 & 0 & 1 & \ddots & 0 & \cdots & & \cdots & & & \cdots & & & \\ & & \ddots & & & & \ddots & & \cdots & \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ & & \vdots & & & \vdots & & & \cdots & \vdots & & -1 & 0 & \vdots & & 1 & 0 \\ 0 & & \cdots & 0 & 0 & \cdots & & 0 & \cdots & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 & 1\end{array}\right]_{n m \times n(m+1)}$
In summary, we can represent all strict inequality constraints conditions in matrix form:

$$
\mathbf{B p}<\mathbf{0},
$$

where

$$
\mathbf{B}=\left[\begin{array}{c}
\mathbf{B}_{0} \\
\mathbf{B}_{1} \\
\mathbf{B}_{2} \mathbf{B}_{3}
\end{array}\right]_{2 n m \times n(m+1)}
$$

Appendix C. Laplace transform of cumulative portfolio default intensity for shot-noise model. Recall that the portfolio default intensity for the shot-noise model is equal to

$$
\lambda_{t}=\eta_{t}+J_{t}
$$

where $\left(\eta_{t}\right)$ and $\left(J_{t}\right)$ are independent. Therefore,

$$
S(\theta, T)=E^{\mathbb{Q}}\left[e^{\theta \int_{0}^{T} \lambda_{t} d t} \mid \mathcal{F}_{0}\right]=E^{\mathbb{Q}}\left[e^{\theta \int_{0}^{T} \eta_{t} d t} \mid \mathcal{F}_{0}\right] E^{\mathbb{Q}}\left[e^{\theta \int_{0}^{T} J_{t} d t} \mid \mathcal{F}_{0}\right] .
$$

Since $\left(\eta_{t}\right)$ is an affine process, we have

$$
E^{\mathbb{Q}}\left[e^{\theta \int_{0}^{T} \eta_{t} d t} \mid \mathcal{F}_{0}\right]=e^{A(\theta, T)+B(\theta, T) \eta_{0}}
$$

where

$$
A(\theta, T)=-\frac{2 \kappa b}{\sigma^{2}} \ln \left(\frac{c+d e^{-\gamma T}}{c+d}\right)+\frac{\kappa b T}{c}, \quad B(\theta, T)=\frac{1-e^{-\gamma t}}{c+d e^{-\gamma T}}
$$

with

$$
\gamma=\sqrt{\kappa^{2}-2 \sigma^{2} \theta}, \quad c=(\kappa+\gamma) / 2 \theta, \quad d=(-\kappa+\gamma) / 2 \theta .
$$

Since $\left(J_{t}\right)$ is an Ornstein-Uhlenbeck process, we know that

$$
E^{\mathbb{Q}}\left[e^{\theta \int_{0}^{T} J_{t} d t} \mid \mathcal{F}_{0}\right]=e^{C(\theta, T)+D(\theta, T) J_{0}}
$$

where

$$
C(\theta, T)=l\left(\int_{0}^{T} \psi\left(\frac{\theta}{\alpha}\left(1-e^{\alpha(s-T)}\right)\right) d s-T\right), \quad D(\theta, T)=\frac{\theta}{\alpha}\left(1-e^{-\alpha T}\right)
$$

with $\psi(u):=E^{\mathbb{Q}}\left[e^{u Y_{1}}\right]$ being the Laplace transform of $Y_{1}$ and $l$ being the intensity of the underlying Poisson process. If $Y_{1}$ is exponentially distributed with mean $\mu$, then

$$
C(\theta, T)=\frac{l \mu}{1-\theta \mu}\left[\theta T-\frac{1}{\mu} \ln \left(1-\theta \mu\left(1-e^{-\alpha T}\right)\right)\right] .
$$

## REFERENCES

[1] M. Arnsdorf and I. Halperin, BSLP: Markovian bivariate spread-loss model for portfolio credit derivatives, J. Comput. Finance, 12 (2008), pp. 77-107.
[2] P. T. Boggs, P. D. Domich, and J. E. Rogers, An interior point method for general large-scale quadratic programming problems, Ann. Oper. Res., 62 (1996), pp. 419-437.
[3] P. Brémaud, Point Processes and Queues, Springer-Verlag, New York, 1981.
[4] R. Cont, Model uncertainty and its impact on the pricing of derivative instruments, Math. Finance, 16 (2006), pp. 519-547.
[5] R. Cont and Y. H. Kan, Dynamic Hedging of Portfolio Credit Derivatives, Financial Engineering Report 2008-08, Columbia University, 2008, SIAM J. Financial Math., submitted.
[6] R. Cont and A. Minca, Recovering Portfolio Default Intensities Implied by CDO Tranches, Financial Engineering Report 2008-01, Columbia University, 2008.
[7] R. Cont and I. Savescu, Forward equations for portfolio credit derivatives, in Frontiers in Quantitative Finance: Volatility and Credit Risk Modeling, R. Cont, ed., Wiley, New York, 2008, pp. 269-288.
[8] R. Cont and P. Tankov, Non-parametric calibration of jump-diffusion option pricing models, J. Comput. Finance, 7 (2004), pp. 1-49.
[9] D. Duffie and N. Gârleanu, Risk and valuation of collateralized debt obligations, Financial Anal. J., 57 (2001), pp. 41-59.
[10] B. Dupire, Pricing with a smile, Risk, 7 (1994), pp. 18-20.
[11] A. Eckner, Computational techniques for basic affine models of portfolio credit risk, J. Comput. Finance, 13 (2009), pp. 63-97.
[12] R. Gaspar and T. Schmidt, CDOs in the light of the current crisis, in Financial Risks: New Developments in Structured Products and Credit Derivatives, M. Jeanblanc and C. Gourieroux, eds., Economica, Paris, France, 2009, pp. 33-48.
[13] A. Herbertsson, Pricing synthetic CDO tranches in a model with default contagion: The matrix-analytic approach, J. Credit Risk, 4 (2008), pp. 3-35.
[14] J.-P. Laurent and A. Cousin, An overview of factor modeling for CDO pricing, in Frontiers in Quantitative Finance: Volatility and Credit Risk Modeling, R. Cont, ed., Wiley, New York, 2008, pp. 185-212.
[15] J.-P. Laurent, A. Cousin, and J.-D. Fermanian, Hedging Default Risks of CDOs in Markovian Contagion Models, Working paper, 2008.
[16] D. Li, Base correlation, in Encyclopedia of Quantitative Finance, R. Cont, ed., Wiley, New York, 2010.
[17] D. Li, On default correlation: A copula function approach, J. Fixed Income, 9 (2000), pp. 43-54.
[18] A. V. Lopatin and T. Misirpashaev, Two-dimensional Markovian model for dynamics of aggregate credit loss, in Econometrics and Risk Management, Adv. Econom. 22, Emerald/JAI, Bingley, UK, 2008, pp. 243-274.
[19] A. Mortensen, Semi-analytical valuation of basket credit derivatives in intensity-based models, J. Derivatives, 13 (2006), pp. 8-26.
[20] P. Schönbucher, Credit Derivatives Pricing Models, Wiley, New York, 2003.
[21] P. Schönbucher, Portfolio Losses and the Term Structure of Loss Transition Rates: A New Methodology for the Pricing of Portfolio Credit Derivatives, Working paper, 2005.
[22] R. J. Vanderbei, LOQO: An interior point code for quadratic programming, Optim. Methods Softw., 11/12 (1999), pp. 451-484.

# Path-Dependence of Leveraged ETF Returns* 

Marco Avellaneda ${ }^{\dagger}$ and Stanley Zhang $^{\dagger}$


#### Abstract

It is well known that leveraged exchange-traded funds (LETFs) do not reproduce the corresponding multiple of index returns over extended (quarterly or annual) investment horizons. For instance, in 2008 and early 2009, most LETFs underperformed the corresponding static strategies. In this paper, we study this phenomenon in detail. We give an exact formula linking the return of a leveraged fund with the corresponding multiple of the return of the unleveraged fund and its realized variance. This formula is tested empirically over quarterly horizons for 56 leveraged funds (44 double-leveraged and 12 triple-leveraged) using daily prices since January 2008 or since inception, according to the fund considered. The results indicate excellent agreement between the formula and the empirical data. The study also shows that leveraged funds can be used to replicate the returns of the underlying index, provided we use a dynamic rebalancing strategy. Empirically, we find that rebalancing frequencies required to achieve this goal are moderate - on the order of one week between rebalancings. Nevertheless, this need for dynamic rebalancing leads to the conclusion that LETFs as currently designed may be unsuitable for buy-and-hold investors.


Key words. ETFs, leveraged ETFs, volatility
AMS subject classifications. $62-07,62 \mathrm{P}-20$
DOI. 10.1137/090760805

1. Introduction. Leveraged exchange-traded funds (LETFs) are relative newcomers to the world of exchange-traded funds (ETFs). ${ }^{1}$ An LETF tracks the value of an index, a basket of stocks, or another ETF, with the additional feature that it uses leverage. For instance, the ProShares Ultra Financial ETF (UYG) offers double exposure to the Dow Jones U.S. Financials index. To achieve this, the manager invests two dollars in a basket of stocks tracking the index per each dollar of UYG's net asset value, borrowing an additional dollar. This is an example of a long LETF. Short LETFs, such as the ProShares UltraShort Financial ETF (SKF), offer a negative multiple of the return of the underlying ETF. In this case, the manager sells short a basket of stocks tracking the Dow Jones U.S. Financials index (or equivalent securities) to achieve a short exposure in the index of two dollars per each dollar of NAV $(\beta=-2)$. In both cases, the fund's holdings are rebalanced daily. ${ }^{2}$
[^48]It has been empirically established that if we consider investments over extended periods of time (e.g, three months, one year, or more), there are significant discrepancies between LETF returns and the returns of the corresponding leveraged buy-and-hold portfolios composed of index ETFs and cash [7]. Since early 2008, the quarterly performance of LETFs over any period of 60 business days has been inferior to the performance of the corresponding static leveraged portfolios for many leveraged/unleveraged pairs tracking the same index. There are a few periods where LETFs actually outperform, so this is not just a one-sided effect.

For example, a portfolio consisting of two dollars invested in I-Shares Dow Jones U.S. Financial Sector ETF (IYF) and short one dollar can be compared with an investment of one dollar in UYG. Figures 1 and 2 compare the returns of UYG and a twice-leveraged buy-andhold strategy with IYF, considering all 60 -day periods (overlapping) since February 2, 2008. For convenience, we present returns in arithmetic and logarithmic scales. Figures 3 and 4 display the same data for SKF and IYF.


Figure 1. 60-day returns for $U Y G$ versus leveraged $60-$ day return of IYF. $(X=2 \operatorname{Ret} .(I Y F) ; Y=$ Ret. $(U Y G)$ ). We considered all 60-day periods (overlapping) between February 2, 2008 and March 3, 2009. The concentrated cloud of points near the 45 -degree line corresponds to 60 -day returns prior to to September 2008, when volatility was relatively low. The remaining points correspond to periods when IYF was much more volatile.

Observing Figures 1 and 3, we see that the returns of the LETFs have predominantly underperformed the static leveraged strategy. This is particularly the case in periods when returns are moderate and volatility is high. LETFs outperform the static leveraged strategy only when returns are large and volatility is small. Another observation is that the historical underperformance is more pronounced for the short LETF (SKF).

These charts can be partially explained by the mismatch between the quarterly investment horizon and the daily rebalancing frequency; yet there are several points that deserve attention.

First, we notice that, due to the daily rebalancing of LETFs, the geometric (continuously compounded) relation

$$
\log \operatorname{ret} .(\operatorname{LETF}) \approx \beta \log \operatorname{ret} .(\mathrm{ETF})
$$



Figure 2. Same as in Figure 1, but returns are logarithmic, i.e., $X=2 \ln \left(I Y F_{t} / I Y F_{t-60}\right), Y=$ $\ln \left(U Y G_{t} / U Y G_{t-60}\right)$.

$-1$

Figure 3. Overlapping 60-day returns of SKF compared with the leveraged returns of the underlying ETF, overlapping, between February 2, 2008 and March 3, 2009. ( $X=-2 \operatorname{Ret} .(I Y F) ; Y=\operatorname{Ret} .(S K F)$.)
is more appropriate than the arithmetic (simply compounded) relation

$$
\operatorname{ret} .(\mathrm{LETF}) \approx \beta \operatorname{ret} .(\mathrm{ETF}), \quad \beta= \pm 2
$$

This explains the apparent alignment of the data points once we pass to logarithmic returns in Figures 2 and 4.

Second, we notice that the data points do not fall on the 45-degree line; they lie for the most part below it. This effect is due to volatility. It can be explained by the fact that


Figure 4. Comparison of logarithmic returns of SKF with the corresponding log-returns of IYF. $X=$ $-2 \ln \left(I Y F_{t} / I Y F_{t-60}\right) ; Y=\ln \left(S K F_{t} / S K F_{t-60}\right)$.
the LETF manager must necessarily "buy high and sell low" in order to enforce the target leverage requirement. Therefore, frequent rebalancing will lead to underperformance for the LETF relative to a static leveraged portfolio. The underperformance will be larger in periods when volatility is high, because daily rebalancing in a more volatile environment leads to more round-trip transactions, all other things being equal.

Thus, we find that holders of LETFs have an inherent exposure to the realized volatility of the underlying index, or to the volatility of the LETF itself. This effect is quantified using a simple model in section 2. We derive an exact formula for the return of an LETF as a function of its expense ratio, the applicable rate of interest, and the return and realized variance of an unleveraged ETF tracking the same index (the "underlying ETF," for short). In particular, we show that the holder of an LETF has a negative exposure to the realized variance of the underlying ETF. Since the expense ratio and the funding costs can be determined in advance with reasonable accuracy, the main factor that affects LETF returns is the realized variance. In section 3, we empirically validate the formula on a set of 56 LETFs with double and triple leverage, using all the data since their inception. The empirical study suggests that the proposed formula is very accurate. ${ }^{3}$

In the last section, we show that it is possible to use LETFs to replicate a predefined multiple of the underlying ETF returns, provided that we use dynamic hedging strategies. Specifically, in order to achieve a specified multiple of the return of the underlying index or ETF using LETFs, we must adjust the portfolio holdings in the LETF dynamically, according

[^49]to the amount of variance realized up to the hedging time by the index, as well as the level of the index. We derive a formula for the dynamic hedge-ratio, which is closely related to the model for LETFs, and we validate it empirically on the historical data for 44 double-leveraged LETFs. This last point, dynamic hedging, provides an interesting connection between LETFs and options.

After completing this paper, we found out that analogous results were obtained independently in a note issued by Barclays Global Investors [4], which contains a formula similar to (10) without including finance and expense ratios. To the best of our knowledge, the empirical testing of the formula over a broad universe of existing LETFs and the application to dynamic hedging using LETFs are new.
2. Modeling LETF returns. We denote the spot price of the underlying ETF by $S_{t}$, the price of the LETF by $L_{t}$, and leverage ratio by $\beta$. For instance, a double-leveraged long ETF will correspond to $\beta=2$, whereas a double-leveraged short ETF corresponds to $\beta=-2$.
2.1. Discrete-time model. Assume a model where there are $N$ trading days, and denote by $R_{i}^{S}$ and $R_{i}^{L}, i=1,2, \ldots, N$, the one-day returns for the underlying ETF and the LETF, respectively. The LETF provides a pro forma daily exposure of $\beta$ dollars of the underlying security per dollar under management. ${ }^{4}$ Accordingly, there is a simple link between $R_{i}^{S}$ and $R_{i}^{L}$. If the LETF is bullish $(\beta>1)$, then

$$
\begin{equation*}
R_{i}^{L}=\beta R_{i}^{S}-\beta r \Delta t-f \Delta t+r \Delta t=\beta R_{i}^{S}-((\beta-1) r+f) \Delta t, \tag{1}
\end{equation*}
$$

where $r$ is the reference interest rate (for instance, 3 -month LIBOR), $f$ is the expense ratio of the LETF, and $\Delta t=1 / 252$ represents one trading day.

If the LETF is bearish $(\beta \leq-1)$, the same equation holds with a small modification, namely,

$$
\begin{equation*}
R_{i}^{L}=\beta R_{i}^{S}-\beta\left(r-\lambda_{t}\right) \Delta t-f \Delta t+r \Delta t=\beta R_{i}^{S}-\left((\beta-1) r+f-\beta \lambda_{i}\right) \Delta t, \tag{2}
\end{equation*}
$$

where $\lambda_{i} \Delta t$ represents the cost of borrowing the components of the underlying index or the underlying ETF on day $i$. By definition, this cost is the difference between the reference interest rate and the "short rate" applied to cash proceeds from short sales of the underlying ETF. If the ETF, or the stocks that it holds, are widely available for lending, the short rate will be approximately equal to the reference rate and the borrowing costs are negligible. ${ }^{5}$

Let $t$ be a period of time (in years) covering several days ( $t=N \Delta$ ). Compounding the returns of the LETF, we have

$$
\begin{equation*}
L_{t}=L_{0} \prod_{i=1}^{N}\left(1+R_{i}^{L}\right) \tag{3}
\end{equation*}
$$

[^50]Substituting the value of $R_{t}^{L}$ into (1) or (2) (according to the sign of $\beta$ ), we obtain a relation between the prices of the LETF and the underlying asset. In fact, we show in the appendix that, under mild assumptions, we have

$$
\begin{equation*}
\frac{L_{t}}{L_{0}} \approx\left(\frac{S_{t}}{S_{0}}\right)^{\beta} \exp \left\{\frac{\beta-\beta^{2}}{2} V_{t}+\beta H_{t}+((1-\beta) r-f) t\right\} \tag{4}
\end{equation*}
$$

where

$$
\begin{equation*}
V_{t}=\sum_{i=1}^{N}\left(R_{i}^{S}-\overline{R^{S}}\right)^{2} \text { with } \overline{R^{S}}=\frac{1}{N} \sum_{i=1}^{N} R_{i}^{S}, \tag{5}
\end{equation*}
$$

i.e., $V_{t}$ is the realized variance of the price over the time-interval of interest, and where

$$
\begin{equation*}
H_{t}=\sum_{i=1}^{N} \lambda_{i} \Delta t \tag{6}
\end{equation*}
$$

represents the accumulated cost of borrowing the underlying stocks or ETF. This cost is obtained by subtracting the average applicable short rate from the reference interest rate each day and accumulating this difference over the period of interest. In addition to these two factors, formula (4) also shows the dependence on the funding rate and the expense ratio of the underlying ETF. The symbol " $\approx$ " in (4) means that the difference is small in relation to the daily volatility of the ETF or LETF. In the following section, we exhibit an exact relation, under the idealized assumptions that the price of the underlying ETF follows an Ito process and that rebalancing is done continuously.
2.2. Continuous-time model. To clarify the sense in which (4) holds, it is convenient to derive a similar formula assuming that the underlying ETF price follows an Ito process. To wit, we assume that $S_{t}$ satisfies the stochastic differential equation

$$
\begin{equation*}
\frac{d S_{t}}{S_{t}}=\sigma_{t} d W_{t}+\mu_{t} d t \tag{7}
\end{equation*}
$$

where $W_{t}$ is a standard Wiener process and $\sigma_{t}, \mu_{t}$ are, respectively, the instantaneous price volatility and drift. The latter processes are assumed to be random and nonanticipative with respect to $W_{t} .{ }^{6}$

Mimicking (1) and (2), we observe that if the LETF is bullish, the model for the return of the leveraged fund is now

$$
\begin{equation*}
\frac{d L_{t}}{L_{t}}=\beta \frac{d S_{t}}{S_{t}}-((\beta-1) r+f) d t \tag{8}
\end{equation*}
$$

If the LETF is bearish, the corresponding equation is

$$
\begin{equation*}
\frac{d L_{t}}{L_{t}}=\beta \frac{d S_{t}}{S_{t}}-\left((\beta-1) r-\beta \lambda_{t}+f\right) d t, \tag{9}
\end{equation*}
$$

[^51]where $\lambda_{t}$ represents the cost of borrowing the underlying ETF or the stocks that make up the ETF. In the appendix, we show that the following formula holds:
\[

$$
\begin{equation*}
\frac{L_{t}}{L_{0}}=\left(\frac{S_{t}}{S_{0}}\right)^{\beta} \exp \left\{((1-\beta) r-f) t+\beta \int_{0}^{t} \lambda_{s} d s+\frac{\left(\beta-\beta^{2}\right)}{2} \int_{0}^{t} \sigma_{s}^{2} d s\right\} \tag{10}
\end{equation*}
$$

\]

where we assume implicitly that $\lambda_{t}=0$ if $\beta>0$.
Formulas (4) and (10) convey essentially the same information if we define

$$
V_{t}=\int_{0}^{t} \sigma_{s}^{2} d s \quad \text { and } \quad H_{t}=\int_{0}^{t} \lambda_{s} d s
$$

The only difference is that (4) is an approximation which is valid for $\Delta t \ll 1$, whereas (10) is exact if the ETF price follows an Ito process. These equations show that the relation between the values of an LETF and its underlying ETF depends on

- the funding rate,
- the expense ratio for the LETF,
- the cost of borrowing shares in the case of short LETFs,
- the convexity (power law) associated with the leverage ratio $\beta$, and
- the realized variance of the underlying ETF.

The first two items require no explanation. The third follows from the fact that the manager of a bearish LETF may incur additional financing costs to maintain short positions if components of the underlying ETF or the ETF itself are hard-to-borrow. The last two items are more interesting: (i) due to daily rebalancing of the beta of the LETF, we find that the prices of an LETF and a nonleveraged ETF are related by a power law with power $\beta$, and (ii) the realized variance of the underlying ETF plays a significant role in determining the LETF returns.

The dependence on the realized variance might seem surprising at first. It turns out that the holder of an LETF has negative exposure to the realized variance of the underlying asset. This holds regardless of the sign of $\beta$. For instance, if the investor holds a double-long LETF, the term corresponding to to the realized variance in formula (10) is

$$
-\frac{\left(2^{2}-2\right)}{2} \int_{0}^{t} \sigma_{s}^{2}=-\int_{0}^{t} \sigma_{s}^{2} .
$$

In the case of a doubly bearish fund, the corresponding term is

$$
-\frac{\left((-2)^{2}-(-2)\right)}{2} \int_{0}^{t} \sigma_{s}^{2}=-3 \int_{0}^{t} \sigma_{s}^{2} .
$$

We note, in particular, that the dependence on the realized variance is stronger in the case of the double-short LETF.
3. Empirical validation. To validate the formula in (10), we consider 56 LETFs which currently trade in the U.S. markets. Of these, we consider 44 LETFs issued by ProShares, consisting 22 Ultra Long and 22 UltraShort ETFs. ${ }^{7}$ Table 1 gives a list of the Proshares

[^52]LETFs, their tickers, together with the corresponding sectors and their ETFs. We consider the evolution of the 44 LETFs from January 2, 2008 to March 20, 2009, a period of 308 business days.

We also consider 12 triple-leveraged ETFs issued by Direxion Funds. ${ }^{8}$ Direxion's LETFs were issued later than the ProShares funds, in November 2008; they provide a shorter historical record to test our formula. Nevertheless, we include the 3X Direxion funds for the sake of completeness and also because they have triple leverage (Table 2).

Table 1
Double-leveraged ETFs considered in the study. ETFs and the corresponding ProShares Ultra Long and UltraShort LETFs.

| Underlying |  |  |  |
| :---: | :---: | :---: | :---: |
| ETF | Proshares Ultra |  |  |
| $(\beta=2)$ | Proshares UltraShort <br> $(\beta=-2)$ | Index/sector |  |
| QQQQ | QLD | QID | Nasdaq 100 |
| DIA | DDM | DXD | Dow 30 |
| SPY | SSO | SDS | S\&P500 Index |
| IJH | MVV | MZZ | S\&P MidCap 400 |
| IJR | SAA | SDD | S\&P Small Cap 600 |
| IWM | UWM | TWM | Russell 2000 |
| IWD | UVG | SJF | Russell 1000 |
| IWF | UKF | SFK | Russell 1000 Growth |
| IWS | UVU | SJL | Russell MidCap Value |
| IWP | UKW | SDK | Russell MidCap Growth |
| IWN | UVT | SJH | Russell 2000 Value |
| IWO | UKK | SKK | Russell 2000 Growth |
| IYM | UYM | SMN | Basic materials |
| IYK | UGE | SZK | Consumer goods |
| IYC | UCC | SCC | Consumer services |
| IYF | UYG | SKF | Financials |
| IYH | RXL | RXD | Health care |
| IYJ | UXI | SIJ | Industrials |
| IYE | DIG | DUG | Oil \& gas |
| IYR | URE | SRS | Real estate |
| IYW | ROM | REW | Technology |
| IDU | UPW | SDP | Utilities |

Table 2
Triple-leveraged ETFs considered in the study. ETFs and corresponding Direxion 3X LETFs.

| Underlying <br> ETF or index | Direxion 3X bull <br> $(\beta=3)$ | Direxion 3X bear <br> $(\beta=-3)$ | Index/sector |
| :---: | :---: | :---: | :---: |
| IWB | BGU | BGZ | Russell 1000 |
| IWM | TNA | TZA | Russell 2000 |
| RIFIN.X | FAS | FAZ | Russell 1000 Financial Serv. |
| RIENG.X | ERX | ERY | Russell 1000 Energy |
| EFA | DZK | DPK | MSCI EAFE Index |
| EEM | EDC | EDZ | MSCI Emerging Markets |

[^53]We define the tracking error

$$
\begin{equation*}
\epsilon(t)=\frac{L_{t}}{L_{0}}-\left(\frac{S_{t}}{S_{0}}\right)^{\beta} \exp \left\{\frac{\beta-\beta^{2}}{2} V_{t}+\beta H_{t}+((1-\beta) r-f) t\right\} \tag{11}
\end{equation*}
$$

where $V_{t}$ is the accumulated variance, $H_{t}$ is the accumulated borrowing costs (in excess of the reference interest rate), $r$ is the interest rate, and $f$ is the management fee for the underlying ETF. The instantaneous volatility is modeled as the standard deviation of the returns of the underlying ETF sampled over a period of 5 days preceding each trading date:

$$
\begin{equation*}
\hat{\sigma}_{s}^{2}=\frac{1}{5} \sum_{i=1}^{5}\left(R_{(s / \Delta t)-i}^{(S)}\right)^{2}-\left(\frac{1}{5} \sum_{i=1}^{5} R_{(t / \Delta t)-i}^{(S)}\right)^{2}, \quad 0 \leq s \leq t \tag{12}
\end{equation*}
$$

For the interest rates and expense ratio, we use the 3 -month LIBOR rate published by the Federal Reserve Bank (H. 15 Report) and the expense ratio for the Proshares LETFs published in the prospectus. In all cases, we set $\lambda_{t}=0$; i.e., we do not take into account stock-borrowing costs explicitly.

The empirical results for ProShares are summarized in Tables 3 and 4 .
In the case of long LETFs, we find that the average of the tracking error $\epsilon(t)$ over the simulation period is typically less than 100 basis points. The standard deviation is also on the order of 100 basis points, with a few slightly higher observations. This suggests that the formula (10), using the model for stochastic volatility in (12), gives a reliable model for the relation between the LETFs and the underlying ETFs across time.

In the case of short ETFs, we also assume that $\lambda_{t}=0$ but expect a slightly higher tracking error, particularly between July and November of 2008, when restrictions for short selling in U.S. stocks were put in place. We observe higher levels for the mean and the standard deviation of the tracking error and some significant departures from the exact formula during the period of October and November of 2008, especially in Financials, which we attribute to short selling constraints. The tracking errors for the Direxion triple-leveraged ETFs have higher standard deviations, which is not surprising given that they have higher leverage (Tables 5 and 6). We note, in particular, that the errors for FAS and FAZ are the largest, which is consistent with the fact that they track financial stocks.

The conclusion of the empirical analysis is that the formula (10) explains well the behavior of the price of LETFs and the deviations between LETF returns and the returns of the underlying ETFs.

## 4. Consequences for buy-and-hold investors.

4.1. Comparison with buy-and-hold: Break-even levels. Formula (10) suggests that an investor who is long an LETF has a "time-decay" associated with the realized variance of the underlying ETF. In other words, if the price of the underlying ETF does not change significantly over the investment horizon, but the realized volatility is large, the investor in the LETF will underperform the corresponding leveraged return on the underlying ETF. On the contrary, if the underlying ETF makes a sufficiently large move in either direction, the investor will outperform the underlying ETF.

Table 3
Double-leveraged Ultra Long ETFs. Average tracking error (11) and standard deviation obtained by applying formula (10) to the Proshares long LETFs from January 2, 2008 to March 20, 2009. Notice that that the average tracking error is for the most part below 100 bps and the standard deviation is comparable. In particular, the standard deviation is inferior to the daily volatility of these assets, which often exceeds 100 basis points as well. This suggests that formula (10) gives the correct relation between the NAV of the LETFs and their underlying ETFs.

| Underlying <br> ETF | Tracking error <br> (average, \%) | Standard deviation <br> $(\%)$ | Leveraged <br> ETF |
| :---: | :---: | :---: | :---: |
| QQQQ | 0.04 | 0.47 | QLD |
| DIA | 0.00 | 0.78 | DDM |
| SPY | -0.06 | 0.40 | SSO |
| IJH | -0.06 | 0.38 | MVV |
| IJR | 1.26 | 0.71 | SAA |
| IWM | 1.26 | 0.88 | UWM |
| IWD | 1.00 | 0.98 | UVG |
| IWF | 0.50 | 0.59 | UKF |
| IWS | -0.33 | 1.20 | UVU |
| IWP | -0.02 | 0.61 | UKW |
| IWN | 2.15 | 1.29 | UVT |
| IWO | 0.50 | 0.74 | UKK |
| IYM | 1.44 | 1.21 | UYM |
| IYK | 1.20 | 0.75 | UGE |
| IYC | 1.56 | 1.04 | UCC |
| IYF | -0.22 | 0.74 | UYG |
| IYH | 0.40 | 0.42 | RXL |
| IYJ | 1.05 | 0.74 | UXI |
| IYE | -0.73 | 1.71 | DIG |
| IYR | 1.64 | 1.86 | URE |
| IYW | 0.51 | 0.55 | ROM |
| IDU | 0.25 | 0.55 | UPW |

Consider an investor who buys one dollar of an LETF and simultaneously shorts $\beta$ dollars of the underlying ETF (where shorting a negative amount means buying). For simplicity, we assume that the interest rate, fees, and borrowing costs are zero.

If we use (10), the equity in the investor's account will be equal to

$$
\begin{equation*}
E(t)=\left(\frac{S_{t}}{S_{0}}\right)^{\beta} e^{-\frac{\left(\beta^{2}-\beta\right)}{2} V_{t}}-\beta \frac{S_{t}}{S_{0}}-(1-\beta), \tag{13}
\end{equation*}
$$

including the cash credit or debit from the initial transaction. To be concrete, we consider the case of a double-long and a double-short separately. Setting $Y=E(t)$ and $X=\frac{S_{t}}{S_{0}}$, we obtain

$$
\begin{align*}
& Y=e^{-V_{t}} X^{2}-2 X+1, \quad \beta=2 \\
& Y=e^{-3 V_{t}} \frac{1}{X^{2}}+2 X-3, \quad \beta=-2 . \tag{14}
\end{align*}
$$

In the case of the double-long ETFs, the equity behaves like a parabola in $X=S_{t} / S_{0}$ with a curvature tending to zero exponentially as a function of the realized variance. The investor

## Table 4

Double-leveraged Ultra Short ETFs. Same as in Table 3, for double-short LETFs. Notice that the tracking error is relatively small, but there are a few funds where the tracking error is superior to 200 basis points. We attribute these errors to the fact that May ETFs, particularly in the Financial and Energy sectors, and the stocks in their holdings were hard-to-borrow from July to November 2008.

| Underlying <br> ETF | Tracking error <br> (average, \%) | Standard deviation <br> $(\%)$ | Leveraged <br> ETF |
| :---: | :---: | :---: | :---: |
| QQQQ | 0.22 | 0.80 | QID |
| DIA | -2.01 | 3.24 | DXD |
| SPY | -1.40 | 2.66 | SDS |
| IJH | 0.69 | 0.64 | MZZ |
| IJR | -0.55 | 0.86 | SDD |
| IWM | 0.94 | 0.91 | TWM |
| IWD | 0.32 | 1.40 | SJF |
| IWF | -0.30 | 1.34 | SFK |
| IWS | -2.06 | 3.03 | SJL |
| IWP | 0.93 | 0.92 | SDK |
| IWN | -2.21 | 1.80 | SJH |
| IWO | -0.19 | 0.79 | SKK |
| IYM | 1.82 | 0.99 | SMN |
| IYK | -0.76 | 1.98 | SZK |
| IYC | 0.79 | 0.92 | SCC |
| IYF | 3.30 | 3.03 | SKF |
| IYH | 1.04 | 0.91 | RXD |
| IYJ | 0.32 | 0.74 | SIJ |
| IYE | 0.43 | 3.09 | DUG |
| IYR | 2.00 | 2.07 | SRS |
| IYW | 0.01 | 0.80 | REW |
| IDU | 1.75 | 1.06 | SDP |

Table 5
Triple-Leveraged bullish ETFs. Average tracking errors and standard deviations for triple-leveraged long ETFs analyzed here, since their inception in November 2008.

| Underlying <br> ETF/index | Tracking error <br> (average, \%) | Standard deviation <br> $(\%)$ | 3X bullish <br> LETF |
| :---: | :---: | :---: | :---: |
| IWB | 0.44 | 0.55 | BGU |
| IWM | 0.81 | 0.75 | TNA |
| RIFIN.X | 3.67 | 2.08 | FAS |
| RIENG.X | 2.57 | 0.70 | ERX |
| EFA | 1.26 | 2.32 | DZK |
| EEM | 1.41 | 1.21 | EDC |

is therefore long convexity (Gamma, in options parlance) and short variance; hence he incurs time-decay (Theta). In the case of double-shorts, the profile is also a convex curve, which has convexity concentrated mostly for $X \ll 1$, and a much faster time-decay. From (14), we find that the break-even levels of $X, V_{t}$ needed for achieving a positive return by time $t$ are as follows.

Table 6
Triple-Leveraged bearish ETFs. Average tracking errors and standard deviations for triple-leveraged short ETFs analyzed here, since their inception in November 2008. Notice again that the errors for financials and energy are slightly higher than the rest.

| Underlying <br> ETF/index | Tracking error <br> average, $\%)$ | Standard deviation <br> $(\%)$ | 3X bearish <br> LETF |
| :---: | :---: | :---: | :---: |
| IWB | -0.08 | 0.64 | BGZ |
| IWM | 0.65 | 0.76 | TZA |
| RIFIN.X | -1.63 | 4.04 | FAZ |
| RIENG.X | -1.41 | 1.01 | ERY |
| EFA | -1.54 | 1.86 | DPK |
| EEM | 0.49 | 1.43 | EDZ |

- Double-long LETF:

$$
\begin{aligned}
& X>X_{+}=e^{V_{t}}\left(1+\sqrt{1-e^{-V_{t}}}\right), \\
& X<X_{-}=e^{V_{t}}\left(1-\sqrt{1-e^{-V_{t}}}\right) .
\end{aligned}
$$

- Double-short LETF:
$X_{+}, X_{-}$are the positive roots of the cubic equation

$$
\begin{equation*}
2 X^{3}-3 X^{2}+e^{-3 V_{t}}=0 . \tag{15}
\end{equation*}
$$

A similar analysis can be made for triple-leveraged ETFs. The main observation is that, regardless of whether the LETFs are long or short, they underperform the static leveraged strategy unless the returns of the underlying ETFs overcome the above volatility-dependent break-even levels. These levels are further away from the initial level as the realized variance increases.
4.2. Targeting investment returns using dynamic strategies with LETFs. A strong motivation for using LETFs is to take advantage of leverage when gaining exposure to an index. ${ }^{9}$ In this section we observe that it is possible to achieve a predetermined target return for the underlying ETF (e.g., IYF) using a dynamic strategy with LETFs (UYG or SKF).

Suppose that an agent is long $n$ shares of the underlying ETF and short $\Delta$ shares of the corresponding $\beta$-leveraged ETF. The one-period return of this portfolio is

$$
\begin{align*}
d \Pi & =n d S-\Delta d L \\
& =n S \frac{d S}{S}-\Delta L \frac{d L}{L} \\
& =n S \frac{d S}{S}-\Delta L \beta \frac{d S}{S}+\text { (carry terms). } \tag{16}
\end{align*}
$$

[^54]It follows that the choice of $\Delta=\frac{n S}{L}$ should give a portfolio without market risk. Based on this observation, consider an investor with an initial endowment of $\Pi_{0}$ dollars and a dynamic strategy which invests

$$
\delta_{t}=\Pi_{0} \frac{1}{\beta} \frac{S_{t}}{S_{0}}
$$

dollars in the LETF with multiplier $\beta$. For simplicity, we assume $\lambda=f=0$. Let us denote the value of his position at time $t$ by $\Pi_{t}$. The change in value of this position across time, funded at the rate $r$, satisfies

$$
\begin{align*}
d \Pi_{t} & =r \Pi_{t} d t+\delta_{t} \frac{d L_{t}}{L_{t}}-r \delta_{t} d t  \tag{17}\\
& =r \Pi_{t} d t+\delta_{t}\left(\beta \frac{d S_{t}}{S_{t}}+(1-\beta) r d t\right)-r \delta_{t} d t \\
& =r \Pi_{t} d t+\Pi_{0} \frac{1}{\beta} \frac{S_{t}}{S_{0}}\left(\beta \frac{d S_{t}}{S_{t}}+(1-\beta) r d t\right)-r \delta_{t} d t \\
& =r \Pi_{t} d t+\Pi_{0} \frac{d S_{t}}{S_{0}}+\Pi_{0} \frac{1}{\beta} \frac{S_{t}}{S_{0}} r d t-\Pi_{0} \frac{S_{t}}{S_{0}} r d t-\Pi_{0} \frac{1}{\beta} \frac{S_{t}}{S_{0}} r d t \\
& =r \Pi_{t} d t+\Pi_{0} \frac{d S_{t}}{S_{0}}-\Pi_{0} \frac{S_{t}}{S_{0}} r d t . \tag{18}
\end{align*}
$$

Integrating this stochastic differential equation from $t=0$ to $t=T$, the reader can verify that (18) implies that

$$
\begin{equation*}
\frac{\Pi_{T}}{\Pi_{0}}=\frac{S_{T}}{S_{0}} \tag{19}
\end{equation*}
$$

for all $T$.
Accordingly, the agent can "replicate" the returns of the underlying stock over any maturity by dynamic hedging with LETFs, with essentially $\frac{1}{|\beta|}$ of the capital required to do the same trade in the underlying ETF.

However, in order to achieve his target return over an extended investment period using LETFs, the investor needs to rebalance his portfolio according to his Delta exposure. Because of this, dynamic replication with LETFs may not be suitable for many retail investors. On the other hand, this analysis will be useful to active traders, investment advisors, or issuers who manage LETFs with longer investment horizons, since we have shown that the latter can be "replicated" dynamically with LETFs which are rebalanced daily.

In Tables 7 and 8 we demonstrate the effectiveness of this dynamic replication method using different rebalancing techniques. We consider dynamic hedging in which we rebalance if the total Delta exceeds a band of $1 \%, 2 \%, 5 \%$, and $10 \%$, and also hedging with fixed timesteps of $1,2,5$, or 15 business days. Table 9 indicates the expected number of days between rebalancing for strategies that are price-dependent. The results indicate that rebalancing when the Delta exposure exceeds $5 \%$ of notional give reasonable tracking errors. The corresponding average intervals between rebalancings can be large, which means that, in practice, one can achieve reasonable tracking errors without necessarily rebalancing the Delta daily or even weekly.

Table 7
Average tracking error, in \% of notional, for the dynamic replication of 6-month returns using double-long LETFs with $m=\beta$. The first four columns correspond to rebalancing when the Delta reaches the edge of a band of $\pm x \%$ around zero. The last four columns correspond to rebalancing at fixed time intervals. The data used to generate this table corresponds, for each ETF, to all overlapping 6-month returns in the year 2008.

| ETF | $1 \%$ | $2 \%$ | $5 \%$ | $10 \%$ | 1 day | 2 day | 5 day | 15 day |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| QQQQ | -0.29 | -0.71 | -1.05 | -1.62 | -0.56 | -0.96 | -1.45 | -1.74 |
| DIA | -0.99 | -0.99 | -1.37 | -1.45 | -0.47 | -0.59 | -0.84 | -0.99 |
| SPY | -0.97 | -0.92 | -1.19 | -1.47 | -0.92 | -1.17 | -1.55 | -1.77 |
| IJH | -0.39 | -0.37 | -0.79 | -0.99 | -0.53 | -0.75 | -1.05 | -1.09 |
| IJR | -0.56 | -0.57 | -1.07 | -2.68 | -0.66 | -0.90 | -1.44 | -1.70 |
| IWM | 0.37 | 0.22 | -0.49 | -1.44 | 0.47 | 0.03 | -0.70 | -0.93 |
| IWD | -0.03 | -0.35 | -0.30 | -0.64 | 0.00 | -0.15 | -0.57 | -0.79 |
| IWF | -0.15 | -0.25 | -0.54 | -1.08 | -0.12 | -0.37 | -0.68 | -0.81 |
| IWS | 0.87 | 0.22 | 0.81 | 0.14 | 0.69 | 0.71 | 0.54 | 0.24 |
| IWP | -0.16 | -0.14 | -0.54 | -1.41 | -0.36 | -0.40 | -0.82 | -0.89 |
| IWN | 0.94 | 0.40 | 0.56 | -0.03 | -0.91 | 0.86 | 0.36 | 0.14 |
| IWO | 0.23 | 0.03 | -1.00 | -1.63 | -0.05 | -0.44 | -1.15 | -1.45 |
| IYM | -0.39 | -0.51 | -0.89 | -2.35 | -0.24 | -0.67 | -1.54 | -1.91 |
| IYK | 0.24 | 0.13 | -0.16 | -0.06 | 0.37 | 0.34 | 0.10 | 0.04 |
| IYC | 0.58 | 0.57 | -0.13 | -0.76 | 0.71 | 0.70 | 0.04 | -0.21 |
| IYF | -0.36 | -0.62 | 0.01 | -0.54 | -0.30 | -0.35 | -1.28 | -2.17 |
| IYH | 0.22 | -0.10 | -0.14 | 0.27 | 0.30 | 0.19 | 0.03 | 0.07 |
| IYJ | 0.12 | -0.09 | -0.36 | -0.92 | 0.14 | -0.04 | -0.30 | -0.61 |
| IYE | -1.44 | -2.02 | -1.90 | -1.76 | -1.19 | -1.82 | -2.21 | -2.07 |
| IYR | -0.43 | 0.58 | -0.80 | -0.95 | -0.61 | 0.55 | -0.74 | -1.48 |
| IYW | -0.50 | -0.46 | -1.67 | -1.39 | -0.37 | -0.85 | -1.41 | -1.76 |
| IDU | 0.75 | 0.45 | 0.73 | 0.11 | 0.83 | 0.78 | 0.46 | 0.52 |

5. Conclusion. This study presents a formula for the value of an LETF in terms of the value of the underlying index or ETF. The formula is validated empirically using end-ofday data on 56 LETFs, of which 44 are double-leveraged and 12 are triple leveraged. This formula validates the fact that on log-scale LETFs will underperform the nominal returns by a contribution that is primarily due to the realized volatility of the underlying ETF. The formula also takes into account financing costs and shows that for short ETFs, the cost of borrowing the underlying stock may play a role as well, as observed in [1].

We also demonstrate that LETFs can be used for hedging and replicating unleveraged ETFs, provided that traders engage in dynamic hedging. In this case, the hedge-ratio depends on the realized accumulated variance as well as on the level of the LETF at any point in time. The path-dependence of leveraged ETFs makes them interesting for the professional investor, since they are linked to the realized variance and the financing costs. However, they may not be suitable for buy-and-hold investors who aim at replicating a particular index taking advantage of the leveraged provided, for the reasons explained above.

Addendum, October 27, 2009. Very recently, regulators have issued notices concerning the suitability of LETFs for buy-and-hold investors which are consistent with our findings. These include a Regulatory Notice by the Financial Instrument Regulatory Authority [8] and an alert by the U.S. Securities and Exchange Commission (July 2009). Also, a major issuer of

Table 8
Standard deviation of tracking error (\%) for dynamic replication of 6-month returns using double-long LETFs.

| ETF | $1 \%$ | $2 \%$ | $5 \%$ | $10 \%$ | 1 day | 2 day | 5 day | 15 day |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| QQQQ | 0.75 | 0.77 | 0.80 | 1.20 | 0.75 | 0.76 | 0.84 | 0.93 |
| DIA | 0.35 | 0.37 | 0.41 | 0.36 | 0.36 | 0.39 | 0.47 | 0.43 |
| SPY | 0.27 | 0.32 | 0.39 | 0.69 | 0.27 | 0.30 | 0.39 | 0.45 |
| IJH | 0.48 | 0.49 | 0.56 | 1.19 | 0.47 | 0.48 | 0.62 | 0.65 |
| IJR | 1.19 | 1.21 | 1.33 | 1.39 | 1.20 | 1.29 | 1.22 | 1.17 |
| IWM | 0.66 | 0.67 | 0.71 | 1.60 | 0.67 | 0.75 | 0.71 | 0.74 |
| IWD | 1.38 | 1.38 | 1.40 | 1.52 | 1.38 | 1.43 | 1.41 | 1.49 |
| IWF | 0.93 | 0.94 | 1.07 | 1.21 | 0.95 | 0.99 | 0.94 | 0.99 |
| IWS | 2.05 | 2.05 | 2.08 | 2.29 | 2.05 | 2.01 | 2.07 | 2.09 |
| IWP | 0.83 | 0.82 | 0.93 | 1.24 | 0.83 | 0.91 | 0.84 | 0.91 |
| IWN | 1.71 | 1.70 | 1.76 | 2.09 | 1.72 | 1.70 | 1.80 | 1.82 |
| IWO | 0.80 | 0.80 | 0.91 | 1.32 | 0.79 | 0.99 | 0.84 | 1.00 |
| IYM | 1.05 | 1.07 | 1.15 | 1.24 | 1.07 | 1.20 | 1.29 | 1.59 |
| IYK | 0.57 | 0.57 | 0.63 | 0.67 | 0.56 | 0.63 | 0.61 | 0.63 |
| IYC | 0.80 | 0.78 | 0.83 | 0.95 | 0.80 | 0.98 | 0.91 | 1.03 |
| IYF | 1.12 | 1.18 | 1.12 | 1.88 | 1.10 | 1.21 | 2.01 | 1.49 |
| IYH | 0.56 | 0.55 | 0.58 | 0.73 | 0.56 | 0.55 | 0.57 | 0.62 |
| IYJ | 0.71 | 0.75 | 0.82 | 0.84 | 0.70 | 0.70 | 0.79 | 0.87 |
| IYE | 0.64 | 0.66 | 0.77 | 1.26 | 0.64 | 0.65 | 1.02 | 1.49 |
| IYR | 1.47 | 1.47 | 1.62 | 2.08 | 1.47 | 1.60 | 1.88 | 1.83 |
| IYW | 1.45 | 1.45 | 1.59 | 2.05 | 1.45 | 1.39 | 1.49 | 1.42 |
| IDU | 0.53 | 0.52 | 0.54 | 0.72 | 0.51 | 0.53 | 0.54 | 0.60 |

LETFs recently issued a paper addressing such concerns [5].

## Appendix.

A.1. Discrete-time model (equation (4)). Set $a_{i}=(\beta-1) r+f+\beta \lambda_{i}$, where $\lambda_{i}$ is the cost of borrowing the underlying asset on day $i\left(\lambda_{i}\right.$ is zero if $\beta>0$.) We assume that

$$
R_{i}^{S}=\xi_{i} \sqrt{\Delta t}+\mu \Delta t,
$$

where $\Delta t=1 / 252$ and $\xi_{i}, i=1,2, \ldots$, is a stationary process such that $\xi_{i}$ has mean equal to zero and finite absolute moments of order 3. This assumption is consistent with many models of equity returns. Notice that we do not assume that successive returns are uncorrelated.

From (1), (2), and (3), we find using Taylor expansion that

$$
\begin{align*}
\ln \left(\frac{L_{t}}{L_{0}}\right) & =\sum_{i} \ln \left(1+R_{i}^{L}\right) \\
& =\sum_{i} \ln \left(1+\beta R_{i}^{S}-a_{i} \Delta t\right) \\
& =\sum_{i}\left(\beta R_{i}^{S}-a_{i} \Delta t-\frac{\beta^{2}}{2}\left(R_{i}^{S}\right)^{2}\right)+\sum_{i}\left(O\left(\left|R_{i}^{S}\right|^{3}\right)+O\left(\left|R_{i}^{S}\right| \Delta t\right)\right) \tag{20}
\end{align*}
$$

## Table 9

Average number of business days between portfolio rebalancing for the 6-month dynamic hedging strategy: The effect of changing the Delta band. Each column shows the average number of days between rebalancing the portfolio, assuming different Delta-bandwidth for portfolio rebalancing. For instance, the column with heading of $1 \%$ corresponds to a strategy that rebalances the portfolio each time the net delta exposure exceeds $1 \%$ of the notional amount. The expected number of days between rebalancing increases as the bandwidth increases.

| ETF | $1 \%$ | $2 \%$ | $5 \%$ | $10 \%$ |
| :---: | :---: | :---: | :---: | :---: |
| QQQQ | 2.03 | 4.14 | 24 | 60 |
| DIA | 2.50 | 5.22 | 30 | 120 |
| SPY | 2.73 | 5.22 | 40 | NR |
| IJH | 2.26 | 4.62 | 24 | NR |
| IJR | 2.03 | 4.29 | 20 | NR |
| IWM | 1.85 | 4.62 | 30 | NR |
| IWD | 2.18 | 5.00 | 30 | 120 |
| IWF | 2.26 | 5.00 | 30 | NR |
| IWS | 2.26 | 6.67 | 17 | NR |
| IWP | 1.85 | 4.14 | 30 | NR |
| IWN | 2.26 | 4.62 | 24 | NR |
| IWO | 1.85 | 4.00 | 20 | 60 |
| IYM | 1.74 | 3.08 | 9 | 40 |
| IYK | 3.16 | 8.57 | 30 | 120 |
| IYC | 1.90 | 3.87 | 30 | NR |
| IYF | 1.45 | 2.93 | 9 | 30 |
| IYH | 2.79 | 10.91 | 30 | 120 |
| IYJ | 2.35 | 4.80 | 17 | NR |
| IYE | 1.79 | 3.43 | 12 | 40 |
| IYR | 1.62 | 3.16 | 17 | 30 |
| IYW | 2.00 | 3.53 | 40 | NR |
| IDU | 2.67 | 6.32 | 20 | 120 |

By the same token, we have

$$
\begin{align*}
\beta \ln \left(\frac{S_{t}}{S_{0}}\right) & =\beta \sum_{i} \ln \left(1+R_{i}^{S}\right) \\
& =\beta \sum_{i}\left(R_{i}^{S}-\frac{1}{2}\left(R_{i}^{S}\right)^{2}\right)+\sum_{t} O\left(\left(R_{i}^{S}\right)^{3}\right) \tag{21}
\end{align*}
$$

Subtracting (22) from (21), we find that

$$
\begin{align*}
\ln \left(\frac{L_{t}}{L_{0}}\right)-\beta \ln \left(\frac{S_{t}}{S_{0}}\right)= & -\sum_{i}\left(a_{i} \Delta t+\frac{\beta^{2}-\beta}{2}\left(R_{i}^{S}\right)^{2}\right)+\sum_{i}\left(O\left(\left|R_{i}^{S}\right|^{3}\right)+O\left(\left|R_{i}^{S}\right| \Delta t\right)\right) \\
= & -\sum_{i}\left(a_{i} \Delta t+\frac{\beta^{2}-\beta}{2}\left(\left(R_{i}^{S}\right)^{2}-\mu^{2}(\Delta t)^{2}\right)\right) \\
& +\sum_{i}\left(O\left(\left|R_{i}^{S}\right|^{3}\right)+O\left(\left|R_{i}^{S}\right| \Delta t\right)+O\left((\Delta t)^{2}\right)\right) \\
= & -\sum_{i} a_{i} \Delta t-\frac{\beta^{2}-\beta}{2} V_{t}+\sum_{i}\left(O\left(\left|R_{i}^{S}\right|^{3}\right)+O\left(\left|R_{i}^{S}\right| \Delta t\right)+O\left((\Delta t)^{2}\right)\right) . \tag{22}
\end{align*}
$$

We show that the remainder in this last equation is negligible. In fact, we have

$$
\begin{align*}
\sum_{i}\left|R_{i}^{S}\right|^{3} & =\sum_{i}\left|\xi_{i}\right|^{3}(\Delta t)^{3 / 2} \\
& =\left(\frac{\sum_{i}\left|\xi_{i}\right|^{3}}{N}\right) t \sqrt{\Delta t} \\
& \approx E\left(\left|\xi_{1}^{3}\right|\right) t \sqrt{\Delta t} \tag{23}
\end{align*}
$$

and, similarly,

$$
\begin{align*}
\sum_{i}\left|R_{i}^{S}\right| \Delta t & =\sum_{i}\left|\xi_{i}\right|(\Delta t)^{3 / 2} \\
& =\left(\frac{\sum_{i}\left|\xi_{i}\right|}{N}\right) t \sqrt{\Delta t} \\
& \approx E\left(\left|\xi_{1}\right|\right) t \sqrt{\Delta t} \tag{24}
\end{align*}
$$

Therefore, the contribution of the last sum in (23) is bounded by the first three moments of $\xi_{1}$, multiplied by the investment horizon, $t$, and by $\sqrt{\Delta t}$. This means that if we neglect the last terms, for investment horizons of less than 1 year, the error is on the order of the standard deviation of the daily returns of the underlying ETF, which is negligible.
A.2. Continuous-time model (equation (10)). The above reasoning is exact for Itô processes, because it corresponds to $\Delta t \rightarrow 0$. To be precise, we consider (7), (8), and (9) and apply Itô's lemma to obtain

$$
\begin{align*}
& d \ln S_{t}=\frac{d S_{t}}{S_{t}}-\frac{\sigma_{t}^{2}}{2} d t  \tag{25}\\
& d \ln L_{t}=\beta \frac{d S_{t}}{S_{t}}-\frac{\beta^{2} \sigma_{t}^{2}}{2} d t+\left((\beta-1) r+f+\beta \lambda_{i}\right) d t \quad\left(\lambda_{i}=0 \text { if } \beta>0\right) \tag{2}
\end{align*}
$$

Multiplying (26) by $\beta$ and subtracting it from (27), we obtain

$$
\begin{equation*}
d \ln L_{t}-\beta d \ln S_{t}=-\frac{\left(\beta^{2}-\beta\right) \sigma_{t}^{2}}{2} d t+\left((\beta-1) r+f+\beta \lambda_{i}\right) d t \tag{27}
\end{equation*}
$$

which implies (10).
Acknowledgment. We thank two anonymous referees for pointing out the parallels between LETFs and over-the-counter CPPI products, as well as for pointing out a simplified proof of the dynamic replication scheme of section 4.2.

## REFERENCES

[1] M. Avellaneda and M. D. Lipkin, A Dynamic Model for Hard-to-Borrow Stocks, RISK Magazine, June 2009. Also available at http://ssrn.com/abstract=1357069.
[2] P. Bertrand and J.-L. Prigent, Portfolio insurance strategies: A comparison of standard methods when the volatility of the stock is stochastic, Internat. J. Business, 8 (2003), pp. 461-472.
[3] F. Black and A. Perold, Theory of constant proportion portfolio insurance, J. Econom., Dynam. Control, 16 (1992), pp. 403-426.
[4] M. Cheng and A. Madhavan, The Dynamics of Leveraged and Inverse Exchange-Traded Funds, Barclays Global Investors; available at: http://www.barclaysglobal.com/secure/repository/publications/usa/ ResearchPapers/Leveraged_ETF.pdf (April 8, 2009).
[5] J. Hill and G. Forster, Understanding Returns of Leveraged and Inverse Funds, J. Indexes (2009), http://www.indexuniverse.com/publications/journalofindexes/.
[6] Leveraged and Inverse ETFs: Specialized Products with Extra Risks for Buy-and-Hold Investors, SEC Alert, http://www.sec.gov/investor/pubs/leveragedetfs-alert.htm (August 18, 2009).
[7] L. Lu, J. Wang, and G. Zhang, Long Term Performance of Leveraged ETFs, available at http://ssrn.com/abstract=1344133 (February 15, 2009).
[8] Non-traditional ETFs, FINRA Regulatory Notice, FINRA Reminds Firms of Sales Practice Obligations Relating to Leveraged and Inverse Exchange-Traded Funds, http://www.finra.org/web/groups/ industry/@ip/@reg/@notice/documents/notices/p118952.pdf (June, 2009).
[9] Non-traditional ETFs, FINRA Regulatory Notice, Increased Margin Requirements for Leveraged Exchange-Traded Funds and Associated Uncovered Options, http://www.finra.org/web/groups/ industry/@ip/@reg/@notice/documents/notices/p119906.pdf (August, 2009).

# Dual Valuation and Hedging of Bermudan Options* 

L. C. G. Rogers ${ }^{\dagger}$


#### Abstract

Some years ago, a different characterization of the value of a Bermudan option was discovered which can be thought of as the viewpoint of the seller of the option, in contrast to the conventional characterization which took the viewpoint of the buyer. Since then, there has been a lot of interest in finding numerical methods which exploit this dual characterization. This paper presents a pure dual algorithm for pricing and hedging Bermudan options.


Key words. optimal stopping, dual valuation, Bermudan option, hedging
AMS subject classifications. 62L15, 60G40, 91G40
DOI. 10.1137/090772198

1. Introduction. This paper derives an algorithm for valuing and hedging a Bermudan ${ }^{1}$ option from a "pure dual" standpoint. Apart from various changes of names, pricing a Bermudan option is the same as solving an optimal stopping problem, which is arguably the simplest possible stochastic optimal control problem. For as long as derivatives have been priced within the Black-Scholes paradigm, the traditional value-function approach, and the associated Bellman equations, have been widely used in the attempt to price Bermudan options; the whole area has been a mathematical playground because of the scarcity of closed-form solutions and the consequent need for approximations, estimates, and asymptotics to come up with prices.

During the last century, the value-function approach was, in effect, the only method available, but in recent years another quite different "dual" approach has been discovered; see Rogers [3] and Haugh and Kogan [2]. The main result is that if the reward process ${ }^{2}$ is denoted $Z$, then the value $Y_{0}^{*}$ of the optimal stopping problem can be alternatively expressed as

$$
\begin{equation*}
Y_{0}^{*}=\sup _{\tau \in \mathcal{T}} E\left[Z_{\tau}\right]=\min _{M \in \mathcal{M}_{0}} E\left[\sup _{0 \leq t \leq T}\left(Z_{t}-M_{t}\right)\right], \tag{1.1}
\end{equation*}
$$

where $\mathcal{T}$ is the set of stopping times bounded by $T$, the time horizon for the problem, and $\mathcal{M}_{0}$ is the set of uniformly integrable martingales vanishing at zero. The minimum is attained, by the martingale $M^{*}$ of the Doob-Meyer decomposition of the Snell envelope process of $Z$, and in that case

$$
\begin{equation*}
\sup _{0 \leq t \leq T}\left(Z_{t}-M_{t}^{*}\right)=Y_{0}^{*} \quad \text { almost surely. } \tag{1.2}
\end{equation*}
$$

[^55]The traditional approach via the value function and the Bellman equations takes the viewpoint of the buyer of the option, who seeks to choose the best stopping time at which to exercise; the first expression for $Y_{0}^{*}$ in (1.1) embodies this. The dual approach is the solution of the problem from the viewpoint of the seller of the option, who seeks a hedging martingale $Y_{0}^{*}+M_{t}$, whose value at all times will be at least the value of the reward process; and the result (1.2) shows that for the perfect choice of $M=M^{*}$ this does indeed happen.

Since the dual approach was discovered, there have been various attempts to apply it in practice, with mixed success; choosing a good martingale is at least as difficult as choosing a good stopping time! The early paper of Andersen and Broadie [1] uses a numerical approximation to the value function to suggest a good martingale to use and in this way obtains quite tight bounds on both sides for a number of test examples. However, at a conceptual level, it has been an outstanding issue to derive a "pure" dual method which solves the optimal stopping problem without the need to calculate a value function using only the dual characterization of (1.1). Let us amplify the distinction. Pure primal methods are well understood; indeed, virtually all solutions of the optimal stopping problem (and all solutions prior to the discovery of the dual characterization in the early 21st century) are of this type. There are hybrid methods, such as that of Andersen and Broadie, but where is the pure dual method? It is the purpose of this short note to demonstrate how the solution may be derived by purely dual methods akin to the backward recursion of dynamic programming. The key observation, which is obvious from (1.1), is that the value of the optimal stopping problem is left unaltered if $Z$ is replaced by $Z-M$, where $M \in \mathcal{M}_{0}$.
2. The algorithm. In this section, we specify the algorithm by which the given Bermudan option is to be hedged. We are given a reward process $\left(Z_{t}\right)_{t=0, \ldots, T}$ adapted to the filtration $\left(\mathcal{F}_{t}\right)_{t=0, \ldots, T}$, and the aim is to find some martingale $M \in \mathcal{M}_{0}$ such that (1.1) holds. If (1.1) holds, then by the earlier result of [3], [2] we also have (1.2). The construction is based on two very simple observations:

1. The value of the stopping problem for $Z$ is the same as the value of the stopping problem for $Z+N$, where $N$ is any martingale in $\mathcal{M}_{0}$.
2. Adding a constant to $Z$ adds a constant to the value.

The proof of the following result shows how to solve the problem recursively, by constructing a sequence of martingales which do an ever better job of hedging. The idea is that the pathwise maximum must become a constant random variable (see (1.2)); while it is not obvious how we shall achieve this in one go, we can easily see how to ensure that the final value of $Z$ is a constant-by subtracting a martingale which is equal to $Z_{T}$ at time $T$. The inductive proof constructs martingales which are constant on some interval $[T-k, T]$ for ever bigger $k$.

Proposition 1. There exist a sequence of constants $a_{j}$ and a sequence of martingales $N^{(j)} \in$ $\mathcal{M}_{0}, j=1, \ldots, T+1$, such that

$$
\begin{equation*}
\max _{T-j<i \leq T} Z_{i}^{(j)}=a_{j}, \tag{2.1}
\end{equation*}
$$

where

$$
\begin{equation*}
Z_{t}^{(j)} \equiv Z_{t}-N_{t}^{(j)} \tag{2.2}
\end{equation*}
$$

Proof. The proof proceeds by induction on $j$. To start the induction, we consider the martingale

$$
M_{t}^{(1)}=E_{t}\left[Z_{T}\right]
$$

and set

$$
N_{t}^{(1)}=M_{t}^{(1)}-E\left[M_{T}^{(1)}\right]
$$

which is clearly in $\mathcal{M}_{0}$ and equally clearly achieves (2.1) for $j=1$, with $a_{1}=E\left[M_{T}^{(1)}\right]=E\left[Z_{T}\right]$.
Now suppose that (2.1) is true for $j \leq k$, and consider the nonnegative martingale

$$
\begin{equation*}
M_{t}^{(k+1)}=E_{t}\left[\left\{Z_{T-k}^{(k)}-a_{k}\right\}^{+}\right] . \tag{2.3}
\end{equation*}
$$

This martingale is constant in $[T-k, T]$; we shall subtract it from $Z^{(k)}$ to form the process $\tilde{Z}_{t}^{(k)} \equiv Z_{t}^{(k)}-M_{t}^{(k+1)}$. Two cases then need to be considered.

1. If $Z_{T-k}^{(k)}>a_{k}$, then we have $\tilde{Z}_{T-k}^{(k)}=a_{k}$, and $\tilde{Z}_{t}^{(k)} \leq Z_{t}^{(k)}$ for all $t>T-k$, since $M^{(k+1)}$ is nonnegative. Thus

$$
\max _{T-k \leq t \leq T} \tilde{Z}_{t}^{(k)}=a_{k}
$$

2. If $Z_{T-k}^{(k)} \leq a_{k}$, then $M^{(k+1)}$ is zero in [T-k,T], and by the inductive hypothesis

$$
\max _{T-k \leq t \leq T} \tilde{Z}_{t}^{(k)}=\max _{T-k<t \leq T} Z_{t}^{(k)}=a_{k}
$$

Either way, the conclusion is the same, namely, that

$$
\begin{equation*}
\max _{T-k \leq t \leq T} \tilde{Z}_{t}^{(k)}=a_{k} . \tag{2.4}
\end{equation*}
$$

We now define

$$
\begin{align*}
N^{(k+1)} & =N^{(k)}+M^{(k+1)}-E\left[M_{T}^{(k+1)}\right],  \tag{2.5}\\
a_{k+1} & =a_{k}+E\left[M_{T}^{(k+1)}\right], \tag{2.6}
\end{align*}
$$

so that

$$
\begin{aligned}
Z_{t}^{(k+1)} & \equiv Z_{t}-N_{t}^{(k+1)} \\
& =\tilde{Z}_{t}^{(k)}+E\left[M_{T}^{(k+1)}\right]
\end{aligned}
$$

satisfies (2.1), taking $j=k+1$.
Remarks. (i) Applying the proposition in the case $j=T+1$, we learn that

$$
\begin{equation*}
\max _{0 \leq i \leq T} Z_{i}^{(T+1)} \equiv \max _{0 \leq i \leq T}\left\{Z_{i}-N_{i}^{(T+1)}\right\}=a_{T+1} \tag{2.7}
\end{equation*}
$$

Hence the value of the optimal stopping problem with reward process $Z^{(T+1)}$ is equal to $a_{T+1}$, and by the first observation, this is actually the value of the optimal stopping problem for the original reward process $Z$.
(ii) The recursive construction generates an increasing sequence $a_{1} \leq a_{2} \leq \cdots \leq a_{T+1}$, increasing the value of the problem as well as the hedging martingale.
(iii) Observe that by adding (2.5) and (2.6) we learn that

$$
\begin{equation*}
N^{(k)}+a_{k}=\sum_{i=1}^{k} M^{(i)} . \tag{2.8}
\end{equation*}
$$

From (2.3) and (2.2) we deduce that

$$
\begin{aligned}
M_{t}^{(k+1)} & =E_{t}\left[\left\{Z_{T-k}^{(k)}-a_{k}\right\}^{+}\right] \\
& =E_{t}\left[\left\{Z_{T-k}-N_{T-k}^{(k)}-a_{k}\right\}^{+}\right] \\
& =E_{t}\left[\left\{Z_{T-k}-\sum_{i=1}^{k} M_{T-k}^{(i)}\right\}^{+}\right]
\end{aligned}
$$

and taking $t=T-k$ leads to the conclusion

$$
\begin{equation*}
M_{T-k}^{(k+1)}=\left\{Z_{T-k}-\sum_{i=1}^{k} M_{T-k}^{(i)}\right\}^{+} \tag{2.9}
\end{equation*}
$$

The statement and proof of Proposition 1 is "pure dual"; there is no mention of the value of the stopping problem; we talked only about the hedging martingales $M^{(k)}$. To tie things together, we shall now show how the constructs from the proof of Proposition 1 relate to the more familiar value process of the optimal stopping problem,

$$
\begin{equation*}
Y_{t}^{*} \equiv \sup _{\tau \in \mathcal{T}, \tau \geq t} E_{t}\left[Z_{\tau}\right] . \tag{2.10}
\end{equation*}
$$

As is well known, $Y^{*}$ is the Snell envelope process of the reward process $Z$, and $Y_{t}^{*}$ is interpreted as the best that can be done if by time $t$ the process has not been stopped.

Proposition 2. For all $k=0,1, \ldots, T$ we have

$$
\begin{equation*}
\sum_{i=1}^{k+1} M_{T-k}^{(i)}=Y_{T-k}^{*} \tag{2.11}
\end{equation*}
$$

Proof. The proof is by induction on $k$. Clearly the statement is true if $k=0$, for both sides of (2.11) are equal to $Z_{T}$. Suppose now that the statement is true for all $k<n$, and
consider

$$
\begin{aligned}
\sum_{i=1}^{n+1} M_{T-n}^{(i)} & =M_{T-n}^{(n+1)}+\sum_{i=1}^{n} M_{T-n}^{(i)} \\
& =M_{T-n}^{(n+1)}+E_{T-n}\left[Y_{T-n+1}^{*}\right] \quad \text { by inductive hypothesis } \\
& =\left\{Z_{T-n}-\sum_{i=1}^{n} M_{T-n}^{(i)}\right\}^{+}+E_{T-n}\left[Y_{T-n+1}^{*}\right] \quad \text { using (2.9) } \\
& =\left\{Z_{T-n}-E_{T-n}\left[Y_{T-n+1}^{*}\right]\right\}^{+}+E_{T-n}\left[Y_{T-n+1}^{*}\right] \quad \text { by inductive hypothesis } \\
& =\max \left\{Z_{T-n}, E_{T-n}\left[Y_{T-n+1}^{*}\right]\right\} \\
& =Y_{T-n}^{*},
\end{aligned}
$$

as required.
Remarks. (i) From (2.6) and Proposition 2 we see that $a_{k}=E\left[Y_{T-k+1}^{*}\right]$, which explains why the sequence $a_{k}$ is increasing and why $a_{T+1}$ is the value of the problem.
(ii) If we restrict our attention to problems where there is some underlying Markovian structure, it is not hard to see that the martingales $M^{(k)}$ constructed are characterized ${ }^{3}$ as $M_{T-k}^{(k+1)}=\varphi_{k}\left(X_{T-k}\right)$ for all $k=0,1, \ldots, T$. In trying to use (2.9) to determine the functions $\varphi_{k}$ recursively, we are faced with the two steps, conditional expectation and pointwise maximization, which are central to the standard dynamic programming approach. Thus it seems unlikely that in problems with Markovian structure the pure dual approach presented here will generate numerical methodologies that differ significantly from existing methodologies for such examples.

## REFERENCES

[1] L. Andersen and M. Broadie, A primal-dual simulation algorithm for pricing multidimensional American options, Management Sci., 50 (2004), pp. 1222-1234.
[2] M. B. Haugh and L. Kogan, Pricing American options: A duality approach, Oper. Res., 52 (2004), pp. 258-270.
[3] L. C. G. Rogers, Monte Carlo valuation of American options, Math. Finance, 12 (2002), pp. 271-286.

[^56]
# Asymptotic Formulas with Error Estimates for Call Pricing Functions and the Implied Volatility at Extreme Strikes* 

Archil Gulisashvili ${ }^{\dagger}$


#### Abstract

In this paper, we obtain asymptotic formulas with error estimates for the implied volatility associated with a European call pricing function. We show that these formulas imply Lee's moment formulas for the implied volatility and the tail-wing formulas due to Benaim and Friz. In addition, we analyze Pareto-type tails of stock price distributions in uncorrelated Hull-White, Stein-Stein, and Heston models and find asymptotic formulas with error estimates for call pricing functions in these models.


Key words. call and put pricing functions, implied volatility, asymptotic formulas, Pareto-type distributions, regularly varying functions

AMS subject classifications. $91 \mathrm{G} 20,91 \mathrm{G} 80$
DOI. 10.1137/090762713

1. Introduction. In this paper, we study the asymptotic behavior of the implied volatility $K \mapsto I(K)$ associated with a European type call pricing function $K \mapsto C(K)$. Here the symbol $K$ stands for the strike price, and it is assumed that the expiry $T$ is fixed. One of the main results obtained in the present paper is the following asymptotic formula:
(1) $I(K)=\frac{\sqrt{2}}{\sqrt{T}}\left[\sqrt{\log K+\log \frac{1}{C(K)}}-\sqrt{\log \frac{1}{C(K)}}\right]+O\left(\left(\log \frac{1}{C(K)}\right)^{-\frac{1}{2}} \log \log \frac{1}{C(K)}\right)$
as $K \rightarrow \infty$. A similar formula holds for $K$ near zero:
(2) $I(K)=\frac{\sqrt{2}}{\sqrt{T}}\left[\sqrt{\log \frac{1}{P(K)}}-\sqrt{\log \frac{1}{P(K)}-\log \frac{1}{K}}\right]+O\left(\left(\log \frac{K}{P(K)}\right)^{-\frac{1}{2}} \log \log \frac{K}{P(K)}\right)$
as $K \rightarrow 0$, where $K \mapsto P(K)$ is the put pricing function corresponding to $C$. Formulas (1) and (2) are valid under very mild restrictions on call and put pricing function. For special stochastic volatility models, sharp asymptotic formulas for the implied volatility were established in the papers of E. M. Stein and the author (see [20, 22, 19, 21]). In sections 4 and 5 , we compare formulas (1) and (2) with known asymptotic formulas for the implied volatility. For instance, it will be shown that Lee's moment formulas (see [29]) and the tail-wing formulas due to Benaim and Friz (see [2]) can be derived using (1) and (2).

We model the random behavior of the stock price by a nonnegative adapted stochastic process $X$ defined on a filtered probability space $\left(\Omega, \mathcal{F}, \mathcal{F}_{t}, \mathbb{P}^{*}\right)$. It is supposed that $X_{0}=x_{0}$ $\mathbb{P}^{*}$-a.s. for some number $x_{0}>0$ and that $\mathbb{E}^{*}\left[X_{t}\right]<\infty$ for every $t \geq 0$. In addition, we assume

[^57]that $\mathbb{P}^{*}$ is a risk-free measure. This means that the discounted stock price process $\left\{e^{-r t} X_{t}\right\}_{t \geq 0}$ is a $\left(\mathcal{F}_{t}, \mathbb{P}^{*}\right)$-martingale. Here $r \geq 0$ denotes the interest rate. It follows from the martingale condition that
\[

$$
\begin{equation*}
x_{0}=e^{-r t} \mathbb{E}^{*}\left[X_{t}\right] \quad \text { for all } \quad t \geq 0 \tag{3}
\end{equation*}
$$

\]

The following standard notation will be used in the paper: $u^{+}=\max \{u, 0\}$, where $u$ is a real number. The next definition introduces European style call and put pricing functions. Recall that we denoted by $r, X, x_{0}$, and $\mathbb{P}^{*}$ the interest rate, the stock price process, the initial condition, and the risk-free measure, respectively.

Definition 1.1. Given $r, X, x_{0}$, and $\mathbb{P}^{*}$, the European call pricing function $C$ at time $t=0$ is defined on $[0, \infty)^{2}$ by the following formula:

$$
\begin{equation*}
C(T, K)=e^{-r T} \mathbb{E}^{*}\left[\left(X_{T}-K\right)^{+}\right] \tag{4}
\end{equation*}
$$

Similarly, the European put pricing function $P$ at time $t=0$ is defined by

$$
\begin{equation*}
P(T, K)=e^{-r T} \mathbb{E}^{*}\left[\left(K-X_{T}\right)^{+}\right] \tag{5}
\end{equation*}
$$

The functions $C$ and $P$ satisfy the put-call parity condition $C(T, K)=P(T, K)+x_{0}-$ $e^{-r T} K$. This formula can be easily derived from (4) and (5).

A popular example of a call pricing function is the function $C_{B S}$ arising in the BlackScholes model. In this model, the stock price process is a geometric Brownian motion, satisfying the stochastic differential equation $d X_{t}=r X_{t} d t+\sigma X_{t} d W_{t}$, where $r \geq 0$ is the interest rate, $\sigma>0$ is the volatility of the stock, and $W$ is a standard Brownian motion. The process $X$ is given by

$$
\begin{equation*}
X_{t}=x_{0} \exp \left\{\left(r-\frac{\sigma^{2}}{2}\right) t+\sigma W_{t}\right\} \tag{6}
\end{equation*}
$$

where $x_{0}>0$ is the initial condition. Black and Scholes found an explicit formula for the pricing function $C_{B S}$. This formula is as follows:

$$
C_{B S}(T, K, \sigma)=x_{0} N\left(d_{1}(K, \sigma)\right)-K e^{-r T} N\left(d_{2}(K, \sigma)\right)
$$

where

$$
d_{1}(K, \sigma)=\frac{\log x_{0}-\log K+\left(r+\frac{1}{2} \sigma^{2}\right) T}{\sigma \sqrt{T}}, \quad d_{2}(K, \sigma)=d_{1}(K, \sigma)-\sigma \sqrt{T}
$$

and

$$
N(z)=\frac{1}{\sqrt{2 \pi}} \int_{-\infty}^{z} \exp \left\{-\frac{y^{2}}{2}\right\} d y
$$

(see, e.g., [28]).
Remark 1.2. It is known that, for fixed $K>0$ and $T>0$, the function $\rho(\sigma)=C_{B S}(T, K, \sigma)$ is strictly increasing on $(0, \infty)$. If $0<K<x_{0} e^{r T}$, then the range of the function $\rho$ is the interval $\left(x_{0}-K e^{-r T}, x_{0}\right)$, while, for $x_{0} e^{r T} \geq K$, the range of $\rho$ coincides with the interval $\left(0, x_{0}\right)$.

Definition 1.3. Let $C$ be a call pricing function. For $(T, K) \in(0, \infty)^{2}$, the implied volatility $I(T, K)$ associated with $C$ is equal to the volatility $\sigma$ in the Black-Scholes model such that $C(T, K)=C_{B S}(T, K, \sigma)$. The implied volatility is defined only if such a number $\sigma$ exists and is unique.

Using Remark 1.2, we see that, for $T>0$ and $K>0$ with $0<K<x_{0} e^{r T}$, a necessary condition for the existence of the implied volatility $I(T, K)$ is as follows: $x_{0}-K e^{-r T}<$ $C(T, K)<x_{0}$. Similarly, $I(T, K)$ is defined for $T>0$ and $K>0$ with $x_{0} e^{r T} \leq K$ if and only if $0<C(T, K)<x_{0}$. Note that the inequality $C(T, K)<x_{0}, T \geq 0, K>0$, is always true. Moreover, if $(T, K) \in[0, \infty)^{2}$, then $\left(x_{0}-K e^{-r T}\right)^{+} \leq C(T, K)$.

In the next definitions, we introduce classes of call pricing functions.
Definition 1.4. The class $P F_{\infty}$ consists of all call pricing functions $C$, for which one of the following equivalent conditions holds:

1. $C(T, K)>0$ for all $T>0$ and $K>0$ with $x_{0} e^{r T} \leq K$.
2. $P(T, K)>e^{-r T} K-x_{0}$ for all $T>0$ and $K>0$ with $x_{0} e^{r T} \leq K$.
3. For every $T>0$ and all $a>0$, the random variable $X_{T}$ satisfies the condition $\mathbb{P}^{*}\left[X_{T}<a\right]<1$.
Definition 1.5. The class $P F_{0}$ consists of all call pricing functions $C$, for which one of the following equivalent conditions holds:
4. $P(T, K)>0$ for all $T>0$ and $K>0$ with $K<x_{0} e^{r T}$.
5. $C(T, K)>x_{0}-e^{-r T} K$ for all $T>0$ and $K>0$ with $K<x_{0} e^{r T}$.
6. For every $T>0$ and all $a>0$, the random variable $X_{T}$ satisfies the condition $0<$ $\mathbb{P}^{*}\left[X_{T}<a\right]$.
Remark 1.6. Suppose that the expiry $T>0$ is fixed, and consider the pricing function $K \mapsto C(K)$ and the implied volatility $K \mapsto I(K)$ as functions of the strike price $K$. If $C \in P F_{\infty}$, then the implied volatility $I(K)$ is defined for large values of $K$. This allows one to study the asymptotic behavior of the implied volatility as $K \rightarrow \infty$. Similarly, if $C \in P F_{0}$, then $I(K)$ exists for small values of $K$. In this case, it is interesting to determine how the implied volatility behaves near zero. Finally, if $C \in P F_{\infty} \cap P F_{0}$, then the implied volatility $I(T, K)$ exists for all $T>0$ and $K>0$.

We refer the interested reader to [13, 14, 17, 24] for additional information on the implied volatility. The asymptotic behavior of the implied volatility for extreme strikes was studied in $[2,3,4,22,21,29]$ (see also sections 10.5 and 10.6 of [24]).

In the present paper, we use various asymptotic relations between functions.
Definition 1.7. Let $\varphi_{1}$ and $\varphi_{2}$ be positive functions on $(a, \infty)$. We define several asymptotic relations by the following:

1. If there exist $\alpha_{1}>0, \alpha_{2}>0$, and $y_{0}>0$ such that $\alpha_{1} \varphi_{1}(y) \leq \varphi_{2}(y) \leq \alpha_{2} \varphi_{1}(y)$ for all $y>y_{0}$, then we write $\varphi_{1}(y) \approx \varphi_{1}(y)$ as $y \rightarrow \infty$.
2. If the condition $\lim _{y \rightarrow \infty}\left[\varphi_{2}(y)\right]^{-1} \varphi_{1}(y)=1$ holds, then we write $\varphi_{1}(y) \sim \varphi_{2}(y)$ as $y \rightarrow \infty$.
3. Let $\rho$ be a positive function on $(0, \infty)$. We use the notation $\varphi_{1}(y)=\varphi_{2}(y)+O(\rho(y))$ as $y \rightarrow \infty$ if there exist $\alpha>0$ and $y_{0}>0$ such that $\left|\varphi_{1}(y)-\varphi_{2}(y)\right| \leq \alpha \rho(y)$ for all $y>y_{0}$.
Similar relations can be defined in the case where $y \downarrow 0$.
We will next give a quick overview of the results obtained in the present paper. In section

2, various asymptotic formulas are established for the implied volatility $K \mapsto I(K)$ as $K \rightarrow \infty$. In section 3, the asymptotic behavior of the implied volatility is studied in the case where $K \rightarrow 0$. We also discuss a symmetry property of the implied volatility and formulate a characterization theorem for call pricing functions (see section 3 ). In section 4 , we give a new proof of Lee's moment formulas for the implied volatility, while in section 5 we compare our asymptotic formulas with the tail-wing formulas due to Benaim and Friz. We also obtain tailwing formulas with error estimates under certain restrictions expressed in terms of smoothly varying functions. In section 6, we talk about Pareto-type tails of stock price distributions in uncorrelated Hull-White, Stein-Stein, and Heston models. For these distributions, we compute the Pareto-type index and provide explicit expressions for the corresponding slowly varying functions. In section 7, we obtain sharp asymptotic formulas for call pricing functions in special stochastic volatility models and find tail-wing formulas with error estimates in the case where the stock price density is equivalent to a regularly varying function. Finally, in the appendix, we prove the characterization theorem for call pricing functions formulated in section 3.
2. Asymptotic behavior of the implied volatility as $K \rightarrow \infty$. In this section, we find sharp asymptotic formulas for the implied volatility $K \mapsto I(K)$ associated with a general call pricing function $C$.

Theorem 2.1. Let $C \in P F_{\infty}$, and let $\psi$ be a positive function with $\lim _{K \rightarrow \infty} \psi(K)=\infty$. Then

$$
\begin{align*}
& I(K)=\frac{1}{\sqrt{T}}\left[\sqrt{2 \log K+2 \log \frac{1}{C(K)}-\log \log \frac{1}{C(K)}}-\sqrt{2 \log \frac{1}{C(K)}-\log \log \frac{1}{C(K)}}\right] \\
& \quad+O\left(\left(\log \frac{1}{C(K)}\right)^{-\frac{1}{2}} \psi(K)\right) \tag{7}
\end{align*}
$$

as $K \rightarrow \infty$.
Theorem 2.1 and the mean value theorem imply the following assertion.
Corollary 2.2. For any call pricing function $C \in P F_{\infty}$,
(8) $I(K)=\frac{\sqrt{2}}{\sqrt{T}}\left[\sqrt{\log K+\log \frac{1}{C(K)}}-\sqrt{\log \frac{1}{C(K)}}\right]+O\left(\left(\log \frac{1}{C(K)}\right)^{-\frac{1}{2}} \log \log \frac{1}{C(K)}\right)$
as $K \rightarrow \infty$.
Proof of Theorem 2.1. The following lemma was established in [22, 21] under certain restrictions on the pricing function. The proof in the general case is similar.

Lemma 2.3. Let $C$ be a call pricing function, and fix a positive continuous increasing function $\psi$ with $\lim _{K \rightarrow \infty} \psi(K)=\infty$. Suppose $\phi$ is a positive function such that $\lim _{K \rightarrow \infty} \phi(K)=\infty$ and

$$
\begin{equation*}
C(K) \approx \frac{\psi(K)}{\phi(K)} \exp \left\{-\frac{\phi(K)^{2}}{2}\right\} \tag{9}
\end{equation*}
$$

Then the following asymptotic formula holds:

$$
I(K)=\frac{1}{\sqrt{T}}\left(\sqrt{2 \log \frac{K}{x_{0} e^{r T}}+\phi(K)^{2}}-\phi(K)\right)+O\left(\frac{\psi(K)}{\phi(K)}\right)
$$

as $K \rightarrow \infty$.
Remark 2.4. It is easy to see that if (9) holds, then we have $C \in P F_{\infty}$.
Let us continue the proof of Theorem 2.1. With no loss of generality, we can assume that the function $\psi(K)$ tends to infinity slower than the function $K \mapsto \log \log \frac{1}{C(K)}$. Put

$$
\phi(K)=\left[2 \log \frac{1}{C(K)}-\log \log \frac{1}{C(K)}+2 \log \psi(K)\right]^{\frac{1}{2}}
$$

Then we have $\phi(K) \approx \sqrt{2 \log \frac{1}{C(K)}}$ as $K \rightarrow \infty$, and it follows that

$$
\psi(K) \exp \left\{-\frac{\phi(K)^{2}}{2}\right\} \phi(K)^{-1} \approx C(K), \quad K \rightarrow \infty
$$

Therefore, Lemma 2.3 gives

$$
\begin{equation*}
I(K)=\frac{1}{\sqrt{T}}\left(\sqrt{2 \log \frac{K}{x_{0} e^{r T}}+\phi(K)^{2}}-\phi(K)\right)+O\left(\left(\log \frac{1}{C(K)}\right)^{-\frac{1}{2}} \psi(K)\right) \tag{10}
\end{equation*}
$$

as $K \rightarrow \infty$. Now it is not hard to see that (10) and the mean value theorem imply (7).
This completes the proof of Theorem 2.1.
Our next goal is to replace the function $C$ in formula (7) by another function $\widetilde{C}$.
Corollary 2.5. Let $C \in P F_{\infty}$, and let $\psi$ be a positive function with $\lim _{K \rightarrow \infty} \psi(K)=\infty$. Suppose that $\widetilde{C}$ is a positive function such that $\widetilde{C}(K) \approx C(K)$ as $K \rightarrow \infty$. Then

$$
\begin{aligned}
& I(K)=\frac{1}{\sqrt{T}}\left[\sqrt{2 \log K+2 \log \frac{1}{\widetilde{C}(K)}-\log \log \frac{1}{\widetilde{C}(K)}}-\sqrt{2 \log \frac{1}{\widetilde{C}(K)}-\log \log \frac{1}{\widetilde{C}(K)}}\right] \\
& \quad+O\left(\left(\log \frac{1}{\widetilde{C}(K)}\right)^{-\frac{1}{2}} \psi(K)\right)
\end{aligned}
$$

as $K \rightarrow \infty$. In addition,

$$
\begin{equation*}
I(K)=\frac{\sqrt{2}}{\sqrt{T}}\left[\sqrt{\log K+\log \frac{1}{\widetilde{C}(K)}}-\sqrt{\log \frac{1}{\widetilde{C}(K)}}\right]+O\left(\left(\log \frac{1}{\widetilde{C}(K)}\right)^{-\frac{1}{2}} \log \log \frac{1}{\widetilde{C}(K)}\right) \tag{12}
\end{equation*}
$$

as $K \rightarrow \infty$.
Formula (11) can be established exactly as (7). Formula (12) follows from (11) and the mean value theorem.

We can also replace a call pricing function $C$ in (7) by a function $\widetilde{C}$ under more general conditions. However, this may lead to a weaker error estimate. Put

$$
\begin{equation*}
\tau(K)=\left|\log \frac{1}{C(K)}-\log \frac{1}{\widetilde{C}(K)}\right| \tag{13}
\end{equation*}
$$

Then the following theorem holds.
Theorem 2.6. Let $C \in P F_{\infty}$, and let $\psi$ be a positive function with $\lim _{K \rightarrow \infty} \psi(K)=\infty$. Suppose $\widetilde{C}$ is a positive function satisfying the following condition: There exist $K_{1}>0$ and $c$ with $0<c<1$ such that

$$
\begin{equation*}
\tau(K)<c \log \frac{1}{\widetilde{C}(K)} \quad \text { for all } \quad K>K_{1} \tag{14}
\end{equation*}
$$

where $\tau$ is defined by (13). Then

$$
\begin{align*}
& I(K)=\frac{1}{\sqrt{T}}\left[\sqrt{2 \log K+2 \log \frac{1}{\widetilde{C}(K)}-\log \log \frac{1}{\widetilde{C}(K)}}-\sqrt{2 \log \frac{1}{\widetilde{C}(K)}-\log \log \frac{1}{\widetilde{C}(K)}}\right] \\
& \quad+O\left(\left(\log \frac{1}{\widetilde{C}(K)}\right)^{-\frac{1}{2}}[\psi(K)+\tau(K)]\right) \tag{15}
\end{align*}
$$

as $K \rightarrow \infty$.
Proof. It is not hard to check that (14) implies the formula $\log \frac{1}{C(K)} \approx \log \frac{1}{C(K)}$ as $K \rightarrow \infty$. Now, using (7), (13), and the mean value theorem, we obtain (15).

The next statement follows from the special case of Theorem 2.6, where $\psi(K)=\log \log \frac{1}{C(K)}$, and from the mean value theorem.

Corollary 2.7. Let $C \in P F_{\infty}$, and suppose $\widetilde{C}$ is a positive function satisfying the following condition: There exist $\nu>0$ and $K_{0}>0$ such that

$$
\begin{equation*}
\left|\log \frac{1}{\widetilde{C}(K)}-\log \frac{1}{C(K)}\right| \leq \nu \log \log \frac{1}{\widetilde{C}(K)}, \quad K>K_{0} \tag{16}
\end{equation*}
$$

Then

$$
I(K)=\frac{\sqrt{2}}{\sqrt{T}}\left[\sqrt{\log K+\log \frac{1}{\widetilde{C}(K)}}-\sqrt{\log \frac{1}{\widetilde{C}(K)}}\right]+O\left(\left(\log \frac{1}{\widetilde{C}(K)}\right)^{-\frac{1}{2}} \log \log \frac{1}{\widetilde{C}(K)}\right)
$$

as $K \rightarrow \infty$.
Remark 2.8. It is not hard to see that if $C(K) \approx \widetilde{C}(K)$ as $K \rightarrow \infty$, or if (16) holds, then $\log \frac{1}{C(K)} \sim \log \frac{1}{C(K)}$ as $K \rightarrow \infty$.

Corollary 2.9. Let $C \in P F_{\infty}$, and suppose $\widetilde{C}$ is a positive function satisfying the condition

$$
\begin{equation*}
\log \frac{1}{C(K)} \sim \log \frac{1}{\widetilde{C}(K)} \tag{17}
\end{equation*}
$$

as $K \rightarrow \infty$. Then

$$
\begin{equation*}
I(K) \sim \frac{\sqrt{2}}{\sqrt{T}}\left[\sqrt{\log K+\log \frac{1}{\widetilde{C}(K)}}-\sqrt{\log \frac{1}{\widetilde{C}(K)}}\right] \tag{18}
\end{equation*}
$$

as $K \rightarrow \infty$.
Proof. It follows from (8) that

$$
\begin{equation*}
I(K) \sim \frac{\sqrt{2}}{\sqrt{T}}\left[\sqrt{\log K+\log \frac{1}{\widetilde{C}(K)}}-\sqrt{\log \frac{1}{\widetilde{C}(K)}}\right] \Lambda(K) \tag{19}
\end{equation*}
$$

where

$$
\Lambda(K)=\frac{\sqrt{\log K+\log \frac{1}{\widetilde{C(K)}}}+\sqrt{\log \frac{1}{\widetilde{C(K)}}}}{\sqrt{\log K+\log \frac{1}{C(K)}}+\sqrt{\log \frac{1}{C(K)}}}
$$

Our next goal is to prove that $\Lambda(K) \rightarrow 1$ as $K \rightarrow \infty$. We have

$$
\Lambda(K)=\frac{\sqrt{\Lambda_{1}(K)+\Lambda_{2}(K)}+\sqrt{\Lambda_{2}(K)}}{\sqrt{\Lambda_{1}(K)+1}+1}
$$

where

$$
\Lambda_{1}(K)=\frac{\log K}{\log \frac{1}{C(K)}} \quad \text { and } \quad \Lambda_{2}(K)=\frac{\log \frac{1}{\widetilde{C(K)}}}{\log \frac{1}{C(K)}}
$$

It is not hard to show that, for all positive numbers $a$ and $b,|\sqrt{a+b}-\sqrt{a+1}| \leq|\sqrt{b}-1|$. Therefore,

$$
\begin{align*}
|\Lambda(K)-1| & =\frac{\left|\sqrt{\Lambda_{1}(K)+\Lambda_{2}(K)}-\sqrt{\Lambda_{1}(K)+1}\right|+\left|\sqrt{\Lambda_{2}(K)}-1\right|}{\sqrt{\Lambda_{1}(K)+1}+1} \\
& \leq\left|\sqrt{\Lambda_{2}(K)}-1\right| \tag{20}
\end{align*}
$$

for $K>K_{0}$. It follows from (17) and (20) that $\Lambda(K) \rightarrow 1$ as $K \rightarrow \infty$. Next, using (19), we see that (18) holds.

This completes the proof of Corollary 2.9.
3. Asymptotic behavior of the implied volatility as $\boldsymbol{K} \rightarrow \mathbf{0}$. In this section, we turn our attention to the behavior of the implied volatility as the strike price tends to zero. Let $C$ be a general call pricing function, and let $X$ be a corresponding stock price process. This process is defined on a filtered probability space $\left(\Omega, \mathcal{F}, \mathcal{F}_{t}, \mathbb{P}^{*}\right)$, where $\mathbb{P}^{*}$ is a risk-free probability measure. As before, we assume that the interest rate $r$, the initial condition $x_{0}$, and the expiry $T$ are fixed, and we denote by $\mu_{T}$ the distribution of the random variable $X_{T}$. The Black-Scholes pricing function $C_{B S}$ satisfies the following condition:

$$
\begin{equation*}
C_{B S}(T, K, \sigma)=x_{0}-K e^{-r T}+\frac{K e^{-r T}}{x_{0}} C_{B S}\left(T,\left(x_{0} e^{r T}\right)^{2} K^{-1}, \sigma\right) \tag{21}
\end{equation*}
$$

A similar formula holds for the pricing function $C$. Indeed, it is not hard to prove using the put/call parity formula that

$$
\begin{equation*}
C(T, K)=x_{0}-K e^{-r T}+\frac{K e^{-r T}}{x_{0}} G\left(T,\left(x_{0} e^{r T}\right)^{2} K^{-1}\right), \tag{22}
\end{equation*}
$$

where the function $G$ is given by

$$
\begin{equation*}
G(T, K)=\frac{K}{x_{0} e^{r T}} P\left(T,\left(x_{0} e^{r T}\right)^{2} K^{-1}\right) . \tag{23}
\end{equation*}
$$

It follows from (22) that

$$
G\left(T,\left(x_{0} e^{r T}\right)^{2} K^{-1}\right)=x_{0} \int_{0}^{K} d \mu_{T}(x)-\frac{x_{0}}{K} \int_{0}^{K} x d \mu_{T}(x) .
$$

Define a family of Borel measures $\left\{\widetilde{\mu}_{t}\right\}_{t \geq 0}$ on $(0, \infty)$ as follows: For any Borel set $A$ in $(0, \infty)$, put

$$
\begin{equation*}
\widetilde{\mu}_{t}(A)=\frac{1}{x_{0} e^{r t}} \int_{\eta^{-1}(A)} x d \mu_{t}(x), \tag{24}
\end{equation*}
$$

where $\eta(K)=\left(x_{0} e^{r T}\right)^{2} K^{-1}, K>0$. It is not hard to see that $\widetilde{\mu}_{t}((0, \infty))=1$ for all $t \geq 0$. Moreover, for every Borel set $A$ in $(0, \infty)$, we have

$$
\begin{equation*}
\int_{\eta(A)} d \widetilde{\mu}(x)=\frac{1}{x_{0} e^{r T}} \int_{A} x d \mu_{T}(x) \quad \text { and } \quad \int_{\eta(A)} x d \widetilde{\mu}_{T}(x)=x_{0} e^{r T} \int_{A} d \mu_{T}(x) . \tag{25}
\end{equation*}
$$

It follows from (23) and (25) that

$$
\begin{equation*}
G(T, K)=e^{-r T} \int_{K}^{\infty} x d \widetilde{\mu}_{T}(x)-e^{-r T} K \int_{K}^{\infty} d \widetilde{\mu}_{T}(x) . \tag{26}
\end{equation*}
$$

Remark 3.1. Suppose that, for every $t>0$, the measure $\mu_{t}$ is absolutely continuous with respect to the Lebesgue measure on $(0, \infty)$. Denote the Radon-Nikodym derivative by $D_{t}$. Then, for every $t>0$, the measure $\widetilde{\mu}_{t}$ admits a density $\widetilde{D}_{t}$ given by

$$
\widetilde{D}_{t}(x)=\left(x_{0} e^{r t}\right)^{3} x^{-3} D_{t}\left(\left(x_{0} e^{r t}\right)^{2} x^{-1}\right), \quad x>0 .
$$

Theorem 3.2. Let $C$ be a call pricing function, and let $P$ be the corresponding put pricing function. Denote by $\left\{\mu_{t}\right\}_{t \geq 0}$ the family of marginal distributions of the stock price process $X$, and define a family of measures $\left\{\widetilde{\mu}_{t}\right\}_{t \geq 0}$ by formula (24) and a function $G$ by formula (26). Then $G$ is a call pricing function with the same interest rate $r$ and the initial condition $x_{0}$ as the pricing function $C$, and it has a stock price process $\widetilde{X}$ having $\left\{\widetilde{\mu}_{t}\right\}_{t \geq 0}$ as the family of its marginal distributions.

We will need the following characterization of call pricing functions.
Theorem 3.3. A nonnegative function $C$ defined on $[0, \infty)^{2}$ is a call pricing function at time $t=0$ with interest rate $r$ and initial condition $x_{0}$ if and only if the following hold:

1. For every $T \geq 0$, the function $K \rightarrow C(T, K)$ is convex.
2. For every $T \geq 0$, the second distributional derivative $\mu_{T}$ of the function $K \mapsto e^{r T} C(T, K)$ is a Borel probability measure such that

$$
\begin{equation*}
\int_{0}^{\infty} x d \mu_{T}(x)=x_{0} e^{r T} \tag{27}
\end{equation*}
$$

3. For every $K \geq 0$, the function $T \rightarrow C\left(T, e^{r T} K\right)$ is nondecreasing.
4. For every $K \geq 0, C(0, K)=\left(x_{0}-K\right)^{+}$.
5. For every $T \geq 0, \lim _{K \rightarrow \infty} C(T, K)=0$.

Remark 3.4. Let us note that similar, but a little less general, characterizations of pricing functions are known. For instance, in section 1 of the paper [7] of Carmona and Nadtochiy, there is a description of conditions which are imposed on call pricing functions by the absence of arbitrage. This description essentially coincides with the necessity part of Theorem 3.3. We also refer the reader to section 3 of the paper [6] of Buehler, where results related to Theorem 3.3 were established. The proof of Theorem 3.3 uses standard methods and is given in the appendix for the sake of completeness.

Proof of Theorem 3.2. It suffices to prove that conditions 1-5 in Theorem 3.3 are valid for the function $G$. We have $\widetilde{\mu}_{T}([0, \infty))=1$. This equality follows from (25) and (3). In addition, equality (27) holds for $\widetilde{\mu}_{t}$ by (25). Put

$$
V(T, K)=\int_{K}^{\infty} x d \widetilde{\mu}_{T}(x)-K \int_{K}^{\infty} d \widetilde{\mu}_{T}(x) .
$$

Then $G(T, K)=e^{-r T} V(T, K)$. Moreover, the function $K \mapsto V(T, K)$ is convex on $[0, \infty)$ since its second distributional derivative coincides with the measure $\widetilde{\mu}_{T}$. This establishes conditions 1 and 2. The equality $G(0, K)=\left(x_{0}-K\right)_{+}$can be obtained using (25) and (26). This gives condition 4. Next, we see that (26) implies

$$
G(T, K) \leq e^{-r T} \int_{K}^{\infty} x d \widetilde{\mu}_{T}(x)
$$

and hence $\lim _{K \rightarrow \infty} G(T, K)=0$. This establishes condition 5 . In order to prove condition 3 for $G$, we notice that (22) gives the following:

$$
G\left(T, e^{r T} K\right)=\frac{K}{x_{0}} C\left(T, e^{r T} \frac{x_{0}^{2}}{K}\right)+x_{0}-K .
$$

Now it is clear that condition 3 for $G$ follows from the same condition for $C$. Therefore, $G$ is a call pricing function.

This completes the proof of Theorem 3.2.
Remark 3.5. It is not hard to see that if the call pricing function $C$ in Theorem 3.2 satisfies $C \in P F_{\infty}$, then $G \in P F_{0}$. Similarly, if $C \in P F_{0}$, then $G \in P F_{\infty}$. Finally, if $C \in P F_{\infty} \cap P F_{0}$, then $G \in P F_{\infty} \cap P F_{0}$.

We will next discuss a certain symmetry condition satisfied by the implied volatility. A similar condition can be found in section 4 of [29] (see also formula 2.9 in [4]). Let $C$ be
a call pricing function such that $C \in P F_{\infty} \cap P F_{0}$. Then $G \in P F_{\infty} \cap P F_{0}$, and hence the implied volatility $I_{C}$ associated with $C$ and the implied volatility $I_{G}$ associated with $G$ exist for all $T>0$ and $K>0$. Replacing $\sigma$ by $I_{C}(K)$ in (21) and taking into account the equality $C_{B S}\left(T, K, I_{C}(T, K)\right)=C(T, K)$ and (22), we see that

$$
C_{B S}\left(T,\left(x_{0} e^{r t}\right)^{2} K^{-1}, I_{C}(T, K)\right)=G\left(T,\left(x_{0} e^{r T}\right)^{2} K^{-1}\right) .
$$

Therefore, the following lemma holds.
Lemma 3.6. Let $C \in P F_{\infty} \cap P F_{0}$, and let $G$ be defined by (26). Then

$$
\begin{equation*}
I_{C}(T, K)=I_{G}\left(T,\left(x_{0} e^{r T}\right)^{2} K^{-1}\right) \tag{28}
\end{equation*}
$$

for all $T>0$ and $K>0$.
Formula (28) can be interpreted as a symmetry condition for the implied volatility. For a certain class of uncorrelated stochastic volatility models, a similar condition was established in [31] (see also [19]). For the models from this class, we have $I(T, K)=I\left(T,\left(x_{0} e^{r T}\right)^{2} K^{-1}\right)$.

Lemma 3.6 and the results obtained in section 2 can be used to find sharp asymptotic formulas for the implied volatility as $K \rightarrow 0$.

Theorem 3.7. Let $C \in P F_{0}$, and let $P$ be the corresponding put pricing function. Suppose that $\tau$ is a positive function with $\lim _{K \rightarrow 0} \tau(K)=\infty$. Suppose also that

$$
\begin{equation*}
P(K) \approx \widetilde{P}(K) \quad \text { as } \quad K \rightarrow 0 \tag{29}
\end{equation*}
$$

where $\widetilde{P}$ is a positive function. Then the following asymptotic formula holds:

$$
\begin{aligned}
I(K)= & \frac{\sqrt{2}}{\sqrt{T}}\left[\sqrt{\log \frac{1}{\widetilde{P}(K)}-\frac{1}{2} \log \log \frac{K}{\widetilde{P}(K)}}-\sqrt{\log \frac{K}{\widetilde{P}(K)}-\frac{1}{2} \log \log \frac{K}{\widetilde{P}(K)}}\right] \\
& +O\left(\left(\log \frac{K}{\widetilde{P}(K)}\right)^{-\frac{1}{2}} \tau(K)\right) .
\end{aligned}
$$

In addition,

$$
I(K)=\frac{\sqrt{2}}{\sqrt{T}}\left[\sqrt{\log \frac{1}{\widetilde{P}(K)}}-\sqrt{\log \frac{K}{\widetilde{P}(K)}}\right]+O\left(\left(\log \frac{K}{\widetilde{P}(K)}\right)^{-\frac{1}{2}} \log \log \frac{K}{\widetilde{P}(K)}\right)
$$

as $K \rightarrow 0$.
An important special case of Theorem 3.7 is as follows.
Theorem 3.8. Let $C \in P F_{0}$, and let $P$ be the corresponding put pricing function. Then

$$
\begin{equation*}
I(K)=\frac{\sqrt{2}}{\sqrt{T}}\left[\sqrt{\log \frac{1}{P(K)}}-\sqrt{\log \frac{K}{P(K)}}\right]+O\left(\left(\log \frac{K}{P(K)}\right)^{-\frac{1}{2}} \log \log \frac{K}{P(K)}\right) \tag{30}
\end{equation*}
$$

as $K \rightarrow 0$.

Proof of Theorem 3.7. Formulas (29) and (23) imply that $G(K) \approx \widetilde{G}(K)$ as $K \rightarrow \infty$, where

$$
\begin{equation*}
\widetilde{G}(K)=K \widetilde{P}\left(\left(x_{0} e^{r T}\right)^{2} K^{-1}\right) . \tag{31}
\end{equation*}
$$

Put $\psi(K)=\psi\left(\left(x_{0} e^{r T}\right)^{2} K^{-1}\right)$. Then, applying Corollary 2.5 to $G$ and $\widetilde{G}$, we get

$$
\begin{align*}
I_{G}(K)= & \frac{\sqrt{2}}{\sqrt{T}}\left[\sqrt{\log K+\log \frac{1}{\widetilde{G}(K)}-\frac{1}{2} \log \log \frac{1}{\widetilde{G}(K)}}-\sqrt{\log \frac{1}{\widetilde{G}(K)}-\frac{1}{2} \log \log \frac{1}{\widetilde{G}(K)}}\right] \\
\text { 2) } & +O\left(\left(\log \frac{1}{\widetilde{G}(K)}\right)^{-\frac{1}{2}} \psi(K)\right) \tag{32}
\end{align*}
$$

as $K \rightarrow \infty$. It follows from (28), (31), (32), and the mean value theorem that

$$
\begin{aligned}
& \frac{\sqrt{T}}{\sqrt{2}} I(K)=\sqrt{\log \frac{\left(x_{0} e^{r T}\right)^{2}}{K}+\log \frac{K}{\left(x_{0} e^{r T}\right)^{2} \widetilde{P}(K)}-\frac{1}{2} \log \log \frac{K}{\left(x_{0} e^{r T}\right)^{2} \widetilde{P}(K)}} \\
& -\sqrt{\log \frac{K}{\left(x_{0} e^{r T}\right)^{2} \widetilde{P}(K)}-\frac{1}{2} \log \log \frac{K}{\left(x_{0} e^{r T}\right)^{2} \widetilde{P}(K)}}+O\left(\left(\log \frac{K}{\widetilde{P}(K)}\right)^{-\frac{1}{2}} \tau(K)\right) \\
& =\sqrt{\log \frac{1}{\widetilde{P}(K)}-\frac{1}{2} \log \log \frac{K}{\widetilde{P}(K)}}-\sqrt{\log \frac{K}{\widetilde{P}(K)}-\frac{1}{2} \log \log \frac{K}{\widetilde{P}(K)}}+O\left(\left(\log \frac{K}{\widetilde{P}(K)}\right)^{-\frac{1}{2}} \tau(K)\right)
\end{aligned}
$$

as $K \rightarrow 0$. Note that formula (31) implies $K[\widetilde{P}(K)]^{-1} \rightarrow \infty$ as $K \rightarrow 0$.
This completes the proof of Theorem 3.7.
4. Sharp asymptotic formulas for the implied volatility and Lee's moment formulas. It will be assumed in this section that the call pricing function satisfies $C \in P F_{\infty} \cap P F_{0}$. In [29], Lee obtained important asymptotic formulas for the implied volatility. Lee's results are as follows.

Theorem 4.1. Let I be the implied volatility associated with a call pricing function C. Define a number $\tilde{p}$ by

$$
\begin{equation*}
\tilde{p}=\sup \left\{p \geq 0: \mathbb{E}^{*}\left[X_{T}^{1+p}\right]<\infty\right\} \tag{33}
\end{equation*}
$$

Then the following equality holds:

$$
\begin{equation*}
\limsup _{K \rightarrow \infty} \frac{T I(K)^{2}}{\log K}=\psi(\tilde{p}) \tag{34}
\end{equation*}
$$

where the function $\psi$ is given by

$$
\begin{equation*}
\psi(u)=2-4\left(\sqrt{u^{2}+u}-u\right), \quad u \geq 0 \tag{35}
\end{equation*}
$$

Theorem 4.2. Let I be the implied volatility associated with a call pricing function $C$. Define a number $\tilde{q}$ by

$$
\begin{equation*}
\tilde{q}=\sup \left\{q \geq 0: \mathbb{E}\left[X_{T}^{-q}\right]<\infty\right\} . \tag{36}
\end{equation*}
$$

Then the following formula holds:

$$
\begin{equation*}
\limsup _{K \rightarrow 0} \frac{T I(K)^{2}}{\log \frac{1}{K}}=\psi(\tilde{q}) . \tag{37}
\end{equation*}
$$

Formulas (34) and (37) are called Lee's moment formulas, and the numbers $1+\tilde{p}$ and $\tilde{q}$ are called the right-tail index and the left-tail index of the distribution of the stock price $X_{T}$, respectively. These numbers show how fast the tails of the distribution of the stock price decay.

We will next show how to derive Lee's moment formula (34) using our formula (8). In order to see how (8) is linked to Lee's formulas, we note that, for every $a>0$,

$$
\begin{equation*}
\sqrt{1+a}-\sqrt{a}=\left(1-2\left(\sqrt{a^{2}+a}-a\right)\right)^{\frac{1}{2}}=\sqrt{2^{-1} \psi(a)} \tag{38}
\end{equation*}
$$

Therefore, Lee's formulas (34) and (37) can be rewritten as follows:

$$
\begin{equation*}
\limsup _{K \rightarrow \infty} \frac{\sqrt{T} I(K)}{\sqrt{2 \log K}}=\sqrt{1+\tilde{p}}-\sqrt{\tilde{p}} \tag{39}
\end{equation*}
$$

and

$$
\begin{equation*}
\limsup _{K \rightarrow 0} \frac{\sqrt{T} I(K)}{\sqrt{2 \log \frac{1}{K}}}=\sqrt{1+\tilde{q}}-\sqrt{\tilde{q}} \tag{40}
\end{equation*}
$$

Our next goal is to establish formula (39).
Lemma 4.3. Let $C \in P F_{\infty} \cap P F_{0}$, and put

$$
\begin{equation*}
l=\liminf _{K \rightarrow \infty}(\log K)^{-1} \log \frac{1}{C(K)} . \tag{41}
\end{equation*}
$$

Then

$$
\begin{equation*}
\limsup _{K \rightarrow \infty} \frac{\sqrt{T} I(K)}{\sqrt{2 \log K}}=\sqrt{1+l}-\sqrt{l} . \tag{42}
\end{equation*}
$$

Proof. Observe that (8) implies

$$
\begin{align*}
& \frac{\sqrt{T} I(K)}{\sqrt{2 \log K}}=\sqrt{1+\frac{\log \frac{1}{C(K)}}{\log K}}-\sqrt{\frac{\log \frac{1}{C(K)}}{\log K}}+O\left((\log K)^{-\frac{1}{2}}\left(\log \frac{1}{C(K)}\right)^{-\frac{1}{2}} \log \log \frac{1}{C(K)}\right) \\
& (43) \quad=\left[\sqrt{1+\frac{\log \frac{1}{C(K)}}{\log K}}+\sqrt{\frac{\log \frac{1}{C(K)}}{\log K}}\right]^{-1}+O\left((\log K)^{-\frac{1}{2}}\left(\log \frac{1}{C(K)}\right)^{-\frac{1}{2}} \log \log \frac{1}{C(K)}\right) \tag{43}
\end{align*}
$$

as $K \rightarrow \infty$. It is clear that (42) follows from (43).
Let us continue the proof of formula (39). Denote by $\rho_{T}$ the complementary distribution function of $X_{T}$ given by $\rho_{T}(y)=\mathbb{P}\left[X_{T}>y\right], y>0$. Then

$$
\begin{equation*}
C(K)=e^{-r T} \int_{K}^{\infty} \rho_{T}(y) d y, \quad K>0 . \tag{44}
\end{equation*}
$$

Define the following numbers:

$$
\begin{equation*}
r^{*}=\sup \left\{r \geq 0: C(K)=O\left(K^{-r}\right) \quad \text { as } \quad K \rightarrow \infty\right\} \tag{45}
\end{equation*}
$$

and

$$
\begin{equation*}
s^{*}=\sup \left\{s \geq 0: \rho_{T}(y)=O\left(y^{-(1+s)}\right) \quad \text { as } \quad y \rightarrow \infty\right\} . \tag{46}
\end{equation*}
$$

Lemma 4.4. The numbers $\tilde{p}, l, r^{*}$, and $s^{*}$ given by (33), (41), (45), and (46), respectively, are all equal.

Proof. If $s^{*}=0$, then the inequality $s^{*} \leq r^{*}$ is trivial. If $s>0$ is such that $\rho_{T}(y)=$ $O\left(y^{-(1+s)}\right)$ as $y \rightarrow \infty$, then

$$
C(K)=O\left(\int_{K}^{\infty} y^{-(1+s)} d y\right)=O\left(K^{-s}\right)
$$

as $K \rightarrow \infty$. Hence $s^{*} \leq r^{*}$.
Next, let $r \geq 0$ be such that $C(K)=O\left(K^{-r}\right)$ as $K \rightarrow \infty$. Then (44) shows that there exists $c>0$, for which

$$
c K^{-r} \geq e^{-r T} \int_{K}^{\infty} \rho_{T}(y) d y \geq e^{-r T} \int_{K}^{2 K} \rho_{T}(y) d y \geq e^{-r T} \rho_{T}(2 K) K, \quad K>K_{0} .
$$

Therefore, $\rho_{T}(K)=O\left(K^{-(r+1)}\right)$ as $K \rightarrow \infty$. It follows that $r^{*} \leq s^{*}$. This proves the equality $r^{*}=s^{*}$.

Suppose that $0<l<\infty$. Then, for every $\varepsilon>0$, there exists $K_{\varepsilon}>0$ such that

$$
\log \frac{1}{C(K)} \geq(l-\varepsilon) \log K, \quad K>K_{\varepsilon} .
$$

Therefore, $C(K) \leq K^{-l+\varepsilon}, K>K_{\varepsilon}$. It follows that $l-\varepsilon \leq r^{*}$ for all $\varepsilon>0$, and hence $l \leq r^{*}$. The inequality $l \leq r^{*}$ also holds if $l=0$ or $l=\infty$. This fact can be established similarly.

To prove the inequality $r^{*} \leq l$, suppose that $r^{*} \neq 0$ and $r<r^{*}$. Then $C(K)=O\left(K^{-r}\right)$ as $K \rightarrow \infty$, and hence $\frac{1}{C(K)} \geq c K^{r}$ for some $c>0$ and all $K>K_{0}$. It follows that $\log \frac{1}{C(K)} \geq \log c+r \log K, K>K_{0}$, and

$$
\frac{\log \frac{1}{C(K)}}{\log K} \geq \frac{\log c}{\log K}+r
$$

Now it is clear that

$$
\begin{equation*}
\liminf _{K \rightarrow \infty} \frac{\log \frac{1}{C(K)}}{\log K} \geq r \tag{47}
\end{equation*}
$$

Using (47), we see that $l \geq r^{*}$. If $r^{*}=0$, then the inequality $l \geq r^{*}$ is trivial. This proves that $l=r^{*}=s^{*}$.

It is clear that, for all $p \geq 0$,

$$
\begin{equation*}
\mathbb{E}\left[X_{T}^{1+p}\right]=(1+p) \int_{0}^{\infty} y^{p} \rho_{T}(y) d y \tag{48}
\end{equation*}
$$

Suppose that $s^{*}=0$; then the inequality $s^{*} \leq \tilde{p}$ is trivial. If, for some $s>0, \rho_{T}(y)=$ $O\left(y^{-(1+s)}\right)$ as $y \rightarrow \infty$, then it is not hard to see using (48) that $\mathbb{E}\left[X_{T}^{1+p}\right]<\infty$ for all $p<s$. It follows that $s^{*} \leq \tilde{p}$.

On the other hand, if $\mathbb{E}\left[X_{T}^{1+p}\right]<\infty$ for some $p \geq 0$, then there exists a number $M>0$ such that

$$
\begin{equation*}
M>\int_{K}^{\infty} y^{p} \rho_{T}(y) d y \geq K^{p} \int_{K}^{\infty} \rho_{T}(y) d y=e^{r T} K^{p} C(K) \tag{49}
\end{equation*}
$$

In the proof of (49), we used (48) and (44). It follows from (49) that $C(K)=O\left(K^{-p}\right)$ as $K \rightarrow \infty$, and hence $\tilde{p} \leq r^{*}$.

This completes the proof of Lemma 4.4.
To finish the proof of formula (34), we observe that (42) and the equality $l=\tilde{p}$ in Lemma 4.4 imply formula (39).

It will be explained next how to obtain formula (40) from formula (30). Taking into account (30), we get the following lemma.

Lemma 4.5. Let $C \in P F_{\infty} \cap P F_{0}$, and define a number by

$$
\begin{equation*}
m=\liminf _{K \rightarrow 0}\left(\log \frac{1}{K}\right)^{-1} \log \frac{1}{P(K)} \tag{50}
\end{equation*}
$$

Then

$$
\begin{equation*}
\limsup _{K \rightarrow 0} \frac{\sqrt{T} I(K)}{\sqrt{2 \log \frac{1}{K}}}=\sqrt{m}-\sqrt{m-1} \tag{51}
\end{equation*}
$$

It is not hard to see that $m \geq 1$, where $m$ is defined by (50). Put

$$
\eta_{T}(y)=\mathbb{P}\left[X_{T} \leq y\right]=1-\rho_{T}(y), \quad y \geq 0 .
$$

Then

$$
P(K)=e^{-r T} \int_{0}^{K} \eta_{T}(y) d y
$$

and

$$
\mathbb{E}\left[X_{T}^{-q}\right]=q \int_{0}^{\infty} y^{-q-1} \eta_{T}(y) d y
$$

for all $q>0$. Note that $\eta_{T}(0)=\mathbb{P}\left[X_{T}=0\right]$.
Consider the following numbers:

$$
\begin{equation*}
u^{*}=\sup \left\{u \geq 1: P(K)=O\left(K^{u}\right) \quad \text { as } \quad K \rightarrow 0\right\} \tag{52}
\end{equation*}
$$

and

$$
\begin{equation*}
v^{*}=\sup \left\{v \geq 0: \eta_{T}(y)=O\left(y^{v}\right) \quad \text { as } \quad y \rightarrow 0\right\} . \tag{53}
\end{equation*}
$$

It is not hard to see, using the same ideas as in the proof of Lemma 4.4, that the following lemma holds.

Lemma 4.6. The numbers $\tilde{q}, m, u^{*}$, and $v^{*}$ defined by (36), (50), (52), and (53), respectively, satisfy the condition $\tilde{q}+1=m=u^{*}=v^{*}+1$.

Now it is clear that formula (40) follows from (51) and Lemma 4.6.
5. Sharp asymptotic formulas for the implied volatility and the tail-wing formulas of Benaim and Friz. In [2], Benaim and Friz studied asymptotic relations between a given call pricing function, the implied volatility associated with it, and the law of the stock returns, under an additional assumption that there exist nontrivial moments of the stock price. We will next give several definitions from the theory of regularly varying functions (these definitions will be needed in the remaining part of the paper) and then formulate some of the results obtained in [2].

Definition 5.1. Let $\alpha \in \mathbb{R}$, and let $f$ be a positive Lebesgue measurable function defined on some neighborhood of infinity. The function $f$ is called regularly varying with index $\alpha$ if the following condition holds: For every $\lambda>0, \frac{f(\lambda x)}{f(x)} \rightarrow \lambda^{\alpha}$ as $x \rightarrow \infty$. The class consisting of all regularly varying functions with index $\alpha$ is denoted by $R_{\alpha}$. Functions belonging to the class $R_{0}$ are called slowly varying.

The following asymptotic formula is valid for all functions $f \in R_{\alpha}$ with $\alpha>0$ :

$$
\begin{equation*}
-\log \int_{K}^{\infty} e^{-f(y)} d y \sim f(K) \quad \text { as } \quad K \rightarrow \infty \tag{54}
\end{equation*}
$$

(see Theorem 4.12.10(i) in [5]). This result is known as Bingham's lemma.
Definition 5.2. Let $\alpha \in \mathbb{R}$, and let $f$ be a positive function defined on some neighborhood of infinity. The function $f$ is called smoothly varying with index $\alpha$ if the function $h(x)=$ $\log f\left(e^{x}\right)$ is infinitely differentiable and $h^{\prime}(x) \rightarrow \alpha, h^{(n)}(x) \rightarrow 0$ for all integers $n \geq 2$ as $x \rightarrow \infty$.

An equivalent definition of the class $S R_{\alpha}$ is as follows:

$$
\begin{equation*}
f \in S R_{\alpha} \Leftrightarrow \lim _{x \rightarrow \infty} \frac{x^{n} f^{(n)}(x)}{f(x)}=\alpha(\alpha-1) \ldots(\alpha-n+1) \tag{55}
\end{equation*}
$$

for all $n \geq 1$.
Definition 5.3. Let $g$ be a function on $(0, \infty)$ such that $g(x) \downarrow 0$ as $g \rightarrow \infty$. A function $l$ defined on $(a, \infty), a \geq 0$, is called slowly varying with remainder $g$ if $l \in R_{0}$ and $\frac{l(\lambda x)}{l(x)}-1=O(g(x))$ as $x \rightarrow \infty$ for all $\lambda>1$.

Definitions 5.1-5.3 can be found in [5]. The theory of regularly varying functions has interesting applications in financial mathematics. Besides [2, 3, 4], such functions appear in the study of Pareto-type tails of distributions of stock returns (see, e.g, [10] and the references therein). We refer the reader to [1] for more information on Pareto-type distributions and their applications. Pareto-type distributions are defined as follows. Let $X$ be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $F$ be the distribution function of $X$ given by $F(y)=\mathbb{P}[X \leq y], y \in \mathbb{R}$. By $\bar{F}$ we denote the complementary distribution function of $X$, that is, the function $\bar{F}(y)=1-F(y), y \in \mathbb{R}$. The distribution $F$ is called a Pareto-type distribution with index $\alpha>0$ if and only if there exists a function $g \in R_{-\alpha}$ such that $\bar{F}(y) \sim g(y)$ as $y \rightarrow \infty$. More precisely, the Pareto-type condition means that there exists a function $h \in R_{0}$ such that $\bar{F}(y) \sim y^{-\alpha} h(y)$ as $y \rightarrow \infty$. Note that the Pareto-type index of $F$ is the negative of the index of regular variation of the function $g$.

We will next formulate some of the results obtained in [2] adapting them to our notation (see Theorem 1 in [2]). Benaim and Friz use a different normalization in the Black-Scholes formula and consider the normalized implied volatility as a function of the log-strike $k$. In the formulation of Theorem 5.4, the function $\psi$ is defined by $\psi(u)=2-4\left(\sqrt{u^{2}+u}-u\right)$. This function has already appeared in section 4. It is clear that $\psi$ is strictly decreasing on the interval $[0, \infty]$ and maps this interval onto the interval $[0,2]$. Note that the conditions in Theorem 5.4 imply $C \in P F_{\infty}$.

Theorem 5.4. Let $C$ be a call pricing function, and suppose that

$$
\begin{equation*}
\mathbb{E}^{*}\left[X_{T}^{1+\varepsilon}\right]<\infty \quad \text { for some } \quad \varepsilon>0 . \tag{56}
\end{equation*}
$$

Then the following statements hold:

1. If $C(K)=\exp \{-\eta(\log K)\}$ with $\eta \in R_{\alpha}, \alpha>0$, then

$$
\begin{equation*}
I(K) \sim \frac{\sqrt{\log K}}{\sqrt{T}} \sqrt{\psi\left(-\frac{\log C(K)}{\log K}\right)} \quad \text { as } \quad K \rightarrow \infty \tag{57}
\end{equation*}
$$

2. If $\bar{F}(y)=\exp \{-\rho(\log y)\}$ with $\rho \in R_{\alpha}, \alpha>0$, then

$$
\begin{equation*}
I(K) \sim \frac{\sqrt{\log K}}{\sqrt{T}} \sqrt{\psi\left(-\frac{\log [K \bar{F}(K)]}{\log K}\right)} \quad \text { as } \quad K \rightarrow \infty \tag{58}
\end{equation*}
$$

3. If the distribution $\mu_{T}$ of the stock price $X_{T}$ admits a density $D_{T}$ and if

$$
D_{T}(x)=\frac{1}{x} \exp \{-h(\log x)\}
$$

as $x \rightarrow \infty$, where $h \in R_{\alpha}, \alpha>0$, then

$$
\begin{equation*}
I(K) \sim \frac{\sqrt{\log K}}{\sqrt{T}} \sqrt{\psi\left(-\frac{\log \left[K^{2} D_{T}(K)\right]}{\log K}\right)} \quad \text { as } \quad K \rightarrow \infty \tag{59}
\end{equation*}
$$

The functions $V(k)$ and $c(k)$ in [2] correspond in our notation to the functions $\sqrt{T} I(K)$ and $e^{r T} C(K)$, respectively. We also take into account that the distribution density $f$ of the stock return at time $T$ is related to the density $D_{T}$ by the formula $f(y)=e^{y} D_{T}\left(e^{y}\right)$.

The formulas contained in Theorem 5.4 are called the right-tail-wing formulas. The idea to express the asymptotic properties of the implied volatility in terms of the behavior of the distribution density of the stock price has also been exploited in [22] and [21] in the case of special stock price models with stochastic volatility.

Our next goal is to derive Theorem 5.4 from Corollary 2.9. The next statement is nothing else but this corollary in disguise.

Corollary 5.5. Let $C \in P F_{\infty}$. Then

$$
\begin{equation*}
I(K)=\frac{\sqrt{\log K}}{\sqrt{T}} \sqrt{\psi\left(-\frac{\log C(K)}{\log K}\right)}+O\left(\left(\log \frac{1}{C(K)}\right)^{-\frac{1}{2}} \log \log \frac{1}{C(K)}\right) \tag{60}
\end{equation*}
$$

as $K \rightarrow \infty$, where $\psi$ is defined by (35).
The equivalence of formulas (8) and (60) can be easily shown using (38).
Remark 5.6. It follows from Corollary 5.5 that formula (57) holds for any call pricing function from the class $P F_{\infty}$, and hence no restrictions are needed in part 1 of Theorem 5.4. Moreover, formula (60) contains an error term, which is absent in formula (57).

We will next briefly explain how to obtain (58) and (59). We will prove a slightly more general statement, assuming that

$$
\begin{equation*}
\bar{F}(y) \approx \exp \{-\rho(\log y)\} \tag{61}
\end{equation*}
$$

as $y \rightarrow \infty$ in part 2 of Theorem 5.4 and

$$
\begin{equation*}
D_{T}(x) \approx x^{-1} \exp \{-h(\log x)\} \tag{62}
\end{equation*}
$$

as $x \rightarrow \infty$ in part 3. Some of the ideas used in the proof below are borrowed from [2] (see the proofs in section 3 of [2]). With no loss of generality, we may suppose that $\alpha \geq 1$. The proof of the tail-wing formulas is based on Bingham's lemma and the following equalities:

$$
C(K)=e^{-r T} \int_{K}^{\infty} \bar{F}(y) d y
$$

and

$$
\begin{equation*}
\bar{F}(y)=\int_{y}^{\infty} D_{T}(x) d x \tag{63}
\end{equation*}
$$

If $\alpha>1$ in parts 2 or 3 of Theorem 5.4, then the moment condition (56) holds, and we have $\rho(u)-u \in R_{\alpha}$ in part 2 and $h(u)-u \in R_{\alpha}$ in part 3. If $\alpha=1$, then the moment condition gives $\rho(u)-u \in R_{1}$ in part 2 and $h(u)-u \in R_{1}$ in part 3 (see section 3 in [2]).

Suppose that (61) holds. Put $\lambda(u)=\rho(u)-u$. Then we have $C(K) \approx \widehat{C}(K)$ as $K \rightarrow \infty$, where

$$
\widehat{C}(K)=\int_{\log K}^{\infty} \exp \{-\lambda(u)\} d u
$$

Applying formula (54) to the function $\lambda$, we obtain

$$
\log \frac{1}{\widehat{C}(K)} \sim \lambda(\log K)=\log \frac{1}{K \bar{F}(K)}
$$

as $K \rightarrow \infty$. Since $C(K) \approx \widehat{C}(K)$, we also have

$$
\log \frac{1}{C(K)} \sim \log \frac{1}{\widehat{C}(K)},
$$

and hence

$$
\log \frac{1}{C(K)} \sim \log \frac{1}{K \bar{F}(K)}
$$

as $K \rightarrow \infty$. Now it clear that formula (58) follows from (18) and (38).
Next, assume that equality (62) holds. Then (63) implies (61) with

$$
\rho(y)=-\log \int_{y}^{\infty} e^{-h(u)} d u .
$$

Applying Bingham's lemma, we see that $\rho \in R_{\alpha}$. This reduces the case of the distribution density $D_{T}$ of the stock price in Theorem 5.4 to that of the complementary distribution function $\bar{F}$.

Remark 5.7. The tail-wing formula (58) also holds, provided that $\alpha=1$ and $\rho(u)-u \in R_{\beta}$ with $0<\beta \leq 1$. A similar statement is true in the case of formula (59). The proof of these assertions does not differ from the proof given above. Interesting examples here are $\rho(u)=u+u^{\beta}$ if $\beta<1$ and $\rho(u)=u+\frac{u}{\log u}$ if $\beta=1$. Note that the moment condition does not hold in these cases.

Formulas (58) and (59) do not include error estimates. Our next goal is to find asymptotic formulas for the implied volatility which contain error estimates. We can do it under certain smoothness assumptions on the functions $\rho$ and $h$ appearing in Theorem 5.4. Note that the conditions in Theorems 5.8 and 5.9 imply $C \in P F_{\infty}$.

Theorem 5.8. Let $\bar{F}$ be the complementary distribution function of the stock price $X_{T}$. Suppose that

$$
\begin{equation*}
\bar{F}(y) \approx \exp \{-\rho(\log y)\} \tag{64}
\end{equation*}
$$

as $y \rightarrow \infty$, where $\rho$ is a function such that either $\rho \in S R_{\alpha}$ with $\alpha>1$, or $\rho \in S R_{1}$ and $\lambda(u)=\rho(u)-u \in R_{\beta}$ for some $0<\beta \leq 1$. Then

$$
\begin{equation*}
I(K)=\frac{\sqrt{2}}{\sqrt{T}}(\sqrt{\rho(\log K)}-\sqrt{\rho(\log K)-\log K})+O\left(\frac{\log [\rho(\log K)]}{\sqrt{\rho(\log K)}}\right) \tag{65}
\end{equation*}
$$

as $K \rightarrow \infty$.
Theorem 5.9. Let $D_{T}$ be the distribution density of the stock price $X_{T}$. Suppose that

$$
\begin{equation*}
D_{T}(x) \approx \frac{1}{x} \exp \{-h(\log x)\} \tag{66}
\end{equation*}
$$

as $x \rightarrow \infty$, where $h$ is a function such that either $h \in S R_{\alpha}$ with $\alpha>1$, or $h \in S R_{1}$ and $g(u)=h(u)-u \in S R_{\beta}$ for some $0<\beta \leq 1$. Then

$$
\begin{equation*}
I(K)=\frac{\sqrt{2}}{\sqrt{T}}(\sqrt{h(\log K)}-\sqrt{h(\log K)-\log K})+O\left(\frac{\log [h(\log K)]}{\sqrt{h(\log K)}}\right) \tag{67}
\end{equation*}
$$

as $K \rightarrow \infty$.
Remark 5.10. Formulas (65) and (67) are equivalent to the formulas

$$
I(K)=\frac{\sqrt{\log K}}{\sqrt{T}} \sqrt{\psi\left(\frac{\rho(\log K)-\log K}{\log K}\right)}+O\left(\frac{\log [\rho(\log K)]}{\sqrt{(\rho(\log K))}}\right), \quad K \rightarrow \infty
$$

and

$$
I(K)=\frac{\sqrt{\log K}}{\sqrt{T}} \sqrt{\psi\left(\frac{h(\log K)-\log K}{\log K}\right)}+O\left(\frac{\log [h(\log K)]}{\sqrt{(h(\log K))}}\right), \quad K \rightarrow \infty
$$

respectively, where the function $\psi$ is defined by (35). If equality holds in (64) and (66), then we get the following tail-wing formulas with error estimates:

$$
I(K)=\frac{\sqrt{\log K}}{\sqrt{T}} \sqrt{\psi\left(-\frac{\log [K \bar{F}(K)]}{\log K}\right)}+O\left(\left(\log \frac{1}{[K \bar{F}(K)]}\right)^{-\frac{1}{2}} \log \log \frac{1}{[K \bar{F}(K)]}\right)
$$

and

$$
I(K)=\frac{\sqrt{\log K}}{\sqrt{T}} \sqrt{\psi\left(-\frac{\log \left[K^{2} D_{T}(K)\right]}{\log K}\right)}+O\left(\left(\log \frac{1}{\left[K^{2} D_{T}(K)\right]}\right)^{-\frac{1}{2}} \log \log \frac{1}{\left[K^{2} D_{T}(K)\right]}\right)
$$

as $K \rightarrow \infty$.
We will next prove Theorem 5.9. The proof of Theorem 5.8 is similar but less complicated. We leave it as an exercise for the reader.

Proof of Theorem 5.9. We borrow several ideas used in the proof of formula (54) (see Theorem 4.12.10(i) in [5]). The following lemma is standard.

Lemma 5.11. Suppose $r \in S R_{\alpha}$ with $\alpha>0$. Then

$$
\int_{x}^{\infty} e^{-r(u)} d u=\frac{e^{-r(x)}}{r^{\prime}(x)}\left(1+O\left(\frac{1}{r(x)}\right)\right)
$$

as $x \rightarrow \infty$.
Proof. Using the integration by parts formula, we see that

$$
\begin{equation*}
\int_{x}^{\infty} e^{-r(u)} d u=\frac{e^{-r(x)}}{r^{\prime}(x)}-\int_{x}^{\infty} e^{-r(u)} \rho_{1}(u) d u \tag{68}
\end{equation*}
$$

where

$$
\rho_{1}(u)=\left(\frac{1}{r^{\prime}(u)}\right)^{\prime}=\frac{r^{\prime \prime}(u)}{r^{\prime}(u)^{2}}
$$

It follows from (55) that $\left|\rho_{1}(u)\right|=O\left(r(u)^{-1}\right)$ as $u \rightarrow \infty$. Using (55) again, we obtain

$$
\begin{align*}
\int_{x}^{\infty} e^{-r(u)} \rho_{1}(u) d u & =O\left(\int_{x}^{\infty} r^{\prime}(u) e^{-r(u)} \frac{1}{r(u) r^{\prime}(u)} d u\right) \\
& =O\left(\int_{x}^{\infty} r^{\prime}(u) e^{-r(u)} \frac{u}{r(u)^{2}} d u\right) \tag{69}
\end{align*}
$$

as $x \rightarrow \infty$. Since, for every $\varepsilon>0$, the function $u \mapsto e^{-\varepsilon r(u)} u r(u)^{-2}$ is ultimately decreasing, (69) implies that

$$
\begin{equation*}
\int_{x}^{\infty} e^{-r(u)} \rho_{1}(u) d u=O\left(e^{-r(x)} \frac{x}{r(x)^{2}}\right) \tag{70}
\end{equation*}
$$

as $x \rightarrow \infty$. Now Lemma 5.11 follows from (68), (70), and (55).
Let us continue the proof of Theorem 5.9. For $h \in S R_{\alpha}$ with $\alpha>1$, we have $g \in S R_{\alpha}$. On the other hand, if $\alpha=1$, we assume that $g \in S R_{\beta}$ with $0<\beta \leq 1$. Consider the following functions:

$$
\widetilde{D}_{T}(x)=\frac{1}{x} \exp \{-h(\log x)\} \quad \text { and } \quad \widehat{C}(K)=K^{2} \widetilde{D}_{T}(K)=\exp \{-g(\log K)\}
$$

We have

$$
\begin{equation*}
C(K) \approx \int_{\log K}^{\infty} e^{-g(u)} d u-K \int_{\log K}^{\infty} e^{-h(u)} d u \tag{71}
\end{equation*}
$$

as $K \rightarrow \infty$. Now, applying Lemma 5.11, we get

$$
\int_{\log K}^{\infty} e^{-g(u)} d u=\frac{K e^{-h(\log K)}}{g^{\prime}(\log K)}\left(1+O\left(\frac{1}{g(\log K)}\right)\right)
$$

and

$$
K \int_{\log K}^{\infty} e^{-h(u)} d u=\frac{K e^{-h(\log K)}}{h^{\prime}(\log K)}\left(1+O\left(\frac{1}{h(\log K)}\right)\right)
$$

as $K \rightarrow \infty$. It follows that

$$
\begin{equation*}
\int_{\log K}^{\infty} e^{-g(u)} d u-K \int_{\log K}^{\infty} e^{-h(u)} d u=\frac{K e^{-h(\log K)}}{h^{\prime}(\log K) g^{\prime}(\log K)}\left(1+O\left(\frac{h(\log K)}{g(\log K) \log K}\right)\right) \tag{72}
\end{equation*}
$$

as $K \rightarrow \infty$. In the proof of (72), we used (55). Next, employing (71), (72), and (55), we obtain

$$
C(K) \approx \widetilde{C}(K), \quad \text { where } \quad \widetilde{C}(K)=\frac{K(\log K)^{2} e^{-h(\log K)}}{h(\log K) g(\log K)}
$$

Next, we see that

$$
\begin{equation*}
\log \frac{1}{\widetilde{C}(K)}-\log \frac{1}{\widehat{C}(K)}=\log \frac{h(\log K) g(\log K)}{(\log K)^{2}} \tag{73}
\end{equation*}
$$

It follows from (73) that there exists $a>0$ such that

$$
\begin{equation*}
\left|\log \frac{1}{\widetilde{C}(K)}-\log \frac{1}{\widehat{C}(K)}\right| \leq a \log \log \frac{1}{\widehat{C}(K)}, \quad K>K_{1} \tag{74}
\end{equation*}
$$

Indeed, if $\alpha>1$, we can take $a>\frac{2 \alpha-2}{\alpha}$ in (74), and if $\alpha=1$ and $0<\beta \leq 1$, we take $a>\frac{1-\beta}{\beta}$. It is not hard to see that an estimate similar to (74) is valid with $C$ instead of $\widetilde{C}$. Now it follows from Corollary 2.7 that formula (67) holds.

The proof of Theorem 5.9 is thus completed.
Similar results can be obtained in the case of the left-tail-wing formulas established in [2]. We will formulate only the following assertion, which is equivalent to Theorem 3.8.

Corollary 5.12. Let $C \in P F_{0}$, and let $P$ be the corresponding put pricing function. Then

$$
I(K)=\frac{\log \frac{1}{K}}{\sqrt{T}} \sqrt{\psi\left(\frac{\log P(K)}{\log K}-1\right)}+O\left(\left(\log \frac{K}{P(K)}\right)^{-\frac{1}{2}} \log \log \frac{K}{P(K)}\right)
$$

as $K \rightarrow 0$, where $\psi(u)=2-4\left(\sqrt{u^{2}+u}-u\right), u \geq 0$.
The equivalence of Theorem 3.8 and Corollary 5.12 can be shown using (38) with $a=$ $(\log K)^{-1} \log P(K)-1$.

We will next apply Theorem 5.9 to the Black-Scholes model (a similar computation was felicitously called a sanity check in [2]). It is clear that $I(K)=\sigma$ for all $K>0$. Let us see what can be obtained for the Black-Scholes model by using formula (67). We have to at least make sure that $\lim _{K \rightarrow \infty} I(K)=\sigma$.

The distribution density of the stock price in the Black-Scholes model is given by

$$
D_{T}(x)=\frac{\sqrt{x_{0} e^{r T}}}{\sqrt{2 \pi T} \sigma} \exp \left\{-\frac{\sigma^{2} t}{8}\right\} x^{-\frac{3}{2}} \exp \left\{-\left(\log \frac{x}{x_{0} e^{r T}}\right)^{2}\left(2 T \sigma^{2}\right)^{-1}\right\} .
$$

This follows from (6). Hence, $D_{T}$ satisfies condition (66) in Theorem 5.9 with

$$
h(u)=\frac{1}{2 T \sigma^{2}}\left(u-\log \left(x_{0} e^{r T}\right)\right)^{2}+\frac{1}{2} u .
$$

It is clear that $h \in S R_{2}$. Applying Theorem 5.9, we get
$I(K)=\frac{1}{\sqrt{T}}\left[\sqrt{\frac{1}{T \sigma^{2}}\left(\log \frac{K}{x_{0} e^{\mu T}}\right)^{2}+\log K}-\sqrt{\frac{1}{T \sigma^{2}}\left(\log \frac{K}{x_{0} e^{\mu T}}\right)^{2}-\log K}\right]+O\left(\frac{\log \log K}{\log K}\right)$
as $K \rightarrow \infty$. It is not hard to see that the right-hand side of the previous asymptotic formula tends to $\sigma$ as $K \rightarrow \infty$.

## 6. Stock price distribution densities and pricing functions in stochastic volatility mod-

 els.Regularly varying stock price distributions. We will next show that, in some models with stochastic volatility, the stock price $X_{t}$ is Pareto-type distributed for every $t>0$ (the definition of a Pareto-type distribution can be found in section 5). The models we are going to consider are the uncorrelated Stein-Stein, Heston, and Hull-White models. We will first formulate several results obtained in $[20,19,21]$. Recall that the stock price process $X_{t}$ and the volatility process $Y_{t}$ in the Hull-White model satisfy the following system of stochastic differential equations:

$$
\left\{\begin{array}{l}
d X_{t}=r X_{t} d t+Y_{t} X_{t} d W_{t}^{*}  \tag{75}\\
d Y_{t}=\nu Y_{t} d t+\xi Y_{t} d Z_{t}^{*}
\end{array}\right.
$$

In (75), $r \geq 0$ is the interest rate, $\nu \in \mathbb{R}^{1}$, and $\xi>0$. The Hull-White model was introduced in [26]. The volatility process in this model is a geometric Brownian motion.

The Stein-Stein model is defined as follows:

$$
\left\{\begin{array}{l}
d X_{t}=r X_{t} d t+\left|Y_{t}\right| X_{t} d W_{t}^{*}  \tag{76}\\
d Y_{t}=q\left(m-Y_{t}\right) d t+\sigma d Z_{t}^{*}
\end{array}\right.
$$

It was introduced and studied in [33]. In this model, the absolute value of an OrnsteinUhlenbeck process plays the role of the volatility of the stock. We assume that $r \geq 0, q \geq 0$, $m \geq 0$, and $\sigma>0$.

The Heston model was developed in [25]. It is given by

$$
\left\{\begin{array}{l}
d X_{t}=r X_{t} d t+\sqrt{Y_{t}} X_{t} d W_{t}^{*},  \tag{77}\\
d Y_{t}=q\left(m-Y_{t}\right) d t+c \sqrt{Y_{t}} d Z_{t}^{*},
\end{array}\right.
$$

where $r \geq 0, m \geq 0$, and $c \geq 0$. The volatility equation in (77) is uniquely solvable in the strong sense, and the solution $Y_{t}$ is a positive stochastic process. This process is called a Cox-Ingersoll-Ross process. We assume that the processes $W_{t}^{*}$ and $Z_{t}^{*}$ in (75), (76), and (77) are independent Brownian motions under a risk-free probability $\mathbb{P}^{*}$. The initial conditions for the processes $X$ and $Y$ are denoted by $x_{0}$ and $y_{0}$, respectively.

We will next formulate sharp asymptotic formulas for the distribution density $D_{t}$ of the stock price $X_{t}$ in special stock price models. These formulas were obtained in [20, 19, 21].

1. Stein-Stein model. The following result was established in [21]: There exist positive constants $B_{1}, B_{2}$, and $B_{3}$ such that

$$
\begin{equation*}
D_{t}\left(x_{0} e^{r t} x\right)=B_{1}(\log x)^{-\frac{1}{2}} e^{B_{2} \sqrt{\log x}} x^{-B_{3}}\left(1+O\left((\log x)^{-\frac{1}{4}}\right)\right) \tag{78}
\end{equation*}
$$

as $x \rightarrow \infty$. The constants in (78) depend on the model parameters. The constant $B_{3}$ satisfies $B_{3}>2$. Explicit formulas for the constants $B_{1}, B_{2}$, and $B_{3}$ can be found in [21]. It follows from (78), the mean value theorem, and the inequality

$$
\begin{equation*}
e^{a}-1 \leq a e^{a}, \quad 0 \leq a \leq 1, \tag{79}
\end{equation*}
$$

that

$$
\begin{equation*}
D_{t}(x)=B_{0}(\log x)^{-\frac{1}{2}} e^{B_{2} \sqrt{\log x}} x^{-B_{3}}\left(1+O\left((\log x)^{-\frac{1}{4}}\right)\right) \tag{80}
\end{equation*}
$$

as $x \rightarrow \infty$, where $B_{0}=B_{1}\left(x_{0} e^{r t}\right)^{B_{3}}$.
2. Heston model. It was shown in [21] that there exist constants $A_{1}>0, A_{2}>0$, and $A_{3}>2$ such that

$$
\begin{equation*}
D_{t}\left(x_{0} e^{r t} x\right)=A_{1}(\log x)^{-\frac{3}{4}+\frac{q m}{c^{2}}} e^{A_{2} \sqrt{\log x}} x^{-A_{3}}\left(1+O\left((\log x)^{-\frac{1}{4}}\right)\right) \tag{81}
\end{equation*}
$$

as $x \rightarrow \infty$. The constants in (81) depend on the model parameters. Explicit expressions for these constants can be found in [21]. It follows from (81), the mean value theorem, and (79) that

$$
\begin{equation*}
D_{t}(x)=A_{0}(\log x)^{-\frac{3}{4}+\frac{q m}{c^{2}}} e^{A_{2} \sqrt{\log x}} x^{-A_{3}}\left(1+O\left((\log x)^{-\frac{1}{4}}\right)\right) \tag{82}
\end{equation*}
$$

as $x \rightarrow \infty$, where $A_{0}=A_{1}\left(x_{0} e^{r t}\right)^{A_{3}} .^{1}$
3. Hull-White model. The following asymptotic formula holds for the distribution density of the stock price in the Hull-White model (see Theorem 4.1 in [19]):

$$
\begin{align*}
& D_{t}\left(x_{0} e^{r t} x\right)=C x^{-2}(\log x)^{\frac{c_{2}-1}{2}}(\log \log x)^{c_{3}}  \tag{83}\\
& \quad \exp \left\{-\frac{1}{2 t \xi^{2}}\left(\log \left[\frac{1}{y_{0}} \sqrt{\frac{2 \log x}{t}}\right]+\frac{1}{2} \log \log \left[\frac{1}{y_{0}} \sqrt{\frac{2 \log x}{t}}\right]\right)^{2}\right\} \\
& \quad\left(1+O\left((\log \log x)^{-\frac{1}{2}}\right)\right)
\end{align*}
$$

as $x \rightarrow \infty$. The constants $C, c_{2}$, and $c_{3}$ have been computed in [19]. It follows from (83), the mean value theorem, and (79) that

$$
\begin{align*}
& D_{t}(x)=C_{0} x^{-2}(\log x)^{\frac{c_{2}-1}{2}}(\log \log x)^{c_{3}}  \tag{84}\\
& \quad \exp \left\{-\frac{1}{2 t \xi^{2}}\left(\log \left[\frac{1}{y_{0}} \sqrt{\frac{2 \log x}{t}}\right]+\frac{1}{2} \log \log \left[\frac{1}{y_{0}} \sqrt{\frac{2 \log x}{t}}\right]\right)^{2}\right\} \\
& \quad\left(1+O\left((\log \log x)^{-\frac{1}{2}}\right)\right)
\end{align*}
$$

as $x \rightarrow \infty$, where $C_{0}=C\left(x_{0} e^{r t}\right)^{2}$.
Equalities (80), (82), and (84) show that stock price distribution densities in the SteinStein, Heston, and Hull-White models are equivalent to regularly varying functions. Indeed, for the Stein-Stein model, we have

$$
\begin{equation*}
D_{t}(x) \sim x^{-\beta_{t}} h_{t}(x), \quad x \rightarrow \infty \tag{85}
\end{equation*}
$$

with

$$
\begin{equation*}
\beta_{t}=B_{3} \quad \text { and } \quad h_{t}(x)=B_{0}(\log x)^{-\frac{1}{2}} e^{B_{2} \sqrt{\log x}} \tag{86}
\end{equation*}
$$

and it is not hard to see that $h_{t} \in R_{0}$. For the Heston model, (85) is valid with

$$
\begin{equation*}
\beta_{t}=A_{3} \quad \text { and } \quad h_{t}(x)=A_{0}(\log x)^{-\frac{3}{4}+\frac{q m}{c^{2}}} e^{A_{2} \sqrt{\log x}} \tag{87}
\end{equation*}
$$

The function $h_{t}$ defined in (87) is a slowly varying function. For the Hull-White model, condition (85) holds with $\beta_{t}=2$ and

$$
\begin{align*}
h_{t}(x)= & C_{0}(\log x)^{\frac{c_{2}-1}{2}}(\log \log x)^{c_{3}}  \tag{88}\\
& \times \exp \left\{-\frac{1}{2 t \xi^{2}}\left(\log \left[\frac{1}{y_{0}} \sqrt{\frac{2 \log x}{t}}\right]+\frac{1}{2} \log \log \left[\frac{1}{y_{0}} \sqrt{\frac{2 \log x}{t}}\right]\right)^{2}\right\} .
\end{align*}
$$

[^58]Here we also have $h_{t} \in R_{0}$. Note that the constants $B_{3}$ and $A_{3}$ in (86) and (87) depend on $t$ (see [21]).

Remark 6.1. It is also true that the functions $h_{t}$ in (86) and (87) are slowly varying with remainder $g(x)=(\log x)^{-\frac{1}{2}}$. To prove this fact, put $h_{a, b}(x)=(\log x)^{a} e^{b \sqrt{\log x}}$, where $b>0$ and $a \in \mathbb{R}$. Then the function $h_{a, b}$ is slowly varying with remainder $g(x)=(\log x)^{-\frac{1}{2}}$. This follows from the following asymptotic formula:

$$
\begin{aligned}
& \left|\frac{h_{a, b}(\lambda x)}{h_{a, b}(x)}-1\right| \\
& =\frac{(\log x+\log \lambda)^{a}[\exp \{b \sqrt{\log x+\log \lambda}-b \sqrt{\log x}\}-1]+(\log x+\log \lambda)^{a}-(\log x)^{a}}{(\log x)^{a}} \\
& \quad+O\left((\log x)^{-\frac{1}{2}}\right), \quad x \rightarrow \infty
\end{aligned}
$$

Remark 6.2. This remark illustrates the smooth variation condition in Theorem 5.9. Here we show that the stock price distribution density $D_{T}$ in the Stein-Stein model or in the Heston model satisfies condition (66). Indeed, for the Stein-Stein model, we can take

$$
h(u)=\left(B_{3}-1\right) u-B_{2} \sqrt{u}+\frac{1}{2} \log u
$$

and using Definition 5.2 prove that $h \in S R_{1}$ and $h(u)-u \in S R_{1}$. Similarly, for the Heston model,

$$
h(u)=\left(A_{3}-1\right) u-A_{2} \sqrt{u}+\left(\frac{3}{4}-\frac{q m}{c^{2}}\right) \log u,
$$

and hence $h \in S R_{1}$ and $h(u)-u \in S R_{1}$. Next, taking into account the previous statements and applying Theorem 5.9, we can obtain tail-wing formulas with error estimates in the SteinStein and the Heston models. However, for the stock price density in the Hull-White model, condition (66) does not hold. This can be shown by using the explicit formula for the function $h$ that can be obtained from (88) and proving that $h \in S R_{1}$ and $h(u)-u \in S R_{0}$. In section 7 , more refined methods will be developed. The tail-wing formulas obtained in section 7 are applicable to all the special stochastic volatility models considered in the present paper.

The next theorem shows that the distribution of the stock price $X_{t}$ in the Stein-Stein, Heston, and Hull-White models is Pareto-type.

Theorem 6.3. Let $\bar{F}_{t}$ be the complementary distribution function of the stock price $X_{t}$ in the Stein-Stein, Heston, or Hull-White models. Then the following formula holds:

$$
\begin{equation*}
\bar{F}_{t}(y) \sim y^{-\alpha_{t}} \tilde{h}_{t}(y) \tag{89}
\end{equation*}
$$

as $y \rightarrow \infty$. For the Stein-Stein model, $\alpha_{t}=B_{3}-1$ and $\tilde{h}_{t}(y)=\frac{1}{B_{3}-1} h_{t}(y)$, where $h_{t}$ is defined in (86). For the Heston model, we have $\alpha_{t}=A_{3}-1$ and $\tilde{h}(y)=\frac{1}{A_{3}-1} h_{t}(y)$, where $h_{t}$ is given by (87). Finally, for the Hull-White model, $\alpha_{t}=1$ and $\tilde{h}_{t}=h_{t}$, where $h_{t}$ is defined by (88).

To prove Theorem 6.3, we integrate equalities (80), (82), and (84) on the interval $[x, \infty)$ and take into account the following theorem due to Karamata: For all $\alpha<-1$ and $l \in R_{0}$,

$$
\begin{equation*}
\frac{x^{\alpha+1} l(x)}{\int_{x}^{\infty} t^{\alpha} l(t) d t} \rightarrow-\alpha-1 \tag{90}
\end{equation*}
$$

as $x \rightarrow \infty$ (see Proposition 1.5.10 in [5]).
It follows from Theorem 6.3 that the Pareto-type index $\alpha_{t}$ of the stock price $X_{t}$ in the Stein-Stein model is equal to $B_{3}-1$. For the Heston model, we have $\alpha_{t}=A_{3}-1$, and, for the Hull-White model, the Pareto-type index satisfies $\alpha_{t}=1$. Theorem 6.3 also describes the corresponding slowly varying functions.
7. Asymptotic behavior of call pricing functions and tail-wing formulas with error estimates. In this section, we find asymptotic formulas with error estimates for call pricing functions and the implied volatility under the assumption that the distribution density of the stock price is asymptotically equivalent to a regularly varying function.

Theorem 7.1. Let $C$ be a call pricing function, and suppose that the distribution of the stock price $X_{T}$ admits a density $D_{T}$. Suppose also that

$$
\begin{equation*}
D_{T}(x)=x^{\beta} h(x)(1+O(\rho(x))) \tag{91}
\end{equation*}
$$

as $x \rightarrow \infty$, where $\beta<-2$, $h$ is a slowly varying function with remainder $g$, and $\rho(x) \downarrow 0$ as $x \rightarrow \infty$. Then

$$
\begin{equation*}
C(K)=e^{-r T} \frac{1}{(\beta+1)(\beta+2)} K^{\beta+2} h(K)[1+O(\rho(K))+O(g(K))] \tag{92}
\end{equation*}
$$

as $K \rightarrow \infty$.
Proof. The following assertion (see Theorem 3.1.1 in [18] or problem 30 on p. 192 in [5]) will be used in the proof.

Theorem 7.2. Let $L$ be a slowly varying function with remainder $g$, and let $v$ be a positive function on $(1, \infty)$ such that

$$
\int_{1}^{\infty} \lambda^{\epsilon} v(\lambda) d \lambda<\infty \quad \text { for some } \quad \epsilon \geq 0
$$

Then

$$
\int_{1}^{\infty} v(\lambda) \frac{L(\lambda x)}{L(x)} d \lambda=\int_{1}^{\infty} v(\lambda) d \lambda+O(g(x))
$$

as $x \rightarrow \infty$.
It follows from (91) and Theorem 7.2 that

$$
\begin{aligned}
C(K) & =e^{-r T} \int_{K}^{\infty}(x-K) D_{T}(x) d x=e^{-r T} \int_{K}^{\infty}(x-K) x^{\beta} h(x) d x(1+O(\rho(K))) \\
& =e^{-r T} K^{\beta+2} h(K) \int_{1}^{\infty}(y-1) y^{\beta} \frac{h(K y)}{h(K)} d y(1+O(\rho(K))) \\
& =e^{-r T} K^{\beta+2} h(K) \int_{1}^{\infty}(y-1) y^{\beta} d y[1+O(\rho(K))+O(g(K))]
\end{aligned}
$$

as $K \rightarrow \infty$. Now it is clear that formula (92) holds, and the proof of Theorem 7.1 is thus completed.

Theorem 7.1 allows us to characterize the asymptotic behavior of the call pricing function $C(K)$ in the Stein-Stein and the Heston models.

Theorem 7.3. (a) The following formula holds for the call pricing function $C(K)$ in the Stein-Stein model:

$$
\begin{equation*}
C(K)=e^{-r T} \frac{B_{0}}{\left(1-B_{3}\right)\left(2-B_{3}\right)}(\log K)^{-\frac{1}{2}} e^{B_{2} \sqrt{\log K}} K^{2-B_{3}}\left(1+O\left((\log K)^{-\frac{1}{4}}\right)\right) \tag{93}
\end{equation*}
$$

as $K \rightarrow \infty$. The constants in (93) are the same as in (80).
(b) The following formula holds for the call pricing function $C(K)$ in the Heston model:

$$
\begin{equation*}
C(K)=e^{-r T} \frac{A_{0}}{\left(1-A_{3}\right)\left(2-A_{3}\right)}(\log K)^{-\frac{3}{4}+\frac{q m}{c^{2}}} e^{A_{2} \sqrt{\log K}} K^{2-A_{3}}\left(1+O\left((\log K)^{-\frac{1}{4}}\right)\right) \tag{94}
\end{equation*}
$$

as $K \rightarrow \infty$. The constants in (94) are the same as in (82).
It is not hard to see that Theorem 7.3 follows from (80), (82), Remark 6.1, and Theorem 7.1.

Next, we turn our attention to the Hull-White model. Note that Theorem 7.3 cannot be applied in this case since for the Hull-White model we have $\beta=-2$. Instead, we will employ the asymptotic formula for fractional integrals established in [19, Theorem 3.7]. A special case of this formula is as follows: Let $b(x)=B(\log x)$ be a positive increasing function on $[c, \infty)$ with $B^{\prime \prime}(x) \approx 1$ as $x \rightarrow \infty$. Then

$$
\begin{equation*}
\int_{K}^{\infty} \exp \{-b(x)\} d x=\frac{\exp \{-b(K)\}}{b^{\prime}(K)}\left(1+O\left((\log K)^{-1}\right)\right) \tag{95}
\end{equation*}
$$

as $K \rightarrow \infty$. Formula (95) will be used in the proof of the following result.
Theorem 7.4. Let $C$ be a call pricing function, and suppose that the distribution of the stock price $X_{T}$ admits a density $D_{T}$. Suppose also that

$$
\begin{equation*}
D_{T}(x)=x^{-2} \exp \{-b(\log x)\}(1+O(\rho(x))) \tag{96}
\end{equation*}
$$

as $x \rightarrow \infty$. Here the function $b$ is positive, is increasing on $[c, \infty)$ for some $c>0$, and is such that the condition $b(x)=B(\log x)$. Moreover, $B^{\prime \prime}(x) \approx 1$ as $x \rightarrow \infty$, and $\rho(x) \downarrow 0$ as $x \rightarrow \infty$. Then

$$
\begin{equation*}
C(K)=e^{-r T} \frac{\exp \{-b(\log K)\} \log K}{B^{\prime}(\log \log K)}\left[1+O\left((\log \log K)^{-1}\right)+O(\rho(K))\right] \tag{97}
\end{equation*}
$$

as $K \rightarrow \infty$.
Proof. We have

$$
\begin{align*}
& C(K)=e^{-r T}\left[\int_{K}^{\infty} x D_{T}(x) d x-K \int_{K}^{\infty} D_{T}(x) d x\right] \\
& =e^{-r T}\left[\int_{K}^{\infty} x^{-1} \exp \{-b(\log x)\} d x-K \int_{K}^{\infty} x^{-2} \exp \{-b(\log x)\} d x\right](1+O(\rho(K)))  \tag{98}\\
& =e^{-r T} \int_{K}^{\infty} x^{-1} \exp \{-b(\log x)\} d x(1+O(\rho(K)))+O(\exp \{-b(\log K)\}) \\
& =e^{-r T} \int_{\log K}^{\infty} \exp \{-b(u)\} d x(1+O(\rho(K)))+O(\exp \{-b(\log K)\}) . \tag{99}
\end{align*}
$$

Using (95), we get
$C(K)=e^{-r T} \frac{\exp \{-b(\log K)\}}{b^{\prime}(\log K)}\left(1+O\left((\log \log K)^{-1}\right)\right)[1+O(\rho(K))]+O(\exp \{-b(\log K)\})$
$=e^{-r T} \frac{\exp \{-b(\log K)\} \log K}{B^{\prime}(\log \log K)}\left[1+O\left((\log \log K)^{-1}\right)+O(\rho(K))\right]+O(\exp \{-b(\log K)\})$.
Since $B^{\prime}(x) \approx x$ as $x \rightarrow \infty,(100)$ implies (97).
The next assertion characterizes the asymptotic behavior of a call pricing function in the Hull-White model.

Theorem 7.5. Let $C$ be a call pricing function in the Hull-White model. Then

$$
\begin{align*}
& C(K)=4 T \xi^{2} C_{0} e^{-r T}(\log K)^{\frac{c_{2}+1}{2}}(\log \log K)^{c_{3}-1}  \tag{101}\\
& \times \exp \left\{-\frac{1}{2 T \xi^{2}}\left(\log \left[\frac{1}{y_{0}} \sqrt{\frac{2 \log K}{T}}\right]+\frac{1}{2} \log \log \left[\frac{1}{y_{0}} \sqrt{\frac{2 \log K}{T}}\right]\right)^{2}\right\}\left(1+O\left((\log \log K)^{-\frac{1}{2}}\right)\right)
\end{align*}
$$

as $K \rightarrow \infty$. The constants in (101) are the same as in formula (80).
Proof. We will employ Theorem 7.4 in the proof. It is not hard to see using (84) that formula (96) holds for the distribution density $D_{T}$ of the stock price in the Hull-White model. Here we choose the functions $b, B$, and $\rho$ as follows:
$b(u)=-\log C_{0}-\frac{c_{2}-1}{2} \log u-c_{3} \log \log u+\frac{1}{2 T \xi^{2}}\left(\log \left[\frac{1}{y_{0}} \sqrt{\frac{2 u}{T}}\right]+\frac{1}{2} \log \log \left[\frac{1}{y_{0}} \sqrt{\frac{2 u}{T}}\right]\right)^{2}$,
$B(u)=-\log C_{0}-\frac{c_{2}-1}{2} u-c_{3} \log u+\frac{1}{2 T \xi^{2}}\left[\log \frac{1}{y_{0}} \sqrt{\frac{2}{T}}+\frac{1}{2} u+\frac{1}{2} \log \left(\log \frac{1}{y_{0}} \sqrt{\frac{2}{T}}+\frac{1}{2} u\right)\right]^{2}$,
and $\rho(x)=(\log \log x)^{-\frac{1}{2}}$. It is clear that $B^{\prime \prime}(u) \approx 1$ and $B^{\prime}(u) \approx u$ as $u \rightarrow \infty$. Moreover, using the mean value theorem, we obtain the following estimate:

$$
\begin{equation*}
\frac{1}{B^{\prime}(\log \log K)}-\frac{4 T \xi^{2}}{\log \log K}=O\left((\log \log K)^{-2}\right) \tag{102}
\end{equation*}
$$

as $K \rightarrow \infty$. Next, taking into account (97) and (102), we see that (101) holds.
This completes the proof of Theorem 7.5.
The following theorem concerns tail-wing formulas with error estimates under the restrictions imposed on the stock price density $D_{T}$ in Theorems 7.1 and 7.4.

Theorem 7.6. Suppose that the distribution of the stock price $X_{T}$ admits a density $D_{T}$ such that

$$
\begin{equation*}
D_{T}(x) \approx x^{\beta} h(x) \tag{103}
\end{equation*}
$$

as $x \rightarrow \infty$, where $\beta<-2$ and $h$ is a slowly varying function. Let $\psi$ be a positive function with $\lim _{K \rightarrow \infty} \psi(K)=\infty$. Then

$$
\begin{aligned}
& I(K) \frac{\sqrt{T}}{\sqrt{2}}=\sqrt{\log K+\log \frac{1}{K^{2} D_{T}(K)}-\frac{1}{2} \log \log \frac{1}{K^{2} D_{T}(K)}} \\
& \quad-\sqrt{\log \frac{1}{K^{2} D_{T}(K)}-\frac{1}{2} \log \log \frac{1}{K^{2} D_{T}(K)}} \\
& \quad+O\left((\log K)^{-\frac{1}{2}} \psi(K)\right) \\
& =\sqrt{\log K+\log \frac{1}{K^{\beta+2} h(K)}-\frac{1}{2} \log \log \frac{1}{K^{\beta+2} h(K)}}-\sqrt{\log \frac{1}{K^{\beta+2} h(K)}-\frac{1}{2} \log \log \frac{1}{K^{\beta+2} h(K)}} \\
& \quad+O\left((\log K)^{-\frac{1}{2}} \psi(K)\right)
\end{aligned}
$$

as $K \rightarrow \infty$.
Proof. Theorem 7.6 can be established as follows. It is not hard to see that (103) and (90) imply the following formula: $\underset{\sim}{C}(K) \approx K^{\beta+2} h(K)$ as $K \rightarrow \infty$. Therefore, $C \in P F_{\infty}$, and we can apply Corollary 2.5 with $\widetilde{C}(K)=K^{2} D_{T}(K)$ or $\widetilde{C}(K)=K^{\beta+2} h(K)$ to finish the proof of Theorem 7.6.

A similar theorem holds for small values of the strike price.
Theorem 7.7. Suppose that the distribution of the stock price $X_{T}$ admits a density $D_{T}$ such that

$$
\begin{equation*}
D_{T}(x) \approx x^{\gamma} h\left(x^{-1}\right) \tag{104}
\end{equation*}
$$

as $x \rightarrow 0$, where $\gamma>-1$ and $h$ is a slowly varying function. Let $\tau$ be a positive function with $\lim _{K \rightarrow 0} \tau(K)=\infty$. Then

$$
\begin{aligned}
& I(K) \frac{\sqrt{T}}{\sqrt{2}}=\sqrt{\log \frac{1}{K^{2} D_{T}(K)}-\frac{1}{2} \log \log \frac{1}{K D_{T}(K)}}-\sqrt{\log \frac{1}{K D_{T}(K)}-\frac{1}{2} \log \log \frac{1}{K D_{T}(K)}} \\
& \quad+O\left(\left(\log \frac{1}{K}\right)^{-\frac{1}{2}} \tau(K)\right) \\
& =\sqrt{\log \frac{1}{K^{\gamma+2} h\left(K^{-1}\right)}-\frac{1}{2} \log \log \frac{1}{K^{\gamma+1} h\left(K^{-1}\right)}} \\
& \quad-\sqrt{\log \frac{1}{K^{\gamma+1} h\left(K^{-1}\right)}-\frac{1}{2} \log \log \frac{1}{K^{\gamma+1} h\left(K^{-1}\right)}} \\
& \quad+O\left(\left(\log \frac{1}{K}\right)^{-\frac{1}{2}} \tau(K)\right)
\end{aligned}
$$

as $K \rightarrow 0$.

Proof. Theorem 7.7 can be derived from Theorem 3.7. Indeed, using (104) and (90), we obtain the following formula: $P(K) \approx K^{\gamma+2} h\left(K^{-1}\right)$ as $K \rightarrow 0$. Next, applying Theorem 3.7 with $\widetilde{P}(K)=K^{2} D_{T}(K)$ or $\widetilde{P}(K)=K^{\gamma+2} h\left(K^{-1}\right)$, we see that Theorem 7.7 holds.

Theorems 7.6 and 7.7 imply sharp asymptotic formulas for the implied volatility in special stochastic volatility models. For instance, the following assertions (they were established in [21]) follow from Theorem 7.6:

1. Let us consider the implied volatility $k \mapsto \hat{I}(k)$ in the uncorrelated Heston model as a function of the $\log$-strike $k=\log \frac{K}{x_{0} e^{r T}}$. Then there exist positive constants $A_{2}$ and $A_{3}$ such that, for every positive function $\psi$ with $\lim _{x \rightarrow \infty} \psi(x)=\infty$,

$$
\begin{equation*}
\hat{I}(k)=\beta_{1} k^{\frac{1}{2}}+\beta_{2}+\beta_{3} \frac{\log k}{k^{\frac{1}{2}}}+O\left(\frac{\psi(k)}{k^{\frac{1}{2}}}\right) \tag{105}
\end{equation*}
$$

as $k \rightarrow \infty$, where

$$
\begin{aligned}
\beta_{1} & =\frac{\sqrt{2}}{\sqrt{T}}\left(\sqrt{A_{3}-1}-\sqrt{A_{3}-2}\right) \\
\beta_{2} & =\frac{A_{2}}{\sqrt{2 T}}\left(\frac{1}{\sqrt{A_{3}-2}}-\frac{1}{\sqrt{A_{3}-1}}\right),
\end{aligned}
$$

and

$$
\beta_{3}=\frac{1}{\sqrt{2 T}}\left(\frac{1}{4}-\frac{a}{c^{2}}\right)\left(\frac{1}{\sqrt{A_{3}-1}}-\frac{1}{\sqrt{A_{3}-2}}\right) .
$$

Explicit expressions for the constants $A_{2}$ and $A_{3}$ can be found in [21].
2. For the uncorrelated Stein-Stein model, there exist positive constants $B_{2}$ and $B_{3}$ such that, for every positive function $\psi$ with $\lim _{x \rightarrow \infty} \psi(x)=\infty$,

$$
\hat{I}(k)=\gamma_{1} k^{\frac{1}{2}}+\gamma_{2}+O\left(\frac{\psi(k)}{k^{\frac{1}{2}}}\right)
$$

as $k \rightarrow \infty$, where

$$
\gamma_{1}=\frac{\sqrt{2}}{\sqrt{T}}\left(\sqrt{B_{3}-1}-\sqrt{B_{3}-2}\right)
$$

and

$$
\gamma_{2}=\frac{B_{2}}{\sqrt{2 T}}\left(\frac{1}{\sqrt{B_{3}-2}}-\frac{1}{\sqrt{B_{3}-1}}\right) .
$$

Explicit expressions for the constants $B_{2}$ and $B_{3}$ can be found in [21].
Next, we turn our attention to the stock price density in the uncorrelated Hull-White model. Since this density is not equivalent to a regularly varying function (this follows from (84)), we cannot use Theorems 7.6 and 7.7 to characterize the asymptotic behavior of the implied volatility. The next assertion provides an asymptotic formula which can be applied to the implied volatility in the Hull-White model.

Theorem 7.8. Suppose that the distribution of the stock price $X_{T}$ admits a density $D_{T}$. Suppose also that

$$
D_{T}(x) \approx x^{-2} \exp \{-b(\log x)\}
$$

as $x \rightarrow \infty$, where the function $b$ is positive, is increasing on $[c, \infty)$ for some $c>0$, and is such that $b(x)=B(\log x)$ with $B^{\prime \prime}(x) \approx 1$ as $x \rightarrow \infty$. Let $\psi$ be a positive function with $\lim _{K \rightarrow \infty} \psi(K)=\infty$. Then

$$
\begin{aligned}
I(K)= & \frac{1}{\sqrt{T}}\left[\sqrt{2 \log K}-\sqrt{2 \log \frac{1}{K^{2} D_{T}(K)}-\log \log \frac{1}{K^{2} D_{T}(K)}+2 \log \left(\frac{\log \log K}{\log K}\right)}\right] \\
& +O\left((\log \log K)^{-1} \psi(K)\right) \\
= & \frac{1}{\sqrt{T}}\left[\sqrt{2 \log K}-\sqrt{2 b(\log K)-\log (b(\log K))+2 \log \left(\frac{\log \log K}{\log K}\right)}\right] \\
& +O\left((\log \log K)^{-1} \psi(K)\right)
\end{aligned}
$$

as $K \rightarrow \infty$.
Remark 7.9. The stock price distribution density in the uncorrelated Hull-White model satisfies the condition in Theorem 7.8. This can be checked using (84). Applying Theorem 7.8 and simplifying the resulting expressions, we can deduce the asymptotic formula for the implied volatility in the Hull-White model that was obtained in [22]. More details can be found in [22].

To prove Theorem 7.8, we first reason as in (99) and (100) to get the relation

$$
C(K) \approx \frac{\exp \{-b(\log K)\} \log K}{\log \log K} \approx K^{2} D_{T}(K) \frac{\log K}{\log \log K}
$$

as $K \rightarrow \infty$. It follows that $C \in P F_{\infty}$, and hence Corollary 2.5 can be applied. Simplifying the resulting expressions, using the mean value theorem, and taking into account that $b(u) \approx$ $(\log u)^{2}$ as $u \rightarrow \infty$, we complete the proof of Theorem 7.8.

## 8. Appendix.

Proof of Theorem 3.3. The following well-known fact from the theory of convex functions will be used in the proof: If $U(x)$ is a convex function on $(0, \infty)$, then the second (distributional) derivative $\mu$ of the function $U$ is a locally finite Borel measure on $(0, \infty)$, and any such measure is the second derivative of a convex function $U$ which is unique up to the addition of an affine function (see, e.g., [32, Appendix, p. 500]). Kellerer's theorem that appears in the proof below concerns marginal distributions of Markov martingales (see [27]; see also [23, 30]). This theorem is often used in the papers devoted to the existence of option pricing models reproducing observed option prices (see $[6,9,11]$ and the references therein; see also [12], where the Sherman-Stein-Blackwell theorem is employed).

Let $C$ be a pricing function, and denote by $\mu_{T}$ the distribution of the random variable $X_{T}$. Then the second distributional derivative of the function $K \mapsto e^{r T} C(T, K)$ coincides with the measure $\mu_{T}$. Our goal is to establish that conditions 1-5 in the formulation of Theorem 3.3 hold. It is clear that conditions 1 and 4 follow from the definitions. Condition 2 can be established using the equivalence of (3) and (27). We will next prove that condition 3 holds. This condition is equivalent to the following inequality:

$$
\begin{equation*}
\int_{0}^{\infty}(x-K)^{+} d \mu_{S}\left(e^{r S} x\right) \leq \int_{0}^{\infty}(x-K)^{+} d \mu_{T}\left(e^{r T} x\right) \tag{106}
\end{equation*}
$$

for all $K \geq 0$ and $0 \leq S \leq T<\infty$. The previous inequality can be established by taking into account the fact that the process $e^{-r t} X_{t}, t \geq 0$, is a martingale and applying Jensen's inequality. Finally, condition 5 follows from the estimate

$$
C(T, K) \leq e^{-r T} \int_{K}^{\infty} x d \mu_{T}(x)
$$

and (27). This proves the necessity part of Theorem 3.3.
To prove the sufficiency, let us assume that $C$ is a function such that conditions $1-5$ in the formulation of Theorem 3.3 hold. Consider the family of Borel probability measures $\left\{\nu_{T}\right\}_{T \geq 0}$ on $\mathbb{R}$ defined as follows: For every $T \geq 0, \nu_{T}(A)=\mu_{T}\left(e^{r T} A\right)$ if $A$ is a Borel subset of $[0, \infty)$, and $\nu_{T}((-\infty, 0))=0$. In this definition, the symbol $\mu_{T}$ stands for the second distributional derivative of the function $K \mapsto e^{r T} C(T, K)$. Since condition 3 is equivalent to (106), we have

$$
\begin{equation*}
\int_{\mathbb{R}}(x-K)^{+} d \nu_{S}(x) \leq \int_{\mathbb{R}}(x-K)^{+} d \nu_{T}(x) \tag{107}
\end{equation*}
$$

for all $K \geq 0$ and $0 \leq S \leq T<\infty$. Let $\varphi$ be a nondecreasing convex function on $[0, \infty)$, and denote by $\eta$ its second distributional derivative. Then we have $\varphi(x)=\int_{0}^{\infty}(x-u)^{+} d \eta(u)+$ $a x+b$ for all $x \geq 0$, where $a$ and $b$ are some constants. Next, using (107) and (27), we obtain

$$
\begin{equation*}
\int_{\mathbb{R}} \varphi(x) d \nu_{S}(x) \leq \int_{\mathbb{R}} \varphi(x) d \nu_{T}(x) \tag{108}
\end{equation*}
$$

for all $K \geq 0,0 \leq S \leq T<\infty$. Hence, the family $\left\{\nu_{T}\right\}_{T \geq 0}$ is increasing in the convex ordering. The reasoning leading from (107) to (108) is known (see [15] or Appendix 1 in [8]). By Kellerer's theorem (Theorem 3 in [27]), (108) and (27) imply the existence of a filtered probability space $\left(\Omega, \mathcal{F}, \mathcal{F}_{t}, \mathbb{P}^{*}\right)$ and of a Markov $\left(\mathcal{F}_{t}, \mathbb{P}^{*}\right)$-martingale $Y$ such that the distribution of $Y_{T}$ coincides with the measure $\nu_{T}$ for every $T \geq 0$. Now put $X_{T}=e^{r T} Y_{T}$, $T \geq 0$. It follows that the measure $\mu_{T}$ is the distribution of the random variable $X_{T}$ for every $T \geq 0$. This produces a stock price process $X$ such that the process $e^{-r t} X_{t}$ is a martingale. Next, we see that condition 4 implies $\mu_{0}=\delta_{x_{0}}$, and hence $X_{0}=x_{0} \mathbb{P}^{*}$-a.s. Taking into account inequality (27), we define the following function:

$$
\begin{equation*}
V(T, K)=\int_{K}^{\infty} x d \mu_{T}(x)-K \int_{K}^{\infty} d \mu_{T}(x)=\mathbb{E}^{*}\left[\left(X_{T}-K\right)^{+}\right] \tag{109}
\end{equation*}
$$

It is clear that the second distributional derivative of the function $K \mapsto V(T, K)$ coincides with the measure $\mu_{T}$. Therefore, $e^{r T} C(T, K)=V(T, K)+a(T) K+b(T)$ for all $T \geq 0$ and $K \geq 0$, where the functions $a$ and $b$ do not depend on $K$. Since $\lim _{K \rightarrow \infty} C(T, K)=0$ (condition 5) and $\lim _{K \rightarrow \infty} V(T, K)=0$, we see that $a(T)=b(T)=0$, and hence $C(T, K)=e^{-r T} V(T, K)$. It follows from (109) that $C$ is a call pricing function.

This completes the proof of Theorem 3.3.
Acknowledgment. The author is indebted to the anonymous referees for their helpful comments, which led to many improvements in the paper.

## REFERENCES

[1] J. Beirlant, Y. Goegebeur, J. Teugels, and J. Segers, Statistics of Extremes: Theory and Applications, John Wiley \& Sons, Chichester, UK, 2004.
[2] S. Benaim and P. Friz, Regular variation and smile asymptotics, Math. Finance, 19 (2009), pp. 1-12.
[3] S. Benaim and P. Friz, Smile asymptotics II: Models with known moment generating function, J. Appl. Probab., 45 (2008), pp. 16-32.
[4] S. Benaim, P. Friz, and R. Lee, On Black-Scholes implied volatility at extreme strikes, in Frontiers in Quantitative Finance: Volatility and Credit Risk Modeling, R. Cont, ed., John Wiley \& Sons, Hoboken, NJ, 2009, pp. 19-45.
[5] N. H. Bingham, C. M. Goldie, and J. L. Teugels, Regular Variation, Cambridge University Press, Cambridge, UK, 1987.
[6] H. Buehler, Expensive martingales, Quant. Finance, 6 (2006), pp. 207-218.
[7] R. Carmona and S. Nadtochiy, Local volatility dynamic models, Finance Stoch., 13 (2009), pp. 1-48.
[8] P. Carr and D. B. Madan, Towards a theory of volatility trading, in Volatility, R. Jarrow, ed., Risk Books, London, 1998, pp. 417-427.
[9] P. Carr and D. B. Madan, A note on sufficient conditions for no arbitrage, Finance Res. Lett., 2 (2005), pp. 125-130.
[10] R. Cont, Empirical properties of asset returns: Stylized facts and statistical issues, Quant. Finance, 1 (2001), pp. 223-236.
[11] L. Cousot, Conditions on option prices for absence of arbitrage and exact calibration, J. Banking and Finance, 31 (2007), pp. 3377-3397.
[12] M. H. A. Davis and D. G. Hobson, The range of traded option prices, Math. Finance, 17 (2007), pp. 1-14.
[13] M. R. Fengler, Semiparametric Modeling of Implied Volatility, Springer-Verlag, Berlin, Heidelberg, 2005.
[14] J.-P. Fouque, G. Papanicolaou, and K. R. Sircar, Derivatives in Financial Markets with Stochastic Volatility, Cambridge University Press, Cambridge, UK, 2000.
[15] H. Föllmer and A. Schied, Stochastic Finance, 2nd ed., Walter de Gruyter, Berlin, 2004.
[16] P. Friz, S Gerhold, A. Gulisashvili, and S. Sturm, On Refined Volatility Smile Expansion in the Heston Model, http://arxiv.org/abs/1001.3003 (2010), Quant. Finance, submitted.
[17] J. Gatheral, The Volatility Surface, A Practitioner's Guide, John Wiley \& Sons, Hoboken, NJ, 2006.
[18] C. M. Goldie and R. L. Smith, Slow variation with remainder: Theory and applications, Quart. J. Math. Oxford, 38 (1987), pp. 45-71.
[19] A. Gulisashvili and E. M. Stein, Asymptotic behavior of distribution densities in models with stochastic volatility, I, Math. Finance, 20 (2010), pp. 447-477.
[20] A. Gulisashvili and E. M. Stein, Asymptotic behavior of the distribution of the stock price in models with stochastic volatility: The Hull-White model, C. R. Math. Acad Sci. Paris, 343 (2006), pp. 519-523.
[21] A. Gulisashvili and E. M. Stein, Asymptotic behavior of the stock price distribution density and implied volatility in stochastic volatility models, Appl. Math. Optim., 61 (2010), pp. 287-315.
[22] A. Gulisashvili and E. M. Stein, Implied volatility in the Hull-White model, Math. Finance, 19 (2009), pp. 303-327.
[23] K. Hamza and F. C. Klebaner, A family of non-Gaussian martingales with Gaussian marginals, J. Appl. Math. Stoch. Anal., (2007), article ID 92723.
[24] P. Henry-Labordère, Analysis, Geometry, and Modeling in Finance: Advanced Methods in Option Pricing, Chapman \& Hall/CRC, Boca Raton, FL, 2009.
[25] S. L. Heston, A closed-form solution for options with stochastic volatility, with applications to bond and currency options, Rev. Financial Stud., 6 (1993), pp. 327-343.
[26] J. Hull and A. White, The pricing of options on assets with stochastic volatilities, J. Finance, 42 (1987), pp. 281-300.
[27] H. G. Kellerer, Markov-Komposition und eine Anwendung auf Martingale, Math. Ann., 198 (1972), pp. 99-122.
[28] D. Lamberton and B. Lapeyre, Introduction to Stochastic Calculus Applied to Finance, 2nd ed., Chapman \& Hall/CRC, Boca Raton, FL, London, New York, 2008.
[29] R. Lee, The moment formula for implied volatility at extreme strikes, Math. Finance, 14 (2004), pp. 469480.
[30] D. B. Madan and M. Yor, Making Markov martingales meet marginals: With explicit constructions, Bernoulli, 8 (2002), pp. 509-536.
[31] E. Renault and N. Touzi, Option hedging and implied volatilities in a stochastic volatility model, Math. Finance, 6 (1996), pp. 279-302.
[32] D. Revuz and M. Yor, Continuous Martingales and Brownian Motion, Springer-Verlag, Berlin, 1991.
[33] E. M. Stein and J. Stein, Stock price distributions with stochastic volatility: An analytic approach, Rev. Financial Stud., 4 (1991), pp. 727-752.

# Affine Point Processes and Portfolio Credit Risk* 

Eymen Errais ${ }^{\dagger}$, Kay Giesecke ${ }^{\ddagger}$, and Lisa R. Goldberg ${ }^{\S}$

Abstract. This paper analyzes a family of multivariate point process models of correlated event timing whose arrival intensity is driven by an affine jump diffusion. The components of an affine point process are self- and cross-exciting and facilitate the description of complex event dependence structures. ODEs characterize the transform of an affine point process and the probability distribution of an integer-valued affine point process. The moments of an affine point process take a closed form. This guarantees a high degree of computational tractability in applications. We illustrate this in the context of portfolio credit risk, where the correlation of corporate defaults is the main issue. We consider the valuation of securities exposed to correlated default risk and demonstrate the significance of our results through market calibration experiments. We show that a simple model variant can capture the default clustering implied by index and tranche market prices during September 2008, a month that witnessed significant volatility.

Key words. self-exciting point process, affine jump diffusion, Hawkes process, transform, portfolio credit derivative, correlated default, index and tranche swap

AMS subject classifications. 60-08, 60J99, 60G55, 90-08, 90B25
DOI. 10.1137/090771272

1. Introduction. The collapse of Lehman Brothers brought the financial system to the brink of a breakdown. The dramatic repercussions point to the existence of feedback phenomena that are channeled through the complex web of informational and contractual relationships in the economy. Lehman was an important node in a network of derivative traders. It had bought and sold default insurance on a large number of firms and was itself a reference entity in countless other insurance contracts. Its downfall triggered payments that forced some insurance sellers into default, leaving the corresponding protection buyers with losses. It also exposed the counterparties to the contracts Lehman itself had written.

This and related episodes motivate the design of models of correlated default timing that incorporate the feedback phenomena that plague credit markets. This paper analyzes a family of computationally tractable self-exciting point processes that can capture event feedback. The future evolution of a self-exciting point process is influenced by the timing of past events and their marks, for example, the financial loss they caused. This feature takes account of the direct impact of events. It also generates a dependence structure between arrival rates and losses, a property that is empirically well documented.

[^59]Our stepping stone is the Hawkes process, perhaps the most parsimonious self-exciting point process. The conditional event arrival rate or intensity of a Hawkes process jumps in response to events and tends toward a target level in the absence of an event. While the Hawkes process is widely used in a range of disciplines, its distributional properties are poorly understood. We develop these properties, exploiting the fact that the two-dimensional process consisting of a Hawkes process and its intensity is Markov. The structure of the associated infinitesimal generator leads to a Dynkin formula and closed expressions for the moments of the Hawkes intensity. We show that a transform of the Hawkes process satisfies a certain partial integral differential equation (PIDE). The solution to that equation turns out to be an exponentially affine function of the initial value of the two-dimensional process, whose coefficients satisfy a system of ODEs. Analysis of the transform leads to ODEs that characterize the probability distribution of a Hawkes process. We obtain closed formulae for the moments of the process.

The transform, distribution, and moment formulae generate computational tractability for a range of applications in portfolio credit risk. To illustrate, we use a Hawkes process to model the cumulative loss due to default in a portfolio of firms. The jump times represent default times, and the jump magnitudes represent the random losses at default. This formulation captures the impact of a default on the surviving names. It also incorporates the negative correlation between default and recovery rates. The transform formulae facilitate the valuation, hedging, and calibration of a portfolio credit derivative, which is a security whose payoff is a specified function of the portfolio loss, and which provides insurance against default losses in the portfolio. An index swap, for example, pays any portfolio losses before the contract maturity. A tranche swap pays a slice of the portfolio loss specified by a lower and an upper attachment point. Our market calibration experiments, which are based on index and tranche price data observed before and after Lehman's collapse, indicate the empirical significance of the self-exciting property of the loss process. An empirical analysis demonstrates that the parsimonious Hawkes process can capture the default correlation implied by credit market rates on each trading day in September 2008, a month that witnessed dramatic volatility. This is a significant improvement over the industry standard copula model.

The insight into the probabilistic structure of the Hawkes process leads us to consider an extension to a multivariate point process whose intensity is driven by an affine jump diffusion process that represents a vector of stochastic risk factors. The variation of the factors generates diffusive and jump volatility of conditional event arrival rates. The point process itself can be a risk factor so that the point process components are self- and cross-exciting. The basic transform, distribution, and moment characterization arguments developed for the Hawkes process extend to this family of affine point processes.

While this article highlights derivatives valuation applications, the self-exciting point processes considered here are also potential tools for the risk management of corporate debt portfolios, for which event feedback and the dependence between default and recovery rates are significant issues. For example, Das et al. [14] test a doubly stochastic model whose features are similar to those of the models endorsed by the regulatory authorities to estimate portfolio credit risk. In a doubly stochastic model event times are conditionally independent. Das et al. [14] find evidence of historical default clustering in excess of that implied by the model they tested. This suggests that doubly stochastic models underestimate risk capital.

A self-exciting point process model implies a more realistic degree of default clustering. It also accounts for the dependence between default and recovery rates, whose importance is emphasized in Basel II's pillar I guidelines; see [5].

The point processes examined in this paper facilitate a top-down approach to portfolio credit risk, in which the portfolio loss process is specified without reference to the constituent names. There are other examples of this approach in the literature. Arnsdorf and Halperin [2] describe arrivals by a nonlinear death process and its generalizations. Brigo, Pallavicini, and Torresetti [9] model events by a mixed Poisson process and its extensions. Cont and Minca [12] describe arrivals by a nonhomogeneous Markov chain. Davis and Lo [15] model events by a piecewise deterministic Markov process. Ding, Giesecke, and Tomecek [16] propose a time-changed linear birth process as a model of arrivals. In [28], defaults are driven by independent Poisson processes that model idiosyncratic, sector specific, and economy-wide events. Lopatin and Misirpashaev [29] introduce a point process model with an intensity whose drift is modulated by the portfolio loss. The family of loss point processes proposed in this paper is distinct from, or in some special cases encompasses, the models introduced in these contributions. This paper features multivariate, interacting point processes with stochastic recoveries, and correlated arrival and recovery rates. Calibration experiments establish the fit of a basic model variant to the market data in September 2008.

The top-down specification of the portfolio loss process generates computational advantages for portfolio derivatives valuation. Constituent name hedging requires the sensitivities of the portfolio derivative price with respect to changes in the prices of the single-name derivatives referenced on the constituents. The sensitivities determine the amount of single-name protection on each portfolio constituent to be bought or sold in order to neutralize portfolio derivative price fluctuations due to small changes in the constituent risks. Giesecke, Goldberg, and Ding [20] develop a random thinning approach to estimate these hedges for a given portfolio loss process. For September 2008, they demonstrate the effectiveness of this approach for the Hawkes process. The empirical results indicate that a top-down model enables better hedging than the industry standard copula model.

Section 2 develops the distributional properties of the Hawkes process. Section 3 applies these results to the valuation of credit derivatives. Market calibration experiments demonstrate the significance of the self-exciting feature and the fit of the Hawkes model. Section 4 provides further parametric examples of multivariate self- and cross-exciting point processes. Section 5 concludes. The appendix contains the proofs.
2. Hawkes process. Consider a sequence of default stopping times $0<T_{1}<T_{2}<\ldots$ that are defined on a complete probability space $(\Omega, \mathcal{F}, P)$ with right-continuous and complete information filtration $\mathbb{F}=\left(\mathcal{F}_{t}\right)_{t \geq 0}$. The nature of the probability measure $P$ depends on the application. In risk management applications, $P$ is the actual or statistical measure. In valuation applications, $P$ is a risk-neutral pricing measure, relative to which the discounted price of a traded security is a martingale. The financial loss at $T_{n}$ is given by a random variable $\ell_{n} \in \mathcal{F}_{T_{n}}$. The sequence $\left(T_{n}, \ell_{n}\right)$ generates a nonexplosive default counting process $N$ given by $N_{t}=\sum_{n \geq 1} 1_{\left\{T_{n} \leq t\right\}}$ and a loss point process $L$ defined by $L_{t}=\sum_{n \geq 1} \ell_{n} 1_{\left\{T_{n} \leq t\right\}}$.

We propose to specify the processes $N$ and $L$ directly through a conditional arrival rate or intensity $\lambda$ and a distribution $\nu$ on $(0, \infty)$ for the loss $\ell_{n}$ at an event. We assume that the


Figure 1. Annual defaults of Moody's rated U.S. firms between January 1970 and November 17, 2008. Source: Moody's Default Risk Service. The peak in 1970 represents a cluster of 24 railway defaults triggered by the collapse of Penn Central Railway on June 21,1970. The fallout of the 1987 crash is indicated by the peak in the early 1990s. The burst of the internet bubble caused many defaults during 2001-2002. From a trough in 2007, default rates increased significantly in 2008.
jump transform $\int e^{\omega z} d \nu(z)$ exists and is finite for complex $\omega$ and admits a finite derivative $\int z e^{\omega z} d \nu(z)$. The intensity follows a strictly positive stochastic process that describes the conditional mean default rate in the sense that $E\left(N_{t+\Delta}-N_{t} \mid \mathcal{F}_{t}\right) \approx \lambda_{t} \Delta$ for small $\Delta>0$. This means that $N-\int_{0}^{c} \lambda_{s} d s$ is a local martingale relative to $P$ and $\mathbb{F}$. The process followed by $\lambda$ completely determines the conditional distribution of $N$. We analyze a family of models for $\lambda$ whose features are empirically motivated.
2.1. Empirical motivation. Empirical observation dictates the properties of the stochastic process followed by $\lambda$. Most importantly, $\lambda$ must replicate the clustering of defaults seen in Figure 1, which is due to the dependence of the default times $T_{n}$. The dependence is a result of the sensitivity of firms to common economic risk factors that vary stochastically through time. It is also caused by the informational and contractual linkages between firms, which provide a channel for the propagation of financial distress from one firm to another. The existence of these feedback phenomena is indicated by the ripple effects associated with the default of Lehman Brothers on September 15, 2008 and is further empirically documented in [4], [11], [17], [24], [27], and others. We propose to model the impact of a default on the other firms by ramping up the intensity at an event. This amounts to including the default process itself as a risk factor.

Another important empirical feature is the negative correlation between default and recovery rates. During periods with elevated default rates, creditors tend to recover less at a default than during periods with relatively few defaults. While documented by Altman et al. [1] and many others, this property has been largely ignored in a theoretical literature that is focused on modeling default timing. We propose to capture this property by modeling the magnitude of the response of the intensity to a default as a linear function of the realized loss


Figure 2. Sample paths of the intensity (1) and the associated loss process L. A jump in the intensity represents the impact of a default. The jump size is the product of the loss at default and the sensitivity parameter $\delta=1$. The loss at default is drawn from a uniform distribution on $\{0.4,0.6,0.8,1\}$. The rate $\kappa=5$ controls the exponential decay after an event. The reversion level $c=\lambda_{0}=0.7$.
at an event. The larger the loss, the larger the impact of an event on the other firms, and the bigger the increase of the intensity at an event.
2.2. Specification. We examine a basic intensity model that incorporates the features described above. We suppose events arrive with intensity $\lambda$ given by

$$
\begin{equation*}
\lambda_{t}=u(t)+\int_{0}^{t} h(t-s) d L_{s} . \tag{1}
\end{equation*}
$$

The first to default intensity $u(t)=c+e^{-\kappa t}\left(\lambda_{0}-c\right)$ is a deterministic function of time, $c>0$, $\lambda_{0}>0$, and the impact of a loss on the intensity is governed by the function

$$
\begin{equation*}
h(v)=\delta e^{-\kappa v}, \quad v \geq 0, \tag{2}
\end{equation*}
$$

with $\kappa \geq 0$ and $\delta \geq 0$. As illustrated in the sample path of $(\lambda, L)$ shown in Figure 2, the randomness in the default rate is driven by two sources of uncertainty: the timing of events and the recovery at these events. At a default, the intensity ramps up by the realized loss scaled with the sensitivity parameter $\delta$. The lower the recovery, the higher the jump of the default rate. The impact of an event decays exponentially over time with rate $\kappa$. The reversion level is $c$. It follows that defaults are positively self-affecting, or self-exciting. Further, default and recovery rates are negatively correlated. Despite its parsimony, in section 3.3 we show that the basic model (1) is very effective at replicating the clustering of defaults implied by market prices.

While parsimonious and empirically motivated, the intensity model (1) and its extensions discussed in section 4 may not be appropriate for relatively small portfolios. This is because (1) generates a point process $(N, L)$ that does not terminate when all firms in the portfolio are


Figure 3. Sample paths of the Poisson and Hawkes processes. A bar indicates the number of arrivals in a given year. While the Poisson arrivals are evenly distributed over time due to the order statistics property, the Hawkes arrivals are clustered thanks to the self-exciting property. For the Hawkes process, $c=\lambda_{0}=1$, $\kappa=1.5, \delta=2$, and the loss at default has a uniform distribution $\nu$ on $\{0.4,0.6,0.8,1\}$. The Poisson process is the special Hawkes process for which $\delta=0$ so that $\lambda_{t}=u(t)$. We choose $c=\lambda_{0}=10.57$ to match the expected number of Hawkes events over 30 years.
in default. This feature tends to be innocuous for the large diversified portfolios in practice, which can have as many as several thousand names, and for which the likelihood of total default is typically negligible. It is fully appropriate only for portfolios in which defaulted names are replaced with new names, as may be the case for collateralized debt obligations with actively managed collateral pools.

The intensity $\lambda$ governs the common event times of $N$ and $L$. While the jumps of the default process $N$ are unit-sized, the jumps of the loss process $L$ are drawn from the distribution $\nu$ on $(0, \infty)$ of loss given default. The two-dimensional process $J=(L, N)^{\top}$ is a Hawkes process; see [22] and [23]. Variants of the Hawkes process have been applied to a range of problems in science and engineering; see [13] for a sample of these. They have only recently been used in financial economics. For example, Bowsher [7] fits different Hawkes models to security trade data. Surprisingly, despite these applications, the distributional properties of Hawkes-type processes are poorly understood. We establish these properties below and extend the Hawkes process to a broad family of self-exciting point processes.

We note two special cases. If $\kappa=0$ in the specification (1), then $\lambda$ is constant between events, and so $J$ is a birth process. If $\delta=0$, then $J$ is a Poisson process whose intensity $u(t)$ evolves deterministically through time and whose interarrival times are independent. Figure 3 contrasts the sample paths of the Poisson and Hawkes processes. It shows how the selfexciting property of the Hawkes process generates arrivals that are overdispersed relative to Poisson arrivals. The resulting event clusters are strikingly similar to those of the empirical default process in Figure 1. The parameter $\delta$ in (1) allows us to directly control the frequency of these clusters. The parameter $\kappa$ governs their magnitude.

While the Hawkes process induces clustered arrivals, it does not explode, and therefore the integral $A_{t}=\int_{0}^{t} \lambda_{s} d s$ is finite for all $t$. To see this, it suffices to consider the case $\kappa=0$, when $\lambda$ has nondecreasing sample paths. Then $N$ is a birth process, which is nonexplosive and integrable. From this we can conclude that $N_{t}$ is integrable also in the case $\kappa>0$. This, in turn, implies that $A_{t}$ is integrable, and that the compensated local jump martingale $M=N-A$ is a martingale. To see this, we can appeal to Corollary 3 in Chapter II. 6 of [30],
which implies that $M$ is a martingale if $[M, M]$ is integrable. But $[M, M]=N$. Further, we obtain that $E\left(M_{t}^{2}\right)=E\left(N_{t}\right)=E\left(A_{t}\right)$.
2.3. Dynkin formula. While the process $J$ itself does not have the Markov property, the process $Y=(\lambda, J)^{\top}$ is a Markov process in a state space $D=\mathbb{R}_{+} \times\left(\mathbb{R}_{+} \times \mathbb{N}\right) . Y$ has an infinitesimal generator $\mathcal{D}$, defined at a function $g: D \rightarrow \mathbb{R}$ with continuous partial derivative $g_{\lambda}(\lambda, x)$, by

$$
\begin{equation*}
(\mathcal{D} g)(\lambda, x)=\kappa(c-\lambda) g_{\lambda}(\lambda, x)+\lambda \int\left[g\left(\lambda+\delta z, x+(z, 1)^{\top}\right)-g(\lambda, x)\right] d \nu(z) \tag{3}
\end{equation*}
$$

Proposition 2.1. If the expectation

$$
\begin{equation*}
E\left(\int_{0}^{t}\left|(\mathcal{D} g)\left(\lambda_{s}, J_{s}\right)-\kappa\left(c-\lambda_{s}\right) g_{\lambda}\left(\lambda_{s}, J_{s}\right)\right| d s\right) \tag{4}
\end{equation*}
$$

is finite for each $t$, then for each $t \leq T$ we have a Dynkin formula

$$
\begin{equation*}
E\left(g\left(\lambda_{T}, J_{T}\right) \mid \mathcal{F}_{t}\right)=g\left(\lambda_{t}, J_{t}\right)+E\left(\int_{t}^{T}(\mathcal{D} g)\left(\lambda_{s}, J_{s}\right) d s \mid \mathcal{F}_{t}\right) . \tag{5}
\end{equation*}
$$

Proposition 2.1 provides a formula expressing the conditional expectation of a function $g$ of the Markov process $Y=(\lambda, J)^{\top}$ in terms of the infinitesimal generator (3). This formula is useful for functions $g$ for which the expectation of $(\mathcal{D} g)\left(\lambda_{s}, J_{s}\right)$ can be calculated. We exploit it to obtain explicit formulae for the first and second moments of the intensity. Consider the function $g(\lambda, x)=\lambda$. We calculate

$$
\begin{aligned}
E\left((\mathcal{D} g)\left(\lambda_{s}, J_{s}\right)\right) & =E\left(\kappa\left(c-\lambda_{s}\right)\right)+E\left(\lambda_{s} \int \delta z d \nu(z)\right) \\
& =\kappa c+\mu E\left(\lambda_{s}\right)
\end{aligned}
$$

where $\mu=\delta \ell-\kappa$ and $\ell=\int z d \nu(z)$ is the expected loss at default. Letting $m(t)=E\left(\lambda_{t}\right)$, equation (5) implies that

$$
\begin{equation*}
m(t)=E\left(g\left(\lambda_{t}, J_{t}\right)\right)=\lambda_{0}+\kappa c t+\mu \int_{0}^{t} m(s) d s \tag{6}
\end{equation*}
$$

In other words, the mean intensity solves the ODE $m^{\prime}(t)=\kappa c+\mu m(t)$ with initial condition $m(0)=\lambda_{0}$. The solution is

$$
\begin{equation*}
m(t)=E\left(\lambda_{t}\right)=\left(\frac{\kappa c}{\mu}+\lambda_{0}\right) e^{\mu t}-\frac{\kappa c}{\mu} \tag{7}
\end{equation*}
$$

provided that $\mu \neq 0$. Equation (7) shows that for $\mu>0$, the mean intensity grows exponentially (without bound) in time. For $\mu \rightarrow 0$, the mean intensity grows linearly with time as $\lim _{\mu \rightarrow 0} E\left(\lambda_{t}\right)=\lambda_{0}+\kappa c t$. For $\mu<0$, the mean intensity decays exponentially with time to the long-run mean

$$
\begin{equation*}
\lim _{t \rightarrow \infty} E\left(\lambda_{t}\right)=-\frac{\kappa c}{\mu}>0 \tag{8}
\end{equation*}
$$

We apply a similar argument to the function $g(\lambda, x)=\lambda^{2}$ to show that the second moment $v(t)=E\left(\lambda_{t}^{2}\right)$ solves the ODE

$$
\begin{equation*}
v^{\prime}(t)=\rho m(t)+2 \mu v(t) \tag{9}
\end{equation*}
$$

with initial condition $v(0)=\lambda_{0}^{2}$, where $\rho=2 \kappa c+\delta^{2} \int z^{2} d \nu(z)$. Here, we assume that $\nu$ has finite second moment. For $\mu \neq 0$, the solution to (9) is given by

$$
\begin{equation*}
v(t)=\frac{1}{2 \mu^{2}}\left[\kappa c \rho\left(e^{\mu t}-1\right)^{2}+2 \mu \lambda_{0} e^{\mu t}\left(\rho\left(e^{\mu t}-1\right)+\mu \lambda_{0} e^{\mu t}\right)\right] . \tag{10}
\end{equation*}
$$

We conclude that for $\mu<0$, the long run variance of the intensity satisfies

$$
\begin{equation*}
\lim _{t \rightarrow \infty} \operatorname{Var}\left(\lambda_{t}\right)=\lim _{t \rightarrow \infty}\left(v(t)-m^{2}(t)\right)=\frac{\kappa c \delta^{2}}{2 \mu^{2}} \int z^{2} d \nu(z) . \tag{11}
\end{equation*}
$$

2.4. Transform. Along with the distribution $\nu$ of the loss at default, the intensity specification (1) determines the distribution of the processes $N$ and $L$. We develop a formula for a transform of the point process $J=(L, N)^{\top}$ that includes as special cases the Laplace and Fourier transforms. The transform can be inverted to obtain the corresponding distribution.

Let $u \in \mathbb{C}_{-}^{2}$, the set of pairs of complex numbers with nonpositive real part. Express the conditional transform $E\left(\exp \left(u \cdot J_{T}\right) \mid \mathcal{F}_{t}\right)$ as $f\left(t, \lambda_{t}, J_{t}\right)$ for some complex-valued function $f$ on $[0, T] \times D$. For $f\left(t, \lambda_{t}, J_{t}\right)$ to be a martingale, its drift must vanish. This means that $f$ must satisfy the PIDE

$$
0=f_{t}(t, \lambda, x)+(\mathcal{D} f)(t, \lambda, x)
$$

with boundary condition $f(T, \lambda, x)=\exp (u \cdot x)$, where $(\mathcal{D} f)(t, \lambda, x)$ is obtained by applying the generator $\mathcal{D}$ to the real and imaginary parts of $f(t, \cdot, \cdot)$. This PIDE reduces to a system of ODEs when $f$ is taken to be an exponentially affine function of $\lambda$ and $x$. This argument is made precise in the next result.

Proposition 2.2. The transform of the point process $J=(L, N)^{\top}$ is given by

$$
\begin{equation*}
E\left(\exp \left(u \cdot J_{T}\right) \mid \mathcal{F}_{t}\right)=\exp \left(a(t)+b(t) \lambda_{t}+u \cdot J_{t}\right) \tag{12}
\end{equation*}
$$

where $t \leq T, u \in \mathbb{C}_{-}^{2}$, and the coefficient functions $a(t)=a(u, t, T)$ and $b(t)=b(u, t, T)$ satisfy the following ODEs:

$$
\begin{align*}
& \partial_{t} b(t)=\kappa b(t)+1-\theta\left(\delta b(t)+u \cdot(1,0)^{\top}\right) \exp \left(u \cdot(0,1)^{\top}\right),  \tag{13}\\
& \partial_{t} a(t)=-\kappa c b(t) \tag{14}
\end{align*}
$$

with boundary conditions $a(T)=b(T)=0$, where $\theta$ is the jump transform

$$
\begin{equation*}
\theta(\omega)=\int e^{\omega z} d \nu(z), \quad \omega \in \mathbb{C} \tag{15}
\end{equation*}
$$

The argument behind Proposition 2.2 immediately extends to more general functionals of $Y=(\lambda, J)^{\top}$. For example, the argument leads to the transform of the vector $Y$ and not
just its component $J$. Section 4 covers this extension in a more general setting with multiple stochastic risk factors driving the intensity $\lambda$ of $J$. We then obtain the joint transform of the point process and the risk factors. This yields, in particular, the transform of the intensity of the point process.

In some cases the ODEs (13) and (14) can be solved analytically-for example, when $\kappa=0$ and $J$ is a birth process. In the general case, the ODEs are quickly solved numerically, using the Runge-Kutta algorithm, for example.
2.5. Moments. By differentiating the conditional transform of $J$ with respect to $u$ and evaluating the derivative at $u=0$, we find the conditional expectation

$$
\begin{equation*}
E\left(v \cdot J_{T} \mid \mathcal{F}_{t}\right)=\mathcal{A}(0, t, T)+\mathcal{B}(0, t, T) \lambda_{t}+v \cdot J_{t} \tag{16}
\end{equation*}
$$

for $v \in \mathbb{R}^{2}$ and $t \leq T$, where the functions $\mathcal{A}(t)=\mathcal{A}(u, t, T)$ and $\mathcal{B}(t)=\mathcal{B}(u, t, T)$ satisfy

$$
\begin{align*}
\partial_{t} \mathcal{B}(t) & =\kappa \mathcal{B}(t)-\left(\delta \mathcal{B}(t)+v \cdot(1,0)^{\top}\right) \theta^{\prime}(\delta b(t))-v \cdot(0,1)^{\top} \theta(\delta b(t)),  \tag{17}\\
\partial_{t} \mathcal{A}(t) & =-\kappa c \mathcal{B}(t) \tag{18}
\end{align*}
$$

with boundary conditions $\mathcal{A}(T)=\mathcal{B}(T)=0$. The derivative of the jump transform is

$$
\theta^{\prime}(\omega)=\int z e^{\omega z} d \nu(z), \quad \omega \in \mathbb{C}
$$

and $a(t)$ and $b(t)$ satisfy (13) and (14) and the relevant boundary conditions. Since $u=0$, we can choose $a(t)=b(t)=0$. Under mild assumptions on the distribution $\nu$ (continuity of $\theta^{\prime}(\omega)$ suffices), this solution is unique and (17) simplifies to

$$
\begin{equation*}
\partial_{t} \mathcal{B}(t)=(\kappa-\delta \ell) \mathcal{B}(t)-v \cdot(\ell, 1)^{\top}, \tag{19}
\end{equation*}
$$

where $\ell=\int z d \nu(z)$ is the expected loss at default. Provided that $\mu=\delta \ell-\kappa \neq 0$, we obtain the following explicit solutions:

$$
\begin{align*}
\mathcal{B}(t) & =\frac{1}{\mu} v \cdot(\ell, 1)^{\top}\left(e^{\mu(T-t)}-1\right)  \tag{20}\\
\mathcal{A}(t) & =\frac{1}{\mu} v \cdot(\ell, 1)^{\top} \kappa c\left(\frac{1}{\mu}\left(e^{\mu(T-t)}-1\right)-T+t\right) . \tag{21}
\end{align*}
$$

For $\mu \neq 0$, the mean number of events takes the form

$$
E\left(N_{t}\right)=c_{1}\left(e^{\mu t}-1\right)+c_{2} t,
$$

where $c_{1}=\left(\kappa c+\mu \lambda_{0}\right) / \mu^{2}$ and $c_{2}=-\kappa c / \mu$. If $\kappa=0$ and $N$ is a birth process, then $E\left(N_{t}\right)$ grows exponentially in $t$ at rate $\delta \ell>0$. If $\kappa>0$, then the growth of $E\left(N_{t}\right)$ is determined by the sign of $\mu$. If $\mu>0$, then we have exponential growth at rate $\mu$ that is counteracted by linear decay at rate $c_{2}$. If $\mu<0$, then we have linear growth at rate $c_{2}=\lim _{t \rightarrow \infty} E\left(\lambda_{t}\right)$ that is counteracted by exponential decay at rate $\mu$.

Higher order conditional moments of $J_{T}$ can be calculated similarly, by successively differentiating the transform. Thus, we obtain a closed form expression for conditional expectations of the form $E\left(f\left(J_{T}\right) \mid \mathcal{F}_{t}\right)$, where $f: \mathbb{R}_{+} \times \mathbb{N} \rightarrow \mathbb{R}$ is an integrable function that is polynomial. We can estimate the conditional expectation $E\left(h\left(J_{T}\right) \mid \mathcal{F}_{t}\right)$ for any continuous, integrable function $h$ on $\mathbb{R}_{+} \times \mathbb{N}$ by approximating $h$ with $f$.
2.6. Transform inversion. The transform of $J_{T}$ can be inverted using fast Fourier transform techniques to obtain the conditional distribution given $\mathcal{F}_{t}$ of $N_{T}$ and $L_{T}$ for all future dates $T$. The distribution of $N_{T}$ can also be obtained directly. Proposition 2.2 implies that the conditional probability generating function of the Hawkes process takes the form

$$
\begin{equation*}
E\left(v^{N_{T}-N_{t}} \mid \mathcal{F}_{t}\right)=\exp \left(c(t)+d(t) \lambda_{t}\right) \tag{22}
\end{equation*}
$$

for $v \in(0,1)$ and $t \leq T$, where the coefficient functions

$$
\begin{aligned}
c(t) & =c(v, t, T) \\
d(t) & =d\left((0, \log v)^{\top}, t, T\right), \\
d, T) & =b\left((0, \log v)^{\top}, t, T\right)
\end{aligned}
$$

satisfy the ODEs

$$
\begin{align*}
\partial_{t} d(t) & =\kappa d(t)+1-v \theta(\delta d(t)),  \tag{23}\\
\partial_{t} c(t) & =-\kappa c d(t) \tag{24}
\end{align*}
$$

with boundary conditions $c(T)=d(T)=0$. Equations (23)-(24) determine the distribution of the Hawkes process. Expansion of the left-hand side of (22) into a power series shows that for $n=0,1,2, \ldots$.

$$
\begin{equation*}
\varphi_{t}(n, T)=P\left(N_{T}-N_{t}=n \mid \mathcal{F}_{t}\right)=\left.\frac{1}{n!} \partial_{v}^{n} \exp \left(c(v, t, T)+d(v, t, T) \lambda_{t}\right)\right|_{v=0} \tag{25}
\end{equation*}
$$

In practice, the calculation of these probabilities calls for the solution of a system of ODEs derived from (23)-(24). In case $n=0$ we get

$$
\varphi_{t}(0, T)=\exp \left(C(t)+D(t) \lambda_{t}\right),
$$

where the functions $D(t)=d(0, t, T)$ and $C(t)=c(0, t, T)$ satisfy the ODEs

$$
\begin{align*}
\partial_{t} D(t) & =\kappa D(t)+1,  \tag{26}\\
\partial_{t} C(t) & =-\kappa c D(t)
\end{align*}
$$

with boundary conditions $C(T)=D(T)=0$. We obtain the formula

$$
\begin{equation*}
\varphi_{t}(0, T)=\exp \left(\frac{c-\lambda_{t}}{\kappa}\left(1-e^{-\kappa(T-t)}\right)-c(T-t)\right) \tag{27}
\end{equation*}
$$

provided that $\kappa>0$. Note that the probability (27) of no events during $(t, T]$ does not explicitly depend on the clustering parameter $\delta$ (this parameter influences only the current intensity value $\lambda_{t}$ ). It agrees with the probability of no events during $(t, T]$ for an inhomogeneous Poisson process with intensity $u(t)=c+\left(\lambda_{0}-c\right) \exp (-\kappa t)$. This is because between events, the Hawkes intensity (1) is deterministic and governed by $u(t)$.

For $n \geq 1$ the probability (25) depends on the clustering parameter $\delta$. Using Faà di Bruno's formula (chain rule for higher derivatives), we get

$$
\begin{equation*}
\varphi_{t}(n, T)=\varphi_{t}(0, T) \sum \frac{1}{m_{1}!\cdots m_{n}!} \prod_{k=1}^{n}\left(\frac{C_{k}(t)+D_{k}(t) \lambda_{t}}{k!}\right)^{m_{k}} \tag{28}
\end{equation*}
$$

where the sum is over all $n$-tuples $\left(m_{1}, \ldots, m_{n}\right)$ of nonnegative integers that satisfy the constraint $m_{1}+2 m_{2}+\cdots+n m_{n}=n$ and the functions

$$
\begin{array}{r}
D_{k}(t)=\left.\partial_{v}^{k} d(v, t, T)\right|_{v=0}, \\
C_{k}(t)=\left.\partial_{v}^{k} c(v, t, T)\right|_{v=0}
\end{array}
$$

satisfy the ordinary differential equations

$$
\begin{align*}
\partial_{t} D_{k}(t) & =\kappa D_{k}(t)-k E_{k-1}(t),  \tag{29}\\
\partial_{t} C_{k}(t) & =-\kappa c D_{k}(t) \tag{30}
\end{align*}
$$

with boundary conditions $C_{k}(T)=D_{k}(T)=0$, where $E_{k}(t)=\left.\partial_{v}^{k} \theta(\delta d(v, t, T))\right|_{v=0}$ can be calculated by another application of Faà di Bruno's formula,

$$
\begin{equation*}
E_{k}(t)=\sum \frac{k!}{m_{1}!\cdots m_{k}!} \theta^{\left(m_{1}+\cdots+m_{k}\right)}(\delta D(t)) \prod_{i=1}^{k}\left(\frac{\delta D_{i}(t)}{i!}\right)^{m_{i}}, \tag{31}
\end{equation*}
$$

and the sum is over all $k$-tuples $\left(m_{1}, \ldots, m_{k}\right)$ of nonnegative integers that satisfy $m_{1}+2 m_{2}+$ $\cdots+k m_{k}=k$. Equations (29) and (30) are solved numerically unless $\delta=0$, in which case $N$ is a time-inhomogeneous Poisson process with intensity $u(t)$. Klimko [25] provides a fast algorithm for the calculation of the summands in Faà di Bruno's formula, which can be applied to treat the computation of (28) and (31) efficiently.
3. Portfolio credit derivatives. We use the results developed above to analyze portfolio credit derivatives, which expose an investor to the risk of correlated default in a reference portfolio of credit-sensitive instruments such as loans or bonds. We consider index and tranche swaps, which are the most popular portfolio derivatives. These are bilateral contracts, in which one party provides default protection on the reference portfolio, and the other party pays a premium for this protection. The basic valuation problem, addressed below, is to determine the fair premium, which is governed by the correlated default timing in the reference portfolio. Through a market calibration, we show that the basic intensity model (1) captures the correlated default risk implied by the premia offered in the market.
3.1. Valuation. Index and tranche swaps are based on a portfolio whose $n$ constituent securities have a notional 1 , maturity date $T$, and premium payment dates $\left(t_{m}\right)$. The loss at the $k$ th default is $\ell_{k} \in[0,1]$. The swap is specified by a lower attachment point $\underline{K} \in[0,1]$ and an upper attachment point $\bar{K} \in(\underline{K}, 1]$. An index swap has attachment points $\underline{K}=0$ and $\bar{K}=1$. The swap notional $K=n(\bar{K}-\underline{K})$. The protection seller covers portfolio losses as they occur, given that the cumulative losses are larger than $\underline{K}$ but do not exceed $\bar{K}$. The cumulative payments at time $t$, denoted $U_{t}$, are given by the "call spread"

$$
\begin{equation*}
U_{t}=\left(L_{t}-\underline{K} n\right)^{+}-\left(L_{t}-\bar{K} n\right)^{+} . \tag{32}
\end{equation*}
$$

The value at time $t \leq T$ of these payments is given by

$$
\begin{equation*}
D_{t}=E\left(\int_{t}^{T} e^{-r(s-t)} d U_{s} \mid \mathcal{F}_{t}\right) \tag{33}
\end{equation*}
$$

where, here and below, the reference measure $P$ is a risk-neutral pricing measure with respect to an interest rate $r>0$. By Stieltjes integration by parts, we can conveniently express the value $D_{t}$ in terms of conditional expectations of $U$ :

$$
\begin{equation*}
D_{t}=e^{-r(T-t)} E\left(U_{T} \mid \mathcal{F}_{t}\right)-U_{t}+r \int_{t}^{T} e^{-r(s-t)} E\left(U_{s} \mid \mathcal{F}_{t}\right) d s \tag{34}
\end{equation*}
$$

The protection buyer receives the loss payments and, in return, makes premium payments to the protection seller. Each premium payment has two parts. The first part is an upfront payment, which is expressed as a fraction $F$ of the tranche notional $K$. For an index swap, $F=0$. The second part consists of payments that are proportional to the premium notional $I_{t}$, which is given by $n-\left(N_{t} \wedge n\right)=n+\left(N_{t}-n\right)^{+}-N_{t}$ for an index swap and $K-U_{t}$ for a tranche swap with $\bar{K}<1$. Let $c_{m}$ be the day count fraction for the period $m$, roughly $1 / 4$ for quarterly payments. Then, with $S$ denoting the running premium rate, the value at time $t \leq T$ of the premium payments is given by

$$
\begin{equation*}
P_{t}(F, S)=F K+S \sum_{t_{m} \geq t} e^{-r\left(t_{m}-t\right)} c_{m} E\left(I_{t_{m}} \mid \mathcal{F}_{t}\right) \tag{35}
\end{equation*}
$$

For a fixed upfront rate $F$, the running spread $S_{t}$ at time $t$ is the solution $S=S_{t}$ to the equation $D_{t}=P_{t}(F, S)$. Setting $D_{t}=P_{t}(F, S)$ for a fixed $S$ gives a value $F=F_{t}$ for the time $t$ upfront rate $F_{t}$. Formulae (34) and (35) indicate that these rates depend only on call options $E\left(\left(v \cdot J_{s}-c\right)^{+} \mid \mathcal{F}_{t}\right)$ with various strikes $c$, maturities $s \in(t, T]$, and values $v \in\left\{(0,1)^{\top},(1,0)^{\top}\right\}$. Thus, to value a swap we need only calculate the values of options on $N$ and $L$. One approach to calculating these values is to integrate the option payoff function $(x-c)^{+}$against the point process distribution. The latter is obtained by inverting the point process transform in Proposition 2.2. In the case of $N$, the inverse transform is given by formula (25). Alternatively, we can invert the transform of the option price, which is a function of the transform of the point process and that of the payoff function; see [26]. We may also apply the saddlepoint approximations developed by Glasserman and Kyoung-kuk [21] for affine jump diffusion models. These are applicable here since the transform of $J_{s}$ is an exponentially affine function of the intensity.

The valuation formulae developed above lead to approximate tranche and index rates to the extent that the distribution of the loss process $L$, which does not terminate at the $n$th default in the portfolio, has nonnegligible mass beyond the total portfolio notional $n$. This is mainly relevant for super senior tranches, which depend on the tail of the loss distribution. As pointed out in section 2.2 above, the approximation tends to be accurate for the large diversified reference portfolios typical in practice. ${ }^{1}$ Note also that in this case, the index swap spread can be expressed in terms of $E\left(v \cdot J_{s} \mid \mathcal{F}_{t}\right)$; from formula (16), we then obtain a closed formula for the index rate.

[^60]

Figure 4. Annualized premium for protection against default losses in a portfolio of 100 equally weighted securities as implied by the model specification (1), stated in basis points ( 1 basis point $=10^{-4}$ ). We assume quarterly premium payments and set $\lambda_{0}=c=1$. The risk-free rate $r=5 \%$. The loss at default $\ell_{n}$ is uniformly distributed on $\{0.24,0.96\}$ with expectation $\ell=0.6$. Left panel: Index swap spread as a function of the maturity date $T$ for each of several values of the clustering parameter $\delta$. The decay rate $\kappa=1$. Right panel: 5-year index spread as a function of $\delta$ for each of several values of $\kappa$.
3.2. Numerical examples. To develop some intuition for the model parameters, we provide numerical examples of index and tranche rates. Based on the approximate index swap spread formula, the left panel of Figure 4 shows the index swap spread $S_{0}$ as a function of the maturity date $T$ for each of several values of the clustering parameter $\delta$. For $\delta=0$, the portfolio loss process is a compound Poisson process, which generates a flat spread term structure. All else fixed, the higher $\delta$, the more frequent the event clusters, the higher the spread, and the steeper the term structure. The right panel of Figure 4 shows the $T=5$ year index swap spread as a function of $\delta$ for each of several values of the decay rate $\kappa$. For $\kappa=0$, the portfolio loss process is a compound birth process. All else equal, the higher $\kappa$, the faster the impact of an event decays, and the smaller the chance of large losses.

The tranche swap rate increases with maturity and decreases with seniority. An increase in the lower attachment point, $\underline{K}$, leads to greater subordination (that is, buffer capital) available to the protection seller, who covers only the losses in excess of $\underline{K}$. The left panel of Figure 5 shows the upfront tranche swap rate ${ }^{2} F_{0}$ as a function of the clustering parameter $\delta$ for each of several sets of standard attachment points, assuming the running spread $S_{0}=0$. The sensitivity of the tranche rate to $\delta$ increases with the seniority of the tranche. This is because senior tranches are most exposed to default clustering, the frequency of which is controlled by $\delta$. The bigger $\delta$, the fatter the tail of the portfolio loss distribution. Since relatively few defaults suffice to wipe out the subordinated tranches, the frequency of clusters is less relevant for the pricing of these tranches. The right panel of Figure 5 shows the upfront tranche rate as a function of the volatility (standard deviation) of the loss at default $\ell_{n}$. The sensitivity to the volatility of the loss increases with seniority.

[^61]

Figure 5. Upfront premium rate for protection against default losses in a tranche of a portfolio of 100 equally weighted securities as implied by the model specification (1). We assume the running spread is equal to zero. The maturity $T=5$ years. We set $\lambda_{0}=c=1$ and $\kappa=1$. The risk-free rate $r=5 \%$. Left panel: Tranche upfront rate as a function of the clustering parameter $\delta$ for each of several sets of standard attachment points. The loss at default $\ell_{n}$ is uniformly distributed on $\{0.24,0.96\}$ with expectation $\ell=0.6$. Right panel: Tranche upfront rate as a function of the volatility of the loss at default $\ell_{n}$ for each of several sets of standard attachment points. The expected loss at default is 0.6 . The clustering parameter $\delta=1$.
3.3. Market calibration. We perform market calibration experiments to illustrate the empirical importance of the self-exciting feature. To this end, we fit the parameters of the specification (1) from index and tranche swap market rates for each of the 21 trading days of September 2008. This month witnessed significant volatility due to the demise of Fannie Mae and Freddy Mac on the 8th, the default of Lehman Brothers on the 15th, the collapse of American International Group on the 16th, the problems appearing at Morgan Stanley on the 18th, and the default of Washington Mutual on the 25th. The swap market rates are for the $T=5$ year maturity, and are obtained from UBS. They reference the CDX High Yield portfolio, which consists of 100 equally weighted names of relatively low credit quality. The tranches have attachment points $(0,10 \%),(10 \%, 15 \%)$, $(15 \%, 25 \%)$, and $(25 \%, 35 \%)$. Thus, together with the quote for the index contract, we have five quotes per calibration date, from which we fit the parameter vector $\theta=\left(c, \kappa, \delta, \lambda_{0}\right)$. In accordance with market practice, we assume that the distribution of the loss at default $\nu=\delta_{0.6}$. The risk-free rate of interest $r$ is set to the five year rate quoted on a calibration date (from Bloomberg). Swap premium payments are made quarterly. Using a gradient-based method, we numerically solve the nonlinear optimization problem

$$
\begin{equation*}
\min _{\theta \in \Theta} \sum_{m} \frac{(\operatorname{Mid}(m)-\operatorname{Model}(m, \theta))^{2}}{\operatorname{Mid}(m)} \tag{36}
\end{equation*}
$$

where $\Theta=(0,8] \times[0,10] \times[0,20] \times(0,30]$ and the sum ranges over the contracts. The market midquote $\operatorname{Mid}(m)$ is the arithmetic average of the market bid and ask quotes for index or tranche $m$. The model rate for index or tranche $m$ is denoted $\operatorname{Model}(m, \theta)$ and is given by
the formulae developed in section 3.1. The optimization is initialized at a set of parameter values drawn from a uniform distribution over the parameter space $\Theta$ and is repeated for each of 100 independent draws. The optimal parameter vector $\theta^{*}$ is the solution to (36) with the minimum objective function value among all 100 runs.

The index and tranche swap pricer and the optimization are implemented in MATLAB. The ODEs (13)-(14) determining the point process transform (12) are solved numerically using the Runge-Kutta method. The transform (12) is inverted numerically using the fast Fourier transform. We use 1100 sample points, a value we found to be a good tradeoff between speed and accuracy. To get a sense of the numerical accuracy of this procedure, we contrast the distribution of $N$ generated by the fast Fourier transform with that based on formula (28). We find that the fast Fourier transform produces very accurate results. It is also quicker than the evaluation of formula (25), which requires high-precision algorithms. The computations are performed on a PC with a 2.66 GHz Intel Processor and 4 GB of RAM.

For each of the 21 calibration dates, we calculate the average absolute percentage pricing error $(\mathrm{AAPE})$, given by $(1 / 5) \sum_{m=1}^{5}\left|\operatorname{Model}\left(m, \theta^{*}\right)-\operatorname{Mid}(m)\right| / \operatorname{Mid}(m)$. The left panel of Figure 6 shows the time series of these errors during September 2008. The model fits the market on each date, with an AAPE over all calibration dates of $4.4 \%$ and a variance of 0.086 . For comparison, the market standard copula model completely failed calibration on several days during September 2008. ${ }^{3}$ The calibration errors are relatively low during the most volatile period, which started on September 15 th with the default of Lehman. This indicates that the model captures the default correlation implied by the market prices, especially in periods of extreme stress. The calibrated values of the clustering parameter $\delta$, shown in the right panel of Figure 6, suggest that this may be related to the self-exciting property of the default process. The time series behavior of the calibrated values of $\delta$ clearly reflects the events in mid-September. It indicates the fear of investors of a cluster of events triggered by the default of Lehman on the 15 th (first peak) and the near collapse of Morgan Stanley on the 18th (second peak).
4. Affine point processes. We extend the basic Hawkes model. Our primary objective is to permit a richer structure for the intensity without reducing the computational tractability of the basic specification. The extension is required for applications in which the intensity is influenced by a set of risk factors that follow stochastic processes on their own. In the basic model (1), the point process itself is the only such risk factor. The extension facilitates the inclusion of exogenous (jump) diffusion risk factors that are relevant to arrivals. This is particularly important for empirical applications, in which the intensity model is estimated from a time series of portfolio derivative market prices, as in [3] and [28]. In these applications, the model must replicate the diffusive fluctuation of market prices, and this requires the presence of diffusive risk factors.

The idea behind the extension is to replace the intensity component $\lambda$ of the Markov process $(\lambda, J)^{\top}$ analyzed in Proposition 2.1 by a Markov process $X$ that represents a vector of stochastic risk factors. The process $X$ drives the intensity of the jumps of $J$. The transform of $(X, J)^{\top}$ is computationally tractable if $X$ is taken to be an affine jump diffusion. This

[^62]

Figure 6. Calibration results for the basic intensity model (1) for the 21 trading days of September 2008, for the CDX High Yield portfolio. Left panel: Average absolute percentage error (AAPE) for each calibration date. Right panel: Value of the clustering parameter $\delta$ for each calibration date.
formulation leads to an affine point process.
4.1. Specification. We call a point process affine if its event arrival intensity is an affine function of an affine jump diffusion and its jump sizes are drawn from a fixed distribution. A Markov process $X$ in a state space $D \subset \mathbb{R}^{d} \times \mathbb{R}_{+}$is an affine jump diffusion in the sense of [19] if $X$ is a strong solution to the SDE

$$
\begin{equation*}
d X_{t}=\mu\left(X_{t}, t\right) d t+\sigma\left(X_{t}, t\right) d W_{t}+\sum_{i=1}^{m} \zeta^{i} d Z_{t}^{i}, \quad X_{0} \in \mathbb{R}^{d}, \tag{37}
\end{equation*}
$$

where $W$ is an $\mathbb{R}^{d}$-valued standard Brownian motion, $\mu: D \rightarrow \mathbb{R}^{d}$ is the drift, $\sigma: D \rightarrow \mathbb{R}^{d \times d}$ is the volatility, and each $Z^{i}$ is a temporally consistent $\mathbb{R}_{+}^{d}$-valued point process. In other words, the component processes of each vector $Z^{i}$ share event times and differ only in jump sizes, and we denote their common intensity by $\lambda^{i}\left(X_{t}, t\right)$ for some $\lambda^{i}: D \rightarrow \mathbb{R}_{+}$. The jump sizes are drawn from a distribution $\nu^{i}$ on $\mathbb{R}_{+}^{d}$ that has no mass zero at 0 . Each parameter $\zeta^{i}$ is a $d$-dimensional diagonal matrix. We assume that for each $t,\{x:(x, t) \in D\}$ contains an open subset of $\mathbb{R}^{d}$. For time-dependent coefficient functions that are bounded and continuous on $\mathbb{R}_{+}$, we assume that

$$
\begin{aligned}
\mu(x, t) & =K_{0}(t)+K_{1}(t) x, \quad K_{0}(t) \in \mathbb{R}^{d}, \quad K_{1}(t) \in \mathbb{R}^{d \times d}, \\
\left(\sigma(x, t) \sigma(x, t)^{\top}\right)_{j k} & =\left(H_{0}\right)_{j k}(t)+\left(H_{1}\right)_{j k}(t) \cdot x, \quad H_{0}(t) \in \mathbb{R}^{d \times d}, \quad H_{1}(t) \in \mathbb{R}^{d \times d \times d}, \\
\lambda^{i}(x, t) & =\Lambda_{0}^{i}(t)+\Lambda_{1}^{i}(t) \cdot x, \quad \Lambda_{0}^{i}(t) \in \mathbb{R}, \quad \Lambda_{1}^{i}(t) \in \mathbb{R}^{d}, \quad i=1,2, \ldots, m .
\end{aligned}
$$

Whether an affine point process $J$ driven by $X$ has the self-exciting property depends on the relation between $J$ and its intensity $\lambda_{t}=\lambda\left(X_{t}, t\right)$, where $\lambda: D \rightarrow \mathbb{R}_{+}$. The self-exciting property holds if $X$, and hence $\lambda$, depends on $J$ itself. A sufficient condition is that at least one
of the component processes of $J$ is temporally consistent with one of the component processes of one of the jump terms $Z^{i}$ of $X$. Since the intensity is a function of the realized loss at default, we also generate a dependence structure between default and recovery rates in the self-exciting case.
4.2. Self-exciting examples. To illustrate the specification of an affine point process, we first consider the two-dimensional case $J=(L, N)^{\top}$. $J$ is driven by a one-dimensional risk factor $X$ with a single jump term $Z$ that we identify with the component process $L$. The remaining component process $N$ is temporally consistent with $Z$. Thus, $L$ and $N$ are onedimensional affine point processes that share common event times that arrive with intensity $\lambda\left(X_{t}, t\right)$. By construction, the jump sizes of $L$ are governed by the distribution $\nu$. The jumps of $N$ are unit-sized since $N$ counts the arrivals. The specification of $J$ as an affine point process is completed with the specification of the coefficient functions of the risk factor $X$.

Example 4.1. Suppose $K_{0}(t)=\kappa c$ for $\kappa \geq 0$ and $c>0, K_{1}(t)=-\kappa, H_{0}(t)$ is a matrix of zeros, $H_{1}(t)$ is a tensor of zeros, $X_{0}=c$, and $\zeta=\delta \geq 0$. Let $\Lambda_{0}(t)=0$ and $\Lambda_{1}(t)=1$. Then the intensity $\lambda=X$ of $J$ satisfies the basic model (1):

$$
d \lambda_{t}=\kappa\left(c-\lambda_{t}\right) d t+\delta d L_{t}
$$

A significant generalization of the basic specification is made by introducing Brownian terms and independent jump terms in the intensity. The Brownian terms model the diffusive fluctuation in the default rate and can be driven by a stochastic volatility. The jump terms model the sensitivity of the intensity to market events, such as macroeconomic shocks or defaults to names that are outside the portfolio.

Example 4.2. Suppose the coefficient functions of $X$ are chosen as in Example 4.1, with the exception of the tensor $H_{1}(t)$, for which $\left(H_{1}\right)_{111}(t)=\sigma^{2}$ for $\sigma \geq 0$; all other elements are zero. Then the intensity $\lambda=X$ of $J$ satisfies the SDE

$$
\begin{equation*}
d \lambda_{t}=\kappa\left(c-\lambda_{t}\right) d t+\sigma \sqrt{\lambda_{t}} d W_{t}+\delta d L_{t} \tag{38}
\end{equation*}
$$

showing that between events, the intensity drifts stochastically toward $c$ with diffusive fluctuations driven by $W$. A sample path of the intensity is in Figure 7. Compare with the sample path of the basic model (1) in Figure 2.

In practice, the impact of a default on the other firms may depend on the characteristics of the defaulter. The next example allows us to distinguish between different types of firms represented in the portfolio. We consider a two-dimensional affine point process $J=\left(L^{1}, L^{2}\right)^{\top}$ whose components record portfolio losses triggered by defaults of two firm types. The portfolio loss process $L$ is the sum $L^{1}+L^{2}$. The process $J$ is driven by a two-dimensional risk factor $X=\left(X^{1}, X^{2}\right)^{\top}=\left(\lambda^{1}, \lambda^{2}\right)^{\top}$ with two jump terms $Z^{1}$ and $Z^{2}$ whose components are identified, respectively, with $L^{1}$ and $L^{2}$. This means that both component processes of $Z^{i}$ are indistinguishable from $L^{i}$.

Example 4.3. Suppose

$$
K_{0}(t)=\left(\kappa^{1} c^{1}, \kappa^{2} c^{2}\right)^{\top}, \quad K_{1}(t)=\operatorname{diag}\left(-\kappa^{1},-\kappa^{2}\right)
$$

$H_{0}(t)$ is a matrix of zeros, $H_{1}(t)$ is a tensor of zeros, and

$$
\zeta^{1}=\operatorname{diag}\left(\delta^{1,1}, \delta^{2,1}\right), \quad \zeta^{2}=\operatorname{diag}\left(\delta^{1,2}, \delta^{2,2}\right)
$$



Figure 7. Sample paths of the intensity (38) and the associated loss process L. A jump in the intensity represents the impact of a default. The jump size is the product of the loss at default and the sensitivity parameter $\delta=1$. The loss at default is drawn from an independent uniform distribution on $\{0.4,0.6,0.8,1\}$. The reversion rate $\kappa=5$ and the reversion level $c=\lambda_{0}=0.7$. The volatility $\sigma=0.2$ controls the diffusive fluctuation of the intensity between events.

Let $\Lambda_{0}^{i}(t)=0$ for $i=1,2, \Lambda_{1}^{1}(t)=(1,0)^{\top}$, and $\Lambda_{1}^{2}(t)=(0,1)^{\top}$. Then the intensities $\lambda^{1}=X^{1}$ and $\lambda^{2}=X^{2}$ satisfy the SDE

$$
d \lambda_{t}^{i}=\kappa^{i}\left(c^{i}-\lambda_{t}^{i}\right) d t+\delta^{i, 1} d L_{t}^{1}+\delta^{i, 2} d L_{t}^{2} .
$$

While the parameters $\delta^{1,1}$ and $\delta^{2,2}$ control the self-excitation, the parameters $\delta^{1,2}$ and $\delta^{2,1}$ control the cross-excitation of the processes $L^{1}$ and $L^{2}$.

The preceding example illustrates the significance of allowing $J$ to be vector-valued. The multidimensional setting facilitates the specification of (univariate) point processes whose arrivals are correlated. The dependence between the components of $J$ can be induced by a diffusion risk factor that is common to the component intensities. It can also be generated by direct interaction terms, as in Example 4.3. In the latter case, an arrival of one component process has an impact on the intensities of other component processes. This facilitates the modeling of cross-excitation phenomena. For example, the component processes can model distinct portfolios. A default causes a loss in the respective portfolio and also has an impact on firms in the other portfolios.
4.3. Transform. A general affine point process $J$ is as computationally tractable as the basic Hawkes model (1). To derive a formula for the transform of $J$, we can apply the PIDE arguments developed in Propositions 2.1 and 2.2, which are based on the generator of the Markov process $Y=(X, J)^{\top}$. An alternative approach is to construct $Y$ as an affine jump diffusion process in an enlarged state space and then to apply to this process the transform characterization for affine jump diffusions developed by [19]. This approach, pursued below, allows us to embed our point process model formulation into the affine jump diffusion setting, which is standard in many areas.

Proposition 4.4. Suppose $J$ is a dimension $k$ affine point process driven by a dimension $d$ affine jump diffusion $X$ with state space $D^{X} \times \mathbb{R}_{+} \subset \mathbb{R}^{d} \times \mathbb{R}_{+}$. Then $Y=(X, J)^{\top}$ is a dimension $(k+d)$ affine jump diffusion with state space $D^{Y} \times \mathbb{R}_{+} \subset \mathbb{R}^{k+d} \times \mathbb{R}_{+}$whose coefficients ( $K_{0}, K_{1}, H_{0}, H_{1}, \Lambda_{0}^{i}, \Lambda_{1}^{i}, \zeta^{i}$ ) and jump distributions $\nu^{i}$ are canonically determined by the corresponding items for $X$.

In Proposition 4.4 and below, we use the same symbols for the coefficients of $X$ and $Y$ to avoid further complicating notation. We illustrate the idea in Example 4.3.

Example 4.5. In the setting of Example 4.3, the four-dimensional affine jump diffusion $Y=(X, J)^{\top}=\left(\lambda^{1}, \lambda^{2}, L^{1}, L^{2}\right)^{\top}$ has two (four-dimensional) jump terms $Z^{1}$ and $Z^{2}$ whose components are indistinguishable from $L^{1}$ and $L^{2}$, respectively. Since the third and fourth components of $Y$ are driftless, the third and fourth rows of the drift coefficients are populated with zeros. These coefficients are given by

$$
K_{0}(t)=\left(\kappa^{1} c^{1}, \kappa^{2} c^{2}, 0,0\right)^{\top}, \quad K_{1}(t)=\operatorname{diag}\left(-\kappa^{1},-\kappa^{2}, 0,0\right) .
$$

The volatility coefficients $H_{0}(t)$ and $H_{1}(t)$ are zero since $X$ has no Brownian term. The sensitivity matrices are

$$
\zeta^{1}=\operatorname{diag}\left(\delta^{1,1}, \delta^{2,1}, 1,0\right), \quad \zeta^{2}=\operatorname{diag}\left(\delta^{1,2}, \delta^{2,2}, 0,1\right)
$$

Finally, we give coefficients of the arrival intensities $\lambda_{t}^{i}=\Lambda_{0}^{i}(t)+\Lambda_{1}^{i}(t) \cdot Y_{t}$. We have $\Lambda_{0}^{1}(t)=$ $\Lambda_{0}^{2}(t)=0$ and

$$
\Lambda_{1}^{1}(t)=(1,0,0,0)^{\top}, \quad \Lambda_{1}^{2}(t)=(0,1,0,0)^{\top} .
$$

Proposition 4.4 facilitates the application of Proposition 1 in [19], which expresses the conditional transform of an affine jump diffusion as an exponentially affine function of its current value. Let $\mathbb{C}^{n}$ denote the set of $n$-tuples of complex numbers. When well defined at $t \leq T$ and $u \in \mathbb{C}^{d+k}$, the conditional transform of the $(d+k)$-dimensional affine jump diffusion $Y=(X, J)^{\top}$ constructed in Proposition 4.4 is given by

$$
\begin{equation*}
\psi\left(u, Y_{t}, t, T\right)=E\left(\exp \left(u \cdot Y_{T}\right) \mid \mathcal{F}_{t}\right) \tag{39}
\end{equation*}
$$

Under technical conditions that are stated in the appendix for completeness,

$$
\begin{equation*}
\psi\left(u, Y_{t}, t, T\right)=\exp \left(\alpha(u, t, T)+\beta(u, t, T) \cdot Y_{t}\right) \tag{40}
\end{equation*}
$$

where the coefficient functions $\beta(t)=\beta(u, t, T)$ and $\alpha(t)=\alpha(u, t, T)$ satisfy the ODEs

$$
\begin{align*}
& \partial_{t} \beta(t)=-K_{1}(t)^{\top} \beta(t)-\frac{1}{2} \beta(t)^{\top} H_{1}(t) \beta(t)-\sum_{i=1}^{m} \Lambda_{1}^{i}(t)\left(\theta^{i}\left(\zeta^{i} \beta(t)\right)-1\right),  \tag{41}\\
& \partial_{t} \alpha(t)=-K_{0}(t) \cdot \beta(t)-\frac{1}{2} \beta(t)^{\top} H_{0}(t) \beta(t)-\sum_{i=1}^{m} \Lambda_{0}^{i}(t)\left(\theta^{i}\left(\zeta^{i} \beta(t)\right)-1\right) \tag{42}
\end{align*}
$$

with boundary conditions $\alpha(T)=0$ and $\beta(T)=u$ and jump transforms

$$
\begin{equation*}
\theta^{i}(\omega)=\int_{\mathbb{R}_{+}^{d+k}} e^{\omega \cdot z} d \nu^{i}(z) \tag{43}
\end{equation*}
$$

Under additional technical conditions stated in the appendix for completeness, the transform (40) can be differentiated with respect to $u$ to get the formula

$$
\begin{equation*}
E\left(e^{u \cdot Y_{T}} v \cdot Y_{T} \mid \mathcal{F}_{t}\right)=\psi\left(u, Y_{t}, t, T\right)\left(A(u, v, t, T)+B(u, v, t, T) \cdot Y_{t}\right), \tag{44}
\end{equation*}
$$

where $t \leq T, u \in \mathbb{C}^{d+k}$, and $v \in \mathbb{R}^{d+k}$, and where the coefficient functions $B(t)=B(u, v, t, T)$ and $A(t)=A(u, v, t, T)$ satisfy the ODEs

$$
\begin{align*}
& \partial_{t} B(t)=-K_{1}(t)^{\top} B(t)-\beta(t)^{\top} H_{1}(t) B(t)-\sum_{i=1}^{m} \Lambda_{1}^{i}(t) \nabla \theta^{i}\left(\zeta^{i} \beta(t)\right) \cdot \zeta^{i} B(t),  \tag{45}\\
& \partial_{t} A(t)=-K_{0}(t) \cdot B(t)-\beta(t)^{\top} H_{0}(t) B(t)-\sum_{i=1}^{m} \Lambda_{0}^{i}(t) \nabla \theta^{i}\left(\zeta^{i} \beta(t)\right) \cdot \zeta^{i} B(t) \tag{46}
\end{align*}
$$

with boundary conditions $A(T)=0$ and $B(T)=v$.
As in the case of the Hawkes process, we get a closed formula for $E\left(v \cdot Y_{T} \mid \mathcal{F}_{t}\right)$. To see this, note that the assumption $u=0$ implies that $\alpha=\beta=0$. Therefore, since $\nabla \theta^{i}(0)$ is equal to the $(d+k)$-vector $\ell^{i}$ of expected losses of the $i$ th jump term, (45) and (46) simplify to ${ }^{4}$

$$
\begin{aligned}
& \partial_{t} B(t)=-K_{1}(t)^{\top} B(t)-\sum_{i=1}^{m} \Lambda_{1}^{i}(t) \otimes \ell^{i} \zeta^{i} B(t), \\
& \partial_{t} A(t)=-K_{0}(t) \cdot B(t)-\sum_{i=1}^{m} \Lambda_{0}^{i}(t) \ell^{i^{\top}} \zeta^{i} \cdot B(t) .
\end{aligned}
$$

We conclude that

$$
\begin{align*}
& B(0, v, t, T)=v \exp \left(\int_{t}^{T}\left(K_{1}(s)^{\top}+\sum_{i=1}^{m} \Lambda_{1}^{i}(s) \otimes \ell^{i} \zeta^{i}\right) d s\right),  \tag{47}\\
& A(0, v, t, T)=\int_{t}^{T}\left(K_{0}(s)+\sum_{i=1}^{m} \Lambda_{0}^{i}(s) \ell^{\top} \zeta^{i}\right) \cdot B(s) d s . \tag{48}
\end{align*}
$$

The transform (40) encodes the joint distribution of the risk factor process $X$ and the point process $J$. It facilitates applications that require the calculation of functionals of the form $E\left(h\left(X_{T}, J_{T}\right) \mid \mathcal{F}_{t}\right)$ for suitable functions $h$. An example application is the valuation of forward or option contracts on index and tranche swaps; see [6] and [16]. Other applications include empirical time-series estimation problems as in [3] and the calculation of the future mark-to-market exposure generated by a portfolio derivative position.

We can follow the argument in section 2.6 to characterize in terms of ODEs derived from (41)-(42) the probability distribution of an integer-valued affine point process that may be driven by a multidimensional risk factor process $X$. Alternatively, we can apply Fourier inversion to the transform, which would also cover multidimensional and real-valued affine point processes.

[^63]4.4. Further extensions. The specification can be generalized to include a stochastic discount rate that is driven by the affine jump diffusion $Y$. Since $Y$ includes the affine point process $J$, this specification would generate a dependence structure among default, recovery, and risk-free rates. Another potential extension is the generalization of the driving risk factor process $X$ to a more general Markov process, such as the general affine process analyzed in [18] or the linear-quadratic jump diffusion process analyzed in [10].
5. Conclusion. We study a family of self-exciting point processes for applications in portfolio credit risk. These processes can capture the feedback from default events and the dependence structure between default and recovery rates, both features that are emphasized in the empirical literature on portfolio credit risk. The processes are also computationally tractable, because ODEs characterize their probability distribution. We illustrate this with an application to the valuation of portfolio credit derivatives, which are securities with payoffs that depend on the cumulative loss due to default in a portfolio of corporate bonds or loans. Market calibration experiments demonstrate the fit of a basic specification and highlight the empirical importance of the self-exciting feature.

## Appendix. Proofs.

Proof of Proposition 2.1. Note that $Y$ is a process with right-continuous paths of finite variation. By a change of variables for Stieltjes integrals,

$$
g\left(\lambda_{t}, J_{t}\right)-g\left(\lambda_{0}, J_{0}\right)=\int_{0}^{t} g_{\lambda}\left(\lambda_{s}, J_{s}\right) \kappa\left(c-\lambda_{s}\right) d s+\sum_{0<s \leq t}\left[g\left(\lambda_{s}, J_{s}\right)-g\left(\lambda_{s-}, J_{s-}\right)\right] .
$$

We can write

$$
\sum_{0<s \leq t}\left[g\left(\lambda_{s}, J_{s}\right)-g\left(\lambda_{s-}, J_{s-}\right)\right]=U_{t}^{g}+\int_{0}^{t} \int\left[g\left(\lambda_{s}+\delta z, J_{s}+(z, 1)^{\top}\right)-g\left(\lambda_{s}, J_{s}\right)\right] d \nu(z) \lambda_{s} d s
$$

where $U^{g}$ is a process defined by

$$
U_{t}^{g}=\int_{0}^{t} \int\left[g\left(\lambda_{s-}+\delta z, J_{s-}+(z, 1)^{\top}\right)-g\left(\lambda_{s-}, J_{s-}\right)\right] d \nu(z) d M_{s}
$$

and $M=N-\int_{0} \lambda_{s} d s$ is a martingale. The integrability condition on the predictable integrand guarantees that $U^{g}$ is a martingale; see Theorem 8 in Chapter II of [8]. We conclude that $g(\lambda, J)$ is a special semimartingale with unique decomposition into a sum of a predictable finite variation process and a martingale:

$$
g\left(\lambda_{t}, J_{t}\right)=g\left(\lambda_{0}, J_{0}\right)+\int_{0}^{t}(\mathcal{D} g)\left(\lambda_{s}, J_{s}\right) d s+U_{t}^{g} .
$$

Since $U^{g}$ is a martingale, so is the process defined by $g\left(\lambda_{t}, J_{t}\right)-g\left(\lambda_{0}, J_{0}\right)-\int_{0}^{t}(\mathcal{D} g)\left(\lambda_{s}, J_{s}\right) d s$, and this yields formula (5).

Proof of Proposition 2.2. For fixed $T$ and $u \in \mathbb{C}_{-}^{2}$, define the function $f:[0, T] \times D \rightarrow \mathbb{C}$ by the conditional expectation

$$
f(t, \lambda, x)=E\left(\exp \left(u \cdot J_{T}\right) \mid \lambda_{t}=\lambda, J_{t}=x\right)
$$

Then $E\left(\exp \left(u \cdot J_{T}\right) \mid \mathcal{F}_{t}\right)=f\left(t, \lambda_{t}, J_{t}\right)$. Since for $u \in \mathbb{C}_{-}^{2}$ the real and imaginary parts of the function $g(\lambda, x)=\exp (u \cdot x)$ are in the domain of the generator $\mathcal{D}$, so are the real and imaginary parts of the function $f(t, \lambda, x)$ for each time $t$. The function $f(t, \lambda, x)$ is also differentiable with respect to $t$, and the derivative is continuous; see Proposition VII.1.2 in [31]. For this, we note that $Y=(\lambda, J)^{\top}$ is a Feller process.

Now, for $\left(f\left(t, \lambda_{t}, J_{t}\right)\right)_{t \leq T}$ to be a martingale, we must have that

$$
\left\{\begin{array}{l}
f_{t}(t, \lambda, x)+(\mathcal{D} f)(t, \lambda, x)=0,  \tag{49}\\
f(T, \lambda, x)=\exp (u \cdot x),
\end{array}\right.
$$

where $(\mathcal{D} f)(t, \lambda, x)$ is obtained by applying $\mathcal{D}$ to the real and imaginary parts of $f(t, \cdot, \cdot)$. We propose the following change of variables: $w(t, \lambda, u)=f(t, \lambda, x) e^{-u \cdot x}$. Then, after rearranging terms and denoting $\bar{z}=(z, 1)^{\top}$, (49) takes the form

$$
\left\{\begin{array}{l}
w_{t}(t, \lambda, u)+\lambda \int[w(t, \lambda+\delta z, u)-w(t, \lambda, u)] e^{u \cdot \bar{z}} d \nu(z) \\
+\lambda \int\left(e^{u \cdot \bar{z}}-1\right) w(t, \lambda, u) d \nu(z)+\kappa(c-\lambda) w_{\lambda}(t, \lambda, u)=0, \\
w(T, \lambda, u)=1 .
\end{array}\right.
$$

We take $w(t, \lambda, u)=e^{a(t)+b(t) \lambda}$. Then,

$$
\lambda \partial_{t} b+\partial_{t} a+\lambda \int\left(e^{\delta b z}-1\right) e^{u \cdot \bar{z}} d \nu(z)+\lambda \int\left(e^{u \cdot \bar{z}}-1\right) d \nu(z)+\kappa(c-\lambda) b=0
$$

so that

$$
\left\{\begin{array}{l}
\partial_{t} b=\kappa b-\int\left(e^{\delta b z+u \cdot \bar{z}}-1\right) d \nu(z), \\
\partial_{t} a=-\kappa c b
\end{array}\right.
$$

with boundary conditions $a(T)=b(T)=0$.
Proof of Proposition 4.4. The process $Y=(X, J)^{\top}$ is Markov since the intensity of $J$ depends only on the current state of $X$, which is Markov, and the jump distribution of $J$ is fixed. We can append the components of $J$ one at a time. Therefore, it is sufficient to consider the case where $J$ is one-dimensional. By assumption, its arrival intensity $\lambda$ can be expressed as a time-dependent affine function $\lambda\left(X_{t}, t\right)=\Lambda_{0}(t)+\Lambda_{1}(t) \cdot X_{t}$ of $X_{t}$.

The first step is to extend the $d$-dimensional jump processes $Z^{i}$ by one dimension. Suppose the affine point process $J$ is self-affecting so that it is temporally consistent with some $Z^{i}$. Renumbering if necessary, assume that $i=1$ and extend $Z^{1}$ by appending the component process $J$. We retain the symbol $\nu^{1}$ for the distribution on $\mathbb{R}^{d+1}$ that governs the jump sizes of $Z^{1}$. The extensions of the remaining $Z^{i}$ s are made by appending temporally consistent counting processes. In other words, the jumps in the $(d+1)$ st component of $Z^{2}, Z^{3}, \ldots, Z^{m}$ are of unit size. It follows that the (degenerate) jump distributions $\nu^{i}$ on $\mathbb{R}^{d+1}$ are canonically determined for $i \geq 2$.

Next, we specify the affine jump diffusion coefficients of $Y$. Since the intensity $\lambda^{i}$ of $Z^{i}$ is an affine function of $X$, it can be expressed as an affine function of $Y$. The onedimensional coefficients $\Lambda_{0}^{i}$ remain the same, and the $d$-dimensional vectors $\Lambda_{1}^{i}$ are extended to dimension $(d+1)$ with a zero. The dimension $(d+1)$ sensitivity matrix of $Y$ to $Z^{1}$ is given by $\zeta^{1}=\operatorname{diag}(0,0, \ldots, 0,1)$. We extend the $d \times d$ sensitivity matrices $\zeta^{2}, \zeta^{3}, \ldots, \zeta^{m}$ by adding one additional row and column with all entries set to zero.

If the affine point process $J$ is not temporally consistent with any $Z^{i}$, then all of the $X^{i} \mathrm{~s}$ are extended by counting processes and all the (degenerate) $\nu^{i}$ s are canonically determined. Here, we add an $(m+1)$ st jump term given by $(d+1)$ copies of $J$. The intensity $\lambda^{m+1}=\lambda$, and so $\Lambda_{0}^{m+1}(t)=\Lambda_{0}(t)$, and the $(d+1)$ vector $\Lambda_{1}^{m+1}(t)$ is obtained from the $d$ vector $\Lambda_{1}(t)$ by appending a zero. Each of the $m$ sensitivity matrices $\zeta^{i}$ is extended from dimension $d$ to dimension one with a row and column of zeros. The $(d+1)$-dimensional sensitivity matrix $\zeta^{m+1}=\operatorname{diag}(0,0, \ldots, 0,1)$.

Since $J$ is a pure jump process, the drift coefficients $K_{0}$ and $K_{1}$ and the covariance matrix coefficients $H_{0}$ and $H_{1}$ are obtained from the corresponding coefficients for $X$ by extending with rows and columns of zeros.

Based on Proposition 4.4, we apply [19, Proposition 1], which implies that the conditional transform of $Y=(X, J)^{\top}$ is given by (40) if the coefficient functions $\beta$ and $\alpha$ uniquely solve the differential equations (41) and (42) and if for $(u, T) \in \mathbb{C}^{d+k} \times \mathbb{R}_{+}$the following expectations are all finite:
(1) $E\left(\left|\Psi_{T}\right|\right)$, where $\Psi_{t}=\exp \left(\alpha(t)+\beta(t) \cdot Y_{t}\right)$.
(2) $E\left(\left(\int_{0}^{T} \eta_{t} \cdot \eta_{t} d t\right)^{1 / 2}\right)$, where $\eta_{t}=\Psi_{t} \beta(t)^{\top} \sigma\left(Y_{t}, t\right)$.
(3) $E\left(\int_{0}^{T}\left|\gamma_{t}\right| d t\right)$, where $\gamma_{t}=\Psi_{t} \sum_{i=1}^{m}\left(\theta^{i}\left(\zeta^{i} \beta(t)\right)-1\right) \lambda^{i}\left(Y_{t}, t\right)$.

Similarly, based on Proposition 4.4, we apply [19, Proposition 3], which implies that the conditional expectation of $e^{u \cdot Y_{T}}\left(v \cdot Y_{T}\right)$ is given by formula (44) if the coefficient functions $\beta$ and $\alpha$ uniquely solve the differential equations (41) and (42), the coefficient functions $B$ and $A$ uniquely solve the differential equations (45) and (46), and if for $(u, v, T) \in \mathbb{C}^{d+k} \times \mathbb{R}^{d+k} \times \mathbb{R}_{+}$ the following expectations are all finite:
(1) $E\left(\left|\Phi_{T}\right|\right)$, where $\Phi_{t}=\Psi_{t}\left(A(t)+B(t) \cdot Y_{t}\right)$.
(2) $E\left(\left(\int_{0}^{T} \bar{\eta}_{t} \cdot \bar{\eta}_{t} d t\right)^{1 / 2}\right)$, where $\bar{\eta}_{t}=\Phi_{t}\left(\beta(t)^{\top}+B(t)^{\top}\right) \sigma\left(Y_{t}, t\right)$.
(3) $E\left(\int_{0}^{T}\left|\bar{\gamma}_{t}\right| d t\right)$, where $\bar{\gamma}_{t}=\sum_{i=1}^{m} \lambda^{i}\left(Y_{t}, t\right)\left(\Phi_{t}\left(\theta^{i}\left(\zeta^{i} \beta(t)\right)-1\right)+\Psi_{t} \nabla \theta^{i}\left(\zeta^{i} \beta(t)\right) \cdot \zeta^{i} B(t)\right)$.

Acknowledgments. We thank Andrew Abrahams, Benjamin Armbruster, Damiano Brigo, Peter Carr, Valdo Durrleman, Andreas Eckner, Bjorn Flesaker, Igor Halperin, Steven Hutt, Peter Jäckel, Rajnish Kamat, Andrei Lopatin, Woujiang Lou, Timur Misirpashaev, George Papanicolaou, Vijay Poduri, David Soronow, Tom Stone, Pascal Tomecek, Nicolas Victoir, Stefan Weber, and Ulrich Zier for helpful discussions and comments. We also thank two anonymous referees for their helpful comments. We are extremely grateful for excellent research assistance from Xiaowei Ding, Baeho Kim, Jack Kim, and Kao-chih Syao.

## REFERENCES

[1] E. Altman, B. Brady, A. Resti, and A. Sironi, The link between default and recovery rates: Theory, empirical evidence and implications, J. Business, 78 (2005), pp. 2203-2227.
[2] M. Arnsdorf and I. Halperin, BSLP: Markovian bivariate spread-loss model for portfolio credit derivatives, J. Comput. Finance, 12 (2008), pp. 77-100.
[3] S. Azizpour, K. Giesecke, and B. Kim, Premia for correlated default risk, Working paper, Stanford University, Stanford, CA, 2008.
[4] S. Azizpour, K. Giesecke, and G. Schwenkler, Exploring the sources of default clustering, Working paper, Stanford University, Stanford, CA, 2008.
[5] Basel Commission on Bank Regulation, International convergence on capital measurement and capital standards, Bank for International Settlements, 2004.
[6] E. Bayraktar and B. Yang, Multi-scale time-changed birth processes for pricing multi-name credit derivatives, Appl. Math. Finance, 16 (2009), pp. 429-449.
[7] C. Bowsher Modelling security market events in continuous time: Intensity based, multivariate point process models, J. Econometrics, 141 (2007), pp. 876-912.
[8] P. Brémaud, Point Processes and Queues - Martingale Dynamics, Springer-Verlag, New York 1981.
[9] D. Brigo, A. Pallavicini, and R. Torresetti, Calibration of CDO tranches with the dynamical generalized Poisson loss model, Risk, 20 (2007), pp. 70-75.
[10] P. Cheng and O. Scaillet, Linear-quadratic jump-diffusion modeling, Math. Finance, 17 (2007), pp. 575-598.
[11] P. Collin-Dufresne, R. Goldstein, and J. Helwege, How large can jump-to-default risk premia be? Modeling contagion via the updating of beliefs, Working paper, Columbia University, New York, NY, 2009.
[12] R. Cont and A. Minca, Extracting portfolio default rates from CDO spreads, Working paper, Columbia University, New York, NY, 2008.
[13] D. Daley and D. Vere-Jones, An Introduction to the Theory of Point Processes, Volume I, SpringerVerlag, New York, 2003.
[14] S. Das, D. Duffie, N. Kapadia, and L. Saita, Common failings: How corporate defaults are correlated, J. Finance, 62 (2007), pp. 93-117.
[15] M. Davis and V. Lo, Modeling default correlation in bond portfolios, in Mastering Risk Volume 2: Applications, C. Alexander, ed., Prentice Hall, Englewood Cliffs, NJ, 2001, pp. 141-151.
[16] X. Ding, K. Giesecke, and P. Tomecek, Time-changed birth processes and multi-name credit derivatives, Oper. Res., 57 (2009), pp. 990-1005.
[17] D. Duffie, A. Eckner, G. Horel, and L. Saita, Frailty correlated default, J. Finance, 64 (2009), pp. 2089-2123.
[18] D. Duffie, D. Filipovic, and W. Schachermayer, Affine processes and applications in finance, Ann. Appl. Probab., 13 (2003), pp. 984-1053.
[19] D. Duffie, J. Pan, and K. Singleton, Transform analysis and asset pricing for affine jump-diffusions, Econometrica, 68 (2000), pp. 1343-1376.
[20] K. Giesecke, L. Goldberg, and X. Ding, A top down approach to multi-name credit, Oper. Res., to apear.
[21] P. Glasserman and K.-K. Kim, Saddlepoint approximations for affine jump diffusion models, J. Economic Dynamics and Control, 33 (2009), pp. 37-52.
[22] A. G. Hawkes Spectra of some self-exciting and mutually exciting point processes, Biometrika, 58 (1971), pp. 83-90.
[23] A. G. Hawkes and D. Oakes, A cluster process representation of a self-exciting process, J. Appl. Probab., 11 (1974), pp. 493-503.
[24] P. Jorion and G. Zhang, Good and bad credit contagion: Evidence from credit default swaps, J. Financial Economics, 84 (2007), pp. 860-883.
[25] E. M. Klimko, An algorithm for calculating indices in Faa di Bruno's formula, BIT, 13 (1973), pp. 38-49.
[26] R. W. Lee, Option pricing by transform methods: Extensions, unification, and error control, J. Comput. Finance, 7 (2004), pp. 51-86.
[27] F. Longstaff, The subprime credit crisis and contagion in financial markets, J. Finance, to appear.
[28] F. Longstaff and A. Rajan, An empirical analysis of collateralized debt obligations, J. Finance, 63 (2008), pp. 529-563.
[29] A. Lopatin and T. Misirpashaev, Two-dimensional Markovian model for dynamics of aggregate credit loss, Adv. Econometrics, 22 (2008), pp. 243-274.
[30] P. Protter, Stochastic Integration and Differential Equations, Springer-Verlag, New York, 2004.
[31] D. Revuz and M. Yor, Continuous Martingales and Brownian Motion, Springer-Verlag, Heidelberg, 2005.

# Real Options Games in Complete and Incomplete Markets with Several Decision Makers* 

Alain Bensoussan ${ }^{\dagger}$, J. David Diltz ${ }^{\ddagger}$, and SingRu Hoe ${ }^{\S}$


#### Abstract

We consider optimal investment policies for irreversible capital investment projects under uncertainty in a monopoly situation and in a Stackelberg leader-follower game. We consider two types of payoffs: lump-sum and cash flows. The decisions are the times to enter into the market. The problems belong to the class of optimal stopping times, for which the right approach is that of variational inequalities (V.I.s). In the case of complete markets, payoffs are expected values with respect to the risk-neutral probability. In the case of incomplete markets, the risk-neutral probability is not defined. We consider an investor maximizing his/her utility function, and we consider the investment in the project as an additional decision, besides portfolio investment and consumption decisions. This decision remains a stopping time, conversely to the portfolio investment and consumption decisions (continuous controls). The game problem raises new difficulties. The leader's V.I. has a nondifferentiable obstacle. The weak formulation of the V.I. handles this difficulty. In some cases, the solution of the V.I. may be continuously differentiable although the obstacle is not. An additional difficulty occurs for lump-sum payoffs in the case of incomplete markets. We cannot compare gains and losses at different times. We propose an alternative approach, using equivalence (indifference) considerations. In the case of payoffs characterized by cash flows, this difficulty does not exist, but an intermediary problem arises which has a nice interpretation as a differential game. The solutions thus obtained for the Stackelberg game are not intuitive. Therefore, competition has important consequences on investment decisions.


Key words. variational inequality, utility-based pricing, optimal stopping, real options
AMS subject classifications. 60G40, 62L15
DOI. 10.1137/090768060

1. Introduction. Advances in the development and implementation of options models to value and manage capital investment projects have progressed at a prodigious rate following Stewart Myers's 1977 assertion that a firm's growth opportunities may be viewed as "real options." Early milestones in real options development included Brennan and Schwartz [5], McDonald and Siegel [34], and Paddock, Siegel, and Smith [44]. Brennan and Schwartz [5] applied option pricing to valuation and operation of a copper mine. The owner of an active mine holds a put option to suspend operation should copper prices fall below a certain threshold, while the owner of a suspended mine has a call option to reopen should prices rise above a

[^64]different, higher threshold. Fixed suspension and resumption costs drive a wedge between the thresholds, creating a "hysteresis" effect. Free boundary conditions determine threshold prices.

Classic real options models did not address the possibility that the investment strategy exercised by one firm could have an impact on the optimal investment policies of a competitor. If a firm fears pre-emption, the delay option value is less significant, and project value approaches the traditional net present value. Notable contributions in this area include papers by Grenadier [15], [16], [17]. Grenadier [15] applied option game theory to the real estate market to explain development cascades and overbuilding. Extensions to the analysis (see Grenadier [16], [17]) consider equilibrium strategies with option exercise games. ${ }^{1}$ While preemption fears erode delay option value, Dixit and Pindyck [10, p. 314] note that the situation is different if the roles of leader and follower are exogenously determined as in the Stackelberg game. In this case, the leader has the ability to wait, recognizing the option value. They discuss the possible option values and investment strategies in narrative terms without rigorous proof. We address the problem rigorously and obtain the optimal strategies for the leader and the follower in both complete and incomplete market cases. The latter case is particularly important for real options for the reasons discussed below.

Capital investment projects are typically nontraded. That is, claims to these assets are not traded in well-developed secondary markets. The result is that real options markets are incomplete. Nevertheless, many articles simply assume markets are complete; i.e., all risk can be hedged completely. In some models the static NPV proxies for the nontraded underlying asset, a strategy that Copeland and Antikarov [7] call the "marketed asset disclaimer." Brennan and Schwartz [3], [4] and Dixit and Pindyck [10] use a capital asset pricing model, and then they estimate a risk-adjusted project return rate and solve the option value as a threshold optimization. Financial mathematicians address the issue by employing strategies such as (1) minimizing tracking error (see, for example, Duffie and Richardson [12]); (2) selecting a martingale measure for pricing using minimal martingale or minimal entropy methods (see Follmer and Sonderman [13], Schweizer [45], Delbaen and Schachermayer [9], and Frittelli [14]); and (3) employing utility functions to estimate an indifference price (see Henderson [20], [21], Henderson and Hobson [23], Musiela and Zariphopoulou [39], [41], Oberman and Zariphopoulou [43], and Hodges and Neuberger [26]). For incomplete markets, we adopt in this paper the utility-based approach, considering utility maximization with joint decisions of stopping times, portfolio investment strategies, and/or consumption rules.

We consider a duopoly model of Stackelberg leader-follower type in an irreversible capital investment under uncertainty. The decision can be made over an infinite horizon with the constraint that the follower be forbidden to invest until the leader has already done so. Leader and follower roles are predetermined. The Stackelberg game has been applied in many areas, including electricity pricing (see Ho, Luh, and Muralidharan [25]), NOx allowances in the electric power industry (see Chen et al. [6]), supply chain management and marketing channels (see He et al. [19]), server-proxy-user systems in computer science (see Han and Xia [18]), and others.

Our framework assumes that uncertainties embedded in the irreversible capital investment

[^65]project arise from the uncertainties of investment payoffs. We consider two types of investment payoffs: lump-sum investment payoffs and investment payoffs as a series of cash flows.

The primary advantage to modeling uncertainty as a lump-sum investment payoff lies in its simplicity. As noted by Dixit and Pindyck [10], modeling uncertainty in terms of project value in many cases leads to the same optimal policies as modeling uncertainty in cash flows. ${ }^{2}$ On the other hand, focusing on cash flows permits comparison between results of our model and more conventional decision making paradigms. We model the uncertainty of lump-sum investment payoffs by assuming that the investment project value is governed by an externally determined geometric Brownian motion. To capture the uncertainty from cash flows, we consider project cash flows to evolve according to an arithmetic Brownian motion. The possibility of negative cash flows deals explicitly with the fact that the project's operations may generate losses. The option we are interested in is the time to invest (when to enter into the market) for both the follower and the leader. These problems are optimal stopping time problems. From Bensoussan and Lions [1], it is known that the right tool to solve these problems is variational inequalities (V.I.s), introduced by Stampacchia in 1964 and Lions and Stampacchia [33] in the context of physical and mechanical applications. The definition of payoffs is easy in the complete market case. There exists a unique risk-neutral probability. So the payoffs are mathematical expectations with respect to this probability. For the follower, as in the case of a single decision maker, the V.I. has an obstacle function (the outcome received at the stopping time to be decided) which is $C^{1}$ (continuously differentiable). There is then a strong formulation for the V.I., and the solution can be obtained by a threshold strategy. This is not the case for the leader's problem. The obstacle function is $C^{0}$, not $C^{1}$. We need to use a weak formulation for the V.I. We prove the existence and uniqueness of the solutions in the weak form. We then study the regularity of the solution and find that it is indeed $C^{1}$ and piecewise $C^{2}$. So the solution of the V.I. is more regular than the obstacle. However, the optimal strategy is not given by a single threshold. This is related to the lack of smoothness of the obstacle. The optimal strategy is obtained by two intervals. For the cash flow model characterized by an arithmetic Brownian motion process, we encounter the additional difficulty of an unbounded obstacle function.

In the incomplete market case, we use a utility-based approach for the valuation. The manager's (investor's) risk preferences are modeled through an exponential utility function. We consider the manager's (investor's) utility maximization with joint decisions of stopping times, portfolio investment strategies, and/or consumption rules. For the case of lump-sum investment payoffs, we do not allow intermediate consumption in order to simplify exposition and to avoid the penalization arising from the case of negative consumption. The manager (investor) maximizes his/her expected utility of wealth taking portfolio investment strategies and stopping times into consideration. The follower's and the single decision maker's value functions are solutions of V.I.s, which can be written in a strong formulation, and the optimal strategies are obtained through a threshold approach. The situation is much more complex for the leader's problem since we are unable to compare gains and losses at different times as in the complete market case. We circumvent the problem by employing equivalence (indifference)

[^66]considerations as a surrogate of the present value used as a comparison benchmark in the complete market situation. In addition, we encounter the same mathematical difficulty as in the complete market case, namely, that the leader's obstacle function is nonsmooth. We thus formulate and solve V.I.s in the weak sense. However, we have neither the situation of a more regular solution nor of a two-interval strategy. We comment on this problem, the full solution of which is open.

For the case of investment payoffs as a series of cash flows, the manager (investor) faces a utility maximization problem in which he/she chooses a portfolio investment strategy, a consumption rate, and a stopping time for market entry (capital investment). The problem can be split into two parts. We first assume that the investment project has already been undertaken. This leads to a control problem of portfolio selection and consumption rate, augmented by the stochastic income stream from the capital project now in place. We then consider the problem of choosing the optimal stopping time, portfolio selection, and consumption rules before the optimal stopping. We get a V.I. with a nonlinear differential operator. The obstacle is defined by the first part. The solution to this V.I. has an interesting interpretation, that of the stochastic differential game, different from the Stackelberg game. For the follower and the single decision maker, we prove the existence and uniqueness of the solution of the V.I., which is $C^{1}$ and piecewise $C^{2}$. For the leader's problem, we first characterize his/her payoff at the optimal stopping time. The leader must take into account the effect of the follower's anticipated entry on his/her portfolio investment and consumption decisions. Again, we encounter a nonsmooth obstacle function, and we must interpret the V.I. in the weak sense. We study the regularity of the solution, employing penalization, and here the challenge is to find suitable a priori estimates. We prove the existence and uniqueness of the solutions in their strong forms. The leader's optimal stopping rule is given by a two-interval strategy.

The remainder of this paper is organized as follows. In section 2, we briefly describe general features of problems and models studied. In sections 3 and 4, we present models and generate results for the case of lump-sum investment payoffs, respectively, for the complete market and the incomplete market cases. In sections 5 and 6 , we present models and results for the case when the investment payoff is characterized as a series of cash flows. We present concluding remarks in section 7 .

## 2. General features of problems and models.

2.1. Asset valuation in complete and incomplete markets. The issue of valuation is key in finance. We have to assign a "fair" value to random outcomes, often called "contingent claims." The expected value is not a fair value because risk is not taken into consideration. The essential property of complete markets is that there exists one and only one fair value. It is the expected value of the random outcome with respect to a new probability measure, called the risk-neutral probability. This risk-neutral probability is uniquely defined. So in the case of complete markets we will measure the value of losses and gains by taking simply the expected value with respect to the risk-neutral probability. For our later purpose, the framework is as follows. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{Q})$ with a Wiener process $W(t)$. We assume that the market is characterized by a single asset, $S(t)$, whose evolution is governed by

$$
\begin{equation*}
d S(t)=r S(t) d t+\sigma S(t)(\lambda d t+d W(t)) \tag{2.1}
\end{equation*}
$$

where $r$ is a risk-free interest rate, $\sigma$ is the volatility, and $\lambda$ is the Sharpe ratio; they are all constants. The market is complete, and there exists a unique risk-neutral probability. It is obtained as follows: we define

$$
\widehat{W}(t)=W(t)+\lambda t
$$

and the process $Z(t)$ by

$$
d Z(t)=-\lambda Z(t) d W(t), \quad Z(0)=1 .
$$

The risk-neutral probability measure $\widehat{\mathbb{Q}}$ is then

$$
\left.\frac{d \widehat{\mathbb{Q}}}{d \mathbb{Q}}\right|_{\mathcal{F}_{t}}=Z(t)
$$

with $\mathcal{F}_{t}=\sigma(W(s), s \leq t)$. Under $\widehat{\mathbb{Q}}, \widehat{W}(t)$ is a standard Wiener process. The fair value of a contingent claim is simply the expected value of the random outcome with respect to the risk-neutral probability.

Let us now consider the situation of incomplete markets. The risk-neutral probability is not unique. The classical valuation approach is inappropriate, and alternatives must be considered. One proposed solution is to introduce a rational utility-maximizing investor who evaluates unhedgeable risk based on the investor's risk preferences. Utility-based valuation in stochastic dynamic market environments derives from the famous work of Merton [35], who developed the original dynamic, stochastic model of expected utility maximization. We adopt the utility-based valuation in our incomplete market model. We assume an exponential utility function to describe the investor's risk preferences. As noted in Musiela and Zariphopoulou [41], there are advantages working with an exponential utility function. For example, the asset value is given in terms of a nonlinear pricing rule that has certainty equivalent characteristics. However, this nonlinear pricing functional is not the static analogue of the certainty equivalent corresponding to the exponential preferences. It is distorted, with a magnitude that depends only on the correlation between the traded and the nontraded assets. Another advantage is that the measure under which the indifference price is computed is a measure under which the traded asset price is a martingale and which has the minimal entropy with respect to the historical one.
2.2. Structures of the market. In addition to the consideration of market completeness, we consider two different types of market structures: one market player or two market players. Market players' decisions are the time to enter into the market. In the case of one market player, the entrepreneur may operate without consideration of any potential competitor's decisions (i.e., a monopolist in the product market). By paying an investment cost $K$ at the market entry time $\tau$, the firm expects to receive the whole operation income.

In the case of two market players, we consider a Stackelberg leader-follower game. The roles of leader and follower have been predetermined, for example, by regulations, by competitive advantages of market powers, etc. The follower is forbidden to invest until the leader has already done so. The leader enters the market at time $\theta$ and the follower at time $\tau \geq \theta$. Each player pays an investment cost $K$ upon entry. The leader receives the whole operation income prior to the follower's entry. Upon the follower's entry, depending on the competitiveness or the regulation, the leader may share the market with the follower evenly or the leader may maintain larger portions of the operation income, leaving smaller portions to the follower.
2.3. Lump-sum and cash flow models. We consider two types of operation incomes from the capital investment undertaken. One is the operation income characterized by a lump-sum, which means that the investment project yields a one time payoff at the time of investment for a given cost $K$. We model this situation in terms of the investment project value, governed by a geometric Brownian motion process. In this model, the leader has to incorporate into his/her objective function his/her surrender project value to the follower upon the follower's optimal entry, so one has to compare values of events taking place at different times. In the case of complete markets, the payoff is the expected value with respect to the unique risk-neutral probability, so gains/losses at different times are comparable (by expected present values with respect to the risk-neutral probability). In the case of incomplete markets, we encounter a major hurdle for the leader's value function. Due to the absence of the unique martingale measure, we choose to work with the investor's utility, and there is no identity relationship for values at different times. We circumvent the problem by employing equivalence (indifference) considerations. However, we then encounter another mathematical difficulty. The "obstacle" function is nonsmooth and, in this case, we are unable to get the result that the solution of the V.I. is $C^{1}$ (i.e., smoother than the obstacle) as in the case of complete markets. By exploring with an explicit calculation for a two-interval solution, we arrive at only a partial answer.

In the other model, the operation income is characterized by cash flows, which means that by paying a given cost $K$ at the investment time the entrepreneur receives a series of cash flows thereafter. The cash flow evolves as an arithmetic Brownian motion process. Unlike the lump-sum model, the problem of comparing gains and losses at different times does not arise in this model, in either the complete or the incomplete case, because it is a cash flow valuation.

We present a detailed problem formulation, models, and valuation procedures for the case of lump-sum investment payoffs in sections 3 and 4, and for the case of the investment payoff characterized as a series of cash flows in sections 5 and 6.

## 3. Lump-sum payoffs and complete market assumption.

3.1. Single player. We assume that the investment opportunity is driven by a stochastic value process:

$$
\begin{align*}
d V(s) & =r V(s) d s+\eta V(s)(\xi d s+d W(s))  \tag{3.1}\\
& =(r+\eta(\xi-\lambda)) V(s) d s+\eta V(s) d \widehat{W}(s) ; \quad s \geq t ; \quad V(t)=v \tag{3.2}
\end{align*}
$$

where $\eta$ and $\xi$ are constants. We denote the solution of (3.2) by $V_{v, t}(s), s \geq t$. Beginning at $t$, the firm undertakes the capital investment project at time $\tau>t$, at which point the firm invests at a cost $K$ in return for the project value $(1-a) V_{v, t}(\tau)$, where $a \in[0,1) .^{3}$ The firm's problem is to find an optimal stopping time to maximize the expected discounted project value:

$$
\begin{equation*}
F(v, t)=\sup _{\tau \geq t} \widehat{E}\left[e^{-r(\tau-t)}\left((1-a) V_{v, t}(\tau)-K\right) \mathbb{1}_{\tau<\infty}\right] \tag{3.3}
\end{equation*}
$$

[^67]where $\widehat{E}[\cdot]$ is the expectation with respect to the risk-neutral probability measure $\widehat{\mathbb{Q}}$.
This is a stationary problem; hence the value function (3.3) becomes
\[

$$
\begin{equation*}
F(v)=\sup _{\tau \geq 0} \widehat{E}\left[e^{-r \tau}\left((1-a) V_{v}(\tau)-K\right) \mathbb{1}_{\tau<\infty}\right], \quad \text { where } \quad V_{v}(t)=V_{v, 0}(t) \tag{3.4}
\end{equation*}
$$

\]

Clearly, since $\tau=\infty$ is possible (i.e., the firm never invests), $F(v) \geq 0$. Assuming that the function $F(v)$ is sufficiently smooth, $F(v)$ solves the following V.I. as a consequence of dynamic programming:

$$
\left\{\begin{array}{l}
F(v) \geq((1-a) v-K),  \tag{3.5}\\
F^{\prime}(v) v(r+\eta(\xi-\lambda))+\frac{1}{2} F^{\prime \prime}(v) v^{2} \eta^{2}-r F(v) \leq 0, \\
{[F(v)-((1-a) v-K)]\left[F^{\prime}(v) v(r+\eta(\xi-\lambda))+\frac{1}{2} F^{\prime \prime}(v) v^{2} \eta^{2}-r F(v)\right]=0,} \\
F(0)=0 ; F(v) \geq 0 ; F(v) \text { has linear growth at infinity. }
\end{array}\right.
$$

From Chapter 5 in Dixit and Pindyck [10], we have the following theorem.
Theorem 3.1. Assume $\xi<\lambda$ :

$$
F(v)= \begin{cases}\frac{K}{\beta-1}\left(\frac{v}{\hat{v}}\right)^{\beta}, & v \leq \hat{v},  \tag{3.6}\\ (1-a) v-K, & v \geq \hat{v},\end{cases}
$$

where $\beta=\frac{1}{2}-\frac{r+\eta(\xi-\lambda)}{\eta^{2}}+\sqrt{\left(\frac{1}{2}-\frac{r+\eta(\xi-\lambda)}{\eta^{2}}\right)^{2}+\frac{2 r}{\eta^{2}}}>1$, and $\hat{v}=\frac{\beta K}{(1-a)(\beta-1)}$.
A proof is needed to check that $F(v)$ in (3.6) is the solution of the V.I. since a threshold is not necessarily a solution of the V.I. We omit the proof.

The optimal stopping rule which achieves the supremum in (3.4) is $\hat{\tau}(v)=\inf \left\{t \mid V_{v}(t) \geq \hat{v}\right\}$. Using $V_{v}(\hat{\tau}(v))=\hat{v}$ if $\hat{\tau}(v)<\infty$ and $v<\hat{v}$, we have the probabilistic representation

$$
\begin{equation*}
F(v)=((1-a) \hat{v}-K) \widehat{E}[\exp \{-r \hat{\tau}(v)\}] \quad \text { if } v<\hat{v} . \tag{3.7}
\end{equation*}
$$

3.2. Two players: A Stackelberg game. We now consider a two firm (i.e., two player) Stackelberg game. The roles of leader and follower have been predetermined. Each firm may invest in the capital project, but the actions, actual and/or anticipated, of one player affect the other's decision. Both players face an optimal stopping problem similar to the monopolist's, albeit with the complications inherent in the leader-follower framework. The follower is forbidden to invest until the leader has already done so. The leader enters the market at time $\theta$ and the follower at time $\tau \geq \theta$, both stopping times of the filtration $\mathcal{F}_{t}$. Again, each player pays an investment cost $K$ upon entry. The leader knows that the follower, acting rationally, will enter at time $\hat{\tau}_{\theta}$, at which time he/she must surrender a portion of the project to the follower.

### 3.2.1. Statement of the leader's problem. We first notice that

$$
\begin{equation*}
\hat{\tau}_{\theta}=\theta+\hat{\tau}\left(V_{v}(\theta)\right) \tag{3.8}
\end{equation*}
$$

with a slight abuse of notation. We consider the function $\hat{\tau}(v)$, in which $v$ is deterministic. It is an entry time for the process $V_{v}(t)$. So it depends on $v$ and on all the values of the process
$V_{v}(t)$. It can be considered as a functional of $v$ and $\widehat{W}($.$) . If we replace the initial time 0$ by a random time $\theta$ which is a stopping time with respect to $\mathcal{F}_{t}$, then we mean the same functional on arguments $V_{v}(\theta)$ and $\widehat{W}(.+\theta)-\widehat{W}(\theta)$. It is a stopping time with respect to $\mathcal{F}_{t+\theta}$. It is important to notice that $\widehat{W}(.+\theta)-\widehat{W}(\theta)$ is independent of $V_{v}(\theta)$. Therefore we have the following formula for any test function $\Psi(x, s)$ :

$$
\begin{equation*}
\widehat{E}\left[\Psi\left(V_{v}\left(\hat{\tau}_{\theta}\right), \hat{\tau}_{\theta}\right) \mid \mathcal{F}_{\theta}\right]=\Psi\left(V_{v}(\theta), \theta\right) \mathbb{1}_{V_{v}(\theta) \geq \hat{v}}+\left.\mathbb{1}_{V_{v}(\theta)<\hat{v}} \widehat{E}[\Psi(\hat{v}, t+\hat{\tau}(v))]\right|_{v=V_{v}(\theta), t=\theta} . \tag{3.9}
\end{equation*}
$$

From (3.9) and (3.7), with $\Psi(x, s)=\exp \{-r(s-t)\}, s \geq t$, we can write explicitly the Laplace transform of the conditional density as

$$
\begin{equation*}
\widehat{E}\left[\exp \left\{-r\left(\hat{\tau}_{\theta}-\theta\right)\right\} \mid \mathcal{F}_{\theta}\right]=\mathbb{1}_{V_{v}(\theta) \geq \hat{v}}+\mathbb{1}_{V_{v}(\theta)<\hat{v}} \frac{F\left(V_{v}(\theta)\right)}{(1-a) \hat{v}-K} . \tag{3.10}
\end{equation*}
$$

When the leader enters at time $\theta<\infty$, by paying cost $K$, he/she receives $V_{v}(\theta)$ minus the expected discounted value that he/she will surrender to the follower at the time of the follower's entry. Assuming the follower is rational, the leader knows that the follower will enter at time $\hat{\tau}_{\theta}$. At time $\hat{\tau}_{\theta}$, the leader surrenders $(1-a) V_{v}\left(\hat{\tau}_{\theta}\right)$. So at time $\theta$, if $\theta<\infty$, the leader receives

$$
\begin{aligned}
& V_{v}(\theta)-K-(1-a) \widehat{E}\left[e^{-r\left(\hat{\tau}_{\theta}-\theta\right)} V_{v}\left(\hat{\tau}_{\theta}\right) \mathbb{1}_{\hat{\tau}_{\theta}<\infty} \mid \mathcal{F}_{\theta}\right] \\
& =a V_{v}(\theta) \mathbb{1}_{V_{v}(\theta) \geq \hat{v}}+\left(V_{v}(\theta)-\beta F\left(V_{v}(\theta)\right) \mathbb{1}_{V_{v}(\theta)<\hat{v}}-K,\right.
\end{aligned}
$$

where we use the fact that $\frac{(1-a) \hat{v}}{(1-a) \hat{v}-K}=\beta$ and $F(v)$ is defined in (3.6). To facilitate the presentation, we define

$$
\begin{equation*}
\Psi(v)=a v \mathbb{1}_{v \geq \hat{v}}+(v-\beta F(v)) \mathbb{1}_{v<\hat{v}}-K . \tag{3.11}
\end{equation*}
$$

The leader's problem can now be expressed as

$$
\begin{equation*}
L(v)=\sup _{\theta \geq 0} \widehat{E}\left[e^{-r \theta} \Psi\left(V_{v}(\theta)\right) \mathbb{1}_{\theta<\infty}\right] . \tag{3.12}
\end{equation*}
$$

The leader's value function $L(v)$ must satisfy

$$
\begin{equation*}
L(v) \geq 0 ; \quad L(v) \geq \Psi(v) \tag{3.13}
\end{equation*}
$$

The obstacle $\Psi(v)$ presents a problem because it is only continuous, not $C^{1}(0, \infty)$. There is, however, only one point of nondifferentiability, $\hat{v}$, which we observe from the following:

$$
\Psi^{\prime}(v)= \begin{cases}a, & v>\hat{v}  \tag{3.14}\\ 1-\beta(1-a)\left(\frac{v}{\hat{v}}\right)^{\beta-1}, & v<\hat{v},\end{cases}
$$

with $\Psi^{\prime}(\hat{v}-0)=1-\beta(1-a)<\Psi^{\prime}(\hat{v}+0)=a$.
We next find bounds associated with the obstacle function, $\Psi(v)$, and the value function, $L(v)$. We first proceed to the bound of $\Psi(v)$. By (3.11), $\Psi(v)$ can be expressed as

$$
\Psi(v)=a v-K+\widetilde{\Psi}(v) \mathbb{1}_{v<\hat{v}},
$$

where $\tilde{\Psi}(v)=(1-a) v-\beta F(v)$ satisfies

$$
\tilde{\Psi}^{\prime}(v)=(1-a)-\beta F^{\prime}(v)= \begin{cases}1-a\left[1-\beta\left(\frac{v}{\hat{v}}\right)^{\beta-1}\right], & v \leq \hat{v} \\ (1-a)(1-\beta), & v \geq \hat{v}\end{cases}
$$

Hence, $\tilde{\Psi}^{\prime}(v) \geq 0$ if $0<v<v^{*}=\hat{v}\left(\frac{1}{\beta}\right)^{\frac{1}{\beta-1}}<\hat{v}$ and $\tilde{\Psi}^{\prime}(v)<0$ if $v>v^{*}$.
Moreover, $\tilde{\Psi}(0)=\tilde{\Psi}(\hat{v})=0$. Therefore,

$$
\begin{equation*}
a v-K \leq \Psi(v) \leq a v \tag{3.15}
\end{equation*}
$$

We now proceed to bounds for $L(v)$. By (3.12) and (3.15), we have

$$
\begin{equation*}
L(v) \leq a \sup _{\theta \geq 0} \widehat{E}\left[e^{-r \theta} V_{v}(\theta) \mathbb{1}_{\theta<\infty}\right] \tag{3.16}
\end{equation*}
$$

From (3.2), we have

$$
\begin{aligned}
d V_{v}(t) e^{-r t} & =\eta(\xi-\lambda) V_{v}(t) e^{-r t} d t+\eta V_{v}(t) e^{-r t} d \widehat{W}(t) \\
& \leq \eta V_{v}(t) e^{-r t} d \widehat{W}(t)
\end{aligned}
$$

which shows the supermartingale property for the discounted value process. By the optional sampling theorem, this implies

$$
\widehat{E}\left[V_{v}(\theta \wedge T) e^{-r(\theta \wedge T)}\right] \leq v, \text { from which it follows that } \widehat{E}\left[V_{v}(\theta) e^{-r \theta} \mathbb{1}_{\theta<\infty}\right] \leq v
$$

Therefore, (3.16) implies $0 \leq L(v) \leq a v$. From (3.12) and (3.15), we have $L(v) \geq \Psi(v) \geq$ $a v-K$; hence we arrive at the bound of $L(v)$ :

$$
\begin{equation*}
(a v-K)^{+} \leq L(v) \leq a v \tag{3.17}
\end{equation*}
$$

We now proceed to formulate the leader's problem in a different way. By (3.12), we need to evaluate $\widehat{E}\left[e^{-r \theta} V_{v}(\theta) \mathbb{1}_{\theta<\infty}\right]$. Using

$$
\widehat{E}\left[V_{v}(\theta \wedge T) e^{-r(\theta \wedge T)}\right]=v+\eta(\xi-\lambda) \widehat{E}\left[\int_{0}^{\theta \wedge T} e^{-r s} V_{v}(s) d s\right]
$$

letting $T \rightarrow \infty$, and using

$$
\widehat{E}\left[V_{v}(T) e^{-r T}\right]=v e^{\eta(\xi-\lambda) T} \rightarrow 0 \quad \text { as } T \rightarrow \infty
$$

we obtain

$$
\begin{equation*}
\widehat{E}\left[V_{v}(\theta) e^{-r \theta} \mathbb{1}_{\theta<\infty}\right]=v+\eta(\xi-\lambda) \widehat{E}\left[\int_{0}^{\theta} e^{-r s} V_{v}(s) d s\right] \tag{3.18}
\end{equation*}
$$

Rewriting (3.12) by making use of (3.18) with some algebraic manipulation and setting $U(v)=L(v)-a v+K$, we have ${ }^{4}$

$$
\begin{align*}
U(V)= & \sup _{\theta \geq 0} \widehat{E}\left\{e^{-r \theta}\left(\Psi\left(V_{v}(\theta)\right)-a V_{v}(\theta)+K\right) \mathbb{1}_{\theta<\infty}\right. \\
& \left.\quad+\int_{0}^{\theta} e^{-r s}\left(a \eta(\xi-\lambda) V_{v}(s)+r K\right) d s\right\} \\
= & \sup _{\theta \geq 0} \widehat{E}\left\{e^{-r \theta} \chi\left(V_{v}(\theta)\right) \mathbb{1}_{\theta<\infty}+\int_{0}^{\theta} e^{-r s} f\left(V_{v}(s)\right) d s\right\} . \tag{3.19}
\end{align*}
$$

This formulation leads to an optimal stopping time problem with an obstacle $\chi(v)=\Psi(v)-$ $a v+K$ and a running profit $f(v)=a \eta(\xi-\lambda) v+r K$.

We have

$$
\begin{equation*}
0 \leq \chi(v) \leq K \quad \text { and } \quad 0 \leq U(v) \leq K \tag{3.20}
\end{equation*}
$$

3.2.2. The leader's problem V.I. We now formulate the analytic problem corresponding to (3.19) by dynamic programming. We cannot proceed as in (3.5) because we cannot guarantee that $U(v)$ is $C^{1}(0, \infty)$ since $\chi(v)$ is not $C^{1}(0, \infty)$. We must establish the variational formulation of (3.19) in the weak sense.

We introduce the following useful functional spaces with a weight function $w(x)=\frac{1}{\left(1+x^{2}\right)^{\varrho}}$ :

$$
\begin{gather*}
L_{\varrho}^{2}(\Omega)=\left\{\Phi(x) \mid \int_{\Omega} w(x) \Phi^{2}(x) d x<\infty\right\}  \tag{3.21}\\
H_{\varrho}^{1}(\Omega)=\left\{\Phi \in L_{\varrho}^{2}(\Omega) \mid x \Phi^{\prime}(x) \in L_{\varrho}^{2}(\Omega) \text { if } \Omega \subseteq \mathbb{R}^{+} ; \Phi^{\prime}(x) \in L_{\varrho}^{2}(\Omega) \text { if } \Omega \subseteq \mathbb{R}\right\}, \tag{3.22}
\end{gather*}
$$

where $L_{\varrho}^{2}(\Omega)$ is a Hilbert space with the weighted scalar product

$$
\begin{equation*}
(\Phi, \tilde{\Phi})_{\varrho}=\int_{\Omega} \Phi(x) \tilde{\Phi}(x) w(x) d x \tag{3.23}
\end{equation*}
$$

and $H_{\varrho}^{1}(\Omega)$ is a Sobolev space with the scalar product

$$
\begin{equation*}
((\Phi, \tilde{\Phi}))_{\varrho}=(\Phi, \tilde{\Phi})_{\varrho}+\int_{\Omega} x^{2} w(x) \Phi^{\prime}(x) \tilde{\Phi}^{\prime}(x) d x \tag{3.24}
\end{equation*}
$$

We work on $L_{\varrho}^{2}(0, \infty)$ and $H_{\varrho}^{1}(0, \infty) .{ }^{5}$ We define on $H_{\varrho}^{1}(0, \infty)$ the bilinear form

$$
\begin{align*}
b(\Phi, \tilde{\Phi})= & \int_{0}^{\infty} v \Phi^{\prime}(v)\left[-(r+\eta(\xi-\lambda))+\eta^{2} \frac{1-v^{2}(\varrho-1)}{1+v^{2}}\right] \tilde{\Phi}(v) w(v) d v \\
& +\frac{1}{2} \int_{0}^{\infty} \Phi^{\prime}(v) \tilde{\Phi}^{\prime}(v) v^{2} \eta^{2} w(v) d v+\int_{0}^{\infty} r \Phi(v) \tilde{\Phi}(v) w(v) d v \tag{3.25}
\end{align*}
$$

[^68]Clearly, $b(\Phi, \Phi)^{6}$ is continuous on $H_{\varrho}^{1}(0, \infty)$. For $\alpha \geq \alpha_{0}$, we have the coercivity property

$$
\begin{equation*}
b(\Phi, \Phi)+\alpha(\Phi, \Phi)_{\varrho} \geq c_{0}\|\Phi\|_{\varrho}^{2}, c_{0}>0 \tag{3.26}
\end{equation*}
$$

We consider the convex subset of $H_{\varrho}^{1}(0, \infty)$ :

$$
\begin{equation*}
\mathcal{K}=\left\{\Phi \in H_{\varrho}^{1}(0, \infty) \mid \Phi(v) \geq \chi(v) \forall v, \Phi(0)=K\right\} . \tag{3.27}
\end{equation*}
$$

Note that $\mathcal{K}$ is not empty because the constant $K$ belongs to $\mathcal{K}$.
The V.I. corresponding to (3.19) is

$$
\begin{equation*}
b(U, \tilde{U}-U) \geq(f, \tilde{U}-U)_{\varrho} \quad \forall \tilde{U} \in \mathcal{K}, U \in \mathcal{K} . \tag{3.28}
\end{equation*}
$$

To study the existence and uniqueness of the solution to (3.28), we introduce the modified problem

$$
\begin{equation*}
b(U, \tilde{U}-U)+\alpha(U, \tilde{U}-U)_{\varrho} \geq(f, \tilde{U}-U)_{\varrho}+\alpha(G, \tilde{U}-U)_{\varrho} \quad \forall \tilde{U} \in \mathcal{K}, U \in \mathcal{K} . \tag{3.29}
\end{equation*}
$$

In (3.29), $G$ is given with

$$
\begin{equation*}
0 \leq G(v) \leq K \tag{3.30}
\end{equation*}
$$

The coercivity property (3.26) guarantees the problem (3.29) has a unique solution that we denote by $\Gamma_{\alpha}(G)$ to emphasize the dependence on $G$.

We first check that the unique solution, $\Gamma_{\alpha}(G)$, to (3.29) is in the interval $[0, K]$, the bound of $U(v)$.

Lemma 3.2. We have the property

$$
\begin{equation*}
0 \leq \Gamma_{\alpha}(G) \leq K \tag{3.31}
\end{equation*}
$$

Proof. See Appendix A.
We next check the contraction mapping property.
Lemma 3.3. We have

$$
\begin{equation*}
\left\|\Gamma_{\alpha}\left(G_{1}\right)-\Gamma_{\alpha}\left(G_{2}\right)\right\|_{L^{\infty}} \leq \frac{\alpha}{\alpha+r}\left\|G_{1}-G_{2}\right\|_{L^{\infty}} . \tag{3.32}
\end{equation*}
$$

Proof. See Appendix B.
Theorem 3.4. Assume $\xi-\lambda<0$. The V.I. (3.28) has a unique solution $U$ such that

$$
\begin{equation*}
0 \leq U \leq K \tag{3.33}
\end{equation*}
$$

Proof. It is a consequence of the fact that a solution to (3.28) satisfying (3.33) is a fixed point of $\Gamma_{\alpha}$, defined as a map from the interval (3.33) into itself.

Theorem 3.5. The solution $U$ of the V.I. (3.28) coincides with (3.19).
Proof. See Appendix C.

[^69]3.2.3. Smoothness of the solution. It turns out that the solution $U(v)$ will be smoother than the obstacle. This is due to the fact that the obstacle is not continuously differentiable in a single point. So the V.I. will have a strong formulation, which is given by the following relations:
\[

\left\{$$
\begin{array}{l}
-\frac{1}{2} U^{\prime \prime}(v) \eta^{2} v^{2}-(r+\eta(\xi-\lambda)) v U^{\prime}(v)+r U(v) \geq a \eta(\xi-\lambda) v+r K,  \tag{3.34}\\
U(v) \geq \chi(v), \\
{[U(v)-\chi(v)]\left[-\frac{1}{2} U^{\prime \prime}(v) \eta^{2} v^{2}-(r+\eta(\xi-\lambda)) v U^{\prime}(v)\right.} \\
\quad+r U(v)-a \eta(\xi-\lambda) v-r K]=0, \\
U(0)=K \quad
\end{array}
$$\right.
\]

with $\chi(v)=((1-a) v-\beta F(v)) \mathbb{1}_{v<\hat{v}}$ satisfying

$$
\left\{\begin{array}{l}
-\frac{1}{2} \chi^{\prime \prime}(v) \eta^{2} v^{2}-(r+\eta(\xi-\lambda)) v \chi^{\prime}(v)+r \chi(v)=-(1-a) \eta(\xi-\lambda) v \text { if } v<\hat{v},  \tag{3.35}\\
\chi(0)=\chi(\hat{v})=(1-a) \hat{v}-\frac{\beta K}{\beta-1}=0
\end{array}\right.
$$

where $\hat{v}=\frac{\beta K}{(1-a)(\beta-1)}$.
We look for a solution to (3.34) in an interval $0 \leq U(v) \leq \bar{U}(v)$, where $\bar{U}(v)$ will be a ceiling function which is $C^{1}$ and vanishes for $v$ sufficiently large. We consider

$$
\begin{equation*}
\bar{U}(v)=-a\left[v-\bar{v}-\frac{\bar{v}}{\beta}\left(\left(\frac{v}{\bar{v}}\right)^{\beta}-1\right)\right], \quad v<\bar{v} . \tag{3.36}
\end{equation*}
$$

This function and its derivative vanish at $\bar{v}$, i.e., $\bar{U}(\bar{v})=\bar{U}^{\prime}(\bar{v})=0$. It is extended by 0 for $v>\bar{v}$.

We first need to check that $\bar{U}(v)$ defined above is an appropriate ceiling function. Using (3.36), we see that

$$
\begin{align*}
-\frac{1}{2} v^{2} \eta^{2} \bar{U}^{\prime \prime}-v \bar{U}^{\prime}(r+\eta(\xi-\lambda))+r \bar{U} & =\operatorname{av\eta }(\xi-\lambda)+\operatorname{ar} \bar{v}\left(1-\frac{1}{\beta}\right) \\
& \geq \operatorname{av\eta }(\xi-\lambda)+r K \tag{3.37}
\end{align*}
$$

for $\bar{v}$ sufficiently large.
We also need to check the property that $\bar{U}(v)>\chi(v)$. Since $\chi(v)=0$ for $v>\hat{v}$, it is sufficient to prove $\bar{U}(v)>\chi(v)$ for $v<\hat{v}$. From (3.37), we have

$$
\operatorname{av\eta }(\xi-\lambda)+\operatorname{ar} \bar{v}\left(1-\frac{1}{\beta}\right)>-(1-a) v \eta(\xi-\lambda),
$$

provided $\operatorname{ar} \bar{v}\left(1-\frac{1}{\beta}\right)>-v \eta(\xi-\lambda)$, which occurs if $\operatorname{ar} \bar{v}\left(1-\frac{1}{\beta}\right)>-\hat{v} \eta(\xi-\lambda)$, i.e., if $\bar{v}$ is sufficiently large. This inequality together with (3.35) implies $\bar{U}(v) \geq \chi(v)$. Combining with the first inequality implies

$$
\begin{equation*}
U(v) \leq \bar{U}(v) . \tag{3.38}
\end{equation*}
$$

Thus, $\bar{U}(v)$ is an appropriate ceiling function.

We now apply penalty approximation techniques to (3.34). We approximate (3.34) by a smoother problem, a penalized problem. That is, we look for $U_{\epsilon}$ such that

$$
\left\{\begin{array}{l}
-\frac{1}{2} U_{\epsilon}^{\prime \prime}(v) \eta^{2} v^{2}-(r+\eta(\xi-\lambda)) v U_{\epsilon}^{\prime}(v)+r U_{\epsilon}(v)=a \eta(\xi-\lambda) v+r K+\frac{1}{\epsilon}\left(\chi-U_{\epsilon}(v)\right)^{+}  \tag{3.39}\\
U_{\epsilon}(0)=K ; \quad U_{\epsilon}(\bar{v})=0
\end{array}\right.
$$

We also have

$$
\begin{equation*}
U_{\epsilon}(v) \leq \bar{U}(v) . \tag{3.40}
\end{equation*}
$$

We can check the property of (3.40). Indeed, setting $\widetilde{U}_{\epsilon}=U_{\epsilon}-\bar{U}$, we have

$$
\left\{\begin{array}{l}
-\frac{1}{2} \widetilde{U}_{\epsilon}^{\prime \prime}(v) \eta^{2} v^{2}-(r+\eta(\xi-\lambda)) v \widetilde{U}_{\epsilon}^{\prime}(v)+r \widetilde{U}_{\epsilon}(v) \leq \frac{1}{\epsilon}\left(\widetilde{U}_{\epsilon}(v)\right)^{-} \\
\widetilde{U}_{\epsilon}(0)=0 ; \quad \widetilde{U}_{\epsilon}(\bar{v})=0
\end{array}\right.
$$

This implies $\widetilde{U}_{\epsilon} \leq 0$ by the maximum principle, and (3.40) is satisfied.
The key point is to check the estimate

$$
\begin{equation*}
\frac{\left(\chi-U_{\epsilon}\right)^{+}}{\epsilon} \leq c_{0} . \tag{3.41}
\end{equation*}
$$

Using the property of $\chi(v)$, we have

$$
\begin{aligned}
-\frac{1}{2} U_{\epsilon}^{\prime \prime}(v) \eta^{2} v^{2}-(r+\eta(\xi-\lambda)) v U_{\epsilon}^{\prime}(v)+\left(r+\frac{1}{\epsilon}\right) U_{\epsilon}(v) & \geq a \eta(\xi-\lambda) v+r K \\
& \geq a \eta(\xi-\lambda) \bar{v}+r K \text { for } v<\bar{v}
\end{aligned}
$$

hence

$$
\begin{equation*}
U_{\epsilon}(v) \geq \epsilon(a \eta(\xi-\lambda) \bar{v}+r K) \tag{3.42}
\end{equation*}
$$

Consider next $(0, \hat{v})$ and set

$$
U_{\epsilon}^{*}(v)=U_{\epsilon}(v)-\chi(v) .
$$

$U_{\epsilon}^{*}(v)$ satisfies the following equations:

$$
\left\{\begin{array}{l}
-\frac{1}{2} U_{\epsilon}^{* \prime \prime}(v) \eta^{2} v^{2}-(r+\eta(\xi-\lambda)) v U_{\epsilon}^{* \prime}(v)+r U_{\epsilon}^{\prime}(v)=a \eta(\xi-\lambda) v+r k+\frac{1}{\epsilon} U_{\epsilon}^{*}(v)^{-},  \tag{3.43}\\
U_{\epsilon}^{*}(0)=K ; \quad U_{\epsilon}^{*}(\hat{v}) \geq \epsilon(a \eta(\xi-\lambda) \bar{v}+r K)
\end{array}\right.
$$

This implies

$$
U_{\epsilon}^{*}(v) \geq \min [\epsilon(\eta(\xi-\lambda) \hat{v}+r K), \epsilon(a \eta(\xi-\lambda) \bar{v}+r K)] .
$$

Therefore, we obtain (3.41) with

$$
\begin{equation*}
c_{0}=-r K-\eta(\xi-\lambda) \max (\hat{v}, a \bar{v})>0 . \tag{3.44}
\end{equation*}
$$

Referring back to (3.37), we have

$$
\begin{equation*}
\int_{0}^{\bar{v}} v^{2}\left|U_{\epsilon}^{\prime \prime}\right| d v \leq C \tag{3.45}
\end{equation*}
$$

Therefore, we can extract from $U_{\epsilon}(v)$ a subsequence, still denoted $U_{\epsilon}(v)$, such that

$$
\left\{\begin{array}{l}
U_{\epsilon}(v) \rightarrow U(v) \text { in } L^{\infty}(0, \bar{v}) \text { weak star }  \tag{3.46}\\
v U_{\epsilon}^{\prime}(v) \rightarrow v U^{\prime}(v) \text { in } L^{2}(0, \bar{v}) \text { weakly } \\
v^{2} U_{\epsilon}^{\prime \prime}(v) \rightarrow v^{2} U^{\prime \prime}(v) \text { in } L^{\infty}(0, \bar{v}) \text { weak star. }
\end{array}\right.
$$

In particular, we can take

$$
U(\bar{v})=0 ; \quad U(0)=K
$$

Passing to the limit in (3.39), we obtain

$$
\left\{\begin{array}{l}
-\frac{1}{2} U^{\prime \prime}(v) \eta^{2} v^{2}-(r+\eta(\xi-\lambda)) v U^{\prime}(v)+r U(v)=a \eta(\xi-\lambda) v+r K+\varphi(v)  \tag{3.47}\\
U(v) \geq \chi(v) \\
(U(v)-\chi(v)) \varphi(v)=0 \\
U(0)=K \\
U(\bar{v})=0 \\
U(v) \in C^{1}(0, \bar{v}), U^{\prime \prime}(v) \in L^{\infty}(0, \bar{v})
\end{array}\right.
$$

The function $U(v)$ is the solution of (3.34). Let us check that

$$
\begin{equation*}
U^{\prime}(\bar{v})=0 \tag{3.48}
\end{equation*}
$$

Indeed, $U^{\prime}(\bar{v}) \leq 0$ since $U(v)>0$ for $v<\bar{v}$. But, from (3.40), we have $U(v) \leq \bar{U}(v)$, yielding

$$
\frac{U(v)-U(\bar{v})}{v-\bar{v}} \geq \frac{\bar{U}(v)-\bar{U}(\bar{v})}{v-\bar{v}} \quad \text { for } v<\bar{v}
$$

and thus $U^{\prime}(\bar{v}) \geq 0$. We see that (3.48) holds. It follows that the function $U(v)$, the solution of (3.47), extended by 0 is the solution of (3.34).

The next interesting result is that we can get an explicit solution for $U(v)$. It is not a threshold solution, which is due to the lack of smoothness of the obstacle $\chi(v)$ in $\hat{v}$. The solution will be referred as a two-interval solution, as explained in the following.

Theorem 3.6. There exist three points $0<v_{1}<v_{2}<\hat{v}<v_{3}$ such that

$$
\left\{\begin{array}{l}
-\frac{1}{2} U^{\prime \prime}(v) \eta^{2} v^{2}-(r+\eta(\xi-\lambda)) v U^{\prime}(v)+r U(v)=a \eta(\xi-\lambda) v+r K  \tag{3.49}\\
\quad 0<v<v_{1} \text { and } v_{2}<v<v_{3} \\
U(v)=\chi(v) \text { for } v_{1} \leq v \leq v_{2} \\
U(v)=0 \text { for } v>v_{3}
\end{array}\right.
$$

with matching conditions $U^{\prime}\left(v_{1}\right)=\chi^{\prime}\left(v_{1}\right), U^{\prime}\left(v_{2}\right)=\chi^{\prime}\left(v_{2}\right)$, and $U^{\prime}\left(v_{3}\right)=0$.
Proof. See Appendix D.

So the solution coincides with the obstacle on two intervals. Knowing this fact, it is possible to make a direct calculation taking into account

$$
\left\{\begin{array}{l}
U(v)=-a v+K+A v^{\beta} \quad \text { if } v<v_{1},  \tag{3.50}\\
U(v)=-a v+K+C v^{\beta}+D v^{\beta_{1}} \quad \text { if } v_{2}<v<v_{3}
\end{array}\right.
$$

where $\beta$ and $\beta_{1}$ are, respectively, the positive and negative roots of the characteristic equation $\frac{1}{2} \eta^{2} \beta^{2}+\left(r+\eta(\xi-\lambda)-\frac{1}{2} \eta^{2}\right)-r=0$, and $A, C$, and $D$ are constants to be defined. We leave it to the reader to perform these calculations.

We can define the leader's optimal stopping rule as

$$
\hat{\theta}(v)=\left\{\begin{array}{l}
\inf \left\{t \mid V_{v}(t) \geq v_{1}\right\} \quad \text { if } 0 \leq v<v_{1},  \tag{3.51}\\
0 \text { if } v_{1} \leq v \leq v_{2}, \\
\inf \left\{t \mid V_{v}(t) \leq v_{2} \text { or } V_{v}(t) \geq v_{3}\right\} \quad \text { if } v_{2}<v<v_{3}, \\
0 \text { if } v \geq v_{3}
\end{array}\right.
$$

3.2.4. Optimal rules in the case of complete markets. We summarize the leader's and the follower's optimal investment rules as follows:

1. If $v<v_{1}$, the leader waits to enter until $v \geq v_{1}$, and the follower enters when $v \geq \hat{v}$.
2. If $v_{1}<v<v_{2}$, the leader enters immediately, and the follower enters when $v \geq \hat{v}$.
3. If $v_{2}<v<v_{3}$, the leader waits to enter until $v$ moves outside the interval bounded by $v_{2}$ and $v_{3}$. The follower's optimal policy is more complicated, and there are two possibilities. One possibility is $v_{2}<v<\hat{v}<v_{3}$. If the leader enters because $v \leq v_{2}$, the follower will make his/her move when $v \geq \hat{v}$. If the leader enters because $v \geq v_{3}$, the follower enters immediately since $v \geq v_{3}>\hat{v}$. The other possibility is $v_{2}<\hat{v}<v<v_{3}$. Although $v \geq \hat{v}$, indicating that the follower should enter, the follower is blocked from doing so until the leader enters. If the leader enters because $v \leq v_{2}$, the follower will make his/her move when $v \geq \hat{v}$ again. If the leader enters because $v \geq v_{3}$, the follower enters immediately since $v \geq v_{3}>\hat{v}$.
4. If $v \geq v_{3}$, the leader invests immediately, and the follower then makes his/her move.

Note that both the leader's and the follower's investment rules deviate from the monopolist's case. Although the follower's optimal investment trigger, $\hat{v}$, is the same as the monopolist's, he/she is forbidden to enter the market until the leader has already done so. The leader's optimal investment trigger, characterized by two intervals, which are $\left[v_{1}, v_{2}\right]$ and $\left[v_{3}, \infty\right)$, differs from the monopolist's, $\hat{v}$, due to the consideration of strategic interactions from the follower's action. By Theorem 3.6, the unique triple for the leader's optimal stopping is such that $0<v_{1}<v_{2}<\hat{v}<v_{3}$; the leader will invest suboptimally if he/she ignores the follower's action.

## 4. Lump-sum payoffs and incomplete market assumption.

4.1. Utility-based pricing model. The asset $S$ representing the market still evolves as (2.1). The project value process $V$ of (3.1), generated by the capital investment project, now evolves as

$$
\begin{equation*}
d V(t)=r V(t) d t+\eta V(t)\left(\xi d t+\rho d W(t)+\sqrt{1-\rho^{2}} d W^{0}(t)\right), \tag{4.1}
\end{equation*}
$$

where $W(t)$ and $W^{0}(t)$ are independent Wiener processes. The market asset $S$ depends on only one of the Wiener processes, but the project value depends on both. The parameter $0<|\rho|<1$ is the correlation coefficient between market uncertainty and project value process uncertainty. The market asset $S$ can span only a portion of the project value risk driven by the Wiener process $W(t)$. This leaves the remaining risk driven by $W^{0}(t)$ unhedgeable. Therefore, the market is incomplete, and there is more than one risk-neutral pricing measure. Alternatives to arbitrage-free pricing must be developed in order to correctly value assets and define the related risk management.

We adopt the utility-based pricing approach where the risk-averse investor's preferences are characterized by an exponential utility function given by

$$
\begin{equation*}
U(x)=-\frac{1}{\gamma} e^{-\gamma x} \tag{4.2}
\end{equation*}
$$

where the argument $x$ is the investor's wealth $(x>0)$, and $\gamma$ is his/her risk aversion parameter, with $\gamma>0$.

The rational investor maximizes his/her expected utility of wealth. To simplify exposition, we ignore consumption. The evolution of wealth is uniquely driven by the investment portfolio strategy. Given initial wealth, $x$, the risk-averse investor optimizes his/her portfolio by dynamically choosing allocations in the market asset $S$ and the riskless bond. The investor's wealth, $X$, evolves as

$$
\begin{equation*}
d X(t)=r X(t) d t+X(t) \pi(t) \sigma(\lambda d t+d W(t)) \tag{4.3}
\end{equation*}
$$

where $\pi(t)$ is the portion of wealth invested in asset $S$. We use the discounted values $\widetilde{V}(t)=$ $V(t) e^{-r t}$ and $\widetilde{X}(t)=X(t) e^{-r t}$, where

$$
\begin{align*}
& d \widetilde{V}(t)=\widetilde{V}(t) \eta\left(\xi d t+\rho d W(t)+\sqrt{1-\rho^{2}} d W^{0}(t)\right),  \tag{4.4}\\
& d \widetilde{X}(t)=\widetilde{X}(t) \pi(t) \sigma(\lambda d t+d W(t)) \tag{4.5}
\end{align*}
$$

Processes $\widetilde{V}(t)$ and $\widetilde{X}(t)$ are positive. Let $\mathcal{F}_{t}=\sigma\left(W(s), W^{0}(s) ; s \leq t\right)$. We consider stopping times $\tau$ with respect to $\mathcal{F}_{t}$ starting with initial values $\widetilde{V}(0)=v, \widetilde{X}(0)=x$. At $\tau$, the individual invests and receives $\widetilde{X}_{x}(\tau)+\left(\widetilde{V}_{v}(\tau)-K\right)^{+} .{ }^{7}$ With a pair $(\pi(\cdot), \tau)$ we associate the objective function

$$
\begin{equation*}
J_{x, v}(\pi(\cdot), \tau)=E\left[e^{\frac{\lambda^{2}}{2} \tau} U\left(\widetilde{X}_{x}(\tau)+\left(\widetilde{V}_{v}(\tau)-K\right)^{+}\right)\right] . \tag{4.6}
\end{equation*}
$$

We assume $\tau$ is finite a.s. The function (4.6) is well defined, but the value $-\infty$ is possible. When $|\rho|=1$, we expect the criterion to be equivalent to $\widehat{E}\left(\widetilde{V}_{v}(\tau)-K\right)^{+}$, which differs from the risk-neutral pricing, $\widehat{E}\left[e^{-r \tau}\left(V_{v}(\tau)-K\right)^{+}\right]$. These changes are motivated by the quest for

[^70]an analytical solution. The positive exponential is used because $U$ defined in (4.2) is negative. The coefficient $\frac{\lambda^{2}}{2}$ permits an essential simplification. ${ }^{8}$

The investor's problem is to maximize his/her expected utility taking portfolio investment strategies and stopping times into consideration. We define the associated value function as

$$
\begin{equation*}
F(x, v)=\sup _{\pi(\cdot), \tau} J_{x, v}(\pi(\cdot), \tau) \tag{4.7}
\end{equation*}
$$

4.2. Single player. In this section, we want to solve the single player problem. We prepare for the sharing of the resource by using a coefficient $a$. At time $\tau$, the individual invests and receives $\widetilde{X}_{x}(\tau)+(1-a) \widetilde{V}_{v}(\tau)-K$.

With a pair $(\pi(\cdot), \tau)$ we associate the objective function

$$
\begin{equation*}
J_{x, v}(\pi(\cdot), \tau)=E\left[e^{\frac{\lambda^{2}}{2} \tau} U\left(\widetilde{X}_{x}(\tau)+\left((1-a) \widetilde{V}_{v}(\tau)-K\right)^{+}\right)\right] . \tag{4.8}
\end{equation*}
$$

We again assume $\tau$ is finite a.s. The associated value function is defined as

$$
\begin{equation*}
F(x, v)=\sup _{\pi(\cdot), \tau} J_{x, v}(\pi(\cdot), \tau) \tag{4.9}
\end{equation*}
$$

We solve the problem by V.I. By Henderson [22], we have the following theorem.
Theorem 4.1. We assume $\xi-\rho \lambda<0$ and set

$$
\begin{equation*}
\beta=1-\frac{2(\xi-\rho \lambda)}{\eta}>1 . \tag{4.10}
\end{equation*}
$$

## Define $\hat{v}$ by

$$
\begin{equation*}
(1-a) \hat{v}=K+\frac{\varpi}{\gamma\left(1-\rho^{2}\right)} \tag{4.11}
\end{equation*}
$$

with $\varpi$ the unique positive solution of

$$
\begin{equation*}
e^{\varpi}-\frac{\varpi}{\beta}=1+\frac{K \gamma\left(1-\rho^{2}\right)}{\beta} . \tag{4.12}
\end{equation*}
$$

We look for a solution $F(x, v)=U(x) \Phi(v)$ and have

$$
F(x, v)= \begin{cases}U(x)(\chi(v))^{\frac{1}{1-\rho^{2}}} & \text { if } 0<v<\hat{v}  \tag{4.13}\\ U(x) e^{-\gamma((1-a) v-K)} & \text { if } v \geq \hat{v}\end{cases}
$$

[^71]where $\chi(v)=1-\left(1-e^{-\varpi}\right)\left(\frac{v}{\hat{v}}\right)^{\beta}$. Note that if $|\rho|$ is close to one or $\gamma$ is close to zero (i.e., risk neutrality), then $\varpi \sim \frac{K \gamma\left(1-\rho^{2}\right)}{\beta-1}$ and $\hat{v}=\frac{K \beta}{(\beta-1)(1-a)}$. It is the value identified in section 3.1, except for a different value of $\beta .{ }^{9}$

Differing from Henderson's verification through a probabilistic argument, we prove that $F(x, v)$ in (4.13) is the solution of the V.I. (which is the equivalent of the Bellman equation for stopping times). We omit the proof. The optimal strategy for the investor is

$$
\begin{equation*}
\hat{\tau}(x, v)=\inf \left\{t \geq 0 \mid \widetilde{V}_{v}(t) \text { is outside }(0, \hat{v})\right\}=\hat{\tau}(v) \tag{4.14}
\end{equation*}
$$

and $\hat{\tau}(v)<\infty$ a.s. This is the optimal stopping time and does not depend on the initial wealth $x$.
4.3. Two players: A Stackelberg game. As in section 3.2, we consider a Stackelberg leader and follower game. The leader enters the market at time $\theta$ and the follower at time $\tau \geq \theta$, both stopping times of the filtration $\mathcal{F}_{t}$. The leader knows that the follower, acting rationally, will enter at time $\hat{\tau}_{\theta}$, at which time he/she must surrender a portion of the project to the follower.

The follower's investment problem is given by solving the one player problem. After the leader's entry, he/she makes the optimal stopping decision. By paying an investment cost $K$ at $\tau$, the follower receives $\widetilde{X}_{x}(\tau)+(1-a) \widetilde{V}(\tau)$. The follower's problem is to maximize the expected discounted utility of wealth by choosing stopping time $\tau$ and investment strategy $\pi$. Thus, the follower's strategy is identical to that described in section 4.2. We have the follower's value function as defined by (4.13), where $\hat{v}$ is given by (4.11) with $\varpi$ the unique value of (4.12). Thus, the optimal stopping strategy for the follower, $\hat{\tau}(v)$, is the same as defined in (4.14).

The stopping time $\hat{\tau}(v)$ is the optimal entry if the follower can enter the market at time zero. Since the follower enters after the leader (who starts at $\theta$ ), for finite $\theta$, the follower will enter at time

$$
\begin{equation*}
\hat{\tau}_{\theta}=\theta+\hat{\tau}\left(\widetilde{V}_{v}(\theta)\right) \tag{4.15}
\end{equation*}
$$

with the same considerations as for (3.9).
4.3.1. The leader's problem. The leader must take into account the fact that the follower will enter according to $\hat{\tau}(v)$. The evolution of the leader's discounted wealth, and of the discounted value process, are as in (4.5) and (4.4), respectively.

Beginning with initial wealth $x$, the leader rebalances his/her portfolio holdings by dynamically choosing the investment allocations, $\pi(\cdot)$, in the asset $S$ and riskless bond. The leader chooses a stopping time $\theta$ (with respect to $\mathcal{F}_{t}$ ), at which time he/she invests. At stopping time $\theta$, the leader receives $\widetilde{X}_{x}(\theta)+\left(\widetilde{V}_{v}(\theta)-K\right)^{+}$. However, he/she must anticipate the follower's optimal entry and the corresponding sharing of the market.

[^72]The follower may enter the market after $\theta$, and he/she will decide to do so according to the rule $\hat{\tau}_{\theta}$ described in (4.15). The leader anticipates the follower's actions. So we have the following rules:

1. If $\widetilde{V}(\theta)>\hat{v}$, the follower enters immediately and the leader gets $\widetilde{X}(\theta)+a \widetilde{V}(\theta)-K$, where $a$ is the leader's market share after entry by the follower. This payoff is larger than $\widetilde{X}(\theta)$ since, by assumption, $a>\frac{1}{2}$.
2. If $\widetilde{V}(\theta)<\hat{v}$, then $\hat{\tau}(\widetilde{V}(\theta))>0$, and the leader gets $\widetilde{X}(\theta)+(\widetilde{V}(\theta)-K)$ immediately. However, he/she surrenders, at time $\theta+\hat{\tau}(\tilde{V}(\theta))$, a percentage $(1-a)$ of project value to the follower.
We encounter a hurdle in the latter scenario. Unlike the solution in the complete market, we cannot compare gains and losses occurring at different times directly. At time $\theta$, the leader must determine the value surrendered to the follower, taking into account the follower's optimal entry at $\hat{\tau}_{\theta}$. In the complete market case, this is straightforward because the valuation is in the monetary unit. In the incomplete market case, we are unable to evaluate cash flows generated (or outflows incurred) at different points in time. We will circumvent this problem converting the surrender value by an equivalence (indifference) consideration. We describe this amount in reference to the follower because this surrender value must be equivalent to an inflow that he/she will receive in his/her portfolio optimization problem after the time of entrance $\hat{\tau}_{\theta}$. The leader has no choice but to take into account a surrender value which is acceptable to the follower.

We proceed by first obtaining the surrender project value with equivalence (indifference) consideration. We can start at the origin, which will later be the time of entrance of the leader. Let $\widetilde{X}(t), \widetilde{V}(t)$ be the wealth of the follower and the value process governed by (4.5) and (4.4) with $\widetilde{X}(0)=x, \widetilde{V}(0)=v$. If $v>\hat{v}$, then everything takes place at time 0 . The follower receives $(1-a) v$. He/she values this operation by $U(x+(1-a) v)$ since his/her wealth is $x+(1-a) v$.

If $0<v<\hat{v}$, then the follower receives, at time $\hat{\tau}(v),(1-a) \widetilde{V}(\hat{\tau}(v))$, and the corresponding value is

$$
e^{\frac{\lambda^{2}}{2} \hat{\tau}(v)} U\left(\widetilde{X}_{x}(\hat{\tau}(v))+(1-a) \widetilde{V}_{v}(\hat{\tau}(v))\right) .
$$

Note that $K$ does not enter this calculation because we are looking at the equivalent only at the initial time (later the entrance time of the leader) of what the follower will receive from the leader at time $\hat{\tau}(v)$. The amount $K$ plays a role in calculating $\hat{\tau}(v)$.

Since he/she can manage his/her portfolio on the interval of time $(0, \hat{\tau}(v))$, the value is in fact

$$
\begin{equation*}
H(x, v)=\max _{\pi(\cdot)} E\left[e^{\frac{\lambda^{2}}{2} \hat{\tau}(v)} U\left(\widetilde{X}_{x}(\hat{\tau}(v))+(1-a) \widetilde{V}_{v}(\hat{\tau}(v))\right)\right] \tag{4.16}
\end{equation*}
$$

and we have the boundary condition

$$
\begin{equation*}
H(x, \hat{v})=U(x+(1-a) \hat{v}) \tag{4.17}
\end{equation*}
$$

If $v=0$, then $\hat{\tau}(v)=0$; therefore

$$
\begin{equation*}
H(x, 0)=U(x) \tag{4.18}
\end{equation*}
$$

To simplify exposition, we define the nonlinear differential operator

$$
\Gamma H(x, v)=\frac{\partial H}{\partial v} v \eta \xi+\frac{1}{2} \frac{\partial^{2} H}{\partial v^{2}} v^{2} \eta^{2}+\frac{1}{2} \lambda^{2} H-\frac{1}{2} \frac{\left(\lambda \frac{\partial H}{\partial x}+\frac{\partial^{2} H}{\partial x \partial v} \rho \eta v\right)^{2}}{\frac{\partial^{2} H}{\partial x^{2}}},
$$

and $H(x, v)$ must satisfy

$$
\begin{cases}\Gamma H(x, v)=0, & x \in \mathbb{R}, v<\hat{v}  \tag{4.19}\\ H(x, 0)=U(x) ; & H(x, \hat{v})=U(x+(1-a) \hat{v})\end{cases}
$$

We look for a solution

$$
\begin{equation*}
H(x, v)=U(x) \Lambda(v) \tag{4.20}
\end{equation*}
$$

and find that

$$
\begin{equation*}
\Lambda(v)=\left(1+B v^{\beta}\right)^{\frac{1}{1-\rho^{2}}} \tag{4.21}
\end{equation*}
$$

where $B=-\frac{1-e^{-\left(\varpi+K \gamma\left(1-\rho^{2}\right)\right)}}{\hat{v}^{\beta}}, \beta$ is defined in (4.10), and $\varpi$ is the unique solution of (4.12). Comparing to the function $\Phi(v)$ defined in Theorem 4.1, we have

$$
\begin{equation*}
\Phi(v) e^{-\gamma K} \leq \Lambda(v) \leq \Phi(v) \tag{4.22}
\end{equation*}
$$

We consider that the follower is indifferent between receiving $H^{e}(v)$ at time 0 and losing his/her right to enter into the market at time $\hat{\tau}(v)$, or to exert his/her right to enter into the market. The leader will use this value to express an equivalent to the commitment he/she has to share the market with the follower in due time, at the time when he/she enters into the market. We define now the indifference value $H^{e}(v)$ by

$$
\begin{align*}
H\left(x+H^{e}(v), 0\right) & =H(x, v) \\
& =U\left(x+H^{e}(v)\right) \\
& =U(x) e^{-\gamma H^{e}(v)} \tag{4.23}
\end{align*}
$$

then

$$
\begin{equation*}
e^{-\gamma H^{e}(v)}=\Lambda(v), \quad 0<v<\hat{v} \tag{4.24}
\end{equation*}
$$

We have the extension

$$
\begin{equation*}
H^{e}(v)=(1-a) v, \quad v \geq \hat{v}, \tag{4.25}
\end{equation*}
$$

since the follower jumps into the market immediately and the leader surrenders $(1-a) v$. For $v<\hat{v}$, we have $\hat{\tau}(v)>0$ and it is necessary that $H^{e}(v)<(1-a) v$ since $H^{e}(v)$ is the time 0 value which makes him/her indifferent by having this amount at time 0 and not having ( $1-a$ ) v at time $\hat{\tau}(v)$. Therefore, we need to have the property

$$
\begin{equation*}
H^{e}(v) \leq(1-a) v \quad \text { if } v<\hat{v} \tag{4.26}
\end{equation*}
$$

which is satisfied since this amounts to

$$
\begin{align*}
G(v) & =\Lambda(v)-e^{-\gamma(1-a) v} \\
& =1+B v^{\beta}-e^{-\gamma(1-a)\left(1-\rho^{2}\right) v} \\
& \geq 0 \quad \text { if } v<\hat{v} \tag{4.27}
\end{align*}
$$

Equation (4.27) is true because $G^{\prime \prime}(v) \leq 0, G^{\prime}(0)=\gamma(1-a)\left(1-\rho^{2}\right)>0$, and

$$
G^{\prime}(\hat{v})=\frac{-\gamma(1-a)\left(1-\rho^{2}\right)\left(1-e^{-\gamma K\left(1-\rho^{2}\right)}\right)}{e^{\varpi}-1}<0
$$

The leader deducts the amount $K+H^{e}(v)$ from the value $v$ that he/she receives at time 0 if he/she decides to invest at time 0 . We can now formulate the leader's problem completely.

For each pair $(\pi(\cdot), \theta)$, the leader's objective function is

$$
\begin{equation*}
J_{x, v}(\pi(\cdot), \theta)=E\left[e^{\frac{\lambda^{2}}{2} \theta} U\left(\widetilde{X}(\theta)+\left(\widetilde{V}(\theta)-K-H^{e}(\tilde{V}(\theta))\right)^{+}\right)\right] \tag{4.28}
\end{equation*}
$$

and the leader's problem is to maximize (4.28) with respect to portfolio investment strategies, $\pi$, and stopping times, $\theta$. Thus, we define the leader's value function:

$$
\begin{equation*}
L(x, v)=\sup _{\pi(\cdot), \theta} J_{x, v}(\pi(\cdot), \theta) \tag{4.29}
\end{equation*}
$$

We impose $\theta<\infty$ a.s.
4.3.2. The leader's problem V.I. We first note from the definition (4.29)

$$
L(x, 0) \geq \sup _{\theta} E e^{\frac{\lambda^{2}}{2} \theta} U(x)=U(x)
$$

Next, we may write the V.I. in the strong sense that $L(x, v)$ in (4.29) must satisfy as a consequence of dynamic programming, assuming sufficient smoothness,

$$
\left\{\begin{array}{l}
\Gamma L(x, v) \leq 0  \tag{4.30}\\
L(x, v) \geq U\left(x+\left(v-K-H^{e}(v)\right)^{+}\right) \\
\Gamma L(x, v)\left[L(x, v)-U\left(x+\left(v-K-H^{e}(v)\right)^{+}\right)\right]=0
\end{array}\right.
$$

We can see that for $v=0, U(x)$ satisfies all the conditions. Therefore we may add the boundary condition

$$
\begin{equation*}
L(x, 0)=U(x) \tag{4.31}
\end{equation*}
$$

We look for a solution of the form

$$
\begin{equation*}
L(x, v)=U(x) L(v) \tag{4.32}
\end{equation*}
$$

and obtain

$$
\left\{\begin{array}{l}
L^{\prime}(\xi-\lambda \rho)+\frac{1}{2} v \eta\left(L^{\prime \prime}-\rho^{2} \frac{L^{\prime 2}}{L}\right) \geq 0  \tag{4.33}\\
L(v) \leq e^{-\gamma\left(v-K-H^{e}(v)\right)^{+}} \\
{\left[L^{\prime}(\xi-\lambda \rho)+\frac{1}{2} v \eta\left(L^{\prime \prime}-\rho^{2} \frac{L^{\prime 2}}{L}\right)\right]\left[L(v)-e^{-\gamma\left(v-K-H^{e}(v)\right)^{+}}\right]=0} \\
L(0)=1
\end{array}\right.
$$

We want a solution such that

$$
\begin{equation*}
0 \leq L(v) \leq 1 \tag{4.34}
\end{equation*}
$$

Setting $L(v)=\Sigma(v)^{\frac{1}{1-\rho^{2}}}$, we obtain the V.I.

$$
\left\{\begin{array}{l}
\Sigma^{\prime}(\xi-\lambda \rho)+\frac{1}{2} v \eta \Sigma^{\prime \prime} \geq 0  \tag{4.35}\\
\Sigma(v) \leq e^{-\gamma\left(1-\rho^{2}\right)\left(v-K-H^{e}(v)\right)^{+}} \\
{\left[\Sigma^{\prime}(\xi-\lambda \rho)+\frac{1}{2} v \eta \Sigma^{\prime \prime}\right]\left[\Sigma(v)-e^{-\gamma\left(1-\rho^{2}\right)\left(v-K-H^{e}(v)\right)^{+}}\right]=0} \\
0 \leq \Sigma(v) \leq 1 ; \quad \Sigma(0)=1
\end{array}\right.
$$

We encounter the same mathematical difficulty as we have for the leader's problem in the complete market. The obstacle

$$
\begin{equation*}
\psi(v)=e^{-\gamma\left(1-\rho^{2}\right)\left(v-K-H^{e}(v)\right)^{+}} \tag{4.36}
\end{equation*}
$$

is continuous but not $C^{1}$. At point $\hat{v}$, we have $\psi(\hat{v})=e^{-\gamma\left(1-\rho^{2}\right)(a \hat{v}-K)}$. Moreover, we note the following property.

Lemma 4.2. The function $v-K-H^{e}(v)$ vanishes at a single point $v^{o}$ such that $K<v^{o}<\hat{v}$. Proof. See Appendix E.
The function $\left(v-K-H^{e}(v)\right)^{+}$is continuous but not $C^{1}$. The derivative is discontinuous at $v^{o}$ and $\hat{v}$. The obstacle $\psi(v)$ has the same property. Since the obstacle is not in $C^{1}$, we must consider (4.35) in a weak sense. Nevertheless the function $\Sigma(v)$ will be the value function of an optimal stopping problem (no continuous control). Namely, we have the state equation ${ }^{10}$

$$
\begin{equation*}
d V(t)=V(t) \eta((\xi-\lambda \rho) d t+d W(t)), \quad V(0)=v \tag{4.37}
\end{equation*}
$$

and the value function

$$
\begin{equation*}
\Sigma(v)=\inf _{\tau_{\Sigma}} J_{v}\left(\tau_{\Sigma}\right) \tag{4.38}
\end{equation*}
$$

with

$$
\begin{equation*}
J_{v}\left(\tau_{\Sigma}\right)=E\left[\psi\left(V_{v}\left(\tau_{\Sigma}\right)\right)\right] \tag{4.39}
\end{equation*}
$$

[^73]and $\tau_{\Sigma}$ must be a stopping time which is a.s. finite. To formulate (4.38) in a weak form sense, we introduce the Sobolev space $H_{\varrho}^{1}(0, \infty)$ with the scalar product (3.24). We define the bilinear form
\[

$$
\begin{align*}
b(\phi, \tilde{\phi})=\int_{0}^{\infty} & \phi^{\prime}(v)\left\{-\eta(\xi-\lambda \rho)+\eta^{2} \frac{1-v^{2}(\varrho-1)}{1-v^{2}}\right\} \frac{\tilde{\phi}(v)}{\left(1+v^{2}\right)^{\varrho}} d v \\
& +\frac{1}{2} \int_{0}^{\infty} \frac{\phi^{\prime}(v) \tilde{\phi}^{\prime}(v) v^{2} \eta^{2}}{\left(1+v^{2}\right)^{\varrho}} d v . \tag{4.40}
\end{align*}
$$
\]

Consider next the convex set

$$
\begin{equation*}
\mathcal{K}=\left\{\phi \in H_{\varrho}^{1}(0, \infty) \mid 0 \leq \phi(v) \leq \psi(v) \forall v\right\}, \tag{4.41}
\end{equation*}
$$

which contains $\psi(v)$; i.e., it is not empty.
The V.I. corresponding to (4.38) is

$$
\begin{equation*}
b(\Sigma, \widetilde{\Sigma}-\Sigma) \geq 0 \quad \forall \widetilde{\Sigma} \in \mathcal{K}, \Sigma \in \mathcal{K} . \tag{4.42}
\end{equation*}
$$

We observe from (4.33), (4.34), and (4.35) that $\tau_{\Sigma}$, the optimal stopping we obtain in (4.38), and $\theta$, the optimal stopping we obtain in (4.29), coincide. To simplify the notation and to make use of the optimal stopping notation we defined for the leader, we slightly abuse the notation by replacing $\tau_{\Sigma}$ with $\theta$.

We turn now to prove the existence and uniqueness of the solution of (4.42), whose solution coincides with that of the value function (4.38). We proceed by first establishing the property that the set of solutions of (4.42) has a maximum and a minimum element within a convenient interval. In Theorem 4.3, we prove that the minimum and the maximum solutions coincide, equal to the value function (4.38).

Theorem 4.3. Assume $\xi-\lambda \rho<0$. Then there exists one and only one solution of (4.42). It coincides with the value function (4.38).

Proof. See Appendix F.
4.3.3. Obtaining a two-interval solution. We can investigate whether the solution $\Sigma(v)$ of the V.I. (4.42) is smoother than the obstacle, as is the case in the situation of complete markets. Unfortunately, here, because of the presence of the term $H^{e}(v)$, we cannot have a general result. Nevertheless, we can explore directly the possibility that a $C^{1}$ solution exists by looking for a two-interval strategy for which we will give only a partial account. Namely, we seek three values $v_{1}, v_{2}, v_{3}$ such that

$$
\begin{equation*}
v_{0}<v_{1}<v_{2}<\hat{v}<v_{3} \tag{4.43}
\end{equation*}
$$

and

$$
\begin{cases}\Sigma^{\prime}(\xi-\lambda \rho)+\frac{1}{2} v \eta \Sigma^{\prime \prime}=0, & 0<v<v_{1},  \tag{4.44}\\ \Sigma(v)=\psi(v), & v_{1}<v<v_{2}, \\ \Sigma^{\prime}(\xi-\lambda \rho)+\frac{1}{2} v \eta \Sigma^{\prime \prime}=0, & v_{2}<v<v_{3}, \\ \Sigma(v)=\psi(v), & v=v_{3},\end{cases}
$$

with the value matching conditions

$$
\begin{cases}\Sigma\left(v_{1}\right)=\psi\left(v_{1}\right) ; & \Sigma^{\prime}\left(v_{1}\right)=\psi^{\prime}\left(v_{1}\right)  \tag{4.45}\\ \Sigma\left(v_{2}\right)=\psi\left(v_{2}\right) ; & \Sigma^{\prime}\left(v_{2}\right)=\psi^{\prime}\left(v_{2}\right), \\ \Sigma\left(v_{3}\right)=\psi\left(v_{3}\right) ; & \Sigma^{\prime}\left(v_{3}\right)=\psi^{\prime}\left(v_{3}\right) .\end{cases}
$$

We consider an explicit calculation. We begin with $v_{1}$. We have $\Sigma(v)=1+D v^{\beta}$ for $0<v<v_{1}$ since $\Sigma(0)=1$. It follows that we have the conditions (by (4.45))

$$
\left\{\begin{array}{l}
1+D v_{1}^{\beta}=e^{-\gamma\left(1-\rho^{2}\right)\left(v_{1}-K-H^{e}\left(v_{1}\right)\right)} \\
\beta D v_{1}^{\beta-1}=-\gamma\left(1-\rho^{2}\right)\left(1-\left(H^{e}\right)^{\prime}\left(v_{1}\right)\right) e^{-\gamma\left(1-\rho^{2}\right)\left(v_{1}-K-H^{e}\left(v_{1}\right)\right)}
\end{array}\right.
$$

hence

$$
\begin{equation*}
e^{\gamma\left(1-\rho^{2}\right)\left(v_{1}-K-H^{e}\left(v_{1}\right)\right)}-\frac{\gamma\left(1-\rho^{2}\right) v_{1}}{\beta}\left(1-\left(H^{e}\right)^{\prime}\left(v_{1}\right)\right)=1 . \tag{4.46}
\end{equation*}
$$

On the other hand, by (4.24), we obtain

$$
-\frac{\gamma\left(1-\rho^{2}\right) v_{1}}{\beta}\left(H^{e}(v)\right)^{\prime}=1-e^{\gamma\left(1-\rho^{2}\right) H^{e}\left(v_{1}\right)},
$$

and using this in (4.46) we obtain the equation for $v_{1}$ :

$$
\begin{equation*}
e^{\gamma\left(1-\rho^{2}\right)\left(v_{1}-K-H^{e}\left(v_{1}\right)\right)}+e^{\gamma\left(1-\rho^{2}\right) H^{e}\left(v_{1}\right)}=2+\frac{\gamma v_{1}\left(1-\rho^{2}\right)}{\beta} . \tag{4.47}
\end{equation*}
$$

Proposition 4.4. Assume

$$
\begin{equation*}
(1-a) \frac{e^{\varpi+\gamma K\left(1-\rho^{2}\right)}-1}{e^{\varpi}-1}<1 ; \tag{4.48}
\end{equation*}
$$

then the solution of (4.47), which exists, belongs to $\left(v_{0}, \hat{v}\right)$. Moreover, with any solution $v_{1}$, one can associate a negative constant $D$ such that the first conditions in (4.45) are satisfied.

Proof. See Appendix G.
Having obtained $v_{1}$, we next look for the values of $v_{2}$ and $v_{3}$. For $v_{2}<v<v_{3}, \Sigma(v)=$ $A+C v^{\beta}$, where $A$ and $C$ are constants. We write the conditions as follows (by (4.45)):

$$
\left\{\begin{array}{l}
A+C v_{2}^{\beta}=e^{-\gamma\left(1-\rho^{2}\right)\left(v_{2}-K-H^{e}\left(v_{2}\right)\right)},  \tag{4.49}\\
C \beta v_{2}^{\beta-1}=-\gamma\left(1-\rho^{2}\right)\left(1-\left(H^{e}\right)^{\prime}\left(v_{2}\right)\right) e^{-\gamma\left(1-\rho^{2}\right)\left(v_{2}-K-H^{e}\left(v_{2}\right)\right)}, \\
A+C v_{3}^{\beta}=e^{-\gamma\left(1-\rho^{2}\right)\left(a v_{3}-K\right)}, \\
C \beta v_{3}^{\beta-1}=-a \gamma\left(1-\rho^{2}\right) e^{-\gamma\left(1-\rho^{2}\right)\left(a v_{3}-K\right)} .
\end{array}\right.
$$

We eliminate the constants $A$ and $C$ to obtain a system of two equations with two unknowns, $v_{2}$ and $v_{3}$ :

$$
\left\{\begin{align*}
&\left(1+\frac{1}{\beta} v_{3}\left(1-\rho^{2}\right) a\right) e^{\gamma\left(1-\rho^{2}\right)\left(v_{2}-K-H^{e}\left(v_{2}\right)\right)}= e^{\gamma\left(1-\rho^{2}\right)\left(a v_{3}-K\right)}  \tag{4.50}\\
& \quad\left[2+\frac{\gamma}{\beta} v_{2}\left(1-\rho^{2}\right)-e^{\gamma\left(1-\rho^{2}\right) H^{e}\left(v_{2}\right)}\right] \\
& \frac{\gamma}{\beta} \frac{a\left(1-\rho^{2}\right) e^{\gamma\left(1-\rho^{2}\right)\left(v_{2}-K-H^{e}\left(v_{2}\right)\right)}}{v_{3}^{\beta-1}}=\frac{e^{\gamma\left(1-\rho^{2}\right)\left(a v_{3}-K\right)}}{v_{2}^{\beta-1}}\left[\frac{\gamma}{\beta}\left(1-\rho^{2}\right)+\frac{1-e^{\gamma\left(1-\rho^{2}\right) H^{e}\left(v_{2}\right)}}{v_{2}}\right]
\end{align*}\right.
$$

Unfortunately this nonlinear system of algebraic equations is not easy to study analytically. Even if we can solve it, it will remain to prove that the two-interval solution is indeed a solution of the V.I.
4.3.4. Optimal rules in the case of incomplete markets. In the incomplete market situation, we need a weak formulation for the leader's problem V.I. The optimal stopping is the first time when the solution and the obstacle coincide. However, we cannot state that the optimal strategy is characterized by two intervals, as is the case in the situation of complete markets. Once the leader has entered into the market, the follower will enter according to the threshold strategy $\hat{v}$ defined by (4.11). The single decision maker's optimal entry time is the same as in the complete market case if the correlation between market risk and investment risk approaches unity (i.e., $|\rho| \rightarrow 1$ ) or $\gamma$ approaches zero (i.e., a risk-neutral investor). We conclude that market completeness and risk aversion are important inputs into the optimal investment policy.
5. Cash flow payoffs and complete market assumption. We now turn to consider the second type of investment operation incomes characterized by a series of cash flows (cf. section 2.3).
5.1. Single player. We assume that the cash flow process $Y(t)$ from investment operation evolves as

$$
\begin{align*}
d Y(t) & =\alpha d t+\varsigma d W(t)  \tag{5.1}\\
& =(\alpha-\lambda \varsigma) d t+\varsigma d \widehat{W}(t), \quad Y(0)=y, \tag{5.2}
\end{align*}
$$

where $\alpha$ and $\varsigma$ are constants. The reason for modeling the stochastic investment cash flow stream as an arithmetic Brownian motion is the recognition of the possibility of loss from operations. Negative values from the arithmetic Brownian motion may be interpreted as a loss generated from operations.

If the firm exploits the investment opportunity by paying cost $K$, it will obtain a continuous cash flow $\delta Y(t)$ per unit time. At time $t$, the project value, $V(t)$, from the operation is the expected discounted cash flow stream under the risk-neutral probability measure given as

$$
\begin{equation*}
V(t)=\delta \widehat{E}\left[\int_{t}^{\infty} e^{-r(s-t)} Y_{y}(s) d s \mid \mathcal{F}_{t}\right]=\delta\left(\frac{Y_{y}(t)}{r}+\frac{\alpha-\lambda \varsigma}{r^{2}}\right) . \tag{5.3}
\end{equation*}
$$

From (5.3), the expected discounted payoff from the capital investment project undertaken at time $\tau$ is

$$
\begin{equation*}
J_{y}(\tau)=\widehat{E}\left[e^{-r \tau}\left(\delta\left(\frac{Y_{y}(\tau)}{r}+\frac{\alpha-\lambda \varsigma}{r^{2}}\right)-K\right) \mathbb{1}_{\tau<\infty}\right] . \tag{5.4}
\end{equation*}
$$

The firm's objective is to find an optimal stopping time to maximize the expected discounted payoff. That is,

$$
\begin{equation*}
F(y)=\sup _{\tau \geq 0} J_{y}(\tau) . \tag{5.5}
\end{equation*}
$$

Assuming $F(y)$ is sufficiently smooth, we can write the V.I. in the strong sense that $F(y)$ must satisfy as a consequence of dynamic programming

$$
\left\{\begin{array}{l}
(\alpha-\lambda \varsigma) F^{\prime}(y)+\frac{1}{2} \varsigma^{2} F^{\prime \prime}(y)-r F(y) \leq 0,  \tag{5.6}\\
F(y) \geq\left(\delta\left(\frac{y}{r}+\frac{\alpha-\lambda \varsigma}{r^{2}}\right)-K\right), \\
{\left[F(y)-\left(\delta\left(\frac{y}{r}+\frac{\alpha-\lambda \varsigma}{r^{2}}\right)-K\right)\right]\left[(\alpha-\lambda \varsigma) F^{\prime}(y)+\frac{1}{2} \varsigma^{2} F^{\prime \prime}(y)-r F(y)\right]=0,} \\
F(y) \geq 0 ; \quad F \text { has linear growth at } \infty
\end{array}\right.
$$

The solution is $C^{1}(-\infty, \infty)$ and piecewise $C^{2}$. This suffices to give a meaning to (5.6) for almost all $y$. Such a solution, $F(y)$, if it exists, will be unique since, by a classical verification argument, it will coincide with (5.5).

Theorem 5.1. Let

$$
\begin{equation*}
\beta=-\frac{\alpha-\lambda \varsigma}{\varsigma^{2}}+\sqrt{\left(\frac{\alpha-\lambda \varsigma}{\varsigma^{2}}\right)^{2}+\frac{2 r}{\varsigma^{2}}}>0 \tag{5.7}
\end{equation*}
$$

and

$$
\begin{equation*}
\hat{y}=\frac{1}{\beta}+\frac{r}{\delta} K-\frac{\alpha-\lambda \varsigma}{r}>0 ; \tag{5.8}
\end{equation*}
$$

then there exists a unique solution $F$ of (5.6), which is $C^{1}$, given by

$$
F(y)= \begin{cases}\frac{\delta}{r \beta} \exp \{-\beta(\hat{y}-y)\} & \text { if } y \leq \hat{y}  \tag{5.9}\\ \delta\left(\frac{y}{r}+\frac{\alpha-\lambda \varsigma}{r^{2}}\right)-K & \text { if } y \geq \hat{y}\end{cases}
$$

Proof. See Appendix H.
We next define the optimal stopping rule $\hat{\tau}(y)$ that achieves the supremum in (5.5). From general results on V.I., we have

$$
\hat{\tau}(y)=\inf \left\{t \left\lvert\, F\left(Y_{y}(t)\right)=\delta\left(\frac{Y_{y}(t)}{r}+\frac{\alpha-\lambda \varsigma}{r^{2}}\right)-K\right.\right\}=\inf \left\{t \mid G\left(Y_{y}(t)\right)=0\right\}
$$

where $G(y)=F(y)-\delta\left(\frac{y}{r}+\frac{\alpha-\lambda \varsigma}{r^{2}}\right)+K$. From the proof of Theorem 5.1, we see that $G(y)$ decreases from $K$ to zero on the interval $(0, \hat{y})$ and remains zero for $y>\hat{y}$. Thus, $Y_{y}(t)=\hat{y}$ is the solution to the above equation. Therefore, we can write

$$
\begin{equation*}
\hat{\tau}(y)=\inf \left\{t \mid Y_{y}(t) \geq \hat{y}\right\} . \tag{5.10}
\end{equation*}
$$

By (5.10), the manager undertakes the capital investment as soon as the cash flow process reaches the threshold $\hat{y}$ from below. Using $Y_{y}(\hat{\tau}(y))=\hat{y}$ if $\hat{\tau}(y)<\infty$ and $y<\hat{y}$, we have the probabilistic representation of value function $F(y)$ :

$$
\begin{equation*}
F(y)=\frac{\delta}{r \beta} \widehat{E}[\exp (-r \hat{\tau}(y))] \quad \text { if } y<\hat{y} \tag{5.11}
\end{equation*}
$$

5.2. Two players: A Stackelberg game. We now consider the situation of section 3.2 but with the investment payoff as a series of cash flows. Upon entry, the leader receives a continuous cash flow $\delta_{1} Y(t)$ per unit time prior to the follower's entry. Once both have entered, each gets a continuous cash flow $\delta_{2} Y(t)$ per unit time, with $\delta_{2}<\delta_{1} .{ }^{11}$
5.2.1. Statement of the leader's problem. As in section 3.2.1, the follower's optimal stopping is

$$
\begin{equation*}
\hat{\tau}_{\theta}=\theta+\hat{\tau}\left(Y_{y}(\theta)\right) \tag{5.12}
\end{equation*}
$$

with considerations similar to those for (3.9) and we can write explicitly the Laplace transform of the conditional density as

$$
\begin{equation*}
\widehat{E}\left[e^{-r\left(\hat{\tau}_{\theta}-\theta\right)} \mid \mathcal{F}_{\theta}\right]=\mathbb{1}_{Y_{y}(\theta) \geq \hat{y}}+\mathbb{1}_{Y_{y}(\theta)<\hat{y}} e^{-\beta\left(\hat{y}-Y_{y}(\theta)\right)} . \tag{5.13}
\end{equation*}
$$

When the leader enters at time $\theta<\infty$, by paying cost $K$, he/she receives a continuous cash flow $\delta_{1} Y_{y}(t)$ per unit time prior to the follower's entry and gets a continuous cash flow $\delta_{2} Y_{y}(t)$ per unit time after the follower's entry. By anticipating the follower's optimal entry at time $\hat{\tau}_{\theta}$, entering at time $\theta<\infty$, the leader's expected discounted payoff is

$$
\begin{aligned}
& -K+\delta_{1} \widehat{E}\left[\int_{\theta}^{\hat{\tau}_{\theta}} e^{-r(s-\theta)} Y(s) d s \mid \mathcal{F}_{\theta}\right]+\delta_{2} \widehat{E}\left[\mathbb{1}_{\hat{\tau}_{\theta}<\infty} \int_{\hat{\tau}_{\theta}}^{\infty} e^{-r(s-\theta)} Y(s) d s \mid \mathcal{F}_{\theta}\right] \\
& =-K+\delta_{2}\left(\frac{Y_{y}(\theta)}{r}+\frac{\alpha-\lambda \varsigma}{r^{2}}\right) \\
& \quad+\left(\delta_{1}-\delta_{2}\right) \mathbb{1}_{Y_{y}(\theta)<\hat{y}}\left[\frac{Y_{y}(\theta)}{r}+\frac{\alpha-\lambda \varsigma}{r^{2}}-\left(\frac{1}{r \beta}+\frac{K}{\delta_{2}}\right) e^{-\beta\left(\hat{y}-Y_{y}(\theta)\right)}\right]
\end{aligned}
$$

where we use the fact that $\frac{\hat{y}}{r}+\frac{\alpha-\lambda \varsigma}{r^{2}}=\frac{1}{r \beta}+\frac{K}{\delta_{2}}$. To facilitate the presentation, we define (5.14)

$$
\Psi(y)=-K+\delta_{2}\left(\frac{y}{r}+\frac{\alpha-\lambda \varsigma}{r^{2}}\right)+\left(\delta_{1}-\delta_{2}\right) \mathbb{1}_{y<\hat{y}}\left[\frac{y}{r}+\frac{\alpha-\lambda \varsigma}{r^{2}}-\left(\frac{1}{r \beta}+\frac{K}{\delta_{2}}\right) e^{-\beta(\hat{y}-y)}\right] .
$$

The leader's objective is to find a stopping time, $\theta$, to maximize the expected discounted payoff:

$$
\begin{equation*}
L(y)=\sup _{\theta \geq 0} \widehat{E}\left[e^{-r \theta} \Psi\left(Y_{y}(\theta)\right) \mathbb{1}_{\theta<\infty}\right] . \tag{5.15}
\end{equation*}
$$

The leader's value function $L(y)$ must satisfy

$$
\begin{equation*}
L(y) \geq 0 ; \quad L(y) \geq \Psi(y) \tag{5.16}
\end{equation*}
$$

The obstacle, $\Psi(y)$, presents a challenge because it is $C^{0}(-\infty, \infty)$, not $C^{1}(-\infty, \infty)$. The only point of nondifferentiability is $\hat{y}$. In addition, $\Psi(y)$ is unbounded as $y \rightarrow \pm \infty$, which we observe from the following. By (5.14), we may express $\Psi(y)$ as

$$
\begin{equation*}
\Psi(y)=\frac{\delta_{2}}{r}\left[\frac{1}{\beta}+y-\hat{y}+\frac{\delta_{1}-\delta_{2}}{\delta_{2}} \mathbb{1}_{y<\hat{y}}\left(y-\hat{y}+\mu\left(1-e^{-\beta(\hat{y}-y)}\right)\right)\right], \tag{5.17}
\end{equation*}
$$

where $\mu=\hat{y}+\frac{\alpha-\lambda \varsigma}{r}=\frac{\delta_{2}+r \beta K}{\beta \delta_{2}}$.

[^74]5.2.2. The leader's problem V.I. As in section 3.2.2, the nondifferentiability point of the obstacle function requires us to write the variational formulation of $(5.15)$ in the weak sense. We introduce the useful functional spaces, the Hilbert space $L_{\varrho}^{2}(-\infty, \infty)$ defined by (3.21), and the Sobolev space $H_{\varrho}^{1}(-\infty, \infty)$ defined by (3.22), with the corresponding scalar products defined by (3.23) and (3.24), respectively. We define on $H_{\varrho}^{1}(-\infty, \infty)$ the bilinear form
\[

$$
\begin{align*}
b(\Phi, \tilde{\Phi})= & -\int_{-\infty}^{\infty}\left(\alpha-\lambda \varsigma+\frac{\varrho y \varsigma^{2}}{1+y^{2}}\right) \Phi^{\prime}(y) \tilde{\Phi}^{\prime}(y) w(y) d y+\frac{1}{2} \int_{-\infty}^{\infty} \varsigma^{2} \Phi^{\prime}(y) \tilde{\Phi}(y) w(y) d y \\
& +r \int_{-\infty}^{\infty} \Phi(y) \tilde{\Phi}(y) w(y) d y \tag{5.18}
\end{align*}
$$
\]

We consider the convex subset

$$
\begin{equation*}
\mathcal{K}=\left\{\Phi \in H_{\varrho}^{1}(-\infty, \infty) \mid \Phi(y) \geq \Psi(y) \forall y\right\} \tag{5.19}
\end{equation*}
$$

where $\mathcal{K}$ is not empty since it contains $\Psi$.
The V.I. corresponding to $(5.15)$ is

$$
\begin{equation*}
b(L, \tilde{L}-L) \geq 0 \quad \forall \tilde{L} \in \mathcal{K}, L \in \mathcal{K} \tag{5.20}
\end{equation*}
$$

Theorem 5.2. The value function (5.15) is the unique solution of the V.I. (5.20). Proof. See Appendix I.
5.2.3. Smoothness of the solution. Because the obstacle is nondifferentiable only at a single point, $\hat{y}$, it turns out that the solution $L(y)$ will be smoother than the obstacle. Therefore the V.I. will have a strong formulation:

$$
\left\{\begin{array}{l}
-\frac{1}{2} \varsigma^{2} L^{\prime \prime}(y)-(\alpha-\varsigma \lambda) L^{\prime}(y)+r L(y) \geq 0  \tag{5.21}\\
L(y) \geq \Psi(y) \\
(L(y)-\Psi(y))\left[-\frac{1}{2} \varsigma^{2} L^{\prime \prime}(y)-(\alpha-\varsigma \lambda) L^{\prime}(y)+r L(y)\right]=0
\end{array}\right.
$$

Note that $\Psi(y)$ satisfies

$$
\left\{\begin{array}{l}
-\frac{1}{2} \varsigma^{2} \Psi^{\prime \prime}(y)-(\alpha-\varsigma \lambda) \Psi^{\prime}(y)+r \Psi=\delta_{1} y-r K, \quad y<\hat{y}  \tag{5.22}\\
\Psi(\hat{y})=\frac{\delta_{2}}{r} \hat{y}+\frac{\delta^{2}}{r^{2}}(\alpha-\varsigma \lambda)-K=\frac{\delta_{2}}{r \beta}
\end{array}\right.
$$

We would like $L \geq 0$.
Consider

$$
\begin{aligned}
& u(y)=L(y)-\left(\frac{\delta_{2}}{r} y+\frac{\delta_{2}}{r^{2}}(\alpha-\varsigma \lambda)\right)+K \\
& m(y)=\Psi(y)-\left(\frac{\delta_{2}}{r} y+\frac{\delta_{2}}{r^{2}}(\alpha-\varsigma \lambda)\right)+K
\end{aligned}
$$

giving us

$$
\left\{\begin{array}{l}
-\frac{1}{2} \varsigma^{2} u^{\prime \prime}(y)-(\alpha-\varsigma \lambda) u^{\prime}(y)+r u \geq-\delta_{2} y+r K  \tag{5.23}\\
u(y) \geq m(y) \\
(u(y)-m(y))\left[-\frac{1}{2} \varsigma^{2} u^{\prime \prime}(y)-(\alpha-\varsigma \lambda) u^{\prime}(y)+r u+\delta_{2} y-r K\right]=0 \\
u(y) \geq-\left(\frac{\delta_{2}}{r} y+\frac{\delta_{2}}{r^{2}}(\alpha-\varsigma \lambda)\right)+K
\end{array}\right.
$$

The function $m$ satisfies

$$
\begin{equation*}
-\frac{1}{2} \varsigma^{2} m^{\prime \prime}(y)-(\alpha-\varsigma \lambda) m(y)+r m(y)=\left(\delta_{1}-\delta_{2}\right) y, \quad y<\hat{y} . \tag{5.24}
\end{equation*}
$$

Proposition 5.3. There exists a unique $u \in C^{1}(-\infty, \infty)$, piecewise $C^{2}$ solution of (5.23). This function vanishes for $y$ sufficiently large.

Proof. This is a special case of the more complex problem presented in Theorem 6.5, and thus we refer the reader to the proof of Theorem 6.5.

Next, we are able to arrive at a two-interval solution stated in the proposition that follows.
Proposition 5.4. There exists a triple $y_{1}<y_{2}<\hat{y}<y_{3}$ such that

$$
\left\{\begin{array}{l}
-\frac{1}{2} \varsigma^{2} u^{\prime \prime}(y)-(\alpha-\varsigma \lambda) u^{\prime}(y)+r u=-\delta_{2} y+r K \text { for } y<y_{1} \text { and } y_{2}<y<y_{3}, \\
u(y) \geq m(y), \\
u(y)=m(y) \text { for } y_{1} \leq y \leq y_{2}, \\
u(y)=0 \text { for } y \geq y_{3}, \\
u^{\prime}\left(y_{1}\right)=m^{\prime}\left(y_{1}\right) ; u^{\prime}\left(y_{2}\right)=m^{\prime}\left(y_{2}\right) ; u^{\prime}\left(y_{3}\right)=0 .
\end{array}\right.
$$

Proof. The proof is similar to the proof of Theorem 6.6, a more complex problem. We can define the leader's optimal stopping rule as

$$
\hat{\theta}(y)=\left\{\begin{array}{l}
\inf \left\{t \mid Y_{y}(t) \geq y_{1}\right\} \quad \text { if } y<y_{1},  \tag{5.25}\\
0 \text { if } y_{1} \leq y \leq y_{2}, \\
\inf \left\{t \mid Y_{y}(t) \leq y_{2} \text { or } Y_{y}(t) \geq y_{3}\right\} \quad \text { if } y_{2}<y<y_{3} \\
0 \text { if } y \geq y_{3}
\end{array}\right.
$$

5.2.4. Optimal rules in the case of complete markets. As in section 3.2.4, the leader's optimal stopping rule (i.e., optimal investment strategy) is characterized as a two-interval solution, which is $\left[y_{1}, y_{2}\right]$ and $\left[y_{3}, \infty\right)$ with the relation such that $y_{1}<y_{2}<\hat{y}<y_{3}$, where $\hat{y}$ is the follower's optimal investment trigger (same as the monopolist's). It is observed that the optimal stopping for both parties differs from the monopolist's. Although the follower's optimal investment trigger coincides with the monopolist's, he/she does not act the same as the monopolist due to the concern of strategic interactions from the leader. The leader will invest nonoptimally if he/she ignores the strategic effect of competition.

## 6. Cash flow payoffs and incomplete market assumption.

6.1. Utility-based pricing model. We consider the model of section 4.1 with a cash flow process rather than a project value process. The asset $S$ representing the market still evolves as (2.1). The cash flow process $Y$ of (5.1), generated by the capital investment project, evolves as

$$
\begin{equation*}
d Y(t)=\alpha d t+\varsigma\left(\rho d W(t)+\sqrt{1-\rho^{2}} d W^{0}(t)\right) \tag{6.1}
\end{equation*}
$$

where $W(t)$ and $W^{0}(t)$ are independent Wiener processes.

The rational utility-maximizing individual investor's exponential (i.e., constant absolute risk aversion (CARA)) utility function is given by

$$
\begin{equation*}
U(C)=-\frac{1}{\gamma} e^{-\gamma C}, \tag{6.2}
\end{equation*}
$$

where the argument $C$ is the investor's consumption.
Remark 6.1.1. We allow for negative consumption. For $C \in \mathbb{R}, U$ increases from $-\infty$ to 0 . As $C \rightarrow-\infty$, it leads to huge negative values. We interpret this effect as a penalty to the utility maximization investor. We could, of course, impose the constraint of nonnegative consumption. However, imposing nonnegativity on the consumption would rule out the analytical solutions for further developments, a property we would like to retain for the full analysis. Therefore, we choose to accept negative consumption which could lead to huge negative utility values (big penalties for our utility maximization investor) instead of imposing the nonnegativity constraint on the consumption. We also note that the negative consumption occurs when $x$ or $y$ becomes very negative and we cannot avoid this situation since $x, y \in \mathbb{R}$.

The rational investor maximizes his/her expected utility of consumption. Given the initial wealth, $x$, the risk-averse investor optimizes his/her portfolio by dynamically choosing allocations in the market asset $S$, the riskless bond, and the consumption rate $C$. The investor's wealth, $X$, evolves as

$$
\left\{\begin{array}{l}
d X(t)=\pi(t) X(t) \sigma(\lambda d t+d W(t))+r X(t) d t-C(t) d t, t<\tau,  \tag{6.3}\\
X(\tau)=X(\tau-0)-K, \\
d X(t)=\pi(t) X(t) \sigma(\lambda d t+d W(t))+r X(t) d t-C(t) d t+\delta Y(t) d t, t>\tau, \\
d Y(t)=\alpha d t+\varsigma\left(\rho d W(t)+\sqrt{1-\rho^{2}} d W^{0}(t)\right), \\
X(0)=x, Y(0)=y, \text { with } x, y \in \mathbb{R},
\end{array}\right.
$$

where $\pi(t)$ is the ratio of wealth invested in asset $S, C(t)$ is the consumption rate, and $\tau$ is a stopping time, chosen optimally by the investor. At $\tau$, he/she invests in the project by paying an investment cost $K$. Thus, there is a jump in the wealth process, making it discontinuous at $\tau$. After $\tau$, the wealth process evolves differently since the investor receives an additional income stream, i.e., a cash flow stream at a rate $\delta Y(t)$ per unit time. Prior to $\tau$, allocations in the market asset are for portfolio investment purposes and provide partial hedges against the uncertain capital investment payoff. After $\tau$, the investor solves the optimal investment/consumption portfolio problem to invest his/her total wealth.

We see from (6.3) that the wealth process has two possible evolution regimes. To facilitate further representation, we introduce the processes, regime $0, X^{0}$ and regime $1, X^{1}$ :

$$
\begin{align*}
& d X^{0}(t)=\pi(t) X^{0}(t) \sigma(\lambda d t+d W(t))+r X^{0}(t) d t-C(t) d t, \quad X^{0}(0)=x,  \tag{6.4}\\
& d X^{1}(t)=\pi(t) X^{1}(t) \sigma(\lambda d t+d W(t))+r X^{1}(t) d t-C(t) d t+\delta Y(t) d t . \tag{6.5}
\end{align*}
$$

The wealth process $X$ corresponds to $X^{0}$ before the stopping time $\tau$ and $X^{1}$ after $\tau$.
After $\tau$, the investor solves a control problem of portfolio selection and consumption rate, augmented by the stochastic income $\delta Y(t)$. With a pair of $(C(\cdot), \pi(\cdot))$ we associate the
objective function

$$
\begin{equation*}
J_{x, y}(C(\cdot))=E\left[\int_{0}^{\infty} e^{-\mu t} U(C(t)) d t\right] \tag{6.6}
\end{equation*}
$$

where $\mu$ is the discount rate. This function may take the value $-\infty$. We define the utility maximization control problem as

$$
\begin{equation*}
F^{1}(x, y)=\sup _{\{\pi(\cdot), C(\cdot)\} \in \mathcal{U}_{x, y}^{1}} J(C(\cdot)), \tag{6.7}
\end{equation*}
$$

where $\mathcal{U}_{x, y}^{1}$ is the set of admissible controls to be defined later (cf. section 6.2.1).
At time $\tau$, the investor stops, receiving $F^{1}\left(X^{0}(\tau)-K, Y(\tau)\right)$. With a triple of $(C(\cdot), \pi(\cdot), \tau)$ we associate the objective function

$$
\begin{equation*}
J_{x, y}(C(\cdot), \pi(\cdot), \tau)=E\left[\int_{0}^{\tau} e^{-\mu t} U(C(t)) d t+F^{1}\left(X^{0}(\tau)-K, Y(\tau)\right) e^{-\mu \tau}\right] \tag{6.8}
\end{equation*}
$$

We assume $\tau$ is finite a.s. The investor's problem is to maximize the expected discounted utility from consumption by choosing stopping time $\tau$, consumption rate $C$, and portfolio investment strategy $\pi$. We define the associated value function as

$$
\begin{equation*}
F(x, y)=\sup _{\{\pi(\cdot), C(\cdot), \tau\} \in \mathcal{U}_{x y}^{0}} J_{x, y}(C(\cdot), \pi(\cdot), \tau) \tag{6.9}
\end{equation*}
$$

with $\mathcal{U}_{x y}^{0}$ a set of admissible controls to be defined later (cf. section 6.2.2).
6.2. Single player. In this section, we solve a single player's problem, and, as such, the manager/investor has no need to consider the actions of any other firm. By paying cost $K$, the firm expects to receive a continuous cash flow $\delta Y(t)$ per unit time. The investor's wealth evolves as (6.3), for which we introduce regimes (6.4) and (6.5) to facilitate exposition.

The investor's problem is to maximize the expected discounted utility from consumption with respect to investment time $\tau$, consumption rate $C$, and investment strategy $\pi$. As described in section 6.1, it is necessary to solve the problem in two steps: post- and preinvestment utility maximization, i.e., after the stopping time $\tau$ and prior to $\tau$. The obstacle used in solving preinvestment utility maximization is defined by the solution of the postinvestment utility maximization. We proceed in detail in the following sections.
6.2.1. Postinvestment utility maximization. Before considering the optimal stopping problem, we begin with a control problem relative to the process $X^{1}(t)$. That is, we assume the capital investment project has been undertaken, and we solve the investor's utility maximization as a control problem of portfolio selections and consumption rules, augmented by a stochastic income stream, $\delta Y(t)$ per unit time.

To facilitate the notation, for $\mathcal{F}_{t}$-adapted processes $\pi(t)$ and $C(t)$, we introduce the local integrability condition $I^{i}, i=0,1$ :

$$
I^{i}=\left\{\begin{array}{l}
E \int_{0}^{T}\left(\pi(t) X^{i}(t)\right)^{2} d t<\infty \quad \forall T  \tag{6.10}\\
E \int_{0}^{T}(C(t))^{2} d t<\infty \quad \forall T
\end{array}\right.
$$

and we also define

$$
\tau_{N}^{i}=\inf \left\{t \mid \mathbb{1}_{i=1}\left(r X^{1}(t)+\delta Y(t)\right)+\mathbb{1}_{i=0} X^{0}(t) \leq-N\right\}, \quad i=0,1
$$

To a pair of $(C(\cdot), \pi(\cdot))$ we introduce the objective function

$$
\begin{equation*}
J_{x, y}(C(\cdot))=E\left[\int_{0}^{\infty} e^{-\mu t} U(C(t)) d t\right] . \tag{6.11}
\end{equation*}
$$

This function may take the value $-\infty$. The investor's problem is to maximize his/her expected discounted utility from consumption. We define the associated value function as

$$
\begin{equation*}
F^{1}(x, y)=\sup _{\{\pi(\cdot), C(\cdot)\} \in \mathcal{U}_{x, y}^{1}} J_{x, y}(C(\cdot)) \tag{6.12}
\end{equation*}
$$

where $\mathcal{U}_{x, y}^{1}=\left\{(\pi, C): I^{1} ; \tau_{N}^{1} \uparrow \infty\right.$ a.s. as $N \uparrow \infty ; e^{-\mu T} E\left[e^{-\gamma\left(r X^{1}(T)+\delta Y(T)\right)}\right] \rightarrow 0$ as $\left.T \rightarrow \infty\right\}$. We associate the value function $F^{1}(x, y)$ with the Bellman equation

$$
\begin{align*}
-\mu F^{1} & +\frac{\partial F^{1}}{\partial x}(r x+\delta y)+\frac{\partial F^{1}}{\partial y} \alpha+\frac{1}{2} \frac{\partial^{2} F^{1}}{\partial y^{2}} \varsigma^{2}+\sup _{C}\left\{U(C)-C \frac{\partial F^{1}}{\partial x}\right\} \\
& +\sup _{\pi}\left\{\pi x \sigma\left(\lambda \frac{\partial F^{1}}{\partial x}+\varsigma \rho \frac{\partial^{2} F^{1}}{\partial x y}\right)+\frac{1}{2} \frac{\partial^{2} F^{1}}{\partial x^{2}} \pi^{2} x^{2} \sigma^{2}\right\}=0 . \tag{6.13}
\end{align*}
$$

Define the feedbacks

$$
\left\{\begin{array}{l}
\widehat{C}(x, y)=-\frac{1}{\gamma} \ln \frac{\partial F^{1}}{\partial x}  \tag{6.14}\\
\widehat{\pi}(x, y) \sigma x=-\frac{\lambda+\frac{\partial F^{1}}{\partial x}+\varsigma \rho \frac{\partial^{2} F^{1}}{\partial x \partial y}}{\frac{\partial^{2} F^{1}}{\partial x^{2}}}
\end{array}\right.
$$

and (6.13) may be rewritten as

$$
\begin{array}{r}
-u F^{1}+\frac{\partial F^{1}}{\partial x}\left(r x+\delta y-\frac{1}{\gamma}+\frac{1}{\gamma} \ln \frac{\partial F^{1}}{\partial x}\right)+\frac{\partial F^{1}}{\partial y} \alpha+\frac{1}{2} \frac{\partial F^{1}}{\partial y^{2}} \varsigma^{2} \\
-\frac{1}{2} \frac{\left(\lambda \frac{\partial F^{1}}{\partial x}+\varsigma \rho \frac{\partial^{2} F^{1}}{\partial x \partial y}\right)^{2}}{\frac{\partial^{2} F^{1}}{\partial x^{2}}}=0 . \tag{6.15}
\end{array}
$$

We look for a solution of (6.15) as

$$
\begin{equation*}
F^{1}(x, y)=-\frac{1}{r \gamma} \exp \left\{-r \gamma\left(x+\frac{\delta}{r} y\right)+1-\frac{\mu+\frac{\lambda^{2}}{2}}{r}-\frac{\delta \gamma}{r}\left(\alpha-\lambda \varsigma \rho-\frac{1}{2} \varsigma^{2} \delta \gamma\left(1-\rho^{2}\right)\right)\right\} . \tag{6.16}
\end{equation*}
$$

Theorem 6.1. The function $F^{1}(x, y)$ defined by (6.16) coincides with the value function (6.12).

Proof. See Appendix J.
6.2.2. Preinvestment utility maximization. After obtaining the solution to postinvestment utility maximization in section 6.2 .1 , we solve the full problem by making use of the value function $F^{1}(x, y)$. That is, we now turn to the problem of optimal stopping. Before the stopping time $\tau$, the wealth process is governed by (6.4), and the cash flow evolves as (6.1). At time $\tau$, the investor stops, receiving $F^{1}\left(X^{0}(\tau)-K, Y(\tau)\right)$. Therefore, the objective function is

$$
\begin{equation*}
J_{x, y}(C(\cdot), \pi(\cdot), \tau)=E\left[\int_{0}^{\tau} e^{-\mu t} U(C(t)) d t+F^{1}\left(X^{0}(\tau)-K, Y(\tau)\right) e^{-\mu \tau}\right] \tag{6.17}
\end{equation*}
$$

and we define the value function

$$
\begin{equation*}
F(x, y)=\sup _{\{\pi(\cdot), C(\cdot), \tau\} \in \mathcal{U}_{x y}^{0}} J_{x, y}(C(\cdot), \pi(\cdot), \tau), \tag{6.18}
\end{equation*}
$$

where $\mathcal{U}_{x y}^{0}=\left\{(\pi, C, \tau): I^{0} ; \tau<\infty\right.$ a.s.; $\tau^{*}=\lim \uparrow \tau_{N}^{0} \geq \tau$ a.s. $\}$.
As a consequence of dynamic programming, assuming sufficient smoothness of $F(x, y)$, we write the V.I. in the strong sense that $F(x, y)$ must satisfy

$$
\left\{\begin{array}{c}
-\mu F+\frac{\partial F}{\partial x} r x+\frac{\partial F}{\partial y} \alpha+\frac{1}{2} \frac{\partial^{2} F}{\partial y^{2}} \varsigma^{2}+\sup _{C}\left(U(C)-C \frac{\partial F}{\partial x}\right)+\sup _{\pi}\left[\pi x \sigma\left(\lambda \frac{\partial F}{\partial x}+\varsigma \rho \frac{\partial^{2} F}{\partial x \partial y}\right)\right.  \tag{6.19}\\
\left.\quad+\frac{1}{2} \pi^{2} x^{2} \sigma^{2} \frac{\partial^{2} F}{\partial x^{2}}\right] \leq 0, \\
F(x, y) \geq F^{1}(x-K, y), \\
\left(F(x, y)-F^{1}(x-K, y)\right)\left[-\mu F+\frac{\partial F}{\partial x} r x+\frac{\partial F}{\partial y} \alpha+\frac{1}{2} \frac{\partial^{2} F}{\partial y^{2}} \varsigma^{2}+\sup _{C}\left(U(C)-C \frac{\partial F}{\partial x}\right)\right. \\
\left.\quad+\sup _{\pi}\left[\pi x \sigma\left(\lambda \frac{\partial F}{\partial x}+\varsigma \rho \frac{\partial^{2} F}{\partial x \partial y}\right)+\frac{1}{2} \pi^{2} x^{2} \sigma^{2} \frac{\partial^{2} F}{\partial x^{2}}\right]\right]=0,
\end{array}\right.
$$

where in the curly brace we find the nonlinear 2 nd order differential operator appearing at the left-hand side of the first inequality. Since $x, y \in \mathbb{R}$, there are no boundary conditions. We look for a solution:

$$
\begin{equation*}
F(x, y)=-\frac{1}{r \gamma} \exp \left[-r \gamma(x+g(y))+1-\frac{\mu+\frac{\lambda^{2}}{2}}{r}\right] . \tag{6.20}
\end{equation*}
$$

Going back to the expression of $F^{1}(x, y)$ in (6.16), it will be convenient to introduce

$$
\begin{equation*}
f(y)=\frac{\delta y}{r}+\frac{\delta}{r^{2}}\left(\alpha-\lambda \varsigma \rho-\frac{1}{2} \varsigma^{2} \rho \gamma\left(1-\rho^{2}\right)\right) \tag{6.21}
\end{equation*}
$$

so that

$$
\begin{equation*}
F^{1}(x, y)=-\frac{1}{r \gamma} \exp \left[-r \gamma(x+f(y))+1-\frac{\mu+\frac{\lambda^{2}}{2}}{r}\right] . \tag{6.22}
\end{equation*}
$$

Define the feedbacks

$$
\left\{\begin{array}{l}
\widehat{C}(x, y)=r(x+g(y))-\frac{1}{\gamma}\left(1-\frac{\mu+\frac{\lambda^{2}}{2}}{r}\right),  \tag{6.23}\\
\widehat{\pi}(x, y)=\frac{\lambda}{r \gamma}-\varsigma \rho g^{\prime}(y)
\end{array}\right.
$$

Substituting (6.23) and the functions $F(x, y)$ and $F^{1}(x, y)$, defined in (6.20) and (6.22), respectively, into (6.19), the V.I. (6.19) that $F(x, y)$ must satisfy reduces to the following V.I. expressed in terms of $g(y)$ :

$$
\left\{\begin{array}{l}
\frac{1}{2} \varsigma^{2} g^{\prime \prime}+g^{\prime}(\alpha-\lambda \varsigma \rho)-\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) g^{\prime 2}-r g \leq 0,  \tag{6.24}\\
g(y) \geq f(y)-K, \\
(g(y)-f(y)+K)\left[\frac{1}{2} \varsigma^{2} g^{\prime \prime}+g^{\prime}(\alpha-\lambda \varsigma \rho)-\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) g^{\prime 2}-r g\right]=0
\end{array}\right.
$$

This V.I. cannot be interpreted simply as a control problem. Indeed, the nonlinear operator is connected to a minimization problem and the inequalities are connected to a maximization problem for a stopping time. Therefore, $g(y)$ is more appropriately the value function of a differential game than of a control problem.

Considering $u(y)=g(y)-f(y)+K$, the V.I. (6.24) becomes

$$
\left\{\begin{array}{l}
-\frac{1}{2} \varsigma^{2} u^{\prime \prime}-u^{\prime}\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta \gamma\left(1-\rho^{2}\right)\right)+\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) u^{\prime 2}+r u \geq-\delta y+r K,  \tag{6.25}\\
u \geq 0, \\
u\left[-\frac{1}{2} \varsigma^{2} u^{\prime \prime}-u^{\prime}\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta \gamma\left(1-\rho^{2}\right)\right)+\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) u^{\prime 2}+r u+\delta y-r K\right]=0 .
\end{array}\right.
$$

We study (6.25) by the threshold approach. For $\hat{y}$ fixed, we solve the Dirichlet problem

$$
\left\{\begin{array}{l}
-\frac{1}{2} \varsigma^{2} u^{\prime \prime}-u^{\prime}\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta \gamma\left(1-\rho^{2}\right)\right)+\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) u^{\prime 2}+r u=-\delta y+r K, \quad y<\hat{y},  \tag{6.26}\\
u(\hat{y})=0
\end{array}\right.
$$

and we require linear growth for $y \rightarrow-\infty$.
Theorem 6.2. For each $\hat{y}$, there exists a unique solution of (6.26) with the estimate

$$
\begin{equation*}
-\frac{(-\delta \hat{y}+r K)^{-}}{r} \leq u(y) \leq-\frac{\delta}{r}(y-\hat{y})+\frac{1}{r}\left(-\delta \hat{y}+r K-\frac{\delta}{r}\left(\alpha-\lambda \varsigma \rho-\gamma \delta \varsigma^{2}\left(1-\rho^{2}\right)\right)\right)^{+} \tag{6.27}
\end{equation*}
$$

The solution is $C^{2}$ in $(-\infty, \hat{y})$.
There exists a unique $\hat{y}$ such that

$$
\begin{equation*}
u^{\prime}(\hat{y})=0, \quad \hat{y} \geq \frac{r K}{\delta} \tag{6.28}
\end{equation*}
$$

The corresponding solution of (6.26) extended by 0 beyond $\hat{y}$ is the unique solution of the V.I. (6.25). It is $C^{1}$ and piecewise $C^{2}$.

Proof. See Appendix K.
Returning to (6.24), from Theorem 6.2, we have proven that there exists a unique $g(y) \in$ $C^{1}$ and piecewise $C^{2}$. Moreover, there exists a unique $\hat{y}$ such that

$$
\left\{\begin{array}{l}
-\frac{1}{2} \varsigma^{2} g^{\prime \prime}-g^{\prime}(\alpha-\lambda \varsigma \rho)+\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) g^{\prime 2}+r g=0, \quad y<\hat{y}  \tag{6.29}\\
g(y)=\frac{\delta}{r} y-K+\frac{\delta}{r^{2}}\left(\alpha-\lambda \varsigma \rho-\frac{1}{2} \varsigma^{2} \delta \gamma\left(1-\rho^{2}\right)\right), \quad y \geq \hat{y} \\
g^{\prime}(\hat{y})=\frac{\delta}{r}
\end{array}\right.
$$

The function $g(y)$ is the value function of the stochastic control problem:

$$
\begin{align*}
g(y)=\inf _{v(\cdot)} E & {\left[\int_{0}^{\theta_{y}(v(\cdot))} e^{-r t} \frac{1}{2} v^{2}(t) d t\right.} \\
& \left.+e^{-r \theta_{y}(v(\cdot))}\left[\frac{\delta}{r} y-K+\frac{\delta}{r^{2}}\left(\alpha-\lambda \varsigma \rho-\frac{1}{2} \varsigma^{2} \delta \gamma\left(1-\rho^{2}\right)\right)\right]\right] \tag{6.30}
\end{align*}
$$

with the state equation

$$
\begin{equation*}
d Y(t)=\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta \gamma\left(1-\rho^{2}\right)+\varsigma \sqrt{r \gamma\left(1-\rho^{2}\right)} v(t)\right) d t+\varsigma d W(t), \quad Y(0)=y<\hat{y}, \tag{6.31}
\end{equation*}
$$

where $v(\cdot)$ is adapted and locally square integrable, and $\theta_{y}(v(\cdot))$ is the first time the process (6.31) reaches $\hat{y}$.

As $y \rightarrow-\infty, \theta_{y}(v(\cdot)) \rightarrow \infty$ for any control $v(\cdot)$; hence,

$$
\begin{equation*}
g(y) \rightarrow 0 \text { as } y \rightarrow-\infty ; \quad g(y) \geq 0 . \tag{6.32}
\end{equation*}
$$

The positivity follows from the property $u(y) \geq-f(y)+K$. Indeed, $u(y)=-f(y)+K$ satisfies the first inequality in (6.25) as an equality.

Theorem 6.3. The function $F(x, y)$ given by (6.20) coincides with the value function (6.18). Proof. See Appendix L.
We next define the optimal stopping rule, which achieves supremum in (6.18), as

$$
\hat{\tau}(y)=\inf \left\{t \mid Y_{y}(t) \geq \hat{y}\right\},
$$

where $\hat{y}$ is the unique value defined by the V.I. (6.26) and (6.28).
Remark 6.2.1. When $\rho^{2} \rightarrow 1$, the investor's investment problem converges to the solution of the optimal stopping problem

$$
g(y)=\sup _{\tau} \widehat{E}\left[e^{-r \tau}(V(\tau)-K) \mathbb{1}_{\tau<\infty}\right] .
$$

Proof. When $\rho^{2} \rightarrow 1$, the solution of $g(y)$ in (6.24) coincides with the solution of $F(y)$ in (5.6).

Remark 6.2.2. When $\gamma \rightarrow 0$, the investor's investment problem converges to the solution of the optimal stopping problem under the minimal martingale measure: ${ }^{12}$

$$
g(y)=\sup _{\tau} \widehat{E}\left[e^{-r \tau}(V(\tau)-K) \mathbb{1}_{\tau<\infty}\right] .
$$

The justification is similar to Remark 6.2.1.
Remark 6.2.3. In the incomplete market, the investor's option value to invest, $g(y)$, decreases with respect to the risk aversion parameter.

Proof. By (6.24), it is obvious that $g(y)$ is decreasing with respect to $\gamma$.

[^75]6.3. Two players: A Stackelberg game. We consider the situation of section 5.2 in the case of incomplete markets. The follower's investment problem is similar to that of the single player. After the leader's entry, he/she makes the optimal stopping decision. By paying an investment cost $K$ at time $\tau$, the follower receives a continuous cash flow stream, $\delta_{2} Y(t)$ per unit time. Like the single player, the follower's problem is to maximize the expected discounted utility from consumption by choosing stopping time $\tau$, consumption rate $C$, and investment strategy $\pi$. Thus, the follower's strategy is identical to that described in section 6.2 with $\delta$ becoming $\delta_{2}$. We have the follower's value function as defined by (6.20) with $g(y)$ satisfying (6.24), where $\hat{y}$ is the unique value defined by the V.I. (6.25) and (6.28). We take $\delta=\delta_{2}$.

The optimal stopping strategy for the follower is

$$
\hat{\tau}(y)=\inf \left\{t \mid Y_{y}(t) \geq \hat{y}\right\},
$$

where $\hat{y}$ is the unique value defined by the V.I. (6.25) and (6.28). Note again that we must take $\delta=\delta_{2}$ and thus $\hat{y} \geq \frac{r K}{\delta_{2}}$. The stopping time $\hat{\tau}(y)$ is the optimal entry if the follower can enter the market at time zero. Since the follower enters after the leader (who starts at $\theta$ ), for finite $\theta$, the follower will enter at time

$$
\begin{equation*}
\hat{\tau}_{\theta}=\theta+\hat{\tau}\left(Y_{y}(\theta)\right) \tag{6.33}
\end{equation*}
$$

with considerations similar to those for (3.9).
6.3.1. The leader's problem. By paying cost $K$, the leader expects to receive a continuous cash flow $\delta_{1} Y(t)$ per unit time prior to the follower's entry and $\delta_{2} Y(t)$ per unit time after the follower's entry. The leader's wealth evolution is similar to (6.3), but it is complicated by the fact that the follower will enter according to the optimal stopping rule $\hat{\tau}_{\theta}$. Thus, the wealth evolves according to three regimes, $X^{0}$ (cf. (6.4)), $X^{1}$ (cf. (6.34)), and $X^{2}$ (cf. (6.34)), which correspond to (1) before the leader's stopping, $\theta$, (2) after the leader's stopping but prior to the follower's optimal entry, and (3) after the follower's optimal entry, respectively.

The leader's problem is to maximize the expected discounted utility from consumption by choosing stopping time $\theta$, consumption rate $C$, and investment strategy $\pi$. Again, as described in section 6.1, we solve the utility maximization problem in two steps: post- and preinvestment utility maximization. The leader's postinvestment utility maximization is complicated by the fact that, upon the follower's optimal entry, the leader's stochastic income stream will be changed from $\delta_{1} Y(t)$ per unit time to $\delta_{2} Y(t)$ per unit time. Consequently, the leader must solve the postinvestment utility maximization problem with respect to consumption rules and investment strategies under two stochastic income streams. As in the single player's preinvestment utility maximization, the obstacle for the preinvestment utility maximization is obtained from the solution of postinvestment utility maximization.
6.3.2. Leader's postinvestment utility maximization. As in section 6.2.1, we begin with a control problem assuming the capital investment project has been undertaken. We solve the investor's utility maximization problem of portfolio strategy with stochastic incomes.

A key aspect of the analysis is to define carefully what the leader receives at time $\theta$ when he/she decides to exploit the cash flow $Y(t)$. Suppose $\theta=0$, his/her wealth is $x$, and the cash
flow $y>0$. The wealth becomes $x-K$ immediately since he/she has to pay the fixed cost of entry. The follower will enter at $\hat{\tau}(y)$. We then have the following evolution of wealth:

$$
\left\{\begin{array}{l}
d X^{1}(t)=\pi(t) X^{1}(t) \sigma(\lambda d t+d W(t))+r X^{1}(t) d t+\delta_{1} Y(t) d t-C(t) d t, 0<t<\hat{\tau}(y),  \tag{6.34}\\
X^{1}(0)=x, \\
d X^{2}(t)=\pi(t) X^{2}(t) \sigma(\lambda d t+d W(t))+r X^{2}(t) d t+\delta_{2} Y(t) d t-C(t) d t, t>\hat{\tau}(y), \\
X^{2}(\hat{\tau}(y))=X^{1}(\hat{\tau}(y)), \\
d Y(t)=\alpha d t+\varsigma\left(\rho d W(t)+\sqrt{1-\rho^{2}} d W^{0}(t)\right), \quad Y(0)=y .
\end{array}\right.
$$

If $\theta=0$ and $y \geq \hat{y}$, the follower enters at time 0 and the leader's problem is exactly the same as the follower's, namely problem (6.12) with $\delta=\delta_{2}$. So we consider the function

$$
\begin{equation*}
L^{2}(x, y)=-\frac{1}{r \gamma} e^{-r \gamma(x+f(y))+1-\frac{\mu+\frac{\lambda^{2}}{2}}{r}} \tag{6.35}
\end{equation*}
$$

with

$$
\begin{equation*}
f(y)=\frac{\delta_{2} y}{r}+\frac{\delta_{2}}{r^{2}}\left(\alpha-\lambda \varsigma \rho-\frac{1}{2} \varsigma^{2} \delta_{2} \gamma\left(1-\rho^{2}\right)\right) . \tag{6.36}
\end{equation*}
$$

If $\theta=0$ and $y<\hat{y}$, the leader's problem is described as follows. The wealth process and cash flow process are governed by $X^{1}(t)$ and $Y(t)$ defined in (6.34).

As in the single player case, to facilitate the notation, for $\mathcal{F}_{t}$-adapted processes $\pi(t)$ and $C(t)$, we introduce the local integrability condition $I^{i}, i=0,1$, which is defined in (6.10) and define

$$
\tau_{N}^{i}=\inf \left\{t \mid \mathbb{1}_{i=1} X^{1}(t)+\mathbb{1}_{i=0} X^{0}(t) \leq-N\right\}, \quad i=0,1
$$

At the follower's entry time, $\hat{\tau}(y)$, the leader gets $L^{2}\left(X^{1}(\hat{\tau}(y)), Y(\hat{\tau}(y))\right)$ (cf. (6.35)) as shown in the above case where both market players are in the market. With a pair $(C(\cdot), \pi(\cdot))$ we associate the objective function

$$
\begin{equation*}
J_{x, y}(C(\cdot), \pi(\cdot))=E\left[\int_{0}^{\hat{\tau}(y)} e^{-\mu t} C(t) d(t)+L^{2}\left(X^{1}(\hat{\tau}(y)), Y(\hat{\tau}(y))\right) e^{-\mu \hat{\tau}(y)}\right] \tag{6.37}
\end{equation*}
$$

where we recall $\hat{\tau}(y)<\infty$ a.s. We consider the value function

$$
\begin{equation*}
L^{1}(x, y)=\sup _{\{\pi(\cdot), C(\cdot)\} \in \mathcal{U}}^{1, y} 1 J_{x, y}(C(\cdot), \pi(\cdot)), \tag{6.38}
\end{equation*}
$$

where $\mathcal{U}_{x, y}^{1}=\left\{(\pi, C): I^{1} ; \tau^{*}=\lim \uparrow \tau_{N}^{1} \geq \hat{\tau}(y)\right.$ a.s. $\}$. We associate the value function with the Bellman equation

$$
\left\{\begin{array}{l}
-\mu L^{1}+\frac{\partial L^{1}}{\partial x}\left(r x+\delta_{1} y\right)+\frac{\partial L^{1}}{\partial y} \alpha+\frac{1}{2} \frac{\partial^{2} L^{1}}{\partial y^{2}} \varsigma^{2}+\sup _{C}\left(U(C)-C \frac{\partial L^{1}}{\partial x}\right)  \tag{6.39}\\
\quad+\sup _{\pi}\left[\pi x \sigma\left(\lambda \frac{\partial L^{1}}{\partial x}+\varsigma \rho \frac{\partial^{2} L^{1}}{\partial x \partial y}\right)+\frac{1}{2} \pi^{2} x^{2} \sigma^{2} \frac{\partial^{2} L^{1}}{\partial x^{2}}\right]=0 \\
L^{1}(x, \hat{y})=L^{2}(x, \hat{y}), \quad y<\hat{y}
\end{array}\right.
$$

Define the feedbacks

$$
\left\{\begin{array}{l}
\widehat{C}(x, y)=-\frac{1}{\gamma} \ln \frac{\partial L^{1}}{\partial x}  \tag{6.40}\\
\widehat{\pi}(x, y) \sigma x=-\frac{\lambda+\frac{\partial L^{1}}{\partial x}+\varsigma \rho \frac{\partial^{2} F^{1}}{\partial x \partial y}}{\frac{\partial^{2} L^{1}}{\partial x^{2}}}
\end{array}\right.
$$

Substituting (6.40) into (6.39), we can rewrite (6.39) as

$$
\begin{align*}
-\mu L^{1}+\frac{\partial L^{1}}{\partial x}\left(r x+\delta_{1} y\right. & \left.-\frac{1}{\gamma}+\frac{1}{\gamma} \ln \frac{\partial L^{1}}{\partial x}\right)+\frac{\partial L^{1}}{\partial y} \alpha+\frac{1}{2} \frac{\partial^{1} L^{1}}{\partial y^{2}} \varsigma^{2} \\
& -\frac{1}{2} \frac{\left(\lambda \frac{\partial L^{1}}{\partial x}+\varsigma \rho \frac{\partial^{2} L^{1}}{\partial x \partial y}\right)^{1}}{\frac{\partial^{2} L^{1}}{\partial x^{2}}}=0, \quad y<\hat{y} \tag{6.41}
\end{align*}
$$

and we look for a solution

$$
\begin{equation*}
L^{1}(x, y)=-\frac{1}{r \gamma} e^{-r \gamma(x+g(y))+1-\frac{\mu+\frac{\lambda^{2}}{2}}{r}} \tag{6.42}
\end{equation*}
$$

with $g$ the solution of

$$
\begin{cases}-\frac{1}{2} \varsigma^{2} g^{\prime \prime}-(\alpha-\varsigma \rho) g^{\prime}+\frac{1}{2} r^{2} \gamma^{2} \varsigma^{2}\left(1-\rho^{2}\right) g^{\prime 2}+r g=\delta_{1} y, & y<\hat{y}  \tag{6.43}\\ g(\hat{y})=f(\hat{y}) ; & g \text { has linear growth at }-\infty\end{cases}
$$

Considering the difference $m=g-f$, we rewrite (6.43) as

$$
\left\{\begin{array}{l}
-\frac{1}{2} \varsigma^{2} m^{\prime \prime}-m^{\prime}\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta_{2} \gamma\left(1-\rho^{2}\right)\right)+\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) m^{2}+r m=\left(\delta_{1}-\delta_{2}\right) y  \tag{6.44}\\
m(\hat{y})=0 ; \quad m \text { has linear growth at } y \rightarrow-\infty
\end{array}\right.
$$

We look for a solution of (6.44) in the interval

$$
\begin{equation*}
\frac{\delta_{1}-\delta_{2}}{r}\left(y-y^{*}\right) \leq m(y) \leq \frac{\delta_{1}-\delta_{2}}{r}\left[y-y_{0}+\left(y_{0}-\hat{y}\right) e^{\beta(y-\hat{y})}\right] \quad \text { for } y<\hat{y} \tag{6.45}
\end{equation*}
$$

where $\beta>0$ is the solution of

$$
-\frac{1}{2} \varsigma^{2} \beta^{2}-\beta\left(\alpha-\lambda \varsigma \rho-\gamma \delta \varsigma^{2}\left(1-\rho^{2}\right)\right)+r=0
$$

and $y_{0}=-\frac{\alpha-\lambda \varsigma \rho-\delta_{2} \gamma \varsigma^{2}\left(1-\rho^{2}\right)}{r}$ with $f\left(y_{0}+\frac{\delta_{2}^{2} \gamma \varsigma^{2}\left(1-\rho^{2}\right)}{r^{2}}\right)=0$, and we take

$$
\begin{equation*}
y^{*}=\max \left(\hat{y}, y_{0}+\frac{\gamma \varsigma^{2}\left(1-\rho^{2}\right)\left(\delta_{1}-\delta_{2}\right)}{2 r}\right) \tag{6.46}
\end{equation*}
$$

Similar to Theorem 6.2, there exists one and only one solution of (6.44) in the interval (6.45), which is $C^{1}$ and piecewise $C^{2}$. We thus have proven there exists a unique $g(y) \in C^{1}$ and piecewise $C^{2}$. As a result, we prove the existence and uniqueness solution of the Bellman equation (6.39). It remains to prove the form of solution defined by (6.42) is indeed the value function defined by (6.38).

Theorem 6.4. The function $L^{1}(x, y)$ defined by (6.42) coincides with the value function given in (6.38).

Proof. See Appendix M.
6.3.3. Leader's preinvestment utility maximization. After obtaining the solution to the leader's postinvestment utility maximization in section 6.3 .2 , we can solve the full problem by making use of the value function $L^{1}(x, y)$. We turn to the leader's optimal stopping problem (i.e., choice of $\theta$ ). Before the stopping time $\theta$, wealth and cash flow evolve as (6.4) and (6.1).

At time $\theta$, the leader stops and receives $L^{1}\left(X^{0}(\theta)-K, Y(\theta)\right)$; the objective of the leader is to maximize

$$
\begin{equation*}
J_{x, y}(C(\cdot), \pi(\cdot), \theta)=E\left[\int_{0}^{\theta} U(C(t)) e^{-\mu t} d t+L^{1}\left(X^{0}(\theta)-K, Y(\theta)\right) e^{-\mu \theta}\right] \tag{6.47}
\end{equation*}
$$

and we define the value function

$$
\begin{equation*}
L(x, y)=\sup _{\{\pi(\cdot), C(\cdot), \theta\} \in \mathcal{U}_{x, y}^{0}} J_{x, y}(C(\cdot), \pi(\cdot), \theta), \tag{6.48}
\end{equation*}
$$

where $\mathcal{U}_{x, y}^{0}=\left\{(\pi, C, \theta): I^{0} ; \theta<\infty\right.$ a.s. $; \tau^{*}=\lim \uparrow \tau_{N}^{0} \geq \theta$ a.s. $\}$. As a consequence of dynamic programming, we write the V.I. in the strong sense that the value function $L(x, y)$ must satisfy

We look for a solution

$$
\begin{equation*}
L(x, y)=-\frac{1}{r \gamma} e^{-r \gamma(x+h(y))+1-\frac{\mu+\frac{\lambda^{2}}{2}}{r}} \tag{6.50}
\end{equation*}
$$

and obtain that $h(y)$ must satisfy the V.I.

$$
\left\{\begin{array}{l}
\frac{1}{2} \varsigma^{2} h^{\prime \prime}+(\alpha-\lambda \varsigma \rho) h^{\prime}-\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) h^{\prime 2}-r h \leq 0,  \tag{6.51}\\
h(y) \geq g(y)-K, \\
(h(y)-g(y)+K)\left[\frac{1}{2} \varsigma^{2} h^{\prime \prime}+(\alpha-\lambda \varsigma \rho) h^{\prime}-\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) h^{\prime 2}-r h\right]=0
\end{array}\right.
$$

We meet with the classical difficulty that the obstacle $g(y)-K$ is $C^{0}$ but not $C^{1}$, so that the V.I. (6.51) must be interpreted in a weak sense. We cannot as in (6.25) consider $u(y)=$ $h(y)-g(y)+K$ since $g(y)$ is not sufficiently smooth. We will nonetheless consider the function

$$
u(y)=h(y)-f(y)+K
$$

which satisfies

$$
\left\{\begin{array}{l}
-\frac{1}{2} \varsigma^{2} u^{\prime \prime}-\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta_{2} \gamma\left(1-\rho^{2}\right)\right) u^{\prime}+\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) u^{\prime 2}+r u \geq-\delta_{2} y+r K,  \tag{6.52}\\
u \geq m, \\
(u-m)\left[-\frac{1}{2} \varsigma^{2} u^{\prime \prime}-\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta_{2} \gamma\left(1-\rho^{2}\right)\right) u^{\prime}+\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) u^{\prime 2}\right. \\
\left.\quad+r u+\delta_{2} y-r K\right]=0 .
\end{array}\right.
$$

The function $m(y)=g(y)-f(y)$ is defined by (6.44), which can be interpreted as, after the leader has entered, the difference between the leader's project value in anticipation of the follower's (optimal) entry and the leader's project value after the follower's entry. We see at once that if $m \leq 0$, then $u$ coincides with the solution of (6.26). The leader and the follower have the same strategy since the leader has no advantage in anticipation. So we will consider the case $m$ is not always negative.

In that case, $m$ is positive near $\hat{y}$. Indeed, otherwise, we consider the first point $y^{*}<\hat{y}$ such that $m\left(y^{*}\right)=0$. Necessarily, $y^{*}<0$. Otherwise, $m$ would have a negative local minimum in the interval $(0, \hat{y})$. This is impossible from the maximum principle since the right-hand side of (6.44) is positive if $y>0$. So we may assume

$$
\begin{equation*}
m^{\prime}(\hat{y}-0)<0 . \tag{6.54}
\end{equation*}
$$

Since $u(y)=-f(y)+K$ satisfies the first inequality in (6.52) as an equality, we have

$$
\begin{equation*}
u(y) \geq-f(y)+K \tag{6.55}
\end{equation*}
$$

which is the same constraint as for the follower; thus $h(y) \geq 0$.
We look for a solution of (6.52) in an interval $0 \leq u(y) \leq \bar{u}(y)$, where $\bar{u}(y)$ will be a ceiling function which is $C^{1}$ and vanishes for $y$ sufficiently large. We consider the ceiling function

$$
\begin{equation*}
\bar{u}(y)=\frac{\delta_{2}}{r}\left[-(y-k \hat{y})+\frac{e^{\beta(y-k \hat{y})}-1}{\beta}\right], \tag{6.56}
\end{equation*}
$$

where $k$ is sufficiently large. This function is such that $\bar{u}(k \hat{y})=\bar{u}^{\prime}(k \hat{y})=0$. We extend it by 0 for $y>k \hat{y}$. Also, $\bar{u}(y)>0$ for $y<k \hat{y}$.

Theorem 6.5. Assume (6.53). There exists a unique $u \in C^{1}(-\infty, \infty)$, piecewise $C^{2}$ solution of (6.52). This function vanishes for $y$ sufficiently large. It is the value function

$$
\begin{equation*}
u(y)=\inf _{v(\cdot)} \sup _{\theta} J_{y}(v(\cdot), \theta)=\sup _{\theta} \inf _{v(\cdot)} J_{y}(v(\cdot), \theta)=J_{y}(\hat{v}(\cdot), \hat{\theta}), \tag{6.57}
\end{equation*}
$$

where

$$
\left\{\begin{array}{l}
J_{y}(v(\cdot), \theta)=E\left[\int_{0}^{\theta}\left(-\delta_{2} Y_{y}(t)+r K+\frac{1}{2} v^{2}(t)\right) e^{-r t} d t+m\left(Y_{y}(\theta)\right) e^{-r \theta} \mathbb{1}_{\theta<\infty}\right] \\
v(\cdot) \in \mathcal{U}_{y}=\left\{\lim _{0} \sup _{T \rightarrow \infty}\left(-E Y_{y}(T) e^{-r T}\right)=0\right\} .
\end{array}\right.
$$

Moreover, $u(y)+\frac{\delta_{2}}{r} y$ is bounded for $y \rightarrow-\infty$, and $u \geq 0$.
Proof. See Appendix N.
We now want to show that the solution to (6.52) is characterized by two intervals. The difficulty is that one cannot interpret $u(y)$ simply as the value function of a control problem. We approach the problem directly.

We must find three points $y_{1}, y_{2}, y_{3}$ such that

$$
\begin{equation*}
y_{1}<y_{2}<\hat{y}<y_{3} \tag{6.58}
\end{equation*}
$$

and

$$
\left\{\begin{array}{l}
-\frac{1}{2} \varsigma^{2} u^{\prime \prime}(y)-\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta_{2} \gamma\left(1-\rho^{2}\right)\right) u^{\prime}(y)+\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) u^{2}(y)+r u(y)  \tag{6.59}\\
\quad=-\delta_{2} y+r K \quad \text { for } y<y_{1} \text { and } y_{2}<y<y_{3} \\
u(y)=m(y) \quad \text { for } y_{1} \leq y \leq y_{2}, \\
u(y)=0 \quad \text { for } y \geq y_{3}, \\
u^{\prime}\left(y_{1}\right)=m^{\prime}\left(y_{1}\right) ; \quad u^{\prime}\left(y_{2}\right)=m^{\prime}\left(y_{2}\right) ; \quad u^{\prime}\left(y_{3}\right)=0
\end{array}\right.
$$

We can take $y_{3}=\bar{y}=k \hat{y}$ with $k$ sufficiently large. We show the existence of a unique two-interval solution in the following theorem.

Theorem 6.6. Assume (6.53). Then the solution $u$ of (6.52) is of the form (6.59). There exists a unique triple $y_{1}, y_{2}, y_{3}$ with $y_{1}<y_{2}<\hat{y}<y_{3}$ such that (6.59) holds.

Proof. See Appendix O.
By Theorem 6.6, for the process $Y(t)$ given in (6.1), we can define the optimal stopping rule as

$$
\hat{\theta}(y)=\left\{\begin{array}{l}
\inf \left\{t \mid Y_{y}(t) \geq y_{1}\right\} \quad \text { if } y<y_{1}  \tag{6.60}\\
0 \quad \text { if } y_{1} \leq y \leq y_{2} \\
\inf \left\{t \mid Y_{y}(t) \leq y_{2} \text { or } Y_{y}(t) \geq y_{3}\right\} \quad \text { if } y_{2}<y<y_{3} \\
0 \quad \text { if } y \geq y_{3}
\end{array}\right.
$$

It remains to show that the function $L(x, y)$ defined by (6.50) is the value function (6.48).
Theorem 6.7. Assume (6.53). Then the function $L(x, y)$ defined by (6.50) is the value function (6.48).

Proof. The feedbacks

$$
\left\{\begin{array}{l}
\widehat{C}(x, y)=r(x+h(y))-\frac{1}{\gamma}\left(1-\frac{\mu+\frac{\lambda^{2}}{2}}{r}\right)  \tag{6.61}\\
\widehat{\pi}(x, y) \sigma x=\frac{\lambda}{r \gamma}-\varsigma \rho h^{\prime}(y)
\end{array}\right.
$$

and the stopping time (6.60) provide the optimal control. The proof is similar to that of Theorem 6.3.
6.3.4. Optimal rules in the case of incomplete markets. We prove that the single player and the follower have a unique optimal stopping strategy characterized by a threshold. The leader's optimal stopping strategy is characterized by a two-interval solution. The optimal stopping rules for the leader and the follower can be characterized as those described in the complete market section. In the single decision maker model, we are able to recover complete market results by allowing $|\rho|$ (the correlation between the traded and the nontraded assets) $\rightarrow 1$. And when $\gamma$ (the risk aversion coefficient) $\rightarrow 0$, the problem becomes one of optimal stopping under a minimal martingale measure. The option value to invest is decreasing in $\gamma$. Market completeness and risk aversion are essential factors for evaluating optimal investment policies. For managers, we caution that naively assuming market completeness may lead to nonoptimal investment decisions.
7. Conclusion. We study optimal investment policies in an irreversible capital investment under uncertainty in a monopoly situation and in a Stackelberg leader-follower game. We consider two types of investment payoffs: lump-sum and cash flows. The decisions are the times to enter into the market. We employ V.I.s for solving these optimal stopping time problems.

In the case of complete markets, we work with the risk-neutral probability measure for the valuation. In the case of incomplete markets, we adopt the utility-based valuation where the investor solves the utility maximization with joint decisions of stopping times, portfolio investment, and/or consumption rules.

For the case of lump-sum payoffs, we proved that the single player and consequently the follower have a unique optimal stopping strategy characterized by a threshold, both for the complete and incomplete capital markets. In the complete market, the leader's optimal investment rule is characterized by a two-interval strategy. In the incomplete market, we need the weak formulation of the leader's problem and can characterize the optimal stopping only as the first time the solution touches the obstacle. For the case of flow payoffs, we prove that the single player and the follower have a unique optimal stopping strategy characterized by a threshold, and that the leader's optimal stopping strategy can be characterized by the two-interval solution in both the complete and the incomplete markets.

In both payoff situations, we find that differences in the investment stopping time depend on the degree of completeness in the market and conclude that market completeness and risk aversion are important inputs into the optimal investment policy. In addition, optimal investment policies for both the leader and the follower deviate from those of the single decision maker. Strategic considerations are shown to be important in the formulation of capital investment project decisions. Use of classical real option rules for a single decision maker will lead to nonoptimal investment.

Appendix A. Proof of Lemma 3.2. We take $\tilde{U}=K-(U-K)^{-}$as a test function in (3.29). Indeed, $\tilde{U} \in \mathcal{K}$ so that

$$
-b\left(U,(U-K)^{+}\right)-\alpha\left(U,(U-K)^{+}\right)_{\varrho} \geq-\left(f,(U-K)^{+}\right)_{\varrho}-\alpha\left(G,(U-K)^{+}\right)_{\varrho} .
$$

Hence,

$$
\begin{aligned}
-b\left((U-K)^{+},(U-K)^{+}\right)-\alpha\left|(U-K)^{+}\right|_{\varrho}^{2} \geq- & \left(f-r K,(U-K)^{+}\right)_{\varrho} \\
& -\alpha\left(G-K,(U-K)^{+}\right)_{\varrho}
\end{aligned}
$$

$$
\geq 0,
$$

which implies $(U-K)^{+}=0$. That is, $\Gamma_{\alpha}(G) \leq K$.
The property $\Gamma_{\alpha}(G) \geq 0$ is obvious since $\Gamma_{\alpha}(G) \in \mathcal{K}$; hence $\Gamma_{\alpha}(G) \geq \chi \geq 0$.
Appendix B. Proof of Lemma 3.3. To simplify notation, let $U_{1}=\Gamma_{\alpha}\left(G_{1}\right)$, and let $U_{2}=\Gamma_{\alpha}\left(G_{2}\right)$. By definition,

$$
\left\{\begin{array}{l}
b\left(U_{1}, \tilde{U}-U_{1}\right)+\alpha\left(U_{1}, \tilde{U}-U_{1}\right)_{\varrho} \geq\left(f, \tilde{U}-U_{1}\right)_{\varrho}+\alpha\left(G_{1}, \tilde{U}-U_{1}\right)_{\varrho}  \tag{B.1}\\
b\left(U_{2}, \tilde{U}-U_{2}\right)+\alpha\left(U_{2}, \tilde{U}-U_{2}\right)_{\varrho} \geq\left(f, \tilde{U}-U_{2}\right)_{\varrho}+\alpha\left(G_{2}, \tilde{U}-U_{2}\right)_{\varrho}
\end{array}\right.
$$

Set $M=\frac{\alpha\left\|G_{1}-G_{2}\right\|_{L^{\infty}}}{\alpha+r}$.
We take $\tilde{U}=U_{1}-\left(U_{1}-U_{2}-M\right)^{+}$in the first relation (B.1) and $\tilde{U}=U_{2}+\left(U_{1}-U_{2}-M\right)^{+}$ in the second relation (B.1). Adding, we deduce

$$
\begin{array}{r}
-b\left(U_{1}-U_{2},\left(U_{1}-U_{2}-M\right)^{+}\right)-\alpha\left(U_{1}-U_{2},\left(U_{1}-U_{2}-M\right)^{+}\right)_{\varrho} \\
\geq-\alpha\left(G_{1}-G_{2},\left(U_{1}-U_{2}-M\right)^{+}\right)_{\varrho}
\end{array}
$$

hence,

$$
\begin{aligned}
-b\left(U_{1}-U_{2},\left(U_{1}-U_{2}-M\right)^{+}\right) & -\alpha\left(U_{1}-U_{2},\left(U_{1}-U_{2}-M\right)^{+}\right)_{\varrho} \\
& -(r+\alpha)\left(M,\left(U_{1}-U_{2}-M\right)^{+}\right) \\
\geq & -\alpha\left(G_{1}-G_{2},\left(U_{1}-U_{2}-M\right)^{+}\right)_{\varrho} .
\end{aligned}
$$

Since $(r+\alpha) M \geq \alpha\left(G_{1}, G_{2}\right)$, we have

$$
b\left(\left(U_{1}-U_{2}-M\right)^{+},\left(U_{1}-U_{2}-M\right)^{+}\right)+\alpha\left|\left(U_{1}-U_{2}-M\right)^{+}\right|^{2} \leq 0 .
$$

Therefore, $U_{1}-U_{2} \leq M$. Similarly, $U_{2}-U_{1} \leq M$, and we conclude that (3.32) holds.
Appendix C. Proof of Theorem 3.5. We can approximate $\chi(v)$ by a sequence of smooth functions $\chi^{\epsilon}(v)$ such that

$$
\sup _{v}\left|\chi^{\epsilon}(v)-\chi(v)\right| \rightarrow 0 \quad \text { as } \epsilon \rightarrow 0 .
$$

In the present context, such an approximation is easily obtained because $\chi(v)$ is smooth except in $\hat{v}$. As a result, the solution $U$ is more regular than $H_{\varrho}^{1}(0, \infty)$. The solution is locally $H^{2}$, and the V.I. can be written as in (3.5), calling $U^{\epsilon}$ the solution of the V.I. (3.28) with $\chi^{\epsilon}$ instead of $\chi$; we have

$$
\left\{\begin{array}{l}
v(r+\eta(\xi-\lambda)) U^{\epsilon \prime}(v)+\frac{1}{2} U^{\epsilon \prime \prime}(v) \eta^{2} v^{2}-r U^{\epsilon}(v)+f(v) \leq 0, \\
U^{\epsilon}(v) \geq \chi^{\epsilon}(v), \\
{\left[U^{\epsilon}(v)-\chi^{\epsilon}(v)\right]\left[v(r+\eta(\xi-\lambda)) U^{\epsilon \prime}(v)+\frac{1}{2} U^{\epsilon \prime \prime}(v) \eta^{2} v^{2}-r U^{\epsilon}(v)+f(v)\right]=0,} \\
U^{\epsilon}(0)=K .
\end{array}\right.
$$

It is standard to check that

$$
U^{\epsilon}(v)=\sup _{\theta \geq 0} \widehat{E}\left[e^{-r \theta} \chi^{\epsilon}\left(V_{v}(\theta)\right) \mathbb{1}_{\theta<\infty}+\int_{0}^{\theta} e^{-r s} f\left(V_{v}(s)\right) d s\right]
$$

and, clearly,

$$
\sup _{v}\left|U^{\epsilon}(v)-U(v)\right| \leq \sup _{v}\left|\chi^{\epsilon}(v)-\chi(v)\right| .
$$

On the other hand, $U^{\epsilon}$ remains in a bounded subset of $H_{\varrho}^{1}(0, \infty)$, and we can extract a subsequence converging to $U(v)$, a solution of (3.28), in $H_{\varrho}^{1}(0, \infty)$ weakly. From the uniqueness of the solution of (3.28), the sequence converges, and thus the solution of (3.28) coincides with (3.19).

Appendix D. Proof of Theorem 3.6. Since $U(0)=K>\chi(0)=0$ and $U(\bar{v})=\chi(\bar{v})=0$, there exists a first point $v_{1} \leq \bar{v}$ such that $U\left(v_{1}\right)=\chi\left(v_{1}\right)$. We must have $v_{1}<\hat{v}$. Otherwise, $v_{1}=\bar{v}$. But, from (3.5), we have

$$
\left\{\begin{array}{l}
F^{\prime}(v) v(r+\eta(\xi-\lambda))+\frac{1}{2} F^{\prime \prime}(v) v^{2} \eta^{2}-r F(v) \leq 0, \\
F(v) \geq(1-a) v-K, \\
{[F(v)-(1-a) v+K]\left[F^{\prime}(v) v(r+\eta(\xi-\lambda))+\frac{1}{2} F^{\prime \prime}(v) v^{2} \eta^{2}-r F(v)\right]=0,} \\
F(0)=0
\end{array}\right.
$$

which, by setting $U(v)=F(v)-(1-a) v+K$, may be re-expressed as
(D.1)

$$
\left\{\begin{array}{l}
-\frac{1}{2} U^{\prime \prime}(v) v^{2} \eta^{2}-U^{\prime}(v) v(r+\eta(\xi-\lambda))+r U(v) \geq(1-a) v \eta(\xi-\lambda)+r K \\
U(v) \geq 0 \\
U(v)\left[-\frac{1}{2} U^{\prime \prime}(v) v^{2} \eta^{2}-U^{\prime}(v) v(r+\eta(\xi-\lambda))+r U(v)-(1-a) v \eta(\xi-\lambda)-r K\right]=0 \\
U(0)=K
\end{array}\right.
$$

We then observe that $U(v)$ coincides with the solution of (D.1) (with ( $1-a$ ) replaced by $a$ ), i.e., the same system of (3.34) with $\chi(v)=0$. But then $\bar{v}=\hat{v}$; hence $v_{1}=\hat{v}$. In this case $\widetilde{U}(v)=U(v)-\chi(v)$ satisfies

$$
\left\{\begin{array}{l}
-\frac{1}{2} \widetilde{U}^{\prime \prime}(v) v^{2} \eta^{2}-\widetilde{U}^{\prime}(v) v(r+\eta(\xi-\lambda))+r \widetilde{U}(v)=v \eta(\xi-\lambda)+r K  \tag{D.2}\\
\widetilde{U}(0)=K ; \quad \widetilde{U}(\hat{v})=0
\end{array}\right.
$$

Since $U^{\prime}(\hat{v})=0$, it follows that $\widetilde{U}^{\prime}(\hat{v}-0)=-\chi^{\prime}(\hat{v}-0)$. This in turn implies $\widetilde{U}^{\prime}(\hat{v}-0)>0$ because $\chi^{\prime}(\hat{v}-0)<0$. It then follows that $\widetilde{U}(v)<0$ for $v$ close to $\hat{v}$, which is impossible. Therefore $v_{1} \leq \hat{v}$. Indeed, setting $\widetilde{U}(v)=U(v)-\chi(v)$, it satisfies

$$
-\frac{1}{2} \widetilde{U}^{\prime \prime}(v) v^{2} \eta^{2}-\widetilde{U}^{\prime}(v) v(r+\eta(\xi-\lambda))+r \widetilde{U}(v)=v \eta(\xi-\lambda)+r K
$$

with $\widetilde{U}(0)=K, \widetilde{U}\left(v_{1}\right)=0, \widetilde{U}^{\prime}\left(v_{1}\right)=0$. The matching of the derivative comes from the fact that $\widetilde{U}(v)$ is $C^{1}$ and $\widetilde{U}(v) \geq 0, \widetilde{U}\left(v_{1}\right)=0$. So $v_{1}$ is the local minimum; hence $\widetilde{U}^{\prime}\left(v_{1}\right)=0$, and we have $\widetilde{U}^{\prime}\left(v_{1}\right)=\chi^{\prime}\left(v_{1}\right)$.

Since $U(\hat{v})>\chi(\hat{v})=0$, there exists an interval in which $\hat{v}$ is contained such that the equation holds in this interval, denoted as $\left(v_{2}, v_{3}\right)$ with $v_{3}=\bar{v}$ and $v_{2}<\hat{v}$. Necessarily, we have $v_{2} \geq v_{1}$; otherwise, $U(v)$ will be the solution of the equation on $\left(0, v_{3}\right)$, which is impossible. So we have $U\left(v_{2}\right)=\chi\left(v_{2}\right)$ and necessarily $U^{\prime}\left(v_{2}\right)=\chi^{\prime}\left(v_{2}\right)$.

On the other hand, on the interval $\left(v_{1}, v_{2}\right), \chi(v)$ satisfies all conditions (3.35) and the right-hand side $-(1-a) \eta(\xi-\lambda) v>a \eta(\xi-\lambda) v+r K$; therefore, $\chi(v)$ satisfies all conditions (3.34) on $\left(v_{1}, v_{2}\right)$. Therefore $U(v)=\chi(v)$ on $\left(v_{1}, v_{2}\right)$.

By the uniqueness of $U(v)$, the triple $v_{1}, v_{2}, v_{3}$ is necessarily unique.

Appendix E. Proof of Lemma 4.2. From the relation $-\gamma\left(1-\rho^{2}\right) H^{e}(v)=\ln \left(1+B v^{\beta}\right)$, we obtain for $v<\hat{v}$

$$
\begin{aligned}
\left(H^{e}(v)\right)^{\prime} & =\frac{-\beta B v^{\beta-1}}{\gamma\left(1-\rho^{2}\right)\left(1+B v^{\beta}\right)} \\
\left(H^{e}(v)\right)^{\prime \prime} & =\frac{-\beta B}{\gamma\left(1-\rho^{2}\right)} \frac{(\beta-1) B v^{\beta-2}-B v^{2 \beta-2}}{\left(1+B v^{\beta}\right)^{2}}>0
\end{aligned}
$$

So $\left(H^{e}(v)\right)^{\prime}$ increases from 0 to

$$
\left(H^{e}\right)^{\prime}(\hat{v}-0)=(1-a) \frac{e^{\varpi+K \gamma\left(1-\rho^{2}\right)}-1}{e^{\varpi}-1}>(1-a)=\left(H^{e}\right)^{\prime}(\hat{v}+0) .
$$

It follows that $v-K-H^{e}(v)$ may either be increasing from 0 to $\hat{v}$ or may pass by a positive maximum. At any rate, there is one and only one 0 .

Appendix F. Proof of Theorem 4.3. We first prove that (4.42) has a set of solutions with a maximum and a minimum element. For that we consider a monotone increasing sequence and monotone decreasing sequence $\Sigma_{n}$ and $\Sigma^{n}$ defined as follows:

$$
\begin{align*}
& b\left(\Sigma^{n+1}, \widetilde{\Sigma}-\Sigma^{n+1}\right)+\alpha\left(\Sigma^{n+1}, \widetilde{\Sigma}-\Sigma^{n+1}\right)_{\varrho} \geq \alpha\left(\Sigma^{n}, \widetilde{\Sigma}-\Sigma^{n+1}\right)_{\varrho},  \tag{F.1}\\
& b\left(\Sigma_{n+1}, \widetilde{\Sigma}-\Sigma_{n+1}\right)+\alpha\left(\Sigma_{n+1}, \widetilde{\Sigma}-\Sigma_{n+1}\right)_{\varrho} \geq \alpha\left(\Sigma_{n}, \widetilde{\Sigma}-\Sigma_{n+1}\right)_{\varrho} . \tag{F.2}
\end{align*}
$$

We start with

$$
\begin{equation*}
\Sigma^{0}=\psi ; \quad \Sigma_{0}=0 \tag{F.3}
\end{equation*}
$$

In (F.1) and (F.2) $\alpha$ is a sufficiently large number, which ensures that the bilinear form $b(\Sigma, \widetilde{\Sigma})+\alpha(\Sigma, \widetilde{\Sigma})_{\varrho}$ is coercive. It follows that the sequences $\Sigma^{n}$ and $\Sigma_{n}$ are well defined. Clearly $\Sigma^{1} \leq \Sigma^{0}$. Let us check that if $\Sigma^{n} \leq \Sigma^{n-1}$, then $\Sigma^{n+1} \leq \Sigma^{n}$. Consider (F.1) with $n-1$ replacing $n$; hence

$$
\begin{equation*}
b\left(\Sigma^{n}, \widetilde{\Sigma}-\Sigma^{n}\right)+\alpha\left(\Sigma^{n}, \widetilde{\Sigma}-\Sigma^{n}\right)_{\varrho} \geq \alpha\left(\Sigma^{n-1}, \widetilde{\Sigma}-\Sigma^{n}\right)_{\varrho} \tag{F.4}
\end{equation*}
$$

We take $\widetilde{\Sigma}=\Sigma^{n+1}-\left(\Sigma^{n+1}-\Sigma^{n}\right)^{+}=\min \left(\Sigma^{n}, \Sigma^{n+1}\right)$ in (F.1) and $\widetilde{\Sigma}=\Sigma^{n}+\left(\Sigma^{n+1}-\Sigma^{n}\right)^{+}=$ $\max \left(\Sigma^{n}, \Sigma^{n+1}\right)$ in (F.4). We get

$$
\begin{gathered}
-b\left(\Sigma^{n+1},\left(\Sigma^{n+1}-\Sigma^{n}\right)^{+}\right)-\alpha\left(\Sigma^{n+1},\left(\Sigma^{n+1}-\Sigma^{n}\right)^{+}\right)_{\varrho} \geq-\alpha\left(\Sigma^{n},\left(\Sigma^{n+1}-\Sigma^{n}\right)^{+}\right)_{\varrho}, \\
b\left(\Sigma^{n},\left(\Sigma^{n+1}-\Sigma^{n}\right)^{+}\right)+\alpha\left(\Sigma^{n},\left(\Sigma^{n+1}-\Sigma^{n}\right)^{+}\right)_{\varrho} \geq \alpha\left(\Sigma^{n-1},\left(\Sigma^{n+1}-\Sigma^{n}\right)^{+}\right)_{\varrho} .
\end{gathered}
$$

Adding, we easily deduce that $\left(\Sigma^{n+1}-\Sigma^{n}\right)^{+}=0$. As $n \rightarrow \infty$, we obtain that $\Sigma^{n} \downarrow \bar{\Sigma}$, the maximum solution of (4.42). Indeed, if $\Sigma$ is a solution such that $\Sigma \leq \Sigma^{0}$, then $\Sigma \leq \Sigma^{1}$. It follows that $\Sigma \leq \Sigma^{n}$; hence $\Sigma \leq \bar{\Sigma}$. Similarly $\Sigma_{n} \uparrow \underline{\Sigma}$, the minimum solution of (4.42).

We next prove that the maximum and the minimum solutions coincide. We first verify that the optimal stopping time $\hat{\theta}$ is finite a.s. and proceed with convergence of solutions from the
smooth function construction. Suppose now $\psi \in C^{1}$. Then we rely on the classical result on V.I.s that the minimum and the maximum solutions coincide with the value function defined in (4.38) and (4.39). The optimal stopping time $\hat{\theta}(v)$ is defined by

$$
\begin{equation*}
\hat{\theta}(v)=\inf \left\{t \mid \Sigma\left(V_{v}(t)\right)=\psi\left(V_{v}(t)\right)\right\} . \tag{F.5}
\end{equation*}
$$

We have $\hat{\theta}(v)<\infty$ a.s. Indeed, we first note that $\Sigma$ is also $C^{1}$ and there exists a number $M>\hat{v}$ such that

$$
\begin{equation*}
\Sigma(v)=\psi(v), \quad v \geq M \tag{F.6}
\end{equation*}
$$

Otherwise, $\Sigma(v)<\psi(v) \forall v$ and the differential equation $\Sigma^{\prime}(\xi-\lambda \rho)+\frac{1}{2} v \eta \Sigma^{\prime \prime}=0$ holds for any $v$. Therefore, we have $\Sigma(v)=1+D V^{\beta}$ with $D$ a negative constant. This is not possible since $\Sigma(v) \geq 0 \forall v$.

If

$$
\begin{aligned}
& \Sigma(M)=e^{-\gamma\left(1-\rho^{2}\right)(a M-K)} \\
& \Sigma^{\prime}(M)=-a \gamma\left(1-\rho^{2}\right) e^{-\gamma\left(1-\rho^{2}\right)(a M-K)}
\end{aligned}
$$

then for $v \geq M, \Sigma(v)=e^{-\gamma\left(1-\rho^{2}\right)(a v-K)}$ satisfies conditions (4.35) since $\Sigma^{\prime}(\xi-\lambda \rho)+\frac{1}{2} v \eta \Sigma^{\prime \prime} \geq 0$ thanks to assumption $\xi-\lambda \rho<0$. Therefore, we have (F.6). Take $v<M$ and define

$$
\theta_{M}(v)=\inf \left\{t \mid V_{v}(t)=0 \text { or } V_{v}(t) \geq M\right\} ;
$$

then $\theta_{M}(v)<\infty$ a.s. But $\hat{\theta}(v)<\theta_{M}(v)$ since $\Sigma\left(V_{v}\left(\theta_{M}(v)\right)\right)=\psi\left(V_{v}\left(\theta_{M}(v)\right)\right)$, and we have $\hat{\theta}<\infty$ a.s.

We next construct smooth functions $\psi_{\epsilon}(v)$ which approximate $\psi(v)$ from below or from above. We consider a sequence of $C^{1}$ approximations $\psi_{\epsilon}(v) \geq 0$ of the obstacle $\psi(v)$. Define a convex subset

$$
\mathcal{K}_{\epsilon}=\left\{\phi \mid 0 \leq \phi \leq \psi_{\epsilon}\right\}
$$

and let $\Sigma_{\epsilon}$ be the unique solution of the V.I.

$$
\begin{equation*}
b\left(\Sigma_{\epsilon}, \widetilde{\Sigma}-\Sigma_{\epsilon}\right) \geq 0 \quad \forall \widetilde{\Sigma} \in \mathcal{K}_{\epsilon}, \Sigma_{\epsilon} \in \mathcal{K}_{\epsilon}, \tag{F.7}
\end{equation*}
$$

which, thanks to the regularity of $\psi_{\epsilon}$, can be written as

$$
\left\{\begin{array}{l}
\Sigma_{\epsilon}^{\prime}(\xi-\lambda \rho)+\frac{1}{2} v \eta \Sigma_{\epsilon}^{\prime \prime} \geq 0  \tag{F.8}\\
\Sigma_{\epsilon}(v) \leq \psi_{\epsilon}(v) \\
{\left[\Sigma_{\epsilon}^{\prime}(\xi-\lambda \rho)+\frac{1}{2} v \eta \Sigma_{\epsilon}^{\prime \prime}\right]\left[\Sigma_{\epsilon}(v)-\psi_{\epsilon}(v)\right]=0,} \\
0 \leq \Sigma_{\epsilon}(v) \leq 1 \\
\Sigma_{\epsilon}(0)=1
\end{array}\right.
$$

and $\Sigma_{\epsilon}(v)$ is the value function:

$$
\begin{equation*}
\Sigma_{\epsilon}(v)=\inf _{\theta} E\left[\psi_{\epsilon}\left(V_{v}(\theta)\right)\right] . \tag{F.9}
\end{equation*}
$$

We begin an approximation such that

$$
\begin{equation*}
\psi_{\epsilon}(v) \uparrow \psi(v) \quad \text { as } \epsilon \downarrow 0 . \tag{F.10}
\end{equation*}
$$

We can define the increasing processes $\Sigma_{\epsilon, n}(v)$ corresponding to $\psi_{\epsilon}(v)$ and $\Sigma_{\epsilon, n}(v) \uparrow \Sigma_{\epsilon}(v)$ as $n \uparrow \infty$. However, if we compare $\Sigma_{\epsilon, n}(v)$ and $\Sigma_{n}(v)$, it is easy to check that

$$
\begin{equation*}
\Sigma_{\epsilon, n}(v) \leq \Sigma_{n}(v) \leq \underline{\Sigma}(v) ; \tag{F.11}
\end{equation*}
$$

therefore, we obtain

$$
\begin{equation*}
\Sigma_{\epsilon}(v) \leq \underline{\Sigma}(v) . \tag{F.12}
\end{equation*}
$$

But from formula (F.9), $\Sigma_{\epsilon}(v)$ converges toward the value function. We obtain that the value function is smaller than the minimum solution $\underline{\Sigma}(v)$ of (4.42).

Next consider an approximation

$$
\begin{equation*}
\psi_{\epsilon}(v) \downarrow \psi(v) \quad \text { as } \epsilon \downarrow 0 \tag{F.13}
\end{equation*}
$$

and the monotone decreasing process $\Sigma_{\epsilon}^{n}$. We have $\Sigma_{\epsilon}^{n} \downarrow \Sigma_{\epsilon}$; therefore, this time $\Sigma_{\epsilon}^{n} \geq \Sigma_{\epsilon} \geq \bar{\Sigma}$, and $\Sigma_{\epsilon} \geq \bar{\Sigma}$. But $\Sigma_{\epsilon}$ converges again towards the value function. Therefore the minimum and maximum solutions of (4.42) coincide and are equal to the value function. This completes the proof.

Appendix G. Proof of Proposition 4.4. To prove that solutions of (4.47) exist, we consider the function

$$
Z(v)=e^{\gamma\left(1-\rho^{2}\right)\left(v-K-H^{e}(v)\right)}+e^{\gamma\left(1-\rho^{2}\right) H^{e}(v)}-\left(2+\frac{\gamma v\left(1-\rho^{2}\right)}{\beta}\right)
$$

and note that $Z(0)=e^{-\gamma\left(1-\rho^{2}\right) K}-1<0$.
On the other hand,

$$
Z(\hat{v})=e^{\varpi+\frac{2 a-1}{1-a}\left(\varpi+\gamma K\left(1-\rho^{2}\right)\right)}+e^{\varpi+\gamma K\left(1-\rho^{2}\right)}-2-\frac{1}{1-a}\left(e^{\varpi}-1\right) .
$$

Setting $u=\frac{2 a-1}{1-a}, u \in(0, \infty)$, and considering the function

$$
h(u ; \varpi)=e^{\varpi+u\left(\varpi+\gamma K\left(1-\rho^{2}\right)\right)}+e^{\varpi+\gamma K\left(1-\rho^{2}\right)}-2-(u+2)\left(e^{\varpi}-1\right),
$$

one can easily check that $h(u ; \varpi)>0$. Hence, $Z(\hat{v})>0$. Therefore, there are values $v_{1}$ such that $Z\left(v_{1}\right)=0$. It remains to verify that $v_{1}>v_{0}$. By (4.46), we can rewrite (4.47) as

$$
e^{\gamma\left(1-\rho^{2}\right)\left(v_{1}-K-H^{e}\left(v_{1}\right)\right)}=1+\frac{\gamma v_{1}\left(1-\rho^{2}\right)}{\beta}\left(1-\left(H^{e}\left(v_{1}\right)\right)^{\prime}\right) .
$$

But from the assumption and Lemma 4.2, we have $1-\left(H^{e}\left(v_{1}\right)\right)^{\prime}>0$; hence $v_{1}-K-H^{e}\left(v_{1}\right)>0$, which implies $v_{1}>v_{0}$. It remains to define $D$ by

$$
D v_{1}^{\beta}=e^{-\gamma\left(1-\rho^{2}\right)\left(v_{1}-K-H^{e}\left(v_{1}\right)\right)}-1
$$

to satisfy the first conditions (4.45). This completes the proof.

Appendix H. Proof of Theorem 5.1. The solution (5.9) is $C^{1}(-\infty, \infty)$ and piecewise $C^{2}$. It satisfies $F(y) \geq 0$ and the condition that $F(y)$ has linear growth at $\infty$. By construction, it satisfies the complimentary slackness condition (i.e., the product condition of (5.6)). To complete the proof, we must verify the inequalities

$$
\begin{equation*}
F(y) \geq \delta\left(\frac{y}{r}+\frac{\alpha-\lambda \varsigma}{r^{2}}\right)-K \quad \text { if } y \leq \hat{y} \tag{H.1}
\end{equation*}
$$

and

$$
\begin{equation*}
(\alpha-\lambda \varsigma) F^{\prime}(y)+\frac{1}{2} \varsigma^{2} F^{\prime \prime}(y)-r F(y) \leq 0 \quad \text { if } y \geq \hat{y} \tag{H.2}
\end{equation*}
$$

To prove (H.1), consider $G(y)=F(y)-\delta\left(\frac{y}{r}+\frac{\alpha-\lambda \varsigma}{r^{2}}\right)+K$. We have, if $y<\hat{y}$,

$$
G^{\prime}(y)=\frac{\delta}{r}\left(e^{-\beta(\hat{y}-y)}-1\right)<0 \quad \text { and } \quad G^{\prime \prime}(y)=\frac{\delta \beta}{r} e^{-\beta(\hat{y}-y)}>0 .
$$

Hence $G^{\prime}(y)$ increases up to zero on $(-\infty, \hat{y})$. Therefore, $G^{\prime}(y) \leq 0$ on $(-\infty, \hat{y})$ and $G(y)$ is decreasing on $(-\infty, \hat{y})$. Since $G(\hat{y})=0, G(y) \geq 0$ on $(-\infty, \hat{v})$, we have shown that (H.1) is true.

The inequality (H.2) is equivalent to $y \geq \frac{r K}{\delta}$ for $y \geq \hat{y}$, which implies $\frac{1}{\beta}>\frac{\alpha-\lambda \varsigma}{r}$. This is true because $r-\beta(\alpha-\lambda \varsigma)=\frac{1}{2} \varsigma^{2} \beta^{2}$.

Appendix I. Proof of Theorem 5.2. Before proceeding to prove the value function (5.15) is the unique solution of the V.I. (5.20), we introduce the approximation function, $\Psi_{M}(y)$, to overcome the problem of the unbounded obstacle, $\Psi(y)$, and define the corresponding approximated value function, $L_{M}(y)$. Note that

$$
\begin{equation*}
\Psi(y)=\frac{\delta_{2}}{r \beta}+\frac{\delta_{2}}{r}(y-\hat{y})^{+}-\frac{\delta_{1}}{r}(y-\hat{y})^{-}+\mu \frac{\delta_{1}-\delta_{2}}{r} \mathbb{1}_{y<\hat{y}}\left(1-e^{-\beta(\hat{y}-y)}\right) . \tag{I.1}
\end{equation*}
$$

We consider the following expression:

$$
\begin{equation*}
\Psi_{M}(y)=\frac{\delta_{2}}{r \beta}+\frac{\delta_{2}}{r} \frac{M(y-\hat{y})^{+}}{M+(y-\hat{y})^{+}}-\frac{\delta_{1}}{r}(y-\hat{y})^{-}+\mu \frac{\delta_{1}-\delta_{2}}{r} \mathbb{1}_{y<\hat{y}}\left(1-e^{-\beta(\hat{y}-y)}\right), \tag{I.2}
\end{equation*}
$$

with

$$
\begin{equation*}
\Psi_{M}(y) \uparrow \Psi(y) \quad \text { as } M \uparrow \infty . \tag{I.3}
\end{equation*}
$$

In correspondence with $\Psi_{M}(y)$, we set

$$
\begin{equation*}
L_{M}(y)=\sup _{\theta \geq 0} \widehat{E}\left[e^{-r \theta} \Psi_{M}\left(Y_{y}(\theta)\right) \mathbb{1}_{\theta<\infty}\right] \tag{I.4}
\end{equation*}
$$

It is easy to check that

$$
\begin{equation*}
L_{M}(y) \uparrow L(y) \quad \text { as } M \uparrow \infty . \tag{I.5}
\end{equation*}
$$

From (I.2), we have $\Psi_{M}(y) \leq \frac{\delta_{2}}{r \beta}+\frac{\delta_{2}}{r} M+\frac{\mu\left(\delta_{1}-\delta_{2}\right)}{r}=\mathcal{V}_{M}$, and therefore we have $0 \leq$ $L_{M}(y) \leq \mathcal{V}_{M}$.

We consider the convex subset of $H_{\varrho}^{1}(-\infty, \infty)$ :

$$
\begin{equation*}
\mathcal{K}_{M}=\left\{\Phi \in H_{\varrho}^{1}(-\infty, \infty) \mid \Phi(y) \geq \Psi_{M}(y) \forall y\right\} \tag{I.6}
\end{equation*}
$$

which is not empty since it contains $\Psi_{M}$.
Lemma I.1. There exists a unique $L_{M}$ in $\mathcal{K}_{M}$ such that

$$
\begin{equation*}
b\left(L_{M}, \tilde{L}-L_{M}\right) \geq 0 \quad \forall \tilde{L} \in \mathcal{K}_{M} \tag{I.7}
\end{equation*}
$$

and

$$
\begin{equation*}
0 \leq L_{M}(y) \leq \mathcal{V}_{M} \tag{I.8}
\end{equation*}
$$

Lemma I.2. The unique solution of (I.7) and (I.8) coincides with the value function (I.4).
Lemma I.3. The function $L(y)$ defined by (5.15) is the solution of the V.I. (5.20) with

$$
\begin{equation*}
0 \leq L(y) \leq \frac{\delta_{2}}{r}(y-\hat{y})^{+}+C, \tag{I.9}
\end{equation*}
$$

where $C$ is a convenient constant.
Lemma I.4. The set of solutions of (5.20) such that

$$
\begin{equation*}
0 \leq L(y) \leq L^{0}(y) \tag{I.10}
\end{equation*}
$$

has a maximum and a minimum element, where

$$
\begin{equation*}
L^{0}(y)=\frac{1}{2} y^{2}+\frac{\alpha-\lambda \varsigma}{r} y+C^{0} \tag{I.11}
\end{equation*}
$$

with

$$
\begin{equation*}
C^{0}>\frac{1}{2 r^{2}}(\alpha-\lambda \varsigma)^{2}+\max \left(\frac{\delta_{1}}{r \beta}+\frac{\delta_{1}-\delta_{2}}{\delta_{2}} K, \frac{\delta_{2}^{2}}{2 r^{2}}\right) \tag{I.12}
\end{equation*}
$$

We are now ready to prove Theorem 5.2. Since the discontinuity of $\Psi^{\prime}$ occurs at one point only, we may construct smooth functions $\Psi_{\epsilon}(y)$ which approximate $\Psi(y)$ from below or above and which, in addition, have linear growth. Since $\Psi_{\epsilon}(y)$ is smooth, the corresponding solution of the V.I. (5.20) is smoother than $H_{\varrho}^{1}(-\infty, \infty)$. In fact, $\Psi_{\epsilon}^{\prime \prime}(y) \in L_{\varrho}^{2}(-\infty, \infty)$ and the V.I. can be written in the classical form

$$
\left\{\begin{array}{l}
(\alpha-\lambda \varsigma) L_{\epsilon}^{\prime}(y)+\frac{1}{2} L_{\epsilon}^{\prime \prime}(y)-r L_{\epsilon}(y) \leq 0 \\
L_{\epsilon}(y) \geq \Psi_{\epsilon}(y) \\
\left(L_{\epsilon}(y)-\Psi_{\epsilon}(y)\right)\left[(\alpha-\lambda \varsigma) L_{\epsilon}^{\prime}(y)+\frac{1}{2} L_{\epsilon}^{\prime \prime}(y)-r L_{\epsilon}(y)\right]=0 \\
L_{\epsilon}(y) \geq 0 ; L_{\epsilon}(y) \text { has linear growth. }
\end{array}\right.
$$

However, it is classical that smooth solutions are unique and coincide with the value function

$$
\begin{equation*}
L_{\epsilon}(y)=\sup _{\theta \geq 0} \widehat{E}\left[e^{-r \theta} \Psi_{\epsilon}\left(Y_{y}(\theta)\right) \mathbb{1}_{\theta<\infty}\right] . \tag{I.13}
\end{equation*}
$$

Consider now the case when

$$
\begin{equation*}
\Psi_{\epsilon}(y) \uparrow \Psi(y) \quad \text { as } \epsilon \downarrow 0 . \tag{I.14}
\end{equation*}
$$

We may define the increasing process $L_{\epsilon, n}(y)$ as defined in Lemma I.4, and naturally $L_{\epsilon, n}(y) \uparrow$ $L_{\epsilon}(y)$ as $n \uparrow \infty$. However, if we compare $L_{\epsilon, n}(y)$ and $L_{n}(y)$, one may verify that

$$
\begin{equation*}
L_{\epsilon, n}(y) \leq L_{n}(y) \leq \underline{L}(y) . \tag{I.15}
\end{equation*}
$$

We therefore obtain

$$
\begin{equation*}
L_{\epsilon}(y) \leq \underline{L}(y) . \tag{I.16}
\end{equation*}
$$

As $\epsilon \downarrow 0 L_{\epsilon}(y) \uparrow$ toward a solution of (5.20) and from (I.16) it is necessarily the smallest solution. From formula (I.13), one may verify that $L_{\epsilon}(y) \uparrow$ toward the value function (5.15). So the value function coincides with the minimum solution. Consider now an approximation such that

$$
\begin{equation*}
\Psi_{\epsilon}(y) \downarrow \Psi(y) \quad \text { as } \epsilon \downarrow 0 . \tag{I.17}
\end{equation*}
$$

We prove in a similar manner that $L_{\epsilon}(y) \downarrow \bar{L}(y)$, the maximum solution. Therefore, the minimum and maximum solutions coincide. We have shown that the value function (5.15) is the unique solution of (5.20). The property (I.9) is a consequence of the properties of the value function.

Appendix J. Proof of Theorem 6.1. We compute the feedbacks

$$
\left\{\begin{array}{l}
\widehat{C}(x, y)=r x+\delta y-\frac{1}{\gamma}\left(1-\frac{\mu+\frac{\lambda^{2}}{2}}{r}\right)+\frac{\delta}{r}\left(\alpha-\lambda \varsigma \rho-\frac{1}{2} \varsigma^{2} \delta \gamma\left(1-\rho^{2}\right)\right),  \tag{J.1}\\
\widehat{\pi}(x, y) x \sigma=\frac{\lambda}{\gamma r}-\frac{\varsigma \rho \delta}{r}
\end{array}\right.
$$

and note that $U(\widehat{C}(x, y))=r F^{1}(x, y)$. We also have

$$
\widehat{\pi}(x, y) x \sigma \lambda+r x-\widehat{C}(x, y)+\delta y=\frac{1}{\gamma}\left(1-\frac{\mu-\frac{\lambda^{2}}{2}}{r}\right)-\frac{\delta}{r}\left(\alpha-\frac{1}{2} \varsigma^{2} \delta \gamma\left(1-\rho^{2}\right)\right) .
$$

So the corresponding wealth is given by

$$
\begin{equation*}
d \widehat{X}^{1}=\left[\frac{1}{\gamma}\left(1-\frac{\mu-\frac{\lambda^{2}}{2}}{r}\right)-\frac{\delta}{r}\left(\alpha-\frac{1}{2} \varsigma^{2} \delta \gamma\left(1-\rho^{2}\right)\right)\right] d t+\left(\frac{\lambda}{r \gamma}-\frac{\varsigma \rho \delta}{r}\right) d W . \tag{J.2}
\end{equation*}
$$

We also note

$$
\left\{\begin{array}{l}
\frac{\partial F^{1}}{\partial x} \widehat{\pi}(x, y) x \sigma+\frac{\partial F^{1}}{\partial y} \varsigma \rho=-\lambda F^{1}(x, y),  \tag{J.3}\\
\frac{\partial F^{1}}{\partial y} \varsigma \sqrt{1-\rho^{2}}=-\gamma \delta \varsigma \sqrt{1-\rho^{2}} F^{1}(x, y)
\end{array}\right.
$$

Using the Itô differential together with the above calculation, we have

$$
\begin{aligned}
d\left(F^{1}\left(\widehat{X}^{1}(t), Y(t)\right) e^{-\mu t}\right)= & -r e^{-\mu t} F^{1}\left(\widehat{X}^{1}(t), Y(t)\right) d t \\
& -e^{-\mu t} F^{1}\left(\widehat{X}^{1}(t), Y(t)\right)\left[\lambda d W(t)+\gamma \delta \varsigma \sqrt{1-\rho^{2}} d W^{0}(t)\right]
\end{aligned}
$$

Integrating between 0 and $T \wedge \hat{\tau}_{N}$, we can write

$$
\begin{align*}
& E\left[F^{1}\left(\widehat{X}^{1}\left(T \wedge \hat{\tau}_{N}\right), Y\left(T \wedge \hat{\tau}_{N}\right)\right) e^{-\mu T \wedge \hat{\tau}_{N}}\right] \\
& =F^{1}(x, y)-r E\left[\int_{0}^{T \wedge \hat{\tau}_{N}} e^{-\mu t} F^{1}\left(\widehat{X}^{1}(t), Y(t)\right) d t\right] \tag{J.4}
\end{align*}
$$

where $\hat{\tau}_{N}=\inf \left\{t \mid r \widehat{X}^{1}(t)+\delta Y(t) \leq-N\right\}$, and

$$
\exp (-\mu T) E\left\{\exp \left[-\gamma\left(r \widehat{X}^{1}\left(T \wedge \hat{\tau}_{N}\right)+\delta Y\left(T \wedge \hat{\tau}_{N}\right)\right)\right]\right\} \leq \exp (-\gamma(r x+\delta y))
$$

Hence, $\exp (-\mu T) \exp (\gamma N) \mathbb{P}\left(\hat{\tau}_{N}<T\right) \leq \exp (-\gamma(r x+\delta y))$.
Since $\hat{\tau}_{N} \uparrow \hat{\tau}^{*}$ a.s. and $\mathbb{P}\left(\hat{\tau}^{*}<T\right) \leq \mathbb{P}\left(\hat{\tau}_{N}<T\right)$, we get $\mathbb{P}\left(\hat{\tau}^{*}<T\right)=0 \forall T$; hence, $\hat{\tau}^{*}=\infty$ a.s.

Passing to the limit as $N \rightarrow \infty$ in (J.4), we obtain

$$
E\left[F^{1}\left(\widehat{X}^{1}(T), Y(T)\right) e^{-\mu T}\right]=F^{1}(x, y)-r E\left[\int_{0}^{T} e^{-\mu t} F^{1}\left(\widehat{X}^{1}(t), Y(t)\right) d t\right]
$$

therefore, $E\left[F^{1}\left(\widehat{X}^{1}(T), Y(T)\right) e^{-\mu T}\right]=F^{1}(x, y) e^{-r T}$ and the control

$$
\left\{\begin{array}{l}
\widehat{C}(t)=\widehat{C}\left(\widehat{X}^{1}(t), Y(t)\right), \\
\widehat{\pi}(t)=\widehat{\pi}\left(\widehat{X}^{1}(t), Y(t)\right)
\end{array}\right.
$$

is admissible. Clearly,

$$
\begin{equation*}
J(\widehat{C}(\cdot))=F^{1}(x, y) \tag{J.5}
\end{equation*}
$$

On the other hand, we easily check that for any admissible control

$$
J(C(\cdot)) \leq F^{1}(x, y)
$$

hence the desired result has been obtained.
Appendix K. Proof of Theorem 6.2. For each $\hat{y}$ fixed, the solution of (6.26) is the value function of a stochastic control problem. The state equation is

$$
\begin{equation*}
d Y(t)=\left(\alpha-\lambda \varsigma \rho-\gamma \delta \varsigma^{2}\left(1-\rho^{2}\right)+\varsigma \sqrt{r \gamma\left(1-\rho^{2}\right)} v(t)\right) d t+\varsigma d W(t), \quad Y(0)=y<\hat{y} \tag{K.1}
\end{equation*}
$$

where $v(\cdot)$ is adapted and locally square integrable. Let $\theta_{y}(v(\cdot))$ be the first time the process reaches $\hat{y}$; it can be $\infty$. We then have

$$
\begin{equation*}
u(y)=\inf _{v(\cdot)} E\left[\int_{0}^{\theta_{y}(v(\cdot))} e^{-r t}\left(-\delta Y_{y}(t)+r K+\frac{1}{2} v^{2}(t)\right) d t\right] \tag{K.2}
\end{equation*}
$$

For $0<t<\theta_{y}(v(\cdot))$, we have $Y_{y}(t)<\hat{y}$, and thus

$$
\begin{aligned}
E\left[\int_{0}^{\theta_{y}(v(\cdot))} e^{-r t}\left(-\delta Y_{y}(t)+r K+\frac{1}{2} v^{2}(t)\right) d t\right] & \geq \frac{-\delta \hat{y}+r K}{r}\left(1-E e^{-r \theta_{y}(v(\cdot))}\right) \\
& \geq-\frac{(-\delta \hat{y}+r K)^{-}}{r}
\end{aligned}
$$

hence we get the left inequality in (6.27). To prove the right inequality, we just notice that $u(y) \leq z(y)$, where $z(y)$ is the solution of

$$
\left\{\begin{array}{l}
-\frac{1}{2} \varsigma^{2} z^{\prime \prime}-z^{\prime}\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta \gamma\left(1-\rho^{2}\right)\right)+r z=-\delta y+r K \\
z(\hat{y})=0 .
\end{array}\right.
$$

The function $z(y)$ is given explicitly by

$$
\begin{equation*}
z(y)=-\frac{\delta}{r}(y-\hat{y})+\left(1-e^{\beta(y-\hat{y})}\right)\left[\frac{-\delta \hat{y}+r K}{r}-\frac{\delta}{r^{2}}\left(\alpha-\lambda \varsigma \rho-\gamma \delta \varsigma^{2}\left(1-\rho^{2}\right)\right)\right] \tag{K.3}
\end{equation*}
$$

where $\beta>0$ is the solution of

$$
-\frac{1}{2} \varsigma^{2} \beta^{2}-\beta\left(\alpha-\lambda \varsigma \rho-\gamma \delta \varsigma^{2}\left(1-\rho^{2}\right)\right)+r=0
$$

which proves the right inequality in (6.27).
We now check the condition (6.28). We denote by $u(y ; \hat{y})$ the unique solution of (6.26) and (6.27) for a given $\hat{y}$ and $y<\hat{y}$. From (6.27), we have $u(y ; \hat{y}) \geq 0$ if $\hat{y} \leq \frac{r K}{\delta}$; hence $u^{\prime}(\hat{y} ; \hat{y}) \leq 0$ for $\hat{y} \leq \frac{r K}{\delta}$.

From (K.3),

$$
z^{\prime}(y ; \hat{y})=-\frac{\delta}{r}-\beta e^{\beta(y-\hat{y})}\left[\frac{-\delta \hat{y}+r K}{r}-\frac{\delta}{r^{2}}\left(\alpha-\lambda \varsigma \rho-\gamma \delta \varsigma^{2}\left(1-\rho^{2}\right)\right)\right] ;
$$

hence,

$$
z^{\prime}(\hat{y} ; \hat{y})=-\frac{\delta}{r}-\beta\left[\frac{-\delta \hat{y}+r K}{r}-\frac{\delta}{r^{2}}\left(\alpha-\lambda \varsigma \rho-\gamma \delta \varsigma^{2}\left(1-\rho^{2}\right)\right)\right] .
$$

Therefore, $z^{\prime}(\hat{y} ; \hat{y})>0$ if $\hat{y}$ is sufficiently large. It follows that $z(y ; \hat{y})<0$ if $y<\hat{y}$ close to $\hat{y}$. Therefore, $u(y ; \hat{y})<0$ if $y<\hat{y}$ close to $\hat{y}$. Hence $u^{\prime}(\hat{y} ; \hat{y})>0$.

Therefore, there exists $\hat{y}$ such that (6.28) holds. We have $\hat{y} \geq \frac{r K}{\delta}$; otherwise $u^{\prime \prime}(\hat{y})<0$, which is impossible.

We then check that the function $u(y)$ is the solution of the V.I. (6.25). First, we check that $u(y) \geq 0$. Since $u^{\prime \prime}(\hat{y})>0$, we have $u(y)>0$ near $\hat{y}$; hence $u^{\prime}(y)<0$ near $\hat{y}$. The function $u(y)$ cannot have a positive maximum on $\left(\frac{r K}{\delta}, \hat{y}\right)$; hence $u(y)$ decreases on $\left(\frac{r K}{\delta}, \hat{y}\right)$. Therefore, $u\left(\frac{r K}{\delta}\right)>0$. Necessarily, $u(y)>0$ on $\left(-\infty, \frac{r K}{\delta}\right)$. Since for $y>\hat{y},-\delta y+r K \leq-\delta \hat{y}+r K \leq 0$, the differential inequality is also verified for $y>\hat{y}$.

Finally, to complete the proof we prove the uniqueness of $\hat{y}$ and the fact that the V.I. (6.25) has a solution of the form (6.26) and (6.28). We first check that $\hat{y}$ is uniquely defined. For
that we will rely on an interesting interpretation of the solution of (6.26) and (6.28). Again call $u(y ; \hat{y})$ this solution, in which $u(y ; \hat{y})=0$ for $y \geq \hat{y}$. If we define

$$
\chi(y ; \hat{y})=\left\{\begin{array}{cl}
-\delta y+K r, & y<\hat{y} \\
0, & y>\hat{y}
\end{array}\right.
$$

which is an $L^{\infty}$ function (not continuous in $\hat{y}$ ), then the function $u$ appears as the solution of

$$
\left\{\begin{array}{l}
-\frac{1}{2} \varsigma^{2} u^{\prime \prime}-u^{\prime}\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta \gamma\left(1-\rho^{2}\right)\right)+\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) u^{\prime 2}+r u=\chi(y ; \hat{y}),  \tag{K.4}\\
u \text { has linear growth as } y \rightarrow-\infty .
\end{array}\right.
$$

But then $u(y ; \hat{y})$ has the probabilistic interpretation

$$
\begin{equation*}
u(y ; \hat{y})=\inf _{v(\cdot)} E\left[\int_{0}^{\theta_{y}(v(\cdot))} e^{-r t}\left(\chi\left(Y_{y}(t) ; \hat{y}\right)+\frac{1}{2} v^{2}(t)\right) d t\right], \tag{K.5}
\end{equation*}
$$

where the state equation $d Y$ is as in (K.1), $v(\cdot)$ is adapted and locally square integrable, and $\theta_{y}(v(\cdot))$ is the first time the process reaches $\hat{y}$.

The function $\chi(y ; \hat{y})$ decreases with $\hat{y}$; hence $u(y ; \hat{y})$ also decreases with $\hat{y}$. Hence, if $\hat{y}_{1}<\hat{y}_{2}$, we have $u\left(y ; \hat{y}_{1}\right) \geq u\left(y ; \hat{y}_{2}\right)$. But then for $\hat{y}_{1}<y<\hat{y}_{2}, u\left(y ; \hat{y}_{2}\right) \leq 0$. Since it is also $\geq 0, u\left(y, \hat{y}_{2}\right)=0$ for $\hat{y}_{1}<y<\hat{y}_{2}$. Necessarily, $u\left(y ; \hat{y}_{1}\right)=u\left(y ; \hat{y}_{2}\right)$ for $y<\hat{y}_{1}$ since they satisfy the same equation. Therefore, $u\left(y ; \hat{y}_{1}\right)=u\left(y ; \hat{y}_{2}\right)$. Therefore, $\hat{y}_{1}=\hat{y}_{2}$ since we naturally interpret $\hat{y}$ in (6.28) as the first point for which $u(\hat{y})=0, u^{\prime}(\hat{y})=0$.

To complete the proof, it remains to check that a solution of the V.I. (6.25) is necessarily of the form (6.26) and (6.28). Indeed, let us consider $\hat{y}^{*}$ to be the first point for which $u\left(\hat{y}^{*}\right)=0$. We claim that $\hat{y}^{*} \neq \infty$. If $\hat{y}^{*}=\infty$, then $u$ satisfies

$$
\left\{\begin{array}{l}
-\frac{1}{2} \varsigma^{2} u^{\prime \prime}-u^{\prime}\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta \gamma\left(1-\rho^{2}\right)\right)+\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) u^{\prime 2}+r u=-\delta y+r K,  \tag{K.6}\\
u \geq 0 ; \quad u \text { has linear growth as } y \rightarrow-\infty,
\end{array}\right.
$$

which is of the form (K.4) with $\chi(y ; \infty)$. We know a function $u(y ; \hat{y})$ for a certain $\hat{y}$ finite and $u(y ; \hat{y}) \geq u(y ; \infty)$. But then $u(y ; \infty)=0$ for $y>\hat{y}$, which is not possible from (K.6). So $\hat{y}^{*}<\infty$. We claim that $u^{\prime}\left(\hat{y}^{*}\right)=0$. We know that $u^{\prime}\left(\hat{y}^{*}\right) \leq 0$. If $u^{\prime}\left(\hat{y}^{*}\right)<0$, then $u(y)$ becomes negative after $\hat{y}^{*}$, which is impossible. But then $u(y)=u\left(y ; \hat{y}^{*}\right)$ for $y \leq \hat{y}^{*}$. Moreover $\hat{y}^{*}>\frac{r K}{\delta}$. We also have $u(y)=0$ for $y>\hat{y}^{*}$. Otherwise $u(y)$ becomes positive after $\hat{y}^{*}$. But then it cannot always stay positive since we would have (K.6) holding for $y>\hat{y}^{*}$, with a right-hand side negative and $u\left(\hat{y}^{*}\right)=0$. Necessarily, $u(y) \leq 0$, which is impossible. Since $u(y)$ must vanish after $\hat{y}^{*}$, it will have a positive maximum after $\hat{y}^{*}$, which is also not possible. Therefore, $u=0$ for $y \geq \hat{y}^{*}$. The proof has been completed.

## Appendix L. Proof of Theorem 6.3. Consider the feedbacks

$$
\left\{\begin{array}{l}
\widehat{C}(x, y)=r(x+g(y))-\frac{1}{\gamma}\left(1-\frac{\mu+\frac{\lambda^{2}}{2}}{r}\right),  \tag{L.1}\\
\widehat{\pi}(x, y) \sigma x=\frac{\lambda}{r \gamma}-\varsigma \rho g^{\prime}(y),
\end{array}\right.
$$

and we have $U(\widehat{C}(x, y))=r F(x, y)$. Let $\widehat{X}^{0}(t)$ be the optimal wealth trajectory given by (6.4), applying the feedbacks (L.1). We have

$$
\widehat{\pi}(x, y) \sigma \lambda x+x r-\widehat{C}(x, y)=-\lambda \varsigma \rho g^{\prime}(y)-r g(y)+\frac{1}{\gamma}\left(1-\frac{\mu-\frac{\lambda^{2}}{2}}{r}\right)
$$

hence

$$
\left\{\begin{array}{l}
d \widehat{X}^{0}(t)=\left[-r g(Y(t))-\lambda \varsigma \rho g^{\prime}(Y(t))+\frac{1}{\gamma}\left(1-\frac{\mu-\frac{\lambda^{2}}{2}}{r}\right)\right] d t+\left(\frac{\lambda}{r \gamma}-\varsigma \rho g^{\prime}(Y(t))\right) d W(t)  \tag{L.2}\\
\widehat{X}^{0}(0)=x \\
d Y(t)=\alpha d t+\varsigma\left(\rho d W(t)+\sqrt{1-\rho^{2}} d W^{0}(t)\right) \\
Y(0)=y
\end{array}\right.
$$

Let $\hat{\tau}(y)=\inf \left\{t \mid Y_{y}(t) \geq \hat{y}\right\}$, which does not depend on $x$. We have $E[\hat{\tau}(y)]=\frac{\hat{y}-y}{\alpha}$ for $y<\hat{y}$; therefore, $\hat{\tau}(y)<\infty$ a.s. As in (J.4), we can write

$$
E\left[F\left(\widehat{X}^{0}\left(\hat{\tau} \wedge \hat{\tau}_{N}\right), Y\left(\hat{\tau} \wedge \hat{\tau}_{N}\right)\right) e^{-\mu \hat{\tau} \wedge \hat{\tau}_{N}}\right]-F(x, y)=-r E\left[\int_{0}^{\hat{\tau} \wedge \hat{\tau}_{N}} e^{-\mu t} F\left(\widehat{X}^{0}(t), Y(t)\right) d t\right]
$$

where $\hat{\tau}_{N}=\inf \left\{t \mid \widehat{X}^{0}(t) \leq-N\right\}$, and we deduce

$$
E\left\{\exp \left[-r \gamma\left(\widehat{X}^{0}\left(\hat{\tau} \wedge \hat{\tau}_{N}\right)+g\left(Y\left(\hat{\tau} \wedge \hat{\tau}_{N}\right)\right)\right)\right] \exp \left(-\mu \hat{\tau} \wedge \hat{\tau}_{N}\right)\right\} \leq \exp (-\gamma r(x+g(y))
$$

Since $Y\left(\hat{\tau} \wedge \hat{\tau}_{N}\right)<\hat{y}, g\left(Y\left(\hat{\tau} \wedge \hat{\tau}_{N}\right)\right)$ is bounded; hence $\exp (r \gamma N) E\left[\exp (-\mu \hat{\tau}) \mathbb{1}_{\hat{\tau}_{N}<\hat{\tau}}\right] \leq$ $\exp (-\gamma r(x+g(y)))$. Therefore, if $\hat{\tau}^{*}=\lim \uparrow \hat{\tau}_{N}$, we obtain $\hat{\tau}^{*} \geq \hat{\tau}_{N}$ a.s. It follows that

$$
\left\{\begin{array}{l}
\widehat{C}(t)=\widehat{C}\left(\widehat{X}^{0}(t), Y(t)\right) \\
\widehat{\pi}(t)=\widehat{\pi}\left(\widehat{X}^{0}(t), Y(t)\right) \\
\hat{\tau}(y)
\end{array}\right.
$$

forms an admissible control. We obtain $F(x, y)=J_{x, y}(\widehat{C}(\cdot), \widehat{\pi}(\cdot), \hat{\tau})$.
On the other hand, for any admissible control, we have

$$
F(x, y) \geq E\left[\int_{0}^{\tau \wedge \tau_{N}} U(C(t)) e^{-\mu t} d t+F\left(X^{0}\left(\tau \wedge \tau_{N}\right), Y\left(\tau \wedge \tau_{N}\right)\right) \exp \left(\tau \wedge \tau_{N}\right)\right]
$$

Letting $N$ tend to $\infty$, we obtain $F(x, y) \geq J_{x y}(C(\cdot), \pi(\cdot), \tau)$, which completes the proof of the desired result.

Appendix M. Proof of Theorem 6.4. We consider the feedbacks

$$
\left\{\begin{array}{l}
\widehat{C}(x, y)=r(x+g(y))-\frac{1}{\gamma}\left(1-\frac{\mu+\frac{\lambda^{2}}{2}}{r}\right)  \tag{M.1}\\
\widehat{\pi}(x, y) \sigma x=\frac{\lambda}{r \gamma}-\varsigma \rho g^{\prime}(y)
\end{array}\right.
$$

We have $U(\widehat{C}(x, y))=r L^{1}(x, y)$ if $y<\hat{y}$. The corresponding wealth process $\widehat{X}^{1}(t)$ has the Itô differential

$$
\begin{equation*}
d \widehat{X}^{1}=\left[\delta_{1} Y-r g(Y)-\lambda \varsigma \rho g^{\prime}(Y)+\frac{1}{\gamma}\left(1-\frac{\mu-\frac{\lambda^{2}}{2}}{r}\right)\right] d t+\left(\frac{\lambda}{r \gamma}-\varsigma \rho g^{\prime}(Y)\right) d W \tag{M.2}
\end{equation*}
$$

Let $\hat{\tau}_{N}$ be the first time when $\widehat{X}^{1}(t)$ reaches $-N$. We get

$$
\begin{aligned}
& E\left[L^{1}\left(\widehat{X}^{1}\left(\hat{\tau}(y) \wedge \hat{\tau}_{N}\right), Y\left(\hat{\tau}(y) \wedge \hat{\tau}_{N}\right)\right) e^{-\mu \hat{\tau}(y) \wedge \hat{\tau}_{N}}\right]-L^{1}(x, y) \\
& =-r E\left[\int_{0}^{\hat{\tau}(y) \wedge \hat{\tau}_{N}} L^{1}\left(\widehat{X}^{1}(t), Y(t)\right) e^{-\mu t} d t\right] .
\end{aligned}
$$

We deduce

$$
E\left[L^{1}\left(\widehat{X}^{1}\left(\hat{\tau}(y) \wedge \hat{\tau}_{N}\right), Y\left(\hat{\tau}(y) \wedge \hat{\tau}_{N}\right)\right) e^{-\mu \hat{\tau}(y) \wedge \hat{\tau}_{N}}\right] \leq e^{-\gamma r(x+g(y))}
$$

Calling $\hat{\tau}^{*}=\lim \uparrow \hat{\tau}_{N}$, we obtain $\hat{\tau}^{*} \geq \hat{\tau}(y)$ a.s. Therefore, the control

$$
\left\{\begin{array}{l}
\widehat{C}(t)=\widehat{C}\left(\widehat{X}^{1}(t), Y(t)\right), \\
\widehat{\pi}(t)=\widehat{\pi}\left(\widehat{X}^{1}(t), Y(t)\right)
\end{array}\right.
$$

is admissible. Moreover,

$$
\begin{aligned}
L^{1}(x, y)= & E\left[\int_{0}^{\hat{\tau}(y) \wedge \hat{\tau}_{N}} U(\widehat{C}(t)) e^{-\mu t} d t\right] \\
& +E\left[L^{1}\left(\widehat{X}^{1}\left(\hat{\tau}(y) \wedge \hat{\tau}_{N}\right), Y\left(\hat{\tau}(y) \wedge \hat{\tau}_{N}\right)\right) e^{-\mu \hat{\tau}(y) \wedge \hat{\tau}_{N}}\right]
\end{aligned}
$$

and, from the boundary condition that $L^{1}(x, y)$ satisfies at $y=\hat{y}$, we get

$$
L^{1}(x, y)=J(\widehat{C}(\cdot), \widehat{\pi}(\cdot))
$$

On the other hand, for any admissible pair $(C(\cdot), \pi(\cdot))$, we have

$$
\begin{aligned}
L^{1}(x, y) \geq E & {\left[\int_{0}^{\hat{\tau}(y) \wedge \tau_{N}} U(C(t)) e^{-\mu t} d t\right.} \\
& \left.+L^{1}\left(X^{1}\left(\hat{\tau}(y) \wedge \tau_{N}\right), Y\left(\hat{\tau}(y) \wedge \tau_{N}\right)\right) e^{-\mu \hat{\tau}(y) \wedge \tau_{N}}\right]
\end{aligned}
$$

and letting $N$ tend to $\infty$, we obtain $L^{1}(x, y) \geq J(C(\cdot), \pi(\cdot))$ and the proof is completed.
Appendix N. Proof of Theorem 6.5. We consider the ceiling function

$$
\begin{equation*}
\bar{u}(y)=\frac{\delta_{2}}{r}\left[-(y-k \hat{y})+\frac{e^{\beta(y-k \hat{y})}-1}{\beta}\right] \tag{N.1}
\end{equation*}
$$

where $k$ is sufficiently large. This function is such that $\bar{u}(k \hat{y})=\bar{u}^{\prime}(k \hat{y})=0$. We extend it by 0 for $y>k \hat{y}$. Also, $\bar{u}(y)>0$ for $y<k \hat{y}$. We check

$$
\begin{equation*}
\bar{u}>m . \tag{N.2}
\end{equation*}
$$

Since $m=0$ for $y>\hat{y}$, it is sufficient to prove (N.2) for $y<\hat{y}$. However,

$$
\begin{align*}
& -\frac{1}{2} \varsigma^{2} \bar{u}^{\prime \prime}-\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta_{2} \gamma\left(1-\rho^{2}\right)^{\prime}\right) \bar{u}^{\prime}+\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) \bar{u}^{\prime 2}+r \bar{u} \\
& \geq \frac{\delta_{2}}{r}\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta_{2} \gamma\left(1-\rho^{2}\right)\right)-\delta_{2}(y-k \hat{y})-\frac{\delta_{2}}{\beta} \\
& \geq\left(\delta_{1}-\delta_{2}\right) y \text { for } y<\hat{y} \tag{N.3}
\end{align*}
$$

provided $\frac{\delta_{2}}{r}\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta_{2} \gamma\left(1-\rho^{2}\right)\right)+\delta_{2} k \hat{y}-\frac{\delta_{2}}{\beta}>\delta_{1} \hat{y}$. This implies (N.2). We also have

$$
\frac{\delta_{2}}{r}\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta_{2} \gamma\left(1-\rho^{2}\right)\right)-\delta_{2}(y-k \hat{y})-\frac{\delta_{2}}{\beta}>\delta_{2} y+r K
$$

provided $\frac{\delta_{2}}{r}\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta_{2} \gamma\left(1-\rho^{2}\right)\right)+\delta_{2} k \hat{y}-\frac{\delta_{2}}{\beta}>r K$. Therefore, $\bar{u}$ satisfies the two inequalities (6.52), so it can be taken as a ceiling function.

Combining with (6.55), we must have

$$
\begin{equation*}
-f(y)+k \leq u(y) \leq \bar{u}(y) \tag{N.4}
\end{equation*}
$$

Since $f(y)=\frac{\delta_{2}}{r}\left(y-\tilde{y}_{0}\right)$ with $\tilde{y}_{0}=-\frac{\alpha-\lambda \varsigma \rho-\delta_{2} \gamma \varsigma^{2}\left(1-\rho^{2}\right)}{r}+\frac{\delta_{2}^{2} \gamma \varsigma^{2}\left(1-\rho^{2}\right)}{r^{2}}$, we have $u+\frac{\delta_{2} y}{r}$ bounded as $y \rightarrow-\infty$.

We use the penalty technique to approximate (6.52). We approximate (6.52) by a smoother problem, a penalized problem. We introduce the problem with $u_{\epsilon}$ such that

$$
\left\{\begin{array}{c}
-\frac{1}{2} \varsigma^{2} u_{\epsilon}^{\prime \prime}-\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta_{2} \gamma\left(1-\rho^{2}\right)\right) u_{\epsilon}^{\prime}+\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) u_{\epsilon}^{\prime 2}+r u_{\epsilon}  \tag{N.5}\\
=-\delta_{2} y+r K+\frac{1}{\epsilon}\left(m-u_{\epsilon}\right)^{+} \\
u_{\epsilon}(\bar{y})=0 ; \quad u_{\epsilon}(y)+f(y) \text { is bounded as } y \rightarrow-\infty
\end{array}\right.
$$

In (N.5), $\bar{y}=k \hat{y}$, where $k$ is sufficiently large.
We have the property

$$
\begin{equation*}
u_{\epsilon}(y) \leq \bar{u}(y) \tag{N.6}
\end{equation*}
$$

Let us define $\tilde{u}_{\epsilon}(y)=u_{\epsilon}(y)-m(y)$. We have $\tilde{u}_{\epsilon}(\bar{y})=0$ and $\tilde{u}_{\epsilon} \rightarrow \infty$ as $y \rightarrow-\infty$. If $\tilde{u}_{\epsilon}(y) \geq 0$ on $(-\infty, \bar{y})$, then $\left(m-u_{\epsilon}\right)^{+}=0$.

Suppose that $\tilde{u}_{\epsilon}(y)$ becomes negative; then we can consider points of negative minimum. We check that $\hat{y}$ cannot be such a point. Indeed, we know that $m^{\prime}(\hat{y}-0)<0$ from (6.54). If $\hat{y}$ were a minimum of $\tilde{u}_{\epsilon}$, we would have $\tilde{u}_{\epsilon}(\hat{y}-0) \leq 0 \leq \tilde{u}_{\epsilon}(\hat{y}+0)$ and since $u_{\epsilon}^{\prime}(\hat{y}-0)=u_{\epsilon}^{\prime}(\hat{y}+0)$, $m^{\prime}(\hat{y}-0) \geq 0$, which is a contradiction. Therefore, if $y_{\epsilon}^{*}$ is a negative minimum of $\tilde{u}_{\epsilon}(y)$, we have $y_{\epsilon}^{*}<\hat{y}$ or $y_{\epsilon}^{*}>\hat{y}$.

If $y_{\epsilon}^{*}>\hat{y}$, it is also a local minimum of $u_{\epsilon}$. But then considering (N.5)

$$
r u_{\epsilon}\left(y_{\epsilon}^{*}\right) \geq-\delta_{2} \bar{y}+r K-\frac{1}{\epsilon} u_{\epsilon}\left(y_{\epsilon}^{*}\right)
$$

which implies $u_{\epsilon}\left(y_{\epsilon}^{*}\right) \geq \epsilon\left(-\delta_{2} \bar{y}+r K\right)$; hence $\tilde{u}_{\epsilon}\left(y_{\epsilon}^{*}\right) \geq \epsilon\left(-\delta_{2} \bar{y}+r K\right)$.
Suppose now $y_{\epsilon}^{*}<\hat{y}$. We note that on $(-\infty, \hat{y})$, $\tilde{u}_{\epsilon}$ satisfies

$$
\begin{array}{r}
-\frac{1}{2} \varsigma^{2} \tilde{u}_{\epsilon}^{\prime \prime}-\left(\alpha-\lambda \varsigma \rho-\left(\frac{\delta_{2}}{r}+m^{\prime}\right) \varsigma^{2} r \gamma\left(1-\rho^{2}\right)\right) \tilde{u}_{\epsilon}^{\prime}+\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) \tilde{u}_{\epsilon}^{\prime 2}+r \tilde{u}_{\epsilon} \\
=-\delta_{1} y+r K+\frac{1}{\epsilon} \tilde{u}_{\epsilon}^{-}
\end{array}
$$

but then $r \tilde{u}_{\epsilon}\left(y_{\epsilon}^{*}\right) \geq-\delta_{1} \hat{y}+r K-\frac{1}{\epsilon} \tilde{u}_{\epsilon}\left(y_{\epsilon}^{*}\right)$, and $\tilde{u}_{\epsilon}\left(y_{\epsilon}^{*}\right) \geq \epsilon\left(-\delta_{1} \hat{y}+r K\right)$. Therefore,

$$
\tilde{u}_{\epsilon}(y)=u_{\epsilon}(y)-m(y) \geq \epsilon\left(r K-\max \left(\delta_{2} \bar{y}, \delta_{1} \hat{y}\right)\right) \geq-\epsilon c_{0}
$$

with $c_{0}=\max \left(\delta_{2} \bar{y}, \delta_{1} \hat{y}\right)>0$, and we get

$$
\begin{equation*}
\frac{\left(m-u_{\epsilon}\right)^{+}}{\epsilon} \leq c_{0} \tag{N.7}
\end{equation*}
$$

Consider next $\underline{u}(y)$ to be the solution of
(N.8)

$$
\left\{\begin{array}{l}
-\frac{1}{2} \varsigma^{2} \underline{u}^{\prime \prime}-\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta_{2} \gamma\left(1-\rho^{2}\right)\right) \underline{u}^{\prime}+\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) \underline{u}^{\prime 2}+r \underline{u}=-\delta_{2} y+r K, \quad y<\underline{y} \\
\underline{u}(\underline{y})=0
\end{array}\right.
$$

Then we have the estimates

$$
\begin{equation*}
-f(y)+K \leq \underline{u}(y) \leq u_{\epsilon}(y) \leq \bar{u}(y) \tag{N.9}
\end{equation*}
$$

Since $\underline{u}(\underline{y})=\bar{u}(\underline{y})=0$, we have

$$
\begin{equation*}
\underline{u}^{\prime}(\underline{y}-0) \geq u_{\epsilon}^{\prime}(\underline{y}-0) \geq \bar{u}^{\prime}(\underline{y})=0 \tag{N.10}
\end{equation*}
$$

From the estimates (N.9), we have

$$
\begin{equation*}
\frac{\left|u_{\epsilon}(y)\right|}{\left(1+y^{2}\right)^{\varrho}} \leq C \quad \forall \varrho>\frac{1}{2} \tag{N.11}
\end{equation*}
$$

We now check

$$
\begin{equation*}
\frac{\left|u_{\epsilon}^{\prime}(y)\right|}{\left(1+y^{2}\right)^{\varrho}} \leq C \quad \forall \varrho>\frac{1}{2}, y \leq \underline{y} \tag{N.12}
\end{equation*}
$$

with $u_{\epsilon}^{\prime}(\underline{y})=u_{\epsilon}^{\prime}(\underline{y}-0)$.
From (N.10), we know that this is the time for $y=\underline{y}$.

Consider $z_{\epsilon}=\frac{u_{\epsilon}}{\left(1+y^{2}\right)^{\varrho}}$; then $z_{\epsilon}$ satisfies

$$
-\frac{1}{2} \frac{\varsigma^{2} z_{\epsilon}^{\prime \prime}}{\left(1+y^{2}\right)^{\varrho}}-z_{\epsilon}^{\prime}\left\{-\frac{2 \varrho r \gamma \varsigma^{2}\left(1-\rho^{2}\right) y}{1+y^{2}}+\frac{2 \varrho y}{\left(1+y^{2}\right)^{\varrho+1}}+\frac{\alpha-\lambda \varsigma \rho-\gamma \delta_{2} \varsigma^{2}\left(1-\rho^{2}\right)}{\left(1+y^{2}\right)^{\varrho}}\right\}
$$

$$
\begin{equation*}
+\frac{1}{2} r \gamma \varsigma^{2}\left(1-\rho^{2}\right) z_{\epsilon}^{\prime 2}=\zeta_{\epsilon}(y), \quad y<\bar{y} \tag{N.13}
\end{equation*}
$$

and $\left|\zeta_{\epsilon}(y)\right| \leq C$ if $\varrho>\frac{1}{2}$.
Consider a point where $z_{\epsilon}^{\prime}$ is maximum positive or minimum negative. Let $y_{\epsilon}$ be such a point. If $y_{\epsilon}<\bar{y}$, then $z_{\epsilon}^{\prime \prime}\left(y_{\epsilon}\right)=0$ and from (N.13) we necessarily have $\left|z_{\epsilon}^{\prime}\left(y_{\epsilon}\right)\right| \leq C$. Now, on $\underline{y}$, we have $z_{\epsilon}^{\prime}(\bar{y}-0)=\frac{u_{\epsilon}^{\prime}(\underline{y-0}-0}{\left(1+y^{2}\right)^{\varrho}}$, and thus $\left|z_{\epsilon}^{\prime}(y)\right| \leq C$. Since $\frac{u_{\epsilon}^{\prime}(y)}{\left(1+y^{2}\right)^{\varrho}}=z_{\epsilon}^{\prime}(y)+\frac{2 \varrho y}{1+y^{2}} z_{\epsilon}(y)$, we obtain (N.12). From (N.5), we deduce

$$
\begin{equation*}
\frac{\left|u_{\epsilon}^{\prime \prime}(y)\right|}{\left(1+y^{2}\right)^{2 \varrho}} \leq C \tag{N.14}
\end{equation*}
$$

Therefore, we can extract from $u_{\epsilon}(y)$ a subsequence, still denoted $u_{\epsilon}(y)$, such that

$$
\begin{equation*}
u_{\epsilon}(y) \rightarrow u(y), u_{\epsilon}^{\prime}(y) \rightarrow u(y) \text { pointwise on }(-\infty, \bar{y}) \tag{N.15}
\end{equation*}
$$

Calling $\chi^{\epsilon}=\frac{1}{\epsilon}\left(m-u_{\epsilon}\right)^{+}$, we also have

$$
\left\{\begin{array}{l}
\chi^{\epsilon} \rightarrow \chi \text { in } L^{\infty}(-\infty, \bar{y}) \text { weak star }  \tag{N.16}\\
\frac{u_{\epsilon}^{\prime \prime}(y)}{\left(1+y^{2}\right)^{2 \varrho}} \rightarrow \frac{u^{\prime \prime}(y)}{\left(1+y^{2}\right)^{2 \varrho}} \text { in } L^{\infty}(-\infty, \bar{y}) \text { weak star. }
\end{array}\right.
$$

The limit $u$ satisfies
(N.17)

$$
\left\{\begin{array}{l}
-\frac{1}{2} \varsigma^{2} u^{\prime \prime}-\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta_{2} \gamma\left(1-\rho^{2}\right)\right) u^{\prime}+\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) u^{\prime 2}+r u=-\delta_{2} y+r K+\chi \\
u \geq m \\
(u-m) \chi=0 \\
u(\underline{y})=0
\end{array}\right.
$$

Then, necessarily,

$$
\begin{equation*}
u^{\prime}(\underline{y}-0)=0 . \tag{N.18}
\end{equation*}
$$

Indeed $u(y) \geq m(y)$ is positive for $y$ close to $\underline{y}$; hence $u^{\prime}(\underline{y}-0) \leq 0$. Moreover, from (N.10), $u^{\prime}(\underline{y}-0) \geq 0$. Hence, we have (N.18). Therefore, $u$ extended by 0 beyond $\underline{y}$ remains $C^{1}$ and is the solution of the V.I. (6.52). We can take $\underline{y}$ to be the first point where $u(y)$ touches 0 . Necessarily, $u(y) \geq 0$.

Consider again the first point $y^{*}<\hat{y}$ such that $m\left(y^{*}\right)=0$. Then $y^{*}<0$ and $m(y)<0$ for $y<y^{*}$. Therefore, for $y<y^{*}$ the following differential equation holds:

$$
\begin{aligned}
-\frac{1}{2} \varsigma^{2} u^{\prime \prime}-\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta_{2} \gamma\left(1-\rho^{2}\right)\right) u^{\prime} & +\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) u^{\prime 2}+r u \\
& =-\delta_{2} y+r K, y<y^{*}
\end{aligned}
$$

There must exist a point $y^{* *}<y^{*}$ such that $u^{\prime}\left(y^{* *}\right)<0$. Otherwise, $u^{\prime}(y) \geq 0 \forall y<y^{*}$; hence $u(y)$ is bounded for $y<y^{*}$, which is impossible since $u(y) \rightarrow \infty$ as $y \rightarrow-\infty$. We claim

$$
\begin{equation*}
u^{\prime}(y)<0 \quad \text { if } y<y^{* *} . \tag{N.20}
\end{equation*}
$$

Otherwise, there would be a point of maximum. From (N.19) the value of this maximum would be negative, which is impossible. Hence, we have (N.20).

Define the feedback $\hat{v}(y)=-u^{\prime}(y) \varsigma \sqrt{r \gamma\left(1-\rho^{2}\right)}$, and from (N.20)

$$
\begin{equation*}
\hat{v}(y) \geq \hat{v}(y) \mathbb{1}_{y^{* *}<y<\bar{y}} \geq v_{0} \tag{N.21}
\end{equation*}
$$

We associate with (N.5) a stochastic differential game. The state equation is governed by

$$
\left\{\begin{array}{l}
d Y(t)=\left(\alpha-\lambda \varsigma \rho-\varsigma^{2} \delta_{2} \gamma\left(1-\rho^{2}\right)+v(t) \varsigma \sqrt{r \gamma\left(1-\rho^{2}\right)}\right) d t+\varsigma d W(t),  \tag{N.22}\\
Y(0)=y, \quad y<\bar{y}
\end{array}\right.
$$

where $v(t)$ is the control (adapted process). We define the cost function

$$
\begin{equation*}
J_{y}^{\epsilon}(v(\cdot), \vartheta(\cdot))=E\left[\int_{0}^{\bar{\theta}}\left(-\delta_{2} Y_{y}(t)+r K+\frac{1}{2} v^{2}(t)+\frac{1}{\epsilon} m\left(Y_{y}(t)\right) \vartheta(t)\right) e^{-\int_{0}^{t}\left(r+\frac{\vartheta(s)}{\epsilon}\right) d s} d t\right] \tag{N.23}
\end{equation*}
$$

where $\bar{\theta}=\inf \left\{t \mid Y_{y}(t)=\bar{y}\right\}$. We then have

$$
\begin{equation*}
u_{\epsilon}(y)=\inf _{v(\cdot)} \sup _{\vartheta(\cdot)} J_{y}^{\epsilon}(v(\cdot), \vartheta(\cdot))=\sup _{\vartheta(\cdot)} \inf _{v(\cdot)} J_{y}^{\epsilon}(v(\cdot), \vartheta(\cdot)) \tag{N.24}
\end{equation*}
$$

Thus, there exists a saddle point. Optimal processes, $\hat{v}_{\epsilon}(t)$ and $\hat{\vartheta}_{\epsilon}$, are obtained from feedbacks

$$
\left\{\begin{array}{l}
\hat{v}_{\epsilon}(y)=-u_{\epsilon}^{\prime}(y) \varsigma \sqrt{r \gamma\left(1-\rho^{2}\right)}, \\
\hat{\vartheta}_{\epsilon}(y)=\mathbb{1}_{m(y)>u_{\epsilon}(y)}
\end{array}\right.
$$

by the formula

$$
\left\{\begin{array}{l}
\hat{v}_{\epsilon}(t)=\hat{v}_{\epsilon}\left(Y_{y}(t)\right), \\
\hat{\vartheta}_{\epsilon}(t)=\hat{\vartheta}_{\epsilon}\left(Y_{y}(t)\right) .
\end{array}\right.
$$

Consider the optimal trajectory $\widehat{Y}_{y}(t)$ given by (see (N.22))

$$
\begin{equation*}
d \hat{Y}=\left(\alpha-\lambda \varsigma \rho-\gamma \delta_{2} \varsigma^{2}\left(1-\rho^{2}\right)+\hat{v}(Y) \varsigma \sqrt{r \gamma\left(1-\rho^{2}\right)}\right) d t+\varsigma d W \tag{N.25}
\end{equation*}
$$

We have from (N.21) $d \widehat{Y} \geq-c_{1} d t+\varsigma d W$; hence $E \widehat{Y}_{y}(t) \geq y-c_{1} t$. Therefore, from the estimates (N.4) and (N.1),

$$
\begin{equation*}
E u\left(\widehat{Y}_{y}(t)\right) \leq \frac{\delta_{2}}{r}\left[-E \widehat{Y}_{y}(t)+c_{2}\right] \leq \frac{\delta_{2}}{r}\left[-y+c_{1} t+c_{2}\right] . \tag{N.26}
\end{equation*}
$$

We next define optimal stopping:

$$
\begin{equation*}
\hat{\theta}=\hat{\theta}(y)=\inf \left\{t \mid u\left(\widehat{Y}_{y}(t)\right)=m\left(\widehat{Y}_{y}(t)\right)\right\} . \tag{N.27}
\end{equation*}
$$

Using Itô's formula, we can write

$$
u(y)=E\left[\int_{0}^{T \wedge \hat{\theta}}\left(-\delta_{2} \widehat{Y}_{y}(t)+r K+\frac{1}{2} \hat{v}^{2}(t)\right) e^{-r t} d t+u\left(Y_{y}(T \wedge \hat{\theta})\right) e^{-r T \wedge \hat{\theta}}\right]
$$

where $\hat{v}(t)=\hat{v}\left(\widehat{Y}_{y}(t)\right)$. Therefore,

$$
\begin{aligned}
u(y)=E[ & \int_{0}^{T \wedge \hat{\theta}}\left(-\delta_{2} \widehat{Y}_{y}(t)+r K+\frac{1}{2} v^{2}(t)\right) e^{-r t} d t+m\left(Y_{y}(\hat{\theta})\right) e^{-r T} \mathbb{1}_{\hat{\theta}<T} \\
& \left.+u\left(Y_{y}(T)\right) e^{-r T} \mathbb{1}_{T \leq \hat{\theta}}\right]
\end{aligned}
$$

Letting $T \uparrow \infty$, making use of the estimate (N.27), we obtain

$$
\begin{equation*}
u(y)=J_{y}(\hat{v}(\cdot), \hat{\theta}) \tag{N.28}
\end{equation*}
$$

Define the feedback $\hat{\vartheta}(y)=\mathbb{1}_{u(y)=m(y)}$.
We define next the cost associated with any control $v(\cdot)$ and the feedback $\hat{\vartheta}(\cdot)$ :

$$
J_{y}(v(\cdot), \hat{\vartheta}(\cdot))=E\left[\int_{0}^{\hat{\theta}}\left(-\delta_{2} Y_{y}(t)+r K+\frac{1}{2} v^{2}(t)\right) e^{-r t} d t+m\left(Y_{y}(\hat{\theta})\right) e^{-r \hat{\theta}} \mathbb{1}_{\hat{\theta}<\infty}\right]
$$

where $\hat{\theta}=\inf \left\{t \mid u\left(Y_{y}(t)\right)=m\left(Y_{y}(t)\right)\right\}$. We check

$$
\begin{aligned}
u(y) \leq E & {\left[\int_{0}^{T \wedge \hat{\theta}}\left(-\delta_{2} Y_{y}(t)+r K+\frac{1}{2} v^{2}(t)\right) e^{-r t} d t+u\left(Y_{y}(T \wedge \hat{\theta})\right) e^{-r T \wedge \hat{\theta}}\right] } \\
=E & {\left[\int_{0}^{T \wedge \hat{\theta}}\left(-\delta_{2} Y_{y}(t)+r K+\frac{1}{2} v^{2}(t)\right) e^{-r t} d t+m\left(Y_{y}(\hat{\theta})\right) e^{-r \hat{\theta}} \mathbb{1}_{\hat{\theta}<T}\right.} \\
& \left.+u\left(Y_{y}(T)\right) e^{-r T}\right] .
\end{aligned}
$$

Again, $E u\left(Y_{y}(T)\right) e^{-r T} \leq \frac{\delta_{2}}{r}\left[-E Y_{y}(T) e^{-r T}+c_{2} e^{-r T}\right]$ and from the transversality condition $E u\left(Y_{y}(T)\right) e^{-r T} \rightarrow 0$ as $T \rightarrow \infty$. Therefore,

$$
u(y) \leq J_{y}(v(\cdot), \hat{\vartheta}(\cdot))
$$

thus $u(y) \leq \sup _{\theta} \inf _{v(\cdot)} J_{y}(v(\cdot), \theta)$.
Now, considering again $\hat{Y}_{y}(t)$, we have for any stopping time $\theta$

$$
\begin{aligned}
u(y) & \geq E\left[\int_{0}^{T \wedge \theta}\left(-\delta_{2} \widehat{Y}_{y}(t)+r K+\frac{1}{2} v^{2}(t)\right) e^{-r t} d t+u\left(\widehat{Y}_{y}(T \wedge \theta)\right) e^{-r T \wedge \theta}\right] \\
& \geq E\left[\int_{0}^{T \wedge \theta}\left(-\delta_{2} \widehat{Y}_{y}(t)+r K+\frac{1}{2} v^{2}(t)\right) e^{-r t} d t+m\left(\widehat{Y}_{y}(\theta)\right) e^{-r e \theta} \mathbb{1}_{\theta<T}\right]
\end{aligned}
$$

and, letting $T \rightarrow \infty$,

$$
u(y) \geq J_{y}(\hat{v}(\cdot), \theta) .
$$

Therefore, $u(y) \geq \inf _{v(\cdot)} \sup _{\theta} J_{y}(v(\cdot), \theta)$. This completes the proof.
Appendix O. Proof of Theorem 6.6. We take $y_{1}$ to be the first point such that $u\left(y_{1}\right)=$ $m\left(y_{1}\right)$. We must have $y_{1}<\hat{y}$; otherwise we are in the case $m \leq 0$, which is excluded. We also have

$$
\begin{equation*}
u^{\prime}\left(y_{1}\right)=m^{\prime}\left(y_{1}\right) \tag{0.1}
\end{equation*}
$$

and $\delta_{1} y_{1} \geq r K$.
Indeed, set $\tilde{u}(y)=u(y)-m(y)$. Then $\tilde{u}(y)$ satisfies

$$
-\frac{1}{2} \varsigma^{2} \tilde{u}^{\prime \prime}-\left(\alpha-\lambda \varsigma \rho-\left(\frac{\delta_{2}}{r}-m^{\prime}\right) \varsigma^{2} r \gamma\left(1-\rho^{2}\right)\right) \tilde{u}^{\prime}+\frac{1}{2} \varsigma^{2} r \gamma\left(1-\rho^{2}\right) \tilde{u}^{\prime 2}+r \tilde{u}
$$

$$
\begin{equation*}
=-\delta_{1} y+r K \tag{O.2}
\end{equation*}
$$

with $\tilde{u}\left(y_{1}\right)=0, \tilde{u}^{\prime}\left(y_{1}\right)=0$. The matching of the derivatives comes from the fact that $\tilde{u}(y)$ is $C^{1}$ and $\tilde{u}(y) \geq 0, \tilde{u}\left(y_{1}\right)=0$. So $y_{1}$ is the local minimum; hence $\tilde{u}^{\prime}\left(y_{1}\right)=0$ and we have (O.1). Now suppose $\delta_{1} y_{1} \leq r K$; from (O.2), we see that $\tilde{u}^{\prime \prime}\left(y_{1}-0\right)<0$; therefore, $\tilde{u}(y)<0$ for $y<y_{1}$ close to $y_{1}$, which is impossible.

Call $y_{2}$ the left end of the interval $\left(y_{2}, y_{3}\right)$ with $y_{3}=\bar{y}$ and $y_{2}<\hat{y}$, on which the equation holds. So we have $u\left(y_{2}\right)=m\left(y_{2}\right)$, and necessarily $u^{\prime}\left(y_{2}\right)=m^{\prime}\left(y_{2}\right)$.

On the other hand, on the interval ( $y_{1}, y_{2}$ ), $m$ satisfies all conditions (6.52). So $u=m$ on the interval $\left(y_{1}, y_{2}\right)$. By the uniqueness of $u$, the triple $y_{1}, y_{2}, y_{3}$ is necessarily unique.

## REFERENCES

[1] A. Bensoussan and J.-L. Lions, Applications of Variational Inequalities in Stochastic Control, NorthHolland, Amsterdam, 1982.
[2] F. Black and M. Scholes, The pricing of corporate liabilities, J. Political Economy, 81 (1973), pp. 637-659.
[3] M. Brennan and E. Schwartz, Regulation and corporate investment policy, J. Finance, 37 (1982), pp. 289-300.
[4] M. Brennan and E. Schwartz, Consistent regulatory policy under uncertainty, Bell J. Econom., 13 (1982), pp. 506-524.
[5] M. Brennan and E. Schwartz, Evaluating natural resource investments, J. Business, 58 (1985), pp. 135-157.
[6] Y. Chen, B. Hobbs, S. Leyffer, and T. Munson, Leader-follower equilibria for electric power and NOx allowances markets, Comput. Manag. Sci., 3 (2006), pp. 307-330.
[7] T. Copeland and V. Antikarov, Real Options: A Practitioner's Guide, TEXERE LLC, New York, London, 2001.
[8] M. H. A. Davis and T. Zariphopoulou, American Options and Transaction Fees, IMA Vol. Math. Appl. 65, Springer-Verlag, New York, 1995.
[9] F. Delbaen and W. Schachermayer, The variance optimal martingale measure for continuous processes, Bernoulli, 2 (1996), pp. 81-105.
[10] A. Dixit and R. S. Pindyck, Investment under Uncertainty, Princeton University Press, Princeton, NJ, 1994.
[11] D. Duffie, Dynamic Asset Pricing Theory, Princeton University Press, Princeton, NJ, 1996.
[12] D. Duffie and H. R. Richardson, Mean-variance hedging in continuous time, Ann. Appl. Probab., 1 (1991), pp. 1-15.
[13] H. Follmer and D. Sonderman, Hedging of nonredundant contingent claims, in Contributions to Mathematical Economics, W. Hildenbrand and A. Mas Collel, eds., North-Holland, Amsterdam, 1986, pp. 205-223.
[14] M. Frittelli, The minimal entropy martingale measure and the valuation problem in incomplete markets, Math. Finance, 10 (2000), pp. 39-52.
[15] S. Grenadier, The strategic exercise of options: Development cascades and overbuilding in real estate markets, J. Finance, 51 (1996), pp. 1653-1679.
[16] S. Grenadier, Information revelation through option exercise, Rev. Financial Stud., 12 (1999), pp. 95129.
[17] S. Grenadier, Option exercise games: An application to the equilibrium investment strategies of firms, Rev. Financial Stud., 15 (2002), pp. 691-721.
[18] H. Han and Z. Xia, A Stackelberg model for server-proxies-users systems, J. Appl. Math. Comput., 17 (2005), pp. 185-194.
[19] X. He, A. Prasad, S. Sethi, and G. Gutierrez, A survey of Stackelberg differential game models in supply and marketing channels, J. Syst. Sci. Syst. Eng., 16 (2007), pp. 385-413.
[20] V. Henderson, Valuation of claims on non-traded assets using utility maximization, Math. Finance, 12 (2002), pp. 351-373.
[21] V. Henderson, Explicit solutions to an optimal portfolio choice problem with stochastic income, J. Econom. Dynam. Control, 29 (2005), pp. 1237-1266.
[22] V. Henderson, Valuing the option to invest in an incomplete market, Math. Financ. Econ., 1 (2007), pp. 103-128.
[23] V. Henderson and D. Hobson, Real options with constant relative risk aversion, J. Econom. Dynam. Control, 27 (2002), pp. 329-355.
[24] V. Henderson and D. Hobson, A note on irreversible investment, hedging and optimal consumption problems, Int. J. Theor. Appl. Finance, 9 (2006), pp. 997-1007.
[25] Y. Ho, P. Luh, and R. Muralidharan, Information structure, Stackelberg games, and incentive controllability, IEEE Trans. Automat. Control, 26 (1981), pp. 454-460.
[26] S. Hodges and A. Neuberger, Optimal replication contingent claims under transaction costs, Rev. Futures Markets, 8 (1989), pp. 222-239.
[27] J. Hugonnier and E. Morellec, Corporate control and real investment in incomplete markets, J. Econom. Dynam. Control, 31 (2007), pp. 1781-1800.
[28] K. J. M. Huisman, Technology Investment: A Game Theoretical Real Options Approach, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001.
[29] I. Karatzas and S. Shreve, Brownian Motion and Stochastic Calculus, Springer, New York, 1991.
[30] I. Karatzas and S. Shreve, Methods of Mathematical Finance, Springer, New York, 1998.
[31] D. Kinderlehrer and G. Stampacchia, An Introduction to Variational Inequalities and Their Applications, Academic Press, New York, London, 1980.
[32] B. Lambrecht and W. Perraudin, Real options and preemption under incomplete information, in Proceedings of the 3rd Annual International Conference on Real Options, 1999.
[33] J.-L. Lions and G. Stampacchia, Variational inequalities, Comm. Pure Appl. Math., 20 (1967), pp. 493-519.
[34] R. McDonald and D. Siegel, The value of waiting to invest, Quart. J. Econom., 101 (1986), pp. 708-728.
[35] R. C. Merton, Lifetime portfolio selection under uncertainty: The continuous time case, Rev. Econom. Statist., 51 (1969), pp. 247-257.
[36] R. C. Merton, Optimum consumption and portfolio rules in a continuous time model, J. Econom. Theory, 3 (1971), pp. 373-413.
[37] R. C. Merton, Theory of rational option pricing, Bell J. Econom. Management Sci., 4 (1973), pp. 141-183.
[38] J. Miao and N. Wang, Investment, consumption and hedging under incomplete markets, J. Financial Econom., 86 (2007), pp. 608-642.
[39] M. Musiela and T. Zariphopoulou, Pricing and Risk Management of Derivatives Written on Nontraded Assets, Working paper, 2001.
[40] M. Musiela and T. Zariphopoulou, Valuation in Incomplete Markets and Optimal Investments, preprint, 2002.
[41] M. Musiela and T. Zariphopoulou, An example of indifference prices under exponential preferences, Finance Stoch., 8 (2004), pp. 229-239.
[42] S. C. Myers and S. Majd, Abandonment value and project life, Adv. Futures Options Res., 4 (1990), pp. 1-21.
[43] A. Oberman and T. Zariphopoulou, Pricing early exercise contracts in incomplete markets, Comput. Manag. Sci., 1 (2003), pp. 75-107.
[44] J. Paddock, D. Siegel, and J. Smith, Option valuation of claims on real assets: The case of offshore petroleum leases, Quart. J. Econom., 103 (1988), pp. 479-508.
[45] M. Schweizer, Option hedging for semimartingales, Stochastic Process. Appl., 37 (1991), pp. 339-363.
[46] F. Smets, Essays on Foreign Direct Investment, Unpublished Doctoral Dissertation, Yale University, New Haven, CT, 1993.
[47] H. T. J. Smit and L. A. Ankum, A real options and game-theoretical approach to corporate investment strategy under competition, Financial Management, Autumn (1993), pp. 241-250.
[48] H. T. J. Smit and L. Trigeorgis, Strategy: Options and Games, Princeton University Press, Princeton, NJ, 2004.
[49] T. Zariphopoulou, A solution approach to valuation with unhedgeable risks, Finance Stoch., 5 (2001), pp. 61-82.

# Storage Costs in Commodity Option Pricing* 

Juri Hinz ${ }^{\dagger}$ and Max Fehr ${ }^{\ddagger}$


#### Abstract

Unlike derivatives of financial contracts, commodity options exhibit distinct particularities owing to physical aspects of the underlying. An adaptation of no-arbitrage pricing to this kind of derivative turns out to be a stress test, challenging the martingale-based models with diverse technical and technological constraints, with storability and short selling restrictions, and sometimes with the lack of an efficient dynamic hedging. In this work, we study the effect of storability on risk neutral commodity price modeling and suggest a model class where arbitrage is excluded for both commodity futures trading and simultaneous dynamical management of the commodity stock. The proposed framework is based on key results from interest rate theory.


Key words. commodity options, theory of storage, futures markets, LIBOR model
AMS subject classifications. 91G20, 91G30, 91G70
DOI. 10.1137/090746586

1. Introduction. Prospering economies are highly dependent on commodities. As a consequence, sustainable commodity supply is a key factor for their future growth. Thus, the commodity price risk becomes increasingly important. In the past, the price outbursts for oil, biofuels, and agricultural products have clearly demonstrated that the commodity price modeling and hedging deserve particular attention.

Despite the success of financial mathematics in many fields, we believe that the quantitative understanding of commodity price risk is behind the state of the art and needs further research. In the area of commodities, the models are less sophisticated, flexible, and consistent than, for instance, in the theory of fixed income markets. Not surprisingly, many important questions in commodity risk management cannot be addressed accordingly. For the sake of concreteness, let us consider two commodities, electricity and gold, which are very different in their nature and in their price behavior. Gold is, as a precious metal, a perfectly storable good. Furthermore, gold is considered as an appreciated investment opportunity, particularly during critical times. As a result, the price behavior of gold shows many similarities to a foreign currency. For instance, for gold loans, an interest rate (paid in gold) is available. On the contrary, electricity is not economically storable. Strictly speaking, electrical energy delivered at different points in time must be considered as different commodities. Electricity spot price

[^76]spikes occur regularly; each price jump is followed by a relatively fast price decay, returning back to the normal price level. Such a pattern is not possible for the gold spot price. Consider now a calendar spread call option, which can be viewed as a regular call written on the price spread between commodity futures with different maturities. Such a contract is obviously sensitive to spot price spikes. Evidently, the pricing and hedging of such an instrument must depend on whether it is written on electricity or on gold. However, such a differentiation is hardly possible within common risk neutral commodity price models. At the present level of the theory, the practitioner is essentially left alone with the problem of how to adapt a given commodity price model to account for a perfect storability or for an absolute nonstorability of the underlying good.

Apparently, the lack of storage parameter in the common commodity price models is traced to the very philosophy of commodity risk hedging. Physical commodities are cumbersome: Their storage could be difficult and expensive, the quality may be deteriorated by storage, and the supply may require a costly transportation. Furthermore, short positions in commodities are almost impossible. Contrary to this, futures are clean financial instruments, predestined to hedge against undesirable price changes. Accordingly, futures are frequently considered as prime underlyings. Thus, the generic approaches in commodity modeling attempt to exclude merely the financial arbitrage (achieved by futures and options trading), losing sight of the physical arbitrage, which may result from the trading of financial contracts in addition to an appropriate inventory management. At the present stage, the valuation of commodity options sticks to the calculation of prices which exclude arbitrage within a given futures market, thus neglecting the existing real storage opportunities. According to this, there is a need for a unified model which encompasses all commodities, distinguishing particular cases by their storability degree. Here, we are confronted with complex situations. The variety of storage cost structures ranges from a simple quality deterioration (agricultural products) and dependence on related commodities (fodder price may depend on livestock prices) to the availability of the inventory capacities. Furthermore, economists argue that negative storage costs are useful for describing the benefit or premium associated with holding an underlying product or physical good rather than a financial contract (convenience yield arguments). This benefit may depend on the inventory levels since the marginal yield of the physical stock decreases as the quantity approaches a level larger than the business requires. To complete the perplexity, we should mention that the inventory levels, in turn, are interrelated with commodity spot prices (the inventories are full when the commodity is cheap) and also could exhibit seasonalities (harvest times for agricultural products). The bottom line is that there is no simple approach to facing the entire range of storage particularities. However, we hope that a simplified cost structure is able to capture those storability aspects which are quantitatively essential for derivatives pricing. An empirical study presented in this contribution supports this assumption.

The connection between spot and futures prices for commodities with restricted storability and the valuation of storage opportunities have attracted research interest for a long time. In this work, we emphasize, among others, the works [3], [5], [7], [8], [9], and [10]. Moreover, the comprehensive book [6] presents a state-of-the-art exposition in the commodity derivatives pricing. More specifically, commodity spread options are discussed in [4], in the recent works [1], [2], and in the literature cited therein.

In what follows, we present an approach where a single parameter controls the maximally
possible slope of contango, thus giving a storability constraint. This should yield commodity option prices more realistic than those obtained from traditional models, especially when the instrument under valuation explicitly addresses the storability aspects (like a calendar spread option, a virtual storage, or a swing-type contract).
2. Risk neutral modeling. Common approaches to the valuation of commodity derivatives (see [7]) are based on the assumptions that the commodity trading takes place continuously in time without transaction costs and taxes and that no arbitrage exists for all commodityrelated trading strategies. In the class of spot price models, the evolution $\left(S_{t}\right)_{t \in[0, T]}$ of the commodity spot price is described by a diffusion dynamics

$$
\begin{equation*}
d S_{t}=S_{t}\left(\left(-\mu_{t}\right) d t+\sigma_{t} d W_{t}\right) \tag{2.1}
\end{equation*}
$$

realized on a filtered probability space $\left(\Omega, \mathcal{F},\left(\mathcal{F}_{t}, P\right)_{t \in[0, T]}\right)$ with the prespecified drift $\left(\mu_{t}\right)_{t \in[0, T]}$ and volatility $\left(\sigma_{t}\right)_{t \in[0, T]}$. The process $\left(W_{t}\right)_{t \in[0, T]}$ stands for a Brownian motion under the so-called spot martingale measure $Q$. This measure represents a vehicle for excluding arbitrage opportunities for the trading of commodity-related financial contracts, predominantly of futures. The assumption therefore is that at any time $t \in[0, \tau] \subset[0, T]$ the price $E_{t}(\tau)$ of the futures contract written on the price of a commodity delivered at $\tau$ is given by the $Q$-martingale

$$
\begin{equation*}
E_{t}(\tau)=E^{Q}\left(S_{\tau} \mid \mathcal{F}_{t}\right) \quad \text { for all } t \in[0, \tau] \text { for each futures maturity } \tau \in[0, T] \tag{2.2}
\end{equation*}
$$

whose terminal value equals the spot price $S_{\tau}$. Beyond this property, futures dynamics has to fulfill a series of reasonable assumptions. First of all, the dynamics (2.1), (2.2) has to be consistent with the futures curve $\left(E_{0}^{*}\left(\tau_{i}\right)\right)_{i=0}^{n+1}$ initially observed at the market in the sense that $E\left(S_{\tau_{i}} \mid \mathcal{F}_{0}\right)=E_{0}^{*}\left(\tau_{i}\right)$ for all listed maturity dates $\tau_{0}, \ldots, \tau_{n+1}$. Next, one has to ensure a certain flexibility of the futures curve, at least in terms of the feasibility for changes between backwardation and contango, which evidently occur in commodity markets. Moreover, some authors have argued that the correct choice of the spot price process has to reflect the frequently noticed mean-reverting property. However, we believe that this observation is disputable since there is no obvious reason why a risk neutral dynamics must inherit the statistical properties evident from the perspective of the objective measure. Overall, the correct choice of the commodity price dynamics turns out to be a challenging task, more so because the following storability requirement needs to be considered:

Given a storage cost structure, the dynamics (2.2) should exclude arbitrage opportunities for futures and physical commodity trading.

Our approach aims to give an appropriate implementation of this principle such that particular commodities may be distinguished by a single parameter which stands for their specific storage costs.

In our approach, we utilize a connection between commodity and money market models (see [8]), which needs to be briefly outlined next. Given the dynamics (2.1), the diffusion parameter $\left(\sigma_{t}\right)_{t \in[0, T]}$ obviously reflects the fluctuation of the spot price, whereas the drift term
$\left(\mu_{t}\right)_{t \in[0, T]}$ needs to be adjusted accordingly, in order to match the observed initial futures curve $\left(E_{0}^{*}\left(\tau_{i}\right)\right)_{i=0}^{n+1}$ and to reflect some of its typical changes. It turns out that by an appropriate change of measure these questions can be naturally carried out in the framework of short rate models. Namely, observe that the solution

$$
S_{\tau}=S_{0} e^{-\int_{0}^{\tau} \mu_{s} d s} e^{\int_{0}^{\tau} \sigma_{s} d W_{s}-\frac{1}{2} \int_{0}^{\tau} \sigma_{s}^{2} d s}, \quad \tau \in[0, T],
$$

to (2.1) satisfies

$$
S_{\tau}=S_{t} e^{-\int_{t}^{\tau} \mu_{s} d_{s}} \Lambda_{\tau} \Lambda_{t}^{-1}, \quad 0 \leq t \leq \tau \leq T,
$$

where, under appropriate assumptions on $\left(\sigma_{t}\right)_{t \in[0, T]}$, the martingale

$$
\Lambda_{\tau}=e^{\int_{0}^{\tau} \sigma_{s} d W_{s}-\frac{1}{2} \int_{0}^{\tau} \sigma_{s}^{2} d s}, \quad \tau \in[0, T],
$$

provides a measure change to a probability measure $\tilde{Q}$ which is equivalent to $Q$ and is given by

$$
d \tilde{Q}=\Lambda_{T} d Q
$$

Using the measure $\tilde{Q}$, we obtain

$$
E_{t}(\tau)=\mathbb{E}_{t}^{Q}\left(S_{\tau}\right)=\mathbb{E}_{t}^{Q}\left(S_{t} e^{-\int_{t}^{\tau} \mu_{s} d s} \Lambda_{\tau} \Lambda_{t}^{-1}\right)=S_{t} \mathbb{E}_{t}^{\tilde{Q}}\left(e^{-\int_{t}^{\tau} \mu_{s} d s}\right)
$$

with the proportion between the futures price and the spot price

$$
E_{t}(\tau) / S_{t}=\mathbb{E}_{t}^{\tilde{Q}}\left(e^{-\int_{t}^{\tau} \mu_{s} d s}\right)=: B_{t}(\tau), \quad t \leq \tau
$$

Obviously, all desired properties of the futures curve evolution can be addressed in terms of the dynamics of $\left(B_{t}(\tau)\right)_{t \in[0, \tau]}, \tau \in[0, T]$. This observation shows that by modeling $\left(\mu_{t}\right)_{t \in[0, T]}$ as a short rate of an appropriate interest rate model (with respect to $\tilde{Q}$ ) one obtains a commodity price model which inherits futures curve properties from the zero bond curve of the underlying interest rate model. More generally, [8] argues that any commodity futures price model can be constructed as

$$
E_{t}(\tau)=S_{t} B_{t}(\tau), \quad 0 \leq t \leq \tau \leq T,
$$

by a separate realization of spot price $\left(S_{t}\right)_{t \in[0, T]}$ and an appropriate zero bond $\left(B_{t}(\tau)\right)_{0<t \leq \tau \leq T}$ dynamics. Although such a rigid framework is not ideal for addressing storage cost issues, we utilize an analogy between commodity and money markets and borrow ideas from LIBOR markets to introduce storage cost restrictions into commodity modeling.
3. A risk neutral approach to storage costs. Let us agree that a commodity market on the time horizon $[0, T]$ is modeled by adapted processes

$$
\begin{equation*}
\left(S_{t}\right)_{t \in[0, T]}, \quad\left(E_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}, \quad i=0, \ldots, n+1, \tag{3.1}
\end{equation*}
$$

realized on $\left(\Omega, \mathcal{F}, P,\left(\mathcal{F}_{t}\right)_{\in[0, T]}\right)$ with the interpretation that $\left(S_{t}\right)_{t \in[0, T]}$ is the spot price process and $\left(E_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ denotes the price evolution of the futures maturing at $\tau_{i} \in\left\{\tau_{0}, \ldots, \tau_{n+1}\right\} \subset$ $[0, T]$. For simplicity, we assume that $\tau_{0}=0, \tau_{n+1}=T$ and that all maturity times differ by a fixed tenor $\Delta=\tau_{i+1}-\tau_{i}$ for all $i=0, \ldots, n$. We shall agree on the following.

Definition 3.1. The price processes (3.1) define a commodity market (which excludes arbitrage for futures trading and simultaneous commodity stock management) if the following conditions are satisfied:
(C0) $\left(S_{t}\right)_{t \in[0, T]},\left(E_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ for $i=0, \ldots, n+1$ are positive-valued processes.
(C1) There is no financial arbitrage in the sense that there exists a measure $Q^{E}$ which is equivalent to $P$ and such that, for each $i=0, \ldots, n+1$, the process $\left(E_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ follows a martingale with respect to $Q^{E}$.
(C2) The initial values of the futures price processes $\left(E_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}, i=0, \ldots, n+1$, fit the observed futures curve $\left.\left(E_{0}^{*}\left(\tau_{i}\right)\right)_{i=0}^{n+1} \in\right] 0, \infty{ }^{n+1}$; i.e., it holds almost surely that $E_{0}\left(\tau_{i}\right)=E_{0}^{*}\left(\tau_{i}\right)$ for all $i=0, \ldots, n+1$.
(C3) The terminal futures price matches the spot price: $E_{\tau_{i}}\left(\tau_{i}\right)=S_{\tau_{i}}$ for $i=0, \ldots, n+1$.
(C4) There exists $\kappa>0$ such that

$$
\begin{equation*}
E_{t}\left(\tau_{i+1}\right)-\kappa \leq E_{t}\left(\tau_{i}\right) \quad \text { for all } t \in\left[0, \tau_{i}\right], i=1, \ldots, n \tag{3.2}
\end{equation*}
$$

Let us explain why assumption ( C 4 ) is a convenient description of storage cost. For any time $t \leq \tau_{i}$, consider the commodity forward prices $F_{t}\left(\tau_{i}\right), F_{t}\left(\tau_{i+1}\right)$ and the prices $p_{t}\left(\tau_{i}\right)$, $p_{t}\left(\tau_{i+1}\right)$ of zero bonds (whose face value is normalized to one) maturing at $\tau_{i}$ and $\tau_{i+1}$, respectively. To exclude arbitrage from physical storage facilities, we derive a relation between these prices and the price $k_{t}\left(\tau_{i+1}\right)$ of a contract which serves as a storage facility for one commodity unit within $\left[\tau_{i}, \tau_{i+1}\right]$. Thereby, we assume that the price $k_{t}\left(\tau_{i+1}\right)$ is agreed at time $t$ and is paid at $\tau_{i+1}$. It turns out that to exclude the cash and carry arbitrage the prices must satisfy

$$
\begin{equation*}
F_{t}\left(\tau_{i+1}\right)-k_{t}\left(\tau_{i}\right)-\left(\frac{p_{t}\left(\tau_{i}\right)}{p_{t}\left(\tau_{i+1}\right)}-1\right) F_{t}\left(\tau_{i}\right) \leq F_{t}\left(\tau_{i}\right), \quad t \in\left[0, \tau_{i}\right], \quad i=1, \ldots, n+1 \tag{3.3}
\end{equation*}
$$

This relation follows from the no-arbitrage assumption by examining a strategy, which fixes the prices at time $t$, buys at time $\tau_{i}>t$ one commodity unit, stores it within $\left[\tau_{i}, \tau_{i+1}\right]$, and sells it at $\tau_{i+1}$. Let us investigate in more detail the revenue from such a strategy.

| Time | $\tau_{i}$-future | $\tau_{i+1}$-future | Storage | $\tau_{i}$-bond | $\tau_{i+1}$-bond |
| :---: | :---: | :---: | :---: | :---: | :---: |
| $t$ | 1 long | 1 short | 1 long | $F_{t}\left(\tau_{i}\right)$ long | $\frac{p_{t}\left(\tau_{i}\right)}{p_{t}\left(\tau_{i+1}\right)} F_{t}\left(\tau_{i}\right)$ short |
| $\tau_{i}$ | supply | 1 short | store | cash flow $F_{t}\left(\tau_{i}\right)$ | $\frac{p_{t}\left(\tau_{i}\right)}{p_{t}\left(\tau_{i+1}\right)} F_{t}\left(\tau_{i}\right)$ short |
| $\tau_{i+1}$ | expired | delivery | pay $k_{t}\left(\tau_{i+1}\right)$ | expired | cash flow $-\frac{p_{t}\left(\tau_{i}\right)}{p_{t}\left(\tau_{i+1}\right)} F_{t}\left(\tau_{i}\right)$ |

Obviously, our agent starts with no initial capital since entering forward positions at time $t$ does not require any cash flow and both bond positions are balanced. Furthermore, the strategy is self-financed. Namely, the capital required to buy one commodity unit at time $\tau_{i}$ is financed by a cash flow from the expiring long bond position. At the end of this strategy, the agent requires a capital $k_{t}\left(\tau_{i+1}\right)$ to pay for the storage and $\frac{p_{t}\left(\tau_{i}\right)}{p_{t}\left(\tau_{i+1}\right)} F_{t}\left(\tau_{i}\right)$ to close the short bond position. However, our agent earns a revenue $F_{t}\left(\tau_{i+1}\right)$ from the delivery of the stored commodity unit. Hence, the terminal capital is known with certainty in advance, at the initial time $t$, and is equal to

$$
F_{t}\left(\tau_{i+1}\right)-k_{t}\left(\tau_{i+1}\right)-\frac{p_{t}\left(\tau_{i}\right)}{p_{t}\left(\tau_{i+1}\right)} F_{t}\left(\tau_{i}\right)
$$

In order to exclude arbitrage, we have to suppose that this terminal wealth cannot be positive. Thus, we obtain (3.3). Now, let us elaborate on the approximation (3.2) of (3.3). Consider the cumulative effect

$$
\begin{equation*}
k_{t}\left(\tau_{i}\right)+\left(\frac{p_{t}\left(\tau_{i}\right)}{p_{t}\left(\tau_{i+1}\right)}-1\right) F_{t}\left(\tau_{i}\right) \quad \text { for all } t \in\left[0, \tau_{i}\right] \text { and } i=1, \ldots, n \tag{3.4}
\end{equation*}
$$

of the storage costs and interest rates. If we propose a model which satisfies

$$
\begin{equation*}
F_{t}\left(\tau_{i+1}\right)-\kappa \leq F_{t}\left(\tau_{i}\right) \quad \text { for all } t \in\left[0, \tau_{i}\right] \text { and } i=1, \ldots, n, \tag{3.5}
\end{equation*}
$$

then the arbitrage by cash and carry is excluded at least in those situations where (3.4) is bounded from below by the parameter $\kappa$. In this context, the accuracy of the estimation

$$
\begin{equation*}
k_{t}\left(\tau_{i}\right)+\left(\frac{p_{t}\left(\tau_{i}\right)}{p_{t}\left(\tau_{i+1}\right)}-1\right) F_{t}\left(\tau_{i}\right) \geq \kappa \quad \text { for all } t \in\left[0, \tau_{i}\right] \text { and } i=1, \ldots, n \tag{3.6}
\end{equation*}
$$

is critical. Whether such an estimate is possible in practice and whether it yields models which capture storability aspects of the commodity price evolution must be explicitly verified in any particular situation. In any case, we believe that for certain commodities the left-hand side of (3.6) can be reasonably approximated by a constant and deterministic parameter $\kappa$, which justifies the assumption (3.5). Finally, we pass from (3.5) to (3.2) by the approximation of the forward prices $F_{t}\left(\tau_{i}\right), F_{t}\left(\tau_{i+1}\right)$ by futures prices $E_{t}\left(\tau_{i}\right), E_{t}\left(\tau_{i+1}\right)$.
4. Storage costs as contango limit. In principle, $\kappa$ can be estimated from the actual physical storage costs and the interest rate effects. Consequently, for certain commodities there exists a maximally possible steepness of the futures curve in contango situations, which is known among traders as the contango limit. That is, such a rough estimate of $\kappa$ could also be obtained from historical data by inspecting the maximal increase of the historical futures curves

$$
\kappa=\max \left\{E_{t}\left(\tau_{i+1}\right)(\omega)-E_{t}\left(\tau_{i}\right)(\omega): \text { for all } t \leq \tau_{i}<\tau_{i+1}\right\}
$$

based on a representative data record. Let us illustrate this method.
Consider the history of soybean trading at the Chicago Board of Trade (CBOT). At this exchange, the soybean futures expire in January, March, May, July, August, September, and November. Each contract is listed one year prior to expiry. Let us suppose a fixed tenor $\Delta$ of two months. Thus, all prices of futures maturing in August are not considered. Within each period $\left[\tau_{i-1}, \tau_{i}\right]$, six futures prices with delivery dates $\tau_{i}, \tau_{i+1}, \ldots, \tau_{i+5}$ are available. Figure 1 shows a typical price evolution of six futures and the period where all six contracts are listed. Moreover, Figure 2 illustrates the entire data set we use in this study. It encompasses the end of the day futures prices ranging from 2000-10-02 to 2007-02-23. Figure 3 shows the behavior of the difference of consecutive contracts in the entire data record. Note that this picture clearly supports our viewpoint since there is a clear contango limit, represented by a price which has never been hit by the difference $E_{t}\left(\tau_{i+1}\right)-E_{t}\left(\tau_{i}\right)$. At the same time, there is no limitation on the backwardation side since the differences $E_{t}\left(\tau_{i+1}\right)-E_{t}\left(\tau_{i}\right)$ tilt seemingly arbitrarily far downwards. To estimate the storage cost parameter, we use the historical data depicted in Figure 2 and calculate $\kappa$ by

$$
\begin{equation*}
\max \left\{\left(E_{t}\left(\tau_{i+1}\right)-E_{t}\left(\tau_{i}\right)\right)(\omega): t, \tau_{i}, \tau_{i+1} \text { where price observations are available }\right\} \tag{4.1}
\end{equation*}
$$



Figure 1. The price evolution of six consecutive futures contracts. Vertical lines separate a two month period where all six contracts are traded.


Figure 2. Soybean closing daily prices from CBOT in cents per bushel.
giving 24 US cents per bushel for two months. Thus, setting $\kappa \approx 26$ could give a reasonable futures price model which excludes cash and carry arbitrage for soybeans. Still, there is no guarantee why the difference $E_{t}\left(\tau_{i+1}\right)-E_{t}\left(\tau_{i}\right)$ in a future trajectory does not exceed 26. As discussed before, a reliable estimation of $\kappa$ should be based on a study of storage costs and on bond prices. However, we believe that (4.1) may serve as a reasonable approximation.

Not surprisingly, similar analysis on other commodities shows that a clear historical contango limit can also be observed for other storable agricultural products and for precious metals but not for assets with limited storability (oil, gas, and electricity) and for perishable goods (like livestock).


Figure 3. The difference $E_{t}\left(\tau_{i+1}\right)-E_{t}\left(\tau_{i}\right)$ shows an upper bound at 24 US cents per bushel for two months.
5. Modeling commodity dynamics. This section is devoted to the construction of commodity markets which satisfy the axioms formulated in Definition 3.1. Here the main task is to establish a dynamics for martingales $\left(E_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}(i=1, \ldots, n+1)$ which obeys the storage restriction ( C 4 ) and, at the same time, possesses a certain flexibility in the movements of the futures curve. Fortunately, similar problems occurred in the theory of fixed income markets and have been treated successfully. As a paradigm, we use the forward LIBOR market model, also known as the BGM approach, named after A. Brace, D. Gatarek, and M. Musiela. In their context, the dynamics of zero bonds $\left(p_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ with the fixed tenor $\Delta=\tau_{i+1}-\tau_{i}$, $i=1, \ldots, n$, is described in terms of the so-called simple rates $\left(L_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ defined by

$$
\begin{equation*}
p_{t}\left(\tau_{i+1}\right)=\frac{p_{t}\left(\tau_{i}\right)}{1+\Delta L_{t}\left(\tau_{i}\right)} \quad \text { for } i=1, \ldots, n, t \in\left[0, \tau_{i}\right] \tag{5.1}
\end{equation*}
$$

whose dynamics is modeled by stochastic differential equations

$$
\begin{equation*}
d L_{t}\left(\tau_{i}\right)=L_{t}\left(\tau_{i}\right)\left(\beta_{t}\left(\tau_{i}\right) d t+\gamma_{t}\left(\tau_{i}\right) d W_{t}\right), \quad i=1, \ldots, n \tag{5.2}
\end{equation*}
$$

where the deterministic volatilities $\left(\gamma_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ are freely chosen for $i=1, \ldots, n$, whereas the drifts $\left(\beta_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ for $i=1, \ldots, n$ are determined by this choice. The importance of the BGM formulation is that each simple rate $\left(L_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ follows a geometric Brownian motion with respect to the forward measure corresponding to the numeraire $\left(p_{t}\left(\tau_{i+1}\right)\right)_{t \in\left[0, \tau_{i}\right]}$. This fact yields explicit formulae for Caplets and therefore provides an important tool, implicit calibration, for this fixed-income model class. We suggest transferring this successful concept to commodity markets by proposing a similar framework, where futures are consecutively interrelated by stochastic exponentials, similarly to simple rates in (5.1). As in the BGM setting, it turns out that this concept provides appropriate tools for the implicit model calibration. The idea is to link the dynamics $\left(E_{t}\left(\tau_{i}\right), E_{t}\left(\tau_{i+1}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ by

$$
\begin{equation*}
E_{t}\left(\tau_{i+1}\right)=\frac{E_{t}\left(\tau_{i}\right)+\kappa}{1+Z_{t}\left(\tau_{i}\right)}, \quad t \in\left[0, \tau_{i}\right], \quad i=1, \ldots, n \tag{5.3}
\end{equation*}
$$

where, likewise to the simple rate (5.2), the simple ratio $\left(Z_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ follows a diffusion

$$
\begin{equation*}
d Z_{t}\left(\tau_{i}\right)=Z_{t}\left(\tau_{i}\right)\left(\alpha_{t}\left(\tau_{i}\right) d t+\sigma_{t}\left(\tau_{i}\right) d W_{t}\right), \quad t \in\left[0, \tau_{i}\right], \quad i=1, \ldots, n, \tag{5.4}
\end{equation*}
$$

where $\left(\sigma_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ and $\left(\alpha_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ denote the volatilities and the drifts, respectively. We will see later that the drifts follow from the choice of simple ratio volatilities and other model ingredients. Before entering the details of the construction, let us emphasize that (5.3) indeed ensures (3.2) by the nonnegativity of $\left(Z_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$, which is a consequence of (5.4), under appropriate conditions.

Now, let us construct a model which fulfills the axioms from Definition 3.1. We begin with a complete filtered probability space $\left(\Omega, \mathcal{F}, Q^{E},\left(\mathcal{F}_{t}\right)_{t \in[0, T]}\right)$ where the filtration is the augmentation (by the null sets in $\mathcal{F}_{T}^{W}$ ) of the filtration $\left(\mathcal{F}_{t}^{W}\right)_{t \in[0, T]}$ generated by the $d$-dimensional Brownian motion $\left(W_{t}\right)_{t \in[0, T]}$. All processes are supposed to be progressively measurable. For equidistant futures maturity dates

$$
0=\tau_{0}<\tau_{1}<\cdots<\tau_{n+1}=T \in[0, T], \quad \Delta=\tau_{i+1}-\tau_{i}, \quad i=0, \ldots, n,
$$

and the initial futures curve $\left.\left(E_{0}^{*}\left(\tau_{i}\right)\right)_{i=1}^{n+1} \in\right] 0, \infty\left[^{n+1}\right.$, we construct a commodity market where futures prices follow

$$
\begin{equation*}
d E_{t}\left(\tau_{i}\right)=E_{t}\left(\tau_{i}\right) \Sigma_{t}\left(\tau_{i}\right) d W_{t}, \quad t \in\left[0, \tau_{i}\right], \quad E_{0}\left(\tau_{i}\right)=E_{0}^{*}\left(\tau_{i}\right), \quad i=1, \ldots, n+1 \tag{5.5}
\end{equation*}
$$

and obey (C0)-(C4) with a given storage cost parameter $\kappa>0$. In a separate section, we discuss how the volatility term structure

$$
\left(\Sigma_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}, \quad i=1, \ldots, n+1,
$$

and its dimension $d \in \mathbb{N}$ are determined from a model calibration procedure.
First, we outline the intuition behind our construction. Given the local $Q^{E}$-martingale $\left(E_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ as in (5.5), the Itô formula shows how, given $\left(\sigma_{t}\left(\tau_{i}\right)\right)_{t \in[0, \tau]}$, to settle the drift $\left(\alpha_{t}\left(\tau_{i}\right)\right)_{t \in[0, \tau]}$ in (5.4) such that (5.3) becomes a martingale. With this principle, we construct $\left(E_{t}\left(\tau_{i+1}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ from given $\left(E_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ and $\left(\sigma_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$. To proceed, we need to extend this price process by $\left(E_{t}\left(\tau_{i+1}\right)\right)_{t \in\left[\tau_{i}, \tau_{i+1}\right]}$ to the expiry date. This is effected by

$$
d E_{t}\left(\tau_{i+1}\right)=E_{t}\left(\tau_{i+1}\right) \Sigma_{t}\left(\tau_{i+1}\right) d W_{t}, \quad t \in\left[\tau_{i}, \tau_{i+1}\right],
$$

where the volatility in front of delivery $\left(\Sigma_{t}\left(\tau_{i+1}\right)\right)_{t \in\left[\tau_{i}, \tau_{i+1}\right]}$ is exogenously given by

$$
\begin{equation*}
\Sigma_{t}\left(\tau_{i+1}\right):=\psi_{t} \quad \text { for all } t \in\left[\tau_{i}, \tau_{i+1}\right], i=0, \ldots, n, \tag{5.6}
\end{equation*}
$$

with a prespecified process $\left(\psi_{t}\right)_{t \in[0, T]}$. Having thus established $\left(E_{t}\left(\tau_{i+1}\right)\right)_{t \in\left[0, \tau_{i+1}\right]}$, the same procedure is applied iteratively to determine all remaining futures $\left(E_{t}\left(\tau_{j+1}\right)\right)_{t \in\left[0, \tau_{j+1}\right]}$ with $j=i+1, \ldots, n$.

In the following lemma, we call a $d$-dimensional process $\left(X_{t}\right)_{t \in[0, \tau]}$ bounded if $\left\|X_{t}\right\|<C$ holds for all $t \in[0, \tau]$ almost surely, for some $C \in[0, \infty[$.

Lemma 5.1. Let $\left(E_{t}\right)_{t \in[0, \tau]}$ be a positive-valued martingale following $d E_{t}=E_{t} \Sigma_{t} d W_{t}$ with a bounded volatility process $\left(\Sigma_{t}\right)_{t \in[0, \tau]}$. If $\left(\sigma_{t}\right)_{t \in[0, \tau]}$ is bounded, then there exists a unique strong solution to

$$
\begin{equation*}
\frac{d Z_{t}}{Z_{t}}=-\sigma_{t}\left(\left(\frac{E_{t} \Sigma_{t}}{E_{t}+\kappa}-\frac{Z_{t} \sigma_{t}}{Z_{t}+1}\right) d t-d W_{t}\right), \quad Z_{0}:=Z_{0}^{*}>0 . \tag{5.7}
\end{equation*}
$$

Moreover,

$$
\begin{equation*}
E_{t}^{\prime}=\frac{E_{t}+\kappa}{1+Z_{t}}, \quad t \in[0, \tau], \tag{5.8}
\end{equation*}
$$

follows the martingale dynamics

$$
\begin{equation*}
d E_{t}^{\prime}=E_{t}^{\prime} \Sigma_{t}^{\prime} d W_{t}, \quad t \in[0, \tau], \tag{5.9}
\end{equation*}
$$

with bounded

$$
\begin{equation*}
\Sigma_{t}^{\prime}=\frac{E_{t} \Sigma_{t}}{E_{t}+\kappa}-\frac{Z_{t} \sigma_{t}}{Z_{t}+1}, \quad t \in[0, \tau] . \tag{5.10}
\end{equation*}
$$

Proof. Write (5.7) as

$$
\begin{equation*}
d Z_{t}=F(Z)_{t} d t+Z_{t} \sigma_{t} d W_{t}, \quad Z_{0}=Z_{0}^{*}>0, \tag{5.11}
\end{equation*}
$$

where the functional $F$ acts on the processes $Z=\left(Z_{t}\right)_{t \in[0, \tau]}$ by

$$
\begin{equation*}
F(Z)_{t}=-Z_{t} \sigma_{t}\left(\frac{E_{t} \Sigma_{t}}{E_{t}+\kappa}-\frac{Z_{t} \sigma_{t}}{Z_{t}+1}\right), \quad t \in[0, \tau] . \tag{5.12}
\end{equation*}
$$

To avoid technical difficulties in (5.7) occurring when the denominator $Z_{t}+1$ vanishes, we first discuss a stochastic differential equation similar to (5.11)

$$
\begin{equation*}
d Z_{t}=\tilde{F}(Z)_{t} d t+Z_{t} \sigma_{t} d W_{t}, \quad Z_{0}=Z_{0}^{*}>0 \tag{5.13}
\end{equation*}
$$

where the functional $\tilde{F}$ acts by $\tilde{F}(Z)_{t}:=F(Z)_{t} 1_{\left\{Z_{t} \geq 0\right\}}$ for $t \in[0, \tau]$ on each process $Z=$ $\left(Z_{t}\right)_{t \in[0, \tau]}$. Since $\sup _{t \in[0, \tau]}\left\|\sigma_{t}\right\| \leq C \in[0, \infty[$ by assumption, the diffusion term in (5.13) is Lipschitz continuous:

$$
\left\|Z_{t} \sigma_{t}-Z_{t}^{\prime} \sigma_{t}\right\| \leq C\left|Z_{t}-Z_{t}^{\prime}\right| \quad \text { for all } t \in[0, \tau]
$$

Thus, to ensure the existence and uniqueness of the strong solution to (5.13) it suffices to verify the Lipschitz continuity of $\tilde{F}$ in the sense that there exists $\tilde{C} \in[0, \infty[$ such that

$$
\begin{equation*}
\left\|\tilde{F}(Z)_{t}-\tilde{F}\left(Z^{\prime}\right)_{t}\right\| \leq \tilde{C}\left|Z_{t}-Z_{t}^{\prime}\right| \quad \text { for all } t \in[0, \tau] \tag{5.14}
\end{equation*}
$$

The decomposition $\tilde{F}(Z)_{t}=\tilde{f}_{1}\left(t, Z_{t}\right)+\tilde{f}_{2}\left(t, Z_{t}\right)+\tilde{f}_{3}\left(t, Z_{t}\right)$ with

$$
\begin{aligned}
& \tilde{f}_{1}(t, z)=-1_{\{z \geq 0\}} z \frac{E_{t}}{E_{t}+\kappa} \sigma_{t} \Sigma_{t} \\
& \tilde{f}_{2}(t, z)=-1_{\{z \geq 0\}} z \frac{1}{z+1} \sigma_{t} \sigma_{t} \\
& \tilde{f}_{3}(t, z)=1_{\{z \geq 0\}} z \sigma_{t} \sigma_{t}
\end{aligned}
$$

for all $t \in[0, \tau], z \geq 0$ shows that $\tilde{C} \geq \sup _{t \in[0, \tau]}\left(\left|\sigma_{t} \Sigma_{t}\right|+2\left|\sigma_{t} \sigma_{t}\right|\right)$ yields a Lipschitz constant in (5.14); here $\tilde{C} \in\left[0, \infty\right.$ [ holds since both $\left(\Sigma_{t}\right)_{t \in[0, \tau]}$ and $\left(\sigma_{t}\right)_{t \in[0, \tau]}$ are bounded processes by assumption.

Let $\left(Z_{t}\right)_{t \in[0, \tau]}$ be the unique strong solution to (5.13). In order to show that this process also solves (5.7), it suffices to verify that $\left.Z_{t} \in\right] 0, \infty[$ holds almost surely for all $t \in[0, \tau]$. Indeed, the positivity follows from the stochastic exponential form

$$
d Z_{t}=Z_{t}\left(\frac{\tilde{F}(Z)_{t}}{Z_{t}} d t+\sigma_{t} d W_{t}\right), \quad Z_{0}=Z_{0}^{*}>0
$$

with bounded drift coefficient

$$
\begin{equation*}
\frac{\tilde{F}(Z)_{t}}{Z_{t}}=-\sigma_{t}\left(\frac{E_{t} \Sigma_{t}}{E_{t}+\kappa}-\frac{Z_{t} \sigma_{t}}{Z_{t}+1}\right) 1_{\left\{Z_{t} \geq 0\right\}}, \quad t \in[0, \tau] . \tag{5.15}
\end{equation*}
$$

To show the uniqueness, we argue that any solution $\left(Z_{t}^{\prime}\right)_{t \in[0, \tau]}$ to (5.11) coincides with $\left(Z_{t}\right)_{t \in[0, \tau]}$ on the stochastic interval prior the first entrance time of $\left(Z_{t}^{\prime}\right)_{t \in[0, \tau]}$ into $\left.]-\infty, 0\right]$ since on this interval $\left(Z_{t}^{\prime}\right)_{t \in[0, \tau]}$ solves (5.13). Furthermore, $\left(Z_{t}^{\prime}\right)_{t \in[0, \tau]}$ is a continuous process, being a strong solution to (5.11) by assumption. The continuity of $\left(Z_{t}^{\prime}\right)_{t \in[0, \tau]}$ shows that $\left(Z_{t}^{\prime}\right)_{t \in[0, \tau]}$ matches $\left(Z_{t}\right)_{t \in[0, \tau]}$ on the entire interval $[0, \tau]$. Finally, (5.9) is verified by a straightforward application of the Itô formula.

Using the common stopping technique, Lemma 5.1 extends in a straightforward way from bounded to continuous processes.

Proposition 5.2. Let $\left(E_{t}\right)_{t \in[0, \tau]}$ be a positive-valued martingale following $d E_{t}=E_{t} \Sigma_{t} d W_{t}$ with a continuous volatility process $\left(\Sigma_{t}\right)_{t \in[0, \tau]}$. If $\left(\sigma_{t}\right)_{t \in[0, \tau]}$ is continuous, then there exists a unique strong solution to

$$
\begin{equation*}
\frac{d Z_{t}}{Z_{t}}=-\sigma_{t}\left(\left(\frac{E_{t} \Sigma_{t}}{E_{t}+\kappa}-\frac{Z_{t} \sigma_{t}}{Z_{t}+1}\right) d t-d W_{t}\right), \quad Z_{0}:=Z_{0}^{*}>0 . \tag{5.16}
\end{equation*}
$$

Moreover,

$$
\begin{equation*}
E_{t}^{\prime}=\frac{E_{t}+\kappa}{1+Z_{t}}, \quad t \in[0, \tau], \tag{5.17}
\end{equation*}
$$

follows the martingale dynamics

$$
\begin{equation*}
d E_{t}^{\prime}=E_{t}^{\prime} \Sigma_{t}^{\prime} d W_{t}, \quad t \in[0, \tau], \tag{5.18}
\end{equation*}
$$

with continuous

$$
\begin{equation*}
\Sigma_{t}^{\prime}=\frac{E_{t} \Sigma_{t}}{E_{t}+\kappa}-\frac{Z_{t} \sigma_{t}}{Z_{t}+1}, \quad t \in[0, \tau] \tag{5.19}
\end{equation*}
$$

Proof. Introduce a sequence of stopping times

$$
\vartheta_{k}=\inf \left\{t \in[0, \tau]: \max \left(\left\|\sigma_{t}\right\|,\left\|\Sigma_{t}\right\|\right) \geq k\right\}, \quad k \in \mathbb{N}
$$

Since both processes $\left(\sigma_{t}\right)_{t \in[0, \tau]}$ and $\left(\Sigma_{t}\right)_{t \in[0, \tau]}$ are continuous, we have $\lim _{k \rightarrow \infty} \vartheta_{k}=\tau$; hence the monotonically increasing sequence of stochastic intervals $\left[0, \vartheta_{k}\right], k \in \mathbb{N}$, covers the entire time horizon:

$$
\begin{equation*}
\bigcup_{k \in \mathbb{N}}\left[0, \vartheta_{k}\right]=\Omega \times[0, \tau] \tag{5.20}
\end{equation*}
$$

Now, by stopping $\left(\sigma_{t}\right)_{t \in[0, \tau]}$ and $\left(\Sigma_{t}\right)_{t \in[0, \tau]}$ at the time $\vartheta_{k}$, one obtains bounded processes

$$
\left(\sigma_{t}^{(k)}:=\sigma_{t \wedge \tau_{k}}\right)_{t \in[0, \tau]}, \quad\left(\Sigma_{t}^{(k)}:=\Sigma_{t \wedge \tau_{k}}\right)_{t \in[0, \tau]} .
$$

Further, define $\left(E_{t}^{(k)}\right)_{[0, \tau]}$ as the solution to

$$
d E_{t}^{(k)}=E_{t}^{(k)} \Sigma_{t}^{(k)} d W_{t}, \quad E_{0}^{(k)}=E_{0}
$$

If follows that for each $k \in \mathbb{N}$ the processes $\left(E_{t}^{(k)}\right)_{t \in[0, \tau]},\left(\Sigma_{t}^{(k)}\right)_{t \in[0, \tau]}$, and $\left(\sigma_{t}^{(k)}\right)_{t \in[0, \tau]}$ satisfy the assumptions of Lemma 5.1, which yields the corresponding processes $\left(Z_{t}^{(k)}\right)_{t \in[0, \tau]},\left(E_{t}^{\prime(k)}\right)_{t \in[0, \tau]}$, and $\left(\Sigma_{t}^{\prime(k)}\right)_{t \in[0, \tau]}$. By construction, the next set of processes coincides with the previous one on the common stochastic interval

$$
\left.\begin{array}{rl}
Z_{t}^{(k)}(\omega) & =Z_{t}^{(k+1)}(\omega) \\
\Sigma_{t}^{\prime(k)}(\omega) & =\Sigma_{t}^{\prime(k+1)}(\omega) \\
E_{t}^{\prime(k)}(\omega) & =E_{t}^{\prime(k+1)}(\omega)
\end{array}\right\} \text { for all } t \in\left[0, \vartheta_{k}(\omega)\right], \omega \in \Omega, k \in \mathbb{N}
$$

Thus, their limits

$$
Z_{t}:=\lim _{k \rightarrow \infty} Z_{t}^{(k)}, \quad \Sigma_{t}^{\prime}:=\lim _{k \rightarrow \infty} \Sigma_{t}^{\prime(k)}, \quad E_{t}^{\prime}:=\lim _{k \rightarrow \infty} E_{t}^{\prime(k)}, \quad t \in[0, \tau],
$$

are well defined on $[0, \tau]$ because of (5.20) and satisfy the assertions (5.16)-(5.19).
Remark. For later use, let us point out that the volatility process $\left(\Sigma_{t}^{\prime}\right)_{t \in[0, \tau]}$ resulting from (5.19) can be written as a function

$$
\begin{equation*}
\Sigma_{t}^{\prime}=s\left(E_{t}^{\prime}, E_{t}, \Sigma_{t}, \sigma_{t}\right):=\frac{E_{t} \Sigma_{t}+E_{t}^{\prime} \sigma_{t}}{E_{t}+\kappa}-\sigma_{t}, \quad t \in[0, \tau] . \tag{5.21}
\end{equation*}
$$

The representation (5.21) follows directly from (5.19) by using

$$
\frac{Z_{t}}{Z_{t}+1}=1-\frac{1}{1+Z_{t}}=1-\frac{E_{t}^{\prime}}{E_{t}+\kappa}, \quad t \in[0, \tau]
$$

where the last equality is a consequence of (5.17). Let us point out that in the limiting case, where the contango limit is almost reached, i.e., $E_{t}^{\prime} \approx E_{t}+\kappa$, the volatility $\Sigma_{t}^{\prime}$ can be approximated as

$$
\Sigma_{t}^{\prime} \approx \frac{E_{t} \Sigma_{t}}{E_{t}+\kappa} \approx \frac{E_{t} \Sigma_{t}}{E_{t}^{\prime}}
$$

thus the dynamics $\left(E_{t}^{\prime}\right)_{t \in[0, T]}$ follows that of $\left(E_{t}\right)_{t \in[0, T]}$, because of

$$
d E_{t}^{\prime}=E_{t}^{\prime} \Sigma_{t}^{\prime} d W_{t} \approx E_{t}^{\prime} \frac{E_{t} \Sigma_{t}}{E_{t}^{\prime}} d W_{t}=E_{t} \Sigma_{t} d W_{t}=d E_{t} .
$$

This observation shows that the restriction $E_{t}+\kappa \geq E_{t}^{\prime}$ necessarily causes strong correlations of the process increments if the prices $E_{t}$ and $E_{t}^{\prime}$ come close to the contango limit. On this account the sensitivity of the model to the choice of storage cost parameter $\kappa$ can be significant. Now consider the entire construction of futures prices. Starting with

$$
\left\{\begin{array}{l}
\left.E_{0}^{*}\left(\tau_{i}\right) \in\right] 0, \infty[\text { for } i=1, \ldots, n+1, \text { initial futures curve, }  \tag{5.22}\\
\left(\psi_{t}\right)_{t \in[0, T]}, \text { in-front-of-maturity futures volatility (continuous), } \\
\left(\sigma_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}, i=1, \ldots, n, \text { simple ratio volatilities (continuous), }
\end{array}\right.
$$

we apply the following procedure.
Initialization. Start with $E_{0}\left(\tau_{1}\right)=E_{0}^{*}\left(\tau_{1}\right), \ldots, E_{0}\left(\tau_{n+1}\right)=E_{0}^{*}\left(\tau_{n+1}\right)$.
Recursion. Given initial values $E_{\tau_{i-1}}\left(\tau_{i}\right), \ldots, E_{\tau_{i-1}}\left(\tau_{n+1}\right)$, define

$$
E_{t}\left(\tau_{i}\right), \ldots, E_{t}\left(\tau_{n+1}\right) \quad \text { for all } t \in\left[\tau_{i-1}, \tau_{i}\right]
$$

successively for all $i=1, \ldots, n$ by the following recursive procedure started at $i:=1$ :
(i) Extend the next maturing futures price to its delivery date $\tau_{i}$ by

$$
\begin{array}{rlrl}
\Sigma_{t}\left(\tau_{i}\right) & =\psi_{t}, & t \in\left[\tau_{i-1}, \tau_{i}\right] ; \\
d E_{t}\left(\tau_{i}\right) & =E_{t}\left(\tau_{i}\right) \Sigma_{t}\left(\tau_{i}\right) d W_{t},
\end{array}
$$

then proceed with the other futures.
(ii) For $j=i, \ldots, n$, starting with the initial condition

$$
Z_{\tau_{i-1}}\left(\tau_{j}\right)=\frac{E_{\tau_{i-1}}\left(\tau_{j}\right)+\kappa}{E_{\tau_{i-1}}\left(\tau_{j+1}\right)}-1
$$

solve the stochastic differential equation

$$
\frac{d Z_{t}\left(\tau_{j}\right)}{Z_{t}\left(\tau_{j}\right)}=-\sigma_{t}\left(\tau_{j}\right)\left(\frac{E_{t}\left(\tau_{j}\right) \Sigma_{t}\left(\tau_{j}\right)}{E_{t}\left(\tau_{j}\right)+\kappa}-\frac{Z_{t}\left(\tau_{j}\right) \sigma_{t}\left(\tau_{j}\right)}{Z_{t}\left(\tau_{j}\right)+1}\right) d t+\sigma_{t}\left(\tau_{j}\right) d W_{t}
$$

for $t \in\left[\tau_{i-1}, \tau_{i}\right]$ and define for all $t \in\left[\tau_{i-1}, \tau_{i}\right]$

$$
\begin{aligned}
\Sigma_{t}\left(\tau_{j+1}\right) & =\frac{E_{t}\left(\tau_{j}\right) \Sigma_{t}\left(\tau_{j}\right)}{E_{t}\left(\tau_{j}\right)+\kappa}-\frac{Z_{t}\left(\tau_{j}\right) \sigma_{t}\left(\tau_{j}\right)}{Z_{t}\left(\tau_{j}\right)+1} \\
E_{t}\left(\tau_{j+1}\right) & =\frac{E_{t}\left(\tau_{j}\right)+\kappa}{1+Z_{t}\left(\tau_{j}\right)}
\end{aligned}
$$

(iii) If $i<n+1$, we set $i:=i+1$ and proceed with the recursion onto $\left[\tau_{i}, \tau_{i+1}\right]$; otherwise we finish the loop.

To see that this procedure is well defined, we apply the results of Lemma 5.1 to the recursion step replacing $E_{t}, \Sigma_{t}, \sigma_{t}, Z_{t}$ by $E_{t}\left(\tau_{i}\right), \Sigma_{t}\left(\tau_{i}\right), \sigma_{t}\left(\tau_{i}\right), Z_{t}\left(\tau_{i}\right)$, respectively. The presented construction yields futures prices $\left(E_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ for $i=1, \ldots, n+1$ which obviously satisfy (C0), (C1), (C2), (C4) from Definition 3.1. The last requirement (C4) follows from

$$
E_{t}\left(\tau_{i+1}\right)-\kappa<E_{t}\left(\tau_{i}\right) \quad \Longleftrightarrow \quad Z_{t}\left(\tau_{i}\right)>0
$$

where the existence of $Z_{t}\left(\tau_{i}\right) \geq 0$ for all $t \in\left[0, \tau_{i}\right]$ is ensured by the assumptions (5.22) and Lemma 5.1. Note that we do not consider (C3) since the spot price is not covered by the above construction. If required, the spot price can be constructed in accordance with assumption (C3). (See (5.24) below.)

Remark. Let us elaborate on simplifying assumptions, which we adopted in the present approach in order to highlight the limitations and possible extensions of the model. First, our framework is easily extendable to a nonequidistant maturity grid. By assuming that the tenors $\Delta_{i}:=\tau_{i+1}-\tau_{i}$ depend on $i=0, \ldots, n$, we have to introduce different storage costs $\left(\kappa_{i}\right)_{i=0}^{n}$, supposing that each $\kappa_{i}$ is valid for the corresponding interval $\left[\tau_{i}, \tau_{i+1}\right]$. At this point, let us mention that our construction also works if the parameter $\kappa_{i}$ is random, provided that it is known with certainty prior to $\tau_{i}$, just before the beginning of corresponding interval $\left[\tau_{i}, \tau_{i+1}\right]$. Next, note that we address the futures prices directly, without reference to the spot price. More precisely, the spot price occurs from the construction merely at the grid points $\left(\tau_{i}\right)_{i=1}^{n+1}$, being the terminal futures price $S_{\tau_{i}}=E_{\tau_{i}}\left(\tau_{i}\right), i=1, \ldots, n+1$. Since the spot price is not quoted for most commodities, we believe this gives almost no limitation of model applicability. However, if for some reason a spot price model is required, then the prices $\left(S_{\tau_{i}}\right)_{i=1}^{n+1}$ can be interpolated accordingly, for instance, linearly:

$$
\begin{equation*}
S_{t}=\frac{t-\tau_{i}}{\tau_{i+1}-\tau_{i}} E_{t}\left(\tau_{i+1}\right)+\frac{\tau_{i+1}-t}{\tau_{i+1}-\tau_{i}} E_{\tau_{i}}\left(\tau_{i}\right), \quad t \in\left[\tau_{i}, \tau_{i+1}\right], \quad i=1, \ldots, n \tag{5.24}
\end{equation*}
$$

Such a choice comes close to the frequently used approximation of the spot price by the price of the futures contract with nearest maturity. Finally, let us mention that by the reconstruction of the spot price through interpolation the model can be extended towards a continuous system of maturities by defining futures price evolution

$$
E_{t}(\tau)=\mathbb{E}^{Q^{E}}\left(S_{\tau} \mid \mathcal{F}_{t}\right) \quad \text { for all } t \in[0, \tau], \tau \in[0, T]
$$

Obviously, such an extension provides at any time $t$ a futures curve $\left(E_{t}(\tau)\right)_{\tau \in[t, T]}$ which is given by the underlying discrete curve $\left(E_{t}\left(\tau_{i}\right)\right)_{\tau_{i} \geq t}$ interpolated by the procedure used in the construction of the spot price. In particular, the linear interpolation (5.24) yields a continuous and piecewise linear futures curve which respects the same contango limit as the discrete model.

Finally, we emphasize that one of the main advantages of our model is a perfect time consistency of the futures curve evolution, whereas to the best of our knowledge all approaches in commodity modeling existing so far suffer from an inconsistency. For instance, defined by few parameters, common spot price based commodity models (see [10]) are not able to match an arbitrary initial futures curve. From this perspective, the futures price based models (discussed in, among others, [4], [8]) are more appropriate since they provide an exact fit
to the futures curve at the beginning. However, starting from such an initial curve, the model-based futures curve evolution is in general not able to capture the real-world futures curve change. This does not occur in our model since we describe a finite number of futures contracts, in accordance with the common market practice. Namely, starting from the initial curve $\left(E_{0}\left(\tau_{i}\right)\right)_{i=1}^{n+1}$, the futures curve $\left(E_{t}\left(\tau_{i}\right)\right)_{i=1}^{n+1}$ at a later time $t \in\left[\tau_{0}, \tau_{1}\right]$ can be arbitrary, with the only restriction that it respect the contango limit. To see this, observe that the distribution of

$$
E_{t}\left(\tau_{1}\right), \quad \frac{E_{t}\left(\tau_{2}\right)}{E_{t}\left(\tau_{1}\right)+\kappa}=\frac{1}{Z_{t}\left(\tau_{1}\right)+1}, \quad \ldots, \quad \frac{E_{t}\left(\tau_{n+1}\right)}{E_{t}\left(\tau_{n}\right)+\kappa}=\frac{1}{Z_{t}\left(\tau_{n}\right)+1}
$$

is equivalent to the Lebesgue measure on $] 0, \infty[\times] 0,1[n$, which is ensured by our construction of the next to maturity future and of the simple ratio processes from geometric Brownian motions, using appropriate volatility structures.
6. Model calibration. This section is devoted to the calibration of the parameters of our model. In the case that an appropriate type of calendar spread is actively traded on the market, an implicit calibration is possible. Otherwise one has to rely on a historical calibration based on principal component analysis. Alternatively, an approximation of calendar spread option prices which is described in section 8 could also be used for an implicit calibration.

Implicit calibration. As mentioned previously, the model inherits the implied calibration features from the BGM paradigm. Namely, for the case where interest rates and simple ratio volatilities are deterministic, the fair prices $\left(C_{s}\right)_{s \in[0, t]}$ of the calendar spread option maturing at $t$ with the terminal payoff

$$
\begin{equation*}
C_{t}=\left(E_{t}\left(\tau_{i}\right)+\kappa-(1+K) E_{t}\left(\tau_{i+1}\right)\right)^{+} \quad\left(t \leq \tau_{i}<\tau_{i+1}\right) \tag{6.1}
\end{equation*}
$$

are given by

$$
\begin{equation*}
C_{s}=e^{-r(t-s)} E_{s}\left(\tau_{i+1}\right) \mathrm{BS}\left(\frac{E_{s}\left(\tau_{i}\right)+\kappa-E_{s}\left(\tau_{i+1}\right)}{E_{s}\left(\tau_{i+1}\right)}, K, t, s, 0, \frac{D\left(s, t, \tau_{i}\right)}{\sqrt{t-s}}\right), \quad s \in[0, t], \tag{6.2}
\end{equation*}
$$

where $\mathrm{BS}(x, k, t, s, \rho, v)$ stands for the standard Black-Scholes formula

$$
\begin{aligned}
\mathrm{BS}(x, k, t, s, \rho, v) & :=x \mathcal{N}\left(d_{+}\right)-e^{-\rho(t-s)} k \mathcal{N}\left(d_{-}\right), \\
d_{+} & =\frac{1}{v \sqrt{t-s}}\left[\log \left(\frac{x}{k}\right)+\left(\rho+\frac{1}{2} v^{2}\right)(t-s)\right], \\
d_{-} & =d_{+}-v \sqrt{t-s}
\end{aligned}
$$

and $D\left(s, t, \tau_{i}\right)=\int_{s}^{t}\left\|\sigma_{u}\left(\tau_{i}\right)\right\|^{2} d u$. Consider the measure $Q^{\tau_{i+1}}$, given by

$$
d Q^{\tau_{i+1}}=\frac{E_{\tau_{i+1}}\left(\tau_{i+1}\right)}{E_{0}\left(\tau_{i+1}\right)} d Q^{E}
$$

By the change of measure technique, the process

$$
Z_{t}\left(\tau_{i}\right)=\frac{E_{t}\left(\tau_{i}\right)+\kappa-E_{t}\left(\tau_{i+1}\right)}{E_{t}\left(\tau_{i+1}\right)}, \quad t \in\left[0, \tau_{i}\right]
$$

follows a martingale with respect to $Q^{\tau_{i+1}}$ with stochastic differential

$$
d Z_{t}\left(\tau_{i}\right)=Z_{t}\left(\tau_{i}\right) \sigma_{t}\left(\tau_{i}\right) d W_{t}^{\tau_{i+1}}
$$

driven by the process $\left(W_{t}^{\tau_{i+1}}\right)_{t \in\left[0, \tau_{i}\right]}$ of Brownian motion with respect to $Q^{\tau_{i+1}}$. Since the simple ratio volatility $\left(\sigma_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ is deterministic by assumption, $\left(Z_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$ follows a geometric Brownian motion with respect to $Q^{\tau_{i+1}}$, which we use to derive

$$
\begin{aligned}
C_{s} & =e^{-r(t-s)} \mathbb{E}^{Q^{E}}\left(\left(E_{t}\left(\tau_{i}\right)+\kappa-(1+K) E_{t}\left(\tau_{i+1}\right)\right)^{+} \mid \mathcal{F}_{s}\right) \\
& =e^{-r(t-s)} \mathbb{E}^{Q^{E}}\left(\left.E_{\tau_{i}}\left(\tau_{i+1}\right)\left(\frac{E_{t}\left(\tau_{i}\right)+\kappa-E_{t}\left(\tau_{i+1}\right)}{E_{t}\left(\tau_{i+1}\right)}-K\right)^{+} \right\rvert\, \mathcal{F}_{s}\right) \\
& =e^{-r(t-s)} E_{s}\left(\tau_{i+1}\right) \mathbb{E}^{Q^{\tau_{i+1}}}\left(\left(Z_{t}\left(\tau_{i}\right)-K\right)^{+} \mid \mathcal{F}_{s}\right) .
\end{aligned}
$$

Note that the observation of the implied volatilities through (6.2) yields information on the term structure of the simple ratio volatilities. This kind of implicit calibration is possible if the market lists a sufficient number of calendar spread calls with appropriate parameters ( $K+1$ ) and $\kappa$ as in (6.1). Realistically, one cannot assume that there is always trading in such specific instruments. Hence, the identification of the volatility structure from historical data may become unavoidable.

Historical calibration. Next, we present a method for the historical model calibration based on principal component analysis (PCA). This methodology has been applied for calibration of fixed income market models (see [4]) and has been successfully adapted to the estimation of futures volatilities in commodity and energy markets. In general, this technique requires appropriate assumptions on time homogeneity. A typical hypothesis here is that the volatility is time dependent through the time to maturity only. Under this condition, the volatility term structure is identified by measuring the quadratic covariation of appropriate processes. Let us adapt this technique to our case.

In our approach, the model is defined by the volatility processes $\left(\psi_{t}\right)_{t \in[0, T]}$ and $\left(\sigma_{t}\left(\tau_{i}\right)\right)_{t \in\left[0, \tau_{i}\right]}$, $i=1, \ldots, n$, giving next to maturity futures prices and the simple ratio processes:

$$
\begin{array}{ll}
d E_{t}\left(\tau_{i}\right)=E_{t}\left(\tau_{i}\right) \psi_{t} d W_{t}, & \\
d \in\left[\tau_{i-1}, \tau_{i}\right], \quad i=1, \ldots, \\
d Z_{t}\left(\tau_{i}\right)=Z_{t}\left(\tau_{i}\right)\left(\alpha_{t}\left(\tau_{i}\right)+\sigma_{t}\left(\tau_{i}\right) d W_{t}\right), & \\
t \in\left[0, \tau_{i}\right], \quad i=1, \ldots, n .
\end{array}
$$

We now consider the following assumption:
There exist $v^{0}, v^{1}, \ldots, v^{m} \in \mathbb{R}^{d}$ such that $\psi_{t}=v^{0}$ for all $t \in[0, T]$ and $\sigma_{t}\left(\tau_{i}\right)=\sum_{k=1}^{m} v^{k} 1_{] \Delta(k-1), \Delta k]}\left(\tau_{i}-t\right)$ holds for all $t \in[0, T]$ and $i=1, \ldots, m$.

In other words, (6.3) states that the in-front-of-maturity futures follow constant and deterministic volatility and that the deterministic simple ratio volatility is piecewise constant and time dependent through the time to maturity only. Under the assumption (6.3), the vectors $v^{0}, v^{1}, \ldots, v^{m} \in \mathbb{R}^{d}$ are recovered from the quadratic covariation

$$
\begin{equation*}
v^{k} v^{l} \Delta=\left[X^{k}(i), X^{l}(i)\right]_{\tau_{i}}-\left[X^{k}(i), X^{l}(i)\right]_{\tau_{i-1}} \tag{6.4}
\end{equation*}
$$

where the processes $\left(X_{t}^{k}(i)\right)_{t \in\left[\tau_{i-1}, \tau_{i}\right]}, i=1, \ldots, n+1, k=0, \ldots, m$, are given by

$$
\begin{align*}
& X_{t}^{0}(i):=\ln E_{t}\left(\tau_{i}\right),  \tag{6.5}\\
& X_{t}^{k}(i):=\ln Z_{t}\left(\tau_{i-1+k}\right)=\ln \left(\frac{E_{t}\left(\tau_{i-1+k}\right)+\kappa}{E_{t}\left(\tau_{i+k}\right)}-1\right), \quad k=1, \ldots, m, \tag{6.6}
\end{align*}
$$

for $t \in\left[\tau_{i-1}, \tau_{i}\right]$. Consider historical futures prices, where within each trading period $\left[\tau_{i-1}, \tau_{i}\right]$ futures prices for $m+1$ subsequent maturity dates $\tau_{i}, \ldots, \tau_{i+m}$ are available. For such data, calculate the observations

$$
\left.\left.X_{t_{j}}^{k}(i)(\omega), \quad t_{j} \in\right] \tau_{i-1}, \tau_{i}\right], \quad k=0, \ldots, m, \quad i=1, \ldots, m
$$

from (6.5), (6.6) at discrete times $t_{j}$ where the corresponding futures prices are available. With this data set, approximate the quadratic covariation (6.4) by

$$
V^{k, l}(i)=\sum_{\left.\left.\left(t_{j}, t_{j+1}\right) \in\right] \tau_{i-1}, \tau_{i}\right]^{2}}\left(X_{t_{j+1}}^{k}(i)-X_{t_{j}}^{k}(i)\right)\left(X_{t_{j+1}}^{l}(i)-X_{t_{j}}^{l}(i)\right), \quad k, l=0, \ldots, m .
$$

For a representative history and sufficiently high data frequency, we can suppose that the empirical quadratic covariation

$$
\begin{equation*}
V^{k, l}:=\frac{1}{\Delta(n+1)} \sum_{i=1}^{n+1} V^{k, l}(i), \quad k, l=0, \ldots, m \tag{6.7}
\end{equation*}
$$

estimates

$$
V^{k, l}=v^{k} v^{l}, \quad k, l=0, \ldots, m,
$$

the Gram's matrix of $v^{0}, \ldots, v^{n}$. The orthonormal eigenvectors of $V=\left(V^{k, l}\right)_{k, l=0}^{m}$ give the diagonalization $V=\Phi \Lambda \Phi^{\top}$ as follows: The columns $\Phi=\left(\Phi_{j, k}\right)_{j, k=0}^{m}$ are given by eigenvectors $\phi^{0}, \ldots, \phi^{m}$ and the corresponding eigenvalues $\lambda_{0} \geq \lambda_{1} \geq \cdots \geq \lambda_{m}$ (which we agree to place in descending order) are the diagonal entries of $\Lambda$. Following the philosophy of PCA, the model based on

$$
\begin{equation*}
\hat{v}^{k}:=\left(\lambda_{j}^{1 / 2} \Phi_{j, k}\right)_{j=0}^{m}, \quad k=0, \ldots, m, \tag{6.8}
\end{equation*}
$$

in (6.3) (instead of $v^{k}$ ) fits the historical observations since the quadratic covariation in (6.4) is correctly reflected due to

$$
\begin{equation*}
\hat{v}^{k} \hat{v}^{l}=V^{k, l} \quad \text { for } k, l=0, \ldots, m \tag{6.9}
\end{equation*}
$$

An important problem is whether the observed historical data could be approximatively described by the same type of model but with a reduced number $d^{\prime}<d$ of stochastic factors. In other words, the question is whether a model driven by $d^{\prime}<d$ Brownian motions would be sufficient to reproduce the observed empirical quadratic covariance. Note that if the user decides to reduce the dimension to $d^{\prime}<d$, then a reasonable approximation of the measured quadratic covariance is attained by dropping $\lambda_{d^{\prime}}, \ldots, \lambda_{m}$, the smallest eigenvalues. In this
case, (6.8) reduces to $\hat{v}^{k}:=\left(\lambda_{j}^{1 / 2} \Phi_{j, k}\right)_{j=0}^{d^{\prime}}$ for $k=0, \ldots, m$ with $\hat{v}^{k} \hat{v}^{l} \approx V^{k, l}$ for $k, l=0, \ldots, m$ instead of (6.9). Clearly, the dimension reduction is a trade-off between the accuracy of the covariance approximation and the dimension of underlying Brownian motion. For instance, a rule of thumb could be to reduce the dimension to $d^{\prime}$ if the remaining factors describe $95 \%$ of the covariance:

$$
\sum_{j=0}^{d^{\prime}-1} \lambda_{j} \geq 0.95 \sum_{j=0}^{m} \lambda_{j} .
$$

7. Empirical results. Let us return to the soybean trading from section 4. There we estimated the historical contango limit $\kappa=26$, which we will use in the following calibration. With this parameter, the realization of the logarithmic simple ratios can be computed and their quadratic covariations can be estimated as explained above.

For the entire soybean data set, from 2000-10-02 to 2007-02-23, the Gram matrix $V$ from (6.7) is obtained as

$$
V \approx\left[\begin{array}{cccccc}
0.06 & -0.01 & 0.00 & 0.02 & 0.04 & 0.04  \tag{7.1}\\
-0.01 & 1.45 & -0.09 & 0.04 & 0.05 & -0.18 \\
0.00 & -0.09 & 0.98 & -0.16 & 0.07 & -0.04 \\
0.02 & 0.04 & -0.16 & 0.96 & -0.40 & 0.10 \\
0.04 & 0.05 & 0.07 & -0.40 & 2.37 & -1.79 \\
0.04 & -0.18 & -0.04 & 0.10 & -1.79 & 5.86
\end{array}\right] .
$$

For this symmetric matrix one obtains the following eigenvalue decomposition,

$$
\Phi \approx\left[\begin{array}{cccccc}
0.00 & 0.02 & 0.00 & -0.02 & -0.05 & 1.00  \tag{7.2}\\
-0.04 & -0.18 & 0.96 & 0.21 & -0.01 & 0.01 \\
-0.01 & 0.15 & -0.16 & 0.84 & -0.49 & -0.01 \\
0.04 & -0.39 & 0.00 & -0.41 & -0.82 & -0.04 \\
-0.39 & 0.81 & 0.19 & -0.28 & -0.27 & -0.04 \\
0.92 & 0.36 & 0.12 & -0.08 & -0.08 & -0.02
\end{array}\right],
$$

where the orthonormal eigenvectors are displayed as the columns and the corresponding eigenvalues are given by

$$
\begin{equation*}
\lambda_{0} \approx 6.63, \quad \lambda_{1} \approx 1.78, \quad \lambda_{2} \approx 1.45, \quad \lambda_{3} \approx 1.01, \quad \lambda_{4} \approx 0.74, \quad \lambda_{5} \approx 0.05 \tag{7.3}
\end{equation*}
$$

As shown in the previous section, the volatility vectors are calculated by (6.8). Based on (7.2) and (7.3) one obtains the following volatility vectors for our soybean market:

$$
\begin{align*}
v^{0} & =\left[\begin{array}{ccccccc}
0.01 & 0.03 & 0.00 & -0.02 & -0.04 & 0.23 & ]^{\top}, \\
v^{1} & =\left[\begin{array}{cccccc}
-0.09 & -0.24 & 1.16 & 0.21 & -0.01 & 0.00
\end{array}\right]^{\top}, \\
v^{2} & =\left[\begin{array}{cccccc}
-0.03 & 0.20 & -0.19 & 0.84 & -0.43 & 0.00
\end{array}\right]^{\top}, \\
v^{3} & =\left[\begin{array}{cccccc}
0.11 & -0.53 & 0.01 & -0.41 & -0.71 & -0.01
\end{array}\right]^{\top}, \\
v^{4} & =\left[\begin{array}{cccccc}
-1.00 & 1.08 & 0.23 & -0.28 & -0.23 & -0.01
\end{array}\right]^{\top}, \\
v^{5} & =\left[\begin{array}{ccccc}
2.37 & 0.48 & 0.14 & -0.08 & -0.07 \\
\hline
\end{array}\right]
\end{array} \begin{array}{c}
0.00
\end{array}\right]^{\top} . \tag{7.4}
\end{align*}
$$

8. Option pricing. This section is devoted to the pricing and hedging of European options written on storable underlyings. First, we show how to compute a hedging strategy, based on a partial differential equation. However, as this equation is typically high dimensional, a Monte Carlo simulation could be the appropriate method to price the options in our framework. As the Monte Carlo simulation is straightforward, we do not discuss this method in detail but apply it to examine the risk neutral distribution of the spread option payoff. We find out that this distribution is close to lognormal, which suggests that calendar spread option prices could be approximated by a Black-Scholes-type formula. Here, we build on the model we calibrated in the previous sections. Hence we assume the settings of Lemma 5.1, i.e.,

$$
\left.\begin{array}{c}
\text { the volatilities }\left(\psi_{u}\right)_{u \in[0, T]} \text { and }  \tag{8.1}\\
\left(\sigma_{u}\left(\tau_{i}\right)\right)_{u \in[0, T]} \text { for all } i=1, \ldots, n+1 \\
\text { are bounded and deterministic }
\end{array}\right\} \text {. }
$$

Consider $k+1$ futures prices

$$
\begin{equation*}
\left(E_{u}\left(\tau_{i}\right), E_{u}\left(\tau_{i+1}\right), \ldots, E_{u}\left(\tau_{i+k}\right)\right), \quad u \in\left[\tau_{i-1}, \tau_{i}\right], \quad i+k \leq n+1, \tag{8.2}
\end{equation*}
$$

listed within the interval $\left[\tau_{i-1}, \tau_{i}\right]$. By construction, all of these processes follow

$$
d E_{u}\left(\tau_{i+j}\right)=E_{u}\left(\tau_{i+j}\right) \Sigma_{u}\left(\tau_{i+j}\right) d W_{u}, \quad u \in\left[\tau_{i-1}, \tau_{i}\right], \quad j=0, \ldots, k,
$$

where, according to (5.21), the volatility is given by a function

$$
\Sigma_{u}\left(\tau_{i+j}\right)=s^{(j)}\left(u, E_{u}\left(\tau_{i}\right), \ldots, E_{u}\left(\tau_{i+j}\right)\right), \quad u \in\left[\tau_{i-1}, \tau_{i}\right],
$$

which can be calculated recursively:
(i) if $j=0$, then $s^{(0)}\left(u, E_{u}\left(\tau_{i}\right)\right)=\psi_{u}$;
(ii) if $j>0$, then

$$
\begin{equation*}
s^{(j)}\left(u,\left(E_{u}\left(\tau_{i+l}\right)\right)_{l=0}^{j}\right)=s\left(E_{u}\left(\tau_{i+j}\right), E_{u}\left(\tau_{i+j-1}\right), s^{(j-1)}\left(u,\left(E_{u}\left(\tau_{i+l}\right)\right)_{l=0}^{j-1}\right), \sigma_{u}\left(\tau_{i+j-1}\right)\right), \tag{8.3}
\end{equation*}
$$

where $s(\cdot)$ is the function introduced in (5.21).
That is, due to Proposition 5.2, the vector of processes (8.2) follows a unique strong solution to

$$
d E_{u}\left(\tau_{i+j}\right)=E_{u}\left(\tau_{i+j}\right) s^{(j)}\left(u, E_{u}\left(\tau_{i}\right), \ldots, E_{u}\left(\tau_{i+j}\right)\right) d W_{u}, \quad u \in\left[\tau_{i-1}, \tau_{i}\right], \quad j=0, \ldots, k ;
$$

hence (8.2) is a Markov process.
Let us now consider the valuation of derivatives, written on commodity futures prices. Given the European option with payoff $f\left(\left(E_{\tau}\left(\tau_{i+j}\right)\right)_{j=0}^{k}\right)$ at maturity $\tau \in\left[\tau_{i-1}, \tau_{i}\right]$, its expected payoff conditioned on time $t \in\left[\tau_{i-1}, \tau\right]$ is given by

$$
\mathbb{E}\left(f\left(\left(E_{\tau}\left(\tau_{i+j}\right)\right)_{j=0}^{k}\right) \mid \mathcal{F}_{t}\right)=: g\left(\left(E_{t}\left(\tau_{i+j}\right)\right)_{j=0}^{k}\right)
$$

with an appropriate function $g(\cdot)$ whose existence follows from the Markov property. Furthermore, Itô's formula shows that this function can be determined as $g(\cdot)=\phi(t, \cdot)$ from the solution to the partial differential equation

$$
\begin{align*}
& \frac{\partial}{\partial u} \phi\left(u, e_{0}, \ldots, e_{k}\right)  \tag{8.4}\\
& =-\frac{1}{2} \sum_{j, l=0}^{k} \frac{\partial^{2}}{\partial e_{j} \partial e_{l}} \phi\left(u, e_{0}, \ldots, e_{k}\right) E_{u}\left(\tau_{i+j}\right) E_{u}\left(\tau_{i+l}\right) s^{(j)}\left(u, e_{0}, \ldots, e_{j}\right) s^{(l)}\left(u, e_{0}, \ldots, e_{l}\right)
\end{align*}
$$

for $\left.\left(u, e_{0}, \ldots, e_{k}\right) \in\right] t, \tau[\times] 0, \infty\left[^{k+1}\right.$ subject to the boundary condition

$$
\left.\phi\left(\tau, e_{0}, \ldots, e_{k}\right)=f\left(e_{0}, \ldots, e_{k}\right) \quad \text { for }\left(e_{0}, \ldots, e_{k}\right) \in\right] 0, \infty{ }^{k+1}
$$

Finally, from the stochastic integral representation one obtains

$$
f\left(\left(E_{\tau}\left(\tau_{i+j}\right)\right)_{j=0}^{k}\right)-g\left(\left(E_{t}\left(\tau_{i+j}\right)\right)_{j=0}^{k}\right)=\sum_{j=0}^{k} \int_{t}^{\tau} \frac{\partial}{\partial e_{j}} \phi\left(u, E_{u}\left(\tau_{i}\right), \ldots, E_{u}\left(\tau_{i+k}\right)\right) d E_{u}\left(\tau_{i+j}\right)
$$

and the hedging strategy for the European contingent claim becomes evident. Holding at any time $u \in[t, \tau]$ the position

$$
h_{u}\left(\tau_{i+j}\right)=\frac{\partial}{\partial e_{j}} \phi\left(u, E_{u}\left(\tau_{i}\right), \ldots, E_{u}\left(\tau_{i+k}\right)\right) p_{u}(\tau)
$$

in the futures contract with maturity $\tau_{i+j}$, the European option payoff can be perfectly replicated starting with the initial endowment

$$
\begin{equation*}
p_{t}(\tau) g\left(\left(E_{t}\left(\tau_{i+j}\right)\right)_{j=0}^{k}\right) \tag{8.5}
\end{equation*}
$$

Namely, by transferring the initial endowment and all cash flows from futures settlements to $\tau$ by a zero bond maturing at $\tau$, we determine the wealth of this strategy as

$$
\sum_{j=0}^{k} \int_{t}^{\tau} \frac{1}{p_{u}(\tau)} h_{u}\left(\tau_{i+j}\right) d E_{u}\left(\tau_{i+j}\right)+\frac{p_{t}(\tau) g\left(\left(E_{t}\left(\tau_{i+j}\right)\right)_{j=0}^{k}\right)}{p_{t}(\tau)}=f\left(\left(E_{\tau}\left(\tau_{i+j}\right)\right)_{j=0}^{k}\right),
$$

which matches the contingent claim of our option. In the case that the replication is required from a date $t$ earlier than $\tau_{i-1}$, the same arguments need to be repeated for the previous intervals.

Note that our model is inherently not complete. For instance, on the very last interval $\left[\tau_{n}, \tau_{n+1}\right]$ only one future is traded, but the number of uncertainty sources is still $d>1$. However, the above considerations show that under the assumption (8.1) a European option, written on futures, can be replicated by appropriate positions in futures traded prior to the expiry date of the option and in zero bonds maturing at this date.

Since the dimension of the partial differential equation (8.4) could be high and there is no evident price approximations even for the simplest plain-vanilla options, we believe that


Figure 4. A simulation of futures prices. Vertical lines separate two month periods.

Monte Carlo simulation could be an appropriate way to price derivative instruments within our model. To discuss an application, we turn to the valuation of a calendar spread option. Here, we utilize the parameters obtained from the soybean example presented above. Note that, as previously, futures prices are expressed in US cents per bushel, and $\Delta$ stands for the two month duration between expiry dates, $\tau_{i}=i \Delta, i=1, \ldots, n+1$. The payoff of a calendar spread option depends on the price difference of commodities delivered at different times. For instance, it may provide the holder with the payment $\left(E_{t}(\tau)-E_{t}\left(\tau^{\prime}\right)-K\right)^{+}$at maturity $t$, where $\tau$ and $\tau^{\prime}$ are future delivery dates satisfying $t<\tau<\tau^{\prime}$. Based on a Monte Carlo simulation, we examine the distribution of the difference $E_{t}\left(\tau^{\prime}\right)-E_{t}(\tau)$ for the parameters $t=\tau=4 \Delta, \tau^{\prime}=6 \Delta$. Having supposed that the initial futures curve is flat $\left(E_{0}\left(\tau_{i}\right)=800\right)_{i=1}^{6}$, 5000 realizations are generated. Figure 4 depicts a typical path resulting from a single run of the Monte Carlo method. The estimated density $E_{t}(\tau)-E_{t}\left(\tau^{\prime}\right)$ is shown in Figure 5. Note the clear similarity to the shifted lognormal distribution. Next, we give a detailed discussion of this observation. First, the storage costs for two periods (four months) is recognized as the correct shift parameter. Indeed, $E_{t}(\tau)-E_{t}\left(\tau^{\prime}\right)>-2 \kappa$ holds by construction, so an appropriate choice of the shift parameter corresponds to the expiry dates difference $\tau^{\prime}-\tau$. To compare now the distribution of $\ln \left(E_{t}(\tau)-E_{t}\left(\tau^{\prime}\right)+2 \kappa\right)$ to the appropriately scaled normal distribution, we plot in Figure 6 the estimated density of

$$
\begin{equation*}
\ln \left(E_{t}(\tau)-E_{t}\left(\tau^{\prime}\right)+2 \kappa\right) \tag{8.6}
\end{equation*}
$$

in comparison to the normal density, whose first two moments are chosen to match those estimated for (8.6). That is, a close approximation in distribution

$$
\begin{equation*}
E_{t}(\tau)-E_{t}\left(\tau^{\prime}\right) \stackrel{d}{\approx} \exp (X)-2 \kappa, \quad \text { where } X \text { is Gaussian, } \tag{8.7}
\end{equation*}
$$



Figure 5. The density of $E_{t}(\tau)-E_{t}\left(\tau^{\prime}\right)$ exhibits a similarity to the lognormal density.


Figure 6. The density of (8.6) (black) in comparison to the normal distribution (red).
seems to be possible. This can be used to derive an approximation for the price of spread options, where by assuming the equality in (8.7) a Black-Scholes-type formula is obtained. This observation is according to the approximative pricing schemes for spread and basket options extensively studied in [1], [2], [4] and in the literature cited therein.
9. Conclusion. In this work, we have addressed the role of storage costs in commodity price modeling. By appropriate interpretation of the no-arbitrage principle, we formulate a minimal set of model assumptions which exclude arbitrage opportunities for futures trading and simultaneous management of a stylized storage facility. In the presented commodity model class, the storage cost plays the role of a constant parameter, which bounds the steepness of the futures curve in any contango situation. This bound, well known as the contango limit in commodity trading, forms an intrinsic ingredient of the proposed martingale-based futures price dynamics. Following the expertise from the interest rate theory, we demonstrate how
to construct and to calibrate commodity models which correspond to our assumptions. An empirical study of soybean futures trading illustrates this concept. Moreover, we discuss the valuation of calendar spread options. Here, numerical experiments raise the hope that an appropriately shifted lognormal distribution may give an excellent approximation for the payoff distributions of calendar spreads. This issue could be important for efficient pricing and hedging of calendar spread options.

## REFERENCES

[1] C. Alexander and A. Venkatramanan, Spread Options as Compound Exchange Options with Applications to American Crack Spreads, preprint, 2007.
[2] S. Borovkova, F. J. Permana, and H. Weide, A closed form approach to the valuation and hedging of basket and spread options, J. Derivatives, 14 (2007), pp. 8-24.
[3] M. J. Brennan, The supply of storage, Am. Econ. Rev., 47 (1958), pp. 50-72.
[4] R. Carmona and V. Durrleman, Pricing and hedging spread options, SIAM Rev., 45 (2003), pp. 627-685.
[5] E. F. Fama and K. R. French, Commodity future prices: Some evidence on forecast power, premiums, and the theory of storage, J. Business, 60 (1987), pp. 55-73.
[6] H. Geman, Commodities and Commodity Derivatives: Modeling and Pricing for Agriculturals, Metals and Energy, John Wiley \& Sons, Chichester, 2005.
[7] J. E. Hilard and J. Reis, Valuation of commodity futures and options under stochastic convenience yields, interest rates and jump diffusions in the spot, J. Financ. Quant. Anal., 33 (1998), pp. 61-68.
[8] J. Hinz and M. Wilhelm, Pricing flow commodity derivatives using fixed income market techniques, Int. J. Theor. Appl. Finance, 9 (2006), pp. 1299-1321.
[9] K. R. Miltersen and E. S. Schwartz, Pricing of options on commodity futures with stochastic term structures of convenience yields and interest rates, J. Financ. Quant. Anal., 33 (1998), pp. 33-59.
[10] E. S. Schwartz, The stochastic behavior of commodity prices: Implication for valuation and hedging, J. Finance, 52 (1997), pp. 923-973.

# Optimal Allocation of a Futures Portfolio Utilizing Numerical Market Phase Detection* 

L. Putzig ${ }^{\dagger}$, D. Becherer ${ }^{\ddagger}$, and I. Horenko ${ }^{\S}$


#### Abstract

This paper presents an application of the recently developed method for simultaneous dimension reduction and metastability analysis of high-dimensional time series in the context of computational finance. Further extensions are included to combine state-specific principal component analysis (PCA) and state-specific regressive trend models to handle the high-dimensional, nonstationary data. The identification of market phases allows one to control the involved phase-specific risk for futures portfolios. The numerical optimization strategy for futures portfolios based on Tikhonovtype regularization is presented. The application of proposed strategies to online detection of the market phases is exemplified first on the simulated data and then on historical futures prices for oil and wheat from 2005-2008. Numerical tests demonstrate the comparison of the presented methods with existing approaches.


Key words. economic time series analysis, portfolio optimization, investment, regression, market phases, clustering

AMS subject classifications. $62 \mathrm{M} 10,91 \mathrm{~B} 28,91 \mathrm{~B} 84$

DOI. 10.1137/090754029

1. Introduction. In today's financial industry, computational finance is used for a wide field of applications. Throughout this paper, we will concentrate on the portfolio optimization problem, where, for return maximization and risk minimization problems, the evolution of security returns is estimated with a variety of different methods, in order to find an allocation that optimizes a suitable performance criterion, while limiting the risk [27, 31]. For practical reasons, computational methods have to address the following points:
2. Time series analysis. Market dynamics are most likely not reflected by a stationary process, but involve (hidden) market phases, such as economic cycles, harvest cycles (for grain, fruits, and other kinds of plants), and yearly holidays, and many more aspects influence the markets. This poses fundamental problems for estimation: To obtain sensible estimates for mean returns even in stationary linear models, hundreds of years of data would be needed. In shorter intervals, estimation risk becomes significant.

[^77]2. Sensitivity analysis. Approaches should be robust w.r.t. model parameter uncertainty. This is a challenge for identification when estimating time-varying parameters.
3. High dimensionality. Financial data are high dimensional, as thousands of securities are available from different asset classes (stocks, futures, etc.), and data are often available at high frequencies with nonequidistant time stepping (due to weekends, market closures, holidays, etc.); e.g., the Wilshire 5000 consists of 5000 stocks. Hence, computational ability to handle high numbers of assets (stocks, futures, etc.) is relevant.
Aspects 2 and 3 cause a high uncertainty of estimates, whereas nonstationarity as it relates to aspect 1 (in observable dimensions, at least) means that estimates should involve timedependent functions, which could not be estimated in a nontrivial way, which is a well-known problem. In 1989 Hamilton [13] suggested a numerical method for identification of piecewise linear functions, called hidden market phases, assuming stationarity and Gaussianity of returns. While Gaussianity is still a commonly used assumption by practitioners, many modern models soften this by using, e.g., heavy tails [29]. Later this was extended to multidimensional data using linear vector autoregressive (VAR) models [23]. Still, standard phase-identification techniques based on the filtering approach (like wavelet-based spectral methods; see [1]) have in general infeasible numerical complexity in high dimensions; other methods (e.g., Kalman filter [21], VAR [23], (G)ARCH [6, 2]) require stronger assumptions about the underlying dynamics or the methods (as in [3]) are based on perfect observability of the Markov process, which cannot be taken for granted in reality. As demonstrated by the actual financial crises, nonrobust reliance on simplified assumptions and ignorance of model risk can (among other factors) lead to misperception of risk and possibly to a total loss of the investment. Therefore, there is a need for a wider class of methods that are computationally capable of dealing with realistic amounts of data, monitoring the evolution of risks, and optimizing the risk-return profile of investments. While recent publications concentrate on robustness by using worstcase scenarios (see, e.g., [10, 7, 11]) or penalty functions (see, e.g., [35]), or simply assuming vanishing mean returns, this paper aims at an estimation of nonstationary trends and volatilities directly from prices of futures contracts. Futures markets constitute broad and liquid markets for several asset classes [5]. Leverage can be quite large for futures if the amplitude of futures price variations is measured relative to the size of the margin requirement. But this also induces a high risk. The minimization of this risk is one of the central points of this paper. The recently developed framework for numerical treatment of aspect 1 of the above list of problems is presented and used for clustering and analysis of nonstationary financial data $[16,17,14,15]$. An extension of the framework utilizing linear regression is introduced, allowing for simultaneous market phase identification, dimension reduction, and estimation of the phase-specific trends in the financial context. The ability of the extended framework to identify change points online is demonstrated, both on simulated and real market data. Finally, the numerical optimization strategy for phase-specific selection of futures portfolios is presented and exemplified on a real 8-dimensional data set of futures data for wheat and oil.

This paper is organized as follows. In section 2, the FEM-based clustering framework is introduced and its extension towards the financial data context is presented. In section 3, we briefly consider the definition of returns and adapt this to futures markets, as the naive definition does not fit in this context. The numerical futures portfolio optimization approach is presented. Section 4 demonstrates the application of the proposed numerical framework
to simulated and real data and offers a comparison to standard approaches (like the classical methods proposed by Markowitz [27], the flexible multivariate $\operatorname{GARCH}(1,1)$ model [24], and a simple $\frac{1}{n}$-portfolio). In the end, a practical side constraint is considered, the notion of neutrality.
2. Finding market phases. In the following, we will briefly describe the clustering framework first introduced in $[16,17,14,15]$. Special emphasis will be placed on the numerical aspects and the interpretation of the framework in the context of high-dimensional financial applications.
2.1. The clustering problem. Let $S_{t}:[0, T] \rightarrow[0, \infty)^{d}$ be the observed $d$-dimensional series of asset prices. We look for $K$ models characterized by $K$ distinct sets of a priori unknown model parameters $\theta_{1}, \ldots, \theta_{K} \in \Omega$, where $\Omega$ is the parameter space. Let $g\left(S_{t}, \theta_{i}\right)$ : $\mathbb{R}^{d} \times \Omega \rightarrow[0, \infty)$ be some distance function of the observation $S_{t}$ and the set of parameters $\theta_{i}$. For the work done throughout this paper, the $\theta_{i}$ correspond to the changing market phases. Then the vector $\Gamma_{t}=\left(\gamma_{1}(t), \ldots, \gamma_{K}(t)\right)$, called the affiliation vector, gives the probability of being in the $i$ th market phase at time $t$. For a given observation $S_{t}$ we want to solve the clustering problem

$$
\begin{equation*}
\sum_{i=1}^{K} \gamma_{i}(t) g\left(S_{t}, \theta_{i}\right) \rightarrow \min _{\Gamma_{t}, \Theta} \tag{2.1}
\end{equation*}
$$

for $\Theta=\left(\theta_{1}, \ldots, \theta_{K}\right)$. Moreover, as the $\gamma_{i}(t)$ shall be probabilities, they are subject to the constraint

$$
\begin{align*}
\sum_{i=1}^{K} \gamma_{i}(t)=1 & \forall t \in[0, T]  \tag{2.2}\\
\gamma_{i}(t) \geq 0 & \forall t \in[0, T], i=1, \ldots, K \tag{2.3}
\end{align*}
$$

As we do not want to solve this problem for a specific time $t$ only, but for the whole time interval $[0, T]$, we introduce the averaged clustering functional $L$ :

$$
\begin{equation*}
L(\Theta, \Gamma)=\int_{0}^{T} \sum_{i=1}^{K} \gamma_{i}(t) g\left(S_{t}, \theta_{i}\right) \rightarrow \min _{\Gamma, \Theta}, \tag{2.4}
\end{equation*}
$$

subject to the constraints $(2.2)-(2.3) .{ }^{1}$ As was demonstrated in $[16,17,14,15]$ there are practical difficulties in solving problem (2.4) numerically.

1. The problem is infinite dimensional since $\gamma_{i}(t)$ is an element of some (not yet specified) function space.

[^78]2. The problem is ill-posed in the definition of Hadamard [12], as neither the uniqueness of the solution nor its continuity w.r.t. the time series data is fulfilled in general. Because of the high number of parameters, additional regularity assumptions will be needed to identify the parameters, and to avoid problems like overfitting, underspecifying, and unstable estimates.
3. As long as $g$ is nonconvex, only locally optimal solutions can be found.

Problems 1 and 2 are addressed within sections 2.2 and 2.3 ; how to evade problem 3 is shown in sections 2.4 and 2.5 .
2.2. Persistent market phases. As shown in $[16,17,14,15]$, one of the possibilities for overcoming the above-mentioned problem is to impose some additional assumptions on $\Gamma$. As we look for hidden market phases, we do assume the process switching between phases is slow: When the market switches to a new state, it will stay there for some time. This suggests a new constraint on $\Gamma$ :

$$
\begin{equation*}
\int_{0}^{T}\left(\partial_{t} \gamma_{i}(t)\right)^{2} d t \leq \gamma_{i}^{\varepsilon}, \quad i=1, \ldots, K, \gamma_{i}^{\varepsilon} \in(0, \infty) \tag{2.5}
\end{equation*}
$$

While (2.5) should reduce the transitions between different states, it is hard to handle this inside the optimization. Instead, we thus expand our problem to the regularized form as suggested by Tikhonov for linear least-squares problems [32]:

$$
\begin{equation*}
L^{\varepsilon}(\Theta, \Gamma)=L(\Theta, \Gamma)+\varepsilon^{2} \sum_{i=1}^{K} \int_{0}^{T}\left(\partial_{t} \gamma_{i}(t)\right)^{2} d t . \tag{2.6}
\end{equation*}
$$

As was shown in [16], the parameter $\varepsilon^{2}$ can be used to control the persistence of the cluster states, by introducing a cost for short-term switching into another market phase.
2.3. Finite elements and discretization. Discrete time is, from a data point of view, not really an additional assumption, as market data, even if given intraday, is ticked data; thus it is discrete by nature. Time continuity of the data is a modeling idealization. This limits the function space from which $\Gamma$ is chosen, addressing problems 1 and 2 . Thus let $\mathcal{T}$ be a finite set of finite intervals on $[0, T]$ :

$$
\begin{equation*}
\mathcal{T}=\left\{0=t_{1}<t_{2}<\cdots<t_{N}=T\right\} . \tag{2.7}
\end{equation*}
$$

Then we can choose a finite element basis $\left\{v_{1}(t), \ldots, v_{N}(t)\right\}$ with

$$
\begin{array}{ll}
(2.8) & v_{1}(t) \neq 0 \text { for } t \in\left(t_{1}, t_{2}\right), \quad v_{1}(t)=0 \text { for } t \notin\left(t_{1}, t_{2}\right), \\
(2.9) & v_{k}(t) \neq 0 \text { for } t \in\left(t_{k-1}, t_{k+1}\right), \quad v_{k}(t)=0 \text { for } t \notin\left(t_{k-1}, t_{k+1}\right), \quad k=2, \ldots, N-1, \\
(2.10) & v_{N}(t) \neq 0 \text { for } t \in\left(t_{N-1}, t_{N}\right), \quad v_{N}(t)=0 \text { for } t \notin\left(t_{N-1}, t_{N}\right) .
\end{array}
$$

This allows us to write $\gamma_{i}$ as

$$
\begin{equation*}
\gamma_{i}=\sum_{k=1}^{N} \hat{\gamma}_{i}^{k} v_{k}+\delta_{i}^{N}, \tag{2.11}
\end{equation*}
$$

where $\delta^{N}=\max \left\{\delta_{i}^{N}, i=1, \ldots, K\right\}$ is the discretization error. Now (2.6) could be written as

$$
\begin{equation*}
L^{\varepsilon}(\Theta, \Gamma)=\sum_{i=1}^{K} \sum_{k=1}^{N} \hat{\gamma}_{i}^{k} \int_{0}^{T} v_{k}(t) g\left(S_{t}, \theta_{i}\right) d t+\varepsilon^{2} \sum_{i=1}^{K} \int_{0}^{T}\left(\sum_{k=1}^{N} \hat{\gamma}_{i}^{k} \partial_{t} v_{k}(t)\right)^{2} d t+O\left(\delta^{N}\right) \tag{2.12}
\end{equation*}
$$

Using the local support of $v_{k}$, we can define

$$
\begin{equation*}
\alpha\left(\theta_{i}\right)=\left(\int_{t_{1}}^{t_{2}} v_{1}(t) g\left(S_{t}, \theta_{i}\right) d t, \ldots, \int_{t_{N-1}}^{t_{N}} v_{N}(t) g\left(S_{t}, \theta_{i}\right) d t\right) \tag{2.13}
\end{equation*}
$$

and

$$
H=\left(\begin{array}{cccc}
\int_{t_{1}}^{t_{2}}\left(\partial_{t} v_{1}\right)^{2}(t) d t & \int_{t_{1}}^{t_{2}} \partial_{t} v_{1}(t) \partial_{t} v_{2}(t) d t & 0 & \ldots  \tag{2.14}\\
\int_{t_{1}}^{t_{2}} \partial_{t} v_{1}(t) \partial_{t} v_{2}(t) d t & \int_{t_{1}}^{t_{3}}\left(\partial_{t} v_{2}\right)^{2}(t) d t & \int_{t_{2}}^{t_{3}} \partial_{t} v_{2}(t) \partial_{t} v_{3}(t) d t & \ldots \\
0 & \ddots & \ddots & \ddots
\end{array}\right)
$$

to create the discrete clustering problem (see, for example, [16])

$$
\begin{equation*}
\widetilde{L^{\varepsilon}}(\Theta, \Gamma)=\sum_{i=1}^{K}\left(\alpha\left(\theta_{i}\right)^{T} \hat{\gamma}_{i}+\varepsilon^{2} \hat{\gamma}_{i}^{T} H \hat{\gamma}_{i}\right) \rightarrow \min _{\Gamma, \Theta} . \tag{2.15}
\end{equation*}
$$

Please note that up to now all assumptions made concern the smoothness of the clustering affiliation $\gamma_{i}(t)$ (see condition (2.5)) only and no probabilistic or statistical assumptions on the data $x_{t}$ are made. This abdication of including a priori assumptions on probabilistic properties of the hidden and observable processes is a major feature of the presented method.
2.4. The distance function. When working with financial data, one might end up with thousands of assets; e.g., the Wilshire 5000 index consists of 5000 stocks. W.r.t. this high dimensionality, one has to think about methods able to handle these data. The popular idea is to use the principal component analysis (PCA; see [33, 20]), where only a linear lowdimensional manifold of the data space is used to represent the original. The drawback here is the linearity of the manifold, as a globally optimal linear subspace might not exist. Thus let $T_{i} \in \mathbb{R}^{d \times n}$ for $i=1, \ldots, K$, let $n<d$ with $T_{i}^{T} T_{i}=$ Id be the projection matrices, and let $x_{t} \in \mathbb{R}^{d}$ be some time series deduced from $S_{t}$. Then we can define the distance function by

$$
\begin{equation*}
g\left(x_{t}, \theta_{i}\right)=\left\|\left(x_{t}-\mu_{i}\right)-T_{i} T_{i}^{T}\left(x_{t}-\mu_{i}\right)\right\|_{2}^{2}, \quad \theta_{i}=\left(\mu_{i}, T_{i}\right) \tag{2.16}
\end{equation*}
$$

This distance function measures the reduction error resulting from the projection of the $d$-dimensional data on the $n$-dimensional linear manifold $i$. The $x_{t}$ could be the returns and $T_{i}$ could be interpreted as the direction of maximal volatility of $x$, while $\mu_{i}$ describes the mean behavior [18]. We can now try to get better results by using the time-dependent function $\mu_{i}(t)$ instead of a constant one, e.g., approximating $\mu_{i}(t)$ with some regression model as in [17]. Therefore let $\varphi_{i}(t)$ be some basis of the functional space (e.g., monomials); then the distance function changes to

$$
\begin{equation*}
g\left(x_{t}, \theta_{i}\right)=\left\|\left(x_{t}-\sum_{j=1}^{\omega} \mu_{i}^{j} \varphi_{j}(t)\right)-T_{i} T_{i}^{T}\left(x_{t}-\sum_{j=1}^{\omega} \mu_{i}^{j} \varphi_{j}(t)\right)\right\|_{2}^{2}, \quad \theta_{i}=\left(\left(\mu_{i}^{1}, \ldots, \mu_{i}^{k}\right), T_{i}\right), \tag{2.17}
\end{equation*}
$$

where $\omega$ is the order of the trend. In fact, we can see $\sum_{j=1}^{\omega} \mu_{i}^{j} \varphi_{j}(t)$ as the smoothed trend, so this analysis can be interpreted as a detrending. Additionally we can estimate the covariance matrix of a clustered set $\left(x_{t}\right)$ of the time series for some $i$ by

$$
\begin{equation*}
\Sigma_{i}^{\bar{T}}=\frac{1}{\sum_{t \leq \bar{T}} \gamma_{i}(t)} \sum_{t \leq \bar{T}} \gamma_{i}(t)\left(x_{t}-\sum_{j=1}^{\omega} \mu_{i}^{j} \varphi_{j}(t)\right)\left(x_{t}-\sum_{j=1}^{\omega} \mu_{i}^{j} \varphi_{j}(t)\right)^{T} \tag{2.18}
\end{equation*}
$$

From a practical viewpoint, one should include an additional weighting further connecting the assets to their liquidity. However, the influence of liquidity on risk is not part of our analysis. Please note that the combination of PCA, trend estimation, and clustering is not just a direct chaining of both algorithms. The estimation of trends and covariance matrices is embedded in the minimization problem (2.15).
2.5. Calculating the parameters. Given the analyzed time series $x_{t}$ and a cluster affiliation $\Gamma$, we can now compute the cluster parameters $\mu_{i}^{j}$ and $T_{i}$. Therefore we have to solve for each $i=1, \ldots, K$ the problem

$$
\begin{align*}
\alpha\left(\theta_{i}\right)^{T} \hat{\gamma}_{i}+\varepsilon^{2} \hat{\gamma}_{i}^{T} H \hat{\gamma}_{i} & \rightarrow \min _{\theta_{i}}  \tag{2.19}\\
\Leftrightarrow \quad \alpha\left(\theta_{i}\right)^{T} \hat{\gamma}_{i} \quad & \rightarrow \min _{\theta_{i}} \tag{2.20}
\end{align*}
$$

Using the definition of $\alpha$ and $g$ we get

$$
\begin{equation*}
\sum_{k=1}^{N} \gamma_{i}^{k} \int_{t_{k-1}}^{t_{k+1}} v_{k}(t)\left\|\left(x_{t}-\sum_{j=1}^{\omega} \mu_{i}^{j} \varphi_{j}(t)\right)-T_{i} T_{i}^{T}\left(x_{t}-\sum_{j=1}^{\omega} \mu_{i}^{j} \varphi_{j}(t)\right)\right\|_{2}^{2} d t \rightarrow \min _{\mu_{i}, T_{i}} \tag{2.21}
\end{equation*}
$$

with $t_{0}=t_{1}$ and $t_{N+1}=t_{N}$. First-order conditions of optimality imply that partial derivatives w.r.t. $\mu_{i}^{j}$ and $T_{i}$ vanish; hence for $\mu$ we obtain that

$$
0=\sum_{k=1}^{N} \gamma_{i}^{k} \int_{t_{k-1}}^{t_{k+1}} v_{k}(t)\left(T_{i} T_{i}^{T} \varphi_{l}(t)-\varphi_{l}(t)\right)^{T}\left(x_{t}-T_{i} T_{i}^{T} x_{t}\right) d t
$$

$$
\begin{equation*}
-\sum_{j=1}^{\omega} \mu_{i}^{j} \sum_{k=1}^{N} \gamma_{i}^{k} \int_{t_{k-1}}^{t_{k+1}} v_{k}(t)\left(T_{i} T_{i}^{T} \varphi_{l}(t)-\varphi_{l}(t)\right)^{T}\left(\varphi_{j}(t)-T_{i} T_{i}^{T} \varphi_{j}(t)\right) d t \tag{2.22}
\end{equation*}
$$

holds, and an analogous equation is obtained for $T_{i}$. The latter yields that $T_{i}$ consists of the eigenvectors corresponding to the $n$ largest eigenvalues of the empirical covariance matrix of the values associated with state $i$. Now the problem (2.15) could be solved by subspace iteration; thus, starting with some arbitrary $\Gamma_{0}$, we solve iteratively

$$
\begin{equation*}
\Theta_{k}=\arg \min _{\Theta} \widetilde{L^{\varepsilon}}\left(\Theta, \Gamma_{k}\right) \tag{2.23}
\end{equation*}
$$

and

$$
\begin{equation*}
\Gamma_{k+1}=\arg \min _{\Gamma} \widetilde{L^{\varepsilon}}\left(\Theta_{k}, \Gamma\right) \tag{2.24}
\end{equation*}
$$

for $k=0,1, \ldots$. Although this algorithm is not guaranteed to converge to the globally optimum, at least a local minimum will be found. By performing this algorithm several times with different starting values $\Gamma_{0}$, we can approximate the global optimal parameters. This stepwise optimization allows us to exploit the structure of the problem. Separating these two steps allows us to split the overall optimization into two numerically conceivable steps performed subsequently until the overall optimization threshold is reached. These two steps are (i) a problem that is (in the PCA case) analytically solvable (2.23) and results in the efficient numerical solution of the eigenvalue problem; and (ii) a convex and sparse quadratic programming problem ((2.24) with (2.2)-(2.3)) [16]. Because of that, the proposed numerical scheme is numerically cheaper and more robust for high-dimensional applications than the standard "out of the box" optimization methods applied to the same problem.
3. Adapting portfolio optimization theory to futures. The usual definition of returns $R_{t}=\frac{S_{t+\Delta t-S_{t}}^{S_{t}}}{S_{t}}$ cannot be naively applied to futures data, as the price for a contract is not $S_{t}$. Instead, no price is charged for entering a futures contract, but an initial amount is paid into the margin account that will hold the accrued gains and losses. So, given an initial margin of $I \in \mathbb{R}^{d}$ and a portfolio $\pi$ in proportion of wealth, we define returns for futures over the period $t$ to $t+\Delta t$ as

$$
\begin{equation*}
R_{t}=\operatorname{sgn}\left(\pi_{t}\right) \frac{S_{t+\Delta t}-S_{t}}{I} \quad \text { (componentwise). } \tag{3.1}
\end{equation*}
$$

In contrast to portfolio optimization approaches for the stock market, it is an important feature of futures that long and short positions can be held equally well. But since the initial margin is dependent only on the absolute contract size, our portfolio constraints will change, and the absolute weights of the components should sum up to one:

$$
\begin{equation*}
\sum_{i=1}^{d}\left|\pi_{i}(t)\right|=1 \quad \forall t \in[0, T] . \tag{3.2}
\end{equation*}
$$

Given suitable estimates $\Sigma, \mu$ for covariances and means of returns, the portfolio problem is posed as

$$
\begin{equation*}
\pi^{T} \Sigma \pi \rightarrow \min _{\pi}, \quad \text { with }\|\pi\|_{1}=1, \pi^{T} \mu \geq C \tag{3.3}
\end{equation*}
$$

for a chosen target mean return $C$. Since this is not yet (as in classical mean variance optimization) a quadratic minimization problem with linear (or quadratic) constraints, we use a splitting into long positions and short positions, similar to the idea explained in [9] in the context of image processing, to get an extended parameter set

$$
\hat{\pi}=\left(\pi^{\text {long }}, \pi^{\text {short }}\right)^{T}, \quad \widehat{\Sigma}=\left(\begin{array}{cc}
\Sigma & -\Sigma  \tag{3.4}\\
-\Sigma & \Sigma
\end{array}\right), \quad \hat{\mu}=(\mu,-\mu)^{T}
$$

and we apply a penalty term. We define the extended problem by

$$
\hat{\pi}^{T}\left(\widehat{\Sigma}+\lambda\left(\begin{array}{cc}
0 & \mathbb{I}_{d}  \tag{3.5}\\
\mathbb{I}_{d} & 0
\end{array}\right)\right) \hat{\pi} \rightarrow \min _{\hat{\pi} \in \mathbb{R}^{2 d}} \quad \text { with } \hat{\pi}^{T} \hat{\mu} \geq C, \sum_{i=1}^{2 d} \hat{\pi}_{i}=1, \hat{\pi}_{i} \geq 0 \forall i
$$

with the regularization parameter $\lambda>0$. Please note that for sufficiently large $\lambda$ the part

$$
\lambda \hat{\pi}^{T}\left(\begin{array}{cc}
0 & \mathbb{I}_{d}  \tag{3.6}\\
\mathbb{I}_{d} & 0
\end{array}\right) \hat{\pi}=2 \lambda\left(\pi^{\text {long }}\right)^{T} \pi^{\text {short }}
$$

and the condition $\hat{\pi}_{i} \geq 0 \forall i$ guarantee that for each asset the position is either long or short, but not both at the same time. Simultaneously, the remainder of the extended problem is just

$$
\begin{equation*}
\hat{\pi}^{T} \widehat{\Sigma} \hat{\pi}=\left(\pi^{\text {long }}-\pi^{\text {short }}\right)^{T} \Sigma \pi^{\text {long }}-\left(\pi^{\text {long }}-\pi^{\text {short }}\right)^{T} \Sigma \pi^{\text {short }}=\pi \Sigma \pi \tag{3.7}
\end{equation*}
$$

and

$$
\begin{equation*}
\hat{\mu}^{T} \hat{\pi}=\mu^{T} \pi^{\text {long }}-\mu^{T} \pi^{\text {short }}=\mu^{T} \pi . \tag{3.8}
\end{equation*}
$$

4. Numerical examples. Throughout the first part of this section, we will make use of a simulated time series. These time series have a length of 1000 with a change point at time $t_{C}=500$. All of the data points are distributed independently according to $\mathcal{N}(0, \Sigma)$, where

$$
\Sigma(t)=\alpha^{T}(t)\left[\begin{array}{cc}
0.85 & 0.4  \tag{4.1}\\
0.4 & 0.2
\end{array}\right] \alpha(t), \quad \alpha(t)= \begin{cases}\mathbb{I}_{2}, & t \leq 500 \\
{\left[\begin{array}{cc}
\cos (\rho) & -\sin (\rho) \\
\sin (\rho) & \cos (\rho)
\end{array}\right],} & t>500\end{cases}
$$

As one can see, after the change point at $t_{C}=500$, the covariance matrix is just rotated by some rotation angle $\rho$.
4.1. Covariance for small samples. As known from the literature, for small sample sizes, the maximum likelihood estimator $\frac{1}{k} \sum_{t=t_{1}}^{t_{k}} x_{t} x_{t}^{T}$ is not robust enough; see, e.g., [30, 4]. This will lead to numerical biassing when it comes to portfolio optimization [28, 25, 26]. We will use the Tyler M-estimator (see, e.g., $[34,8]$ ) instead, which is defined by

$$
\begin{equation*}
\hat{\Sigma}^{k}=\frac{d}{k} \sum_{t=t_{1}}^{t_{k}} \frac{x_{t} x_{t}^{T}}{x_{t}^{T}\left(\hat{\Sigma}^{k}\right)^{-1} x_{t}} . \tag{4.2}
\end{equation*}
$$

Although the solution $\hat{\Sigma}^{k}$ to (4.2) is unique only up to a multiplicative constant, this is not a problem for the clustering method (2.15), as neither the result of the minimization nor the eigenvectors (for PCA) is dependent on this constant. For Figure 1, the first 100 points of 100 time series were used (using the distribution stated above) and for every time step $t \in\{2, \ldots, 100\}$ the maximum likelihood estimator and the Tyler M-estimator were calculated. Then the largest singular value of the difference $(\hat{\Sigma}-\Sigma)$ was used as a distance measure, and the average and standard deviation were calculated and plotted. As one can see from this example, the standard deviation of the Tyler M-estimator is much smaller than the one for the maximum likelihood estimator (although both converge to the same limit).
4.2. Number of clusters. In realistic applications, the optimal number of metastable clusters is usually a priori unknown. Nevertheless, one can find out the number of statistically distinguishable clusters (see [16]). Therefore we start the clustering process with a rather high number of clusters, estimate the cluster parameters, and get the error bounds for these


Figure 1. Robustness of the maximum likelihood estimator and the Tyler M-estimator for short normally distributed time series. Parameters as in (4.1).
parameters by bootstrapping the elements of the cluster. If the parameter confidence sets of different clusters overlap, the statistical difference between those clusters is not high enough; thus the number of clusters should be reduced and the process should be started over again [22]. Using this technique for different $\varepsilon$ leads to a picture like Figure 2. The decrease in the number of clusters is due to the fact that for the increasing $\varepsilon$ the algorithm favors metastable clusters and merges the rapidly mixing ones together. Thus, while increasing $\varepsilon$, similar clusters will get merged into one. The plateau results from the high difference between the last two remaining clusters. For a real data set we will observe similar behavior when we look for the number of clusters.
4.3. Detecting a change point. Another important problem is to find the actual position of the change points. The number of detected change points depends on $\varepsilon$ as one can see in Figure 3. However, a high number of change points usually indicates some rapidly mixing clusters that could be united to a metastable cluster (e.g., using half the data in Figure 3). Decreasing the number of clusters and increasing $\varepsilon$ might help to solve this problem. On the other hand, a high value for $\varepsilon$ will lead to a longer delay between occurrence and online detection of the change point; this can be seen in Figure 4. The same problem appears when the rotation angle becomes too small; see Figure 5.


Figure 2. Number of distinguishable clusters depending on $\varepsilon^{2}$. Parameters as in (4.1); $\rho=90^{\circ}$.
4.4. Looking at real data. For the calculations in this paper, we used the futures prices from 2005-2008 for wheat and oil futures (four times to maturity each, quarterly for wheat and monthly for oil, daily closings) from http://www.kcbt.com/historical_data.asp (Kansas City Board of Trade) and http://tonto.eia.doe.gov/dnav/pet/pet_pri_fut_s1_d.htm (U.S. Department of Energy). For both commodities the four futures with the lowest time to maturity were taken into account. Applying the introduced clustering algorithm to these data gives the result shown in Figure 6. Implied by this picture, an optimal number of metastable clusters to use should be two. Then the cluster allocation for the whole time series (see Figure 7) produces a kind of seasonal cycle that could be due to a harvest or heating period. The phases differ in the following points:

- The phase shown in Figure 7, and thus phase 1, is characterized by lesser overall trend in the price than phase 2.
- The general direction of the maximum volatility (described by the projection operator $T_{i}$ from (2.15)) in both phases is similar; the first eigenvectors of the covariance matrices are nearly equal, and the second eigenvectors span an angle of approximately 5 degrees.


Figure 3. Number of detected change points depending on $\varepsilon^{2}$ for half of the (artificial) data and the complete data set. Parameters as in (4.1); $K=2, \rho=90^{\circ}$.

The effective investment for each commodity (as the sum of the portfolio weights over the different maturities) for the phase-dependent portfolio can be seen in Figure 8. The variance of the positions can be considered as a result of the parameter estimation error. Additionally, Figure 9 shows box-whisker plots of the relative notionals, defined by

$$
\begin{equation*}
\frac{\pi_{i}(t)}{\sum_{i \in I_{u}}\left|\pi_{i}(t)\right|}, \tag{4.3}
\end{equation*}
$$

where $I_{u}$ is the set of indices belonging to one commodity. The online version of the cluster affiliations is presented; thus the affiliations at time $t$ are calculated "on the fly" using only the information $\forall s, s<t$ during the portfolio optimization. The high variance in the resulting portfolio weights (resp., the relative notionals) is a direct result of nonrobust parameter estimates. For practical purposes, further steps in creating robust estimates or optimizing robustly w.r.t. uncertain estimates should be considered in further research.
4.5. Comparing portfolios. When looking at the interest rate for U.S. treasury bonds, an investment in January 2005 would have given a yearly interest rate of approximately $3.4 \%$. Thus after four years we have a gain of $14.31 \%$ (that is a daily rate of 1.415 bps ). We use this rate as a target rate $C$ (as in (3.5)). As stated earlier, we compare our algorithm with the Markowitz approach, the FlexM algorithm from [24], and a simple $\frac{1}{n}$-portfolio, where half the assets are bought and the other half sold. For the cluster-based approach, the basis $\varphi$ with polynomials up to order two was taken. The clustering was done w.r.t. $\mu$ and $T$. The FlexM algorithm provided by the authors of [24] was used utilizing the parameter demean; thus the estimation and subtraction of the mean from the data prior to the portfolio optimization is done by the FlexM function. The $\frac{1}{n}$-portfolio was built by buying for $\frac{1}{8}$ of the wealth the two

Detection time of change point depending on $\varepsilon^{2}$


Figure 4. Relative error in detection time of the single change point depending on $\varepsilon^{2}$. Parameters as in (4.1); $K=2, \rho=90^{\circ}$.
long maturity futures of each commodity and selling the remaining two for each commodity equally. Additionally, the Markowitz approach was used by estimating mean and covariance from the whole time series up to the actual point. All algorithms were supplied only with information available at time $t$, so no future knowledge is included. In each case, the portfolio was recalculated and/or rebalanced every 10 business days starting with day 10 of the time series. As one can see in Figure 10, all algorithms fall short of the target $14.31 \%$ mean return. While the $\frac{1}{n}$-portfolio is doing better (in terms of return rate) than the other algorithms, this comes with a higher risk, as one can see in Figure 11. We define the "intrinsic risk estimate" for portfolio returns from $t$ to $t+\Delta t$ by

$$
\begin{equation*}
\mathcal{R}_{t}=\sqrt{\pi_{t}^{T} \operatorname{Cov}\left(R_{t}\right) \pi_{t}}, \tag{4.4}
\end{equation*}
$$

where the empirical returns $R_{t}$ are given by (3.1), and the $\operatorname{Cov}\left(R_{t}\right)$ is an estimate for the covariance matrix of the data (based upon data only up to $t$ ), which is given by

- a linear combination of the empirical cluster covariance matrices given by (2.18) for the cluster-based algorithm, and thus

Detection time of change point depending on $\rho$


Figure 5. Relative error in detection time of the single change point depending on $\rho$. Parameters as in (4.1); $K=2, \varepsilon^{2}=1$.

$$
\begin{equation*}
\operatorname{Cov}\left(R_{t}\right)=\sum_{i=1}^{K} \gamma_{i}(t) \Sigma_{i}^{t} ; \tag{4.5}
\end{equation*}
$$

- the $\operatorname{GARCH}(1,1)$ estimate for the FlexM algorithm (see [24]);
- the standard sample covariance matrix for the Markowitz approach and the $\frac{1}{n}$-portfolio. Although those intrinsic risks shown in Figure 11 are estimates of the instantaneous variance of the portfolio returns, they are not directly comparable, as the measures differ from each other and the underlying assumptions are different, so in Figure 12 risk is meant in the sense of the standard deviation of the daily returns of the portfolios from the empirical mean; the cluster-based approach still produces the strategy with the lowest risk. And even if we use the deviation from the target mean return $C$

$$
\begin{equation*}
\sqrt{\frac{1}{k-1} \sum_{t=t_{1}}^{t_{k-1}}\left(\pi_{t}^{T} R_{t}-C\right)^{2}} \tag{4.6}
\end{equation*}
$$

as a risk measure (see the results in Figure 13), the cluster-based approach reaches the best result. The similarity of this risk measure and the standard deviation is easy to see: If the vari-


Figure 6. Number of distinguishable clusters depending on $\varepsilon^{2}$. Data set: Prices of oil and wheat futures from 2005-2008.
ance minimal portfolio does not satisfy the condition $\hat{\pi}_{\min }^{T} \hat{\mu} \geq C$, i.e., the target mean return is not reached, due to the convexity of the problem the solution of (3.5) will satisfy $\hat{\pi}^{T} \hat{\mu}=C$. This guarantees that for sufficiently large $C$ the expected mean is $C$, and thus (4.6) would give the standard deviation. The advantage of (4.6) over the empirical standard deviation is the fact that the deviation from the target mean return does not depend on the empirical expectation of the portfolio returns. Indeed, as stated in Table 1, the empirical standard deviation of the portfolio returns and the empirical deviation from the target mean return are quite close to each other. The cluster-based approach reaches $69.85 \%$ of the performance of the $\frac{1}{n}$-portfolio, while it takes only $40.33 \%$ of the risk (in the sense of the empirical standard deviation). Here, rate of return denotes the total return minus one and average intrinsic is defined by

$$
\begin{equation*}
\frac{1}{N-1} \sum_{t=t_{1}}^{t_{N-1}} \mathcal{R}_{t} \tag{4.7}
\end{equation*}
$$

For different target returns the results are similar, although the cluster-based portfolio is less sensible to the target return rate (see Figures 14 and 15).

## Cluster allocation, Prices of oil and wheat futures 2005-2008



Figure 7. Cluster allocation for time series of futures; $K=2, \varepsilon^{2}=1$. Data set: Prices of oil and wheat futures from 2005-2008.
4.6. Adaption for notional neutral investment. A common relevant strategy of portfolio managers is so-called market neutral investment. The aim is to bet only from views on relative changes between asset prices but remain neutral w.r.t. overall market moves. To avoid assumptions about an appropriated market indicator or a market portfolio that characterizes the whole market, a more pragmatic approach is chosen: We define notional neutral investments as investments where long and short positions are balanced such that equal notionals are invested long and short for each security class. This is done by an additional constraint to ensure that the accrued notional of contracts within each class $u$ (here wheat and oil) is equal to zero:

$$
\begin{equation*}
\sum_{i \in I_{u}} \pi_{i}=0 \quad \forall u, \tag{4.8}
\end{equation*}
$$

where $I_{u}$ is an index set combining all assets for each selected class of futures within which accrued notionals are requested to be zero. For the extended problem, this transforms to

$$
\begin{equation*}
\sum_{i \in I_{u}}\left(\hat{\pi}_{i}-\hat{\pi}_{i+d}\right)=0 \quad \forall u . \tag{4.9}
\end{equation*}
$$

Effective Investments


Figure 8. Effective investments per commodity as proportion of wealth; $K=2, \varepsilon^{2}=1$. Data set: Prices of oil and wheat futures from 2005-2008.

Thus the notional neutral version of the problem is
(4.10)
$\hat{\pi}^{T}\left(\widehat{\Sigma}+\lambda\left(\begin{array}{cc}0 & \mathbb{I}_{d} \\ \mathbb{I}_{d} & 0\end{array}\right)\right) \hat{\pi} \rightarrow \min _{\tilde{\pi} \in \mathbb{R}^{2} d^{2}} \quad$ with $\hat{\pi}^{T} \hat{\mu} \geq C, \sum_{i=1}^{2 d} \hat{\pi}_{i}=1, \hat{\pi}_{i} \geq 0 \forall i, \sum_{i \in I_{u}}\left(\hat{\pi}_{i}-\hat{\pi}_{i+d}\right)=0$.
The additional constraint was used for all algorithms (except the $\frac{1}{n}$-portfolio). As before, the realized return for each algorithm falls short of the mean target return (see Figure 16), but the presented cluster-based method produces the result with the least risk (see Figures 17, 18, and 19) and, for this data, even the best return rate, next to the $\frac{1}{n}$-portfolio, where no risk minimization is done.
5. Conclusion. We adapted and expanded a method for simultaneous clustering, dimension reduction, and metastability analysis of multidimensional time series to financial data. A main advantage of the method (compared to standard HMM/GMM methods of phase identi-


Figure 9. Spread of the relative notionals per commodity and phase; $K=2, \varepsilon^{2}=1$. Data set: Prices of oil and wheat futures from 2005-2008.
fication $[13,23])$ is that no a priori probabilistic assumptions about the hidden and observed market processes are necessary in the context of the presented method. Furthermore, section 4 indicates that the method is suitable for finding a low-dimensional representation of the data and estimating mean and variance adaptively while detecting market phases, delivering information that can be further used for portfolio optimization. The presented framework allows one to make use of fast and numerically robust FEM solvers (developed for numerical solution of partial differential equations) in a new context of computational time series analysis.

The general feasibility of identification of market phases was demonstrated, and it was shown that an expansion of classical portfolio theory utilizing market phase detection can be used to further decrease the intrinsic risk. Additionally, the behavior of the change point detection when applied online to a data stream was demonstrated to be robust enough to qualify for practical purposes. As a standard stationary risk measure, the standard deviation of portfolio returns was also used to compare different portfolio optimization strategies on real

Total return evolution of portfolios


Figure 10. Value evolution of different portfolio-optimization algorithms.
financial data. It was demonstrated that application of the presented numerical scheme based on the identification of market phases can decrease the risk of resulting portfolios in terms of this risk measure. Strictly speaking, the empirical standard deviation of portfolio returns is not an equitable risk measure, in that it assumes the underlying data to be stationary. If we assume the underlying market process to be nonstationary (as is done in the context of the presented numerical approach), the evaluation and comparison of risks should be done in a measure that reflects this. Numerical studies of robust portfolio optimization methods and empirical quantification of risk by methods that do not rely on stationarity assumptions are a matter of future research.

Having demonstrated the applicability of the presented computational scheme to a realistic financial data example, a more extensive online data study is clearly needed for validation. Also, transaction costs should be included and different risk measures should be taken into account. In contrast to climatological [14, 17, 15, 18] or biophysical applications [19], repeating phases seem to be rare in financial time series, and additional time weighting (e.g., exponential decay) should be further investigated.

Risk (intrinsic) evolution of portfolios


Figure 11. Intrinsic risk (see (4.4)) evolution of different portfolio-optimization algorithms.
Risk (standard deviation) evolution of portfolios


Figure 12. Risk (empirical standard deviation) evolution of different portfolio-optimization algorithms.


Figure 13. Risk (deviation from target mean return) evolution of different portfolio-optimization algorithms.

Table 1
Rate of return and different risk measures.

| Algorithm | Rate of return | Averaged intrinsic <br> risk | Standard deviation <br> of returns | Deviation from tar- <br> get mean return |
| :--- | :---: | :---: | :---: | :---: |
| Cluster-based | $4.01 \%$ | $1.416410^{-3}$ | $2.461310^{-3}$ | $2.461910^{-3}$ |
| FlexM | $-0.55 \%$ | $7.511310^{-3}$ | $7.896810^{-3}$ | $7.893510^{-3}$ |
| Markowitz | $0.30 \%$ | $1.518010^{-3}$ | $3.932410^{-3}$ | $3.932510^{-3}$ |
| $\frac{1}{n}$-portfolio | $5.74 \%$ | $3.702210^{-3}$ | $6.103310^{-3}$ | $6.100310^{-3}$ |



Figure 14. Total return depending on the yearly target return rate $C$ in (3.5).


Figure 15. Risk (empirical standard deviation) depending on the yearly target return rate $C$ in (3.5).


Figure 16. Value evolution of different notional neutral portfolio-optimization algorithms.

Risk (intrinsic) evolution of portfolios


Figure 17. Intrinsic risk (4.4) evolution of different notional neutral portfolio-optimization algorithms.

Risk (standard deviation) evolution of notional neutral portfolios


Figure 18. Risk (empirical standard deviation) evolution of different notional neutral portfolio-optimization algorithms.


Figure 19. Risk (deviation from target mean return) evolution of different notional neutral portfoliooptimization algorithms.

## REFERENCES

[1] A. Akansu and R. Haddad, Multiresolution Signal Decomposition: Transforms, Subbands, and Wavelets, Academic Press, San Diego, CA, 1992.
[2] T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, J. Econometrics, 31 (1986), pp. 307-327.
[3] U. Çelikyurt and S. Özekicı, Multiperiod portfolio optimization models in stochastic markets using the mean-variance approach, European J. Oper. Res., 179 (2007), pp. 186-202.
[4] V. Demiguel and F. Nogales, Portfolio selection with robust estimation, Oper. Res., 57 (2009), pp. 560-577.
[5] D. Duffie, Futures Markets, Prentice-Hall, Englewood Cliffs, NJ, 1989.
[6] R. Engle, Autoregressive conditional heteroscedasticity with estimates of variance of United Kingdom inflation, Econometrica, 50 (1982), pp. 987-1008.
[7] F. Fabozzi, P. Kolm, D. Pachamanova, and S. Focardi, Robust Portfolio Optimization and Management, John Wiley \& Sons, New York, 2007.
[8] G. Frahm and U. Jaekel, Tyler's M-estimator, random matrix theory, and generalized elliptical distributions with applications to finance, Phys. A, submitted.
[9] H. Fu, M. K. Ng, M. Nikolova, and J. L. Barlow, Efficient minimization methods of mixed l2-l1 and l1-l1 norms for image restoration, SIAM J. Sci. Comput., 27 (2006), pp. 1881-1902.
[10] L. Garlappi, R. Uppal, and T. Wang, Portfolio selection with parameter and model uncertainty: A multi-prior approach, Rev. Financ. Stud., 20 (2007), pp. 41-81.
[11] D. Goldfarb and G. Iyengar, Robust portfolio selection problems, Math. Oper. Res., 28 (2003), pp. 1-38.
[12] J. Hadamard, Sur les problèmes aux dérivées partielles et leur signification physique, Bull. Univ. Princeton 13, Princeton University, Princeton, NJ, 1902, pp. 49-52.
[13] J. Hamilton, A new approach to the economic analysis of nonstationary time series and the business cycle, Econometrica, 57 (1989), pp. 357-384.
[14] I. Horenko, On simultaneous data-based dimension reduction and hidden phase identification, J. Atmospheric Sci., 65 (2008), pp. 1941-1954.
[15] I. Horenko, On robust estimation of low-frequency variability trends in discrete Markovian sequences of atmospheric circulation patterns, J. Atmospheric Sci., 66 (2009), pp. 2059-2072.
[16] I. Horenko, Finite element approach to clustering of multidimensional time series, SIAM J. Sci. Comput., 32 (2010), pp. 62-83.
[17] I. Horenko, On clustering of non-stationary meteorological time series, Dyn. Atmos. Oceans, 49 (2010), pp. 164-187.
[18] I. Horenko, R. Klein, S. Dolaptchiev, and Ch. Schütte, Automated generation of reduced stochastic weather models I: Simultaneous dimension and model reduction for time series analysis, Multiscale Model. Simul., 6 (2008), pp. 1125-1145.
[19] I. Horenko and Ch. Schütte, Likelihood-based estimation of multidimensional Langevin models and its application to biomolecular dynamics, Multiscale Model. Simul., 7 (2008), pp. 731-773.
[20] I. Jolliffe, Principal Component Analysis, Springer, New York, 2002.
[21] R. Kalman, A new approach to linear filtering and prediction problems, Trans. ASME J. Basic Eng., 82 (1960), pp. 35-45.
[22] B. Kedem and K. Fokianos, Regression Models for Time Series Analysis, Wiley Ser. Probab. Stat., John Wiley \& Sons, New York, 2002.
[23] H. Krolzig, Predicting Markov-switching vector autoregressive processes, J. Forecasting, to appear.
[24] O. Ledoit, P. Santa-Clara, and M. Wolf, Flexible multivariate GARCH modeling with an application to international stock markets, Rev. Econ. Statist., 3 (2003), pp. 735-747.
[25] O. Ledoit and M. Wolf, Improved estimation of the covariance matrix of stock returns with an application to portfolio selection, J. Empirical Finance, 10 (2003), pp. 603-621.
[26] O. Ledoit and M. Wolf, Honey, I shrunk the sample covariance matrix, J. Portfolio Management, 30 (2004), pp. 110-119.
[27] H. Markowitz, Portfolio selection, J. Finance, 7 (1952), pp. 77-91.
[28] R. Michaud, The Markowitz optimization enigma: Is optimized optimal?, Financial Analysts Journal, 45 (1989), pp. 31-42.
[29] T. Rolski, H. Schmidli, V. Schmidt, and J. Teugels, Stochastic Processes for Insurance and Finance, Wiley Ser. Probab. Stat., John Wiley \& Sons, Chichester, 1999.
[30] J. Schäfer and K. Strimmer, A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, Stat. Appl. Genet. Mol. Biol., 4 (2005), article 32.
[31] W. Sharpe, Capital asset prices: A theory of market equilibrium under conditions of risk, J. Finance, 9 (1964), pp. 425-442.
[32] A. Tikhonov, On the stability of inverse problems, Dokl. Akad. Nauk SSSR, 39 (1943), pp. 195-198.
[33] R. S. Tsay, Analysis of Financial Time Series, John Wiley \& Sons, New York, 2005.
[34] D. E. Tyler, A distribution-free M-estimator of multivariate scatter, Ann. Statist., 15 (1987), pp. 223251.
[35] R. Welsch and X. Zhou, Application of robust statistics to asset allocation models, REVSTAT, 5 (2007), pp. 97-114.

# Trend Following Trading under a Regime Switching Model* 

M. Dai ${ }^{\dagger}$, Q. Zhang ${ }^{\ddagger}$, and Q. J. Zhu ${ }^{\S}$


#### Abstract

This paper is concerned with the optimality of a trend following trading rule. The idea is to catch a bull market at its early stage, ride the trend, and liquidate the position at the first evidence of the subsequent bear market. We characterize the bull and bear phases of the markets mathematically using the conditional probabilities of the bull market given the up to date stock prices. The optimal buying and selling times are given in terms of a sequence of stopping times determined by two threshold curves. Numerical experiments are conducted to validate the theoretical results and demonstrate how they perform in a marketplace.


Key words. optimal stopping time, regime switching model, Wonham filter, trend following trading rule
AMS subject classifications. 91G80, 93E11, 93E20
DOI. 10.1137/090770552

1. Introduction. Trading in organized exchanges has increasingly become an integrated part of our life. Big moves of market indices of major stock exchanges all over the world are often the headlines of news media. By and large, active market participants can be classified into two groups according to their trading strategies: those who trade contra-trend and those who follow the trend. On the other hand, there are also passive market participants who simply buy and hold for a long period of time (often indirectly through mutual funds). Within each group of strategies there are numerous technical methods. Much effort has been devoted to theoretical analysis of these strategies.

Using optimal stopping time to study optimal exit strategy for stock holdings has become the standard textbook method. For example, Øksendal [24, Examples 10.2.2 and 10.4.2] considered optimal exit strategy for stocks whose price dynamics were modeled by a geometric Brownian motion. To maximize an expected return discounted by the risk-free interest rate, the analysis in [24] showed that if the drift of the geometric Brownian motion was not high enough in comparison to the discount of interest rate, then one should sell at a given threshold. Although the model of a single geometric Brownian motion with a constant drift was somewhat too simplistic, this result well illustrated the flaw of the so-called buy and hold strategy, which worked only in limited situations. Stock selling rules under more realistic models have gained increasing attention. For example, Zhang [31] considered a selling rule determined by two

[^79]threshold levels, a target price and a stop-loss limit. Under a regime switching model, optimal threshold levels were obtained by solving a set of two-point boundary value problems. In Guo and Zhang [13], the results of Øksendal [24] were extended to incorporate a model with regime switching. In addition to these analytical results, various mathematical tools have been developed to compute these threshold levels. For example, a stochastic approximation technique was used in Yin, Liu, and Zhang [29]; a linear programming approach was developed in Helmes [14]; and the fast Fourier transform was used in Liu, Yin, and Zhang [20]. See also Beibel and Lerche [1] for a different approach to stopping time problems. Furthermore, consideration of capital gain taxes and transaction costs in connection with selling can be found in Cadenillas and Pliska [2], Constantinides [3], and Dammon and Spatt [8], among others.

Recently, there has been an increasing volume of literature concerning trading rules that involve both buying and selling. For instance, Zhang and Zhang [30] studied the optimal trading strategy in a mean reverting market, which validated a well-known contra-trend trading method. In particular, they established two threshold prices (buy and sell) that maximized overall discounted return if one traded at those prices. In addition to the results obtained in [30] along this line of research, an investment capacity expansion/reduction problem was considered in Merhi and Zervos [22]. Under a geometric Brownian motion market model, the authors used the dynamic programming approach and obtained an explicit solution to the singular control problem. A more general diffusion market model was treated by Løkka and Zervos [21] in connection with an optimal investment capacity adjustment problem. More recently, Johnson and Zervos [16] studied an optimal timing of investment problem under a general diffusion market model. The objective was to maximize the expected cash flow by choosing when to enter an investment and when to exit the investment. An explicit analytic solution was obtained in [16].

However, a theoretical justification of trend following trading methods is missing despite the fact that they are widely used among professional traders (see, e.g., [26]). It is the purpose of this paper to fill this void. We adopt a finite horizon regime switching model for the stock price dynamics. In this model the price of the stock follows a geometric Brownian motion whose drift switches between two different regimes representing the up trend (bull market) and down trend (bear market), respectively, and the exact switching times between the different trends are not directly observable as in the real markets. We model the switching as an unobservable Markov chain. Our trading decisions are based on current information represented by both the stock price and the historical information with the probability in the bull phase conditioning to all available historical price levels as a proxy. Assuming trading one share with a fixed percentage transaction cost, we show that the strategy that optimizes the discounted expected return is a simple implementable trend following system. This strategy is characterized by two time-dependent thresholds for the conditional probability in a bull regime signaling buy and sell, respectively. The main advantage of this approach is that the conditional probability in a bull market can be obtained directly using actual historical stock price data through a differential equation.

The derivation of this result involves a number of different technical tools. One of the main difficulties in handling the regime switching model is that the Markov regime switching process is unobservable. Following Rishel and Helmes [25] we use the optimal nonlinear
filtering technique (see, e.g., $[18,28]$ ) regarding the conditional probability in a bull regime as an observation process. Combining with the stock price process represented in terms of this observing process, we obtain an optimal stopping problem with complete observation. Our model involves possibly infinitely many buy and sell operations represented by sequences of stopping times, and it is not a standard stopping time problem. As in Zhang and Zhang [30] we introduce two optimal value functions that correspond to the starting net position being either flat or long. Using a dynamic programming approach, we can formally derive a system of two variational inequalities. A verification theorem justifies that the solutions to these variational inequalities are indeed the optimal value functions. It is interesting that we can show that this system of variational inequalities leads to a double obstacle problem satisfied by the difference of the two value functions. Since the solution and properties of double obstacle problems are well understood, this conversion simplifies the analysis of our problem considerably. An accompanying numerical procedure is also established to determine the thresholds involved in our optimal trend following strategy.

Numerical experiments have been conducted for a simple trend following trading strategy that approximates the optimal one. We test our strategy using both simulation and actual market data for the NASDAQ, SP500, and DJIA indices. Our trend following trading strategy outperforms the buy and hold strategy with a huge advantage in simulated trading. This strategy also significantly prevails over the buy and hold strategy when tested with the real historical data for the NASDAQ, SP500, and DJIA indices.

The rest of this paper is arranged as follows. We formulate our problem and present its theoretical solutions in the next section. Numerical results for optimal trading strategy are presented in section 3 . We conduct extensive simulations and tests on market data in section 4 and conclude in section 5 . Details of data and results related to simulations and market tests are collected in the appendix.
2. Problem formulation. Let $S_{r}$ denote the stock price at time $r$ satisfying the equation

$$
d S_{r}=S_{r}\left[\mu\left(\alpha_{r}\right) d r+\sigma d B_{r}\right], \quad S_{t}=S, \quad t \leq r \leq T<\infty
$$

where $\mu(i)=\mu_{i}, i=1,2$, are the expected return rates, $\alpha_{r} \in\{1,2\}$ is a two-state Markov chain, $\sigma>0$ is the volatility, $B_{r}$ is a standard Brownian motion, and $T$ is a finite time.

The process $\alpha_{r}$ represents the market mode at each time $r$ : $\alpha_{r}=1$ indicates a bull market and $\alpha_{r}=2$ a bear market. Naturally, we assume $\mu_{1}>\mu_{2}$. Let

$$
Q=\left(\begin{array}{cc}
-\lambda_{1} & \lambda_{1} \\
\lambda_{2} & -\lambda_{2}
\end{array}\right) \quad\left(\lambda_{1}>0, \lambda_{2}>0\right)
$$

denote the generator of $\alpha_{r}$. We assume that $\left\{\alpha_{r}\right\}$ and $\left\{B_{r}\right\}$ are independent.
Let

$$
t \leq \tau_{1} \leq v_{1} \leq \tau_{2} \leq v_{2} \leq \cdots \leq T \text { a.s. }
$$

denote a sequence of stopping times. Note that one may construct a sequence of stopping times satisfying the above inequalities from any monotone sequence of stopping times truncated at time $T$. A buying decision is made at $\tau_{n}$ and a selling decision at $v_{n}, n=1,2, \ldots$.

We consider the case that the net position at any time can be either flat (no stock holding) or long (with one share of stock holding). Let $i=0,1$ denote the initial net position. If initially
the net position is long $(i=1)$, then one should sell the stock before acquiring any share. The corresponding sequence of stopping times is denoted by $\Lambda_{1}=\left(v_{1}, \tau_{2}, v_{2}, \tau_{3}, \ldots\right)$. Likewise, if initially the net position is flat $(i=0)$, then one should first buy a stock before selling any shares. The corresponding sequence of stopping times is denoted by $\Lambda_{0}=\left(\tau_{1}, v_{1}, \tau_{2}, v_{2}, \ldots\right)$.

Let $0<K<1$ denote the percentage of slippage (or commission) per transaction. Given the initial stock price $S_{t}=S$, initial market trend $\alpha_{t}=\alpha \in\{1,2\}$, and initial net position $i=0,1$, the reward functions of the decision sequences, $\Lambda_{0}$ and $\Lambda_{1}$, are given as follows:

$$
J_{i}\left(S, \alpha, t, \Lambda_{i}\right)= \begin{cases}E_{t}\left\{\sum_{n=1}^{\infty}\left[e^{-\rho\left(v_{n}-t\right)} S_{v_{n}}(1-K)-e^{-\rho\left(\tau_{n}-t\right)} S_{\tau_{n}}(1+K)\right] I_{\left\{\tau_{n}<T\right\}}\right\} & \text { if } i=0, \\ E_{t}\left\{e^{-\rho\left(v_{1}-t\right)} S_{v_{1}}(1-K)\right. \\ \left.\quad+\sum_{n=2}^{\infty}\left[e^{-\rho\left(v_{n}-t\right)} S_{v_{n}}(1-K)-e^{-\rho\left(\tau_{n}-t\right)} S_{\tau_{n}}(1+K)\right] I_{\left\{\tau_{n}<T\right\}}\right\} & \text { if } i=1,\end{cases}
$$

where $\rho>0$ is the discount factor. Here, term $E \sum_{n=1}^{\infty} \xi_{n}$ is interpreted as $\lim \sup _{N \rightarrow \infty} E \sum_{n=1}^{N} \xi_{n}$ for random variables $\xi_{n}$. Our goal is to maximize the reward function.

To exclude trivial cases, ${ }^{1}$ we always assume

$$
\mu_{2}<\rho<\mu_{1}
$$

Remark 1. Note that the indicator function $I_{\left\{\tau_{n}<T\right\}}$ is used in the definition of the reward functions $J_{i}$. This is to ensure that if the last buy order is entered at $t=\tau_{n}$, then the position will be sold at $v_{n} \leq T$.

The indicator function $I$ confines the effective part of the sum to a finite time horizon so that the reward functions are bounded above.

Note that only the stock price $S_{r}$ is observable at time $r$ in the marketplace. The market trend $\alpha_{r}$ is not directly available. Thus, it is necessary to convert the problem into a completely observable one. One way to accomplish this is to use the Wonham filter [28]; see also [18, 32] for recent development and the references therein in connection with Wonham filters.

Let $p_{r}=P\left(\alpha_{r}=1 \mid \mathcal{S}_{r}\right)$ denote the conditional probability of $\alpha_{r}=1$ (bull market) given the filtration $\mathcal{S}_{r}=\sigma\left\{S_{u}: 0 \leq u \leq r\right\}$. Then we can show (see Wonham [28]) that $p_{r}$ satisfies the following SDE:

$$
d p_{r}=\left[-\left(\lambda_{1}+\lambda_{2}\right) p_{r}+\lambda_{2}\right] d r+\frac{\left(\mu_{1}-\mu_{2}\right) p_{r}\left(1-p_{r}\right)}{\sigma} d \widehat{B}_{r},
$$

where $\widehat{B}_{r}$ is the innovation process (a standard Brownian motion; see, e.g., Øksendal [24]) given by

$$
d \widehat{B}_{r}=\frac{d \log \left(S_{r}\right)-\left[\left(\mu_{1}-\mu_{2}\right) p_{r}+\mu_{2}-\sigma^{2} / 2\right] d r}{\sigma} .
$$

[^80]Given $S_{t}=S$ and $p_{t}=p$, the problem is to choose $\Lambda_{i}$ to maximize the discounted return

$$
J_{i}\left(S, p, t, \Lambda_{i}\right)=J_{i}\left(S, \alpha, t, \Lambda_{i}\right)
$$

subject to

$$
\left\{\begin{aligned}
d S_{r}=S_{r}\left[\left(\mu_{1}-\mu_{2}\right) p_{r}+\mu_{2}\right] d r+S_{r} \sigma d \widehat{B}_{r}, & S_{t}=S, \\
d p_{r}=\left[-\left(\lambda_{1}+\lambda_{2}\right) p_{r}+\lambda_{2}\right] d r+\frac{\left(\mu_{1}-\mu_{2}\right) p_{r}\left(1-p_{r}\right)}{\sigma} d \widehat{B}_{r}, & p_{t}=p .
\end{aligned}\right.
$$

Indeed, this new problem is a completely observable one because the conditional probability can be obtained using the stock price up to time $r$.

For $i=0,1$, let $V_{i}(S, p, t)$ denote the value functions with the states ( $S, p$ ) and net positions $i=0,1$ at time $t$. That is,

$$
V_{i}(S, p, t)=\sup _{\Lambda_{i}} J_{i}\left(S, p, t, \Lambda_{i}\right)
$$

The following lemma gives the upper bounds of the value functions.
Lemma 2.1. We have

$$
\begin{aligned}
& 0 \leq V_{0}(S, p, t) \leq S\left[e^{\left(\mu_{1}-\rho\right)(T-t)}-1\right] \\
& 0 \leq V_{1}(S, p, t) \leq S\left[2 e^{\left(\mu_{1}-\rho\right)(T-t)}-1\right] .
\end{aligned}
$$

Proof. It is clear that the nonnegativity of $V_{i}$ follows from their definition. It remains to show their upper bounds. First, given $\Lambda_{0}$, we have

$$
e^{-\rho v_{n}} S_{v_{n}}-e^{-\rho \tau_{n}} S_{\tau_{n}}=\int_{\tau_{n}}^{v_{n}} e^{-\rho r} S_{r}\left(\left(\mu_{1}-\mu_{2}\right) p_{r}+\mu_{2}-\rho\right) d r+\int_{\tau_{n}}^{v_{n}} e^{-\rho r} S_{r} \sigma d \widehat{B}_{r} .
$$

Note that

$$
E\left[I_{\left\{\tau_{n}<T\right\}} \int_{\tau_{n}}^{v_{n}} e^{-\rho r} S_{r} \sigma d \widehat{B}_{r}\right]=E\left[I_{\left\{\tau_{n}<T\right\}} E\left[\int_{\tau_{n}}^{v_{n}} e^{-\rho r} S_{r} \sigma d \widehat{B}_{r} \mid \tau_{n}\right]\right]=0 .
$$

It follows that

$$
\begin{aligned}
& E\left(e^{-\rho v_{n}} S_{v_{n}}-e^{-\rho \tau_{n}} S_{\tau_{n}}\right) I_{\left\{\tau_{n}<T\right\}} \\
& =E\left[I_{\left\{\tau_{n}<T\right\}} \int_{\tau_{n}}^{v_{n}} e^{-\rho r} S_{r}\left(\left(\mu_{1}-\mu_{2}\right) p_{r}+\mu_{2}-\rho\right) d r\right] \\
& \leq\left(\mu_{1}-\rho\right) \int_{\tau_{n}}^{v_{n}} e^{-\rho r} E S_{r} d r .
\end{aligned}
$$

Using the definition of $J_{0}\left(S, p, t, \Lambda_{0}\right)$, we have

$$
\begin{aligned}
J_{0}\left(S, p, t, \Lambda_{0}\right) & \leq \sum_{n=1}^{\infty} E\left(e^{-\rho\left(v_{n}-t\right)}\left(S_{v_{n}}-e^{-\rho\left(\tau_{n}-t\right)} S_{\tau_{n}}\right)\right) I_{\left\{\tau_{n}<T\right\}} \\
& \leq\left(\mu_{1}-\rho\right) \sum_{n=1}^{\infty} e^{\rho t} E \int_{\tau_{n}}^{v_{n}} e^{-\rho r} S_{r} d r \\
& \leq\left(\mu_{1}-\rho\right) e^{\rho t} \int_{t}^{T} e^{-\rho r} E S_{r} d r .
\end{aligned}
$$

It is easy to see using Gronwall's inequality that $E S_{r} \leq S e^{\mu_{1}(r-t)}$. It follows that

$$
J_{0}\left(S, p, t, \Lambda_{0}\right) \leq\left(\mu_{1}-\rho\right) e^{\rho t} S \int_{t}^{T} e^{-\rho r+\mu_{1}(r-t)} d r=S\left[e^{\left(\mu_{1}-\rho\right)(T-t)}-1\right]
$$

This implies that $0 \leq V_{0}(x) \leq S\left[e^{\left(\mu_{1}-\rho\right)(T-t)}-1\right]$.
Similarly, we have the inequality

$$
J_{1}\left(S, p, t, \Lambda_{1}\right) \leq E e^{-\rho\left(v_{1}-t\right)} S_{v_{1}}(1-K)+S\left[e^{\left(\mu_{1}-\rho\right)(T-t)}-1\right]
$$

Moreover, note that

$$
E e^{-\rho\left(v_{1}-t\right)} S_{v_{1}}-S \leq\left(\mu_{1}-\rho\right) e^{\rho t} \int_{t}^{T} e^{-\rho r} S_{r} d r \leq S\left[e^{\left(\mu_{1}-\rho\right)(T-t)}-1\right]
$$

This implies that

$$
V_{1}(S, p, t) \leq S+2 S\left[e^{\left(\mu_{1}-\rho\right)(T-t)}-1\right]=S\left[2 e^{\left(\mu_{1}-\rho\right)(T-t)}-1\right]
$$

Therefore, $0 \leq V_{1}(S, p, t) \leq S\left[2 e^{\left(\mu_{1}-\rho\right)(T-t)}-1\right]$. This completes the proof.
Let

$$
\begin{aligned}
\mathcal{A}= & \frac{1}{2}\left(\frac{\left(\mu_{1}-\mu_{2}\right) p(1-p)}{\sigma}\right)^{2} \partial_{p p}+\frac{1}{2} \sigma^{2} S^{2} \partial_{S S}+S\left[\left(\mu_{1}-\mu_{2}\right) p(1-p)\right] \partial_{S p} \\
& +\left[-\left(\lambda_{1}+\lambda_{2}\right)+\lambda_{2}\right] \partial_{p}+S\left[\left(\mu_{1}-\mu_{2}\right) p+\mu_{2}\right] \partial_{S}-\rho
\end{aligned}
$$

Then the HJB equations associated with our optimal stopping time problem can be given formally as follows:

$$
\begin{align*}
& \min \left\{-\partial_{t} V_{0}-\mathcal{A} V_{0}, V_{0}-V_{1}+S(1+K)\right\}=0  \tag{2.1}\\
& \min \left\{-\partial_{t} V_{1}-\mathcal{A} V_{1}, V_{1}-V_{0}-S(1-K)\right\}=0
\end{align*}
$$

in $(0,+\infty) \times(0,1) \times[0, T)$, with the terminal conditions

$$
\begin{align*}
& V_{0}(S, p, T)=0 \\
& V_{1}(S, p, T)=S(1-K) \tag{2.2}
\end{align*}
$$

The terminal condition implies that at the terminal time $T$ the net position must be flat.
Remark 2. In this paper, we restrict the state space of $p$ to $(0,1)$ because both $p=0$ and $p=1$ are entrance boundaries (see Karlin and Taylor [17] for the definition and discussions). Such boundaries cannot be reached from the interior of the state space. If the process begins there, it quickly moves to the interior and never returns. We next show that $p=0$ is indeed an entrance boundary. The case when $p=1$ is similar. It is easy to see that the speed density (see [17]) can be given by

$$
m(x)=\frac{\sigma^{2}}{\left(\mu_{1}-\mu_{2}\right)^{2} x^{2}(1-x)^{2} s(x)}
$$

where

$$
s(x)=\exp \left\{\frac{2 \sigma^{2}}{\left(\mu_{1}-\mu_{2}\right)^{2}}\left[\frac{\lambda_{2}}{x}+\frac{\lambda_{1}}{1-x}+\left(\lambda_{2}-\lambda_{1}\right) \log \frac{x}{1-x}\right]\right\}
$$

To show that $p=0$ is the entrance boundary, it suffices [17] to show that, for any $0<a<1$,

$$
\begin{equation*}
\lim _{\delta \rightarrow 0^{+}} \int_{\delta}^{a}\left(\int_{\delta}^{\xi} s(\eta) d \eta\right) m(\xi) d \xi=\infty \tag{2.3}
\end{equation*}
$$

and

$$
\begin{equation*}
\lim _{\delta \rightarrow 0^{+}} \int_{\delta}^{a}\left(\int_{\xi}^{a} s(\eta) d \eta\right) m(\xi) d \xi<\infty \tag{2.4}
\end{equation*}
$$

Now, (2.3) follows the fact that $\lim _{\delta \rightarrow 0^{+}} \int_{\delta}^{\xi} s(\eta) d \eta=\infty$ for any $\xi>0$. Moreover, note that for any $A>0$, after a change of variables,

$$
\int_{0}^{a}\left(\int_{\xi}^{a} e^{A / \eta} d \eta\right) \frac{e^{-A / \xi}}{\xi^{2}} d \xi=\int_{1 / a}^{\infty}\left(\int_{1 / \xi}^{1 / a} e^{A / \eta} d \eta\right) e^{-A \xi} d \xi=\int_{1 / a}^{\infty}\left(\int_{a}^{\xi} \frac{e^{A / \eta}}{\eta^{2}} d \eta\right) e^{-A \xi} d \xi<\infty
$$

Using this estimate, it is not difficult to see (2.4).
It is easy to show that the value functions $V_{0}$ and $V_{1}$ are linear in $S$. This motivates us to adopt the following transformation: $U_{0}(p, t)=V_{0}(S, p, t) / S$ and $U_{1}(p, t)=V_{1}(S, p, t) / S$. Then the HJB equations (2.1) with the terminal condition (2.2) can be reduced to

$$
\begin{align*}
& \min \left\{-\partial_{t} U_{0}-\mathcal{L} U_{0}, U_{0}-U_{1}+(1+K)\right\}=0  \tag{2.5}\\
& \min \left\{-\partial_{t} U_{1}-\mathcal{L} U_{1}, U_{1}-U_{0}-(1-K)\right\}=0
\end{align*}
$$

in $(0,1) \times[0, T)$, with the terminal conditions

$$
\begin{align*}
& U_{0}(p, T)=0  \tag{2.6}\\
& U_{1}(p, T)=1-K
\end{align*}
$$

where

$$
\begin{aligned}
\mathcal{L}= & \frac{1}{2}\left(\frac{\left(\mu_{1}-\mu_{2}\right) p(1-p)}{\sigma}\right)^{2} \partial_{p p} \\
& +\left[-\left(\lambda_{1}+\lambda_{2}\right) p+\lambda_{2}+\left(\mu_{1}-\mu_{2}\right) p(1-p)\right] \partial_{p}+\left(\mu_{1}-\mu_{2}\right) p+\mu_{2}-\rho
\end{aligned}
$$

Thanks to Lemma 2.1, we will focus on the bounded solutions of problem (2.5)-(2.6).
Lemma 2.2. Problem (2.5)-(2.6) has a unique bounded strong solution $\left(U_{0}, U_{1}\right)$, where $U_{i} \in$ $W_{q}^{2,1}([\varepsilon, 1-\varepsilon] \times[0, T])$ for any $\varepsilon \in(0,1 / 2), q \in[1,+\infty)$. Moreover,

$$
\begin{align*}
& -\partial_{t} U_{0}-\mathcal{L} U_{0}=\left(-\partial_{t} Z-\mathcal{L} Z\right)^{-}  \tag{2.7}\\
& -\partial_{t} U_{1}-\mathcal{L} U_{1}=\left(-\partial_{t} Z-\mathcal{L} Z\right)^{+} \tag{2.8}
\end{align*}
$$

where $Z(p, t) \equiv U_{1}(p, t)-U_{0}(p, t)$ is the unique strong solution to the following double obstacle problem:

$$
\min \left\{\max \left\{-\partial_{t} Z-\mathcal{L} Z, Z-(1+K)\right\}, Z-(1-K)\right\}=0
$$

or, equivalently,

$$
\begin{array}{ll}
-\partial_{t} Z-\mathcal{L} Z=0 & \text { if } 1-K<Z<1+K \\
-\partial_{t} Z-\mathcal{L} Z \geq 0 & \text { if } Z=1-K \\
-\partial_{t} Z-\mathcal{L} Z \leq 0 & \text { if } Z=1+K \tag{2.11}
\end{array}
$$

in $(0,1) \times[0, T)$, with the terminal condition $Z(p, T)=1-K$. Furthermore,

$$
\begin{equation*}
\partial_{p} Z \geq 0 \tag{2.12}
\end{equation*}
$$

and

$$
\begin{equation*}
\partial_{t} Z \leq 0 \tag{2.13}
\end{equation*}
$$

Proof. For any strong solution $\left(U_{0}, U_{1}\right)$ of problem (2.5)-(2.6), we first show that ${ }^{2} Z(p, t) \equiv$ $U_{1}(p, t)-U_{0}(p, t)$ satisfies (2.9)-(2.11). Indeed, if $1-K<Z(p, t)<1+K$, then

$$
-\partial_{t} U_{0}-\left.\mathcal{L} U_{0}\right|_{(p, t)}=-\partial_{t} U_{1}-\left.\mathcal{L} U_{1}\right|_{(p, t)}=0
$$

which gives $-\partial_{t} Z-\left.\mathcal{L} Z\right|_{(p, t)}=0$. If $Z(p, t)=1-K$ or $U_{1}(p, t)-U_{0}(p, t)-(1-K)=0$, then $U_{0}(p, t)-U_{1}(p, t)+1+K=2 K>0$, from which we infer that $-\partial_{t} U_{0}-\left.\mathcal{L} U_{0}\right|_{(p, t)}=0$. On the other hand, we always have $-\partial_{t} U_{1}-\left.\mathcal{L} U_{1}\right|_{(p, t)} \geq 0$, so that $-\partial_{t} Z-\left.\mathcal{L} Z\right|_{(p, t)} \geq 0$. Similarly we can deduce that $-\partial_{t} Z-\mathcal{L} Z \leq 0$ if $Z=1+K$.

By the penalization method (cf. Friedman [12]), we can show that the double obstacle problem has a unique strong solution

$$
Z(p, t) \in W_{q}^{2,1}([\varepsilon, 1-\varepsilon] \times[0, T])
$$

for any $\varepsilon \in(0,1 / 2), q \in[1,+\infty)$. To show the regularity and uniqueness of bounded solution to problem (2.5)-(2.6), it suffices to show that the solution $\left(U_{0}, U_{1}\right)$ to problem (2.5)-(2.6) satisfies (2.7)-(2.8). Let us prove (2.7) first. When $U_{0}(p, t)-U_{1}(p, t)>-(1+K)$ or $Z(p, t)<$ $1+K$, we have

$$
\begin{array}{r}
-\partial_{t} U_{0}-\mathcal{L} U_{0}=0 \\
-\partial_{t} Z-\mathcal{L} Z \geq 0
\end{array}
$$

from which we can see that (2.7) holds. As a result, it suffices to show that (2.7) remains valid when $U_{0}(p, t)-U_{1}(p, t)=-Z(p, t)=-(1+K)$. In this case $U_{1}(p, t)-U_{0}(p, t)=Z(p, t)>1-K$ and

$$
\begin{array}{r}
-\partial_{t} U_{1}-\mathcal{L} U_{1}=0 \\
-\partial_{t} Z-\mathcal{L} Z \leq 0
\end{array}
$$

[^81]It follows that

$$
\begin{aligned}
0 & \geq-\partial_{t} Z-\mathcal{L} Z=\left(-\partial_{t} U_{1}-\mathcal{L} U_{1}\right)-\left(-\partial_{t} U_{0}-\mathcal{L} U_{0}\right) \\
& =-\left(-\partial_{t} U_{0}-\mathcal{L} U_{0}\right),
\end{aligned}
$$

which implies the desired result (2.7). Equation (2.8) can be proved in a similar way.
Now let us prove (2.12). We need only restrict our attention to

$$
\begin{equation*}
N T \equiv\{(p, t) \in(0,1) \times[0, T): 1-K<Z(p, t)<1+K\} \tag{2.14}
\end{equation*}
$$

in which $-\partial_{t} Z-\mathcal{L} Z=0$. Differentiating the equation w.r.t. $p$, we have

$$
-\partial_{t}\left(\partial_{p} Z\right)-\mathcal{T}\left[\partial_{p} Z\right]=\left(\mu_{1}-\mu_{2}\right) Z,
$$

where $\mathcal{T}$ is another differential operator. Owing to $\left(\mu_{1}-\mu_{2}\right) Z \geq 0$ in $N T$ and $\partial_{p} Z=0$ on the boundary of $N T \backslash\{p=0,1\},{ }^{3}$ we then get (2.12) by using the maximum principle. It remains to prove (2.13). Clearly $\left.\partial_{t} Z\right|_{t=T} \leq 0$, which yields the desired result again by the maximum principle.

The region $N T$ defined in (2.14) refers to the no-trading region. In terms of the solution $Z(p, t)$ to the double obstacle problem (2.9)-(2.11), with the terminal condition $Z(p, T)=$ $1-K$, we can define the buying region $(B R)$ and the selling region $(S R)$ as follows:

$$
\left.\begin{array}{rl}
B R & =\{(p, t) \in(0,1) \times[0, T): Z(p, t)
\end{array}=1+K\right\}, ~ 子, ~=\{(p, t) \in(0,1) \times[0, T): Z(p, t)=1-K\} .
$$

We aim to characterize these regions through the study of the double obstacle problem. To begin with, we prove the connectivity of any $t$-section of $B R$ or $S R$.

Lemma 2.3. For any $t \in[0, T)$, the following hold:
(i) if $\left(p_{1}, t\right) \in B R$ and $p_{2} \geq p_{1}$, then $\left(p_{2}, t\right) \in B R$;
(ii) if $\left(p_{1}, t\right) \in S R$ and $p_{2} \leq p_{1}$, then $\left(p_{2}, t\right) \in S R$.

Proof. We prove only part (i) since the proof of part (ii) is similar. Since $Z\left(p_{1}, t\right)=1+K$, we infer by $(2.12)$ that $Z\left(p_{2}, t\right) \geq Z\left(p_{1}, t\right)=1+K$. On the other hand, $Z\left(p_{2}, t\right) \leq 1+K$. So we must have

$$
Z\left(p_{2}, t\right)=1+K,
$$

i.e., $\left(p_{2}, t\right) \in B R$, as desired.

The lemma below implies that $B R$ shrinks and $S R$ expands as $t$ approaches the terminal time $T$.

Lemma 2.4. For any $p \in(0,1)$, the following hold:
(i) if $\left(p, t_{1}\right) \in B R$ and $t_{2} \leq t_{1}$, then $\left(p, t_{2}\right) \in B R$;
(ii) if $\left(p, t_{1}\right) \in S R$ and $t_{2} \geq t_{1}$, then $\left(p, t_{2}\right) \in S R$.

Moreover,

$$
\begin{align*}
& B R \subset\left\{(p, t) \in(0,1) \times[0, T): p \geq \frac{\rho-\mu_{2}}{\mu_{1}-\mu_{2}}\right\}  \tag{2.15}\\
& S R \subset\left\{(p, t) \in(0,1) \times[0, T): p \leq \frac{\rho-\mu_{2}}{\mu_{1}-\mu_{2}}\right\} \tag{2.16}
\end{align*}
$$

[^82]Proof. In view of (2.13), the proofs of parts (i) and (ii) are similar to those of Lemma 2.3. It remains to show (2.15) and (2.16). Due to (2.10), we infer that if $(p, t) \in B R$, then

$$
0 \leq\left(-\partial_{t}-\mathcal{L}\right)(1-K)=-\left[\left(\mu_{1}-\mu_{2}\right) p+\mu_{2}-\rho\right],
$$

namely,

$$
p \leq \frac{\rho-\mu_{2}}{\mu_{1}-\mu_{2}}
$$

which gives (2.15). By (2.11), we can similarly get (2.16).
Combining $\overline{B R} \cap \overline{S R}=\emptyset$ with (2.15) and (2.16), we deduce that any $t$-section of $N T$ is nonempty. In view of Lemma 2.3, we can define two free boundaries:

$$
\begin{align*}
& p_{s}^{*}(t)=\inf \{p \in(0,1):(p, t) \in N T\},  \tag{2.17}\\
& p_{b}^{*}(t)=\sup \{p \in(0,1):(p, t) \in N T\} \tag{2.18}
\end{align*}
$$

for any $t \in[0, T)$. They are the thresholds for sell and buy, respectively.
Theorem 2.5. Let $p_{s}^{*}(t)$ and $p_{b}^{*}(t)$ be as given in (2.17)-(2.18). Then the following hold:
(i) $p_{b}^{*}(t)$ and $p_{s}^{*}(t)$ are monotonically increasing in $t$, and

$$
p_{b}^{*}(t) \geq \frac{\rho-\mu_{2}}{\mu_{1}-\mu_{2}} \geq p_{s}^{*}(t)
$$

for all $t \in[0, T)$. Moreover, $p_{b}^{*}(t), p_{s}^{*}(t) \in C^{\infty}(0, T)$;
(ii) $p_{s}^{*}(T):=\lim _{t \rightarrow T^{-}} p_{s}^{*}(t)=\frac{\rho-\mu_{2}}{\mu_{1}-\mu_{2}}$;
(iii) there is a $\delta>0$ such that $(1, t) \in N T$ for all $t \in(T-\delta, T)$, namely,

$$
p_{b}^{*}(t)=1 \quad \text { for } t \in(T-\delta, T)
$$

Moreover,

$$
\begin{equation*}
\delta \geq \frac{1}{\mu_{1}-\rho} \log \frac{1+K}{1-K} . \tag{2.19}
\end{equation*}
$$

Proof. The monotonicity in part (i) is a corollary of Lemmas 2.3 and 2.4. The proof for the smoothness of $p_{b}^{*}(t)$ and $p_{s}^{*}(t)$ is somewhat technical and is placed in the appendix. To show part (ii), we use the method of contradiction. Suppose not. Due to $p_{s}^{*}(t) \leq \frac{\rho-\mu_{2}}{\mu_{1}-\mu_{2}}$, we would have

$$
p_{s}^{*}(T)<\frac{\rho-\mu_{2}}{\mu_{1}-\mu_{2}}
$$

Then, for any $p \in\left(p_{s}^{*}(T), \frac{\rho-\mu_{2}}{\mu_{1}-\mu_{2}}\right)$,

$$
\left.\partial_{t} Z\right|_{t=T}=-\left.\mathcal{L} Z\right|_{t=T}=-\mathcal{L}(1-K)=-\left[\left(\mu_{1}-\mu_{2}\right) p+\mu_{2}-\rho\right]>0,
$$

which is in contradiction with (2.13).
It remains to show part (iii). Owing to $\left.Z\right|_{t=T}=1-K$ and (2.12), the existence of $\delta$ is apparent. We need only show (2.19). On $p=1$, problem (2.9)-(2.11) is reduced to

$$
\left\{\begin{array}{l}
-Z_{t}+\left.\left(\rho-\mu_{1}\right) Z\right|_{p=1}=-\lambda_{1} \partial_{p} Z \quad \text { if } Z<1+K  \tag{2.20}\\
-Z_{t}+\left.\left(\rho-\mu_{1}\right) Z\right|_{p=1} \leq-\lambda_{1} \partial_{p} Z \quad \text { if } Z=1+K
\end{array}\right.
$$

in $t \in(0, T)$, with the terminal condition $Z(1, T)=1-K$, where we have excluded the lower obstacle according to (2.16). Due to $-\lambda_{1} \partial_{p} Z \leq 0$, problem (2.20) has a supersolution $\bar{Z}(t)$ satisfying

$$
\left\{\begin{array}{l}
-\bar{Z}_{t}+\left(\rho-\mu_{1}\right) \bar{Z}=0 \text { if } \bar{Z}(t)<1+K,  \tag{2.21}\\
-\bar{Z}_{t}+\left(\rho-\mu_{1}\right) \bar{Z} \leq 0 \text { if } \bar{Z}(t)=1+K
\end{array}\right.
$$

in $t \in(0, T)$, with $\bar{Z}(T)=1-K$. It is easy to verify

$$
\bar{Z}(t)=\left\{\begin{array}{l}
e^{\left(\mu_{1}-\rho\right)(T-t)}(1-K) \text { if } t>T-\frac{1}{\mu_{1}-\rho} \log \frac{1+K}{1-K}, \\
1+K \text { if } t \leq T-\frac{1}{\mu_{1}-\rho} \log \frac{1+K}{1-K} .
\end{array}\right.
$$

We then infer that

$$
Z(1, t) \leq \bar{Z}(t)=e^{\left(\mu_{1}-\rho\right)(T-t)}(1-K)<1+K \quad \text { for } t>T-\frac{1}{\mu_{1}-\rho} \log \frac{1+K}{1-K}
$$

which implies that $(1, t) \in N T$ for any $t>T-\frac{1}{\mu_{1}-\rho} \log \frac{1+K}{1-K}$. Then (2.19) follows.
Remark 3. Part (iii) indicates that there is a critical time after which it is never optimal to buy stock. This is an important feature when transaction costs are involved in a finite horizon model. Similar results were obtained in the study of finite horizon portfolio selection with transaction costs (cf. [4, 6, 7, 19]). The intuition is that if the investor does not have a long enough time horizon to recover at least the transaction costs, then she/he should not initiate a long position (bear in mind that the terminal position must be flat).

Remark 4. Using the maximum principle, it is not hard to show that $Z\left(\cdot, \cdot ; \lambda_{1}, \lambda_{2}, \rho\right)$ is a decreasing function of $\lambda_{1}$ and $\rho$ and an increasing function of $\lambda_{2}$. As a consequence, $p_{s}^{*}\left(\cdot ; \lambda_{1}, \lambda_{2}, \rho\right)$ and $p_{b}^{*}\left(\cdot ; \lambda_{1}, \lambda_{2}, \rho\right)$ are also increasing functions of $\lambda_{1}$ and $\rho$ and decreasing functions of $\lambda_{2}$.

By Lemma 2.2, the Sobolev embedding theorem (cf. [12]), and the smoothness of free boundaries, the solutions $U_{0}$ and $U_{1}$ of problem (2.5)-(2.6) belong to $C^{1}$ in $(0,1) \times[0, T)$. Furthermore, it is easy to show that the solutions are sufficiently smooth (i.e., at least $C^{2}$ ) except at the free boundaries $p_{s}^{*}(t)$ and $p_{b}^{*}(t)$. These enable us to establish a verification theorem to show that the solutions $U_{0}$ and $U_{1}$ of problem (2.5)-(2.6) are equal to the value functions $V_{0} / S$ and $V_{1} / S$, respectively, and sequences of optimal stopping times can be constructed by using $\left(p_{s}^{*}, p_{b}^{*}\right)$.

This theorem gives sufficient conditions for the optimality of the trading rules in terms of the stopping times $\left\{\tau_{n}, v_{n}\right\}$. The construction procedure will be used in the next sections to develop numerical solutions in various scenarios.

Theorem 2.6 (verification theorem). Let $\left(U_{0}, U_{1}\right)$ be the unique bounded strong solution to problem (2.5)-(2.6) and $p_{b}^{*}(t)$ and $p_{s}^{*}(t)$ be the associated free boundaries. Then, $w_{0}(S, p, t) \equiv$ $S U_{0}(p, t)$ and $w_{1}(S, p, t) \equiv S U_{1}(p, t)$ are equal to the value functions $V_{0}(S, p, t)$ and $V_{1}(S, p, t)$, respectively.

Moreover, let

$$
\Lambda_{0}^{*}=\left(\tau_{1}^{*}, v_{1}^{*}, \tau_{2}^{*}, v_{2}^{*}, \ldots\right)
$$

where the stopping times $\tau_{1}^{*}=T \wedge \inf \left\{r \geq t: p_{r} \geq p_{b}^{*}(r)\right\}, v_{n}^{*}=T \wedge \inf \left\{r \geq \tau_{n}^{*}: p_{r} \leq p_{s}^{*}(r)\right\}$, and $\tau_{n+1}^{*}=T \wedge \inf \left\{r>v_{n}^{*}: p_{r} \geq p_{b}^{*}(r)\right\}$ for $n \geq 1$, and let

$$
\Lambda_{1}^{*}=\left(v_{1}^{*}, \tau_{2}^{*}, v_{2}^{*}, \tau_{3}^{*}, \ldots\right)
$$

where the stopping times $v_{1}^{*}=T \wedge \inf \left\{r \geq t: p_{r}^{*} \leq p_{s}^{*}(r)\right\}, \tau_{n}^{*}=T \wedge \inf \left\{r>v_{n-1}^{*}: p_{r} \geq p_{b}^{*}(r)\right\}$, and $v_{n}^{*}=T \wedge \inf \left\{r \geq \tau_{n}^{*}: p_{r} \leq p_{s}^{*}(r)\right\}$ for $n \geq 2$. If $v_{n}^{*} \rightarrow T$ a.s., as $n \rightarrow \infty$, then $\Lambda_{0}^{*}$ and $\Lambda_{1}^{*}$ are optimal.

Proof. The proof is divided into two steps. In the first step, we show that $w_{i}(S, p, t) \geq$ $J_{i}\left(S, p, t, \Lambda_{i}\right)$ for all $\Lambda_{i}$. Then, in the second step, we show that $w_{i}(S, p, t)=J_{i}\left(S, p, t, \Lambda_{i}^{*}\right)$. Therefore, $w_{i}(S, p, t)=V_{i}(S, p, t)$ and $\Lambda_{i}^{*}$ is optimal.

Using $\left(-\partial_{t} w_{i}-\mathcal{L} w_{i}\right) \geq 0$, Dynkin's formula, and Fatou's lemma as in Øksendal [24, p. 226], we have, for any stopping times $t \leq \theta_{1} \leq \theta_{2}$, a.s.,

$$
\begin{equation*}
E e^{-\rho\left(\theta_{1}-t\right)} w_{i}\left(S_{\theta_{1}}, p_{\theta_{1}}, \theta_{1}\right) I_{\left\{\theta_{1}<a\right\}} \geq E e^{-\rho\left(\theta_{2}-t\right)} w_{i}\left(S_{\theta_{2}}, p_{\theta_{2}}, \theta_{2}\right) I_{\left\{\theta_{1}<a\right\}} \tag{2.22}
\end{equation*}
$$

for any $a$ and $i=0,1$.
Note that $w_{0} \geq w_{1}-S(1+K)$. Given $\Lambda_{0}=\left(\tau_{1}, v_{1}, \tau_{2}, v_{2}, \ldots\right)$, by (2.22), and noting that $w_{0}(S, p, T)=0$, we have

$$
\begin{aligned}
w_{0}(S, p, t) & \geq E e^{-\rho\left(\tau_{1}-t\right)} w_{0}\left(S_{\tau_{1}}, p_{\tau_{1}}, \tau_{1}\right) \\
& =E e^{-\rho\left(\tau_{1}-t\right)} w_{0}\left(S_{\tau_{1}}, p_{\tau_{1}}, \tau_{1}\right) I_{\left\{\tau_{1}<T\right\}} \\
& \geq E e^{-\rho\left(\tau_{1}-t\right)}\left(w_{1}\left(S_{\tau_{1}}, p_{\tau_{1}}, \tau_{1}\right)-S_{\tau_{1}}(1+K)\right) I_{\left\{\tau_{1}<T\right\}} \\
& =E e^{-\rho\left(\tau_{1}-t\right)} w_{1}\left(S_{\tau_{1}}, p_{\tau_{1}}, \tau_{1}\right) I_{\left\{\tau_{1}<T\right\}}-E e^{-\rho\left(\tau_{1}-t\right)} S_{\tau_{1}}(1+K) I_{\left\{\tau_{1}<T\right\}}
\end{aligned}
$$

Using (2.22) again with $i=1$ and noticing that $v_{1} \geq \tau_{1}$ and $w_{1} \geq w_{0}+S(1-K)$, we have

$$
\begin{aligned}
w_{0}(S, p, t) \geq & E e^{-\rho\left(v_{1}-t\right)} w_{1}\left(S_{v_{1}}, p_{v_{1}}, v_{1}\right) I_{\left\{\tau_{1}<T\right\}}-E e^{-\rho\left(\tau_{1}-t\right)} S_{\tau_{1}}(1+K) I_{\left\{\tau_{1}<T\right\}} \\
\geq & E e^{-\rho\left(v_{1}-t\right)}\left(w_{0}\left(S_{v_{1}}, p_{v_{1}}, v_{1}\right)+S_{v_{1}}(1-K)\right) I_{\left\{\tau_{1}<T\right\}} \\
& -E e^{-\rho\left(\tau_{1}-t\right)} S_{\tau_{1}}(1+K) I_{\left\{\tau_{1}<T\right\}} \\
= & E e^{-\rho\left(v_{1}-t\right)} w_{0}\left(S_{v_{1}}, p_{v_{1}}, v_{1}\right) I_{\left\{\tau_{1}<T\right\}} \\
& +E\left[e^{-\rho\left(v_{1}-t\right)} S_{v_{1}}(1-K)-e^{-\rho\left(\tau_{1}-t\right)} S_{\tau_{1}}(1+K)\right] I_{\left\{\tau_{1}<T\right\}} \\
= & E e^{-\rho\left(v_{1}-t\right)} w_{0}\left(S_{v_{1}}, p_{v_{1}}, v_{1}\right) \\
& +E\left[e^{-\rho\left(v_{1}-t\right)} S_{v_{1}}(1-K)-e^{-\rho\left(\tau_{1}-t\right)} S_{\tau_{1}}(1+K)\right] I_{\left\{\tau_{1}<T\right\}}
\end{aligned}
$$

Continue this way and recall that $w_{i}(S, p, t) \geq 0$ to obtain

$$
w_{0}(S, p, t) \geq E \sum_{n=1}^{N}\left[e^{-\rho\left(v_{n}-t\right)} S_{v_{n}}(1-K)-e^{-\rho\left(\tau_{n}-t\right)} S_{\tau_{n}}(1+K)\right] I_{\left\{\tau_{n}<T\right\}}
$$

Sending $N \rightarrow \infty$, we have $w_{0}(S, p, t) \geq J_{0}\left(S, p, t, \Lambda_{0}\right)$ for all $\Lambda_{0}$. This implies that $w_{0}(S, p, t) \geq$ $V_{0}(S, p, t)$. Similarly, we can show that $w_{1}(S, p, t) \geq V_{1}(S, p, t)$.

We next establish the equalities. For given $t$, define

$$
\tau_{1}^{*}= \begin{cases}t & \text { if } p \geq p_{b}^{*}(t) \\ T \wedge \inf \left\{r \geq t: p_{r}=p_{b}^{*}(r)\right\} & \text { if } p<p_{b}^{*}(t)\end{cases}
$$

Using Dynkin's formula, we have

$$
\begin{aligned}
w_{0}(S, p, t) & =E e^{-\rho\left(\tau_{1}^{*}-t\right)} w_{0}\left(S_{\tau_{1}^{*}}, p_{\tau_{1}^{*}}, \tau_{1}^{*}\right) \\
& =E e^{-\rho\left(\tau_{1}^{*}-t\right)} w_{0}\left(S_{\tau_{1}^{*}}, p_{\tau_{1}^{*}}, \tau_{1}^{*}\right) I_{\left\{\tau_{1}^{*}<T\right\}} \\
& =E e^{-\rho\left(\tau_{1}^{*}-t\right)}\left(w_{1}\left(S_{\tau_{1}^{*}}, p_{\tau_{1}^{*}}, \tau_{1}^{*}\right)-S_{\tau_{1}^{*}}(1+K)\right) I_{\left\{\tau_{1}^{*}<T\right\}} \\
& =E e^{-\rho\left(\tau_{1}^{*}-t\right)} w_{1}\left(S_{\tau_{1}^{*}}, p_{\tau_{1}^{*}}, \tau_{1}^{*}\right) I_{\left\{\tau_{1}^{*}<T\right\}}-E e^{-\rho\left(\tau_{1}^{*}-t\right)} S_{\tau_{1}^{*}}(1+K) I_{\left\{\tau_{1}^{*}<T\right\}}
\end{aligned}
$$

Let $v_{1}^{*}=T \wedge \inf \left\{r \geq \tau_{1}^{*}: p_{r}=p_{s}^{*}(r)\right\}$. Noticing that $w_{1}(S, p, T)=S(1-K)$, we have

$$
\begin{aligned}
& E e^{-\rho\left(\tau_{1}^{*}-t\right)} w_{1}\left(S_{\tau_{1}^{*}}, p_{\tau_{1}^{*}}, \tau_{1}^{*}\right) I_{\left\{\tau_{1}^{*}<T\right\}} \\
&= E e^{-\rho\left(v_{1}^{*}-t\right)} w_{1}\left(S_{v_{1}^{*}}, p_{v_{1}^{*}}, v_{1}^{*}\right) I_{\left\{\tau_{1}^{*}<T\right\}} \\
&= E e^{-\rho\left(v_{1}^{*}-t\right)} w_{1}\left(S_{v_{1}^{*}}, p_{v_{1}^{*}}, v_{1}^{*}\right) I_{\left\{\tau_{1}^{*}<T\right\}} I_{\left\{v_{1}^{*}<T\right\}}+E e^{-\rho T} S_{T-t}(1-K) I_{\left\{\tau_{1}^{*}<T\right\}} I_{\left\{v_{1}^{*}=T\right\}} \\
&= E e^{-\rho\left(v_{1}^{*}-t\right)}\left(w_{0}\left(S_{v_{1}^{*}}, p_{v_{1}^{*}}, v_{1}^{*}\right)+S_{v_{1}^{*}}(1-K)\right) I_{\left\{\tau_{1}^{*}<T\right\}} I_{\left\{v_{1}^{*}<T\right\}} \\
&+E e^{-\rho T} S_{T-t}(1-K) I_{\left\{\tau_{1}^{*}<T\right\}} I_{\left\{v_{1}^{*}=T\right\}} \\
&= E e^{-\rho\left(v_{1}^{*}-t\right)} w_{0}\left(S_{v_{1}^{*}}, p_{v_{1}^{*}}, v_{1}^{*}\right) I_{\left\{\tau_{1}^{*}<T\right\}} I_{\left\{v_{1}^{*}<T\right\}}+E e^{-\rho\left(v_{1}^{*}-t\right)} S_{v_{1}^{*}}(1-K) I_{\left\{\tau_{1}^{*}<T\right\}} I_{\left\{v_{1}^{*}<T\right\}} \\
&+E e^{-\rho(T-t)} S_{T}(1-K) I_{\left\{\tau_{1}^{*}<T\right\}} I_{\left\{v_{1}^{*}=T\right\}} \\
&= E e^{-\rho\left(v_{1}^{*}-t\right)} w_{0}\left(S_{v_{1}^{*}}, p_{v_{1}^{*}}, v_{1}^{*}\right)+E e^{-\rho\left(v_{1}^{*}-T\right)} S_{v_{1}^{*}}(1-K) I_{\left\{\tau_{1}^{*}<T\right\}} I_{\left\{v_{1}^{*}<T\right\}} \\
&+E e^{-\rho(T-t)} S_{T}(1-K) I_{\left\{\tau_{1}^{*}<T\right\}} I_{\left\{v_{1}^{*}=T\right\}} \\
&= E e^{-\rho\left(v_{1}^{*}-t\right)} w_{0}\left(S_{v_{1}^{*}}, p_{v_{1}^{*}}, v_{1}^{*}\right)+E e^{-\rho\left(v_{1}^{*}-t\right)} S_{v_{1}^{*}}(1-K) I_{\left\{\tau_{1}^{*}<T\right\}}
\end{aligned}
$$

It follows that $w_{0}(S, p, t)=E e^{-\rho\left(v_{1}^{*}-t\right)} w_{0}\left(S_{v_{1}^{*}}, p_{v_{1}^{*}}, v_{1}^{*}\right)+E\left[e^{-\rho\left(v_{1}^{*}-t\right)} S_{v_{1}^{*}}(1-K)-e^{-\rho\left(\tau_{1}^{*}-t\right)} S_{\tau_{1}^{*}}(1+K)\right] I_{\left\{\tau_{1}^{*}<T\right\}}$.
Continue the procedure to obtain

$$
\begin{aligned}
w_{0}(S, p, t)= & E e^{-\rho\left(v_{n}^{*}-t\right)} w_{0}\left(S_{v_{n}^{*}}, p_{v_{n}^{*}}, v_{n}^{*}\right) \\
& +E \sum_{k=1}^{n}\left[e^{-\rho\left(v_{k}^{*}-t\right)} S_{v_{k}^{*}}(1-K)-e^{-\rho\left(\tau_{k}^{*}-t\right)} S_{\tau_{k}^{*}}(1+K)\right] I_{\left\{\tau_{k}^{*}<T\right\}}
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
w_{1}(S, p, t)= & E e^{-\rho\left(v_{n}^{*}-t\right)} w_{0}\left(S_{v_{n}^{*}}, p_{v_{n}^{*}}, v_{n}^{*}\right) \\
& +E e^{-\rho\left(v_{1}^{*}-t\right)} S_{v_{1}^{*}}(1-K) \\
& +E \sum_{k=2}^{n}\left[e^{-\rho\left(v_{k}^{*}-t\right)} S_{v_{k}^{*}}(1-K)-e^{-\rho\left(\tau_{k}^{*}-t\right)} S_{\tau_{k}^{*}}(1+K)\right] I_{\left\{\tau_{k}^{*}<T\right\}}
\end{aligned}
$$

Recall that $v_{n}^{*} \rightarrow T$. Sending $n \rightarrow \infty$ and noticing $w_{0}(S, p, T)=0$, we obtain the equalities. This completes the proof.

Table 1
Parameter values.

| $\lambda_{1}$ | $\lambda_{2}$ | $\mu_{1}$ | $\mu_{2}$ | $\sigma$ | $K$ | $\rho$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0.36 | 2.53 | 0.18 | -0.77 | 0.184 | 0.001 | 0.0679 |



Figure 1. Optimal trading strategy.
3. Numerical results for optimal trading strategy. The theoretical analysis in section 2 shows that $p_{s}^{*}(t)$ and $p_{b}^{*}(t)$ are thresholds for the optimal trend following trading strategy: buy the stock when $p_{t}$ crosses $p_{b}^{*}(t)$ from below and sell the stock when $p_{t}$ crosses $p_{s}^{*}(t)$ from above. Knowing the parameters of our regime switching model, we can numerically solve the double obstacle problem (2.9)-(2.11) to derive approximations of those thresholds. To do that, we employ the penalization method with a finite difference discretization (see Dai, Kwok, and You [5] and Forsyth and Vetzal [11]). The penalized approximation to the double obstacle problem is

$$
-\partial_{t} Z-\mathcal{L} Z=\beta(1-K-Z)^{+}-\beta(Z-1-K)^{+},
$$

where $\beta$ is the penalty parameter. In our numerical examples, we choose $\beta=10^{7}$. The righthand side of the approximation is linearized by using a nonsmooth version of the Newton iteration. Then the resulting equations are discretized by the standard finite difference method.

We take $T=1$ and use the model parameters in Table 1 based on the statistics for DJIA. Figure 1 represents $p_{s}^{*}(\cdot)$ and $p_{b}^{*}(\cdot)$ as functions of time $t$. We see that $p_{s}^{*}(t)$ approaches the theoretical value $\left(\rho-\mu_{2}\right) /\left(\mu_{1}-\mu_{2}\right)=(0.0679+0.77) /(0.18+0.77)=0.882$ as $t \rightarrow T=1$. Also, we observe that there is a $\delta>0$ such that $p_{b}^{*}(t)=1$ for $t \in[T-\delta, T]$, which indicates that it is never optimal to buy stock when $t$ is very close to $T$. Using Theorem 2.5, the lower bound of $\delta$ is estimated as $\frac{1}{\mu_{1}-\rho} \log \frac{1+K}{1-K}=\frac{1}{0.18-0.0679} \log \frac{1.001}{0.999}=0.0178$, which is consistent with the numerical result.

Figure 2 illustrates the impact of parameters $\lambda$ and $\rho$ on the optimal thresholds $p_{s}^{*}(\cdot)$ and $p_{b}^{*}(\cdot)$. We can see that they are increasing functions of $\rho$ and decreasing functions of $\lambda_{2}$ as indicated in Remark 2. Moreover, properties (i)-(iii) of $p_{s}^{*}(\cdot)$ and $p_{b}^{*}(\cdot)$ stated in Theorem 2.5 are also visible.


Figure 2. Effects of parameters $\lambda$ and $\rho$ on optimal trading strategy.


Figure 3. Effects of transaction cost $K$ on optimal trading strategy.

In Figure 3, we examine the impact of transaction cost $K$ on the optimal thresholds. It is observed that the no-trading regions expands as we increase the transaction cost from 0.001 to 0.005 . This is consistent with our intuition that increasing transaction costs can decrease trading frequency.
4. Simulations and market tests. To evaluate how well our trend following trading strategy would work in practice, we conduct experiments using both simulations and market historical data. The tests are oriented towards the practicality of using $p_{t}$ to detect the regime switches. Thus, we focus on an implementable trend following strategy that approximates the optimal one and test its robustness.
4.1. Method. From Figures $1-3$ we can see that the thresholds $p_{s}^{*}(\cdot)$ and $p_{b}^{*}(\cdot)$ are almost constant except when $t$ gets very close to the terminal time $T$. In our experiments, $T$ is always more than 10 years, which is relatively large. As a result, the contribution of the
trading near $T$ to the reward function is small. Thus, we will approximate $p_{s}^{*}(\cdot)$ and $p_{b}^{*}(\cdot)$ with two constant threshold values $p_{s}^{*}=\lim _{t \rightarrow 0+} p_{s}^{*}(t)$ and $p_{b}^{*}=\lim _{t \rightarrow 0+} p_{b}^{*}(t)$. Then we estimate $p_{t}$, the conditional probability in a bull market at time $t$, and check it against the thresholds to determine whether to buy, hold, or sell. The trend following trading strategy we will use is to buy the stock when $p_{t}$ crosses $p_{b}^{*}$ from below for the first time and convert to the bond when $p_{t}$ crosses $p_{s}^{*}$ from above for the first time. Moreover, we always liquidate our holdings of stock or bond at $T$.

Some qualitative analysis of $p_{t}$ is helpful before experiments. Using the observation SDE equations, we see that $p_{t}$ is related to the stock price $S_{t}$ by

$$
\begin{equation*}
d p_{t}=f\left(p_{t}\right) d t+\frac{\left(\mu_{1}-\mu_{2}\right) p_{t}\left(1-p_{t}\right)}{\sigma^{2}} d \log \left(S_{t}\right) \tag{4.1}
\end{equation*}
$$

where $f$ is a third-order polynomial of $p_{t}$ given by

$$
f\left(p_{t}\right)=-\left(\lambda_{1}+\lambda_{2}\right) p_{t}+\lambda_{2}-\frac{\left(\mu_{1}-\mu_{2}\right) p_{t}\left(1-p_{t}\right)\left(\left(\mu_{1}-\mu_{2}\right) p_{t}+\mu_{2}-\sigma^{2} / 2\right)}{\sigma^{2}}
$$

It is easy to check that $f(0)=\lambda_{2}>0$ and $f(1)=-\lambda_{1}<0$. Moreover, as $p_{t}$ approaches $\pm \infty$, so does $f\left(p_{t}\right)$. Thus, $f$ has exactly one root $\xi$ in $[0,1]$. When the stock prices stay constant, $p_{t}$ is attracted to $\xi$. This attractor is an unbiased choice for $p_{0}$. Since $\left(\mu_{1}-\mu_{2}\right) p_{t}\left(1-p_{t}\right) / \sigma^{2} \geq 0$, $p_{t}$ moves in the same direction as the stock prices. This is also intuitive since the stock price movements indicate trends. The magnitude of the impact varies, though. Fixing all the parameters, the impact of stock price movement becomes relatively small when $p_{t}$ is getting close to 0 or 1 . Among the parameters $\mu_{1}, \mu_{2}$, and $\sigma$ the latter has a larger impact since it appears in a square in the formula. A smaller $\sigma$ magnifies the impact of stock movement and tends to cause more frequent trading, and a larger $\sigma$ does the opposite. We will estimate $p_{t}$ simply by replacing the differential in (4.1) with a difference using the trading day as the step size on a finite time horizon $[0, N d t]$ :

$$
p_{t+1}=p_{t}+f\left(p_{t}\right) d t+\frac{\left(\mu_{1}-\mu_{2}\right) p_{t}\left(1-p_{t}\right)}{\sigma^{2}} \log \left(\frac{S_{t+1}}{S_{t}}\right)
$$

where $d t=1 / 250$ and $t=0, d t, 2 d t, \ldots, N d t$.
If everything changes continuously, then $p_{t} \in[0,1]$. Indeed, simply changing the differential equation to a difference equation works extremely well for simulated paths. However, this could be violated in the approximation when $S_{t}$ jumps. This would happen, for example, in testing the SP500: during the 1987 crash on the downside and recently on the upside. We have to restrict $p_{t} \in[0,1]$ in implementing the approximation. Thus, in simulation and testing, we calculate $p_{t}$ iteratively with

$$
p_{t+1}=\min \left(\max \left(p_{t}+f\left(p_{t}\right) d t+\frac{\left(\mu_{1}-\mu_{2}\right) p_{t}\left(1-p_{t}\right)}{\sigma^{2}} \log \left(\frac{S_{t+1}}{S_{t}}\right), 0\right), 1\right)
$$

4.2. Simulation. In our analysis the choice of the objective function is partly due to tractability. Nevertheless, this gives us a way of recognizing trends by relatively high probabilities in the bull or bear regimes, respectively. How effective is this trend recognizing method? We check it by simulation. We use the same parameter values as in Table 1.


Figure 4. Sample path.
Table 2
Simulation results.

| No. of simulations | TF | BH | TF/BH | No. of trades |
| :---: | :---: | :---: | :---: | :---: |
| 1000 | 77.28 | 6.04 | 12.79 | 37.56 |
| 2000 | 73.97 | 6.00 | 12.33 | 37.66 |
| 5000 | 75.48 | 5.64 | 13.38 | 37.56 |
| 10000 | 73.6 | 5.45 | 13.50 | 37.64 |
| 50000 | 74.5 | 5.59 | 13.33 | 37.56 |

Numerically solving the obstacle problem for $t \rightarrow 0+$ yields thresholds $p_{s}^{*}=0.768$ and $p_{b}^{*}=0.934$ for down and up trends, respectively. Assuming no prior knowledge on the trend of the market, we set $p_{0}=0.8$ roughly in the middle of the two thresholds. A typical 20 -year sample path is given in Figure 4, in which for the prices $S_{r}$ when we are long the stocks according to our strategy $\left(p(t) \in B R \cup N T\right.$ after crossing $p_{b}^{*}$ from below) are marked in blue and those when we are flat $\left(p(t) \in S R \cup N T\right.$ after crossing $p_{s}^{*}$ from above) in red. We can see that on this typical path the signals are quite effective in detecting the regime switching.

A natural trading strategy of using these signals is to buy stock in the beginning of an up trend as signaled by $p_{t}$ crossing the upper threshold $p_{b}^{*}$ from below and switch to bond when $p_{t}$ crosses the lower threshold $p_{s}^{*}$ from above. We simulate this trend following strategy against the buy and hold strategy by using a large number of simulated paths. Again we use $p_{0}=0.8$. The average returns of the trend following (TF) strategy on one unit invested on simulation paths are listed in Table 2 along with the average number of trades on each path. We also list the average return of the buy and hold ( BH ) strategy for comparison.

Judging from the ratio $\mathrm{TF} / \mathrm{BH}$, the simulation tends to stabilize around 5000 rounds. We run the 5000 -round simulation 10 times and summarize the mean and standard deviation in

Table 3
Statistics of 105000 -round simulations.

|  | TF | BH | No. of trades |
| :---: | :---: | :---: | :---: |
| Mean | 74.6 | 5.72 | 37.55 |
| Stdev | 1.64 | 0.31 | 0.15 |

Table 4
Thresholds corresponding to different parameters.

| $\lambda_{1}$ | $\lambda_{2}$ | $\mu_{1}$ | $\mu_{2}$ | $\sigma$ | $\rho$ | $p_{s}^{*}$ | $p_{b}^{*}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0.36 | 2.53 | 0.18 | -0.77 | 0.184 | 0.067 | 0.768 | 0.934 |
| 0.36 | 2.53 | 0.18 | -0.77 | 0.184 | $\mathbf{0 . 0 6 2}$ | 0.761 | 0.931 |
| 0.36 | 2.53 | 0.18 | -0.77 | 0.184 | $\mathbf{0 . 0 7 2}$ | 0.775 | 0.938 |
| 0.36 | 2.53 | 0.18 | -0.77 | $\mathbf{0 . 1 7 4}$ | 0.067 | 0.762 | 0.936 |
| 0.36 | 2.53 | 0.18 | -0.77 | $\mathbf{0 . 1 9 4}$ | 0.067 | 0.773 | 0.933 |
| 0.36 | 2.53 | $\mathbf{0 . 1 7}$ | $\mathbf{- 0 . 7 1}$ | 0.184 | 0.067 | 0.774 | 0.935 |
| 0.36 | 2.53 | $\mathbf{0 . 1 9}$ | $\mathbf{- 0 . 8 3}$ | 0.184 | 0.067 | 0.762 | 0.934 |
| 0.36 | $\mathbf{2}$ | 0.18 | -0.77 | 0.184 | 0.067 | 0.769 | 0.935 |
| 0.36 | $\mathbf{3}$ | 0.18 | -0.77 | 0.184 | 0.067 | 0.767 | 0.935 |
| $\mathbf{0 . 3}$ | 2.53 | 0.18 | -0.77 | 0.184 | 0.067 | 0.767 | 0.934 |
| $\mathbf{0 . 4 2}$ | 2.53 | 0.18 | -0.77 | 0.184 | 0.067 | 0.769 | 0.935 |

Table 5
Shifting the thresholds.

| $p_{s}^{*}$ | $p_{b}^{*}$ | TF | BH | TF/BH | No. of trades |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 0.75900 | 0.92500 | 74.407 | 5.6478 | 13.174 | 37.039 |
| 0.76300 | 0.92900 | 76.523 | 5.9231 | 12.919 | 37.139 |
| 0.76700 | 0.93300 | 75.467 | 5.8328 | 12.938 | 37.371 |
| 0.77100 | 0.93700 | 74.232 | 5.6367 | 13.169 | 37.648 |
| 0.77500 | 0.94100 | 75.986 | 5.5975 | 13.575 | 37.766 |
| 0.77900 | 0.94500 | 77.029 | 6.0613 | 12.708 | 37.923 |

Table 3, which confirms our observation.
These simulations show that the trend following strategy has a distinctive advantage over the buy and hold strategy and is quite stable in both return and trading frequency. What if the parameters are perturbed? It turns out that the thresholds are not sensitive to the parameters as summarized in Table 4.

Next we test the robustness of the trend following trading strategy against the perturbation of the thresholds. This is important because we are using the limits of the optimal thresholds $p_{s}^{*}(\cdot)$ and $p_{b}^{*}(\cdot)$ as $t \rightarrow 0+$ to approximate them. We perturb the constant thresholds $p_{s}^{*}$ and $p_{b}^{*}$ both by shifting and by altering the spreads. The results are summarized in Tables 5 and 6 .

We can see from Table 5 that shifting the thresholds has little impact. Table 6 shows that the average number of trades in the trend following trading strategy is inversely correlated to the spreads of the thresholds. However, the relative advantage of the trend following strategy over the buy and hold strategy is not sensitive to the perturbation of the thresholds.

In the tests discussed above, the trading cost $K=0.001$ is fixed. Increasing $K$, the spread of the optimal thresholds will also increase as indicated in Figure 3. Simulations summarized in Table 6 indicate that this will reduce the trading frequency, which is consistent with intuition.

Table 6
Changing the spreads of the thresholds.

| $p_{s}^{*}$ | $p_{b}^{*}$ | TF | BH | TF/BH | No. of trades |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 0.75900 | 0.94500 | 74.892 | 6.1736 | 12.131 | 34.148 |
| 0.76300 | 0.94100 | 73.561 | 5.4182 | 13.577 | 35.591 |
| 0.76700 | 0.93700 | 75.480 | 5.7035 | 13.234 | 36.743 |
| 0.77100 | 0.93300 | 77.059 | 6.0427 | 12.753 | 38.122 |
| 0.77500 | 0.92900 | 74.010 | 5.5380 | 13.364 | 39.861 |
| 0.77900 | 0.92500 | 76.442 | 5.7994 | 13.181 | 41.375 |

Table 7
Changing the trading cost.

| $K$ | $p_{s}^{*}$ | $p_{b}^{*}$ | TF | BH | No. of trades |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 0.001 | 0.768 | 0.934 | 75.8 | 5.9 | 37.3 |
| 0.005 | 0.64 | 0.954 | 55.77 | 5.75 | 20.93 |
| 0.01 | 0.545 | 0.962 | 43.37 | 5.9 | 15.88 |
| 0.02 | 0.422 | 0.969 | 30.6 | 5.25 | 11.92 |

Table 8
Statistics of bull and bear markets.

| Index | $\lambda_{1}$ | $\lambda_{2}$ | $\mu_{1}$ | $\mu_{2}$ | $\sigma_{1}$ | $\sigma_{2}$ | $\sigma$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| SP500 (62-08) | 0.353 | 2.208 | 0.196 | -0.616 | 0.135 | 0.211 | 0.173 |
| DJIA (62-08) | 0.36 | 2.53 | 0.18 | -0.77 | 0.144 | 0.223 | 0.184 |
| NASDAQ (91-08) | 2.158 | 2.3 | 0.875 | -1.028 | 0.273 | 0.35 | 0.31 |

Simulating 5000 rounds for each of the trading cost levels $K=0.005,0.01,0.02$ shows that the advantage of the trend following methods decreases as $K$ increases as expected. However, even at $K=0.02$, the advantage is still quite obvious. The testing results are summarized in Table 7.

The simulations convince us that the trend following trading system of using $p_{t}$ crossing the constant thresholds $p_{s}^{*}$ and $p_{b}^{*}$ to detect the trend of the stock price movement in a regime switching model is effective and robust.
4.3. Testing in stock markets. Does this trend following strategy work in real markets? We test it on the historical data of the SP500, DJIA, and NASDAQ indices. The SP500 index started active trading in 1962 and NASDAQ in 1991, and we test them up to the end of 2008. We also test DJIA from 1962-2008.

First we need to determine the parameters. We regard a decline of more than $19 \%$ as a bear market and a rally of $24 \%$ or more as a bull market. Statistics of bull and bear markets for SP500 index and DJIA in the 47 years from 1962-2008 and NASDAQ from 1991-2008 (see Tables 11,12 , and 13 in the appendix) are shown in Table 8 . Here $\sigma_{1}$ and $\sigma_{2}$ are the average annualized standard deviation corresponding to the bull and bear markets, respectively, and $\sigma=\left(\sigma_{1}+\sigma_{2}\right) / 2$. Currently, retail discount brokers usually charge $\$ 2.5-10$ per trade for unlimited number of shares (e.g., Just2Trade $\$ 2.5$ per trade, ScotTrade $\$ 7$ per trade, ETrade $\$ 10$ per trade, and TDAmeritrade $\$ 10$ per trade). For professional traders who deal with clearing houses directly, trading a $\$ 100000$ standard lot usually cost less than $\$ 1$. Assuming

Table 9
Thresholds for indices.

| Index | Lower thresholds | Upper thresholds |
| :---: | :---: | :---: |
| SP500 $\sigma_{1}$ | 0.69 | $\mathbf{0 . 9 1}$ |
| SP500 $\sigma_{2}$ | $\mathbf{0 . 7 4}$ | 0.90 |
| DJIA $\sigma_{1}$ | 0.74 | $\mathbf{0 . 9 4}$ |
| DJIA $\sigma_{2}$ | $\mathbf{0 . 7 8}$ | 0.93 |
| NASDAQ $\sigma_{1}$ | 0.43 | $\mathbf{0 . 6 9}$ |
| NASDAQ $\sigma_{2}$ | $\mathbf{0 . 4 5}$ | 0.67 |

$\$ 10$ per trade for an account of size $\$ 10000$, we choose $K=0.001$ to simulate our strategy. This is close to the actual cost for a typical individual investor now or for a professional trader before the emergence of online discount brokers. We choose 10-year treasury bonds as the alternative risk-free investment instrument and use the annual yield released on the Federal Reserve Statistical Release web site [10]. This gives us an average yield of $6.7 \%$ per year from 1962-2008 and $5.4 \%$ from 1991-2008. However, to be more realistic we use the actual yield (see Table 14 in the appendix) when holding the bonds. We also take advantage of knowing $\sigma_{1}$ and $\sigma_{2}$ for the bull and bear markets. Solving the obstacle problems we derive for each index two sets of thresholds corresponding to $\sigma_{1}$ and $\sigma_{2}$, respectively, as listed in Table 9. Since when holding bonds and looking for a signal to switch to a stock position, we anticipate entering a bull market whose volatility is better represented by $\sigma_{1}$. Therefore, it is reasonable to choose the upper threshold related to $\sigma_{1}$. Similarly, for a signal to exit a stock position, we should use the lower threshold corresponding to $\sigma_{2}$. Thus, in conducting our test, we use the boldfaced thresholds in Table 9.

The discussions so far are based on the raw index values. This is appropriate since when pursuing market trend the raw price is what market agents can directly observe and react upon. However, for a long investment horizon, we also need to consider the effect of dividend. In this aspect the three indices are different. DJIA already includes dividend in its computation while NASDAQ and SP500 do not. The dividend paid by companies listed on NASDAQ is small; for example, the estimated 2009 average dividend for the 100 largest companies (who usually pay more dividend compared to smaller ones) listed in NASDAQ is only $0.68 \%$ (see [15]). Moreover, NASDAQ is tested for a shorter period of time. Thus, omitting dividend in the test for the NASDAQ index does not skew the comparison much. This is not the case for the SP500 index, which averages an annual dividend of about $2 \%$. We use the annual dividend compiled in [9] to compensate the raw gains for using the trend following strategy to trade the SP500 index.

The testing results for trading the NASDAQ, SP500, and DJIA indices are summarized in Table 10, and the trading details for the trend following strategy are contained in Tables 15, 16 , and 17 in the appendix, respectively. Taking NASDAQ as an example, using our trend following trading strategy, one dollar invested in the beginning of 1991 returns 8.82 at the end of 2008. By comparison, one dollar invested in the NASDAQ index using the buy and hold strategy in the same period returns only 4.24 while when invested in 10-year bonds it returns 2.63. The stories for the SP500 and DJIA are similar.

The average percentage gains per trade listed in Table 10 indicate that there is room for

Table 10
Testing results for trend following trading strategies. Legend: TF-trend following, BH—buy and hold, and $G$-average $\%$ gain per trade.

| Index (time frame) | TF | BH | $10 y$ bonds | No. trades | G |
| :---: | :---: | :---: | :---: | :---: | :---: |
| NASDAQ (1991-2008) | 8.82 | 4.24 | 2.63 | 66 | 3.35 |
| SP500 (1962-2008) | 64.98 | 33.5 | 23.44 | 80 | 5.36 |
| DJIA (1962-2008) | 26.03 | 12.11 | 23.44 | 80 | 4.16 |

the trend following method to absorb a higher trading cost and still outperform the buy and hold method. Using the NASDAQ test as an example, the same 66 trades with a trading cost of $1 \%$ per trade, the total return of the trend following method will be 4.64 , still higher than the 4.24 from the buy and hold method. In theory, changing $K$ to recalculate the buy and sell thresholds will yield even better returns, and this is confirmed also by the simulation reported in Table 7. However, this not the case here. In fact, using thresholds corresponding to $K=0.01$ to test the NASDAQ data, we get a total return of only 4.04 with 22 trades. On the other hand, the same test with $K=0.0001$ yields a total return of 10.8 with 130 trades. This corresponds to an average gain of $1.847 \%$ per trade. Had a 0.001 trading cost been charged on these same trades, we would have ended up with a total return of 9.5 , which is higher than the return of 8.82 tested with the "optimized" thresholds corresponding to $K=0.001$. Moreover, the relative advantage of the trend following method over the buy and hold method in the testing results for the stock indices is not as good as those from the simulations in the previous subsection. These indicate that the regime switching geometric Brownian motion model with those parameter values is only an approximation of the real markets.
5. Conclusion. We show that under a regime switching model a trend following trading system can be justified as an optimal trading strategy with a discounted reward function of trading one share of stock with a fixed percentage transaction cost. The optimal trading strategy has a simple implementable approximation. Extensive simulations and tests on historical stock market data show that this "optimal" trend following trading strategy (using constant thresholds $p_{b}^{*}(t)$ and $\left.p_{s}^{*}(t)\right)$ significantly outperforms the buy and hold strategy and is robust when parameters are perturbed. This investigation provides a useful theoretical framework for the widely used trend following trading methods.

## Appendix.

A.1. The smoothness of the free boundaries $p_{b}^{*}(t)$ and $p_{s}^{*}(t)$. By changing variables

$$
y=\log \left(\frac{p}{p-1}\right), \quad w(y, t)=Z(p, t),
$$

the double obstacle problem (2.9)-(2.11) becomes

$$
\min \left\{\max \left\{-\partial_{t} w-\mathcal{L}_{1} w, w-(1+K)\right\}, w-(1-K)\right\}=0
$$

in $(-\infty,+\infty) \times[0, T)$, with the terminal condition $w(y, T)=(1-K)$, where

$$
\begin{aligned}
\mathcal{L}_{1}= & \frac{\left(\mu_{1}-\mu_{2}\right)^{2}}{2 \sigma^{2}} \partial_{y y}+\left[\left(\mu_{1}-\mu_{2}-\lambda_{2}+\lambda_{2}-\frac{\left(\mu_{1}-\mu_{2}\right)^{2}}{2 \sigma^{2}}\right)+\frac{\left(\mu_{1}-\mu_{2}\right)^{2}}{\sigma^{2}} \frac{e^{y}}{e^{y}+1}\right] \partial_{y} \\
& +\left(\mu_{1} \frac{e^{y}}{e^{y}+1}+\mu_{2} \frac{1}{e^{y}+1}-\rho\right)
\end{aligned}
$$

To show the smoothness of the free boundaries, it suffices to verify the so-called cone property (cf. [6, 27]):

$$
(T-t) \partial_{t} w+C \partial_{y} w \geq 0
$$

with some constant $C>0$, locally uniformly for $y \in(-\infty,+\infty)$. For illustration, ${ }^{4}$ let us restrict our attention to the region $\{(y, t): 1-K<w(y, t)<1+K\}$, in which $-\partial_{t} w-\mathcal{L}_{1} w=0$. Differentiating the above equation w.r.t. $y$, we have

$$
\begin{aligned}
\left(-\partial_{t}-\mathcal{L}_{2}\right)\left(\partial_{y} w\right) & =\frac{e^{y}}{\left(e^{y}+1\right)^{2}}\left(\frac{\left(\mu_{1}-\mu_{2}\right)^{2}}{\sigma^{2}} \partial_{y} w+\left(\mu_{1}-\mu_{2}\right) w\right) \\
& \geq \frac{e^{y}}{\left(e^{y}+1\right)^{2}}\left(\mu_{1}-\mu_{2}\right)(1-K)
\end{aligned}
$$

where $\mathcal{L}_{2}=\mathcal{L}_{1}-\left(\lambda_{1} e^{y}+\lambda_{2} e^{-y}\right)$. On the other hand,

$$
\left(-\partial_{t}-\mathcal{L}_{2}\right)\left((T-t) \partial_{t} w\right)=\left[\left(\lambda_{1} e^{y}+\lambda_{2} e^{-y}\right)(T-t)+1\right] \partial_{t} w
$$

It is easy to see that $\partial_{t} w$ is uniformly bounded and $\frac{e^{y}}{\left(e^{y}+1\right)^{2}}$ has a local positive lower bound. By using the auxiliary function $\psi\left(y, t ; y_{0}\right)=e^{C_{1} t}\left(y-y_{0}\right)^{2}$, with some positive constant $C_{1}$, as adopted in $[6,27]$, we can infer from the maximum principle that

$$
(T-t) \partial_{t} w+C \partial_{y} w+\psi\left(y, t ; y_{0}\right) \geq 0
$$

for an appropriate $C>0$. The desired result follows by taking $y=y_{0}$.

[^83]
## A.2. Statistics for bull and bear markets of NASDAQ, SP500, and DJIA indices.

Table 11
Statistics of NASDAQ bull and bear markets (1991-2009).

| Top/bottom | Index | \% Move | Mean | Stdev | Duration |
| :---: | :---: | :---: | :---: | :---: | :---: |
| $08 / 22 / 1990$ | 374.84 |  |  |  |  |
| $07 / 20 / 1998$ | 2014.25 | 4.37362608 | 0.000841593 | 0.009228481 | 1998 |
| $10 / 08 / 1998$ | 1419.12 | -0.295459849 | -0.005991046 | 0.024869575 | 57 |
| $03 / 10 / 2000$ | 5048.62 | 2.557570889 | 0.003450954 | 0.018083041 | 358 |
| $04 / 04 / 2001$ | 1638.8 | -0.675396445 | -0.004165933 | 0.032639469 | 269 |
| $05 / 22 / 2001$ | 2313.85 | 0.411917257 | 0.00953807 | 0.031383788 | 33 |
| $09 / 21 / 2001$ | 1423.19 | -0.384925557 | -0.005883363 | 0.020987512 | 81 |
| $01 / 04 / 2002$ | 2059.38 | 0.447016913 | 0.004609728 | 0.020812769 | 72 |
| $10 / 09 / 2002$ | 1114.11 | -0.45900708 | -0.003144999 | 0.021500913 | 192 |
| $01 / 26 / 2004$ | 2153.83 | 0.933229214 | 0.001980534 | 0.015351658 | 325 |
| $08 / 12 / 2004$ | 1752.49 | -0.186337826 | -0.00138275 | 0.011873996 | 138 |
| $10 / 31 / 2007$ | 2859.12 | 0.63146152 | 0.000581948 | 0.008861377 | 811 |
| $03 / 09 / 2009$ | 1268.64 | -0.556283052 | -0.002345953 | 0.024687197 | 339 |

Table 12
Statistics of SP500 bull and bear markets (1962-2009).

| Top/bottom | Index | \% Move | Mean | Stdev | Duration |
| :---: | :---: | :---: | :---: | :---: | :---: |
| $01 / 03 / 1962$ | 71.13 |  |  |  |  |
| $06 / 26 / 1962$ | 52.32 | -0.264445382 | -0.002538268 | 0.011710656 | 121 |
| $02 / 09 / 1966$ | 94.06 | 0.797782875 | 0.000639029 | 0.005328904 | 913 |
| $10 / 07 / 1966$ | 73.2 | -0.221773336 | -0.001460123 | 0.007778645 | 167 |
| $11 / 29 / 1968$ | 108.37 | 0.480464481 | 0.000736578 | 0.005908178 | 516 |
| $05 / 26 / 1970$ | 69.29 | -0.360616407 | -0.00119353 | 0.007244905 | 369 |
| $01 / 11 / 1973$ | 120.24 | 0.735315341 | 0.000806951 | 0.00682808 | 665 |
| $10 / 03 / 1974$ | 62.28 | -0.482035928 | -0.001489909 | 0.011230379 | 436 |
| $09 / 21 / 1976$ | 107.83 | 0.731374438 | 0.001067083 | 0.010019283 | 497 |
| $03 / 06 / 1978$ | 86.9 | -0.194101827 | -0.000549579 | 0.00600745 | 366 |
| $11 / 28 / 1980$ | 140.52 | 0.61703107 | 0.00068538 | 0.008509688 | 691 |
| $08 / 12 / 1982$ | 102.6 | -0.269854825 | -0.000728014 | 0.008772591 | 430 |
| $08 / 25 / 1987$ | 336.77 | 2.282358674 | 0.000932206 | 0.00905962 | 1274 |
| $12 / 04 / 1987$ | 223.92 | -0.335095169 | -0.005525613 | 0.035700189 | 71 |
| $07 / 16 / 1990$ | 368.95 | 0.647686674 | 0.000747921 | 0.009679323 | 659 |
| $10 / 11 / 1990$ | 295.46 | -0.199186882 | -0.003455122 | 0.012748722 | 62 |
| $07 / 17 / 1998$ | 1186.75 | 3.016618155 | 0.000699894 | 0.007671028 | 1962 |
| $08 / 31 / 1998$ | 957.28 | -0.193360017 | -0.006642962 | 0.017942563 | 31 |
| $03 / 24 / 2000$ | 1527.46 | 0.595625104 | 0.001002091 | 0.013475437 | 395 |
| $10 / 09 / 2002$ | 776.76 | -0.491469498 | -0.001059809 | 0.014432882 | 637 |
| $10 / 05 / 2007$ | 1557.59 | 1.005239714 | 0.000531502 | 0.008593452 | 1256 |
| $03 / 09 / 2009$ | 676.53 | -0.565655917 | -0.001833476 | 0.023888376 | 357 |

Table 13
Statistics of DJIA bull and bear markets (1962-2009).

| Top/bottom | Index | \% Move | Mean | Stdev | Duration |
| :---: | :---: | :---: | :---: | :---: | :---: |
| $01 / 02 / 1962$ | 724.7 |  |  |  |  |
| $06 / 26 / 1962$ | 535.7 | -0.260797571 | -0.002476914 | 0.010625788 | 122 |
| $02 / 09 / 1966$ | 995.2 | 0.857756207 | 0.000674868 | 0.005654829 | 914 |
| $10 / 07 / 1966$ | 744.3 | -0.252110129 | -0.00170399 | 0.007878357 | 167 |
| $12 / 03 / 1968$ | 985.2 | 0.323659815 | 0.000525586 | 0.006114526 | 519 |
| $05 / 26 / 1970$ | 631.2 | -0.359317905 | -0.001204596 | 0.007520823 | 367 |
| $01 / 11 / 1973$ | 1051.7 | 0.666191381 | 0.000742507 | 0.007302779 | 665 |
| $12 / 06 / 1974$ | 577.6 | -0.450793953 | -0.001389029 | 0.012603452 | 481 |
| $09 / 21 / 1976$ | 1014.8 | 0.756925208 | 0.0012054 | 0.009654537 | 453 |
| $02 / 28 / 1978$ | 742.1 | -0.268722901 | -0.000804275 | 0.006918213 | 363 |
| $04 / 27 / 1981$ | 1024 | 0.379867942 | 0.0003929 | 0.008882315 | 797 |
| $08 / 12 / 1982$ | 776.9 | -0.241308594 | -0.00082839 | 0.008056537 | 328 |
| $08 / 25 / 1987$ | 2722.42 | 2.504209036 | 0.000983971 | 0.009436921 | 1273 |
| $10 / 19 / 1987$ | 1738.74 | -0.361325585 | -0.011256546 | 0.043065972 | 38 |
| $07 / 17 / 1990$ | 2999.75 | 0.725243567 | 0.000416497 | 0.015688021 | 693 |
| $10 / 11 / 1990$ | 2365.1 | -0.211567631 | -0.00383401 | 0.013142105 | 61 |
| $07 / 17 / 1998$ | 9337.97 | 2.948234747 | 0.000690436 | 0.007984466 | 1962 |
| $08 / 31 / 1998$ | 7539.07 | -0.192643583 | -0.006654443 | 0.016969817 | 31 |
| $01 / 14 / 2000$ | 11722.98 | 0.554963676 | 0.001079512 | 0.012092007 | 347 |
| $10 / 09 / 2002$ | 7286.27 | -0.378462643 | -0.000675652 | 0.014062882 | 685 |
| $10 / 09 / 2007$ | 14164.53 | 0.944002899 | 0.000504876 | 0.008380232 | 1258 |
| $03 / 09 / 2009$ | 6547.05 | -0.537785581 | -0.002149751 | 0.021943976 | 355 |

A.3. Yield for 10 -year US treasury bonds. Table 14 summarizes the annual yield of 10-year US treasury bonds as released on the Federal Reserve Statistical Release web site [10].

Table 14
Yield of 10-year bonds (1962-2008).

| Year | Yield | Year | Yield | Year | Yield |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 1962 | 3.95 | 1978 | 8.41 | 1994 | 7.09 |
| 1963 | 4 | 1979 | 9.43 | 1995 | 6.57 |
| 1964 | 4.19 | 1980 | 11.43 | 1996 | 6.44 |
| 1965 | 4.28 | 1981 | 13.92 | 1997 | 6.35 |
| 1966 | 4.93 | 1982 | 13.01 | 1998 | 5.26 |
| 1967 | 5.07 | 1983 | 11.1 | 1999 | 5.65 |
| 1968 | 5.64 | 1984 | 12.46 | 2000 | 6.03 |
| 1969 | 6.67 | 1985 | 10.62 | 2001 | 5.02 |
| 1970 | 7.35 | 1986 | 7.67 | 2002 | 4.61 |
| 1971 | 6.16 | 1987 | 8.39 | 2003 | 4.01 |
| 1972 | 6.21 | 1988 | 8.85 | 2004 | 4.27 |
| 1973 | 6.85 | 1989 | 8.49 | 2005 | 4.29 |
| 1974 | 7.56 | 1990 | 8.55 | 2006 | 4.8 |
| 1975 | 7.99 | 1991 | 7.86 | 2007 | 4.63 |
| 1976 | 7.61 | 1992 | 7.01 | 2008 | 3.66 |
| 1977 | 7.42 | 1993 | 5.87 |  |  |

## A.4. Transactions for the trend following strategy on NASDAQ, SP500, and DJIA indices.

Table 15
NASDAQ investment test (1991-2008).

| Symbol | Buy date | Buy price | Sell date | Sell price | Gain |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Bond | 01/02/1991 | 1 | 01/23/1991 | 1.004522 | 1.004522 |
| COMPQX | 01/23/1991 | 383.91 | 06/24/1991 | 475.23 | 1.235393 |
| Bond | 06/24/1991 | 1 | 08/13/1991 | 1.010767 | 1.010767 |
| COMPQX | 08/13/1991 | 514.4 | 03/27/1992 | 604.67 | 1.173135 |
| Bond | 03/27/1992 | 1 | 09/14/1992 | 1.032841 | 1.032841 |
| COMPQX | 09/14/1992 | 594.21 | 10/05/1992 | 565.21 | 0.9492933 |
| Bond | 10/05/1992 | 1 | 11/05/1992 | 1.005954 | 1.005954 |
| COMPQX | 11/05/1992 | 614.08 | 02/16/1993 | 665.39 | 1.081389 |
| Bond | 02/16/1993 | 1 | 05/26/1993 | 1.015921 | 1.015921 |
| COMPQX | 05/26/1993 | 704.09 | 11/22/1993 | 738.13 | 1.046249 |
| Bond | 11/22/1993 | 1 | 08/24/1994 | 1.053418 | 1.053418 |
| COMPQX | 08/24/1994 | 751.72 | 11/22/1994 | 741.21 | 0.9840468 |
| Bond | 11/22/1994 | 1 | 02/09/1995 | 1.01422 | 1.01422 |
| COMPQX | 02/09/1995 | 785.44 | 10/04/1995 | 1002.27 | 1.27351 |
| Bond | 10/04/1995 | 1 | 02/01/1996 | 1.021173 | 1.021173 |
| COMPQX | 02/01/1996 | 1069.46 | 06/18/1996 | 1183.08 | 1.104028 |
| Bond | 06/18/1996 | 1 | 09/13/1996 | 1.01535 | 1.01535 |
| COMPQX | 09/13/1996 | 1188.67 | 02/27/1997 | 1312.66 | 1.102101 |
| Bond | 02/27/1997 | 1 | 05/02/1997 | 1.011134 | 1.011134 |
| COMPQX | 05/02/1997 | 1305.33 | 10/27/1997 | 1535.09 | 1.173665 |
| Bond | 10/27/1997 | 1 | 02/02/1998 | 1.014123 | 1.014123 |
| COMPQX | 02/02/1998 | 1652.89 | 05/26/1998 | 1778.09 | 1.073595 |
| Bond | 05/26/1998 | 1 | 06/24/1998 | 1.004179 | 1.004179 |
| COMPQX | 06/24/1998 | 1877.76 | 08/03/1998 | 1851.1 | 0.9838306 |
| Bond | 08/03/1998 | 1 | 11/04/1998 | 1.013402 | 1.013402 |
| COMPQX | 11/04/1998 | 1823.57 | 05/25/1999 | 2380.9 | 1.303015 |
| Bond | 05/25/1999 | 1 | 06/18/1999 | 1.003715 | 1.003715 |
| COMPQX | 06/18/1999 | 2563.44 | 08/04/1999 | 2540 | 0.9888744 |
| Bond | 08/04/1999 | 1 | 08/24/1999 | 1.003096 | 1.003096 |
| COMPQX | 08/24/1999 | 2752.37 | 10/18/1999 | 2689.15 | 0.9750766 |
| Bond | 10/18/1999 | 1 | 10/28/1999 | 1.001548 | 1.001548 |
| COMPQX | 10/28/1999 | 2875.22 | 04/03/2000 | 4223.68 | 1.466056 |
| Bond | 04/03/2000 | 1 | 07/12/2000 | 1.016521 | 1.016521 |
| COMPQX | 07/12/2000 | 4099.59 | 07/27/2000 | 3842.23 | 0.9353486 |
| Bond | 07/27/2000 | 1 | 08/31/2000 | 1.005782 | 1.005782 |
| COMPQX | 08/31/2000 | 4206.35 | 09/08/2000 | 3978.41 | 0.9439188 |
| Bond | 09/08/2000 | 1 | 05/02/2001 | 1.032458 | 1.032458 |
| COMPQX | 05/02/2001 | 2220.6 | 05/11/2001 | 2107.43 | 0.9471381 |
| Bond | 05/11/2001 | 1 | 05/21/2001 | 1.001375 | 1.001375 |
| COMPQX | 05/21/2001 | 2305.59 | 05/30/2001 | 2084.5 | 0.9022987 |
| Bond | 05/30/2001 | 1 | 10/25/2001 | 1.020355 | 1.020355 |
| COMPQX | 10/25/2001 | 1775.47 | 10/29/2001 | 1699.52 | 0.9553082 |
| Bond | 10/29/2001 | 1 | 11/06/2001 | 1.0011 | 1.0011 |
| COMPQX | 11/06/2001 | 1835.08 | 02/04/2002 | 1855.53 | 1.009122 |

NASDAQ investment test (1991-2008) (continued).

| Symbol | Buy date | Buy price | Sell date | Sell price | Gain |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Bond | $02 / 04 / 2002$ | 1 | $11 / 01 / 2002$ | 1.034101 | 1.034101 |
| COMPQX | $11 / 01 / 2002$ | 1360.7 | $01 / 27 / 2003$ | 1325.27 | 0.9720141 |
| Bond | $01 / 27 / 2003$ | 1 | $03 / 17 / 2003$ | 1.005383 | 1.005383 |
| COMPQX | $03 / 17 / 2003$ | 1392.27 | $02 / 23 / 2004$ | 2007.52 | 1.43902 |
| Bond | $02 / 23 / 2004$ | 1 | $04 / 05 / 2004$ | 1.004913 | 1.004913 |
| COMPQX | $04 / 05 / 2004$ | 2079.12 | $04 / 16 / 2004$ | 1995.74 | 0.9579766 |
| Bond | $04 / 16 / 2004$ | 1 | $10 / 01 / 2004$ | 1.019654 | 1.019654 |
| COMPQX | $10 / 01 / 2004$ | 1942.2 | $01 / 20 / 2005$ | 2045.88 | 1.051276 |
| Bond | $01 / 20 / 2005$ | 1 | $05 / 23 / 2005$ | 1.014457 | 1.014457 |
| COMPQX | $05 / 23 / 2005$ | 2056.65 | $09 / 21 / 2005$ | 2106.64 | 1.022258 |
| Bond | $09 / 21 / 2005$ | 1 | $11 / 11 / 2005$ | 1.005994 | 1.005994 |
| COMPQX | $11 / 11 / 2005$ | 2202.47 | $05 / 11 / 2006$ | 2272.7 | 1.029823 |
| Bond | $05 / 11 / 2006$ | 1 | $09 / 05 / 2006$ | 1.015386 | 1.015386 |
| COMPQX | $09 / 05 / 2006$ | 2205.7 | $02 / 27 / 2007$ | 2407.86 | 1.08947 |
| Bond | $02 / 27 / 2007$ | 1 | $07 / 12 / 2007$ | 1.017125 | 1.017125 |
| COMPQX | $07 / 12 / 2007$ | 2701.73 | $07 / 26 / 2007$ | 2599.34 | 0.9601779 |
| Bond | $07 / 26 / 2007$ | 1 | $09 / 26 / 2007$ | 1.007865 | 1.007865 |
| COMPQX | $09 / 26 / 2007$ | 2699.03 | $11 / 09 / 2007$ | 2627.94 | 0.9717135 |
| Bond | $11 / 09 / 2007$ | 1 | $04 / 24 / 2008$ | 1.016746 | 1.016746 |
| COMPQX | $04 / 24 / 2008$ | 2428.92 | $06 / 11 / 2008$ | 2394.01 | 0.9836561 |
| Bond | $06 / 11 / 2008$ | 1 | $08 / 11 / 2008$ | 1.006117 | 1.006117 |
| COMPQX | $08 / 11 / 2008$ | 2439.95 | $09 / 03 / 2008$ | 2333.73 | 0.9545534 |
| Bond | $09 / 03 / 2008$ | 1 | $12 / 31 / 2008$ | 1.011933 | 1.011933 |

Table 16
SP500 investment test (1962-2008).

| Symbol | Buy date | Buy price | Sell date | Sell price | Gain | Dividend |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Bond | $01 / 03 / 1962$ | 1 | $11 / 12 / 1962$ | 1.033873 | 1.033873 |  |
| SP500 | $11 / 12 / 1962$ | 59.59 | $06 / 09 / 1965$ | 85.04 | 1.424231 | 0.085 |
| Bond | $06 / 09 / 1965$ | 1 | $09 / 10 / 1965$ | 1.010905 | 1.010905 |  |
| SP500 | $09 / 10 / 1965$ | 89.12 | $03 / 02 / 1966$ | 89.15 | 0.998336 | 0.021 |
| Bond | $03 / 02 / 1966$ | 1 | $11 / 16 / 1966$ | 1.034983 | 1.034983 |  |
| SP500 | $11 / 16 / 1966$ | 82.37 | $11 / 03 / 1967$ | 91.78 | 1.112012 | 0.033 |
| Bond | $11 / 03 / 1967$ | 1 | $04 / 03 / 1968$ | 1.023487 | 1.023487 |  |
| SP500 | $04 / 03 / 1968$ | 93.47 | $01 / 07 / 1969$ | 101.22 | 1.080748 | 0.023 |
| Bond | $01 / 07 / 1969$ | 1 | $05 / 05 / 1969$ | 1.021563 | 1.021563 |  |
| SP500 | $05 / 05 / 1969$ | 104.37 | $06 / 09 / 1969$ | 101.2 | 0.967688 | 0.003 |
| Bond | $06 / 09 / 1969$ | 1 | $08 / 24 / 1970$ | 1.088804 | 1.088804 |  |
| SP500 | $08 / 24 / 1970$ | 80.99 | $06 / 22 / 1971$ | 97.59 | 1.202554 | 0.028 |
| Bond | $06 / 22 / 1971$ | 1 | $09 / 03 / 1971$ | 1.01232 | 1.01232 |  |
| SP500 | $09 / 03 / 1971$ | 100.69 | $10 / 20 / 1971$ | 95.65 | 0.9480455 | 0.004 |
| Bond | $10 / 20 / 1971$ | 1 | $12 / 17 / 1971$ | 1.009789 | 1.009789 |  |
| SP500 | $12 / 17 / 1971$ | 100.26 | $02 / 27 / 1973$ | 110.9 | 1.103912 | 0.034 |
| Bond | $02 / 27 / 1973$ | 1 | $07 / 24 / 1973$ | 1.027588 | 1.027588 |  |
| SP500 | $07 / 24 / 1973$ | 108.14 | $08 / 13 / 1973$ | 103.71 | 0.9571165 | 0.002 |
| Bond | $08 / 13 / 1973$ | 1 | $10 / 05 / 1973$ | 1.009947 | 1.009947 |  |
| SP500 | $10 / 05 / 1973$ | 109.85 | $11 / 14 / 1973$ | 102.45 | 0.9307701 | 0.004 |
| Bond | $11 / 14 / 1973$ | 1 | $03 / 13 / 1974$ | 1.024648 | 1.024648 |  |
| SP500 | $03 / 13 / 1974$ | 99.74 | $03 / 28 / 1974$ | 94.82 | 0.9487704 | 0.002 |

SP500 investment test (1962-2008) (continued).

| Symbol | Buy date | Buy price | Sell date | Sell price | Gain | Dividend |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Bond | 03/28/1974 | 1 | 01/10/1975 | 1.063044 | 1.063044 |  |
| SP500 | 01/10/1975 | 72.61 | 08/04/1975 | 87.15 | 1.197847 | 0.023 |
| Bond | 08/04/1975 | 1 | 10/13/1975 | 1.015323 | 1.015323 |  |
| SP500 | 10/13/1975 | 89.46 | 04/06/1977 | 97.91 | 1.092267 | 0.063 |
| Bond | 04/06/1977 | 1 | 04/14/1978 | 1.085943 | 1.085943 |  |
| SP500 | 04/14/1978 | 92.92 | 10/20/1978 | 97.95 | 1.052024 | 0.028 |
| Bond | 10/20/1978 | 1 | 01/12/1979 | 1.021702 | 1.021702 |  |
| SP500 | 01/12/1979 | 99.93 | 10/22/1979 | 100.71 | 1.00579 | 0.043 |
| Bond | 10/22/1979 | 1 | 11/26/1979 | 1.009043 | 1.009043 |  |
| SP500 | 11/26/1979 | 106.8 | 03/17/1980 | 102.26 | 0.9555756 | 0.015 |
| Bond | 03/17/1980 | 1 | 05/23/1980 | 1.020981 | 1.020981 |  |
| SP500 | 05/23/1980 | 110.62 | 08/27/1981 | 123.51 | 1.114292 | 0.067 |
| Bond | 08/27/1981 | 1 | 11/03/1981 | 1.025933 | 1.025933 |  |
| SP500 | 11/03/1981 | 124.8 | 11/16/1981 | 120.24 | 0.9615346 | 0.002 |
| Bond | 11/16/1981 | 1 | 04/23/1982 | 1.056317 | 1.056317 |  |
| SP500 | 04/23/1982 | 118.64 | 05/26/1982 | 113.11 | 0.9514816 | 0.005 |
| Bond | 05/26/1982 | 1 | 08/20/1982 | 1.030654 | 1.030654 |  |
| SP500 | 08/20/1982 | 113.02 | 02/08/1984 | 155.85 | 1.376202 | 0.067 |
| Bond | 02/08/1984 | 1 | 08/02/1984 | 1.060081 | 1.060081 |  |
| SP500 | 08/02/1984 | 157.99 | 09/17/1985 | 181.36 | 1.145625 | 0.047 |
| Bond | 09/17/1985 | 1 | 11/05/1985 | 1.014257 | 1.014257 |  |
| SP500 | 11/05/1985 | 192.37 | 10/16/1987 | 282.7 | 1.466625 | 0.071 |
| Bond | 10/16/1987 | 1 | 02/29/1988 | 1.032975 | 1.032975 |  |
| SP500 | 02/29/1988 | 267.82 | 08/06/1990 | 334.43 | 1.246214 | 0.089 |
| Bond | 08/06/1990 | 1 | 12/03/1990 | 1.027875 | 1.027875 |  |
| SP500 | 12/03/1990 | 324.1 | 01/09/1991 | 311.49 | 0.95917 | 0.004 |
| Bond | 01/09/1991 | 1 | 01/18/1991 | 1.001938 | 1.001938 |  |
| SP500 | 01/18/1991 | 332.23 | 03/30/1994 | 445.55 | 1.338407 | 0.096 |
| Bond | 03/30/1994 | 1 | 08/24/1994 | 1.028554 | 1.028554 |  |
| SP500 | 08/24/1994 | 469.03 | 07/15/1996 | 629.8 | 1.340086 | 0.045 |
| Bond | 07/15/1996 | 1 | 08/22/1996 | 1.006705 | 1.006705 |  |
| SP500 | 08/22/1996 | 670.68 | 08/31/1998 | 957.28 | 1.424473 | 0.032 |
| Bond | 08/31/1998 | 1 | 09/23/1998 | 1.003314 | 1.003314 |  |
| SP500 | 09/23/1998 | 1066.09 | 10/01/1998 | 986.39 | 0.9233904 | 0.000 |
| Bond | 10/01/1998 | 1 | 10/22/1998 | 1.003026 | 1.003026 |  |
| SP500 | 10/22/1998 | 1078.48 | 08/10/1999 | 1281.43 | 1.185805 | 0.009 |
| Bond | 08/10/1999 | 1 | 08/25/1999 | 1.002322 | 1.002322 |  |
| SP500 | 08/25/1999 | 1381.79 | 08/30/1999 | 1324.02 | 0.9562755 | 0.000 |
| Bond | 08/30/1999 | 1 | 10/28/1999 | 1.009133 | 1.009133 |  |
| SP500 | 10/28/1999 | 1342.44 | 10/11/2000 | 1364.59 | 1.014467 | 0.012 |
| Bond | 10/11/2000 | 1 | 10/31/2000 | 1.003304 | 1.003304 |  |
| SP500 | 10/31/2000 | 1429.4 | 11/10/2000 | 1365.98 | 0.9537205 | 0.000 |
| Bond | 11/10/2000 | 1 | 01/23/2001 | 1.010177 | 1.010177 |  |
| SP500 | 01/23/2001 | 1360.4 | 02/20/2001 | 1278.94 | 0.9382402 | 0.001 |
| Bond | 02/20/2001 | 1 | 05/16/2001 | 1.01169 | 1.01169 |  |
| SP500 | 05/16/2001 | 1284.99 | 06/18/2001 | 1208.43 | 0.938539 | 0.001 |
| Bond | 06/18/2001 | 1 | 11/19/2001 | 1.02118 | 1.02118 |  |
| SP500 | 11/19/2001 | 1151.06 | 07/22/2002 | 819.85 | 0.7108319 | 0.012 |
| Bond | 07/22/2002 | 1 | 07/29/2002 | 1.000884 | 1.000884 |  |
| SP500 | 07/29/2002 | 898.96 | 09/24/2002 | 819.29 | 0.9095526 | 0.003 |

SP500 investment test (1962-2008) (continued).

| Symbol | Buy date | Buy price | Sell date | Sell price | Gain | Dividend |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Bond | $09 / 24 / 2002$ | 1 | $10 / 11 / 2002$ | 1.002147 | 1.002147 |  |
| SP500 | $10 / 11 / 2002$ | 835.32 | $01 / 18 / 2008$ | 1325.19 | 1.583273 | 0.096 |
| Bond | $01 / 18 / 2008$ | 1 | $03 / 18 / 2008$ | 1.006016 | 1.006016 |  |
| SP500 | $03 / 18 / 2008$ | 1330.74 | $07 / 07 / 2008$ | 1252.31 | 0.9391808 | 0.009 |
| Bond | $07 / 07 / 2008$ | 1 | $08 / 08 / 2008$ | 1.003209 | 1.003209 |  |
| SP500 | $08 / 08 / 2008$ | 1296.31 | $10 / 09 / 2008$ | 909.92 | 0.700527 | 0.005 |
| Bond | $10 / 09 / 2008$ | 1 | $10 / 30 / 2008$ | 1.002106 | 1.002106 |  |
| SP500 | $10 / 30 / 2008$ | 954.09 | $12 / 31 / 2008$ | 903.25 | 0.9467136 | 0.005 |

Table 17
DJIA investment test (1962-2008).

| Symbol | Buy date | Buy price | Sell date | Sell price | Gain |
| :---: | :---: | :---: | :---: | :---: | :---: |
| DJ-30 | $11 / 02 / 1962$ | 604.6 | $07 / 22 / 1963$ | 688.7 | 1.136822 |
| Bond | $07 / 22 / 1963$ | 1 | $09 / 24 / 1963$ | 1.007014 | 1.007014 |
| DJ-30 | $09 / 24 / 1963$ | 746 | $11 / 22 / 1963$ | 711.5 | 0.9518458 |
| Bond | $11 / 22 / 1963$ | 1 | $12 / 05 / 1963$ | 1.001425 | 1.001425 |
| DJ-30 | $12 / 05 / 1963$ | 763.9 | $06 / 09 / 1965$ | 879.8 | 1.149418 |
| Bond | $06 / 09 / 1965$ | 1 | $09 / 16 / 1965$ | 1.011609 | 1.011609 |
| DJ-30 | $09 / 16 / 1965$ | 931.2 | $03 / 01 / 1966$ | 938.2 | 1.005502 |
| Bond | $03 / 01 / 1966$ | 1 | $01 / 12 / 1967$ | 1.044033 | 1.044033 |
| DJ-30 | $01 / 12 / 1967$ | 830 | $10 / 31 / 1967$ | 879.7 | 1.05776 |
| Bond | $10 / 31 / 1967$ | 1 | $04 / 08 / 1968$ | 1.024723 | 1.024723 |
| DJ-30 | $04 / 08 / 1968$ | 884.4 | $01 / 08 / 1969$ | 921.3 | 1.03964 |
| Bond | $01 / 08 / 1969$ | 1 | $05 / 06 / 1969$ | 1.021563 | 1.021563 |
| DJ-30 | $05 / 06 / 1969$ | 962.1 | $06 / 06 / 1969$ | 924.8 | 0.9593082 |
| Bond | $06 / 06 / 1969$ | 1 | $07 / 17 / 1970$ | 1.081756 | 1.081756 |
| DJ-30 | $07 / 17 / 1970$ | 735.1 | $06 / 21 / 1971$ | 876.5 | 1.18997 |
| Bond | $06 / 21 / 1971$ | 1 | $09 / 07 / 1971$ | 1.013164 | 1.013164 |
| DJ-30 | $09 / 07 / 1971$ | 916.5 | $10 / 18 / 1971$ | 872.4 | 0.9499784 |
| Bond | $10 / 18 / 1971$ | 1 | $12 / 29 / 1971$ | 1.012151 | 1.012151 |
| DJ-30 | $12 / 29 / 1971$ | 893.7 | $02 / 07 / 1973$ | 968.3 | 1.081306 |
| Bond | $02 / 07 / 1973$ | 1 | $07 / 25 / 1973$ | 1.031529 | 1.031529 |
| DJ-30 | $07 / 25 / 1973$ | 933 | $08 / 10 / 1973$ | 892.4 | 0.9545715 |
| Bond | $08 / 10 / 1973$ | 1 | $09 / 24 / 1973$ | 1.008445 | 1.008445 |
| DJ-30 | $09 / 24 / 1973$ | 936.7 | $11 / 14 / 1973$ | 869.9 | 0.9268284 |
| Bond | $11 / 14 / 1973$ | 1 | $02 / 27 / 1974$ | 1.021748 | 1.021748 |
| DJ-30 | $02 / 27 / 1974$ | 863.4 | $07 / 08 / 1974$ | 770.6 | 0.8907329 |
| Bond | $07 / 08 / 1974$ | 1 | $01 / 09 / 1975$ | 1.040497 | 1.040497 |
| DJ-30 | $01 / 09 / 1975$ | 645.3 | $08 / 05 / 1975$ | 810.2 | 1.253029 |
| Bond | $08 / 05 / 1975$ | 1 | $10 / 13 / 1975$ | 1.015104 | 1.015104 |
| DJ-30 | $10 / 13 / 1975$ | 837.8 | $10 / 12 / 1976$ | 932.4 | 1.110689 |
| Bond | $10 / 12 / 1976$ | 1 | $12 / 24 / 1976$ | 1.01522 | 1.01522 |
| DJ-30 | $12 / 24 / 1976$ | 996.1 | $02 / 08 / 1977$ | 942.2 | 0.9439972 |
| Bond | $02 / 08 / 1977$ | 1 | $04 / 14 / 1978$ | 1.099077 | 1.099077 |
| DJ-30 | $04 / 14 / 1978$ | 795.1 | $10 / 26 / 1978$ | 821.1 | 1.030635 |
| Bond | $10 / 26 / 1978$ | 1 | $01 / 15 / 1979$ | 1.020927 | 1.020927 |
| DJ-30 | $01 / 15 / 1979$ | 848.7 | $10 / 19 / 1979$ | 814.7 | 0.9580188 |
| Bond | $10 / 19 / 1979$ | 1 | $01 / 10 / 1980$ | 1.025992 | 1.025992 |
|  |  |  |  |  |  |
|  |  | 1 |  | 10 |  |

DJIA investment test (1962-2008) (continued).

| Symbol | Buy date | Buy price | Sell date | Sell price | Gain |
| :---: | :---: | :---: | :---: | :---: | :---: |
| DJ-30 | 01/10/1980 | 858.9 | 03/10/1980 | 818.9 | 0.9515219 |
| Bond | 03/10/1980 | 1 | 05/20/1980 | 1.022234 | 1.022234 |
| DJ-30 | 05/20/1980 | 832.5 | 12/11/1980 | 908.5 | 1.089109 |
| Bond | 12/11/1980 | 1 | 01/05/1981 | 1.009534 | 1.009534 |
| DJ-30 | 01/05/1981 | 992.7 | 07/21/1981 | 934.5 | 0.9394892 |
| Bond | 07/21/1981 | 1 | 04/23/1982 | 1.098377 | 1.098377 |
| DJ-30 | 04/23/1982 | 862.2 | 05/28/1982 | 819.5 | 0.9485745 |
| Bond | 05/28/1982 | 1 | 08/17/1982 | 1.028872 | 1.028872 |
| DJ-30 | 08/17/1982 | 831.2 | 02/03/1984 | 1197 | 1.437206 |
| Bond | 02/03/1984 | 1 | 08/02/1984 | 1.061788 | 1.061788 |
| DJ-30 | 08/02/1984 | 1166.1 | 10/16/1987 | 2246.74 | 1.92286 |
| Bond | 10/16/1987 | 1 | 02/29/1988 | 1.032975 | 1.032975 |
| DJ-30 | 02/29/1988 | 2071.62 | 08/06/1990 | 2716.34 | 1.308593 |
| Bond | 08/06/1990 | 1 | 12/05/1990 | 1.028344 | 1.028344 |
| DJ-30 | 12/05/1990 | 2610.4 | 01/09/1991 | 2470.3 | 0.9444374 |
| Bond | 01/09/1991 | 1 | 01/18/1991 | 1.001938 | 1.001938 |
| DJ-30 | 01/18/1991 | 2646.78 | 03/30/1994 | 3626.75 | 1.367509 |
| Bond | 03/30/1994 | 1 | 08/24/1994 | 1.028554 | 1.028554 |
| DJ-30 | 08/24/1994 | 3846.73 | 08/31/1998 | 7539.07 | 1.955945 |
| Bond | 08/31/1998 | 1 | 10/15/1998 | 1.006485 | 1.006485 |
| DJ-30 | 10/15/1998 | 8299.36 | 02/25/2000 | 9862.12 | 1.185922 |
| Bond | 02/25/2000 | 1 | 03/16/2000 | 1.003304 | 1.003304 |
| DJ-30 | 03/16/2000 | 10630.6 | 10/12/2000 | 10034.58 | 0.9420457 |
| Bond | 10/12/2000 | 1 | 12/05/2000 | 1.008921 | 1.008921 |
| DJ-30 | 12/05/2000 | 10898.72 | 03/22/2001 | 9389.48 | 0.8597984 |
| Bond | 03/22/2001 | 1 | 04/05/2001 | 1.001925 | 1.001925 |
| DJ-30 | 04/05/2001 | 9918.05 | 07/10/2001 | 10175.64 | 1.02392 |
| Bond | 07/10/2001 | 1 | 11/14/2001 | 1.017467 | 1.017467 |
| DJ-30 | 11/14/2001 | 9823.61 | 07/22/2002 | 7784.58 | 0.7908509 |
| Bond | 07/22/2002 | 1 | 07/29/2002 | 1.000884 | 1.000884 |
| DJ-30 | 07/29/2002 | 8711.88 | 09/24/2002 | 7683.13 | 0.8801503 |
| Bond | 09/24/2002 | 1 | 10/11/2002 | 1.002147 | 1.002147 |
| DJ-30 | 10/11/2002 | 7850.29 | 02/12/2003 | 7758.17 | 0.9862888 |
| Bond | 02/12/2003 | 1 | 03/17/2003 | 1.003626 | 1.003626 |
| DJ-30 | 03/17/2003 | 8141.92 | 04/15/2005 | 10087.51 | 1.236482 |
| Bond | 04/15/2005 | 1 | 11/11/2005 | 1.024682 | 1.024682 |
| DJ-30 | 11/11/2005 | 10686.04 | 01/22/2008 | 11971.19 | 1.118024 |
| Bond | 01/22/2008 | 1 | 03/18/2008 | 1.005615 | 1.005615 |
| DJ-30 | 03/18/2008 | 12392.66 | 06/26/2008 | 11453.42 | 0.9223616 |
| Bond | 06/26/2008 | 1 | 08/28/2008 | 1.006317 | 1.006317 |
| DJ-30 | 08/28/2008 | 11715.18 | 09/04/2008 | 11188.23 | 0.9531099 |
| Bond | 09/04/2008 | 1 | 09/30/2008 | 1.002607 | 1.002607 |
| DJ-30 | 09/30/2008 | 10850.66 | 10/06/2008 | 9955.5 | 0.9156668 |
| Bond | 10/06/2008 | 1 | 12/31/2008 | 1.008624 | 1.008624 |

Acknowledgments. This collaboration started when Zhang and Zhu were visiting the Department of Mathematics and Risk Management Institute at NUS, respectively. They thank their hosts at NUS for creating an excellent atmosphere for academic discussions. The
authors thank two anonymous referees for their constructive comments, which have led to a much improved version of the paper.

## REFERENCES

[1] M. Beibel and H. R. Lerche, A new look at optimal stopping problems related to mathematical finance, Statist. Sinica, 7 (1997), pp. 93-108.
[2] A. Cadenillas and S. R. Pliska, Optimal trading of a security when there are taxes and transaction costs, Finance Stoch., 3 (1999), pp. 137-165.
[3] G. M. Constantinides, Capital market equilibrium with personal tax, Econometrica, 51 (1983), pp. 611-636.
[4] M. Dai, L. Jiang, P. Li, and F. Yi, Finite horizon optimal investment and consumption with transaction costs, SIAM J. Control Optim., 48 (2009), pp. 1134-1154.
[5] M. Dai, Y. K. Kwok, and H. You, Intensity-based framework and penalty formulation of optimal stopping problems, J. Econom. Dynam. Control, 31 (2007), pp. 3860-3880.
[6] M. Dai, Z. Q. Xu, and X. Y. Zhou, Continuous-time Markowitz's model with transaction costs, SIAM J. Financial Math., 1 (2010), pp. 96-125.
[7] M. Dai and F. Yi, Finite horizontal optimal investment with transaction costs: A parabolic double obstacle problem, J. Differential Equations, 246 (2009), pp. 1445-1469.
[8] R. M. Dammon and C. S. Spatt, The optimal trading and pricing of securities with asymmetric capital gains taxes and transaction costs, Rev. Financ. Stud., 9 (1996), pp. 921-952.
[9] A. Damodaran, Damodaran Online, http://pages.stern.nyu.edu/~adamodar/.
[10] Federal Reserve, Federal Reserve Statistic Release Website, http://www.federalreserve.gov/releases/ h15/data/Annual/H15_TCMNOM_Y10.txt.
[11] P. A. Forsyth and K. R. Vetzal, Quadratic convergence for valuing American options using a penalty method, SIAM J. Sci. Comput., 23 (2002), pp. 2095-2122.
[12] A. Friedman, Variational Principles and Free-Boundary Problems, Wiley, New York, 1982.
[13] X. Guo and Q. Zhang, Optimal selling rules in a regime switching model, IEEE Trans. Automat. Control, 50 (2005), pp. 1450-1455.
[14] K. Helmes, Computing optimal selling rules for stocks using linear programming, in Mathematics of Finance, Contemp. Math. 351, G. Yin and Q. Zhang, eds., American Mathematical Society, Providence, RI, 2004, pp. 187-198.
[15] IndexArb, http://www.indexarb.com/dividendYieldSortedn1.html.
[16] T. C. Johnson and M. Zervos, The Optimal Timing of Investment Decisions, manuscript.
[17] S. Karlin and H. M. Taylor, A Second Course in Stochastic Processes, Academic Press, New York, 1981.
[18] R. S. Liptser and A. N. Shiryayev, Statistics of Random Processes I. General Theory, Springer-Verlag, Berlin, New York, 1977.
[19] H. Liu and M. Loewenstein, Optimal portfolio selection with transaction costs and finite horizons, Rev. Financ. Stud., 15 (2002), pp. 805-835.
[20] R. Liu, G. Yin, and Q. Zhang, Option pricing in a regime switching model using the fast Fourier transform, J. Appl. Math. Stoch. Anal., 2006 (2006), article ID 18109.
[21] A. Løkka and M. Zervos, Long-Term Optimal Real Investment Strategies in the Presence of Adjustment Costs, manuscript.
[22] A. Merhi and M. Zervos, A model for reversible investment capacity expansion, SIAM J. Control Optim., 46 (2007), pp. 839-876.
[23] O. Nakoulima, Sur une notion de solution faible pour les inéquations variationnelles d'évolution à deux obstacles, C. R. Acad. Sci. Paris Sér. A-B, 284 (1977), pp. A1037-A1040.
[24] B. Øksendal, Stochastic Differential Equations, 6th ed., Springer-Verlag, Berlin, New York, 2003.
[25] R. Rishel and K. Helmes, A variational inequality sufficient condition for optimal stopping with application to an optimal stock selling problem, SIAM J. Control Optim., 45 (2006), pp. 580-598.
[26] J. D. Schwager, Market Wizards: Interviews with Top Traders, HarperBusiness, New York, 1990.
[27] H. M. Soner and S. E. Shreve, A free boundary problem related to singular stochastic control: The parabolic case, Comm. Partial Differential Equations, 16 (1991), pp. 373-424.
[28] W. M. Wonham, Some applications of stochastic differential equations to optimal nonlinear filtering, J. SIAM Control, 2 (1965), pp. 347-369.
[29] G. Yin, R. H. Liu, and Q. Zhang, Recursive algorithms for stock liquidation: A stochastic optimization approach, SIAM J. Optim., 13 (2002), pp. 240-263.
[30] H. Zhang and Q. Zhang, Trading a mean-reverting asset: Buy low and sell high, Automatica J. IFAC, 44 (2008), pp. 1511-1518.
[31] Q. Zhang, Stock trading: An optimal selling rule, SIAM J. Control Optim., 40 (2001), pp. 64-87.
[32] Q. Zhang, G. Yin, and J. B. Moore, Two-time scale approximation for Wonham filters, IEEE Trans. Inform. Theory, 53 (2007), pp. 1706-1715.

# Representations for Optimal Stopping under Dynamic Monetary Utility Functionals* 

Volker Krätschmer ${ }^{\dagger}$ and John Schoenmakers ${ }^{\ddagger}$


#### Abstract

In this paper we consider the optimal stopping problem for general dynamic monetary utility functionals. Sufficient conditions for the Bellman principle and the existence of optimal stopping times are provided. Particular attention is paid to representations which allow for a numerical treatment in real situations. To this aim, generalizations of standard evaluation methods like policy iteration and dual and consumption based approaches are developed in the context of general dynamic monetary utility functionals. As a result, it turns out that the possibility of a particular generalization depends on specific properties of the utility functional under consideration.


Key words. monetary utility functionals, optimal stopping, duality, policy iteration
AMS subject classifications. 49L20, 60G40, 91B16
DOI. 10.1137/090775841

1. Introduction. Dynamic monetary utility functionals, or DMU functionals for short, can be seen as generalizations of the ordinary conditional expectation, the usual functional which is to be maximized in standard stopping problems, which occur, for instance, in the theory of pricing of American (Bermudan) options in a complete market. It is well known that in an incomplete market the price of an American option is determined by the so-called upper and lower Snell envelopes which in turn are obtained via optimal stopping of the reward process w.r.t. two particular mutually conjugate DMU functionals (cf., e.g., [17]). From an economic point of view, DMU functionals may be seen as representations of dynamic preferences in terms of utilities of financial investors.

By changing sign, a DMU functional becomes a dynamic risk measure (e.g., in [25]) which represents preferences in terms of losses instead of utilities in fact. Therefore, technically, the study of DMU functionals is basically equivalent to the study of dynamic risk measures, which became a growing field of research in recent years. A realistic dynamic risk assessment of financial positions should allow for updating as time evolves, taking into account new information. The notion of dynamic risk measures has been established to provide a proper framework (cf., e.g., [3], [10], [12], [13], [16]). It is based on an axiomatic characterization extending the classical axioms for the concept of one-period risk measures in [2] to the dynamic multiperiod setting. From the very beginning one crucial issue was to find reasonable

[^84]conditions of mutual relationships between the risk functionals, so-called dynamic consistency, leading to different concepts (cf., e.g., [3], [10], [12], [13], [35], [34], [38], [39]). The most used one is often called strong time consistency, and it is linked with a technical condition for dynamic risk measures known as recursiveness. This condition will play an important technical role in our investigations.

Recently, DMU functionals (being dynamic risk measures with changed sign) have been incorporated into different topics, such as, for example, the dynamics of indifference prices (see [25], [11]) and the pricing of derivatives in incomplete financial markets (cf., e.g., [35], [17], [32]). In this respect we want to emphasize the contributions in [17], [32] as being the starting point of this paper. There the superhedging of American options is analyzed as solutions of optimal stopping problems in the context of coherent DMU functionals. We want to extend these considerations to more general monetary utility functionals. For instance, we will not necessarily assume translation invariance, which has recently been questioned as a suitable condition for risk assessment since it tacitly supposes certainty on discounting factors by the investors (cf. [15]).

Within a time discrete setting we shall look for a set of conditions for DMU functionals which is as minimal as possible in a sense, while keeping the presentation compact and such that solutions for the related optimal stopping problems may be guaranteed. For classical stopping problems w.r.t. ordinary conditional expectations the starting point for any solution representation is the Bellman principle. This suggests investigating when the Bellman principle holds for the general optimal stopping problems. The above-mentioned condition of recursiveness in connection with a specific regularity condition will turn out to be sufficient.

Beyond the considerations of the general optimal stopping, the main contribution of this paper is the development of iterative methods and other representations for solving them. Based on these methods we naturally construct simulation based solution algorithms which allow for solving such stopping problems in practice. In contrast, meanwhile, to industrial standard approaches for Bermudan options, and hence the ordinary stopping problem in discrete time (see, among others, [1], [8], [26], [28], [37]), we have not seen yet a comprehensive generic approach for treating generalized optimal stopping problems numerically. In this respect this paper intends to be a first step in this direction.

This paper is organized as follows. In section 2 the concept of DMU functionals is introduced. In section 3 we investigate the Bellman principle and the existence of optimal stopping strategies. In section 4 a generalization of the policy iteration method of [26] is presented. Section 5, section 6, and section 7 generalize, respectively, the additive dual method of [33], [20], the multiplicative dual of [22], and the consumption based approach in [4], [5]. In section 8 we shall provide a simulation setting to utilize the results of sections $4,5,6$, and 7 to construct approximations to the optimal values of the investigated stopping problems. More technical proofs are given in the appendix.
2. Dynamic monetary utility functionals. Let $\left(\Omega,\left(\mathcal{F}_{t}\right)_{t \in\{0, \ldots, T\}}, \mathcal{F}, \mathrm{P}\right)$ be a filtered probability space with $\{0,1\}$-valued $\mathrm{P} \mid \mathcal{F}_{0}$, and let $\mathfrak{X}$ be a real vector subspace of $L^{0}(\Omega, \mathcal{F}, \mathrm{P})$ containing the indicator mappings $1_{A}$ of subsets $A \in \mathcal{F}$. It is assumed that for any $X \in \mathfrak{X}$ and $A \in \mathcal{F}$ it holds that $1_{A} X \in \mathfrak{X}$. Moreover, $X \wedge Y \in \mathfrak{X}$, and $X \vee Y \in \mathfrak{X}$ is valid for any $X, Y \in \mathfrak{X}$. Hence, in particular, $\mathfrak{X}$ is a vector lattice.

A family of mappings $\Phi:=\left(\Phi_{t}\right)_{t \in\{0, \ldots, T\}}$ with $\Phi_{t}: \mathfrak{X} \rightarrow \mathfrak{X} \cap L^{0}\left(\Omega, \mathcal{F}_{t}, \mathrm{P}\right)$ being monotone,
i.e., $\Phi_{t}(X) \leq \Phi_{t}(Y)$ for $X, Y \in \mathfrak{X}$ with $X \leq Y$ P-a.s., is called a dynamic monetary utility functional or DMU functional for short.

We say that $\left(\Phi_{t}\right)_{t \in\{0, \ldots, T\}}$ is recursively generated if there is some family $\left(\Psi_{t}\right)_{t \in\{0, \ldots, T\}}$ of mappings $\Psi_{t}: \mathfrak{X} \cap L^{0}\left(\Omega, \mathcal{F}_{t+1}, \mathrm{P}\right) \rightarrow \mathfrak{X} \cap L^{0}\left(\Omega, \mathcal{F}_{t}, \mathrm{P}\right)$ with $\mathcal{F}_{T+1}:=\mathcal{F}$ such that

$$
\Psi_{T}=\Phi_{T}, \quad \text { and } \quad \Phi_{t}=\Psi_{t} \circ \Phi_{t+1} \quad \text { for } \quad t=0, \ldots, T-1 .
$$

In this case the mappings $\Psi_{t}$ will be given the name generators of $\left(\Phi_{t}\right)_{t \in\{0, \ldots, T\}}$.
Let us introduce some further notation. Henceforth $\mathcal{T}_{t}$ will stand for the set of finite stopping times $\tau$ with $\tau \geq t$ P-a.s., whereas $\mathcal{H}$ will denote the set of adapted processes $Z:=\left(Z_{t}\right)_{t \in\{0, \ldots, T\}}$ such that $Z_{t} \in \mathfrak{X} \cap L^{0}\left(\Omega, \mathcal{F}_{t}, \mathrm{P}\right)$ for $t \in\{0, \ldots, T\}$.

The following conditions on $\left(\Phi_{t}\right)_{t \in\{0, \ldots, T\}}$ will play an important role in the context of optimal stopping of DMU functionals studied later.
(C1) $\Phi_{t}(X) \leq \Phi_{t}(Y)$ P-a.s. for $t \in\{0, \ldots, T-1\}$, and $X, Y \in \mathfrak{X}$ with $\Phi_{t+1}(X) \leq \Phi_{t+1}(Y)$ P-a.s. (time consistency).
(C2) $\Phi_{t}\left(1_{A} X\right)=1_{A} \Phi_{t}(X)$ P-a.s. for $t \in\{0, \ldots, T\}, A \in \mathcal{F}_{t}$, and $X \in \mathfrak{X}$ (regularity).
(C3) $\Phi_{t}(X+Y)=\Phi_{t}(X)+Y$ P-a.s. for $t \in\{0, \ldots, T\}$, and $X, Y \in \mathfrak{X}$ with $Y$ being $\mathcal{F}_{t}$-measurable (conditional translation invariance).
(C4) $\Phi_{t}=\Phi_{t} \circ \Phi_{t+1}$ P-a.s. for $t \in\{0, \ldots, T-1\}$ (recursiveness).
(C5) $\Phi_{t}(0)=0$ P-a.s. for $t \in\{0, \ldots, T\}$ (normalization).
(C6) $\Phi_{t}(Y X)=Y \Phi_{t}(X)$ P-a.s. for $t \in\{0, \ldots, T\}, X \in \mathfrak{X}$, and $Y \in \mathfrak{X} \cap L^{0}\left(\Omega, \mathcal{F}_{t}, \mathrm{P}\right)$ with $Y \geq 0$ P-a.s. as well as $X Y \in \mathfrak{X}$ (conditional positive homogeneity).
(C7) For each $X \in \mathfrak{X}$ with $X \geq 0$ P-a.s. there exists a function $g:[0, \infty) \rightarrow \mathbb{R}_{+}$such that $\lim _{\varepsilon \downarrow 0} g(\varepsilon)=0$, and

$$
\begin{equation*}
\Phi_{t}(X \vee \varepsilon) \leq \Phi_{t}(X)+g(\varepsilon) \quad \text { for } \quad t \in\{0, \ldots, T\} . \tag{2.1}
\end{equation*}
$$

Remark 2.1. The construction of DMU functionals via generators opens up the possibility of obtaining functionals with desired properties by just imposing them on the generators.

In this paper we frequently use one of the following implications. Their proofs are simple and therefore omitted.

- Recursiveness implies that $\left(\Phi_{t}\right)_{t \in\{0, \ldots, T\}}$ is recursively generated, where the generators are the restrictions $\Phi_{t} \mid \mathfrak{X} \cap L^{0}\left(\Omega, \mathcal{F}_{t+1}, \mathrm{P}\right)$ for $t=0, \ldots, T$.
- Let $\left(\Phi_{t}\right)_{t \in\{0, \ldots, T\}}$ be recursively generated by $\left(\Psi_{t}\right)_{t \in\{0, \ldots, T\}}$. Then the following hold:
- If $\Phi_{t}(X)=X$ P-a.s. for $t \in\{0, \ldots, T\}$ and $X \in \mathfrak{X} \cap L^{0}\left(\Omega, \mathcal{F}_{t}, P\right)$, then $\left(\Phi_{t}\right)_{t \in\{0, \ldots, T\}}$ is recursive.
- If $\Psi_{t}(X)=X$ P-a.s. for $t \in\{0, \ldots, T\}$ and $X \in \mathfrak{X} \cap L^{0}\left(\Omega, \mathcal{F}_{t}, \mathrm{P}\right)$, then $\left(\Phi_{t}\right)_{t \in\{0, \ldots, T\}}$ is recursive.
- If for any $X \in \mathfrak{X}$ and $A \in \mathcal{F}_{t}$ it holds that $\Psi_{t}\left(1_{A} X\right)=1_{A} \Psi_{t}(X)$, then $\Phi$ is regular.

Example 2.2. The functional $\Phi$ given by the conditional expectations $\Phi_{t}:=E\left[\cdot \mid \mathcal{F}_{t}\right]$, defined on $\mathfrak{X}:=L^{1}(\Omega, \mathcal{F}, \mathrm{P})$, is a basic example for a DMU functional. It satisfies all of the conditions (C1)-(C7).

It is natural to generalize the usual martingale concept to the notion of " $\Phi$-martingale" for a given DMU functional $\Phi$ as defined below. The notion of $\Phi$-martingales will be used for different representations of optimal stopping problems in sections 5 and 6 .

Definition 2.3. $M:=\left(M_{t}\right)_{t \in\{0, \ldots, T\}} \in \mathcal{H}$ is said to be a $\Phi$-martingale if $\Phi_{t}\left(M_{t+1}\right)=M_{t}$ P -a.s. for every $t \in\{0, \ldots, T-1\}$. Note that for recursive $\Phi, M \in \mathcal{H}$ is a $\Phi$-martingale if and only if $\Phi_{t}\left(M_{s}\right)=M_{t} \mathrm{P}$-a.s. for every $s, t \in\{0, \ldots, T-1\}$ with $s>t$.

Let us discuss some further examples of DMU functionals. First of all we want to consider the relationship with the so-called dynamic risk measures.

Remark 2.4. DMU functionals may be viewed as generalizations of dynamic risk measures. Recall that a family $\left(\rho_{t}\right)_{t \in\{0, \ldots, T\}}$ is a dynamic risk measure if and only if $\left(-\rho_{t}\right)_{t \in\{0, \ldots, T\}}$ is a conditionally translation invariant monetary utility functional. The property of translation invariance suggests restricting considerations to normalized functionals because for a translation invariant $\Phi$ we have $\Phi_{t}\left(X-\Phi_{t}(X)\right)=0$. In the normalized case it then holds that $\Phi_{t}(Y)=Y$ P-a.s. for every $t \in\{0, \ldots, T\}$ and any $Y \in \mathfrak{X} \cap L^{0}\left(\Omega, \mathcal{F}_{t}, \mathrm{P}\right)$, and in view of Remark 2.1, $\Phi$ is recursively generated if and only if it is recursive.

We shall call a normalized conditional translation invariant $\Phi$ convex/concave if the mappings $\Phi_{t}(t \in\{0, \ldots, T\})$ are simultaneously convex/concave. If $\Phi$ is convex/concave, then

$$
\bar{\Phi}_{t}: \mathfrak{X} \rightarrow \mathfrak{X} \cap L^{0}\left(\Omega, \mathcal{F}_{t}, \mathrm{P}\right), \quad X \mapsto-\Phi_{t}(-X),
$$

defines a concave/convex normalized conditional translation invariant DMU functional called the conjugate of $\Phi$. The conditions of recursiveness and regularity are satisfied by $\Phi$ if and only if its conjugate $\bar{\Phi}$ fulfills them. Conditional translation invariance of convex/concave $\Phi$ implies the regularity condition for the restriction of $\Phi$ to $\mathfrak{X} \cap L^{\infty}(\Omega, \mathcal{F}, \mathrm{P})$ (cf. [25], where this restriction is essential for the proof). Moreover, regularity is even valid on the entire space $\mathfrak{X}$ if $\lim _{n \rightarrow \infty} \Phi_{t}\left((X-n)^{+}\right)=0$ P-a.s. for every $t \in\{0, \ldots, T\}$ and any nonnegative $X \in \mathfrak{X}$. Indeed, one may conclude from Lemma 6.5 along with Proposition 6.6 in [27] that

$$
\Phi_{t}(X)=\underset{m \in \mathbb{N}}{\operatorname{ess} \inf } \operatorname{ess} \sup \Phi_{n \in \mathbb{N}}\left(X^{+} \wedge n-X^{-} \wedge m\right)
$$

holds for $t \in\{0, \ldots, T\}$ and $X \in \mathfrak{X}$.
In the context of dynamic risk measures the property of recursiveness plays an important role. On the one hand, it is intimately linked with the property of time consistency, which has a specific meaning in expressing dynamic preferences of investors. For a thorough study the reader may consult, e.g., [16] or [3]. On the other hand, optimal stopping with dynamic risk measures may be related to specific financial applications.

In the following we shall give some examples of recursive regular DMU functionals. Let us start with DMU functionals related to time consistent coherent dynamic risk measures.

Example 2.5 (DMU functionals related to time consistent coherent dynamic risk measures). A large class of convex, regular, and recursive conditional translation invariant DMU functionals is provided by

$$
\Phi_{t}(X):=\underset{\mathbb{Q} \in \mathcal{Q}}{\operatorname{ess} \sup } E_{\mathbb{Q}}\left[X \mid \mathcal{F}_{t}\right],
$$

where $\mathcal{Q}$ denotes a set of probability measures which are equivalent with P , and $\mathcal{Q}$ should be stable in the sense of [17]. In order to keep the presentation compact we here assume for simplicity that such a DMU functional $\Phi$ is defined on $\mathfrak{X}:=L^{\infty}(\Omega, \mathcal{F}, \mathrm{P})$ (cf. Theorem 6.53 in [17]). Extensions to functionals defined on suitable sets of unbounded random variables may be found in [32] or [17], and a combination of both results might yield further refinements.

Notice that in addition $\Phi$ fulfills conditional positive homogeneity, which means that $\rho(X):=\Phi_{t}(-X)$ defines a so-called coherent dynamic risk measure $\left(\rho_{t}\right)_{t \in\{0, \ldots, T\}}$, which is time consistent.

We may extend the class of time consistent coherent dynamic risk measures to convex ones.

Example 2.6 (DMU functionals related to time consistent convex dynamic risk measures). Let $\mathcal{Q}$ be a set of probability measures which are equivalent with P , and let $\left(\alpha_{t}\right)_{t \in\{0, \ldots, T\}}$ be a family of mappings $\alpha_{t}$ from $\mathcal{Q}$ into the space of $\mathcal{F}_{t}$-measurable random variables with values in $\mathbb{R} \cup\{\infty\}$ fulfilling ess $\inf _{\mathrm{Q} \in \mathcal{Q}} \alpha_{t}(\mathrm{Q})=0$ P-a.s. Then we may define on $\mathfrak{X}:=L^{\infty}(\Omega, \mathcal{F}, \mathrm{P})$ a DMU functional $\Phi$ via the generators $\left(\Psi_{t}\right)_{t \in\{0, \ldots, T\}}$ of the following form:

$$
\begin{equation*}
\Psi_{t}(X):=\underset{\mathrm{Q} \in \mathcal{Q}}{\operatorname{ess} \sup }\left(E_{\mathrm{Q}}\left[X \mid \mathcal{F}_{t}\right]-\alpha_{t}(\mathrm{Q})\right) . \tag{2.2}
\end{equation*}
$$

The obtained DMU functional is normalized, convex, and conditionally translation invariant, and therefore regular (see Remark 2.4) as well as recursive (cf. Remark 2.1). Moreover, it defines a time consistent convex dynamic risk measure $\left(\rho_{t}\right)_{t \in\{0, \ldots, T\}}$ by $\rho_{t}(X):=\Phi_{t}(-X)$. Notice, furthermore, that by a standard argument (see the proof of Theorem 1 in [13]) every generator $\Psi_{t}$ of the form (2.2) satisfies $\Psi_{t}\left(X_{n}\right) \nearrow \Psi_{t}(X)$ whenever $X_{n} \nearrow X$ P-a.s. From this we may conclude by routine backward induction that every $\Phi_{t}$ fulfills $\Phi_{t}\left(X_{n}\right) \nearrow \Phi_{t}(X)$ for $X_{n} \nearrow X$ P-a.s., and then in view of Theorem 2.3 from [16] we have the following representation of $\Phi$ :

$$
\Phi_{t}(X)=\underset{\mathrm{Q} \in \mathcal{Q}_{t}}{\operatorname{ess} \sup }\left(E_{\mathrm{Q}}\left[X \mid \mathcal{F}_{t}\right]-\alpha_{t}^{\min }(\mathrm{Q})\right) .
$$

Here $\mathcal{Q}_{t}$ denotes the set of all probability measures on $\mathcal{F}$ which are absolutely continuous w.r.t. P and coincide with P on $\mathcal{F}_{t}$, and $\alpha_{t}^{\min }$ is a certain mapping from $\mathcal{Q}_{t}$ into the space of $\mathcal{F}_{t}$-measurable random variables with values in $\mathbb{R} \cup\{\infty\}$, which may generally differ from the initial $\alpha_{t}$ in the representation of the generator (2.2). In particular the DMU functional $\Phi$ does not in general have the same representation as its generators.

A prominent special case is obtained by using conditional relative entropies as a choice for the mappings $\alpha_{t}$; i.e., $\mathcal{Q}$ consists of all probability measures Q with some strictly positive P-Radon-Nikodým derivative $\frac{d \mathrm{Q}}{d \mathrm{P}}$ satisfying $E_{\mathrm{P}}\left[\frac{d \mathrm{Q}}{d \mathrm{P}} \ln \left(\frac{d \mathrm{Q}}{d \mathrm{P}}\right)\right]<\infty$, and

$$
\alpha_{t}(\mathrm{Q}):=\left\{\begin{array}{cl}
\frac{1}{\gamma} E_{\mathrm{P}}\left[\left.\frac{d \mathrm{Q}}{d \mathrm{P}} \ln \left(\frac{d \mathrm{Q}}{d \mathrm{P}}\right) \right\rvert\, \mathcal{F}_{t}\right] & : \mathrm{Q}\left|\mathcal{F}_{t}=\mathrm{P}\right| \mathcal{F}_{t}, \\
\infty & : \quad \text { otherwise }
\end{array} \quad \text { for some } \gamma>0 .\right.
$$

This leads to the so-called entropic DMU functional $\Phi$ with

$$
\Phi_{t}(X):=\frac{1}{\gamma} \ln \left(E\left[\exp (\gamma X) \mid \mathcal{F}_{t}\right]\right)
$$

(see [13] along with Theorem 4.5 in [16]). The entropic DMU functional is not conditionally positive homogeneous like the ones in Example 2.5.

The next large class of DMU functionals concerns the so-called $g$-expectations. They are prominent examples of nonlinear functionals satisfying martingale-type properties like recursiveness but need not be conditionally translation invariant as the examples before.

Example 2.7. Let $\left(\mathcal{G}_{s}\right)_{s \geq 0}$ be the augmented filtration on $\Omega$ associated with the filtration generated by a standard $d$-dimensional Brownian motion $\left(B_{s}\right)_{s \geq 0}$ with $B_{0}:=0$, and let for $S>0$ the function $g: \Omega \times[0, S] \times \mathbb{R} \times \mathbb{R}^{d} \rightarrow \mathbb{R}$ satisfy some certain technical conditions such that it can be used as the driver of a backward stochastic differential equation (BSDE):

$$
Y_{s}=X+\int_{s}^{S} g\left(\cdot, r, Y_{r}, Z_{r}\right) d r-\int_{s}^{S} Z_{r} d B_{r} \quad \text { for } \quad s \in[0, S],
$$

where $X \in L^{2}\left(\Omega, \mathcal{G}_{S}, \mathrm{P}\right)$ (cf. [29]). Moreover, there always exists a unique couple $\left(Y_{s}^{X}\right)_{s \in[0, S]}$ and $\left(Z_{s}^{X}\right)_{s \in[0, S]}$ of adapted 1- and $d$-dimensional processes, respectively, with square integrable paths solving the BSDE (cf. [29] again). Now it is natural to define the family $\left(\mathcal{E}_{g}\left[\cdot \mid \mathcal{G}_{s}\right]\right)_{s \in[0, S]}$ via

$$
\mathcal{E}_{g}\left[\cdot \mid \mathcal{G}_{s}\right]: L^{2}\left(\Omega, \mathcal{G}_{S}, \mathrm{P}\right) \rightarrow L^{2}\left(\Omega, \mathcal{G}_{s}, \mathrm{P}\right), \quad X \mapsto Y_{s}^{X}
$$

known as (a family of) conditional $g$-expectations, where $\mathcal{E}_{g}\left[\cdot \mid \mathcal{G}_{0}\right]$ is just called $g$-expectation. For $g \equiv 0$ we retrieve the usual (conditional) expectation of a square integrable random variable. For applications of conditional $g$-expectations in finance the reader is referred to [14], [30].

Let us now pick some observation times $0=$ : $s_{0}<s_{1}<\cdots<s_{T}:=S$, and define $\left(\Omega,\left(\mathcal{F}_{t}\right)_{t \in\{0, \ldots, T\}}, \mathcal{F}, \mathrm{P}\right)$ and $\Phi:=\left(\Phi_{t}\right)_{t \in\{0, \ldots, T\}}$ by $\mathcal{F}_{t}:=\mathcal{G}_{s_{t}}, \mathcal{F}:=\mathcal{F}_{T}$, and $\Phi_{t}:=\mathcal{E}_{g}\left[\cdot \mid \mathcal{G}_{s_{t}}\right]$. Drawing on basic properties of conditional $g$-expectation as derived by Peng in [29], $\Phi$ is always a regular recursive DMU functional fulfilling $\Phi_{t}(X)=X$ P-a.s. for $t \in\{0, \ldots, T\}$ and $\mathcal{F}_{t}$-measurable $X$.

Furthermore, $\Phi$ is conditional translation invariant if and only if $g(\omega, s, \cdot, z)$ is constant for every $\omega \in \Omega, s \in[0, S]$, and $z \in \mathbb{R}^{d}$ (for the if part see [29]; for the only if part cf. [23]). In this case $\Phi$ is even a convex normalized conditionally translation invariant DMU functional if and only if in addition

$$
g\left(\cdot, \cdot, \cdot, \lambda z_{1}+(1-\lambda) z_{2}\right) \leq \lambda g\left(\cdot, \cdot, \cdot, z_{1}\right)+(1-\lambda) g\left(\cdot, \cdot, \cdot, z_{2}\right) \quad(\mathrm{P} \otimes d t) \text {-a.s. }
$$

for $z_{1}, z_{2} \in \mathbb{R}^{d}$ and $\lambda \in[0,1]$ (cf. [23]). Note that $\rho_{t}(X):=\Phi_{t}(-X)$ defines a convex dynamic risk measure $\left(\rho_{t}\right)_{t \in\{0, \ldots, T\}}$. The relationship between $g$-expectations and convex dynamic risk measures has been observed in [36] too.

We shall finish the section with some simple nonstandard examples.
Examples 2.8. Let $\mathfrak{X}=L^{\infty}(\Omega, \mathcal{F}, \mathrm{P})$, and let $\mathrm{Q}_{1}, \ldots, \mathrm{Q}_{K}$ denote probability measures which are equivalent with P .

1. For strictly increasing $U_{1}, \ldots, U_{K}: \mathbb{R} \rightarrow \mathbb{R}$ with $U_{1}(0)=\cdots=U_{K}(0)=0$ and positive $\alpha_{1}, \ldots, \alpha_{K}$, let $\Phi$ be recursively generated with generators $\left(\Psi_{t}\right)_{t \in\{0, \ldots, T\}}$ defined by

$$
\Psi_{t}(X):=\sum_{k=1}^{K} \alpha_{k} U_{k}^{-1}\left(E_{\mathrm{Q}_{k}}\left[U_{k}(X) \mid \mathcal{F}_{t}\right]\right) \text { for } t \in\{0, \ldots, T\} \text { and } X \in L^{\infty}\left(\Omega, \mathcal{F}_{t+1}, \mathrm{P}\right)
$$

Obviously the functional $\Phi$ is regular. Moreover, if $\sum_{k=1}^{K} \alpha_{k}=1$, it satisfies $\Phi_{t}(X)=$ $X$ P-a.s. for $X \in L^{\infty}\left(\Omega, \mathcal{F}_{t}, \mathrm{P}\right), t \in\{0, \ldots, T\}$; hence $\Phi$ is recursive. In the case of $K=\alpha_{1}=1, \Phi_{t}$ is defined in literally the same way as its generator $\Psi_{t}$, leading to a
so-called conditional certainty equivalent (cf. [18]) which may be viewed as a dynamic version of a premium principle for insurance contracts known as mean value premium principle (cf., e.g., [24]). We note that $\Phi$ is not conditionally translation invariant in general.
2. For nondecreasing $U_{1}, \ldots, U_{K}: \mathbb{R} \rightarrow \mathbb{R}$ with $U_{1}(0)=\cdots=U_{K}(0)=0$ and positive $\alpha_{1}, \ldots, \alpha_{K}$, let $\Phi$ be recursively generated with generators $\left(\Psi_{t}\right)_{t \in\{0, \ldots, T\}}$ defined by

$$
\Psi_{t}(X):=\sum_{k=1}^{K} \alpha_{k} E_{\mathrm{Q}_{k}}\left[U_{k}(X) \mid \mathcal{F}_{t}\right] \text { for } t \in\{0, \ldots, T\} \text { and } X \in L^{\infty}\left(\Omega, \mathcal{F}_{t+1}, \mathrm{P}\right)
$$

In general, this $\Phi$ will be neither recursive nor conditionally translation invariant, but still regular.
3. The optimal stopping problem. We will study the following stopping problem:

$$
\begin{equation*}
Y_{t}^{*}:=\underset{\tau \in \mathcal{T}_{t}}{\operatorname{esssup}} \Phi_{t}\left(Z_{\tau}\right), \quad t \in\{0, \ldots, T\}, \tag{3.1}
\end{equation*}
$$

for $Z \in \mathcal{H}$. We refer to the process $Y^{*}$ as the ( $\Phi$-)Snell envelope of $Z$. Below we consider two important aspects. First, we investigate the existence of optimal stopping times and, second, we try to find Bellman principles. The crucial step for guaranteeing optimal stopping times is provided by the following lemma.

Lemma 3.1. Let $Z:=\left(Z_{t}\right)_{t \in\{0, \ldots, T\}} \in \mathcal{H}$, and let for some fixed $t \in\{0, \ldots, T-1\}$ some $\tau_{t+1}^{*} \in \mathcal{T}_{t+1}$ exist such that $\Phi_{t+1}\left(Z_{\tau_{t+1}^{*}}\right)=\operatorname{ess}_{\sup }^{\tau \in \mathcal{T}_{t+1}}{ }_{t+1}\left(Z_{\tau}\right)$. Defining the event

$$
B_{t}:=\left[\Phi_{t}\left(Z_{t}\right)-\Phi_{t}\left(Z_{\tau_{t+1}^{*}}\right) \geq 0\right]
$$

and $\tau_{t}^{*}:=t 1_{B_{t}}+\tau_{t+1}^{*} 1_{\Omega \backslash B_{t}}$, we obtain $B_{t} \in \mathcal{F}_{t}, \tau_{t}^{*} \in \mathcal{T}_{t}$, and under the conditions of time consistency and regularity

$$
\Phi_{t}\left(Z_{\tau_{t}^{*}}\right)=\underset{\tau \in \mathcal{T}_{t}}{\operatorname{ess} \sup } \Phi_{t}\left(Z_{\tau}\right)=\Phi_{t}\left(Z_{t}\right) \vee \Phi_{t}\left(Z_{\tau_{t+1}^{*}}\right) .
$$

Proof. $B_{t} \in \mathcal{F}_{t}, \tau_{t}^{*} \in \mathcal{T}_{t}$ follows from $\mathcal{F}_{t}$-measurability of the outcomes of $\Phi_{t}$. Furthermore, we may observe $Z_{\tau_{t}^{*}}=1_{B_{t}} Z_{t}+1_{\Omega \backslash B_{t}} Z_{\tau_{i+1}^{*}}$. Then the application of (C2) yields

$$
\begin{aligned}
\Phi_{t}\left(Z_{\tau_{t}^{*}}\right)=1_{B_{t}} \Phi_{t}\left(Z_{\tau_{t}^{*}}\right)+1_{\Omega \backslash B_{t}} \Phi_{t}\left(Z_{\tau_{t}^{*}}\right) & \stackrel{(\mathrm{C} 2)}{=} \Phi\left(1_{B_{t}} Z_{t}\right)+\Phi\left(1_{\Omega \backslash B_{t}} Z_{\tau_{t+1}^{* *}}\right) \\
& \stackrel{(\mathrm{C} 2)}{=} 1_{B_{t}} \Phi_{t}\left(Z_{t}\right)+1_{\Omega \backslash B_{t}} \Phi_{t}\left(Z_{\tau_{t+1}^{*}}\right) \\
& =\Phi_{t}\left(Z_{t}\right) \vee \Phi_{t}\left(Z_{\tau_{t+1}^{*}}\right) .
\end{aligned}
$$

Next let us define the mapping $\sigma: \mathcal{T}_{t} \rightarrow \mathcal{T}_{t+1}$ by $\sigma(\tau):=(t+1) 1_{[\tau=t]}+\tau 1_{[\tau>t]}$. Then we obtain for $\tau \in \mathcal{T}_{t}$

$$
\Phi_{t}\left(Z_{\tau}\right)=\Phi_{t}\left(1_{[\tau=t]} Z_{t}+1_{[\tau>t]} Z_{\sigma(\tau)}\right) \stackrel{(\mathrm{C} 2)}{=} 1_{[\tau=t]} \Phi_{t}\left(Z_{t}\right)+1_{[\tau>t]} \Phi_{t}\left(Z_{\sigma(\tau)}\right) \leq \Phi_{t}\left(Z_{t}\right) \vee \Phi_{t}\left(Z_{\sigma(\tau)}\right)
$$

By assumption $\Phi_{t+1}\left(Z_{\sigma(\tau)}\right) \leq \Phi_{t+1}\left(Z_{\tau_{t+1}^{*}}\right)$ P-a.s. so that condition (C1) implies

$$
\Phi_{t}\left(Z_{\tau}\right) \leq \Phi_{t}\left(Z_{t}\right) \vee \Phi_{t}\left(Z_{\sigma(\tau)}\right) \stackrel{\mathrm{P}}{\leq} \Phi_{t}\left(Z_{t}\right) \vee \Phi_{t}\left(Z_{\tau_{t+1}^{*}}\right)=\Phi_{t}\left(Z_{\tau_{t}^{*}}\right),
$$

which completes the proof.

Since $\tau: \equiv T$ is always the optimal stopping time in $\mathcal{F}_{T}$, we may apply sequentially Lemma 3.1 to obtain the following result concerning the existence of optimal stopping times.

Theorem 3.2. Let $Z:=\left(Z_{t}\right)_{t \in\{0, \ldots, T\}} \in \mathcal{H}$. Then under conditions of time consistency and regularity there exists for any $t \in\{0, \ldots, T\}$ some $\tau_{t}^{*} \in \mathcal{T}_{t}$ such that

$$
\Phi_{t}\left(Z_{\tau_{t}^{*}}\right)=\underset{\tau \in \mathcal{T}_{t}}{\operatorname{ess} \sup } \Phi_{t}\left(Z_{\tau}\right) .
$$

The sequence $\left(\tau_{t}^{*}\right)_{t \in\{0, \ldots, T\}}$ of optimal stopping times may be chosen such that $\tau_{T}^{*}=T$, and

$$
1_{\left[\tau_{t}^{*}>t\right]} \tau_{t}^{*}=1_{\left[\tau_{t}^{*}>t\right]} \tau_{t+1}^{*} \quad \text { for any } \quad t \in\{0, \ldots, T-1\} .
$$

From now on we consider exclusively recursively generated DMU functionals. The following theorem gathers the main results concerning optimal stopping w.r.t. such functionals for our purposes later.

Theorem 3.3. Let $\left(\Phi_{t}\right)_{t \in\{0, \ldots, T\}}$ be regular and recursively generated with generators satisfying the property $\Psi_{t}(X) \leq \Psi_{t}(Y) \mathrm{P}$-a.s. for $t \in\{0, \ldots, T-1\}$ and $X, Y \in \mathfrak{X} \cap L^{0}\left(\Omega, \mathcal{F}_{t+1}, \mathrm{P}\right)$ with $X \leq Y$ P-a.s. Then Theorem 3.2 may be restated, and we have the Bellman principle, which means that

$$
\underset{\tau \in \mathcal{T}_{t}}{\operatorname{esssup}} \Phi_{t}\left(Z_{\tau}\right)=\Phi_{t}\left(Z_{t}\right) \vee \Psi_{t}\left(\underset{\sigma \in \mathcal{T}_{t+1}}{\operatorname{ess} \sup } \Phi_{t+1}\left(Z_{\sigma}\right)\right)
$$

is valid for any $Z \in \mathcal{H}$ and every $t \in\{0, \ldots, T-1\}$. If, moreover, $\Phi$ is recursive, then it holds that

$$
\underset{\tau \in \mathcal{T}_{t}}{\operatorname{esssup}} \Phi_{t}\left(Z_{\tau}\right)=\Phi_{t}\left(Z_{t}\right) \vee \Phi_{t}\left(\underset{\sigma \in \mathcal{T}_{t+1}}{\operatorname{ess} \sup } \Phi_{t+1}\left(Z_{\sigma}\right)\right)
$$

for arbitrary $Z \in \mathcal{H}$ and $t \in\{0, \ldots, T-1\}$.
Proof. The assumptions on the generators $\left(\Psi_{t}\right)_{t \in\{0, \ldots, T\}}$ imply condition (C1) so that we may apply Theorem 3.2 immediately, whereas we may conclude directly from Lemma 3.1 that the Bellman principle holds. The last statement of Theorem 3.3 is an easy consequence of the observation that in case of recursive $\Phi$ the generators are just the restrictions $\Phi_{t} \mid \mathfrak{X} \cap L^{0}\left(\Omega, \mathcal{F}_{t+1}, \mathrm{P}\right)$ for $t \in\{0, \ldots, T-1\}$.

Example 3.4. Let us consider the optimal stopping problem for the regular and recursively generated DMU functionals from Remark 2.4. There thus exists a family $\left(\tau_{t}^{*}\right)_{t \in\{0, \ldots, T\}}$ of optimal stopping times as in Theorem 3.2, and the Bellman principle is fulfilled due to Theorem 3.3. In particular this applies to functionals from Example 2.6, as well as to the functional

$$
\Phi_{t}: \mathfrak{X} \rightarrow \mathfrak{X} \cap L^{0}\left(\Omega, \mathcal{F}_{t}, \mathrm{P}\right), \quad X \mapsto \underset{\mathrm{Q} \in \mathcal{Q}}{\operatorname{ess} \sup } E_{\mathrm{Q}}\left[X \mid \mathcal{F}_{t}\right],
$$

and its conjugate $\bar{\Phi}$,

$$
\bar{\Phi}_{t}(X)=\underset{Q \in \mathcal{Q}}{\operatorname{essinf}} E_{\mathbb{Q}}\left[X \mid \mathcal{F}_{t}\right],
$$

where $\mathcal{Q}$ is a stable set of probability measures which are equivalent w.r.t. P , and $\mathfrak{X}:=$ $L^{\infty}(\Omega, \mathcal{F}, \mathrm{P})$ (cf. Example 2.5).

As a special case let $\mathcal{Q}$ denote the set of equivalent martingale measures of an arbitragefree incomplete financial market with reference measure P . It is known that $\mathcal{Q}$ is stable (cf.
[17, Theorem 6.45]) so that $\Phi$ and $\bar{\Phi}$ are recursive. Both DMU functionals play a key role for the issue of pricing and hedging American contingent claims in the following sense: For any nonnegative $Z \in \mathcal{H}$ the stopping problems (3.1) according to $\Phi$ and $\bar{\Phi}$ correspond to the upper and lower Snell envelopes of $Z$ w.r.t. $\mathcal{Q}$, respectively. Moreover, the initial values of the lower and upper Snell envelopes are just the lower and upper hedging prices, respectively. Further, the optimal stopping time according to the lower hedging prices corresponds to an optimal exercise strategy for the buyer of the option. For details see, for example, [17, Theorems 7.13, 7.14].

Example 3.5. Let $\Phi$ be a finite subfamily of conditional $g$-expectations. Then in view of Example 2.7 combined with Theorem 3.3 we may find for any $Z \in \mathcal{H}$ some family $\left(\tau_{t}^{*}\right)_{t \in\{0, \ldots, T\}}$ of stopping times $\tau_{t}^{*} \in \mathcal{T}_{t}$ satisfying $\tau_{T}^{*}=T$ as well as $1_{\left[\tau_{t}^{*}>t\right]} \tau_{t}^{*}=1_{\left[\tau_{t}^{*}>t\right]} \tau_{t+1}^{*}$, and

$$
\Phi_{t}\left(Z_{\tau_{t}^{*}}\right)=\underset{\tau \in \mathcal{T}_{t}}{\operatorname{esssup}} \Phi_{t}\left(Z_{\tau}\right)=\Phi_{t}\left(Z_{t}\right) \vee \Phi_{t}\left(\underset{\tau \in \mathcal{T}_{t+1}}{\operatorname{ess} \sup } \Phi_{t+1}\left(Z_{\tau}\right)\right)=\Phi_{t}\left(Z_{t}\right) \vee \Phi_{t}\left(Z_{\tau_{t+1}^{*}}\right)
$$

for $t \in\{0, \ldots, T-1\}$.
Example 3.6. The DMU functionals introduced in Examples 2.8 admit families of optimal stopping times as in Theorem 3.2 and satisfy the Bellman principle due to Theorem 3.3.

Remark 3.7. Stopping problems for certain specific DMU functionals have already been studied in the literature. Let us give a short review and comment.

Existence of optimal stopping families and the verification of the Bellman principle have been obtained for recursive DMU functionals $\Phi$ of the form $\Phi_{t}(X):=\operatorname{esss}_{\sup }^{\mathcal{Q} \in \mathcal{Q}} E_{\mathrm{Q}}\left[X \mid \mathcal{F}_{t}\right]$ or $\Phi_{t}(X):=\operatorname{ess}_{\inf _{\mathcal{Q} \in \mathcal{Q}}} E_{\mathcal{Q}}\left[X \mid \mathcal{F}_{t}\right]$, where $\mathcal{Q}$ is some stable set of probability measures which are equivalent with P (see, e.g., [17], [32]). In view of Example 3.4 these results may be extended to a considerably larger class of conditionally translation invariant DMU functionals.

Recently, Bayraktar and Yao [7] solved the optimal stopping problem for recursive conditionally translation invariant DMU functionals within a time continuous setting w.r.t. a Brownian filtration. In [6] the studies have been detailed for DMU functionals which are additionally concave.

Theorem 3.3 applies to DMU functionals which are not necessarily conditionally translation invariant or even not recursive; see Examples 3.5 and 3.6. We may thus avoid conditions like conditional translation invariance and recursiveness, which are questioned in the literature anyway (see, e.g., [15] and, e.g., [16], [34], [38], [39], respectively).
4. Iterative solution of optimal stopping problems. In this section we develop an iterative procedure for solving the optimal stopping problem. In fact we shall generalize the policy iteration method in [26] for classical optimal stopping with conditional expectations to optimal stopping of regular recursively generated DMU functionals. As such we obtain a procedure for approximating the optimal stopping time and hence the optimal value from below.

Throughout we fix a recursively generated regular DMU functional $\left(\Phi_{t}\right)_{t \in\{0, \ldots, T\}}$ with generators $\left(\Psi_{t}\right)_{t \in\{0, \ldots, T\}}$ satisfying the conditions in Theorem 3.3. So, for any $Z \in \mathcal{H}$ there exists a family $\left(\tau_{t}^{*}\right)_{t \in\{0, \ldots, T\}}$ of optimal stopping times $\tau_{t}^{*} \in \mathcal{T}_{t}$ satisfying

$$
\begin{equation*}
\tau_{T}^{*}=T, \quad \text { and } \quad 1_{\left[\tau_{t}^{*}>t\right]} \tau_{t}^{*}=1_{\left[\tau_{t}^{*}>t\right]} \tau_{t+1}^{*} \quad \text { for any } \quad t \in\{0, \ldots, T-1\} \tag{4.1}
\end{equation*}
$$

and

$$
\begin{equation*}
Y_{t}^{*}=\underset{\tau \in \mathcal{T}_{t}}{\operatorname{ess} \sup } \Phi_{t}\left(Z_{\tau}\right)=\Phi_{t}\left(Z_{\tau_{t}^{*}}\right) \quad \text { for every } \quad t \in\{0, \ldots, T\} \tag{4.2}
\end{equation*}
$$

Let us define $\left(\tau_{t}\right)_{t \in\{0, \ldots, T\}}$ to be a time consistent stopping family if

$$
\tau_{t} \in \mathcal{T}_{t}, \quad \tau_{T}=T, \quad \text { and } \quad 1_{\left[\tau_{t}>t\right]} \tau_{t}=1_{\left[\tau_{t}>t\right]} \tau_{t+1} \quad \text { for } \quad t \in\{0, \ldots, T-1\}
$$

The policy iteration step starts with any time consistent stopping family $\left(\tau_{t}\right)_{t \in\{0, \ldots, T\}}$ and corresponding process $\left(Y_{t}\right)_{t \in\{0, \ldots, T\}}$ with $Y_{t}:=\Phi_{t}\left(Z_{\tau_{t}}\right)$ being an approximation of $\left(Y_{t}^{*}\right)_{t \in\{0, \ldots, T\}}$. In order to improve this approximation we consider the process $\left(\tilde{Y}_{t}\right)_{t \in\{0, \ldots, T\}}$ defined by $\tilde{Y}_{t}:=\max _{t \leq s \leq T} \Phi_{t}\left(Z_{\tau_{s}}\right)$ and the new stopping family

$$
\begin{equation*}
\widehat{\tau}_{T}:=T, \quad \widehat{\tau}_{t}:=\inf \left\{s \in\{t, \ldots, T\} \mid \Phi_{s}\left(Z_{s}\right) \geq \max _{s+1 \leq u \leq T} \Phi_{s}\left(Z_{\tau_{u}}\right)\right\}, \quad 0 \leq t \leq T-1 \tag{4.3}
\end{equation*}
$$

Obviously, the stopping family $\left(\hat{\tau}_{t}\right)_{t \in\{0, \ldots, T\}}$ is also time consistent. By the next theorem, a generalization of Theorem 3.1 in [26] in fact, the process $\left(\widehat{Y}_{t}\right)_{t \in\{0, \ldots, T\}}$, defined by $\widehat{Y}_{t}:=\Phi_{t}\left(Z_{\hat{\tau}_{t}}\right)$, improves the initial approximation $\left(Y_{t}\right)_{t \in\{0, \ldots, T\}}$ of (4.2).

Theorem 4.1. We have the inequalities

$$
Y_{t} \leq \tilde{Y}_{t} \leq \widehat{Y}_{t} \leq Y_{t}^{*}, \quad t \in\{0, \ldots, T\}
$$

The proof of Theorem 4.1 is similar to the proof of Theorem 3.1 in [26]. However, it has to be emphasized that it is sufficient that the DMU functional under consideration is regular and recursively generated. For the convenience of the reader the proof is therefore provided in the appendix (while also comprising the structure of argumentation in [26] slightly).

In view of Theorem 4.1 the idea is to construct recursively a sequence of pairs

$$
\left(\left(\tau_{t}^{(m)}\right)_{t \in\{0, \ldots, T\}},\left(Y_{t}^{(m)}\right)_{t \in\{0, \ldots, T\}}\right)_{m \in \mathbb{N}_{0}}
$$

where $\left(\tau_{t}^{(m)}\right)_{t \in\{0, \ldots, T\}}$ is a time consistent stopping family for any $m \in \mathbb{N}_{0}$ such that $Y_{t}^{(m)}=$ $\Phi_{t}\left(Z_{\tau_{t}^{(m)}}\right)$, and $\tau_{t}^{(m+1)}=\inf \left\{s \in\{t, \ldots, T\} \mid \Phi_{s}\left(Z_{s}\right) \geq \max _{s+1 \leq u \leq T} \Phi_{s}\left(Z_{\tau_{u}^{(m)}}\right)\right\}$ for $t \in$ $\{0, \ldots, T-1\}$.

Next we start with some time consistent stopping family $\left(\tau_{t}^{(0)}\right)_{t \in\{0, \ldots, T\}}$; for example, a canonical choice is $\tau_{t}^{(0)}:=t$. Then, due to Theorem 4.1, we have

$$
\begin{equation*}
Y_{t}^{(0)} \leq Y_{t}^{(m)} \leq \tilde{Y}_{t}^{(m+1)} \leq Y_{t}^{(m+1)} \leq Y_{t}^{*} \quad \text { for } \quad m \in \mathbb{N}_{0}, \quad t \in\{0, \ldots, T\} \tag{4.4}
\end{equation*}
$$

where $\widetilde{Y}_{t}^{(m+1)}:=\max _{t \leq s \leq T} \Phi_{t}\left(Z_{\tau_{s}^{(m)}}\right)$.
The iteration procedure may be stopped after at most $T$ iterations, yielding an optimal stopping family.

Proposition 4.2. For $t \in\{0, \ldots, T\}$ we have

$$
Y_{t}^{(m)}=Y_{t}^{*} \quad \text { if } \quad m \geq T-t
$$

Hence $\tau_{t}^{(m)}$ is an optimal stopping time for the corresponding stopping problem at time $t$, if $m \geq T-t$, and in particular $\left(\tau_{t}^{(m)}\right)_{t \in\{0, \ldots, T\}}$ is an optimal stopping family for $m \geq T$.

Proof. The proof may be done by adapting the proof of Proposition 4.4 in [26] in a way similar to proving Theorem 4.1 and is therefore omitted. Indeed, a closer inspection of the proof of Proposition 4.4 in [26] shows that only regularity, the fact that the DMU functional is recursively generated by a monotonic system $\left(\Psi_{t}\right)$, and the Bellman principle (see Theorem 3.3) are essential.

Examples 4.3.

1. Referring to Example 3.4, Proposition 4.2 guarantees that the proposed iteration method provides a scheme to calculate superhedging prices and optimal exercises of discounted American options.
2. In view of Example 2.7 and Examples 2.8 the associated stopping problems may be solved iteratively by the introduced method. In particular we have a numerical scheme for optimal stopping with $g$-expectations.
3. Additive dual upper bounds. In this section a method for approximating the optimal value of the stopping problem from above will be developed for DMU functionals $\Phi$ which are regular, conditional translation invariant, and recursive. For such a $\Phi$ we propose an additive dual representation for the stopping problem (3.1), in terms of $\Phi$-martingales introduced in Definition 2.3. As such this generalization may be seen as a generalization of the representation of [33], [20] for the standard stopping problem. We first extend the classical additive Doob decomposition theorem.

Lemma 5.1. Let $\Phi$ be a conditional translation invariant DMU functional. Then for any stochastic process $Z:=\left(Z_{t}\right)_{t \in\{0, \ldots, T\}} \in \mathcal{H}$ there exists a unique pair $(M, A) \in \mathcal{H} \times \mathcal{H}$ of a $\Phi$-martingale $M$ and a predictable process $A$, such that $M_{0}=A_{0}=0$, and

$$
\begin{equation*}
Z_{t}=Z_{0}+M_{t}+A_{t} \text { for } t \in\{0, \ldots, T\} \quad \mathrm{P}-a . s \tag{5.1}
\end{equation*}
$$

Proof. Define $A$ recursively by $A_{0}:=0$, and $A_{t+1}:=A_{t}+\Phi_{t}\left(Z_{t+1}\right)-Z_{t}$ for $t \in\{0, \ldots$, $T-1\}$. Then of course $A \in \mathcal{H}$, and $A$ is predictable. Next define $M \in \mathcal{H}$ via $M_{t}:=$ $Z_{t}-Z_{0}-A_{t}$ for $t \in\{0, \ldots, T\}$. Obviously $M_{0}=0$, and by conditional translation invariance (property (C3)),
$\Phi_{t}\left(M_{t+1}\right) \stackrel{(\mathrm{C} 3)}{=} \Phi_{t}\left(Z_{t+1}\right)-Z_{0}-A_{t+1}=\Phi_{t}\left(Z_{t+1}\right)-Z_{0}-\left(A_{t}+\Phi_{t}\left(Z_{t+1}\right)-Z_{t}\right)=Z_{t}-Z_{0}-A_{t}=M_{t}$.
So $M$ is a $\Phi$-martingale and (5.1) holds. Now let $\left(M^{\prime}, A^{\prime}\right) \in \mathcal{H} \times \mathcal{H}$ be another pair as stated. Then for $t \in\{0, \ldots, T-1\}$ we may conclude by conditional translation invariance that

$$
0=\Phi_{t}\left(M_{t+1}^{\prime}-M_{t}^{\prime}\right)=\Phi_{t}\left(Z_{t+1}\right)-Z_{t}+A_{t}^{\prime}-A_{t+1}^{\prime}
$$

in particular, $A_{t+1}^{\prime}=A_{t}^{\prime}+\Phi_{t}\left(Z_{t+1}\right)-Z_{t}$. Hence by induction $A^{\prime}=A$, and so $M^{\prime}=M$.
Remark 5.2. The additive Doob decomposition has already been shown for the functional $\bar{\Phi}$ from Example 3.4 (cf. [32]).

The next lemma may be regarded as a generalization of Doob's optional sampling theorem. It is proved in the appendix.

Lemma 5.3. Let $\Phi$ be a regular, conditional translation invariant, and recursive DMU functional, and let $M$ be any $\Phi$-martingale. Then for every $Z:=\left(Z_{t}\right)_{t \in\{0, \ldots, T\}} \in \mathcal{H}$, each $t \in\{0, \ldots, T\}$, and each stopping time $\tau \in \mathcal{T}_{t}$, we have

$$
\Phi_{t}\left(Z_{\tau}\right)=\Phi_{t}\left(Z_{\tau}+M_{T}-M_{\tau}\right) .
$$

Remark 5.4. Under the assumptions of Lemma 5.3 the statement

$$
\Phi_{t}\left(Z_{\tau}\right)=\Phi_{t}\left(Z_{\tau}+M_{\tau}\right)-M_{t}
$$

which one might expect at a first glance, does not hold.
The Doob-type lemmas (Lemmas 5.1 and 5.3), and the Bellman principle theorem (Theorem 3.3), provide the ingredients to establish the following additive dual representation. In the case of $K=\alpha_{1}=1, \Phi_{t}$ is defined in literally the same way as its generator $\Psi_{t}$ leading to the following result.

Theorem 5.5. Let $\Phi$ be a regular, conditional translation invariant, and recursive DMU functional, and let $\mathcal{M}_{0}^{\Phi}$ be the set of all $\Phi$-martingales $M$ with $M_{0}=0$. For $Z:=\left(Z_{t}\right)_{t \in\{0, \ldots, T\}}$ $\in \mathcal{H}$ let $M^{*} \in \mathcal{M}_{0}^{\Phi}$ be the $\Phi$-martingale of the decomposition of $Y^{*}$ in (3.1) according to Lemma 5.1. Then

$$
\begin{aligned}
Y_{t}^{*} & =\underset{\tau \in \mathcal{T}_{t}}{\operatorname{ess} \sup } \Phi_{t}\left(Z_{\tau}\right)=\underset{M \in \mathcal{M}_{0}^{\Phi}}{\operatorname{ess} \inf } \Phi_{t}\left(\max _{t \leq j \leq T}\left(Z_{j}-M_{j}+M_{T}\right)\right) \\
& =\Phi_{t}\left(\max _{t \leq j \leq T}\left(Z_{j}-M_{j}^{*}+M_{T}^{*}\right)\right) \text { for } \quad t \in\{0, \ldots, T\} .
\end{aligned}
$$

Proof. Let $A^{*}:=\left(A_{t}^{*}\right)_{t \in\{0, \ldots, T\}}$ denote the predictable part of the decomposition of $Y^{*}$ according to Lemma 5.1. Since $M^{*}$ is a $\Phi$-martingale, we have for $t \in\{0, \ldots, T\}$

$$
0=\Phi_{t}\left(M_{t+1}^{*}\right)-M_{t}^{*} \stackrel{(\mathrm{C} 3)}{=} \Phi_{t}\left(M_{t+1}^{*}-M_{t}^{*}\right) \stackrel{(\mathrm{C} 3)}{=} \Phi_{t}\left(Y_{t+1}^{*}\right)-Y_{t}^{*}-\left(A_{t+1}^{*}-A_{t}^{*}\right) .
$$

This implies $A_{t+1}^{*}-A_{t}^{*}=\Phi_{t}\left(Y_{t+1}^{*}\right)-Y_{t}^{*} \leq 0$ due to the Bellman principle. Hence $A^{*}$ has nonincreasing paths. Furthermore, by the Bellman principle, $\Phi_{t}\left(Z_{t}\right)=Z_{t} \leq Y_{t}^{*}$ holds for every $t \in\{0, \ldots, T\}$. We thus have

$$
Z_{t}-M_{t}^{*}+M_{T}^{*}=Z_{t}+Y_{T}^{*}-Y_{t}^{*}+A_{t}^{*}-A_{T}^{*} \leq Y_{T}^{*}+A_{t}^{*}-A_{T}^{*} \quad \text { for } \quad t \in\{0, \ldots, T\} .
$$

Since $A^{*}$ is nonincreasing, $\Phi$ is conditional translation invariant and recursive, and $M^{*}$ is a $\Phi$-martingale, it follows that
$\Phi_{t}\left(\max _{t \leq j \leq T}\left(Z_{j}-M_{j}^{*}+M_{T}^{*}\right) \stackrel{(\mathrm{C} 3)}{\leq} \Phi_{t}\left(Y_{T}^{*}-A_{T}^{*}\right)+A_{t}^{*} \stackrel{(\mathrm{C} 3)}{=} Y_{0}^{*}+\Phi_{t}\left(M_{T}^{*}\right)+A_{t}^{*}=Y_{0}^{*}+M_{t}^{*}+A_{t}^{*}=Y_{t}^{*}\right.$
for $t \in\{0, \ldots, T\}$. Finally, using Lemma 5.3 and (5.2), we have for any $t \in\{0, \ldots, T\}$ and $M \in \mathcal{M}_{0}^{\Phi}$

$$
\begin{aligned}
Y_{t}^{*}=\underset{\tau \in \mathcal{T}_{t}}{\operatorname{ess} \sup } \Phi_{t}\left(Z_{\tau}+M_{T}-M_{\tau}\right) & \leq \underset{M \in \mathcal{M}_{0}^{\Phi}}{\operatorname{ess}} \Phi_{t}\left(\max _{t \leq j \leq T}\left(Z_{j}-M_{j}+M_{T}\right)\right) \\
& \leq \Phi_{t}\left(\max _{t \leq j \leq T}\left(Z_{j}-M_{j}^{*}+M_{T}^{*}\right)\right) \leq Y_{t}^{*}
\end{aligned}
$$

Remark 5.6. Let us emphasize that the structure of the generalized dual representation in Theorem 5.5 cannot be obtained by simply replacing conditional expectations with DMU functionals in the standard dual representation of [33], [20]. This remark applies to the generalized sampling lemma (Lemma 5.1) as well.

Example 5.7. Let $\mathcal{Q}$ denote the set of equivalent martingale measures w.r.t. some arbitragefree financial market, and let $\mathfrak{X}:=L^{\infty}(\Omega, \mathcal{F}, \mathrm{P})$. Fix any nonnegative $Z:=\left(Z_{t}\right)_{t \in\{0, \ldots, T\}} \in \mathcal{H}$. It may be viewed as a discounted American option. Then the DMU functionals $\Phi$ and $\bar{\Phi}$ from Example 3.4 meet the requirements of Theorem 5.5, and thus the superhedging price and the lowest arbitrage-free price of $Z$ may be represented by

$$
\inf _{X \in \mathcal{X}_{0}} \Phi_{0}\left[\max _{t \in\{0, \ldots, T\}}\left(Z_{t}-\Phi_{t}(X)+X\right)\right] \quad \text { and } \quad \inf _{X \in \mathcal{X}_{0}} \bar{\Phi}_{0}\left[\max _{t \in\{0, \ldots, T\}}\left(Z_{t}-\bar{\Phi}_{t}(X)+X\right)\right]
$$

respectively. Here $\mathcal{X}_{0}:=\left\{X \in \mathfrak{X} \mid \sup _{\mathrm{Q} \in \mathcal{Q}} E_{\mathrm{Q}}[X]=0\right\}$.
Examples 5.8. Theorem 5.5 may be applied immediately to the following regular, conditional translation invariant, and recursive functionals (see also Remark 2.1):

1. Any family of $g$-expectations as in Example 2.7 with driver $g: \Omega \times[0, S] \times \mathbb{R} \times \mathbb{R}^{d} \rightarrow \mathbb{R}$ such that $g(\omega, s, \cdot, z)$ is constant for $(\omega, s, z) \in \Omega \times[0, S] \times \mathbb{R}^{d}$.
2. The DMU functional $\Phi$ recursively defined as in Example 2.6.
3. Multiplicative dual upper bounds. The additive dual representation for constructing upper bounds for the standard stopping problem has a multiplicative version which is due to [22]. We will develop in this section a multiplicative dual representation for the stopping problem (3.1) when the DMU functional $\Phi$ is recursive and positively homogeneous. Note that from any positively homogeneous recursively generated DMU functional we may obtain a recursive one, by multiplication with a constant.

To our aim we need an extension of the multiplicative Doob decomposition theorem. As we do not want to burden the presentation with too many technicalities, we restrict ourselves in this section to the case where $\mathfrak{X}=L^{\infty}(\Omega, \mathcal{F}, \mathrm{P})$.

Lemma 6.1. Let $\Phi:=\left(\Phi_{t}\right)_{t \in\{0, \ldots, T\}}$ be a positively homogeneous recursive DMU functional. Let $\delta>0$, and let $Z:=\left(Z_{t}\right)_{t \in\{0, \ldots, T\}} \in \mathcal{H}$ with $Z_{t} \geq \delta \mathrm{P}$-a.s. for any $t \in\{0, \ldots, T\}$. Then there exists a unique pair $(N, U) \in \mathcal{H} \times \mathcal{H}$ of some $\Phi$-martingale $N$ and a predictable process $U$ such that $N_{0}=U_{0}=1$ and

$$
Z_{t}=Z_{0} N_{t} U_{t} \quad \mathrm{P}-a . s
$$

for $t \in\{0, \ldots, T\}$.
Proof. Define processes $U$ and $N$ recursively by $U_{0}:=N_{0}:=1$ and

$$
U_{t+1}:=U_{t} \frac{\Phi_{t}\left(Z_{t+1}\right)}{Z_{t}}, \quad N_{t+1}:=N_{t} \frac{Z_{t+1}}{\Phi_{t}\left(Z_{t+1}\right)} \quad \text { for } \quad t \in\{0, \ldots, T-1\}
$$

Observe that $U$ and $N$ are well defined since by assumption $\Phi_{t}\left(Z_{t}\right) \geq \Phi_{t}(\delta)=\delta$ due to monotonicity of $\Phi$. Obviously, $U$ is predictable, $N$ is a $\Phi$-martingale, and it follows easily by induction that $Z_{t}=Z_{0} N_{t} U_{t}$ for all $t \in\{0, \ldots, T\}$.

Now let $\left(N^{\prime}, U^{\prime}\right) \in \mathcal{H} \times \mathcal{H}$ be another pair as stated. We will show that $N_{t}^{\prime}=N_{t}, U_{t}^{\prime}=U_{t}$ P-a.s. for $t \in\{0, \ldots, T\}$ by induction. The case $t=0$ is trivial. So let $t \in\{0, \ldots, T-1\}$ such
that $N_{t}^{\prime}=N_{t}, U_{t}^{\prime}=U_{t}$ P-a.s. First, $\Phi_{t}\left(N_{t+1}\right)=N_{t}=N_{t}^{\prime}=\Phi_{t}\left(N_{t+1}^{\prime}\right)$ P-a.s. since $N, N^{\prime}$ are $\Phi$-martingales. Therefore by conditional positive homogeneity (C6)

$$
Z_{0} U_{t+1} N_{t}=Z_{0} U_{t+1} \Phi_{t}\left(N_{t+1}\right) \stackrel{(\mathrm{C} 6)}{=} \Phi_{t}\left(Z_{t+1}\right) \stackrel{(\mathrm{C} 6)}{=} Z_{0} U_{t+1}^{\prime} \Phi_{t}\left(N_{t+1}^{\prime}\right)=Z_{0} U_{t+1}^{\prime} N_{t} .
$$

Thus $U_{t+1}^{\prime}=U_{t+1}$ P-a.s. due to $Z_{0} N_{t}>0$ P-a.s., and

$$
Z_{0} U_{t+1} N_{t+1}=Z_{t+1}=Z_{0} U_{t+1}^{\prime} N_{t+1}^{\prime}=Z_{0} U_{t+1} N_{t+1}^{\prime} \quad \text { P-a.s. }
$$

Since $Z_{0} U_{t+1}>0$ P-a.s. we have $N_{t+1}=N_{t+1}^{\prime}$ P-a.s.
The next lemma is a multiplicative version of Lemma 5.3. For a proof see the appendix.
Lemma 6.2. Let $\Phi:=\left(\Phi_{t}\right)_{t \in\{0, \ldots, T\}}$ be a positively homogeneous recursive DMU functional, and let $Z:=\left(Z_{t}\right)_{t \in\{0, \ldots, T\}} \in \mathcal{H}$ with $Z_{t} \geq 0$ P-a.s. for any $t \in\{0, \ldots, T\}$. If $N:=\left(N_{t}\right)_{t \in\{0, \ldots, T\}}$ denotes any $\Phi$-martingale satisfying $N_{t}>0 \mathrm{P}$-a.s., then

$$
\Phi_{t}\left(Z_{\tau}\right)=\Phi_{t}\left(\frac{Z_{\tau} N_{T}}{N_{\tau}}\right) \quad \text { for } \quad t \in\{0, \ldots, T\} \quad \text { and } \quad \tau \in \mathcal{T}_{t}
$$

Obviously, under the assumptions of this section, $\Phi$ satisfies the Bellman principle (see Theorem 3.3), which allows us to establish a multiplicative dual representation for the stopping problem (3.1).

Theorem 6.3. Let the DMU functional $\Phi$ be as in Lemma 6.1, let $\mathcal{M}_{+1}^{\Phi}$ be the set of all $\Phi$-martingales $N$ with $N>0$ and $N_{0}=1$, and let $Z \in \mathcal{H}$ with $Z \geq 0$. We then may state for every $t \in\{0, \ldots, T\}$ the following:
(i)

$$
Y_{t}^{*}=\underset{\tau \in \mathcal{T}_{t}}{\operatorname{esssup}} \Phi_{t}\left(Z_{\tau}\right) \leq \inf _{N \in \mathcal{M}_{+1}^{\Phi}} \Phi_{t}\left(\max _{t \leq j \leq T} \frac{Z_{j} N_{T}}{N_{j}}\right) .
$$

(ii) If $\Phi$ satisfies, in addition, condition (C7), we have

$$
Y_{t}^{*}=\underset{N \in \mathcal{M}_{+1}^{\Phi}}{\operatorname{ess} \inf } \Phi_{t}\left(\max _{t \leq j \leq T} \frac{Z_{j} N_{T}}{N_{j}}\right) .
$$

(iii) If $Z$ is as in Lemma 6.1, we have

$$
Y_{t}^{*}=\underset{N \in \mathcal{M}_{+1}^{+}}{\operatorname{ess} \inf } \Phi_{t}\left(\max _{t \leq j \leq T} \frac{Z_{j} N_{T}}{N_{j}}\right)=\Phi_{t}\left(\max _{t \leq j \leq T} \frac{Z_{j} N_{T}^{*}}{N_{j}^{*}}\right),
$$

where $N^{*} \in \mathcal{M}_{0}^{\Phi}$ is the $\Phi$-martingale in the multiplicative decomposition of $Y^{*}$ in Lemma 6.1.
Proof. Statement (i) is an immediate consequence of Lemma 6.2.
For the proof of statement (ii) let us consider an arbitrary $\varepsilon>0$. The process $Z^{\varepsilon}$, defined by $Z_{t}^{\varepsilon}:=Z_{t} \vee \varepsilon$, induces the process $Y^{\varepsilon *}$ via $Y_{t}^{\varepsilon *}:=\operatorname{ess}^{\sup }{ }_{\tau \in \mathcal{T}_{t}} \Phi_{t}\left(Z_{\tau}^{\varepsilon}\right)$, which fulfills the assumptions of Lemma 6.1. Therefore we may find a pair $\left(U^{\varepsilon}, N^{\varepsilon}\right)$ consisting of a predictable process $U^{\varepsilon}$ and a $\Phi$-martingale $N^{\varepsilon} \in \mathcal{M}_{+1}^{\Phi}$ satisfying

$$
Y_{t}^{\varepsilon *}=Y_{0}^{\varepsilon *} N_{t}^{\varepsilon} U_{t}^{\varepsilon} \quad \text { P-a.s. } \quad \text { for } \quad t \in\{0, \ldots, T\} .
$$

Due to the conditional positive homogeneity of $\Phi$ and the predictability of $U^{\varepsilon}$, and since $N^{\varepsilon}$ is a $\Phi$-martingale, we may conclude that

$$
1=\Phi_{t}\left(\frac{N_{t+1}^{\varepsilon}}{N_{t}^{\varepsilon}}\right)=\Phi_{t}\left(\frac{Y_{t+1}^{\varepsilon *} U_{t}^{\varepsilon}}{Y_{t}^{\varepsilon *} U_{t+1}^{\varepsilon}}\right)=\frac{U_{t}^{\varepsilon}}{U_{t+1}^{\varepsilon}} \frac{\Phi_{t}\left(Y_{t+1}^{\varepsilon *}\right)}{Y_{t}^{\varepsilon *}} \quad \text { for } \quad t \in\{0, \ldots, T-1\} .
$$

In view of the Bellman principle this implies

$$
\frac{U_{t+1}^{\varepsilon}}{U_{t}^{\varepsilon}}=\frac{\Phi_{t}\left(Y_{t+1}^{\varepsilon *}\right)}{Y_{t}^{\varepsilon *}} \leq 1 \quad \text { for } \quad t \in\{0, \ldots, T-1\}
$$

Hence $U^{\varepsilon}$ has nonincreasing paths. Furthermore, $Z_{t}^{\varepsilon}=\Phi_{t}\left(Z_{t}^{\varepsilon}\right)$ and so, in particular, $Z_{t}^{\varepsilon} \leq Y_{t}^{\varepsilon *}$ due to the Bellman principle. Combining, we obtain for $t \in\{0, \ldots, T\}$

$$
\begin{align*}
\Phi_{t}\left(\max _{t \leq j \leq T} \frac{Z_{j}^{\varepsilon} N_{T}^{\varepsilon}}{N_{j}^{\varepsilon}}\right) & \leq \Phi_{t}\left(\max _{t \leq j \leq T} \frac{Y_{j}^{\varepsilon *} N_{T}^{\varepsilon}}{N_{j}^{\varepsilon}}\right)=\Phi_{t}\left(\max _{t \leq j \leq T} \frac{Y_{T}^{\varepsilon *} U_{j}^{\varepsilon}}{U_{T}^{\varepsilon}}\right) \\
& \leq U_{t}^{\varepsilon} \Phi_{t}\left(\frac{Y_{T}^{\varepsilon *}}{U_{T}^{\varepsilon}}\right)=U_{t}^{\varepsilon} \Phi_{t}\left(Y_{0}^{\varepsilon *} N_{T}^{\varepsilon}\right)=U_{t}^{\varepsilon} Y_{0}^{\varepsilon *} \Phi_{t}\left(N_{T}^{\varepsilon}\right)=Y_{t}^{\varepsilon *} \tag{6.1}
\end{align*}
$$

Now let a common function $g$ satisfy (2.1) in condition (C7) for all $Z_{j}^{\varepsilon}, j=0, \ldots, T$. By regularity and condition (C7) it then holds that

$$
\begin{aligned}
Y_{t}^{\varepsilon *} & :=\underset{\tau \in \mathcal{T}_{t}}{\operatorname{ess} \sup } \sum_{j=t}^{T} 1_{[\tau=j]} \Phi_{t}\left(Z_{\tau}^{\varepsilon}\right)=\underset{\tau \in \mathcal{T}_{t}}{\operatorname{ess} \sup } \sum_{j=t}^{T} 1_{[\tau=j]} \Phi_{t}\left(Z_{j}^{\varepsilon}\right) \\
& \leq \underset{\tau \in \mathcal{T}_{t}}{\operatorname{ess} \sup } \sum_{j=t}^{T} 1_{[\tau=j]}\left(\Phi_{t}\left(Z_{j}\right)+g(\varepsilon)\right)=Y_{t}^{*}+g(\varepsilon) .
\end{aligned}
$$

Hence with (6.1) we obtain

$$
Y_{t}^{*}+g(\varepsilon) \geq \underset{N \in \mathcal{M}_{+1}^{\infty}}{\operatorname{essinf}} \Phi_{t}\left(\max _{t \leq j \leq T} \frac{Z_{j} N_{T}}{N_{j}}\right) \stackrel{(\mathrm{i})}{\geq} Y_{t}^{*} \quad \text { for every } \quad t \in\{0, \ldots, T\} .
$$

The proof of (ii) is completed by sending $\varepsilon \rightarrow 0$.
Now let $Z$ and $\delta>0$ be as in Lemma 6.1, and take $\varepsilon$ such that $0<\varepsilon<\delta$. We thus have $Z^{\varepsilon}=Z$, and then statement (iii) follows from statement (i) and using (6.1) in the proof of (ii) (which holds independently of condition (C7)).

Examples 6.4. Theorem 6.3 may be applied in the following situations:

1. Let $\mathcal{Q}$ denote the set of equivalent martingale measures w.r.t. some arbitrage-free financial market, and let $Z:=\left(Z_{t}\right)_{t \in\{0, \ldots, T\}}$ be a nonnegative adapted process of P-essentially bounded random variables. The process $Z$ may be viewed as a discounted American option w.r.t. the recursive conditional positive homogeneous DMU functional $\Phi_{t}(\cdot):=\operatorname{ess} \sup _{\mathrm{Q} \in \mathcal{Q}} E_{\mathrm{Q}}\left[\cdot \mid \mathcal{F}_{t}\right]$ on $L^{\infty}(\Omega, \mathcal{F}, \mathrm{P})$. Furthermore, let us denote by $\mathcal{X}_{+1}$ the set of $X \in L^{\infty}\left(\Omega, \mathcal{F}_{T}, \mathrm{P}\right)$ with $X>0$ P-a.s. such that $\sup _{\mathrm{Q} \in \mathcal{Q}} E_{\mathrm{Q}}[X]=1$.

Then the superhedging price and the lowest arbitrage-free price of $Z$ may be represented by

$$
\inf _{X \in \mathcal{X}_{+1}} \Phi_{0}\left(\max _{t \in\{0, \ldots, T\}} \frac{Z_{t} X}{\Phi_{t}(X)}\right) \quad \text { and } \quad \inf _{X \in \mathcal{X}_{+1}} \bar{\Phi}_{0}\left(\max _{t \in\{0, \ldots, T\}} \frac{Z_{t} X}{\bar{\Phi}_{t}(X)}\right)
$$

respectively (see also Example 5.7).
2. Any family of $g$-expectations as in Example 2.7 with driver $g: \Omega \times[0, S] \times \mathbb{R} \times \mathbb{R}^{d} \rightarrow \mathbb{R}$ such that $g(\omega, s, \cdot, \cdot)$ is homogeneous of degree one for each $(\omega, s) \in \Omega \times[0, S]$ (cf. [23]).
7. Consumption based representation. We now propose a representation for the optimal stopping problem (3.1) which can be seen as a generalization of the consumption upper bound approach in [4], [5]. By this method one may infer from lower approximations upper approximations, and vice versa. In this context, we assume that $\Phi$ is a regular conditional translation invariant recursive DMU functional.

Using the Bellman principle we may prove the following theorem.
Theorem 7.1. For any $Z \in \mathcal{H}$ we have

$$
Y_{t}^{*}:=\underset{\tau \in \mathcal{T}_{t}}{\operatorname{esssup}} \Phi_{t}\left(Z_{\tau}\right)=\Phi_{t}\left(Z_{T}+\sum_{j=t}^{T-1}\left(Z_{j}-\Phi_{j}\left(Y_{j+1}^{*}\right)\right)^{+}\right), \quad t \in\{0, \ldots, T\},
$$

with empty sums being defined by zero.
Proof. We shall proceed by backward induction over $t$. The case $t=T$ is trivial. So let us assume for any $t \in\{1, \ldots, T\}$ that $Y_{t}^{*}=\Phi_{t}\left(Z_{T}+\sum_{j=t}^{T-1}\left(Z_{j}-\Phi_{j}\left(Y_{j+1}^{*}\right)\right)^{+}\right)$is valid. Then due to the Bellman principle $Y_{t-1}^{*}=\left(Z_{t-1}-\Phi_{t-1}\left(Y_{t}^{*}\right)\right)^{+}+\Phi_{t-1}\left(Y_{t}^{*}\right)$, which implies by assumption and recursiveness property (C4)

$$
\begin{aligned}
Y_{t-1}^{*} & =\left(Z_{t-1}-\Phi_{t-1}\left(Y_{t}^{*}\right)\right)^{+}+\Phi_{t-1}\left(\Phi_{t}\left(Z_{T}+\sum_{j=t}^{T-1}\left(Z_{j}-\Phi_{j}\left(Y_{j+1}^{*}\right)\right)^{+}\right)\right) \\
& =\left(Z_{t-1}-\Phi_{t-1}\left(Y_{t}^{*}\right)\right)^{+}+\Phi_{t-1}\left(Z_{T}+\sum_{j=t}^{T-1}\left(Z_{j}-\Phi_{j}\left(Y_{j+1}^{*}\right)\right)^{+}\right) .
\end{aligned}
$$

Then the application of conditional translation invariance yields

$$
\begin{aligned}
Y_{t-1}^{*} & =\Phi_{t-1}\left(\left(Z_{t-1}-\Phi_{t-1}\left(Y_{t}^{*}\right)\right)^{+}+Z_{T}+\sum_{j=t}^{T-1}\left(Z_{j}-\Phi_{j}\left(Y_{j+1}^{*}\right)\right)^{+}\right) \\
& =\Phi_{t-1}\left(Z_{T}+\sum_{j=t-1}^{T-1}\left(Z_{j}-\Phi_{j}\left(Y_{j+1}^{*}\right)\right)^{+}\right)
\end{aligned}
$$

The interesting feature of the representation in Theorem 7.1 is that if we replace $Y^{*}$ on the right-hand side by a lower (upper) approximation, we obtain an upper (lower) bound for $Y^{*}$ on the left-hand side.
8. Numerical approaches for optimal stopping of some specific DMU functionals. In this section we sketch how the different representations developed in sections 4, 5, 6, and 7 may be utilized for constructing (upper and/or lower) approximations to the optimal value of stopping problem (3.1). In order to enable a feasible algorithm or simulation procedure for optimal stopping of a particular DMU functional, we naturally presume that we have a feasible algorithm or simulation procedure for the functional itself at hand. In this respect we underline that numerical (simulation) methods for specific DMU functionals are an interesting issue in their own right but considered to be beyond the scope of this article. Another natural assumption is that we have some underlying process with some kind of Markovian structure which can be simulated straightforwardly. More specifically, we assume that we are in the following setting.

## Setting for solving general optimal stopping problems by simulation.

(i) The filtration $\left(\mathcal{F}_{t}\right)_{t \in\{0, \ldots, T\}}$ is generated by some underlying stochastic process $S:=$ $\left(S_{t}\right)_{t \in\{0, \ldots, T\}}$ in some multidimensional state space, e.g., $\mathbb{R}^{d}$.
(ii) The process $Z:=\left(Z_{t}\right)_{t \in\{0, \ldots, T\}}$ under consideration satisfies $Z_{t}=h\left(t, S_{t}\right)$ for some known nonnegative measurable function $h$. For ease of exposition, $h$ is assumed to be bounded.
(iii) The DMU functional $\Phi=\left(\Phi_{t}\right)_{t \in\{0, \ldots, T\}}$ is regular and recursively generated by $\left(\Psi_{t}\right)_{t \in\{0, \ldots, T\}}$ with generators satisfying $\Psi_{t}(X)=X$ if $X \in \mathcal{F}_{t}$ for any $t \in\{0, \ldots, T\}$. Hence, in particular, $\Phi$ is recursive with $\Phi_{t}(X)=X$ if $X \in \mathcal{F}_{t}$ for $t \in\{0, \ldots, T\}$.
(iv) For any $t \in\{0, \ldots, T\}$, we have $\Phi_{t}(X)$ being $\sigma\left\{S_{t}\right\}$-measurable if $X$ is $\sigma\left\{S_{t}, \ldots, S_{T}\right\}$ measurable (we might think of $S$ being Markovian w.r.t. the functional $\Phi$ ). This condition is, e.g., guaranteed in the case that for any $u, t \in\{0, \ldots, T\}$ with $u \leq t$ we have that $\Psi_{t}(X)$ is $\sigma\left\{S_{u}, \ldots, S_{t}\right\}$-measurable whenever $X$ is $\sigma\left\{S_{u}, \ldots, S_{t+1}\right\}$-measurable.
(v) For any $t \in\{0, \ldots, T\}$, we may compute $\Phi_{t}(X) \in \sigma\left\{S_{t}\right\}$ if $X \in \sigma\left\{S_{t}, \ldots, S_{T}\right\}$ by some kind of simulation method.
In the standard case, where $\Phi$ represents the ordinary conditional expectation and $S$ is Markovian in the ordinary sense, (iii), (iv), and (v) are obviously fulfilled. A canonical way of evaluating conditional expectations is (Monte Carlo) simulation from a particular state $\left(t, S_{t}\right)$ (particularly in higher dimensions). In general there are many interesting examples, for instance, within the class of $g$-expectations.

Example 8.1. Let $\Phi$ be a family of $g$-expectations as in Example 2.7 with Brownian motion $B=\left(B_{s}\right)_{s \geq 0}$ and driver $g: \Omega \times[0, S] \times \mathbb{R} \times \mathbb{R}^{n} \rightarrow \mathbb{R}$ being of the form $g(\omega, s, y, z):=f\left(S_{s}, y, z\right)$. Here $f: \mathbb{R}^{n} \times \mathbb{R} \times \mathbb{R}^{d} \rightarrow \mathbb{R}$ is any Lipschitz function with $f(\cdot, 0) \equiv 0$, and $\left(S_{s}\right)_{s \geq 0}$ is an $n$-dimensional diffusion process with dynamics given by the SDE

$$
d S_{s}=\mu\left(S_{s}\right) d t+\sigma\left(S_{s}\right) d B_{s} .
$$

Under some further conditions of regularity for $\mu, \sigma$, and $f$, it may be verified that $\Phi$ satisfies assumption (iv) (cf. [21, Theorem 6.2]). Furthermore, simulation algorithms as required in assumption (v) are already available (see, e.g., [19], [31]).

Moreover, if $f$ does not depend on $y$, and is sublinear in $z$, then there is some set $\mathcal{Q}$ of probability measures which are absolutely continuous w.r.t. P such that $\Phi$ admits the following
robust representation:

$$
\Phi_{t}(X)=\underset{\mathrm{Q} \in \mathcal{Q}}{\operatorname{ess} \sup } E_{\mathrm{Q}}\left[X \mid \sigma\left\{S_{1}, \ldots, S_{t}\right\}\right],
$$

where the essential supremum is attained (see [9, proof of Theorem 3.1]).
Below we will outline the implementation of the above simulation setting for different solution representations proposed in sections 4, 5, 6, and 7 .

Policy iteration. The policy iteration method in section 4 may be readily applied if the time consistent stopping family $\left(\tau_{t}\right)$ we start with is such that $\left\{\tau_{t}=t\right\} \in \sigma\left\{S_{t}\right\}$. For example we just take the trivial family $\tau_{t}=t, t \in\{0, \ldots, T\}$. Then the iteration procedure will be an analogue to the one spelled out in [26]. In short, given an input stopping family $\left(\tau_{t}\right)$, simulate a set of $N$ (outer) trajectories $S^{(n)}, n=1, \ldots, N$, from $t=0$ to $T$. Determine on each outer trajectory $S^{(n)}$ the improved stopping time $\widehat{\tau}_{0}$. For this, one needs to simulate for each time $s=0,1, \ldots$ a set of $M$ (inner) trajectories $\left({ }_{m} S_{u}^{(n)}\right)_{u=s, \ldots, T}, m=1, \ldots, M$, to check by simulation whether the event

$$
\left\{\Phi_{s}\left(Z_{s}\right) \geq \max _{s+1 \leq u \leq T} \Phi_{s}\left(Z_{\tau_{u}}\right)\right\}
$$

in (4.3) is true. If $s^{(n)}$ is the first time where this is true, we put $\widehat{\tau}_{0}^{(n)}=s^{(n)}$ on trajectory $n$. Finally, we compute $\Phi_{0}\left(Z_{\widehat{\tau}_{0}}\right)$ from the sample $Z_{\widehat{\tau}_{0}}^{(n)}, n=1, \ldots, N$.

Dual upper bounds. We consider the construction of an additive dual upper bound for a regular, recursive DMU functional, which is translation invariant. Let us assume that we are given a proxy $Y_{t}=U\left(t, S_{t}\right)$ of the Snell envelope $Y_{t}^{*}=U^{*}\left(t, S_{t}\right)$. Note that the Snell envelope is indeed of this form due to assumptions (i), (ii), and (iv). For instance, for the DMU functional in Example 3.4, a proxy may be constructed by approximating the Snell envelope w.r.t. a simpler functional, replacing the representing set $\mathcal{Q}$ of probability measures by a smaller subset or even a singleton. Let $M^{Y}$ be the Doob $\Phi$-martingale of $Y$, and consider the upper bound

$$
\begin{aligned}
Y_{0}^{u p} & =\Phi_{0}\left(\max _{0 \leq t \leq T}\left(Z_{t}+M_{T}^{Y}-M_{t}^{Y}\right)\right) \\
& =\Phi_{0}\left(\max _{0 \leq t \leq T}\left(h\left(t, S_{t}\right)+\sum_{s=t}^{T-1}\left[U\left(s+1, S_{s+1}\right)-\Phi_{s}\left(U\left(s+1, S_{s+1}\right)\right)\right]\right)\right)
\end{aligned}
$$

Similarly as in [1] we are going to construct an approximation of this upper bound by a nested simulation. We simulate $N$ (outer) trajectories $S^{(n)}, n=1, \ldots, N$, from $t=0$ to $T$, and for each outer trajectory $n$, and time $s, s<T$, a set of $M$ (inner) two step trajectories $\left({ }_{m} S_{u}^{(n)}\right)_{u=s, s+1}, m=1, \ldots, M$. On a fixed outer trajectory $S^{(n)}$ we then construct for each $s$ an approximation of $\Phi_{s}^{(n)}\left(U\left(s+1, S_{s+1}\right)\right)$ by the inner sample $U\left(s+1,{ }_{1} S_{s+1}^{(n)}\right), \ldots, U\left(s+1,{ }_{M} S_{s+1}^{(n)}\right)$ and next determine

$$
\zeta^{(n)}:=\max _{0 \leq t \leq T}\left(h\left(t, S_{t}^{(n)}\right)+\sum_{s=t}^{T-1}\left[U\left(s+1, S_{s+1}^{(n)}\right)-\Phi_{s}^{(n)}\left(U\left(s+1, S_{s+1}\right)\right)\right]\right)
$$

We thus end up with the sample $\zeta^{(1)}, \ldots, \zeta^{(N)}$ of the random variable $\zeta:=\max _{0 \leq t \leq T}\left(Z_{t}+\right.$ $\left.M_{T}^{Y}-M_{t}^{Y}\right)$, from which finally $Y_{0}^{u p}=\Phi_{0}(\zeta)$ may be estimated.

Multiplicative and consumption upper bounds. From the simulation methods sketched above it will be clear in principle how to construct a multiplicative upper bound for a positively homogeneous DMU functional, and how to construct an upper (lower) bound due to the consumption representation in Theorem 7.1 for a translation invariant functional when a lower (upper) bound of the Snell envelope is given.

Concluding remark. In this article different representations for the optimal stopping problem w.r.t. general DMU functionals are presented. It is shown that these representations allow for a numerical treatment of the generalized stopping problem. A detailed analysis of the numerical algorithms sketched in section 8, which will depend on particular properties of the functional under consideration, remains to be done in future work.

## Appendix.

Proof of Theorem 4.1. The inequalities $Y_{t} \leq \widetilde{Y}_{t}$ and $\widehat{Y}_{t} \leq Y_{t}^{*}$ are obvious for any $t \in$ $\{0, \ldots, T\}$. So inequality $\widetilde{Y}_{t} \leq \widehat{Y}_{t}$ is left to show. We shall use backward induction.

Due to the definitions of $\widetilde{Y}$ and $\widehat{Y}$, we have $\widetilde{Y}_{T}=\widehat{Y}_{T}=\Phi_{T}\left(Z_{T}\right)$. Suppose that $\widetilde{Y}_{t} \leq \widehat{Y}_{t}$ holds for any $t \in\{1, \ldots, T\}$. We then have to show that $\widetilde{Y}_{t-1} \leq \widehat{Y}_{t-1}$. For this we first show sequentially the following:
(1) $1_{\left[\hat{\tau}_{t-1}=t-1\right]} \widehat{Y}_{t-1}=1_{\left[\hat{\tau}_{t-1}=t-1\right]} \Phi_{t-1}\left(Z_{t-1}\right)$.
(2) $1_{\left[\hat{\tau}_{t-1}>t-1\right]} \widehat{Y}_{t-1} \geq 1_{\left[\hat{\tau}_{t-1}>t-1\right]} \max _{t \leq s \leq T} \Phi_{t-1}\left(Z_{\tau_{s}}\right)$.
(3) $\Phi_{t-1}\left(Z_{\tau_{t-1}}\right) \leq \max \left\{\Phi_{t-1}\left(Z_{t-1}\right), \max _{t \leq s \leq T} \Phi_{t-1}\left(Z_{\tau_{s}}\right)\right\}$.

Due to the definition of $\hat{\tau}_{t-1}$, we have, on the set $\left\{\hat{\tau}_{t-1}=t-1\right\}, \Phi_{t-1}\left(Z_{t-1}\right) \geq \max _{t \leq s \leq T} \Phi_{t-1}\left(Z_{\tau_{s}}\right)$ and, on the set $\left\{\hat{\tau}_{t-1}>t-1\right\}, \Phi_{t-1}\left(Z_{t-1}\right)<\max _{t \leq s \leq T} \Phi_{t-1}\left(Z_{\tau_{s}}\right)$. Thus we may conclude immediately from (1)-(3) that

$$
\widehat{Y}_{t-1} \geq \max \left\{\Phi_{t-1}\left(Z_{t-1}\right), \max _{t \leq s \leq T} \Phi_{t-1}\left(Z_{\tau_{s}}\right)\right\} \geq \max \left\{\Phi_{t-1}\left(Z_{\tau_{t-1}}\right), \max _{t \leq s \leq T} \Phi_{t-1}\left(Z_{\tau_{s}}\right)\right\}=\widetilde{Y}_{t-1}
$$

as required.
Proof of (1). By regularity condition (C2) we may find sequentially

$$
1_{\left[\hat{\tau}_{t-1}=t-1\right]} \widehat{Y}_{t-1}=\Phi_{t-1}\left(1_{\left[\hat{\tau}_{t-1}=t-1\right]} Z_{\hat{\tau}_{t-1}}\right)=\Phi_{t-1}\left(1_{\left[\hat{\hat{t}}_{t-1}=t-1\right]} Z_{t-1}\right)=1_{\left[\hat{\tau}_{t-1}=t-1\right]} \Phi_{t-1}\left(Z_{t-1}\right),
$$

which proves (1).
Proof of (2). $1_{\left[\hat{\tau}_{t-1}>t-1\right]} Z_{\hat{\tau}_{t-1}}=1_{\left[\hat{\tau}_{t-1}>t-1\right]} Z_{\hat{\tau}_{t}}$ due to the time consistency of $\left(\hat{\tau}_{t}\right)_{t \in\{0, \ldots, T\}}$. Hence the regularity condition implies

$$
\begin{aligned}
1_{\left[\hat{\tau}_{t-1}>t-1\right]} \widehat{Y}_{t-1} & =\Phi_{t-1}\left(1_{\left[\hat{\tau}_{t-1}>t-1\right]} Z_{\hat{t}_{t-1}}\right)=\Phi_{t-1}\left(1_{\left[\hat{\tau}_{t-1}>t-1\right]} Z_{\hat{\tau}_{t}}\right)=1_{\left[\hat{\tau}_{t-1}>t-1\right]} \Phi_{t-1}\left(Z_{\hat{\tau}_{t}}\right) \\
& =1_{\left[\hat{\tau}_{t-1}>t-1\right]} \Psi_{t-1}\left(\widehat{Y}_{t}\right) .
\end{aligned}
$$

By the induction hypothesis we have $\widehat{Y}_{t} \geq \widetilde{Y}_{t}$, so we may conclude by monotonicity of $\Psi_{t-1}$ that

$$
\begin{aligned}
1_{\left[\hat{\tau}_{t-1}>t-1\right]} \widehat{Y}_{t-1} & \geq 1_{\left[\hat{\tau}_{t-1}>t-1\right]} \Psi_{t-1}\left(\widetilde{Y}_{t}\right) \geq 1_{\left[\hat{\tau}_{t-1}>t-1\right]} \max _{t \leq s \leq T} \Psi_{t-1}\left(\Phi_{t}\left(Z_{\tau_{s}}\right)\right) \\
& =1_{\left[\hat{\tau}_{t-1}>t-1\right]} \max _{t \leq s \leq T} \Phi_{t-1}\left(Z_{\tau_{s}}\right) .
\end{aligned}
$$

Thus (2) is shown.
Proof of (3). Using regularity condition (C2) we obtain

$$
\begin{aligned}
\Phi_{t-1}\left(Z_{\tau_{t-1}}\right) & =1_{\left[\tau_{t-1}=t-1\right]} \Phi_{t-1}\left(Z_{\tau_{t-1}}\right)+1_{\left[\tau_{t-1}>t-1\right]} \Phi_{t-1}\left(Z_{\tau_{t-1}}\right) \\
& =\Phi_{t-1}\left(1_{\left[\tau_{t-1}=t-1\right]} Z_{t-1}\right)+\Phi_{t-1}\left(1_{\left[\tau_{t-1}>t-1\right]} Z_{\tau_{t-1}}\right) .
\end{aligned}
$$

$1_{\left[\tau_{t-1}>t-1\right]} Z_{\tau_{t-1}}=1_{\left[\tau_{t-1}>t-1\right]} Z_{\tau_{t}}$ due to the time consistency of $\left(\tau_{t}\right)_{t \in\{0, \ldots, T\}}$. Hence the application of regularity again yields

$$
\begin{aligned}
\Phi_{t-1}\left(Z_{\tau_{t-1}}\right) & =\Phi_{t-1}\left(1_{\left[\tau_{t-1}=t-1\right]} Z_{t-1}\right)+\Phi_{t-1}\left(1_{\left[\tau_{t-1}>t-1\right]} Z_{\tau_{t}}\right) \\
& =1_{\left[\tau_{t-1}=t-1\right]} \Phi_{t-1}\left(Z_{t-1}\right)+1_{\left[\tau_{t-1}>t-1\right]} \Phi_{t-1}\left(Z_{\tau_{t}}\right),
\end{aligned}
$$

obviously implying (3), and hence completing the proof.
Proof of Lemma 5.3. We shall show the statement of Lemma 5.3 via backward induction. The case $t=T$ is trivial since $\mathcal{T}_{T}=\{T\}$. So let us assume that for any $t \in\{1, \ldots, T\}$ we have $\Phi_{t}\left(Z_{\sigma}\right)=\Phi_{t}\left(Z_{\sigma}+M_{T}-M_{\sigma}\right)$ for every $\sigma \in \mathcal{T}_{t}$. Let us fix an arbitrary $\tau \in \mathcal{T}_{t-1}$ and define $\sigma(\tau):=t 1_{[\tau=t-1]}+\tau 1_{[\tau>t-1]} \in \mathcal{T}_{t}$. Then by assumption $\Phi_{t}\left(Z_{\sigma(\tau)}\right)=\Phi_{t}\left(Z_{\sigma(\tau)}+M_{T}-M_{\sigma(\tau)}\right)$, which implies, via regularity and recursiveness,

$$
\begin{aligned}
1_{[\tau>t-1]} \Phi_{t-1}\left(Z_{\tau}\right) & =1_{[\tau>t-1]} \Phi_{t-1}\left(Z_{\sigma(\tau)}\right)=1_{[\tau>t-1]} \Phi_{t-1} \circ \Phi_{t}\left(Z_{\sigma(\tau)}\right) \\
& =1_{[\tau>t-1]} \Phi_{t-1}\left(\Phi_{t}\left(Z_{\sigma(\tau)}+M_{T}-M_{\sigma(\tau)}\right)\right) \\
& =1_{[\tau>t-1]} \Phi_{t-1}\left(Z_{\sigma(\tau)}+M_{T}-M_{\sigma(\tau)}\right) \\
& \stackrel{(\mathrm{C} 2)}{=} 1_{[\tau>t-1]} \Phi_{t-1}\left(Z_{\tau}+M_{T}-M_{\tau}\right) .
\end{aligned}
$$

Moreover, by regularity, conditional translation invariance, and the $\Phi$-martingale property of $M$, we have

$$
\begin{aligned}
1_{[\tau=t-1]} \Phi_{t-1}\left(Z_{\tau}+M_{T}-M_{\tau}\right) & \stackrel{(\mathrm{C} 2)}{=} 1_{[\tau=t-1]} \Phi_{t-1}\left(Z_{t-1}+M_{T}-M_{t-1}\right) \\
& \stackrel{(\mathrm{C} 3)}{=} 1_{[\tau=t-1]}\left(Z_{t-1}-M_{t-1}+\Phi_{t-1}\left(M_{T}\right)\right) \\
& =1_{[\tau=t-1]} Z_{t-1}=1_{[\tau=t-1]} \Phi_{t-1}\left(Z_{t-1}\right),
\end{aligned}
$$

which completes the proof.
Proof of Lemma 6.2. We shall show the statement of the lemma by backward induction. The case $t=T$ is trivial since $\mathcal{T}_{T}=\{T\}$. Let us assume that for $t \in\{1, \ldots, T\}$ the equality $\Phi_{t}\left(Z_{\tau}\right)=\Phi_{t}\left(\frac{Z_{\tau} N_{T}}{N_{\tau}}\right)$ is valid for every $\tau \in \mathcal{T}_{t}$.

Consider an arbitrary $\tau \in \mathcal{T}_{t-1}$, and define $\sigma(\tau):=1_{\tau=t-1} t+1_{\tau>t-1} \tau \in \mathcal{T}_{t}$. By the induction assumption we have $\Phi_{t}\left(\frac{Z_{\sigma(\tau)} N_{T}}{N_{\sigma(\tau)}}\right)=\Phi_{t}\left(Z_{\sigma(\tau)}\right)$, so that regularity condition (C2) and recursiveness imply

$$
\begin{aligned}
& 1_{[\tau>t-1]} \Phi_{t-1}\left(\frac{Z_{\tau} N_{T}}{N_{\tau}}\right) \stackrel{(\mathrm{C} 2)}{=} 1_{[\tau>t-1]} \Phi_{t-1}\left(\frac{Z_{\sigma(\tau)} N_{T}}{N_{\sigma(\tau)}}\right)=1_{[\tau>t-1]} \Phi_{t-1}\left(\Phi_{t}\left(\frac{Z_{\sigma(\tau)} N_{T}}{N_{\sigma(\tau)}}\right)\right) \\
&=1_{[\tau>t-1]} \Phi_{t-1}\left(\Phi_{t}\left(Z_{\sigma(\tau)}\right)\right)=1_{[\tau>t-1]} \Phi_{t-1}\left(Z_{\sigma(\tau)}\right) \\
& \stackrel{(\mathrm{C} 2)}{=} 1_{[\tau>t-1]} \Phi_{t-1}\left(Z_{\tau}\right) .
\end{aligned}
$$

Moreover, by regularity (C2), conditional positive homogeneity (C6), and the fact that $N$ is a $\Phi$-martingale, it holds that

$$
\begin{aligned}
1_{[\tau=t-1]} \Phi_{t-1}\left(\frac{Z_{\tau} N_{T}}{N_{\tau}}\right) & \stackrel{(\mathrm{C} 2)}{=} 1_{[\tau=t-1]} \Phi_{t-1}\left(\frac{Z_{t-1} N_{T}}{N_{t-1}}\right) \\
& \stackrel{(\mathrm{C} 6)}{=} 1_{[\tau=t-1]} \frac{Z_{t-1}}{N_{t-1}} \Phi_{t-1}\left(N_{T}\right)=1_{[\tau=t-1]} Z_{t-1}=1_{[\tau=t-1]} \Phi_{t-1}\left(Z_{\tau}\right)
\end{aligned}
$$

Acknowledgments. We would like to thank two perceptive referees and the associated editor for their valuable comments, which have led to a substantial improvement of the paper.

## REFERENCES

[1] L. Andersen and M. Broadie, A primal-dual simulation algorithm for pricing multidimensional American options, Management Sci., 50 (2004), pp. 1222-1234.
[2] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, Coherent measures of risk, Math. Finance, 9 (1999), pp. 203-228.
[3] P. Artzner, F. Delbaen, J.-M. Eber, D. Heath, and H. Ku, Coherent multiperiod risk adjusted values and Bellman's principle, Ann. Oper. Res., 152 (2007), pp. 5-22.
[4] D. Belomestny and G. N. Milstein, Monte Carlo evaluation of American options using consumption processes, Int. J. Theor. Appl. Finance, 9 (2006), pp. 1-27.
[5] D. Belomestny, G. N. Milstein, and V. Spokoiny, Regression methods in pricing American and Bermudan options using consumption processes, Quant. Finance, 9 (2009), pp. 315-327.
[6] E. Bayraktar, I. Karatzas, and S. Yao, Optimal Stopping for Dynamic Convex Risk Measures, http://arxiv.org/abs/0909.4948 (2009).
[7] E. Bayraktar and S. Yao, Optimal Stopping for Non-linear Expectations, http://arxiv.org/abs/0905. 3601 (2009).
[8] M. Broadie and P. Glasserman, A stochastic mesh method for pricing high-dimensional American options, J. Comput. Finance, 7 (2004), pp. 35-72.
[9] Z. Chen, R. Kulperger, and L. Jiang, Jensen's inequality for g-expectations, Part 1, C. R. Math. Acad. Sci. Paris, 337 (2003), pp. 725-730.
[10] P. Cheridito, F. Delbaen, and M. Kupper, Dynamic monetary risk measures for bounded discretetime processes, Electron. J. Probab., 11 (2006), pp. 57-106.
[11] P. Cheridito and M. Kupper, Recursiveness of indifference prices and translation invariant preferences, Math. Financ. Econ., 2 (2009), pp. 173-188.
[12] F. Delbaen, The structure of $m$-stable sets and in particular of the set of risk neutral measures, in In Memoriam Paul-André Meyer: Séminaire de Probabilités XXXIX, Lecture Notes in Math. 1874, Springer, Berlin, 2006, pp. 215-258.
[13] K. Detlefsen and G. Scandolo, Conditional and dynamic convex risk measures, Finance Stoch., 9 (2005), pp. 539-561.
[14] N. El Karoui, S. Peng, and M. C. Quenez, Backward stochastic differential equations in finance, Math. Finance, 7 (1997), pp. 1-71.
[15] N. El Karoui and C. Ravanelli, Cash subadditive risk measures under interest rate ambiguity, Math. Finance, 19 (2009), pp. 561-590.
[16] H. Föllmer and I. Penner, Convex risk measures and the dynamics of their penalty functions, Statist. Decisions, 24 (2006), pp. 61-96.
[17] H. Föllmer and A. Schied, Stochastic Finance, 2nd ed., de Gruyter, Berlin, New York, 2004.
[18] M. Frittelli and M. Maggis, Conditional certainty equivalent, Int. J. Theor. Appl. Finance, to appear.
[19] E. Gobet, J. P. Lemor, and X. Warin, Rate of convergence of an empirical regression method for solving generalized backward stochastic differential equations, Bernoulli, 12 (2006), pp. 889-916.
[20] M. B. Haugh and L. Kogan, Pricing American options: A duality approach, Oper. Res., 52 (2004), pp. 258-270.
[21] Y. Hu and J. Ma, Nonlinear Feynman-Kac formula and discrete-functional-type BSDEs with continuous coefficients, Stochastic Process. Appl., 112 (2004), pp. 23-51.
[22] F. Jamshidian, The duality of optimal exercise and domineering claims: A Doob-Meyer decomposition approach to the Snell envelope, Stochastics, 79 (2007), pp. 27-60.
[23] L. Jiang, Convexity, translation invariance and subadditivity for $g$-expectations and related risk measures, Ann. Appl. Probab., 18 (2008), pp. 245-258.
[24] R. Kaas, M. Goovaerts, J. Dhaene, and M. Denuit, Modern Actuarial Risk Theory, 2nd ed., Springer, Berlin, Heidelberg, 2008.
[25] S. Klöppel and M. Schweizer, Dynamic indifference valuation via convex risk measure, Math. Finance, 17 (2007), pp. 599-627.
[26] A. Kolodko and J. Schoenmakers, Iterative construction of the optimal Bermudan stopping time, Finance Stoch., 10 (2006), pp. 27-49.
[27] V. Krätschmer, On $\sigma$-Additive Robust Representation of Convex Risk Measures for Unbounded Financial Positions in the Presence of Uncertainty about the Market Model, SFB 649 discussion paper 2007-010; available online from http://sfb649.wiwi.hu-berlin.de.
[28] F. A. Longstaff and R. S. Schwartz, Valuing American options by simulation: A simple least-square approach, Rev. Financ. Stud., 14 (2001), pp. 113-147.
[29] S. Peng, Backward SDE and related g-expectations, in Backward Stochastic Differential Equations, N. El Karoui and L. Mazliak, eds., Pitman Res. Notes Math. Ser. 364, Longman, Harlow, 1997, pp. 141-159.
[30] S. Peng, Modelling Derivative Pricing Mechanisms with Their Generating Functions, http://arxiv.org/ abs/math/0605599 (2006).
[31] S. Peng and M. Xu, Numerical algorithms for backward stochastic differential equations with 1-d Brownian motion: Convergence and simulations, M2AN Math. Model. Numer. Anal., to appear.
[32] F. Riedel, Optimal stopping with multiple priors, Econometrica, 77 (2009), pp. 857-908.
[33] L. C. G. Rogers, Monte Carlo valuation of American options, Math. Finance, 12 (2002), pp. 271-286.
[34] B. Roorda and J. M. Schumacher, Time consistency conditions for acceptability measures, with an application to Tail Value at Risk, Insurance Math. Econom., 40 (2007), pp. 209-230.
[35] B. Roorda, J. M. Schumacher, and J. Engwerda, Coherent acceptability measures in multiperiod models, Math. Finance, 15 (2005), pp. 589-612.
[36] E. Rosazza Gianin, Risk measures via g-expectations, Insurance Math. Econom., 39 (2006), pp. 19-34.
[37] J. Tsitsiklis and B. Van Roy, Regression methods for pricing complex American style options, IEEE Trans. Neural Networks, 12 (2001), pp. 694-703.
[38] S. Tutsch, Update rules for convex risk measures, Quant. Finance, 8 (2008), pp. 833-843.
[39] S. Weber, Distribution-invariant risk measures, information, and dynamic consistency, Math. Finance, 16 (2006), pp. 419-441.

# Parametrix Approximation of Diffusion Transition Densities* 

Francesco Corielli ${ }^{\dagger}$, Paolo Foschi ${ }^{\ddagger}$, and Andrea Pascucci ${ }^{\S}$


#### Abstract

A new analytical approximation tool, derived from the classical PDE theory, is introduced in order to build approximate transition densities of diffusions. The tool is useful for approximate pricing and hedging of financial derivatives and for maximum likelihood and method of moments estimates of diffusion parameters. The approximation is uniform with respect to time and space variables. Moreover, easily computable error bounds are available in any dimension.


Key words. diffusions, transition densities, option pricing, analytic approximations, parabolic equations, parametrix method

AMS subject classifications. $35 \mathrm{~K} 57,35 \mathrm{~K} 65,35 \mathrm{~K} 70$
DOI. 10.1137/080742336

1. Introduction and motivation. The ability of approximating, with some explicit measure of error, the fundamental solution of a parabolic partial differential equation (PDE) has become of paramount relevance for economical and financial applications. The origin of this can be traced back to the 1960s with the widespread introduction of diffusion-based modeling. In financial applications, transition densities play a central role in several instances from pricing and hedging of financial derivatives to parameter estimation and calibration (by likelihood or moment method based techniques) and solution of optimal control problems.

The early examples of diffusion models considered only very simple specifications, usually in the class of linear diffusions, for which explicit transition densities are known. However, the requirement of more statistical realism and the need for a multidimensional framework stimulated more complex models whose explicit solutions are unavailable. This favored the implementation of standard (in other fields) numerical procedures, e.g., finite differences or finite elements methods, and the development of a number of very useful Monte Carlo techniques.

An alternative or, more precisely, a complement to the use of numerical methods is given by the powerful machinery of analytical approximations. These, in fields like physics and engineering, are the tools of choice for the study of the qualitative properties of a model and for comparison between models. Moreover, these techniques provide the basis for efficient numerical algorithms specific to each model. Today the literature concerning asymptotic expansions and singular perturbation theory applied to finance is vast. We quote, among

[^85]others, the papers by Whalley and Wilmott [34] in the study of transaction costs; Hagan and Woodward [19] and Hagan et al. [18] for the implied volatility in CEV and SABR models; Fouque, Papanicolaou, and Sircar [15], Fouque et al. [16], Howison [21], Widdicks et al. [35], and Svoboda-Greenwood [32] for other local and stochastic volatility models; Barone-Adesi and Elliott [3], Broadie and Detemple [5], and Kuske and Keller [27] in the study of American options; Widdicks et al. [36] in multiasset option pricing; Turnbull and Wakeman [33], Zhang [37], and Dewynne and Shaw [10] concerning average (Asian) options; and Broadie, Glasserman, and Kou [6] in the study of discrete barrier options. For other analytical or partially analytical approximation methods we refer the reader to Aït-Sahalia [1] and [2].

In this paper we propose a new technique, the parametrix method, for the analytical approximation of the fundamental solution of a general parabolic PDE. As far as we know, contrary to perturbation techniques, the parametrix has not been previously examined as a numerical method, nor has it been already employed for approximation purposes in finance or other fields. Indeed, the classical parametrix technique was introduced by Levi [28] as a theoretical method to prove the existence of the fundamental solution of a parabolic PDE: ${ }^{1}$ as such, it is not optimized for the purpose of numerical computation of solutions.

The first main contribution of this paper is the conversion and adjustment of this classical tool in PDE analysis as a numerical method. In particular, we modify the parametrix technique to yield an approximation useful for computational purposes. Moreover, we introduce the new concept of the backward parametrix and we show how it is both more appropriate for computations and more useful for model interpretation purposes than the standard parametrix. As we will see, the backward parametrix lends itself to a direct probabilistic interpretation which is not readily available for the standard parametrix. This interpretation is very useful when the backward parametrix is applied to the problem of pricing financial derivatives as it allows a financial interpretation of the leading and the correction terms in the expansion.

The second main contribution is the derivation, in the new context, of easy-to-compute, a priori (i.e., independent of the solution), uniform bounds on the approximation error both for the PDE solution and for its derivatives. The hypotheses under which the series approximation uniformly converges and under which the evaluation of the error term for a truncated series holds are very general. They can be checked by the sole knowledge of some general property of the diffusion model and are substantially the same hypotheses commonly required for a diffusion problem and a parabolic problem to be equivalent. While perhaps the most easy to check, these hypotheses are not the most general. Indeed the proposed method also works in cases where these hypotheses are not true: we show an example of this in section 3. The results of the paper are then applied to a number of relevant particular cases and compared with other numerical and seminumerical evaluation procedures known in the literature.

The rest of the paper is organized as follows. In section 2, we derive and slightly generalize the classical parametrix expansion in the one-dimensional setting; moreover, we introduce the

[^86]new backward parametrix. We also give a financial interpretation of the derivation of the parametrix. In section 3 we find closed form approximate solutions in a general local volatility model. In section 4 we perform some numerical tests and compare the performance of the parametrix with other known approximate methods. In section 5 , we derive a priori bounds on the approximation error for both the forward parametrix and the backward parametrix. The appendix contains a number of lemmas used in the proofs.
2. Forward and backward parametrix approximations. The aim of this section is to give the main ideas by presenting our results in the simple case of a one-dimensional model. In section 5 the parametrix is derived in its full generality and complete proofs are given. Our contribution is twofold: first, we compute explicit error bounds for the classical parametrix (hereafter called the forward parametrix) approximation introduced by Levi [28]; second, we introduce the so-called backward parametrix, an alternative expansion that is more significant for the financial interpretation and from the computational point of view.

In what follows we denote by $z=(x, t), \zeta=(\xi, \tau)$, and $w=(y, s)$ the points in $\mathbb{R} \times \mathbb{R}$ and consider the parabolic PDE

$$
\begin{equation*}
L u(z):=a(z) \partial_{x x} u(z)+b(z) \partial_{x} u(z)+c(z) u(z)-\partial_{t} u(z)=0 . \tag{2.1}
\end{equation*}
$$

The fundamental solution $\Gamma=\Gamma(z ; \zeta)$ of $L$ is a function such that the following hold:
(i) $L \Gamma(\cdot ; \zeta)=0$ in $\mathbb{R}^{2} \backslash\{\zeta\}$ for any $\zeta$;
(ii) for every bounded and continuous function $\varphi=\varphi(x)$ and $\tau \in \mathbb{R}$, a classical solution to the Cauchy problem

$$
\begin{cases}L u(x, t)=0, & x \in \mathbb{R}, t>\tau  \tag{2.2}\\ u(x, \tau)=\varphi(x), & x \in \mathbb{R}\end{cases}
$$

is given by

$$
\begin{equation*}
u(x, t)=\int_{\mathbb{R}} \Gamma(x, t ; \xi, \tau) \varphi(\xi) d \xi . \tag{2.3}
\end{equation*}
$$

Under the assumption that $L$ is uniformly parabolic (i.e., the coefficient $a$ is greater than a constant $a_{0}>0$ ) and has bounded and Hölder continuous coefficients, it is well known that a fundamental solution for $L$ exists: this theoretical result can be proved by the parametrix method. We now aim to investigate the potentiality of the parametrix as a numerical method.

Coming back to the financial interpretation, formula (2.3) gives the forward price at time to maturity $t-\tau$ of a European option with payoff $\varphi$. Suppose now that problem (2.2) cannot be solved explicitly. It is then inviting to find an approximation formula for (2.3) whose principal term is given by (or is at least similar to) the Black-Scholes formula. This is what the parametrix method allows us to do.
2.1. Forward parametrix. The classical forward parametrix method is based on two ideas. The first is to approximate $\Gamma(z ; \zeta)$ by the so-called parametrix defined by

$$
Z(z ; \zeta)=\Gamma_{\zeta}(z ; \zeta)
$$

where, for fixed $w \in \mathbb{R}^{2}$ and for $\bar{b}, \bar{c}$ arbitrarily fixed real constants, $\Gamma_{w}$ is the fundamental solution to the constant coefficient operator

$$
\begin{equation*}
L_{w} u(z):=a(w) \partial_{x x} u(z)+\bar{b} \partial_{x} u(z)+\bar{c} u(z)-\partial_{t} u(z) . \tag{2.4}
\end{equation*}
$$

Note that $L_{w}$ is a heat operator and the explicit expression of $\Gamma_{w}$ is known:

$$
\begin{equation*}
\Gamma_{w}(z ; \zeta)=\frac{1}{\sqrt{4 \pi a(w)(t-\tau)}} \exp \left(-\frac{(x-\xi)^{2}}{4 a(w)(t-\tau)}-\frac{\bar{b}}{2 a(w)}(x-\xi)-\left(\frac{\bar{b}^{2}}{4 a(w)}-\bar{c}\right)(t-\tau)\right) \tag{2.5}
\end{equation*}
$$

for $t>\tau$.
In the standard parametrix method (cf., for instance, [17]) the constants $\bar{b}$ and $\bar{c}$ are chosen to be null. However, the context of this paper suggests using as a parametrix the fundamental solution of the heat equation which is "most similar" to the equation under analysis. This flexibility in the choice of the operator may result in considerably sharp approximations; see, for instance, section 4, where the some local volatility models are examined.

The second idea is that of supposing that the fundamental solution $\Gamma$ of $L$ is in the form

$$
\begin{equation*}
\Gamma(z ; \zeta)=Z(z ; \zeta)+\int_{\tau}^{t} \int_{\mathbb{R}} Z(z ; w) \Phi(w ; \zeta) d w . \tag{2.6}
\end{equation*}
$$

In view of the financial applications, in what follows we assume $\zeta=(\xi, 0)$. In order to identify $\Phi$ in (2.6), we notice that, since

$$
L \Gamma(\cdot ; \zeta)=0 \quad \text { in } \mathbb{R} \times] 0,+\infty[
$$

for any $\zeta$, we get

$$
\begin{equation*}
0=L Z(z ; \zeta)+L \int_{0}^{t} \int_{\mathbb{R}} Z(z ; w) \Phi(w ; \zeta) d w \tag{2.7}
\end{equation*}
$$

But formally we have

$$
\begin{align*}
L \int_{0}^{t} \int_{\mathbb{R}} Z(z ; w) \Phi(w ; \zeta) d w & =\int_{0}^{t} \int_{\mathbb{R}} L Z(z ; w) \Phi(w ; \zeta) d w-\partial_{t} \int_{0}^{t} \int_{\mathbb{R}} Z(z ; w) \Phi(w ; \zeta) d w \\
& =\int_{0}^{t} \int_{\mathbb{R}} L Z(z ; w) \Phi(w ; \zeta) d w-\Phi(z ; \zeta) \tag{2.8}
\end{align*}
$$

so that

$$
\begin{equation*}
\Phi(z ; \zeta)=L Z(z ; \zeta)+\int_{0}^{t} \int_{\mathbb{R}} L Z(z ; w) \Phi(w ; \zeta) d w \tag{2.9}
\end{equation*}
$$

Notation 2.1. To avoid confusion, when necessary, we write $L^{(z)}$ instead of $L$ in order to indicate that the operator $L$ is acting in the variable $z$.

Formula (2.9) can be solved iteratively and yields

$$
\begin{equation*}
\Gamma(z ; \zeta)=\sum_{n=0}^{+\infty} Z_{n}(z ; \zeta) \tag{2.10}
\end{equation*}
$$

with $Z_{0}(z ; \zeta)=Z(z ; \zeta)$ and

$$
Z_{n}(z ; \zeta)=\int_{0}^{t} \int_{\mathbb{R}} Z(z ; w)(L Z)_{n}(w ; \zeta) d w, \quad n \in \mathbb{N},
$$

where, recalling the notation $w=(y, s)$,

$$
\begin{aligned}
(L Z)_{1}(w ; \zeta) & =L^{(w)} Z(w ; \zeta), \\
(L Z)_{n+1}(w ; \zeta) & =\int_{0}^{s} \int_{\mathbb{R}} L^{(w)} Z\left(w ; z_{0}\right)(L Z)_{n}\left(z_{0} ; \zeta\right) d z_{0}, \quad n \in \mathbb{N} .
\end{aligned}
$$

As we foretold, our first main result consists in the computation of explicit global error bounds of the parametrix approximation. These bounds, provided in Theorem 5.2, are of the following form: for any $T>0$ there exist two positive constants $C, M$ such that

$$
\begin{equation*}
\left|\Gamma(z ; \zeta)-\sum_{k=0}^{n} Z_{k}(z ; \zeta)\right| \leq C \frac{(t-\tau)^{\frac{n}{2}}}{\left(\frac{n}{2}\right)!} \Gamma_{M}(z ; \zeta), \quad n \geq 0 \tag{2.11}
\end{equation*}
$$

for any $z, \zeta \in \mathbb{R}^{2}$ such that $0<t-\tau<T$, where $\Gamma_{M}$ is the fundamental solution of the heat operator

$$
M \partial_{x x}-\partial_{t} .
$$

Moreover, the constants $C$ and $M$ can be explicitly estimated.
2.2. Backward parametrix. Before examining the financial interpretation, we introduce what we call the backward parametrix, which is based on the use of the adjoint operator of $L$. As we shall see shortly, the backward parametrix is more convenient than the forward parametrix from several points of view: first, it allows us to derive an approximating expansion whose first term is given exactly by the Black-Scholes formula, while the subsequent terms can be expressed as solutions to suitable Cauchy problems related to constant coefficient operators that have a clear financial interpretation as well. Second, the approximating terms generated in this way are convolutions of a Gaussian function, and this is convenient from a numerical point of view since we may rely upon several known efficient numerical techniques.

Remark 2.2. The backward parametrix method does not simply consist in the standard parametrix method applied to the backward PDE: indeed, that would give the same problems as in the forward case. On the contrary, the idea is to use the backward parametrix as an approximation for the forward PDE.

The formal adjoint operator of $L$ in (2.1), acting in the variable $\zeta$, is defined as

$$
\begin{equation*}
\widetilde{L}^{(\zeta)}=a(\zeta) \partial_{\xi \xi}+\widetilde{b}(\zeta) \partial_{\xi}+\widetilde{c}(\zeta)+\partial_{\tau}, \tag{2.12}
\end{equation*}
$$

where

$$
\widetilde{b}(\zeta)=-b(\zeta)+2 \partial_{\xi} a(\zeta), \quad \widetilde{c}(\zeta)=c(\zeta)+\partial_{\xi \xi} a(\zeta)-\partial_{\xi} b(\zeta) .
$$

It is known that, under suitable assumptions, $\widetilde{L}$ has a fundamental solution $\widetilde{\Gamma}$ and the following duality formula holds:

$$
\widetilde{\Gamma}(\zeta ; z)=\Gamma(z ; \zeta) ;
$$

in particular, $\widetilde{L}^{(\zeta)} \Gamma(z ; \zeta)=0$ for $z \neq \zeta$. We define the backward parametrix as the fundamental solution of a constant coefficient dual operator: more precisely, for $w \in \mathbb{R}^{2}$ we set

$$
\begin{equation*}
\widetilde{L}_{w}^{(\zeta)}=a(w) \partial_{\xi \xi}+\bar{b} \partial_{\xi}+\bar{c}+\partial_{\tau}, \tag{2.13}
\end{equation*}
$$

where $\bar{b}$ and $\bar{c}$ are arbitrarily fixed constants, and consider its fundamental solution

$$
\begin{equation*}
\widetilde{\Gamma}_{w}(\zeta ; z)=\frac{1}{\sqrt{4 \pi a(w)(t-\tau)}} \exp \left(-\frac{(x-\xi)^{2}}{4 a(w)(t-\tau)}+\frac{\bar{b}}{2 a(w)}(x-\xi)-\left(\frac{\bar{b}^{2}}{4 a(w)}-\bar{c}\right)(t-\tau)\right) \tag{2.14}
\end{equation*}
$$

for $t>\tau$. Then we define the backward parametrix as

$$
P(z ; \zeta)=\widetilde{\Gamma}_{z}(\zeta ; z)
$$

so that in particular we have

$$
\begin{equation*}
\widetilde{L}_{z}^{(\zeta)} P(z ; \zeta)=0 \tag{2.15}
\end{equation*}
$$

As before, we set $\tau=0$ so that $\zeta=(\xi, 0)$, and, proceeding as in the forward case, we have

$$
\begin{equation*}
\Gamma(z ; \zeta)=\widetilde{\Gamma}(\zeta ; z)=P(z ; \zeta)+\sum_{n=1}^{+\infty} \int_{0}^{t} \int_{\mathbb{R}} P(w ; \zeta)(\widetilde{L} P)_{n}(z ; w) d w \tag{2.16}
\end{equation*}
$$

where, recalling the notation $w=(y, s)$,

$$
\begin{aligned}
(\widetilde{L} P)_{1}(z ; w) & =\widetilde{L}^{(w)} P(z ; w) \\
(\widetilde{L} P)_{n+1}(z ; w) & =\int_{s}^{t} \int_{\mathbb{R}} \widetilde{L}^{(w)} P\left(z_{0} ; w\right)(L Z)_{n}\left(z ; z_{0}\right) d z_{0}, \quad n \geq 1
\end{aligned}
$$

In Theorem 5.7 we give explicit global error estimates, completely analogous to (2.11), for the backward approximation truncated at the $n$th term.
2.3. Parametrix expansions. In section 5 , Theorems 5.4 and 5.7 , we prove that the solution to the Cauchy problem (2.2) has "forward and backward" expansions of the following form.

- The expansion obtained using the forward parametrix is given by

$$
\begin{equation*}
u(z)=\sum_{n=0}^{\infty} u_{n}(z), \tag{2.17}
\end{equation*}
$$

where

$$
\begin{equation*}
u_{0}(z)=\int_{\mathbb{R}} Z(z ; \xi, 0) \varphi(\xi) d \xi \tag{2.18}
\end{equation*}
$$

and, in general, for $n \in \mathbb{N}$,

$$
\begin{equation*}
u_{n}(z)=\int_{0}^{t} \int_{\mathbb{R}} Z(z ; \zeta) L U_{n-1}(\zeta) d \zeta, \quad U_{n-1}(z):=\sum_{k=0}^{n-1} u_{k}(z) . \tag{2.19}
\end{equation*}
$$

- Similarly the expansion obtained using the backward parametrix is of the form (2.17), where now

$$
\begin{equation*}
\widetilde{u}_{0}(z)=\int_{\mathbb{R}} P(z ; \xi, 0) \varphi(\xi) d \xi, \tag{2.2}
\end{equation*}
$$

and, in general, for $n \in \mathbb{N}$,

$$
\begin{equation*}
\widetilde{u}_{n}(z)=\int_{0}^{t} \int_{\mathbb{R}} P(z ; \zeta) L \widetilde{U}_{n-1}(\zeta) d \zeta, \quad \widetilde{U}_{n-1}(z):=\sum_{k=0}^{n-1} u_{k}(z), \quad n \in \mathbb{N} \tag{2.21}
\end{equation*}
$$

The main differences between the two parametrix expansions hinted at before are now clear. Each term in the expansion is an "expected value" with respect to the distributions with density $Z(z ; \zeta)$ or $P(z ; \zeta)$. But, while $P(z ; \zeta)$ is the same Gaussian density for each value of the integration variable ( $z$ is frozen and the integration is performed varying $\zeta$ ) and so is a true PDF, $Z(z ; \zeta)$ is a different Gaussian (different variance) for each value of the integration variable $\zeta$.

More precisely, let us examine the first term of the expansion for the forward parametrix $Z$ :

$$
\begin{equation*}
u_{0}(z)=\int_{\mathbb{R}} Z(z ; \xi, 0) \varphi(\xi) d \xi \tag{2.22}
\end{equation*}
$$

Since the explicit expression of $Z(z ; \xi, 0)=\Gamma_{(\xi, 0)}(z ; \xi, 0)$ is known, we see that $u_{0}$ in (2.22) is very similar to the solution of a Cauchy problem for a constant coefficient operator. On the other hand, the integration in (2.22) is performed with respect to the variable $\xi$ which also appears in $L_{(\xi, 0)}$ as the point where the operator $L$ is frozen. Hence, roughly speaking, the first term of the expansion is an "expected value" of the terminal payoff which uses as density a Gaussian with a different volatility (corresponding to the "true" diffusion coefficient) for each point in the integration range. This seems a quite sensible starting point and can obviously be compared with standard "implied volatility" approximations where a different Gaussian distribution (for $\log S$ ) for each strike is used. Here the suggestion is to use the same distribution but with a different volatility for each terminal value of the stock.

Let us now pass to the backward parametrix expansion zero order term:

$$
\widetilde{u}_{0}(z)=\int_{\mathbb{R}^{N}} P(z ; \xi, 0) \varphi(\xi) d \xi
$$

Here the interpretation is straightforward: the zero order term is simply a Black-Scholes option price. Indeed, since

$$
P(z ; \zeta)=\widetilde{\Gamma}_{z}(\zeta ; z),
$$

the parametrix $P(z ; \zeta)$ is the terminal log-price density corresponding to starting point $(x, t)$. Notice, however, that a different "volatility" value is used for each initial pair $(x, t)$. Accordingly, if we compute the derivative of the option with respect to the price $S=e^{x}$, we have that the Delta for the zero order approximation is given, with the obvious notation, by

$$
\Delta=\Delta_{\mathrm{BS}}+\operatorname{veg} a * \frac{\partial \sigma}{\partial S}
$$

This Delta computation derived from the parametrix expansion is interesting because it is a direct reinterpretation of one of the ad hoc modifications of the Black-Scholes model used by practitioners in order to get the "correct" answer from the "wrong" model. In fact, this is an example of a "skew correction" for the computation of Delta.

This correction, which can take various shapes (see, e.g., [9]), tries to account for the change of implied volatility which may accompany the change in moneyness of a given option. In a particular specification, often termed "sticky delta," if $\Delta_{\mathrm{BS}}$ is the Black-Scholes delta, vega is the standard Black-Scholes vega, and $\sigma(K / S)_{S}$ is the derivative, with respect to the price, of the volatility used for evaluating the option with strike $K$ with moneyness $K / S$, then we have a skew corrected delta computed as

$$
\Delta=\Delta_{\mathrm{BS}}+\operatorname{vega} * \sigma(K / S)_{S}
$$

Strictly speaking, this correction is inconsistent as it implies different risk neutral distributions for different strikes at the same date. However, if we read this correction in the framework of the parametrix, we can justify it consistently as an approximation of the "true" delta by a truncated expansion.

Next we consider the subsequent terms in the expansions. Both expansions are similar in that each new term can be interpreted approximately as an expected value. The difference that makes the backward parametrix more readable is that each term in the backward expansion is a true expected value (with respect to the same Gaussian function $P(z ; \zeta)$ ), while in the case of the standard parametrix, $Z(z ; \xi, 0)$ does not correspond to a real density.

Since each new term can be read as the value (exact or approximate) of a new option in a Black-Scholes world, it is interesting to understand the meaning of such options. To this end, it suffices to recall (2.19) and (2.21) and note that the operator $L$ in the term of order $n$ acts on the "option approximation" derived up to order $n-1$. Therefore, each action of the $L$ operator can be interpreted as a check of the fact that the approximation of order $n-1$ satisfies

$$
\begin{equation*}
L \widetilde{U}_{n-1}=0 \tag{2.23}
\end{equation*}
$$

In other words $L \widetilde{U}_{n-1}$ is a measure of the error implied in supposing that $\widetilde{U}_{n-1}$ satisfies (2.23). This error term is known (as it depends on the $n-1$ approximation) and if added to the original PDE as an inhomogeneous term makes the $n-1$ approximation exact. In the classic literature concerning parabolic equations and, in particular, the heat equation, such terms model the existence of additional heat sources (or sinks). In financial theory similar terms may arise as the result of transaction costs: we refer, for instance, to the asymptotic expansion setting for optimal hedging under transaction costs in [34].

We see how the parametrix expansion partitions the value of a given option computed in a non-Black-Scholes world into a series of option values each computed in the Black-Scholes world. This is exact in the case of the backward parametrix and approximately exact, if we recall that $Z$ is not a density, in the classical forward parametrix case. In section 5 we prove how it is possible to bound the overall error derived by truncating the series at the $n$th term with explicit and easily computable bounds uniformly decreasing as $n$ goes to infinity and as time to maturity decreases.
2.4. Computing the second term in the backward parametrix expansion. Formula (2.14) gives the first term of the backward approximation of the fundamental solution of $L$ in (2.1). We now illustrate a method for approximating the second term of the expansion (2.16). Recalling that $\zeta=(\xi, 0)$, we have

$$
P_{1}(z ; \zeta):=\int_{0}^{t} \int_{\mathbb{R}} P(w ; \zeta) \widetilde{L}^{(w)} P(z ; w) d w .
$$

It turns out that a convenient choice of the coefficients $\bar{b}, \bar{c}$ in (2.13) is

$$
\begin{equation*}
\bar{b}=-b(z), \quad \bar{c}=c(z), \tag{2.24}
\end{equation*}
$$

and therefore we set

$$
\begin{equation*}
\widetilde{L}_{z}^{(\zeta)}=a(z) \partial_{\xi \xi}-b(z) \partial_{\xi}+c(z)+\partial_{\tau} . \tag{2.25}
\end{equation*}
$$

Note that, by (2.24), $\widetilde{L}_{z}^{(\zeta)}$ is the adjoint of the frozen (forward) operator

$$
\begin{equation*}
L_{z}^{(\zeta)}=a(z) \partial_{\xi \xi}+b(z) \partial_{\xi}+c(z)-\partial_{\tau} . \tag{2.26}
\end{equation*}
$$

This fact will be used in section 3 . We also performed numerical tests that show that, for a local volatility model, the choice (2.24) is indeed optimal in the sense that it minimizes the pricing error among all other possible choices of $\bar{b}, \bar{c}$.

Recalling the notation $w=(y, s)$ and setting

$$
\begin{equation*}
I(s)=\int_{\mathbb{R}} P(y, s ; \zeta) \widetilde{L}^{(y, s)} P(z ; y, s) d y \tag{2.27}
\end{equation*}
$$

the idea is to use the trapezoidal method to approximate

$$
P_{1}(z ; \zeta)=\int_{0}^{t} I(s) d s \simeq \frac{t}{2}(I(0)+I(t)) .
$$

This allows us to exploit the fact that

$$
I(0)=\widetilde{L}^{(\zeta)} P(z ; \zeta)
$$

since $P(y, 0 ; \zeta)$ is a Dirac delta centered at $\zeta$. Note that this approximation avoids the computation of the spatial integral $I(s)$ in $(2.27)$ for any $s$ : this results in a significant simplification especially for high dimensional models.

By (2.15) we have

$$
\widetilde{L}^{(y, s)} P(z ; y, s)=\left(\widetilde{L}^{(y, s)}-\widetilde{L}_{z}^{(y, s)}\right) P(z ; y, s)
$$

for $t>s$; then

$$
I(s)=\int_{\mathbb{R}} P(w ; \zeta)\left(\widetilde{L}^{(w)}-\widetilde{L}_{z}^{(w)}\right) P(z ; w) d y
$$

(by parts, for $L_{z}^{(w)}$ as in (2.26))

$$
=\int_{\mathbb{R}} P(z ; w)\left(L^{(w)}-L_{z}^{(w)}\right) P(w ; \zeta) d y
$$

and, passing to the limit as $s$ goes to $t$, thanks to the choice (2.24), we obtain

$$
I(t)=0 .
$$

In conclusion, we have the following explicit formula for the backward parametrix approximation with two terms:

$$
\begin{equation*}
\Gamma(z ; \zeta) \simeq P(z ; \zeta)+P_{1}(z ; \zeta) \simeq P(z ; \zeta)+\frac{t-\tau}{2} \widetilde{L}^{(\zeta)} P(z ; \zeta) \tag{2.28}
\end{equation*}
$$

By using this technique, it is not difficult to determine explicit expressions for higher order approximations. However, we do not report them here since the preliminary experiments we performed in local volatility models (cf. section 4) show a negligible contribution of the terms of order higher than two.
3. Analytic formulae in local volatility models. By using the parametrix method, in this section we derive analytic (closed form) approximation formulae for one-dimensional local volatility models. This result seems significant on its own; however, we would like to emphasize that analogous results are valid for general local or stochastic volatility models even in high dimensions and possibly in a degenerate setting (i.e., for instance, for Asian options): we refer the reader to the forthcoming paper [12] for more results in this direction.

Let us consider a local volatility model where the dynamic of the underlying asset is given by the SDE

$$
\begin{equation*}
d S_{t}=\mu\left(S_{t}, t\right) S_{t} d t+\sigma\left(S_{t}, t\right) S_{t} d W_{t} \tag{3.1}
\end{equation*}
$$

where $W$ is a one-dimensional Brownian motion and $\mu, \sigma$ are sufficiently regular coefficients. Assuming a constant riskless interest rate $r$, the price $V(S, t)$ of a European call option with strike $K$ and maturity $T$ is the solution to the Cauchy problem

$$
\begin{cases}\frac{\sigma^{2}(S, t) S^{2}}{2} \partial_{S S} V(S, t)+r S \partial_{S} V(S, t)-r V(S, t)+\partial_{t} V(S, t)=0, & S>0, t \in] 0, T[,  \tag{3.2}\\ V(S, T)=(S-K)^{+}, & S>0 .\end{cases}
$$

By the standard change of variables

$$
\begin{equation*}
V(S, t)=e^{-r(T-t)} u(r(T-t)+\log S, T-t) \tag{3.3}
\end{equation*}
$$

we have that $V$ solves (3.2) if and only if $u$ solves

$$
\begin{cases}a(x, t)\left(\partial_{x x} u(x, t)-\partial_{x} u(x, t)\right)-\partial_{t} u(x, t)=0, & x \in \mathbb{R}, t \in] 0, T[,  \tag{3.4}\\ u(x, 0)=\left(e^{x}-K\right)^{+}, & x \in \mathbb{R},\end{cases}
$$

where

$$
\begin{equation*}
a(x, t)=\frac{1}{2} \sigma^{2}\left(e^{x-r t}, T-t\right) . \tag{3.5}
\end{equation*}
$$

In particular, the option price at $t=0$ is given by

$$
\begin{equation*}
V(S, 0)=e^{-r T} u(r T+\log S, T), \tag{3.6}
\end{equation*}
$$

where

$$
\begin{equation*}
u(x, T)=\int_{\mathbb{R}}\left(e^{\xi}-K\right)^{+} \Gamma(x, T ; \xi, 0) d \xi \tag{3.7}
\end{equation*}
$$

and $\Gamma$ is the fundamental solution of the PDE in (3.4).
Proceeding as in subsection 2.4, we consider the frozen operator corresponding to (2.25)

$$
\begin{equation*}
\widetilde{L}_{z}^{(\zeta)}=a(z)\left(\partial_{\xi \xi}+\partial_{\xi}\right)+\partial_{\tau} . \tag{3.8}
\end{equation*}
$$

Then the related backward parametrix is given by

$$
\begin{equation*}
P(x, t ; \xi, \tau)=\frac{1}{\sqrt{4 \pi a(x, t)(t-\tau)}} \exp \left(-\frac{(x-\xi)^{2}}{4 a(x, t)(t-\tau)}+\frac{1}{2}(x-\xi)-\frac{1}{4} a(x, t)(t-\tau)\right) \tag{3.9}
\end{equation*}
$$

for $t>\tau$.
With this choice the first term in the backward parametrix expansion (2.20)-(2.21) corresponds exactly to the Black-Scholes price computed with (constant) volatility:

$$
\begin{equation*}
\widetilde{u}_{0}(x, t)=\int_{\mathbb{R}}\left(e^{\xi}-K\right)^{+} P(x, t ; \xi, 0) d \xi=e^{x} \Phi\left(d_{+}\right)-K \Phi\left(d_{-}\right), \tag{3.10}
\end{equation*}
$$

where

$$
d_{ \pm}=\frac{x-\log K \pm a(x, t) t}{\sqrt{4 a(x, t) t}}
$$

and

$$
\begin{equation*}
\Phi(x)=\frac{1}{\sqrt{\pi}} \int_{-\infty}^{x} e^{-y^{2}} d y \tag{3.11}
\end{equation*}
$$

Next we derive the explicit expression of the backward expansion with two terms.
Theorem 3.1. The second order parametrix approximation of the call price in (3.6) in the local volatility model (3.1) is given by

$$
\begin{equation*}
u(x, t) \simeq \widetilde{u}_{0}(x, t)+\frac{t K}{2}(a(\log K, 0)-a(x, t)) P(x, t ; \log K, 0), \tag{3.12}
\end{equation*}
$$

where $a$ is as in (3.5) and $\widetilde{u}_{0}$ is as defined in (3.10).

Proof. By (2.28) we have

$$
\Gamma(x, t ; \xi, \tau) \simeq P(x, t ; \xi, \tau)+\frac{t-\tau}{2} \widetilde{L}^{(\zeta)} P(x, t ; \xi, \tau)
$$

where $\widetilde{L}^{(\zeta)}$ is the adjoint operator of $L$, acting in the variable $\zeta=(\xi, \tau)$, and $P$ is the backward parametrix in (3.9). We first remark that

$$
\begin{equation*}
\left(L^{(\zeta)}-L_{z}^{(\zeta)}\right)\left(e^{\xi}-K\right)=(a(\zeta)-a(z))\left(\partial_{\xi \xi}-\partial_{\xi}\right)\left(e^{\xi}-K\right)=0 \tag{3.13}
\end{equation*}
$$

Then, recalling the notation $z=(x, t)$, the second term in the parametrix expansion (2.21) is

$$
u_{1}(x, t)=\frac{t}{2} \int_{\mathbb{R}}\left(e^{\xi}-K\right)^{+} \widetilde{L}^{(\zeta)} P(x, t ; \xi, 0) d \xi
$$

(by (2.15))

$$
=\frac{t}{2} \int_{\log K}^{+\infty}\left(e^{\xi}-K\right)\left(\widetilde{L}^{(\zeta)}-\widetilde{L}_{z}^{(\zeta)}\right) P(x, t ; \xi, 0) d \xi
$$

(integrating by parts)

$$
\begin{aligned}
= & \frac{t}{2} \int_{\log K}^{+\infty}\left(L^{(\zeta)}-L_{z}^{(\zeta)}\right)\left(e^{\xi}-K\right) P(x, t ; \xi, 0) d \xi \\
& -\frac{t}{2}\left[P(z, \zeta) \partial_{\xi}\left((a(\zeta)-a(z))\left(e^{\xi}-K\right)\right)\right]_{\xi=\log K}^{\xi=+\infty}
\end{aligned}
$$

(by (3.13))

$$
\begin{aligned}
& =-\frac{t}{2}\left[P(z, \zeta)\left(e^{\xi}(a(\zeta)-a(z))+\partial_{\xi} a(\zeta)\left(e^{\xi}-K\right)\right)\right]_{\xi=\log K}^{\xi=+\infty} \\
& =\frac{t K}{2}(a(\log K, 0)-a(x, t)) P(x, t ; \log K, 0) .
\end{aligned}
$$

4. Numerical experiments. Although the results of this paper, i.e., the introduction of the backward parametrix and the estimation of error bounds, are mainly theoretical, in this section we aim to present some numerical experiment which should convince the reader of the effectiveness of the parametrix method. While more complicated models could have been considered, here we aim only to present some preliminary tests and refer the reader to a forthcoming paper for a more detailed and extensive analysis of the numerical efficiency of the parametrix method for computing option prices and the related sensitivities or Greeks.

In this section we consider the parametrix approximation with only two terms and evaluate its performance by comparing it with several other numerical and analytical approximations in some local volatility models. Specifically we consider three classes of models:

- the constant elasticity of variance (CEV) model by Cox and Ross [8];
- local volatility (LV) models of quadratic and hyperbolic form (cf., for instance, Iacus [23], Jäckel [24], and Kahl and Jäckel [25]); and
- the path-dependent (two-dimensional) Hobson-Rogers model [20].
4.1. CEV model. We first consider a particular one-dimensional local volatility model, the well-known CEV model, where the dynamics of the underlying asset, under the risk neutral measure, is given by

$$
\begin{equation*}
d S_{t}=r S_{t} d t+\sigma S_{t}^{1-\alpha} d W_{t} \tag{4.1}
\end{equation*}
$$

Here $r$ and $\sigma$ are constants and $\alpha \in[0,1]$. Then (4.1) corresponds to (3.1) with $\mu(S, t) \equiv r$ and $\sigma(S, t)=\sigma S^{-\alpha}$ : in this particular case we have

$$
\begin{equation*}
a(x, t)=\frac{\sigma^{2} e^{-2 \alpha(x-r t)}}{2} \tag{4.2}
\end{equation*}
$$

in the approximation formula (3.12).
We opted for this model since there is a vast literature concerning the approximation of CEV option prices so that we may compare the parametrix performance with several other techniques. Moreover, very accurate formulae are available, and therefore we have reference numbers for an (almost) exact comparison with our approximation. Let us remark explicitly that the CEV pricing PDE is not uniformly parabolic, and, in particular, it does not satisfy the nondegeneracy condition (5.3); nevertheless, at least formally, the parametrix method applies, and we shall see that, as a matter of fact, for a wide range of values of the parameters, it provides very accurate approximations.

We compare the parametrix with six different approximation techniques: note that some of these techniques were introduced specifically for the CEV model, while the parametrix is a quite general method.

We distinguish numerical from closed form approximations. In the first group we consider - Cox [7] (see also Hsu, Lin, and Lee [22]; Cox and Ross [8]; and Schroder [30]);

- Shaw [31] (cf. Chapter 28); and
- Monte Carlo (MC).

In the second group, that of analytic approximations, we consider

- Hagan and Woodward [19] (see also Obloj [29]);
- Howison [21]; and
- Svoboda-Greenwood [32].

We aim to compare the performance of the parametrix with respect to the above methods in the pricing of European call options for different values of the parameters $\alpha, \sigma, T$, and $K$ : typically we consider $\alpha$ ranging from $\frac{1}{4}$ to $\frac{3}{4}$ and the maturity $T$ from one week to one year. We also consider different values of the strike price $K$, from 1 to 100 .

Remark 4.1. In the CEV model, the strike $K$ and the volatility coefficient $\sigma$ are inversely correlated. Indeed, by the transformation $Y_{t}=\frac{S_{t}}{K}$, from (4.1) we get

$$
d Y_{t}=r Y_{t} d t+\frac{\sigma}{K^{\alpha}} Y_{t}^{1-\alpha} d W_{t}
$$

which shows that $K$ and $\sigma$ are inversely proportional quantities: increasing the value of the strike corresponds to decreasing the value of the volatility.

Regarding the numerical approximations, we recall that the formulae by Cox [7] express the price of a call option as the sum of a series of Gamma cumulative distribution functions.


Figure 1. CEV-expansion option price by Cox [7] in the case $\alpha=\frac{1}{4}$ and $T=\frac{1}{3}$ with a number $n$ of terms in the series expansion equal to $n=400,420,440,460$.

It is known that these formulae give a good local (at-the-money) approximation of the option price. For instance, Figure 1 shows the Cox option prices in the case $\alpha=\frac{1}{4}$ and $T=\frac{1}{3}$ with a number $n$ of terms in the series expansion equal to $n=400,420,440,460$ : it is evident that for far-from-the-money options this approximation gives wrong prices unless we consider a high number of terms in the series expansion. This is particularly sensible for short times to maturity.

On the other hand, the approximation by Shaw [31] expresses the payoff random variable in terms of Bessel functions and then uses numerical integration to provide the option price. Since it is an adaptive method, the representation of prices is valid globally even if the method may become computationally expensive when we have to compute deep out-of- or in-themoney option prices. We implemented both methods and found that in most cases they are essentially equivalent and provide reliable reference prices. However, as we shall see shortly, the approximation by Hagan and Woodward [19] also seems to be very accurate for all the values of the parameters: since this last approximation gives closed form solutions, we decided to use it to produce the reference values for the computation of the errors.

Concerning Monte Carlo (MC), it seems that it is not competitive with any of the other approximations. In particular, the parametrix and the other analytic approximations we considered, when suitably used, are generally much more accurate than MC solutions, and, what is more, they give explicit algebraic formulae for option prices. For instance, in Figure 2 we represent the Hagan-Woodward price, obtained by (4.4), minus the MC price (continuous line) and the parametrix price minus the MC price (dashed line) scaled to the at-the-money price in the CEV model with $\alpha=\frac{1}{4}$ (left) and $\alpha=\frac{1}{2}$ (right). In the experiments, 500.000 MC simulations, combined with an Euler discretization of the SDE with 100 steps, are performed. The values of the other parameters are $T=0.25, \sigma=30 \%, r=5 \%$, and $K=1$. We marked by green (respectively, red) the confidence regions where prices differ no more than 2 (respectively, 2.57) standard deviations from the simulated MC prices: this means that with probability $95 \%$ (respectively, $99 \%$ ) the true price (scaled to the at-the-money price) belongs to the green (respectively, red) region. According to the standard interpretation of a


Figure 2. Hagan-Woodward minus MC price (continuous line) and parametrix minus MC price (dashed line) scaled to the at-the-money price in the CEV model with $\alpha=\frac{1}{4}$ (left) and $\alpha=\frac{1}{2}$ (right). A 100-step Euler discretization of the SDE and an MC with $N=500.000$ simulations have been used. Moreover, $T=0.25$, $\sigma=20 \%, r=5 \%$, and $K=1$. The $95 \%$ and $99 \%$ confidence regions for MC estimates are marked, respectively, by green and red.
confidence interval, any price inside the bands can be the true price and is compatible (at the given level of confidence) with the MC estimate.

Next we consider the analytic approximations. We recall that Hagan and Woodward in [19], by using singular perturbation techniques, obtain explicit formulae for the approximated implied volatility $\sigma_{B}$ in a local volatility model where the forward price $F_{t}$ of the asset obeys an SDE of the form

$$
\begin{equation*}
d F_{t}=\gamma(t) A\left(F_{t}\right) d W_{t} \tag{4.3}
\end{equation*}
$$

for some deterministic and suitably regular functions $\gamma$ and $A$. Equation (4.1) can be reduced to (4.1) through the transformation $F_{t}=e^{r(T-t)} S_{t}$ : in this case we also have

$$
\gamma(t)=\sigma e^{r \alpha(T-t)} \quad \text { and } \quad A(F)=F^{1-\alpha} .
$$

Therefore, the Hagan-Woodward approximation formula reads

$$
\begin{equation*}
\sigma_{B}=\frac{a}{f^{\alpha}}\left(1+\frac{1}{24} \alpha(3-\alpha)\left(\frac{e^{r T} S_{0}-K}{f}\right)^{2}+\frac{1}{24} \frac{\alpha^{2} a^{2} T}{f^{2 \alpha}}\right), \tag{4.4}
\end{equation*}
$$

where

$$
a=\sqrt{\frac{1}{T} \int_{0}^{T} \gamma(t)^{2} d t}=\sigma \sqrt{\frac{e^{2 r \alpha T}-1}{2 r \alpha T}} \quad \text { and } \quad f=\frac{e^{r T} S_{0}+K}{2} .
$$

It is then sufficient to insert the implied volatility $\sigma_{B}$ into the Black-Scholes formula to get the option prices. Figure 3 shows the Hagan-Woodward call value minus the Cox call value, scaled with the at-the-money value, as a function of $S$ for different times to maturity and values of $\alpha$. Since the Hagan-Woodward approximation seems to be quite accurate, hereafter, as already mentioned, we shall use it for computing the "true" (or reference) values in the experiments.


Figure 3. Hagan-Woodward approximate value minus Cox approximate value (with $n=1000$ terms in the series expansion), scaled with the at-the-money value, as a function of $S$, for $T=2,6,12$ months and $\alpha=\frac{1}{4}$ (left), $\alpha=\frac{3}{4}$ (right), $\sigma=30 \%, r=5 \%$, and $K=1$.


Figure 4. Scaled errors of the parametrix approximation, as a function of $S \in[0,2]$, for $T=7,15,45$ days (left) and $T=4,8,12$ months (right). $\alpha=\frac{1}{4}, \sigma=30 \%, r=5 \%$, and $K=1$.


Figure 5. Scaled errors of the parametrix approximation as a function of $S \in[0.2,2]$ for $T=7,15,45$ days (left) and $T=4,8,12$ months (right). $\alpha=\frac{3}{4}, \sigma=30 \%, r=5 \%$, and $K=1$.

We next test the computational performance of the parametrix approximation. Figure 4 compares the error relative to the at-the-money option value (in short, the scaled error) in the case $\alpha=\frac{1}{4}$ for different times to maturity: $T=7,15,45$ days in the left panel and $T=4,8,12$ months in the right panel.

Figure 5 exhibits analogous results for $\alpha=\frac{3}{4}$ : in this case, we have bigger errors than in the case $\alpha=\frac{1}{4}$ which is "closer" to the standard Black-Scholes model. For $\alpha=\frac{3}{4}$, we also separate the case $S \in\left[\frac{2}{10}, 2\right]$ from the case $S \in\left[0, \frac{2}{10}\right]$, which we represent in Figure 6. The reason is that, as previously remarked, the CEV diffusion operator is not uniformly parabolic and


Figure 6. Scaled errors of the parametrix approximation as a function of $S \in[0,0.2]$ for $T=7,15,45$ days (left) and $T=4,8,12$ months (right). $\alpha=\frac{3}{4}, \sigma=30 \%, r=5 \%$, and $K=1$.



Figure 7. Relative errors as a function of $K \in[1,10]$ (left) and $K \in[10,100]$ (right) of the at-the-money prices in the parametrix approximation. Here $T=\frac{1}{2}, \sigma=30 \%, r=5 \%$, and $\alpha=\frac{1}{4}, \frac{1}{2}, \frac{3}{4}$.
degenerates at the boundary. In particular, in this case, the price process reaches the origin with positive probability and this can be difficultly captured by a nondegenerate diffusion. However, as reported in Figure 6, this fact seems to produce relevant errors only for extremely out-of-the-money options; that is, $S \sim \frac{1}{100}$ when $K=1$.

Finally, we examine the accuracy of the parametrix approximation as the strike $K$ varies or, equivalently, by Remark 4.1, the volatility $\sigma$ varies. Figure 7 shows the relative errors, defined as

$$
\frac{C^{H W}-C^{P}}{C^{H W}}
$$

where $C^{P}, C^{H W}$ are, respectively, the call prices given by the parametrix expansion and by the Hagan-Woodward formula. Here we consider at-the-money options prices as functions of $K \in[10,100]$. Moreover, $T=\frac{1}{2}, \sigma=30 \%, r=5 \%$, and the values of $\alpha=\frac{1}{4}, \frac{1}{2}, \frac{3}{4}$ are considered. Experiments with other choices of the parameters give essentially the same results and show that the parametrix prices basically coincide with the Hagan-Woodward prices for $K$ large (or $\sigma$ small): in this case the approximation for $\alpha=\frac{3}{4}$ is even better than it is for $\alpha=\frac{1}{4}$.

Next we consider the matched asymptotic expansion (MAE) proposed by Howison [21]. In this case we can directly employ the approximation formula in [21, page 392] to compute option prices. We also consider the recent paper by Svoboda-Greenwood [32], where another approximated formula for CEV prices is obtained by performing a small time expansion of


Figure 8. Scaled errors of parametrix (dashed line), MAE (thick line), and Svoboda-Greenwood (continuous line) as a function of $S$, with $\alpha=\frac{1}{4}$; at-the-money volatility is $\sigma=30 \%$ (left) and $\sigma=15 \%$ (right). Moreover, $T=0.25, r=5 \%$, and $K=1$.


Figure 9. Scaled errors of parametrix (dashed line), MAE (thick line), and Svoboda-Greenwood (continuous line) as a function of $S$, with $\alpha=\frac{1}{2}$; at-the-money volatility is $\sigma=30 \%$ (left) and $\sigma=15 \%$ (right). Moreover, $T=0.25, r=5 \%$, and $K=1$.
the option prices around the forward-at-the-money value of the underlying. More precisely, we consider the SDE

$$
d Y_{t}=\gamma(t) f\left(Y_{t}\right) d W_{t}
$$

which corresponds to (4.1) after the transformation $Y_{t}=e^{-r t} S_{t}$, with $\gamma(t)=\sigma e^{-r \alpha t}$ and $f(Y)=Y^{1-\alpha}$. Actually, in [32], only the case of time independent $\gamma=\gamma(t)$ is considered. However, by using the same technique we generalize the result in [32, section 2.2] and obtain the following formula for the approximated call option price with strike $K$ and maturity $T$ :

$$
\begin{aligned}
C(S, T) \sim & \left(S-y_{0}\right) \Phi\left(\frac{S-y_{0}}{y_{0}^{1-\alpha} \sigma_{0}}\right)+\left(\left(1+\frac{\left(S-y_{0}\right)^{1-\alpha}}{2 y_{0}}\right) y_{0}^{1-\alpha} \sigma_{0}+\frac{(1-\alpha)^{2}}{4} \frac{\left(S-y_{0}\right)^{4}}{\sigma_{0} y_{0}^{3-\alpha}}\right. \\
& \left.+\frac{1}{6}\left(S-y_{0}\right)^{2} \sigma_{0} y_{0}^{-1-\alpha}(1+\alpha(2 \alpha-3))+\frac{1}{12} \sigma_{0}^{3} y_{0}^{1-3 \alpha}(1-\alpha)(9 \alpha-8)\right) \Phi^{\prime}\left(\frac{S-y_{0}}{y_{0}^{1-\alpha} \sigma_{0}}\right),
\end{aligned}
$$

where $y_{0}=K e^{-r t}, \sigma_{0}=\sigma \sqrt{\frac{1-e^{-2 r \alpha T}}{2 r \alpha}}$, and $\Phi$ is the standard normal cumulative density function in (3.11).

In Figures 8, 9, and 10 we compare the scaled errors of the parametrix (dashed line), MAE (thick line), and Svoboda-Greenwood (continuous line) approximations. It turns out


Figure 10. Scaled errors of parametrix (dashed line), MAE (thick line), and Svoboda-Greenwood (continuous line) as a function of $S$ with $\alpha=\frac{3}{4}$; at-the-money volatility is $\sigma=30 \%$ (left) and $\sigma=15 \%$ (right). Moreover, $T=0.25, r=5 \%$, and $K=1$.
that, at least for $\alpha \leq \frac{2}{3}$, the parametrix gives the best results: this is confirmed also by other experiments that we do not report here. For $\alpha=\frac{3}{4}$, parametrix and Svoboda-Greenwood errors are of the same order even if the latter seems slightly better. In general the MAE is not competitive with the other two approximations; this is particularly evident when $\sigma$ is small or $K$ is large. We also remark that the MAE is not correct asymptotically for large $S$ (see also Figure 3 in [21]).
4.2. Parabolic and hyperbolic LV models. We consider two specifications of the volatility function in the general local volatility (LV) model

$$
d S_{t}=r S_{t} d t+\sigma\left(S_{t}, t\right) S_{t} d W_{t}
$$

namely, the quadratic $L V$

$$
\sigma(S, t)=\sigma_{0} \min \left\{2, \sqrt{1+(S-\beta)^{2}}\right\}
$$

and the hyperbolic $L V$, as defined in [24],

$$
\sigma(S, t)=\sigma_{0}\left(\frac{1-\beta+\beta^{2}}{\beta} S+\frac{\beta-1}{\beta}\left(\sqrt{S^{2}+\beta^{2}(1-S)^{2}}-\beta\right)\right)
$$

where $\sigma_{0}$ and $\beta$ are suitable parameters.
We remark that, if $r$ is strictly positive, then both models cannot be reduced in the form (4.3) where the coefficient $A$ is independent of time. Consequently in this case the HaganWoodward and Svoboda-Greenwood approximations do not apply. Therefore, we compare the parametrix with an accurate MC method.

In Figure 11 the black line represents the parametrix call price minus an MC call price, scaled to the at-the-money price, in the quadratic LV model (left) and hyperbolic LV model (right). We set $T=0.25, r=5 \%$, and $K=1$. Moreover, in the quadratic LV, we put $\beta=1$ and $\sigma_{0}=20 \%$ so that in this model the volatility varies from the minimum $20 \%$ at-the-money to the maximum $40 \%$. In the hyperbolic LV, we consider the typical values (cf. [24]) $\sigma=20 \%$ and $\beta=\frac{1}{2}$.


Figure 11. The black line represents the difference, scaled to the at-the-money price, between the parametrix call price and an MC call price in the quadratic (left) and hyperbolic (right) local volatility models. The $95 \%$ and $99 \%$ MC confidence regions are marked, respectively, by green and red. A 100-step Euler discretization of the SDEs and an MC with 500.000 simulations have been used. Moreover, $T=0.25, r=5 \%$, and $K=1$.

In the experiments, a 100 -step Euler discretization of the SDEs and an MC with 500.000 simulations have been used. As before, we marked by green (respectively, red) the confidence regions where prices differ no more than 2 (respectively, 2.57) standard deviations from the simulated MC prices; in other words, the true price (scaled to at-the-money price) belongs to the green (red) region with probability $95 \%$ ( $99 \%$ ).

Since in Figure 11 some errors appear corresponding to in-the-money options, for a more comprehensive comparison, we report in Table 1 the MC and parametrix prices (not scaled) for $S \in\{1.2 ; 1.3 ; 1.4 ; 1.5 ; 1.6\}$.

Table 1
$M C$ and parametrix call prices in quadratic and hyperbolic LV models with $\sigma_{0}=20 \%, T=0.25, r=5 \%$, and $K=1$. Moreover, $\beta=1$ in the quadratic $L V$ and $\beta=\frac{1}{2}$ in the hyperbolic $L V$.

|  | Call prices |  |  |  |
| :--- | :--- | :--- | :--- | :--- |
|  | Quadratic LV |  | Hyperbolic LV |  |
|  | MC | Parametrix | MC | Parametrix |
| $S=1.2$ | 0.21352 | 0.213547 | 0.318564 | 0.318545 |
| $S=1.3$ | 0.312533 | 0.312535 | 0.417773 | 0.417748 |
| $S=1.4$ | 0.412432 | 0.412433 | 0.517571 | 0.517548 |
| $S=1.5$ | 0.512423 | 0.512423 | 0.61752 | 0.617503 |
| $S=1.6$ | 0.612422 | 0.612422 | 0.717507 | 0.717495 |

Clearly, in view of an extensive use of the parametrix approximation, a deeper analysis of the performance for a wide range of parameters is in order. However, these preliminary results seem promising.
4.3. Hobson-Rogers model. The Hobson-Rogers model [20] was introduced as an extension of the local volatility. In this model the volatility is defined as a function of the whole trajectory of the underlying asset and not only in terms of the spot price. The model was further generalized to a more flexible path dependent volatility model by two of the authors [13]. The main feature is that it generally leads to a complete market. We refer the reader to [14]
for an empirical analysis which shows the effectiveness of the model and compares the hedging performance with respect to standard stochastic volatility models.

We consider an average weight $\psi$ that is a nonnegative, piecewise continuous, and integrable function on $]-\infty, T]$. We assume that $\psi$ is strictly positive in $[0, T]$, and we set

$$
\Psi(t)=\int_{-\infty}^{t} \psi(s) d s
$$

Then we define the average process as

$$
\left.\left.A_{t}=\frac{1}{\Psi(t)} \int_{-\infty}^{t} \psi(s) Z_{s} d s, \quad t \in\right] 0, T\right]
$$

where $Z_{t}=\log \left(e^{-r t} S_{t}\right)$ denotes the log-discounted price process. The standard HobsonRogers model corresponds to the specification $\psi(t)=e^{\lambda t}$ for some positive parameter $\lambda$. Then by the Itô formula we have

$$
d A_{t}=\frac{\varphi(t)}{\Phi(t)}\left(Z_{t}-A_{t}\right) d t
$$

In a path dependent volatility model the $\log$-price $Z_{t}=\log S_{t}$ has the dynamics

$$
d Z_{t}=\mu\left(Z_{t}-A_{t}\right) d t+\nu\left(Z_{t}-A_{t}\right) d W_{t},
$$

where $\mu=\mu(\cdot)$ and $\nu=\nu(\cdot)$ are suitable functions. Then, by usual no-arbitrage arguments, we obtain the pricing operator

$$
\begin{equation*}
\left.L u=\frac{\nu^{2}(z-a)}{2}\left(\partial_{z z} u-\partial_{z} u\right)+\frac{\varphi(t)}{\Phi(t)}(z-a) \partial_{a} u+\partial_{t} u, \quad(t, z, a) \in\right] 0, T\left[\times \mathbb{R}^{2} .\right. \tag{4.5}
\end{equation*}
$$

We remark that $L$ is not uniformly parabolic (i.e., does not satisfy condition (5.3) in $\mathbb{R}^{2}$ ). However, if we assume that $\nu$ is smooth and bounded from above and below by positive constants, then $L$ is a hypoelliptic operator belonging to the general class of Kolmogorov operators for which the parametrix method has been successfully employed in [11] to construct a fundamental solution. Therefore, it is possible to extend all the theoretical results of this paper to include $L$ in (4.5).

As in the previous examples, we tested the parametrix against the MC method, and in the experiments a 100 -step Euler discretization of the SDEs and 500.000 simulations have been used. In Figure 12 we compare the parametrix and MC approximations for typical values of the parameters, assuming different values for the initial average $A_{0}$-specifically $A_{0}=Z_{0}$ in the left panel and $A_{0}=Z_{0}-\frac{1}{5}$ in the right panel. In Table 2 we also report some of the corresponding MC and parametrix prices.

The last experiment in Figure 13 is similar, but we consider different maturities, namely, 1 month in the left panel and 6 months in the right panel.


Figure 12. The black line represents the difference, scaled to the at-the-money price, between the parametrix call price and an MC call price in the Hobson-Rogers model for $S_{0} \in\left[\frac{7}{10}, \frac{13}{10}\right]$ and $\psi(t)=e^{t}, \nu(x)=\sigma \sqrt{2 x^{2}+1}$, $\sigma=20 \%, T=0.25, r=5 \%$, and $K=1$. The $95 \%$ and $99 \% M C$ confidence regions are marked, respectively, by green and red. A 100-step Euler discretization of the SDEs and an MC with 500.000 simulations have been used. Moreover, $A_{0}=Z_{0}$ (left) and $A_{0}=Z_{0}-\frac{1}{5}$ (right).

Table 2
MC and parametrix call prices in the Hobson-Rogers model with $\psi(t)=e^{t}, \nu(x)=\sigma \sqrt{2 x^{2}+1}, \sigma=20 \%$, $T=0.25, r=5 \%$, and $K=1$ in the case $A_{0}=Z_{0}$ (left) and $A_{0}=Z_{0}-\frac{1}{5}$ (right).

| Call prices |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- |
|  | $A_{0}=Z_{0}$ |  |  | $A_{0}=Z_{0}-\frac{1}{5}$ |
|  | MC | Parametrix | MC | Parametrix |
| $S=0.8$ | 0.000646705 | 0.000679473 | 0.000612704 | 0.000696547 |
| $S=0.9$ | 0.00820041 | 0.00819487 | 0.00904307 | 0.00926125 |
| $S=1$ | 0.0400307 | 0.0399139 | 0.0430224 | 0.0429977 |
| $S=1.1$ | 0.0998243 | 0.0998026 | 0.102566 | 0.102507 |
| $S=1.2$ | 0.168063 | 0.168108 | 0.16929 | 0.169278 |
| $S=1.3$ | 0.230935 | 0.230946 | 0.231299 | 0.231269 |



Figure 13. The black line represents the difference, scaled to the at-the-money price, between the parametrix call price and an MC call price in the Hobson-Rogers model for $S_{0} \in\left[\frac{8}{10}, \frac{12}{10}\right]$ and $\psi(t)=e^{t}, \nu(x)=\sigma \sqrt{2 x^{2}+1}$, $\sigma=10 \%, A_{0}=Z_{0}, r=5 \%$, and $K=1$. The $95 \%$ and $99 \% M C$ confidence regions are marked, respectively, by green and red. A 100-step Euler discretization of the SDEs and an MC with 500.000 simulations have been used. Moreover, $T=\frac{1}{12}$ (left) and $T=\frac{6}{12}$ (right).
5. Error bounds. In this section we present the parametrix expansion in its full generality and derive easily computable error estimates. We consider a parabolic differential equation in the form

$$
\begin{equation*}
L u:=\sum_{i, j=1}^{N} a_{i j}(z) \partial_{x_{i} x_{j}} u+\sum_{i=1}^{N} b_{i}(z) \partial_{x_{i}} u+c(z) u-\partial_{t} u=0, \quad z=(x, t) \in \mathbb{R}^{N} \times \mathbb{R}, \tag{5.1}
\end{equation*}
$$

where $A(z)=\left(a_{i j}(z)\right)$ is a symmetric and positive definite matrix. Throughout this section we systematically denote by $z=(x, t)$ and $\zeta=(\xi, \tau)$ the points in $\mathbb{R}^{N+1}$. We also denote by $\lambda_{1}(z), \ldots, \lambda_{N}(z)$ the eigenvalues of $A(z)$ and $\operatorname{set}^{2}$

$$
m:=\inf _{\substack{i=1, \ldots, N \\ z \in \mathbb{R}^{N+1}}} \lambda_{i}(z), \quad M:=\sup _{\substack{i=1, \ldots, N \\ z \in \mathbb{R}^{N+1}}} \lambda_{i}(z) \mu(z) .
$$

Our main hypotheses are the following.
[H1] $m, M$ are positive and real numbers.
[H2] The coefficients of $L$ are bounded functions and

$$
\begin{equation*}
\left|a_{i j}(x, t)-a_{i j}(\xi, \tau)\right| \leq \alpha\left(|x-\xi|+|t-\tau|^{\frac{1}{2}}\right), \quad(x, t),(\xi, \tau) \in \mathbb{R}^{N+1} \tag{5.2}
\end{equation*}
$$

for $i, j=1, \ldots, N$ and for some positive constant $\alpha$.
As a consequence of [H1] we have the usual uniform parabolicity condition:

$$
\begin{equation*}
m|\xi|^{2} \leq \sum_{i, j=1}^{N} a_{i j}(z) \xi_{i} \xi_{j} \leq M|\xi|^{2}, \quad \xi \in \mathbb{R}^{N}, z \in \mathbb{R}^{N+1} \tag{5.3}
\end{equation*}
$$

5.1. Forward parametrix. In this section for brevity we consider only the classical case, corresponding to $\bar{b}=\bar{c}=0$ in (2.4). Given $w \in \mathbb{R}^{N+1}$, we denote by $\Gamma_{w}(z ; \zeta)$ the fundamental solution to the frozen operator $L_{w}$ defined by

$$
\begin{equation*}
L_{w}=\sum_{i, j=1}^{N} a_{i j}(w) \partial_{x_{i} x_{j}}-\partial_{t} \tag{5.4}
\end{equation*}
$$

We recall that $\Gamma_{w}(z ; \zeta)=\Gamma_{w}(z-\zeta)$, where

$$
\begin{equation*}
\Gamma_{w}(x, t):=\Gamma_{w}(x, t ; 0,0)=\frac{(4 \pi t)^{-\frac{N}{2}}}{\sqrt{\operatorname{det} A(w)}} \exp \left(-\frac{\left\langle A^{-1}(w) x, x\right\rangle}{4 t}\right), \quad x \in \mathbb{R}^{N}, t>0 \tag{5.5}
\end{equation*}
$$

We define the forward parametrix

$$
\begin{equation*}
Z(z ; \zeta)=\Gamma_{\zeta}(z ; \zeta) \tag{5.6}
\end{equation*}
$$

[^87]We recall Notation 2.1 and remark explicitly that

$$
\begin{equation*}
L_{\zeta}^{(z)} Z(z ; \zeta)=0 \quad \text { for } z \neq \zeta . \tag{5.7}
\end{equation*}
$$

The following classical result (cf., for instance, Theorem 1.4 in [11]) states the existence of a fundamental solution $\Gamma$ of operator $L$.

Theorem 5.1. Assume hypotheses $[\mathbf{H 1}]$ and $[\mathbf{H 2}]$. Then for every $\zeta \in \mathbb{R}^{N+1}$, the function defined by

$$
\begin{equation*}
\Gamma(z ; \zeta)=Z(z ; \zeta)+\sum_{n=1}^{+\infty} \int_{\tau}^{t} \int_{\mathbb{R}^{N}} Z(z ; w)(L Z)_{n}(w ; \zeta) d w, \quad t>\tau \tag{5.8}
\end{equation*}
$$

is a fundamental solution of $L$ in (5.1). In (5.8) we have

$$
\begin{aligned}
(L Z)_{1}(w ; \zeta) & =L^{(w)} Z(w ; \zeta), \quad w=(y, s), \\
(L Z)_{n+1}(w ; \zeta) & =\int_{\tau}^{s} \int_{\mathbb{R}^{N}} L^{(w)} Z\left(w ; z_{0}\right)(L Z)_{n}\left(z_{0} ; \zeta\right) d z_{0}, \quad n \geq 1,
\end{aligned}
$$

and, for every $T>0$, the series

$$
\begin{equation*}
\Phi(w ; \zeta):=\sum_{n=1}^{+\infty}(L Z)_{n}(w ; \zeta) \tag{5.9}
\end{equation*}
$$

converges uniformly for $\left.w \in \mathbb{R}^{N} \times\right] \tau, \tau+T[$.
Our first result is the following global estimate for the parametrix approximation truncated at the $n$th term.

Theorem 5.2. Under the assumptions of Theorem 5.1, for every positive $\varepsilon$ we have

$$
\begin{align*}
\mid \Gamma(z ; \zeta)-Z(z ; \zeta) & -\sum_{k=1}^{n-1} \int_{\tau}^{t} \int_{\mathbb{R}^{N}} Z(z ; w)(L Z)_{k}(w ; \zeta) d w \mid \\
& \leq \sqrt{\frac{2}{\pi}}\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}} f_{n}\left(\eta_{\varepsilon, T} \sqrt{2 \pi(t-\tau)}\right) \Gamma^{M+\varepsilon}(z ; \zeta) \tag{5.10}
\end{align*}
$$

for $t \in] \tau, \tau+T\left[\right.$, where $\Gamma^{M+\varepsilon}$ is the Gaussian density defined in (A.1)-(A.2), $\eta_{\varepsilon, T}$ is the constant defined in (A.7), and

$$
\begin{equation*}
f_{n}(\eta)=e^{\frac{\eta^{2}}{2}}(\eta+1) \frac{\left(\frac{\eta^{2}}{2}\right)^{\left[\frac{n+1}{2}\right]}}{\left[\frac{n+1}{2}\right]!} \tag{5.11}
\end{equation*}
$$

with $[a]$ denoting the integer part of $a \in \mathbb{R}$.
Remark 5.3. We remark explicitly that when $\eta=\eta_{\varepsilon, T} \sqrt{2 \pi(t-\tau)}<1$ in (5.11),

$$
f_{n}(\eta) \leq C \frac{\eta^{n}}{\left(\frac{n}{2}\right)!}, \quad n \in \mathbb{N}
$$

for some positive constant $C$ so that the convergence of the parametrix approximation is extremely fast. This is the case, for instance, when $t-\tau \ll 1$, i.e., for short time to maturity. Also note that (5.10) is a global estimate with respect to the spatial variables.

Proof of Theorem 5.2. Theorem 5.2 is based on several results whose proofs are postponed to the appendix. We have

$$
\left|\Gamma(z ; \zeta)-Z(z ; \zeta)-\sum_{k=1}^{n-1} \int_{\tau}^{t} \int_{\mathbb{R}^{N}} Z(z ; w)(L Z)_{k}(w ; \zeta) d w\right| \leq \sum_{k=n}^{\infty} \int_{\tau}^{t} \int_{\mathbb{R}^{N}} Z(z ; w)\left|(L Z)_{k}(w ; \zeta)\right| d w
$$

(by Lemma A.1, estimate (A.8), and the reproduction property)

$$
\leq\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}} \Gamma^{M+\varepsilon}(z ; \zeta) \sum_{k=n}^{\infty} \int_{\tau}^{t} \frac{\Gamma_{E}\left(\frac{1}{2}\right)^{k}}{\Gamma_{E}\left(\frac{k}{2}\right)} \frac{\eta_{\varepsilon, T}^{k}}{(s-\tau)^{1-\frac{k}{2}}} d s
$$

(using the properties of the Gamma function ${ }^{3}$ )

$$
\begin{equation*}
=\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}} \Gamma^{M+\varepsilon}(z ; \zeta) \sqrt{\frac{2}{\pi}} \sum_{k=n}^{\infty} \frac{\left(\eta_{\varepsilon, T} \sqrt{2 \pi(t-\tau)}\right)^{k}}{k!!} . \tag{5.12}
\end{equation*}
$$

Then estimate (5.10) follows from some elementary computation. Indeed, if $n$ is even, then $\left[\frac{n+1}{2}\right]=\frac{n}{2}$ and we have

$$
\sum_{k=n}^{\infty} \frac{\eta^{k}}{k!!}=\sum_{k=\frac{n}{2}}^{\infty} \frac{\eta^{2 k}}{(2 k)!!}+\sum_{k=\frac{n}{2}+1}^{\infty} \frac{\eta^{2 k-1}}{(2 k-1)!!} \leq \sum_{k=\frac{n}{2}}^{\infty} \frac{\eta^{2 k}}{(2 k)!!}+\sum_{k=\frac{n}{2}+1}^{\infty} \frac{\eta^{2 k-1}}{(2 k-2)!!}
$$

$\left(\right.$ since $\left.(2 k)!!=2^{k} k!\right)$

$$
\leq \sum_{k=\frac{n}{2}}^{\infty} \frac{1}{k!}\left(\frac{\eta^{2}}{2}\right)^{k}+\sum_{k=\frac{n}{2}}^{\infty} \frac{\eta^{2 k+1}}{2^{k} k!}=f_{n}(\eta)
$$

with $f_{n}$ as in (5.11) and using the fact that

$$
\sum_{k=n}^{\infty} \frac{\eta^{k}}{k!}=\frac{e^{\eta} \eta^{n}}{n!}
$$

The case of $n$ odd can be treated analogously and is omitted.
From Theorem 5.2 we deduce the following forward parametrix expansion for solutions to the Cauchy problem for $L$.

[^88]Theorem 5.4. The solution to the Cauchy problem

$$
\begin{cases}L u(x, t)=0, & x \in \mathbb{R}^{N}, t>0  \tag{5.13}\\ u(x, 0)=\varphi(x), & x \in \mathbb{R}^{N}\end{cases}
$$

has an expansion of the form (2.17)-(2.18)-(2.19).
Proof. For simplicity let us consider only the one-dimensional case. By formulae (2.3) and (5.8) we have

$$
u(z)=\int_{\mathbb{R}} Z(z ; \xi, 0) \varphi(\xi) d \xi+\sum_{n=1}^{+\infty} \int_{\mathbb{R}}\left(\int_{\tau}^{t} \int_{\mathbb{R}} Z(z ; w)(L Z)_{n}(w ; \xi, 0) d w\right) \varphi(\xi) d \xi
$$

Using the expression of $(L Z)_{1}$ in (5.1), we have

$$
u_{1}(z)=\int_{\mathbb{R}} \varphi(\xi) \int_{0}^{t} \int_{\mathbb{R}} Z\left(z ; z_{0}\right) L Z\left(z_{0} ; \xi, 0\right) d z_{0} d \xi=\int_{0}^{t} \int_{\mathbb{R}} Z\left(z ; z_{0}\right) L \underbrace{\int_{\mathbb{R}} \varphi(\xi) Z\left(z_{0} ; \xi, 0\right) d \xi d z_{0}}_{=u_{0}\left(z_{0}\right)}
$$

Moreover,

$$
u_{2}(z)=\int_{\mathbb{R}} \varphi(\xi) \int_{0}^{t} \int_{\mathbb{R}} Z\left(z ; z_{1}\right) \int_{0}^{t_{1}} \int_{\mathbb{R}} L Z\left(z_{1} ; z_{0}\right) L Z\left(z_{0} ; \xi, 0\right) d z_{0} d z_{1} d \xi
$$

(changing the order of integration)

$$
\begin{aligned}
& =\int_{0}^{t} \int_{\mathbb{R}} Z\left(z ; z_{1}\right) \int_{0}^{t_{1}} \int_{\mathbb{R}} L Z\left(z_{1} ; z_{0}\right) L \underbrace{\int_{\mathbb{R}} \varphi(\xi) Z\left(z_{0} ; \xi, 0\right) d \xi}_{=u_{0}\left(z_{0}\right)} d z_{0} d z_{1} \\
& =\int_{0}^{t} \int_{\mathbb{R}} Z\left(z ; z_{1}\right)(L \underbrace{\int_{0}^{t_{1}} \int_{\mathbb{R}} Z\left(z_{1} ; z_{0}\right) L u_{0}\left(z_{0}\right) d z_{0}}_{=u_{1}\left(z_{1}\right)}+L u_{0}\left(z_{1}\right)) d z_{1}
\end{aligned}
$$

and this proves $(2.19)$ for $n=2$. The general case can be proved by induction.
As a byproduct of the parametrix method, we also obtain the following upper Gaussian estimate of the fundamental solution.

Corollary 5.5. For every $\varepsilon, T>0$, we have

$$
\Gamma(z ; \zeta) \leq\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}}\left(1+\eta_{\varepsilon, T} \sqrt{2 \pi(t-\tau)}\right) e^{\pi(t-\tau) \eta_{\varepsilon, T}^{2}} \Gamma^{M+\varepsilon}(z ; \zeta)
$$

for $\left.z, \zeta \in \mathbb{R}^{N+1}, t \in\right] \tau, \tau+T\left[\right.$, where $\Gamma^{M+\varepsilon}$ is the Gaussian density defined in (A.1)-(A.2) and $\eta_{\varepsilon, T}$ is the constant in (A.7).

Proof. By Theorem 5.2 we have

$$
\Gamma(z ; \zeta)=Z(z ; \zeta)+\sum_{k=1}^{\infty} \int_{\tau}^{t} \int_{\mathbb{R}^{N}} Z(z ; w)(L Z)_{k}(w ; \zeta) d w
$$

therefore, as in (5.12), we get

$$
\Gamma(z ; \zeta) \leq\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}} \Gamma^{M+\varepsilon}(z ; \zeta) \sum_{k=0}^{\infty} \frac{\left(\eta_{\varepsilon, T} \sqrt{2 \pi(t-\tau)}\right)^{k}}{k!!}
$$

and the thesis follows since

$$
\sum_{k=0}^{\infty} \frac{\eta^{k}}{k!!} \leq(1+\eta) e^{\frac{\eta^{2}}{2}}
$$

for $\eta>0$.
5.2. Backward parametrix. We assume the following additional hypothesis, which allows us to introduce the adjoint operator of $L$.
[H3] The derivatives $\partial_{x_{i}} a_{i j}, \partial_{x_{i} x_{j}} a_{i j}, \partial_{x_{i}} b_{i}$ are bounded functions.
We define the adjoint operator $\widetilde{L}$ of $L$ as usual:

$$
\begin{equation*}
\widetilde{L} u=\sum_{i, j=1}^{N} a_{i j} \partial_{x_{i} x_{j}} u+\sum_{i=1}^{N} \widetilde{b}_{i} \partial_{x_{i}} u+\widetilde{c} u+\partial_{t} u \tag{5.14}
\end{equation*}
$$

where

$$
\begin{equation*}
\widetilde{b}_{i}=-b_{i}+2 \sum_{j=1}^{N} \partial_{x_{i}} a_{i j}, \quad \widetilde{c}=c+\sum_{i, j=1}^{N} \partial_{x_{i} x_{j}} a_{i j}-\sum_{i=1}^{N} \partial_{x_{i}} b_{i} \tag{5.15}
\end{equation*}
$$

Thus we have

$$
\int_{\mathbb{R}^{N+1}} \varphi L \psi=\int_{\mathbb{R}^{N+1}} \psi \widetilde{L} \varphi, \quad \varphi, \psi \in C_{0}^{\infty}\left(\mathbb{R}^{N+1}\right)
$$

and the following classical result holds (cf., for instance, [17, Chap. 1, Theor. 15]).
Theorem 5.6. There exists a fundamental solution $\widetilde{\Gamma}$ of $\widetilde{L}$, and we have

$$
\begin{equation*}
\Gamma(z ; \zeta)=\widetilde{\Gamma}(\zeta ; z), \quad z, \zeta \in \mathbb{R}^{N+1}, z \neq \zeta \tag{5.16}
\end{equation*}
$$

For fixed $w \in \mathbb{R}^{N+1}$, we define the frozen operator

$$
\widetilde{L}_{w}^{(\zeta)}=\sum_{i, j=1}^{N} a_{i j}(w) \partial_{\xi_{i} \xi_{j}}+\partial_{\tau}
$$

and denote by $P(z ; \zeta)$ the backward parametrix defined as the fundamental solution of $\widetilde{L}_{w}^{(\zeta)}$ with $w=z$, or, more precisely,

$$
\begin{equation*}
P(z ; \zeta)=\widetilde{\Gamma}_{z}(\zeta ; z)=\Gamma_{z}(z-\zeta) \tag{5.17}
\end{equation*}
$$

for $\Gamma_{z}$ as in (5.5). Analogously to (5.7), we have

$$
\widetilde{L}_{z}^{(\zeta)} P(z ; \zeta)=0 \quad \text { for } z \neq \zeta
$$

Our main result reads as follows.
Theorem 5.7. Assume hypotheses [H1], [H2], and [H3]. Then, for every $\zeta \in \mathbb{R}^{N+1}$, the following expansion of the fundamental solution $\Gamma$ holds:

$$
\begin{equation*}
\Gamma(z ; \zeta)=P(z ; \zeta)+\int_{\tau}^{t} \int_{\mathbb{R}^{N}} P(z ; w) \Psi(w ; \zeta) d w, \quad t>\tau, \tag{5.18}
\end{equation*}
$$

where

$$
\begin{equation*}
\Psi(z ; \zeta)=\sum_{k=1}^{+\infty}(L P)_{k}(z ; \zeta), \tag{5.19}
\end{equation*}
$$

with

$$
\begin{aligned}
(L P)_{1}(z ; \zeta) & =L^{(z)} P(z, \zeta) \\
(L P)_{k+1}(z ; \zeta) & =\int_{\tau}^{t} \int_{\mathbb{R}^{N}} L^{(z)} Z(z ; w)(L P)_{k}(w ; \zeta) d w
\end{aligned}
$$

and, for every $T>0$, the series in (5.9) converges uniformly in the strip $\left.\mathbb{R}^{N} \times\right] \tau, \tau+T[$. Moreover, for every positive $\varepsilon$, we have the following estimate for the approximation truncated at the nth term:

$$
\begin{align*}
& \left|\Gamma(z ; \zeta)-P(z ; \zeta)-\sum_{k=1}^{n-1} \int_{\tau}^{t} \int_{\mathbb{R}^{N}} P(z ; w)(L P)_{k}(w ; \zeta) d w\right|  \tag{5.20}\\
& \leq \sqrt{\frac{2}{\pi}}\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}} f_{n}\left(\widetilde{\eta}_{\varepsilon, T} \sqrt{2 \pi(t-\tau)}\right) \Gamma^{M+\varepsilon}(z ; \zeta)
\end{align*}
$$

for $t \in] \tau, \tau+T\left[\right.$, where $\widetilde{\eta}_{\varepsilon, T}$ is defined in (A.10) and $f_{n}$ in (5.11). As a consequence, the solution to the Cauchy problem (5.13) has an expansion of the form (2.17)-(2.20)-(2.21).

Proof. Proceeding as in the forward case, one can prove that

$$
\begin{equation*}
\Gamma(z ; \zeta)=\widetilde{\Gamma}(\zeta ; z)=\widetilde{\Gamma}_{z}(\zeta ; z)+\int_{\tau}^{t} \int_{\mathbb{R}^{N}} \widetilde{\Gamma}_{w}(\zeta ; w) \widetilde{\Phi}(w ; z) d w, \quad t>\tau \tag{5.21}
\end{equation*}
$$

where

$$
\begin{equation*}
\widetilde{\Phi}(\zeta ; z)=\sum_{k=1}^{+\infty} I_{k}(\zeta ; z) \tag{5.22}
\end{equation*}
$$

with

$$
\begin{aligned}
I_{1}(\zeta ; z) & =\widetilde{L}^{(\zeta)} \widetilde{\Gamma}_{z}(\zeta ; z) \\
I_{k+1}(\zeta ; z) & =\int_{\tau}^{t} \int_{\mathbb{R}^{N}} \widetilde{L}^{(\zeta)} \widetilde{\Gamma}_{w}(\zeta ; w) I_{k}(w ; z) d w
\end{aligned}
$$

and the series converges uniformly on the strips. Moreover, error estimate (5.20) holds true. In order to conclude the proof, it suffices to invoke Theorem 5.6 to prove that the terms of the expansions (5.18)-(5.19) and (5.21)-(5.22) coincide, that is,

$$
\begin{equation*}
\int_{\tau}^{t} \int_{\mathbb{R}^{N}} P(z ; w)(L P)_{k}(w ; \zeta) d w=\int_{\tau}^{t} \int_{\mathbb{R}^{N}} \widetilde{\Gamma}_{w}(\zeta ; w) I_{k}(w ; z) d w \tag{5.23}
\end{equation*}
$$

for every $k \in \mathbb{N}$.
For $k=1$, recalling (5.17), we have

$$
\int_{\tau}^{t} \int_{\mathbb{R}^{N}} \widetilde{\Gamma}_{w}(\zeta ; w) I_{1}(w ; z) d w=\int_{\tau}^{t} \int_{\mathbb{R}^{N}} P(w ; \zeta) \widetilde{L}^{(w)} P(z ; w) d w,
$$

so that the thesis follows immediately by integrating by parts since we have no contribution at borders. Indeed, denoting $w=(y, s)$, formally we have

$$
\int_{\tau}^{t} \int_{\mathbb{R}^{N}} \Gamma_{w}(w ; \zeta) \partial_{s} \Gamma_{z}(z ; w) d w=\bar{I}-\int_{\tau}^{t} \int_{\mathbb{R}^{N}} \partial_{s} \Gamma_{w}(w ; \zeta) \Gamma_{z}(z ; w) d w
$$

where

$$
\bar{I}=\int_{\mathbb{R}^{N}} \Gamma_{(y, t)}(y, t ; \xi, \tau) \Gamma_{(x, t)}(x, t ; y, t) d y-\int_{\mathbb{R}^{N}} \Gamma_{(y, \tau)}(y, \tau ; \xi, \tau) \Gamma_{(x, t)}(x, t ; y, \tau) d y=0
$$

since $\Gamma_{(x, t)}(x, t ; y, t)=\delta_{x}(y)$ and $\Gamma_{(y, \tau)}(y, \tau ; \xi, \tau)=\delta_{\xi}(y)$. On the other hand, the above argument can be made rigorous by performing the integration by parts on a thinner strip $S_{\tau+\delta, t-\delta}$ and then applying the dominated convergence theorem as $\delta \rightarrow 0^{+}$combined with the summability estimate (A.9).

For $k=2$, we have

$$
\begin{aligned}
& \int_{\tau}^{t} \int_{\mathbb{R}^{N}} \Gamma_{z_{0}}\left(z_{0} ; \zeta\right) \int_{t_{0}}^{t} \int_{\mathbb{R}^{N}} \widetilde{L}^{\left(z_{0}\right)} \Gamma_{z_{1}}\left(z_{1} ; z_{0}\right) \widetilde{L}^{\left(z_{1}\right)} \Gamma_{z}\left(z ; z_{1}\right) d z_{1} d z_{0} \\
& =\int_{\tau}^{t} \int_{\mathbb{R}^{N}} \Gamma_{z_{0}}\left(z_{0} ; \zeta\right)\left(\widetilde{L}^{\left(z_{0}\right)} \int_{t_{0}}^{t} \int_{\mathbb{R}^{N}} \Gamma_{z_{1}}\left(z_{1} ; z_{0}\right) \widetilde{L}^{\left(z_{1}\right)} \Gamma_{z}\left(z ; z_{1}\right) d z_{1}\right. \\
& \left.+\int_{\mathbb{R}^{N}} \Gamma_{\left(y, t_{0}\right)}\left(y, t_{0} ; z_{0}\right) \widetilde{L}^{\left(y, t_{0}\right)} \Gamma_{z}\left(z ; y, t_{0}\right) d y\right) d z_{0} \equiv J_{1}+J_{2},
\end{aligned}
$$

where, using again that $\Gamma_{\left(y, t_{0}\right)}\left(y, t_{0} ; z_{0}\right)=\delta_{x_{0}}(y)$, we get

$$
J_{2}=\int_{\tau}^{t} \int_{\mathbb{R}^{N}} \Gamma_{z_{0}}\left(z_{0} ; \zeta\right) \widetilde{L}^{\left(z_{0}\right)} \Gamma_{z}\left(z ; z_{0}\right) d z_{0}
$$

(proceeding as in the case $k=1$ )

$$
=\int_{\tau}^{t} \int_{\mathbb{R}^{N}} L^{\left(z_{0}\right)} \Gamma_{z_{0}}\left(z_{0} ; \zeta\right) \Gamma_{z}\left(z ; z_{0}\right) d z_{0}
$$

on the other hand,

$$
J_{1}=\int_{\tau}^{t} \int_{\mathbb{R}^{N}} \Gamma_{z_{0}}\left(z_{0} ; \zeta\right) \widetilde{L}^{\left(z_{0}\right)} \int_{t_{0}}^{t} \int_{\mathbb{R}^{N}} \Gamma_{z_{1}}\left(z_{1} ; z_{0}\right) \widetilde{L}^{\left(z_{1}\right)} \Gamma_{z}\left(z ; z_{1}\right) d z_{1} d z_{0}
$$

(by parts as before)

$$
\begin{aligned}
& =\int_{\tau}^{t} \int_{\mathbb{R}^{N}} L^{\left(z_{0}\right)} \Gamma_{z_{0}}\left(z_{0} ; \zeta\right) \int_{t_{0}}^{t} \int_{\mathbb{R}^{N}} L^{\left(z_{1}\right)} \Gamma_{z_{1}}\left(z_{1} ; z_{0}\right) \Gamma_{z}\left(z ; z_{1}\right) d z_{1} d z_{0} \\
& -\int_{\mathbb{R}^{N}} \Gamma_{(y, \tau)}(y, \tau ; \xi, \tau) \int_{\tau}^{t} \int_{\mathbb{R}^{N}} L^{\left(z_{1}\right)} \Gamma_{z_{1}}\left(z_{1} ; y, \tau\right) \Gamma_{z}\left(z ; z_{1}\right) d z_{1} d y
\end{aligned}
$$

$\left(\operatorname{since} \Gamma_{(y, \tau)}(y, \tau ; \xi, \tau)=\delta_{\xi}(y)\right)$

$$
\begin{aligned}
& =\int_{\tau}^{t} \int_{\mathbb{R}^{N}} L^{\left(z_{0}\right)} \Gamma_{z_{0}}\left(z_{0} ; \zeta\right) \int_{t_{0}}^{t} \int_{\mathbb{R}^{N}} L^{\left(z_{1}\right)} \Gamma_{z_{1}}\left(z_{1} ; z_{0}\right) \Gamma_{z}\left(z ; z_{1}\right) d z_{1} d z_{0} \\
& -\int_{\tau}^{t} \int_{\mathbb{R}^{N}} L^{\left(z_{1}\right)} \Gamma_{z_{1}}\left(z_{1} ; \zeta\right) \Gamma_{z}\left(z ; z_{1}\right) d z_{1} .
\end{aligned}
$$

Combining the expressions of $J_{1}$ and $J_{2}$, eventually we obtain

$$
\begin{aligned}
& \int_{\tau}^{t} \int_{\mathbb{R}^{N}} \Gamma_{z_{0}}\left(z_{0} ; \zeta\right) \int_{t_{0}}^{t} \int_{\mathbb{R}^{N}} \widetilde{L}^{\left(z_{0}\right)} \Gamma_{z_{1}}\left(z_{1} ; z_{0}\right) \widetilde{L}^{\left(z_{1}\right)} \Gamma_{z}\left(z ; z_{1}\right) d z_{1} d z_{0} \\
& =\int_{\tau}^{t} \int_{\mathbb{R}^{N}} L^{\left(z_{0}\right)} P\left(z_{0} ; \zeta\right) \int_{t_{0}}^{t} \int_{\mathbb{R}^{N}} L^{\left(z_{1}\right)} P\left(z_{1} ; z_{0}\right) P\left(z ; z_{1}\right) d z_{1} d z_{0},
\end{aligned}
$$

which concludes the proof. As before, the previous argument should be made rigorous by some approximating procedure. The general case can be achieved by induction.

Appendix. We collect several lemmas that are preliminary to the proofs of Theorems 5.2 and 5.7. These lemmas are essentially estimates of $\Gamma_{w}$ in (5.5) and its derivatives in terms of the fundamental solution of the heat equation.

Given a constant $\mu>0$, we denote by $\Gamma^{\mu}$ the fundamental solution to the heat operator

$$
\begin{equation*}
\mu \sum_{i=1}^{N} \partial_{x_{i} x_{i}}-\partial_{t} . \tag{A.1}
\end{equation*}
$$

Lemma A.1. For every $z, \zeta, w \in \mathbb{R}^{N+1}$ with $z \neq \zeta$, we have

$$
\left(\frac{m}{M}\right)^{\frac{N}{2}} \Gamma^{m}(z ; \zeta) \leq \Gamma_{w}(z ; \zeta) \leq\left(\frac{M}{m}\right)^{\frac{N}{2}} \Gamma^{M}(z ; \zeta)
$$

Proof. We prove only the second inequality in the case $\zeta=0$. Keeping in mind formula (5.5), we see that the thesis follows directly from condition (5.3). Indeed, we have

$$
\begin{equation*}
\Gamma_{w}(z) \leq \frac{1}{(4 \pi t m)^{\frac{N}{2}}} \exp \left(-\frac{|x|^{2}}{4 t M}\right)=\left(\frac{M}{m}\right)^{\frac{N}{2}} \Gamma^{M}(z) \tag{A.2}
\end{equation*}
$$

Lemma A.2. For every $\varepsilon, \mu>0$ and $n \in \mathbb{N} \cup\{0\}$ we have

$$
\left(\frac{|x|}{\sqrt{t}}\right)^{n} \Gamma^{\mu}(x, t) \leq\left(\frac{n}{\varepsilon}\right)^{\frac{n}{2}}(\mu+\varepsilon)^{n}\left(\frac{\mu+\varepsilon}{\mu}\right)^{\frac{N}{2}} \Gamma^{\mu+\varepsilon}(x, t)
$$

for any $x \in \mathbb{R}^{N}$ and $t>0$.
Proof. Setting $a=\frac{|x|}{\sqrt{t}}$, we have

$$
\left(\frac{|x|}{\sqrt{t}}\right)^{n} \Gamma^{\mu}(z, 0)=a^{n}(4 \pi \mu t)^{-\frac{N}{2}} \exp \left(-\frac{a^{2}}{4 \mu}\right) \leq(4 \pi \mu t)^{-\frac{N}{2}} \exp \left(-\frac{a^{2}}{4(\mu+\varepsilon)}\right) \sup _{\mathbb{R}_{+}} G
$$

where

$$
\begin{equation*}
G(a)=a^{n} \exp \left(-\left(\frac{1}{4 \mu}-\frac{1}{4(\mu+\varepsilon)}\right) a^{2}\right) \tag{A.3}
\end{equation*}
$$

The thesis follows by a straightforward computation, since $G$ attains a global maximum at $\bar{a}=\sqrt{\frac{2 n \mu(\mu+\varepsilon)}{\varepsilon}}$ and

$$
G(\bar{a})=\left(\frac{2 n \mu(\mu+\varepsilon)}{e \varepsilon}\right)^{\frac{n}{2}} \leq\left(\frac{n}{\varepsilon}\right)^{\frac{n}{2}}(\mu+\varepsilon)^{n}
$$

Lemma A.3. For every $\varepsilon>0$ and $i, j=1, \ldots, N$ we have

$$
\begin{align*}
\left|\partial_{x_{i}} \Gamma_{w}(z ; \zeta)\right| & \leq \frac{1}{2 \sqrt{\varepsilon(t-\tau)}}\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}+1} \Gamma^{M+\varepsilon}(z ; \zeta),  \tag{A.4}\\
\left|\partial_{x_{i} x_{j}} \Gamma_{w}(z ; \zeta)\right| & \leq \frac{1}{\varepsilon(t-\tau)}\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}+2} \Gamma^{M+\varepsilon}(z ; \zeta) \tag{A.5}
\end{align*}
$$

for any $z, \zeta, w \in \mathbb{R}^{N+1}$ with $t>\tau$.
Proof. For the sake of simplicity, we prove the above estimates in the case $\zeta=0$. We have

$$
\left|\partial_{x_{i}} \Gamma_{w}(z)\right|=\frac{1}{2} \frac{\left|\left(A^{-1}(w) x\right)_{i}\right|}{t} \Gamma_{w}(z)
$$

(by Lemma A.1)

$$
\leq \frac{1}{2 m \sqrt{t}}\left(\frac{M}{m}\right)^{\frac{N}{2}} \frac{|x|}{\sqrt{t}} \Gamma^{M}(z)
$$

and (A.5) follows by applying Lemma A. 2 with $\mu=M$ and $n=1$.
Moreover,

$$
\left|\partial_{x_{i} x_{j}} \Gamma_{w}(z)\right|=\frac{1}{2 t}\left|A^{-1}(w)_{i j}+\frac{1}{2 t}\left(A^{-1}(w) x\right)_{i}\left(A^{-1}(w) x\right)_{j}\right| \Gamma_{w}(z) \leq \frac{1}{2 t}\left(\frac{1}{m}+\frac{|x|^{2}}{2 m^{2} t}\right) \Gamma_{w}(z)
$$

and (A.4) easily follows by Lemmas A. 1 and A. 2 with $\mu=M$.

Lemma A.4. For every positive $\varepsilon$ and $T$, we have

$$
\begin{equation*}
\left.\left|L^{(z)} Z(z ; \zeta)\right| \leq \frac{\eta_{\varepsilon, T}}{\sqrt{t-\tau}} \Gamma^{M+\varepsilon}(z ; \zeta), \quad z, \zeta \in \mathbb{R}^{N+1}, t \in\right] \tau, \tau+T[ \tag{A.6}
\end{equation*}
$$

where

$$
\begin{equation*}
\eta_{\varepsilon, T}:=\alpha N^{2}\left(\frac{2}{\varepsilon}\right)^{\frac{3}{2}}\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}+2}\left(M+\varepsilon+\sqrt{\frac{\varepsilon}{2}}\right)+\beta \frac{N}{2 \sqrt{\varepsilon}}\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}+1} \tag{A.7}
\end{equation*}
$$

$$
+\gamma\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}} \sqrt{T}
$$

and

$$
\beta:=\sup _{\substack{i=1, \ldots, N \\ z \in \mathbb{R}^{N+1}}}\left|b_{i}(z)\right|, \quad \gamma:=\sup _{z \in \mathbb{R}^{N+1}}|c(z)|
$$

and $\alpha$ is the constant in (5.2).
Proof. For $t>\tau$, we have

$$
|L Z(z ; \zeta)|=\left|\left(L-L_{\zeta}\right) Z(z ; \zeta)\right| \leq I_{1}+I_{2}+I_{3},
$$

where

$$
I_{1}=\sum_{i, j=1}^{N}\left|a_{i j}(z)-a_{i j}(\zeta)\right|\left|\partial_{x_{i} x_{j}} Z(z ; \zeta)\right|
$$

(by (5.2))

$$
\leq \alpha N^{2}(|x-\xi|+\sqrt{t-\tau}) \max _{i, j}\left|\partial_{x_{i} x_{j}} Z(z ; \zeta)\right|
$$

(by Lemma A.3)

$$
\leq \frac{\alpha N^{2}}{\varepsilon}\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}+2}\left(1+\frac{|x-\xi|}{\sqrt{t-\tau}}\right) \Gamma^{M+\varepsilon}(z ; \zeta)
$$

(by Lemma A.2)

$$
\begin{aligned}
& \leq \frac{\alpha N^{2}}{\varepsilon}\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}+2}\left(\frac{M+2 \varepsilon}{M+\varepsilon}\right)^{\frac{N}{2}}\left(1+\frac{M+2 \varepsilon}{\sqrt{\varepsilon}}\right) \Gamma^{M+2 \varepsilon}(z ; \zeta) \\
& \leq \frac{\alpha N^{2}}{\varepsilon^{\frac{3}{2}}}\left(\frac{M+2 \varepsilon}{m}\right)^{\frac{N}{2}+2}(M+2 \varepsilon+\sqrt{\varepsilon}) \Gamma^{M+2 \varepsilon}(z ; \zeta)
\end{aligned}
$$

Moreover, by Lemma A.3, we have

$$
I_{2}=\sum_{i=1}^{N}\left|b_{i}(z)\right|\left|\partial_{x_{i}} Z(z ; \zeta)\right| \leq \beta \frac{N}{2 \sqrt{\varepsilon(t-\tau)}}\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}+1} \Gamma^{M+\varepsilon}(z ; \zeta)
$$

finally, by Lemma A.1, we have

$$
I_{3}=|c(z)| Z(z ; \zeta) \leq \gamma\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}} \Gamma^{M+\varepsilon}(z ; \zeta)
$$

Lemma A.5. For every $\varepsilon>0$ and $k \geq 1$ the following estimate for the term $(L Z)_{k}$ in (5.9) holds:

$$
\begin{equation*}
\left|(L Z)_{k}(z ; \zeta)\right| \leq \frac{\Gamma_{E}\left(\frac{1}{2}\right)^{k}}{\Gamma_{E}\left(\frac{k}{2}\right)} \frac{\eta_{\varepsilon, T}^{k}}{(t-\tau)^{1-\frac{k}{2}}} \Gamma^{M+\varepsilon}(z ; \zeta), \quad z, \zeta \in \mathbb{R}^{N+1}, t>\tau \tag{A.8}
\end{equation*}
$$

where $\eta_{\varepsilon, T}$ is defined in (A.7) and $\Gamma_{E}$ denotes Euler's Gamma function.
Proof. We prove (A.8) by induction on $k$. The case $k=1$ was proved in Lemma A.4. Let us now assume that (A.8) holds for $k$ and prove it for $k+1$. We have

$$
\left|(L Z)_{k+1}(z ; \zeta)\right|=\left|\int_{\tau}^{t} \int_{\mathbb{R}^{N}} L^{(z)} Z(z ; w)(L Z)_{k}(w ; \zeta) d w\right|
$$

(by Lemma A.4, the inductive hypothesis, and denoting $(y, s)=w$ )

$$
\leq \eta_{\varepsilon, T}^{k+1} \frac{\Gamma_{E}\left(\frac{1}{2}\right)^{k}}{\Gamma_{E}\left(\frac{k}{2}\right)} \int_{\tau}^{t} \frac{1}{\sqrt{t-s}(s-\tau)^{1-\frac{k}{2}}} \int_{\mathbb{R}^{N}} \Gamma^{M+\varepsilon}(x, t ; y, s) \Gamma^{M+\varepsilon}(y, s ; \xi, \tau) d y d s
$$

(by the reproduction property ${ }^{4}$ for $\Gamma^{M+\varepsilon}$ and by the change of variable $s=(1-r) \tau+r t$ )

$$
=\frac{\eta_{\varepsilon, T}^{k+1}}{(t-\tau)^{1-\frac{k+1}{2}}} \frac{\Gamma_{E}\left(\frac{1}{2}\right)^{k}}{\Gamma_{E}\left(\frac{k}{2}\right)} \int_{0}^{1} \frac{1}{r^{1-\frac{k}{2}} \sqrt{1-r}} d r \Gamma^{M+\varepsilon}(z ; \zeta)
$$

and the thesis follows by the known properties ${ }^{5}$ of Euler's Gamma function.
Finally, we recall Notation 2.1 and state the dual version of Lemmas A. 3 and A.4. Proofs are omitted since they are analogous.

Lemma A.6. For every $\varepsilon>0$ and $i, j=1, \ldots, N$ we have

$$
\begin{aligned}
\left|\partial_{\xi_{i}} \Gamma_{w}(z ; \zeta)\right| & \leq \frac{1}{2 \sqrt{\varepsilon(t-\tau)}}\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}+1} \Gamma^{M+\varepsilon}(z ; \zeta), \\
\left|\partial_{\xi_{i} \xi_{j}} \Gamma_{w}(z ; \zeta)\right| & \leq \frac{1}{\varepsilon(t-\tau)}\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}+2} \Gamma^{M+\varepsilon}(z ; \zeta)
\end{aligned}
$$

for any $z, \zeta, w \in \mathbb{R}^{N+1}$ with $t>\tau$.
Lemma A.7. Under hypotheses $[\mathbf{H 1}]-[\mathbf{H 3}]$, for every positive $\varepsilon$, we have

$$
\begin{equation*}
\left|\widetilde{L}^{(\zeta)} P(z ; \zeta)\right| \leq \frac{\widetilde{\eta}_{\varepsilon, T}}{\sqrt{t-\tau}} \Gamma^{M+\varepsilon}(z ; \zeta), \quad z, \zeta \in \mathbb{R}^{N+1}, t>\tau \tag{A.9}
\end{equation*}
$$

${ }^{4}$ For every $x, \xi \in \mathbb{R}^{N}$ and $\tau<s<t$, we have

$$
\int_{\mathbb{R}^{N}} \Gamma^{M+\varepsilon}(z ; y, s) \Gamma^{M+\varepsilon}(y, s ; \zeta) d y=\Gamma^{M+\varepsilon}(z ; \zeta)
$$

${ }^{5}$ It holds that

$$
\int_{0}^{1} \frac{1}{r^{1-\frac{k}{2}} \sqrt{1-r}} d r=\frac{\Gamma_{E}\left(\frac{1}{2}\right) \Gamma_{E}\left(\frac{k}{2}\right)}{\Gamma_{E}\left(\frac{k+1}{2}\right)} .
$$

where
(A.10)

$$
\begin{aligned}
\widetilde{\eta}_{\varepsilon, T}:=\alpha N^{2}\left(\frac{2}{\varepsilon}\right)^{\frac{3}{2}}\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}+2}\left(M+\varepsilon+\sqrt{\frac{\varepsilon}{2}}\right) & +\widetilde{\beta} \frac{N}{2 \sqrt{\varepsilon}}\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}+1} \\
& +\widetilde{\gamma}\left(\frac{M+\varepsilon}{m}\right)^{\frac{N}{2}} \sqrt{T}
\end{aligned}
$$

where

$$
\widetilde{\beta}:=\sup _{\substack{i=1, \ldots, N \\ z \in \mathbb{R}^{N+1}}}\left|\widetilde{b}_{i}(z)\right|, \quad \widetilde{\gamma}:=\sup _{z \in \mathbb{R}^{N+1}}|\widetilde{c}(z)|
$$

Acknowledgment. We thank the two anonymous referees for valuable remarks that improved the paper.

## REFERENCES

[1] Y. Aїт-Sahalia, Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach, Econometrica, 70 (2002), pp. 223-262.
[2] Y. Аїт-Sahalia, Estimating continuous-time models using discretely sampled data, in Advances in Economics and Econometrics, Theory and Applications, Ninth World Congress, R. Blundell, P. Torsten, and W. K. Newey, Econometric Society Monographs, Cambridge University Press, Cambridge, UK, 2007.
[3] G. Barone-Adesi and R. J. Elliott, Approximations for the values of American options, Stochastic Anal. Appl., 9 (1991), pp. 115-131.
[4] I. Bouchouev and V. Isakov, Uniqueness, stability and numerical methods for the inverse problem that arises in financial markets, Inverse Problems, 15 (1999), pp. R95-R116.
[5] M. Broadie and J. Detemple, The valuation of American options on multiple assets, Math. Finance, 7 (1997), pp. 241-286.
[6] M. Broadie, P. Glasserman, and S. Kou, A continuity correction for discrete barrier options, Math. Finance, 7 (1997), pp. 325-349.
[7] J. Cox, Notes on option pricing I: Constant elasticity of variance diffusions, Working paper, Stanford University, Stanford, CA, 1975.
[8] J. Cox and S. Ross, The valuation of options for alternative stochastic processes, J. Financial Economics, 3 (1976), pp. 145-166.
[9] E. Derman, Regimes of volatility, Quantitative Strategies Research Notes, Goldman Sachs, New York, 1999.
[10] J. N. Dewynne and W. T. Shaw, Differential equations and asymptotic solutions for arithmetic Asian options: "Black-Scholes formulae" for Asian rate calls, European J. Appl. Math., 19 (2008), pp. 353-391.
[11] M. Di Francesco and A. Pascucci, On a class of degenerate parabolic equations of Kolmogorov type, AMRX Appl. Math. Res. Express, 2005 (2005), pp. 77-116.
[12] P. Foschi and A. Pascucci, Analytic formulas for local volatility models by the parametrix method, Preprint AMS Acta, University of Bologna, Bologna, Italy, 2010.
[13] P. Foschi and A. Pascucci, Path dependent volatility, Decis. Econ. Finance, 31 (2008), pp. 13-32.
[14] P. Foschi and A. Pascucci, Calibration of a path-dependent volatility model: Empirical tests, Comput. Statist. Data Anal., 53 (2009), pp. 2219-2235.
[15] J.-P. Fouque, G. Papanicolaou, and K. R. Sircar, Derivatives in Financial Markets with Stochastic Volatility, Cambridge University Press, Cambridge, UK, 2000.
[16] J.-P. Fouque, G. Papanicolaou, R. Sircar, and K. Solna, Singular perturbations in option pricing, SIAM J. Appl. Math., 63 (2003), pp. 1648-1665.
[17] A. Friedman, Partial Differential Equations of Parabolic Type, Prentice-Hall, Englewood Cliffs, NJ, 1964.
[18] P. Hagan, D. Kumar, A. Lesniewski, and D. Woodward, Managing smile risk, Wilmott, September (2002), pp. 84-108.
[19] P. Hagan and D. Woodward, Equivalent Black volatilities, Appl. Math. Finance, 6 (1999), pp. 147-159.
[20] D. G. Hobson and L. C. G. Rogers, Complete models with stochastic volatility, Math. Finance, 8 (1998), pp. 27-48.
[21] S. Howison, Matched asymptotic expansions in financial engineering, J. Engrg. Math., 53 (2005), pp. 385-406.
[22] Y. L. Hsu, T. I. Lin, and C. F. Lee, Constant elasticity of variance (CEV) option pricing model: Integration and detailed derivation, Math. Comput. Simulation, 79 (2008), pp. 60-71.
[23] S. M. IAcus, Simulation and Inference for Stochastic Differential Equations. With $R$ Examples, Springer Series in Statistics, Springer, New York, 2008.
[24] P. JÄCkel, Hyperbolic local volatility, http://www.jaeckel.org/HyperbolicLocalVolatility.pdf, 2008.
[25] C. Kahl and P. JÄckel, Fast strong approximation Monte Carlo schemes for stochastic volatility models, Quant. Finance, 6 (2006), pp. 513-536.
[26] V. Konakov, S. Menozzi, and S. Molchanov, Explicit parametrix and local limit theorems for some degenerate diffusion processes, Ann. Inst. H. Poincaré (B) Probab. Statist., to appear.
[27] R. Kuske and J. Keller, Optimal exercise boundary for an American put option, Appl. Math. Finance, 5 (1999), pp. 107-116.
[28] E. E. Levi, Sulle equazioni lineari totalmente ellittiche alle derivate parziali, Rend. Circ. Mat. Palermo, 24 (1907), pp. 275-317.
[29] J. ObloJ, Fine-tune your smile: Correction to Hagan et al., http://arxiv.org/abs/0708.0998, 2008.
[30] M. Schroder, Computing the constant elasticity of variance option pricing formula, J. Finance, 44 (1989), pp. 211-219.
[31] W. T. Shaw, Modelling Financial Derivatives with Mathematica, Cambridge University Press, Cambridge, UK, 1998.
[32] S. Svoboda-Greenwood, The displaced diffusion as an approximation of the CEV, Appl. Math. Finance, 66 (2009), pp. 269-286.
[33] S. Turnbull and L. Wakeman, A quick algorithm for pricing European average options, J. Fin. and Quant. Anal., 26 (1991), pp. 377-389.
[34] A. E. Whalley and P. Wilmott, An asymptotic analysis of an optimal hedging model for option pricing with transaction costs, Math. Finance, 7 (1997), pp. 307-324.
[35] M. Widdicks, P. W. Duck, A. D. Andricopoulos, and D. P. Newton, The Black-Scholes equation revisited: Asymptotic expansions and singular perturbations, Math. Finance, 15 (2005), pp. 373-391.
[36] M. Widdicks, P. W. Duck, Y. Chao, and D. P. Newton, Singular perturbation techniques applied to multiasset option pricing, Math. Finance, 19 (2009), pp. 457-486.
[37] J. Zhang, Pricing continuously sampled Asian options with perturbation method, J. Futures Markets, 26 (2003), pp. 535-560.

# Exact and Efficient Simulation of Correlated Defaults* 

K. Giesecke ${ }^{\dagger}$, H. Kakavand ${ }^{\dagger}$, M. Mousavi ${ }^{\dagger}$, and H. Takada ${ }^{\ddagger}$


#### Abstract

Correlated default risk plays a significant role in financial markets. Dynamic intensity-based models, in which a firm default is governed by a stochastic intensity process, are widely used to model correlated default risk. The computations in these models can be performed by Monte Carlo simulation. The standard simulation method, which requires the discretization of the intensity process, leads to biased simulation estimators. The magnitude of the bias is often hard to quantify. This paper develops an exact simulation method for intensity-based models that leads to unbiased estimators of credit portfolio loss distributions, risk measures, and derivatives prices. In a first step, we construct a Markov chain that matches the marginal distribution of the point process describing the binary default state of each firm. This construction reduces the original estimation problem to one involving a Markov chain expectation. In a second step, we estimate the Markov chain expectation using a simple acceptance/rejection scheme that facilitates exact sampling. To address rare event situations, the acceptance/rejection scheme is embedded in an overarching selection/mutation scheme, in which a selection mechanism adaptively forces the chain into the regime of interest. Numerical experiments demonstrate the effectiveness of the method for a self-exciting model of correlated default risk.


Key words. portfolio credit risk, Markovian projection, rare-event simulation, acceptance/rejection, selection/mutation

AMS subject classifications. 60G55, 60J27, 60J75, 90-08, 65C05
DOI. 10.1137/090778055

1. Introduction. Correlated default risk is one of the most pervasive threats in financial markets. Confronting this threat is a daily business for credit investors such as banks making loans to individuals and corporations or fixed income managers allocating assets in the credit markets. These investors must measure the aggregate default risk in their asset portfolios and devise strategies to mitigate that risk. These tasks typically involve estimating the risk capital, to cushion potential default losses at high confidence levels, and estimating the prices of portfolio credit derivatives, which are financial instruments that provide insurance against correlated default risk.

Risk management and derivatives pricing applications require a stochastic model of correlated default timing. Intensity-based models are widely used for this purpose. In these models, a portfolio constituent firm defaults at an inaccessible stopping time whose stochastic structure is governed by an intensity, or conditional default rate. The intensity follows

[^89]a stochastic process that reflects the information revealed over time, including the value of exogenous risk factors and the state of other firms in the economy. The intensity processes are correlated across firms, to incorporate the dependence between firm defaults. While many intensity models have been developed in the literature, model computation remains challenging. The scope of semianalytical transform methods is limited, including mainly doubly stochastic models. In these models, firm intensities are driven by common risk factors. Conditional on a realization of these factors, default times are independent of one another.

Monte Carlo simulation is an alternative tool for performing computations in intensitybased models. It can be applied to models outside the scope of transform methods or to address applications for which transform methods are unsuitable. A standard simulation method, which applies to most intensity models and is thus widely used, exploits a timechange result for point processes due to [58]. Meyer showed that a nonexplosive counting process can be transformed into a standard Poisson process by a change of time given by the counting process compensator, or cumulative intensity. This implies that the first jump time of the process is equal in law to the first hitting time of the compensator process to a standard exponential variable. This insight provides a recipe for simulating a default time with given intensity process: generate a path of the cumulative intensity and record its first hitting time to a level drawn independently from a standard exponential distribution. The resulting default times have the correct joint distribution, as implied by the correlated evolution of firm intensity processes.

While widely applicable, the time-scaling scheme may lead to biased simulation estimators. This is because it may not be possible to construct the full path of the continuous-time stochastic process followed by the time integral of the intensity. Often, the path must be approximated on a discrete-time grid. Further, it may be difficult to draw exact samples of the values of the integrated intensity at the grid points, because the joint distribution of the integrated intensities across firms, from which one needs to sample, is rarely computationally tractable. In this case, the values of an integrated intensity must be approximated by first simulating the continuous-time intensity process on the discrete-time grid, and then integrating the discretized values. If the intensity values cannot be sampled from their joint transition law, then the SDEs that describe the joint dynamics must be discretized by the Euler or some higher order scheme. Due to the multiple layers of approximations, it is hard to quantify the magnitude of the discretization bias in the resulting simulation estimators. While the bias can be reduced by increasing the number of discretization time steps, this comes at the expense of increasing the time required to generate a replication. Since the additional computational effort per firm replication is scaled by the number of firms in the portfolio, this can quickly lead to a computational burden that is prohibitive for the large portfolios that occur in practice.

This paper develops an exact sampling method that leads to unbiased simulation estimators for intensity-based models. The scope of the method is more limited than that of the time-scaling method but wider than that of the transform methods. It comprises a broad range of models proposed in the literature, including doubly stochastic and self-exciting formulations. The method has two parts. We first construct an inhomogeneous, continuous-time Markov chain $M$ whose value at $t$ has the same distribution as the value at $t$ of the point process $N$ describing the binary default state of each portfolio constituent. The construction reduces the problem of estimating the expectation of $f\left(N_{t}\right)$ to that of estimating the expec-
tation of $f\left(M_{t}\right)$. Unlike $N$, the mimicking chain $M$ has deterministic interarrival intensities, and this facilitates the exact sampling of $M$ using a simple acceptance/rejection scheme. As a result, we obtain an unbiased estimator of the expectation of $f\left(M_{t}\right)$.

For portfolios of high-quality names, most replications produce few if any defaults, so the computational effort required to obtain accurate estimators may be very large. This is especially true for estimators of large loss probabilities or tail risk measures, which are at the center of risk management applications. To address this problem, we embed the acceptance/ rejection scheme in an overarching selection/mutation scheme developed by [21]. On a discretetime grid we evolve a collection of particles, i.e., copies of the mimicking Markov chain $M$, using the acceptance/rejection scheme. At each time step, particles are randomly selected by sampling with replacement, placing more weight on particles with a larger number of transitions in the previous period. The new generation of selected particles is then evolved over the next period, at whose end a selection takes place again. The selection procedure adaptively forces the chain into the regime of interest and therefore reduces variance. The resulting estimators inherit the unbiasedness of the plain acceptance/rejection estimators.

Numerical experiments demonstrate the effectiveness of the method for a portfolio of 100 names. We analyze a self-exciting model, in which firm intensities follow correlated Feller jump-diffusion processes that jump whenever a default event occurs. We find that the exact method requires significantly less computation time than the conventional time-scaling method, for all levels of accuracy. The root mean square errors of the simulation estimators converge much faster for the exact method. The selection/mutation scheme is found to offer substantial variance reduction.
1.1. Related literature. While many alternative intensity-based models of correlated default risk have been developed in the literature, there is surprisingly little work on simulation methods for these models. The authors of [27] review time-scaling and other approaches. [24] provides an inverse transform method for simulating the first to occur of a given set of events. This scheme is exact and can be used to sample the default times sequentially; it leads to unbiased estimators of an expectation of a function of the vector of default times of the constituent names. The method developed in this paper leads to unbiased estimators of an expectation of a function of the future value of the vector point process indicating the default state of the constituent names.

Giesecke, Kakavand, and Mousavi [38] develop an exact method for the related problem of simulating a one-dimensional, real-valued point process. They project the point process onto its own filtration and then sample it in this coarser filtration. The sampling is based on the intensity in the subfiltration, which is deterministic between arrival times and therefore facilitates the use of exact schemes. This projection method leads to unbiased estimators of an expectation of a function of the path of the one-dimensional point process and a skeleton of the driving state process.

Bassamboo and Jain [5] propose an asymptotically optimal importance sampling scheme to estimate the probability of large portfolio losses in a doubly stochastic intensity model with affine risk factor processes. Their approach exploits the conditional independence of firm defaults in the doubly stochastic setting. The implementation of the estimators relies on the time-scaling scheme.

Carmona, Fouque, and Vestal [11] use a selection/mutation scheme to estimate the distri-
bution of portfolio loss in a structural model of correlated default risk. Here, a firm defaults when its market value hits a given barrier. Firm values follow correlated stochastic volatility processes. Carmona and Crepey [10] numerically contrast the performance of the selection/ mutation and importance sampling schemes when estimating the distribution of portfolio loss in a Markov chain model.

There are several papers on variance reduction schemes for copula-based models of correlated default risk. In a copula-based model the firm dependence structure is specified by a copula function that maps firm-level default probabilities to the joint default probability. [6], [14], [49], [40], [42] develop importance sampling schemes that exploit the conditional independence of firm defaults that is also a feature of the copula models. Glasserman and Li [41] examine a related scheme for a mixed Poisson model of portfolio credit risk.
1.2. Structure of this article. Section 2 formulates the problem, reviews conventional simulation approaches, and outlines the exact method. Section 3 discusses the construction of the mimicking Markov chain. Section 4 develops two algorithms for estimating expectations associated with the mimicking chain. Section 5 constructs the mimicking Markov chain for a broad range of models proposed in the literature. Section 6 provides a numerical case study that demonstrates the effectiveness of the method. An appendix contains the proofs.

## 2. Preliminaries.

2.1. Default point processes. Consider a portfolio of $n$ firms that are subject to default risk. The random default times of these firms are modeled by almost surely distinct stopping times $\tau^{i}>0$, which are defined on a complete probability space $(\Omega, \mathcal{F}, P)$ with right-continuous and complete information filtration $\mathbb{F}=\left(\mathcal{F}_{t}\right)_{t \geq 0}$. In risk management applications, $P$ is the statistical probability, while in derivatives pricing applications, $P$ is a risk-neutral pricing measure. Associated with the $\tau^{i}$ are indicator processes $N^{i}$ given by $N_{t}^{i}=I\left(\tau^{i} \leq t\right)$, where $I(A)$ is the indicator function of an event $A \in \mathcal{F}$. For each $i$, there is a strictly positive, integrable, and progressively measurable process $\lambda^{i}$ such that the variables

$$
\begin{equation*}
N_{t}^{i}-\int_{0}^{t} I\left(\tau^{i}>s\right) \lambda_{s}^{i} d s \tag{1}
\end{equation*}
$$

form a martingale relative to $\mathbb{F}$. This means that $\lambda_{t}^{i}$ is the $\mathcal{F}_{t}$-conditional default rate of firm $i$ at time $t<\tau^{i}$, measured in events per unit of time. We refer to the process $\lambda^{i}$ as the default intensity of firm $i$, recognizing that this may involve an innocuous abuse of terminology, as $\lambda^{i}$ need not drop to 0 at $\tau^{i}$. The intensities follow correlated stochastic processes that need not be specified at this point. The correlation among the intensities reflects the default dependence structure of the portfolio constituent firms.

Our goal is to calculate $E\left[f\left(N_{T}\right)\right]$ for a suitable real-valued function $f$ on $\{0,1\}^{n}$ and a horizon $T>0$, where $N=\left(N^{1}, \ldots, N^{n}\right)$ is the vector of firm default indicator processes. Examples include the single-name probabilities $P\left(\tau^{i}>T\right)$, joint probabilities $P\left(\cap_{i \in S}\left\{\tau^{i}>T\right\}\right)$ for subsets $S$ of firms, the distribution $P\left(C_{T}=k\right)$ of the default counting process $C=1_{n} \cdot N$, where $1_{n}$ is an $n$-vector of ones, and the option $E\left[\left(C_{T}-K\right)^{+}\right]$. If taken under the statistical probability measure, $P\left(C_{T}=k\right)$ is the fundamental quantity for the risk management of portfolios of corporate debt. If evaluated under a risk-neutral pricing measure, the option
$E\left[\left(C_{T}-K\right)^{+}\right]$is a key quantity required to price portfolio credit derivatives, as illustrated in [31].

For clarity in the exposition, our formulation has not made explicit the role of the financial loss at a default. This is without loss of generality in case the loss $\ell=\left(\ell^{1}, \ldots, \ell^{n}\right)$ is deterministic as in [5], [6], [10], [11], [14], [15], [49], [40] and others. In that case, $\ell$ can be incorporated into the function $f$, to express the probability that the portfolio loss $\ell \cdot N_{T}$ at $T$ exceeds a level, or an option on $\ell \cdot N_{T}$, as $E\left[f\left(N_{T}\right)\right]$. To treat a loss $\ell$ that is random but independent of $N$ as in [42], the algorithms developed below require only minor modifications. Independence is a reasonable assumption for lack of data bearing on the correlation between $\ell$ and $N$.
2.2. Conventional simulation schemes. To estimate $E\left[f\left(N_{T}\right)\right]$ by simulation, we need to sample the variable $N_{T}$. This is straightforward if the intensity $\lambda$ is deterministic and $N$ is a Poisson process. It is also straightforward if $\lambda$ follows a stochastic process that is adapted to the subfiltration of $\mathbb{F}$ generated by $N$, in which case $\lambda$ is deterministic between default times. In these cases, the classical thinning, or acceptance/rejection (A/R) scheme of [54], can be used to sample the jump times of $N$ exactly. This scheme leads to unbiased simulation estimators of $E\left[f\left(N_{T}\right)\right]$. It involves the sampling of candidate arrival times from a dominating Poisson process, and an acceptance test. However, if $\lambda$ is not adapted to the filtration generated by $N$, then it evolves randomly also between events, and a dominating Poisson process may not exist. In this case, one may be able to use an inverse transform scheme to sample the default times exactly. This requires an explicit expression of the probabilities $P\left(\tau^{i}>T\right)$. It may also be possible to apply the inverse scheme sequentially, as in [24] and [27]. This requires that the conditional distributions of the interarrival times of $C$ and the defaulter identities be tractable and also the ability to sample the state variables determining these distributions.

A more widely applicable method exploits a time-change result for point processes due to [58]. Consider a counting process $Z$ with jumps of size one and compensator $\widehat{Z}$ that is continuous and increases to $\infty$ almost surely. Meyer proved that $Z$ is a standard Poisson process under a change of time defined by $\widehat{Z}$, relative to the time-changed filtration. Thus, the first jump time of $Z$ is equal in law to $\inf \left\{t: \widehat{Z}_{t} \geq \mathcal{E}\right\}$, where $\mathcal{E}$ is a standard exponential random variable. This provides a recipe for simulating the first jump time of $Z$ : generate a path of $\widehat{Z}$ and record its hitting time to the level $\mathcal{E}$, drawn independently from a standard exponential distribution. To apply this recipe to generating a path of $N^{i}$ for a given $\lambda^{i}$, we let $Z$ be a counting process with compensator $\widehat{Z}_{t}=\int_{0}^{t} \lambda_{s}^{i} d s$ and set $N^{i}=\min (Z, 1)$. We generate a path of $\int_{0}^{t} \lambda_{s}^{i} d s$ and draw $\mathcal{E}$ to obtain a sample of $\tau^{i}$ as the hitting time $\inf \left\{t: \int_{0}^{t} \lambda_{s}^{i} d s \geq \mathcal{E}\right\}$. As the $\lambda^{i}$ are correlated across $i$, the $\int_{0}^{t} \lambda_{s}^{i} d s$ must be drawn from the appropriate joint distribution.

While the time-scaling scheme has a wide scope, it suffers from an important shortcoming: it usually leads to biased estimators of $E\left[f\left(N_{T}\right)\right]$. This is because, aside from very special cases, it is impossible to generate the full path of the continuous-time stochastic process $\int_{0}^{t} \lambda_{s}^{i} d s$. The path must be approximated on a discrete-time grid. Even worse, it may not be possible to draw exact samples of the values of $\int_{0}^{t} \lambda_{s}^{i} d s$ at the grid points, because the distribution of $\left(\int_{0}^{t} \lambda_{s}^{1} d s, \ldots, \int_{0}^{t} \lambda_{s}^{n} d s\right)$ from which one needs to sample is rarely known or computationally tractable. This forces one to approximate the values of $\int_{0}^{t} \lambda_{s}^{i} d s$ by first
approximating the continuous-time intensity process $\lambda^{i}$ on the discrete-time grid and then integrating the discretized values. If the intensity values cannot be sampled exactly from their joint transition law, then the SDE that describes the dynamics of $\left(\lambda^{1}, \ldots, \lambda^{n}\right)$ must be discretized by the Euler or some other scheme.

Due to the multiple approximations and the multiple dimensions $n$ of $N$, it is hard to quantify the magnitude of the discretization bias in the estimators of $E\left[f\left(N_{T}\right)\right]$. The bias can be reduced by increasing the number of discretization time steps, but this also increases the computational cost of a replication. Reducing the bias to an acceptable level may require a prohibitively large computational effort, since the dimension $n$ of $N$ is often large in practice, and the effort scales with $n$. Further, it is hard to determine the optimal trade-off between the number of discretization time steps and the number of simulation trials because the convergence rate of the bias is unknown.
2.3. Exact simulation. Below we develop an alternative simulation method that eliminates the need to discretize the vector process $\left(\lambda^{1}, \ldots, \lambda^{n}\right)$ and that leads to unbiased estimators of $E\left[f\left(N_{T}\right)\right]$. The method has two parts. We first construct a time-inhomogeneous continuous-time Markov chain $M=\left(M^{1}, \ldots, M^{n}\right) \in\{0,1\}^{n}$ with the property that $M_{T}=N_{T}$ in distribution for each fixed $T$. This construction is explained in section 3. It reduces the problem of estimating the general point process expectation $E\left[f\left(N_{T}\right)\right]$ to the simpler problem of estimating the Markov chain expectation $E\left[f\left(M_{T}\right)\right]$. Estimators of $E\left[f\left(M_{T}\right)\right]$ are obtained by exact sampling of $M_{T}$ using a thinning scheme, as explained in section 4 .

## 3. Mimicking Markov chain.

3.1. Construction. Throughout, we let $0_{n}$ denote an $n$-vector of zeros.

Proposition 3.1. Suppose that the default indicator process $N$ has intensity $\lambda$. Let $M$ be $a$ Markov chain on $[0, \infty)$ that takes values in $\{0,1\}^{n}$, starts at $0_{n}$, has no joint transitions in any of its components, and whose ith component has transition rate $h^{i}(\cdot, M)$, where

$$
\begin{equation*}
h^{i}(t, B)=E\left(\lambda_{t}^{i} I\left(\tau^{i}>t\right) \mid N_{t}=B\right), \quad B \in\{0,1\}^{n} . \tag{2}
\end{equation*}
$$

Then $M_{T}=N_{T}$ in distribution for each $T \geq 0$.
Proposition 3.1 shows that a component transition function $h^{i}(t, B)$ of the mimicking Markov chain $M$ is given by the projection of the primitive firm intensity $\lambda_{t}^{i} I\left(\tau^{i}>t\right)$ onto the value of the default process $N_{t}=B$, which indicates the state at time $t$ of each firm in the portfolio. The indicator $I\left(\tau^{i}>t\right)$ guarantees that $h^{i}(t, B)$ vanishes if $B^{i}=1$, i.e., in a state where firm $i$ is in default. In the special case where $N$ is a priori a Markov point process, $\lambda_{t}^{i}=q^{i}\left(t, N_{t}\right)$ for some function $q^{i}$ on $\mathbb{R}_{+} \times\{0,1\}^{n}$, and $h^{i}(t, B)=q^{i}(t, B)\left(1-B^{i}\right)$. In general, $N$ is not a priori a Markov point process. Then the conditional expectation (2) is nontrivial, and further steps are required to calculate it for the given process $\lambda$; see section 5 . In any case, Proposition 3.1 leads to a Markov point process $M$ whose value at $t$ has the same distribution as $N_{t}$.

The fact that the construction leads to a Markov process is a consequence of the conditioning set in the conditional expectation (2). Rather than conditioning on the sigma-field $\sigma\left(N_{s}: s \leq t\right)$ generated by the path of $N$ during $[0, t]$, which seems natural at first, the conditional expectation (2) is taken with respect to the final value $N_{t}$ only. The conditioning on
the path does not in general produce a Markov point process. Nevertheless, the conditional expectation $E\left(\lambda_{t}^{i} I\left(\tau^{i}>t\right) \mid \sigma\left(N_{s}: s \leq t\right)\right)$ is meaningful: it defines the intensity of $N^{i}$ in the subfiltration generated by $N$. As shown in [38], this observation can be used to develop an alternative projection scheme for the exact simulation of a point process. That scheme samples a point process in its own filtration, based on the intensity in the subfiltration. It is appropriate for estimating an expectation $E\left[g\left(U_{t}: t \leq T\right)\right]$ that involves the path of a univariate point process $U \in \mathbb{R}$. For computational reasons, the scheme is less well suited to sampling a vector point process such as $N=\left(N^{1}, \ldots, N^{n}\right) \in\{0,1\}^{n}$. The method developed in this paper targets the vector process $N$ but is restricted to expectations of the form $E\left[f\left(N_{T}\right)\right]$. This is because the auxiliary chain $M$ matches the distribution of $N_{T}$ for fixed $T$ only.

Proposition 3.1 extends a univariate construction in [8, Chapter II, exercise E8], which is refined and applied by [2], [18], and [56] to the calibration of (univariate) intensity-based, top-down models of portfolio credit risk. These papers construct a mimicking Markov chain for a nonterminating counting process taking values in $\mathbb{N}_{0}$. Their setting is different from ours even if $n=1$, because the counting processes $N^{i}$ that we consider take values in $\{0,1\}$. Lopatin [55] suggests a multivariate version. Bentata and Cont [7] analyze the construction of a mimicking Markov process for a semimartingale that may be discontinuous.
3.2. Markov point process. To prepare the design of simulation algorithms for $M$ in section 4 below, we consider the mimicking chain $M$ as a Markov point process relative to its own right-continuous and complete filtration $\mathbb{G}=\left(\mathcal{G}_{t}\right)_{t \geq 0}$ generated by $M$. The construction of $M$ implies that, for a suitable real-valued function $g$ on $\{0,1\}^{n}$, the process $g(M)-\int_{0}^{n} \mathscr{A}_{s} g\left(M_{s}\right) d s$ is a martingale in the filtration $\mathbb{G}$, where

$$
\mathscr{A}_{t} g(B)=\sum_{i=1}^{n} h^{i}(t, B)\left(g\left(B^{[i]}\right)-g(B)\right), \quad B \in\{0,1\}^{n},
$$

is the generator of $M$ at $t$. Here, $B^{[i]}$ denotes the vector $B$ whose $i$ th element $B^{i}$ is replaced by $1-B^{i}$. It follows that the process

$$
\begin{equation*}
M^{i}-\int_{0} h^{i}\left(s, M_{s}\right) d s \tag{3}
\end{equation*}
$$

is a martingale with respect to $\mathbb{G}$. Thus, the component counting process $M^{i}$ has intensity $h^{i}(\cdot, M)$ in the filtration $\mathbb{G}$. Recall that $h^{i}(\cdot, B)$ vanishes for any $B \in\{0,1\}^{n}$ whose $i$ th element is equal to 1 , and compare with the Doob-Meyer decomposition (1) of the firm default indicator process $N^{i}$ in the reference filtration $\mathbb{F}$. By Proposition 3.1, the distributions of $N_{t}^{i}$ and $M_{t}^{i}$ agree, and so do the distributions of $C_{t}=1_{n} \cdot N_{t}$ and $1_{n} \cdot M_{t}$. Let $h$ be the $n$-vector of the functions $h^{i}$. From the martingale property of (3),

$$
\begin{equation*}
1_{n} \cdot M-\int_{0} 1_{n} \cdot h\left(s, M_{s}\right) d s \tag{4}
\end{equation*}
$$

is a $\mathbb{G}$-martingale as well. Therefore, the counting process $1_{n} \cdot M$ has intensity $1_{n} \cdot h(\cdot, M)$ relative to the filtration $\mathbb{G}$. As indicated in Table 1 , that intensity is the counterpart to the

Table 1
Indicator and counting processes and their Markovian counterparts.

| Filtration | Indicator <br> process | Component <br> process | Component <br> intensity | Counting <br> process | Counting <br> intensity |
| :---: | :---: | :---: | :---: | :---: | :---: |
| $\mathbb{F}$ | $N$ | $N^{i}$ | $\lambda^{i}\left(1-N^{i}\right)$ | $1_{n} \cdot N$ | $\sum_{i=1}^{n} \lambda^{i}\left(1-N^{i}\right)$ |
| $\mathbb{G}$ | $M$ | $M^{i}$ | $h^{i}(\cdot, M)$ | $1_{n} \cdot M$ | $1_{n} \cdot h(\cdot, M)$ |

intensity $\Lambda$ of $C=1_{n} \cdot N$ in the reference filtration $\mathbb{F}$, given by

$$
\begin{equation*}
\Lambda=\sum_{i=1}^{n}\left(1-N^{i}\right) \lambda^{i} \tag{5}
\end{equation*}
$$

We have

$$
\begin{equation*}
1_{n} \cdot h\left(t, M_{t}\right)=E\left(\Lambda_{t} \mid N_{t}=M_{t}\right)=\sum_{k=0}^{n-1} H(t, k) I\left(T_{k} \leq t<T_{k+1}\right) \tag{6}
\end{equation*}
$$

almost surely. Here, $\left(T_{k}\right)_{k=0,1, \ldots, n}$ is the sequence of event times of $1_{n} \cdot M$ starting at 0 , which is strictly increasing almost surely because there are no joint transitions in $M$ almost surely (Proposition 3.1), and $H(t, k)$ is the $\mathcal{G}_{T_{k}}$-measurable interarrival intensity function given by

$$
\begin{equation*}
H(t, k)=1_{n} \cdot h\left(t, M_{T_{k}}\right), \quad t \geq T_{k}, k=0,1, \ldots, n \tag{7}
\end{equation*}
$$

Formula (7) implies that the interarrival intensities of the mimicking Markov counting process $1_{n} \cdot M$ evolve deterministically through time. This is a key property that our simulation algorithms are going to exploit. Note that the original model, which is formulated in the filtration $\mathbb{F}$, has more complicated interarrival intensity dynamics. This can be seen from formula (5), which indicates that the interarrival $\mathbb{F}$-intensities of $C=1_{n} \cdot N$ follow stochastic processes whenever the firm $\mathbb{F}$-intensities $\lambda^{i}$ do.

Because $H(t, k)$ is $\mathcal{G}_{T_{k}}$-measurable, the random variable $T_{k+1}-T_{k}$ is equal in $\mathcal{G}_{T_{k}}$-conditional distribution to the first jump time of a time-inhomogeneous Poisson process starting at $T_{k}$ with intensity function $H(t, k)$ for $t \geq T_{k}$. Thus, for all $s>0$ the conditional survival function of the interarrival times of $1_{n} \cdot M$ satisfies

$$
\begin{align*}
P\left(T_{k+1}-T_{k}>s \mid \mathcal{G}_{T_{k}}\right) & =P\left(1_{n} \cdot M_{T_{k}+s}=k \mid M_{T_{k}}\right) \\
& =\exp \left(-\int_{T_{k}}^{T_{k}+s} H(t, k) d t\right), \quad k=0,1, \ldots, n-1 \tag{8}
\end{align*}
$$

Let $I_{k} \in\{1,2, \ldots, n\}$ be the $\mathcal{G}_{T_{k}}$-measurable random variable identifying the component of $M$ in which the $k$ th transition takes place, $k=1, \ldots, n$. Noting that the sigma-field $\mathcal{G}_{T_{k}-}$ is generated by the random variables $\left(T_{m}, I_{m}\right)_{m \leq k-1}$ and $T_{k}$, and using an argument similar to the one applied by [8, Theorem II.15], we see that

$$
\begin{equation*}
P\left(I_{k}=i \mid \mathcal{G}_{T_{k}-}\right)=\frac{h^{i}\left(T_{k}, M_{T_{k-1}}\right)}{H\left(T_{k}, k-1\right)}, \quad i, k=1,2, \ldots, n \tag{9}
\end{equation*}
$$

4. Exact simulation algorithms. We wish to evaluate the expectation $E\left[f\left(N_{T}\right)\right]$ for suitable functions $f$ on $\{0,1\}^{n}$ and fixed $T>0$. The key insight is that Proposition 3.1 reduces the problem of evaluating $E\left[f\left(N_{T}\right)\right]$ to the problem of evaluating $E\left[f\left(M_{T}\right)\right]$ for the mimicking Markov chain $M$. Since the number of portfolio constituents $n$ can be 100 or even larger, we estimate $E\left[f\left(M_{T}\right)\right]$ by Monte Carlo simulation of $M$ rather than through alternative numerical methods that would be plagued by the high dimensionality of the state space $\{0,1\}^{n}$. This section discusses two exact simulation algorithms for this purpose.
4.1. Sequential acceptance/rejection scheme. We simulate the mimicking chain $M$ by sequentially generating the event times and identities ( $T_{k}, I_{k}$ ) introduced above. The generation of event times is based on the interarrival intensities (7) of $1_{n} \cdot M$, which evolve deterministically through time. Given an event time, the corresponding identity is drawn from the discrete distribution (9).

The inverse or time-scaling methods can be used to generate $T_{k+1}$ from formula (8) for the conditional survival function $S_{k}(t)$ of the interarrival time $T_{k+1}-T_{k}$. Draw $U \sim U(0,1)$ and calculate the inverse $\inf \left\{s>0: S_{k}(s) \leq U\right\}$. While allowing for exact sampling, this procedure requires us to evaluate the function $S_{k}(t)$ at many points $t$ in order to determine the inverse at $U$. Depending on the structure of the function $H(s, k)$, this may be numerically intensive since $S_{k}(t)$ involves the time-integral of $H(s, k)$.

We prefer an alternative acceptance/rejection (A/R) scheme, which is based on the classical A/R, or thinning, scheme of [54]. This scheme requires the evaluation of $H(\cdot, k)$ only at a set of candidate times for $T_{k+1}$. The candidate times are generated from a Poisson process whose rate dominates $H\left(T_{k}+s, k\right)$ for $s$ in some interval. A candidate time $c$ is accepted with a probability given by the ratio of $H\left(T_{k}+c, k\right)$ to the dominating Poisson rate. The tighter the dominating bound on $H\left(T_{k}+s, k\right)$, the fewer candidate times need to be generated. Therefore, the dominating Poisson process is redetermined at least at each acceptance or rejection of a candidate time.

Algorithm 4.1. To generate a sample path of $M$ over $[0, T]$, perform the following:

1. Initialize $t=0, k=0, T_{0}=0$, and $M_{0}=0_{n}$.
2. Stop if $t \geq T$.
3. Find $J(k)=J(t, k)$ and $K(k)=K(t, k)$ such that $H(t+s, k) \leq J(k), 0 \leq s \leq K(k)$.
4. Draw a random variable $\mathcal{E}$ from the exponential distribution with parameter $J(k)$.

- If $\mathcal{E}>K(k)$, then set $t=t+K(k)$ and go to step 2 .
- If $\mathcal{E} \leq K(k)$, then draw $U \sim U(0,1)$. If $U J(k) \leq H(t+\mathcal{E}, k)$, then set $k=k+1$ and $t=T_{k}=t+\mathcal{E}$. Else set $t=t+\mathcal{E}$ and go to step 2 .

5. Draw a random variable I from the discrete distribution

$$
P(I=i)=\frac{h^{i}\left(T_{k}, M_{T_{k-1}}\right)}{H\left(T_{k}, k-1\right)}, \quad i=1,2, \ldots, n
$$

and let $Q_{k}$ be the $n$-vector with the $I$ th component equal to one and the rest equal to zero. Set $M_{T_{k}}=Q_{1}+\cdots+Q_{k}$. Go to step 2 .
Algorithm 4.1 is applied to generate a collection of $R$ sample paths $M^{1}, \ldots, M^{R}$ of the mimicking Markov chain $M$ over $[0, T]$. Thanks to Proposition 3.1, for suitable functions $f$ on $\{0,1\}^{n}$ we can estimate the expectation $E\left[f\left(N_{T}\right)\right]$ by $\frac{1}{R} \sum_{r=1}^{R} f\left(M_{T}^{r}\right)$. In particular, the
distribution of the portfolio default count $P\left(C_{T}=k\right)$ is estimated by

$$
\begin{equation*}
P_{R}\left(C_{T}=k\right)=\frac{1}{R} \sum_{r=1}^{R} I\left(1_{n} \cdot M_{T}^{r}=k\right), \quad k=0,1, \ldots, n . \tag{10}
\end{equation*}
$$

Algorithm 4.1 also leads to estimators of firm-level probabilities $P\left(\tau^{i}>T\right)$ and related quantities, such as $P\left(\cap_{i \in S}\left\{\tau^{i}>T\right\}\right)$ for subsets $S$ of firms. The setting may simply be $n=1$, with only a single firm of interest. In this case, (10) is an estimator of $P\left(N_{T}^{1}=k\right)$ for $k=0,1$, while step 5 of the algorithm is redundant. In the general case of $n>1$, estimates of $P\left(\tau^{i}>T\right)$ are obtained as a byproduct: with $e^{i}$ denoting an $n$-vector with its $i$ th component equal to one and the rest equal to zero, we take $f\left(M_{T}^{r}\right)=I\left(e^{i} \cdot M_{T}^{r}=0\right)$.

The estimators $E\left[f\left(N_{T}\right)\right]$ generated by Algorithm 4.1 are unbiased, because the A/R scheme generates exact samples of the mimicking chain $M$, and $E\left[f\left(N_{T}\right)\right]=E\left[f\left(M_{T}\right)\right]$. The A/R scheme applies to $M$ because this process has deterministic interarrival intensities that are usually easy to bound. The A/R scheme does not generally apply to the original default indicator process $N$, because this process has stochastic interarrival intensities that are usually not bounded by a constant almost surely.

Algorithm 4.1 may be inefficient in some situations. This occurs, for example, when the portfolio constituents have small default probabilities, which is typical for investment grade portfolios of highly rated issuers, and we are interested in estimating tail probabilities of $N_{T}$. In this situation, a prohibitively large number of replications may be required to estimate these probabilities accurately with Algorithm 4.1.
4.2. Selection/mutation scheme. To reduce variance, we embed the sequential thinning mechanism into a selection/mutation $(\mathrm{S} / \mathrm{M})$ scheme. Let $T>0$ be the simulation horizon. Partition the interval $[0, T]$ into $m$ subintervals of length $T / m$. Let $V$ be the discrete-time Markov chain given by

$$
V_{p}=M_{p T / m}, \quad p=0,1, \ldots, m .
$$

We consider a collection of "particles" $\left\{V_{p}^{r}\right\}_{r=1,2, \ldots, R}$ that are evolved on the discretetime grid $p=0,1, \ldots, m$, all starting from the same state $0_{n}$ at $p=0$. At a time step $p$, we use the sequential A/R Algorithm 4.1 to independently mutate (evolve) each particle $V_{p}^{r}$ during ( $p, p+1$ ] according to the transition rates determined by Proposition 3.1. Then, before entering the next mutation step, we select particles according to the number of transitions during $(p, p+1]$. The selection is done probabilistically, by sampling with replacement. The selection probability increases with the number of transitions during ( $p, p+1$ ], so the selection favors particles with transitions. The total number of particles $R$ is kept constant, and the selected particles are then evolved over the next period. The final estimator "corrects" for the selections performed at each time step.

Algorithm 4.2. To generate an estimate of $P\left(C_{T}=k\right)$, perform the following:

1. Initialize $V_{0}^{r}=W_{0}^{r}=0_{n}$ for $1 \leq r \leq R$.
2. For each $0 \leq p \leq m-1$, repeat the following steps:

Selection.
Fix a constant $\delta>0$. From the set of particles $\left(W_{p}^{r}, V_{p}^{r}\right)_{1 \leq r \leq R}$ at $p$

- Compute the normalizing constant $\eta_{p}=\frac{1}{R} \sum_{r=1}^{R} \exp \left[\delta 1_{n} \cdot\left(V_{p}^{r}-W_{p}^{r}\right)\right]$.
- Select, independently and with replacement, $R$ new particles as follows. Select particle $r$ with probability

$$
\begin{equation*}
\frac{1}{R \eta_{p}} \exp \left[\delta 1_{n} \cdot\left(V_{p}^{r}-W_{p}^{r}\right)\right] \tag{11}
\end{equation*}
$$

Denote the $R$ selected particles by $\left(\hat{W}_{p}^{r}, \hat{V}_{p}^{r}\right)_{1 \leq r \leq R}$.

## Mutation.

Evolve the particle $\left(\hat{W}_{p}^{r}, \hat{V}_{p}^{r}\right)$ to $\left(W_{p+1}^{r}, V_{p+1}^{r}\right)$, independently for each $1 \leq r \leq R$.

- Set $W_{p+1}^{r}=\hat{V}_{p}^{r}$.
- Obtain $V_{p+1}^{r}$ by generating transitions from $\hat{V}_{p}^{r}$ using Algorithm 4.1.

3. For $k=0,1, \ldots, n$, calculate the estimator of $P\left(C_{T}=k\right)$ as

$$
\begin{equation*}
P_{\delta, m, R}\left(C_{T}=k\right)=\frac{\eta_{0} \cdots \eta_{m-1}}{R} \sum_{r=1}^{R} I\left(1_{n} \cdot V_{m}^{r}=k\right) \exp \left(-\delta 1_{n} \cdot W_{m}^{r}\right) . \tag{12}
\end{equation*}
$$

Algorithm 4.2 is a variant of an $\mathrm{S} / \mathrm{M}$ scheme for estimating rare-event probabilities for time-inhomogeneous Markov chains developed and analyzed by [21]. From their Theorem 2.3 and the fact that the $\mathrm{A} / \mathrm{R}$ Algorithm 4.1 facilitates exact sampling from the marginal distribution of $M$ in the mutation step of the scheme, we conclude that the estimator (12) is unbiased in the sense that $E\left[P_{\delta, m, R}\left(C_{T}=k\right)\right]=P\left(C_{T}=k\right)$ for fixed $\delta, m, R$.

The algorithm requires the selection of the number of particles $R$, the number of time steps $m$, and the value of $\delta$. The parameter $\delta$ specifies the exponential weight function $\exp \left[\delta 1_{n} \cdot\left(V_{p}^{r}-W_{p}^{r}\right)\right]$, which determines the probability distribution used for the sampling with replacement in the particle selection step of the algorithm. ${ }^{1}$ Here, $1_{n} \cdot\left(V_{p}^{r}-W_{p}^{r}\right)$ is the number of transitions (defaults) of the Markov chain particle $r$ during period $p$. For $\delta=0$, each particle has the same probability of being selected. For $\delta>0$, the selection probability increases with the number of transitions. The larger $\delta$ is, the relatively greater is the focus on particles with a larger number of transitions. In the extreme case, the particle with the largest number of transitions is selected $R$ times.

Thus, for a positive $\delta$ the selection step favors particles with a greater number of transitions. It tends to replace particles with few transitions with those that experienced more transitions. As a result, with each selection step the particles are forced further into the regime of interest, i.e., a scenario with a large number of defaults during the simulation interval $[0, T]$. The number of time steps $m$ determines the number of selections performed during $[0, T]$. All else being equal, the larger the $m$, the faster the particles transition to the rare event regime. ${ }^{2}$ The estimator (12) accounts for the selections performed at each time step: compare with the estimator (10) generated by the A/R scheme. The required adjustment to the estimator (10) follows from formula (3.13) in [21] and is governed by the form of the weight function.

The parameters $\delta$ and $m$ need to be chosen appropriately to guarantee variance reduction for the event of interest. However, the optimal configuration problem for a general setting

[^90]has not yet been addressed to our knowledge. In practice, simple experiments can lead to a reasonable configuration for a given setting. We explain this in the context of our numerical case study in section 6 .

There are alternative approaches to variance reduction. Relative to an importance sampling (IS) scheme, the $\mathrm{S} / \mathrm{M}$ algorithm eliminates the need to determine and simulate from an importance measure under which the event of interest is not rare anymore. In the $\mathrm{S} / \mathrm{M}$ scheme, we simulate the chain $M$ under the reference measure using the $\mathrm{A} / \mathrm{R}$ algorithm and the transition rate functions $h^{i}$. The chain is adaptively forced into the regime of interest by the selection mechanism. As shown by [21], the selection mechanism can be interpreted as a twisting of Feynman-Kac particle path measures, a measure change analogous to the one underpinning IS. ${ }^{3}$ An advantage of an IS formulation is that it often enables one to establish certain asymptotic optimality properties of the IS estimator. ${ }^{4}$ These properties formally prove the effectiveness of the estimator in a rare-event regime and lead to an optimal configuration of the algorithm. However, for the time-inhomogeneous Markov chain considered here, the optimal IS scheme has not yet been worked out, to our knowledge. ${ }^{5}$
5. Calculating the projection. The practical implementation of the exact simulation method requires the construction of the mimicking Markov chain $M$ for the intensity model $\lambda$ at hand. This construction amounts to the calculation of the conditional expectation $E\left(\lambda_{t}^{i} I\left(\tau^{i}>t\right) \mid N_{t}=B\right)$ defining the transition rates $h^{i}(t, B)$ of $M$; see Proposition 3.1. This section calculates this expectation for a range of doubly stochastic, frailty, and self-exciting models and therefore extends the scope of the exact simulation method to many models proposed in the literature. The calculation relies, intuitively speaking, on Bayes' rule and leads to an explicit expression for the transition rate $h^{i}(t, B)$ in terms of the transform

$$
\phi(t, u, z, Z)=E\left[\exp \left(-u \int_{0}^{t} Z_{s} d s-z Z_{t}\right)\right]
$$

where $Z$ is a nonnegative stochastic process and $u$ and $z$ are reals. This transform can be computed in closed form for a wide range of processes $Z$, including affine jump-diffusion processes. For any choice of $Z$ that we consider below, we assume that $\left.\partial_{z} \phi(t, u, z, Z)\right|_{z=0}$ exists and is finite. Below, $B$ denotes $\left(B^{1}, \ldots, B^{n}\right) \in\{0,1\}^{n}$.
5.1. Doubly stochastic models. We begin with a simple model in which firms default independently of one another. The calculation of the corresponding Markov transition rate $h^{i}(t, B)$ serves as a stepping stone for the calculation in models with a nontrivial default dependence structure.

Proposition 5.1. Suppose that $N$ is doubly stochastic ${ }^{6}$ with intensities $\lambda^{i}=X^{i}$ for mutually independent nonnegative adapted processes $X^{i}$. Then, for $B^{i}=0$, we have

$$
h^{i}(t, B)=-\frac{\left.\partial_{z} \phi\left(t, 1, z, X^{i}\right)\right|_{z=0}}{\phi\left(t, 1,0, X^{i}\right)}
$$

[^91]We generalize to a model in which a firm is exposed to an idiosyncratic risk factor $X^{i}$ and a systematic risk factor $Y$ that is common to all firms. The random variation of $Y$ generates correlated movements in firms' conditional default probabilities. Conditional on a realization of $Y$, firms default independently of one another. This and related formulations have been used extensively in theoretical and empirical analyses of correlated default risk; see, for example, [12], [13], [19], [23], [25], [28], [30], [32], [46], [47], [51], [59], and [60].

Proposition 5.2. Suppose that $N$ is doubly stochastic with intensities $\lambda^{i}=X^{i}+\alpha^{i} Y$ for mutually independent nonnegative adapted processes $X^{1}, \ldots, X^{n}, Y$ and nonnegative factor loadings $\alpha^{i}$. Then, for $B^{i}=0$, we have

$$
\begin{equation*}
h^{i}(t, B)=-\frac{\left.\partial_{z} \phi\left(t, 1, z, X^{i}\right)\right|_{z=0}}{\phi\left(t, 1,0, X^{i}\right)}-\alpha^{i} \frac{\left.\sum_{k=0}^{2^{n}-1} c_{k}(t) \partial_{z} \phi\left(t, b_{k}, z, Y\right)\right|_{z=0}}{\sum_{k=0}^{2^{n}-1} c_{k}(t) \phi\left(t, b_{k}, 0, Y\right)}, \tag{13}
\end{equation*}
$$

where the deterministic functions $c_{k}(t)$ and the constants $b_{k}$ are determined by the relation

$$
\begin{equation*}
\sum_{k=0}^{2^{n}-1} c_{k}(t) e^{-b_{k} v}=\prod_{j=1}^{n}\left[B^{j}-\left(2 B^{j}-1\right) \phi\left(t, 1,0, X^{j}\right) e^{-\alpha^{j} v}\right], \quad v>0 \tag{14}
\end{equation*}
$$

The multiplication of the $n$ terms on the right-hand side of (14) results in a sum of at most $2^{n}$ terms, and typically fewer as $B^{j}=0$ for some $j$. The $c_{k}(t)$ are the coefficients of the summands. Each constant $b_{k}$ is a sum of values $\alpha^{j}$ for certain $j$; note that $b_{0}=0$ and $b_{2^{n}-1}=\sum_{j=1}^{n} \alpha^{j}$. An algorithm for the efficient computation of the $c_{k}(t)$ and the $b_{k}$ is based on the recursive scheme of [1].

We can extend to a doubly stochastic model with multiple common factors, allowing the description of a more sophisticated firm dependence structure.

Proposition 5.3. Suppose that $N$ is doubly stochastic with intensities $\lambda^{i}=X^{i}+\alpha^{i} \cdot Y$ for mutually independent nonnegative adapted processes $X^{1}, \ldots, X^{n}$ and $Y=\left(Y_{1}, \ldots, Y_{q}\right)$ and nonnegative factor loadings $\alpha^{i}=\left(\alpha_{1}^{i}, \ldots, \alpha_{q}^{i}\right)$. Then, for $B^{i}=0$, we have

$$
\begin{align*}
h^{i}(t, B)= & -\frac{\left.\partial_{z} \phi\left(t, 1, z, X^{i}\right)\right|_{z=0}}{\phi\left(t, 1,0, X^{i}\right)} \\
& -\frac{\left.\sum_{l=1}^{q} \alpha_{l}^{i} \sum_{k=0}^{2^{n}-1} c_{k}(t) \partial_{z_{l}} \phi\left(t, b_{k 1}, z_{1}, Y_{1}\right) \cdots \phi\left(t, b_{k q}, z_{q}, Y_{q}\right)\right|_{z_{1}=z_{2}=\cdots=0}}{\sum_{k=0}^{2^{n}-1} c_{k}(t) \phi\left(t, b_{k 1}, 0, Y_{1}\right) \cdots \phi\left(t, b_{k q}, 0, Y_{q}\right)}, \tag{15}
\end{align*}
$$

where the deterministic functions $c_{k}(t)$ and the $q$-vector $b_{k}$ of constants are determined by the relation (14), where $v$ is a $q$-vector of positive constants and the products $b_{k} v$ and $\alpha^{j} v$ are interpreted as dot products.
5.2. Frailty models. Doubly stochastic models ignore the impact of a default on the intensities of the surviving firms. This impact is channeled through the network of informational and contractual relationships in the economy. For instance, for U.S. corporate defaults, Duffie et al. [29] find strong evidence for the presence of frailty, or unobservable common or correlated risk factors. The uncertainty about the value of a frailty generates an additional channel of default correlation, above and beyond the "doubly stochastic channel." It also leads to additional dynamical effects in constituent intensities, in that a default causes a jump in the
intensities of any other firms that depend on the same frailty. These jump effects are due to Bayesian learning in the reference filtration $\mathbb{F}$.

We generalize the complete information model of Proposition 5.3 to include a firm's exposure to unobservable frailty risk factors. This extends the reach of the exact method to the model specifications in a substantial frailty literature, which includes [16], [22], [29], [34], [35], [50], [57], [62], and others.

Proposition 5.4. Suppose that, relative to a complete information filtration $\mathbb{H} \supset \mathbb{F}, N$ is doubly stochastic with intensities $X^{i}+\alpha^{i} \cdot Y$ for mutually independent nonnegative processes $X^{1}, \ldots, X^{n}$ and $Y=\left(Y_{1}, \ldots, Y_{q}\right)$ that are adapted to $\mathbb{H}$, and nonnegative factor loadings $\alpha^{i}=\left(\alpha_{1}^{i}, \ldots, \alpha_{q}^{i}\right)$. Assume that all but one risk factor $Y_{m}$ are adapted to the observation filtration $\mathbb{F} .^{7}$ Then $h^{i}(t, B)$ satisfies (15).

Proposition 5.4 states that the transition rates of the mimicking Markov chain in a model with an unobservable common risk factor $Y_{m}$ agree with those in the corresponding complete information doubly stochastic model. Then, by Proposition 3.1, the distributions of the default indicator $N_{t}$ must agree in these two model specifications. This may seem surprising at first: the specifications generate different intensity processes, so one would expect that they imply different distributions for $N_{t}$. However, while the intensities in the two models are different, they admit the same projections onto $N$. This is because the $\mathbb{F}$-intensity $\lambda^{i}$ in the frailty model is the optional projection onto $\mathbb{F}$ of the complete information $\mathbb{H}$-intensity $X^{i}+\alpha^{i} \cdot Y$. Now the conclusion follows from iterated expectations. The proof of Proposition 5.4 formalizes this intuition.

While the presence of frailty makes no difference for the unconditional distribution $P\left(N_{t}=\right.$ $B$ ) of the default indicator $N_{t}$, it is important to note that it does influence the conditional distributions $P\left(N_{t}=B \mid \mathcal{F}_{s}\right)$ for $t>s>0$. The reason is that the sigma-fields $\mathcal{F}_{s}$ representing the observable information available at time $s$ are different for frailty and complete information models. In the frailty model of Proposition 5.4, $\mathcal{F}_{s}$ does not contain the path of the frailty risk factor $Y_{s}$ over $[0, s]$, while in the complete information model of Proposition 5.3 it does.
5.3. Self-exciting models. The impact of a default on the intensities of the surviving firms can also be attributed to contagion, by which the distress of a firm is propagated to other firms. The authors of [4] find strong evidence for the presence of contagion in U.S. corporate defaults, after controlling for other channels of default correlation, including exposure to observable and unobservable risk factors. This empirical evidence can be addressed with a self-exciting model, in which the intensity of a firm responds to the default of another firm. Formulations of this type have been considered by [17], [20], [33], [36], [37], [45], [48], [52], [53], [55], [63], and others.

Proposition 5.5. Suppose that $N$ has intensities $\lambda^{i}=X^{i}+c^{i}(\cdot, N)$, where $X=\left(X^{1}, \ldots, X^{2}\right)$ solves $d X_{t}=\mu\left(t, X_{t}\right) d t+\sigma\left(t, X_{t}\right) d W_{t}+d J_{t}$ for a standard Brownian motion $W$, a point process $J$ with arrival intensity $\gamma\left(t, X_{t}\right)$, and fixed jump sizes and suitable functions ( $\mu, \sigma, \gamma$ ) such that the $X^{i}$ are nonnegative and independent, and where $c: \mathbb{R}_{+} \times\{0,1\}^{n} \rightarrow \mathbb{R}_{+}$is a

[^92]bounded function. Assume that $\phi\left(t,-1,0, X^{i}\right)$ is finite. Then, for $B^{i}=0$, we have
$$
h^{i}(t, B)=-\frac{\left.\partial_{z} \phi\left(t, 1, z, X^{i}\right)\right|_{z=0}}{\phi\left(t, 1,0, X^{i}\right)}+c^{i}(t, B) .
$$

The impact function $c^{i}$ specifies the dependence of firm $i$ 's intensity on the state of the other firms in the portfolio. By convention, $c^{i}(\cdot, B)=g^{i}\left(\cdot, B^{1}, \ldots, B^{i-1}, B^{i+1}, \ldots, B^{n}\right)$ for some bounded $g^{i}: \mathbb{R}_{+} \times\{0,1\}^{n-1} \rightarrow \mathbb{R}_{+}$. Examples of the impact function include $c^{i}(t, B)=$ $\sum_{j \neq i} \beta^{i j}(t) B^{j}$ for nonnegative, deterministic, and bounded functions $\beta^{i j}(t)$ that model the impact on firm $i$ of firm $j$ 's default. A parsimonious "mean-field" model is obtained by setting $\beta^{i j}(t)=\frac{1}{n-1}$. To allow for nonlinear dependence on events, we can specify a bounded nonnegative function $\varphi^{i}$ and take $c^{i}(t, B)=\varphi^{i}\left(\sum_{j \neq i} \beta^{i j}(t) B^{j}\right)$ for deterministic functions $\beta^{i j}(t)$ that are not required to be nonnegative.

The proof of Proposition 5.5 indicates that we can also treat an alternative multiplicative formulation $\lambda^{i}=X^{i} c^{i}(\cdot, N)$. In this case the impact function acts as a scaling to the "baseline hazard" $X^{i}$. We could take, for instance, $c^{i}(t, B)=\exp \left(\sum_{j \neq i} \beta^{i j}(t) B^{j}\right)$ for deterministic real-valued functions $\beta^{i j}(t)$.
6. Numerical results. This section demonstrates the utility of the exact method through numerical experiments. We consider a variant of the self-exciting model treated by Proposition 5.5.
6.1. Model. For nonnegative constants $\beta^{i j}$, consider the specification

$$
\begin{equation*}
\lambda_{t}^{i}=X_{t}^{i}+\sum_{j \neq i} \beta^{i j} N_{t}^{j} \tag{16}
\end{equation*}
$$

where the risk factors $X^{i}$ follow mutually independent Feller diffusions:

$$
\begin{equation*}
d X_{t}^{i}=\kappa_{i}\left(\theta_{i}-X_{t}^{i}\right) d t+\sigma_{i} \sqrt{X_{t}^{i}} d W_{t}^{i}, \quad X_{0}^{i}>0 . \tag{17}
\end{equation*}
$$

Here, $\kappa_{i}$ is a parameter controlling the speed of mean-reversion of $X^{i}, \theta_{i}$ is the level of mean reversion, and $\sigma_{i}$ controls the diffusive volatility of $X^{i}$. The process $\left(W^{1}, \ldots, W^{n}\right)$ is a standard Brownian motion. The parameter $\beta^{i j}$ determines the impact on firm $i$ of firm $j$ 's default. The corresponding jump terms generate correlation between the firm intensities. The matrix ( $\beta^{i j}$ ) governs the default dependence structure.

For each constituent firm $i=1, \ldots, n$, we initialize the risk factor $X_{0}^{i}$ at its long-run mean $\theta_{i}$. The parameters are selected randomly. We draw $\kappa_{i}$ from $U[0.5,1.5]$ and $\theta_{i}$ from $U[0.001,0.051]$. We take $\sigma_{i}=\min \left(\sqrt{2 \kappa_{i} \theta_{i}}, \bar{\sigma}_{i}\right)$, where $\bar{\sigma}_{i}$ is drawn from $U[0,0.2]$. We draw $\beta^{i j}$ from $U[0,0.01]$ for each $j=1, \ldots, n$. In practice, the parameters are calibrated from market rates of derivatives referenced on the constituent issuers and on the portfolio, as in [30] or [59].

The formulation (16)-(17) generalizes the specifications in [48], [53], and [63] to include a diffusion term that modulates the intensity of a firm between arrivals. In the absence of the diffusion term, the intensity is piecewise deterministic, so that $N$ can be simulated exactly using the classical $\mathrm{A} / \mathrm{R}$ scheme of [54]. The exact method developed here extends the reach of this scheme to the richer model (16)-(17).
6.2. Mimicking chain. Proposition 5.5 determines the rate $h^{i}(t, B)$ of the mimicking Markov chain in terms of the risk factor transform $\phi$, its partial derivative $\partial_{z} \phi$, and the parameters $\beta^{i j}$. The transform $\phi$ takes the well-known exponentially affine form

$$
\begin{equation*}
\phi\left(t, u, z, X^{i}\right)=\exp \left(a^{i}(t, u, z)+b^{i}(t, u, z) X_{0}^{i}\right), \tag{18}
\end{equation*}
$$

where, for $\gamma_{i}=\sqrt{\kappa_{i}^{2}+2 \sigma_{i}^{2} u}$, the coefficient functions

$$
\begin{align*}
a^{i}(t, u, z) & =\frac{2 u\left(1-\exp \left(\gamma_{i} t\right)\right)-z\left(\gamma_{i}+\kappa_{i}+\left(\gamma_{i}-\kappa_{i}\right) \exp \left(\gamma_{i} t\right)\right)}{\sigma_{i}^{2} z\left(\exp \left(\gamma_{i} t\right)-1\right)+\gamma_{i}-\kappa_{i}+\left(\gamma_{i}+\kappa_{i}\right) \exp \left(\gamma_{i} t\right)},  \tag{19}\\
b^{i}(t, u, z) & =\frac{2 \kappa_{i} \theta_{i}}{\sigma_{i}^{2}} \log \frac{2 \gamma_{i} \exp \left(\left(\gamma_{i}+\kappa_{i}\right) t / 2\right)}{\sigma_{i}^{2} z\left(\exp \left(\gamma_{i} t\right)-1\right)+\gamma_{i}-\kappa_{i}+\left(\gamma_{i}+\kappa_{i}\right) \exp \left(\gamma_{i} t\right)} . \tag{20}
\end{align*}
$$

The derivative $\partial_{z} \phi$ of $\phi$ is also available in closed form. Proposition 5.5 then implies, for $B^{i}=0$ and evaluating $\gamma_{i}$ at $u=1$, the formula

$$
h^{i}(t, B)=\frac{4 X_{0}^{i} \gamma_{i}^{2} \exp \left(\gamma_{i} t\right)}{\left(\gamma_{i}-\kappa_{i}+\left(\gamma_{i}+\kappa_{i}\right) \exp \left(\gamma_{i} t\right)\right)^{2}}-\frac{\theta_{i} \kappa_{i}}{\sigma_{i}^{2}} \frac{\left(\kappa_{i}^{2}-\gamma_{i}^{2}\right)\left(\exp \left(\gamma_{i} t\right)-1\right)}{\gamma_{i}-\kappa_{i}+\left(\gamma_{i}+\kappa_{i}\right) \exp \left(\gamma_{i} t\right)}+\sum_{j \neq i} \beta^{i j} B^{j}
$$

With parameter values selected as explained above, the function $h^{i}(t, \cdot)$ is decreasing. This suggests an adaptive rule for setting the bound $J(k)$ and the interval length $K(k)$ in step 3 of Algorithm 4.1. The first candidate time $S$ for $T_{k+1}$ is generated using $H\left(T_{k}, k\right)=$ $\sum_{i=1}^{n} h^{i}\left(T_{k}, M_{T_{k}}\right)$ as a bound, where the interval length is taken to be the time to the simulation horizon. If that time is rejected, we generate the next candidate time using the value $H(S, k)$ as a bound, taking the interval length to be the remaining time to the simulation horizon. The value $H(S, k)$ is computed in any case for the acceptance test. We proceed according to this rule until a candidate time is accepted.
6.3. Estimators. We contrast the estimators generated by the exact scheme with those generated by the time-scaling method described in section 2.2. The time-scaling method requires paths of the continuous-time stochastic processes $\int_{0}^{t} \lambda_{s}^{i} d s$ for $i=1, \ldots, n$, where the $\lambda^{i}$ follow jump-diffusion processes that are correlated through common jumps. We must discretize the time interval and simulate the joint integral process dynamics on this discretetime grid. Since the joint law of the integrals is not known, we first simulate the joint intensity dynamics on the discrete-time grid and then integrate. To generate the values of $\lambda^{i}$, we generate the values of $X^{i}$ by sampling from the noncentral chi-squared distribution that describes the transition law of $X^{i}$, and add the value $\beta^{i j}$ when another firm $j$ defaults. While the sampling from the chi-squared distribution leads to exact values of the $X^{i}$, it tends to be more time-consuming than an alternative Euler scheme. On the other hand, the Euler scheme introduces additional discretization bias because it does not facilitate exact sampling. To analyze the trade-off between computation time and bias, we implement both the exact sampling from the chi-squared transition law and the modified Euler scheme in [39, (3.66)].

To compare the estimators generated by the different simulation methods, we consider the root mean square error ( RMSE ), given by $\sqrt{\mathrm{SE}^{2}+\mathrm{Bias}^{2}}$. The standard error SE is estimated as the sample standard deviation of the simulation output divided by the square root of the

## Table 2

Simulation results under the self-exciting model (16)-(17) for $E\left[\left(C_{1}-3\right)^{+}\right]$. "Time-scaling $\left(\chi^{2}\right)$ " refers to the time-scaling method using the exact sampling of the values of $X^{i}$ from the noncentral transition law. "Time-scaling (Euler)" refers to the time-scaling method using the modified Euler scheme to sample the $X^{i}$. The true value was estimated to be 1.0145 , based on $5,000,000$ trials with the exact scheme.

| Method | Trials | Steps | Estimate | Bias | SE | RMSE | Time (min) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | 5 K | N/A | 1.0394 | 0 | 0.0239 | 0.0239 | 0.10 |
|  | 7.5 K | N/A | 1.0247 | 0 | 0.0193 | 0.0193 | 0.15 |
| Exact | 10 K | N/A | 0.9854 | 0 | 0.0165 | 0.0165 | 0.20 |
| A/R | 50 K | N/A | 1.0161 | 0 | 0.0073 | 0.0073 | 1.01 |
|  | 100 K | N/A | 1.0083 | 0 | 0.0052 | 0.0052 | 2.03 |
|  | 1 M | N/A | 1.0138 | 0 | 0.0016 | 0.0016 | 20.29 |
|  | 1 K | 32 | 0.8744 | 0.0893 | 0.0650 | 0.1104 | 21.20 |
| Time- | 2.5 K | 50 | 0.9400 | 0.0796 | 0.0420 | 0.0899 | 85.30 |
| scaling | 5 K | 71 | 1.0538 | 0.0735 | 0.0246 | 0.0775 | 257.30 |
| $\left(\chi^{2}\right)$ | 7.5 K | 87 | 1.0873 | 0.0697 | 0.0199 | 0.0725 | 483.50 |
|  | 10 K | 100 | 1.0805 | 0.0660 | 0.0171 | 0.0682 | 767.60 |
|  | 1 K | 32 | 0.8570 | 0.2094 | 0.0474 | 0.2147 | 4.83 |
| Time- | 2.5 K | 50 | 0.8448 | 0.1763 | 0.0277 | 0.1785 | 12.27 |
| scaling | 5 K | 71 | 0.9254 | 0.1324 | 0.0232 | 0.1344 | 36.57 |
| (Euler) | 7.5 K | 87 | 0.9394 | 0.1070 | 0.0193 | 0.1087 | 86.54 |
|  | 10 K | 100 | 0.9354 | 0.0860 | 0.0186 | 0.0880 | 194.27 |

number of trials. The bias is given by the difference between the expectation of the estimator and the true value. The bias of the estimator generated by the exact method is zero. The bias of the estimator generated by the time-scaling method with a specific number of time steps can be estimated using a large number of trials to estimate the expectation of the estimator, and then taking the difference with the true value, estimated using the exact method with a large number of trials.

We estimate the (undiscounted) value of a call $E\left[\left(C_{T}-K\right)^{+}\right]$on the default count at $T$. This option is a basic building block for the valuation of portfolio credit derivatives; see [31]. We take the number of reference names $n=100$, which is the standard portfolio size for many traded portfolio derivatives, $T=1$ year, and $K=3$. Table 2 reports the simulation results. The bias in the table is estimated using 50,000 trials. The number of discretization time steps in the time-scaling method is set equal to the square-root of the number of simulation trials. ${ }^{8}$ The experiments were performed on a desktop PC with an Intel 3.4 GHz processor and 1 GB of RAM, running Windows XP Professional. The methods were implemented in Matlab.

Figure 1, which shows the convergence of the RMSE graphically, indicates the substantial performance advantages of the exact method. The exact method requires the shortest computation time to achieve a given accuracy. It also has the fastest convergence rate. This rate is of order $O(1 / \sqrt{t})$, where $t$ is the computation time, the optimal rate of unbiased schemes.

[^93]

Figure 1. Convergence of the RMSEs of the exact $A / R$ and time-scaling methods under the self-exciting model (16)-(17) for $E\left[\left(C_{1}-3\right)^{+}\right]$.

There is an alternative exact scheme for the model (16)-(17). This scheme is based on the repeated application of the first-to-default time and identity simulation algorithm developed by [24]. This scheme requires greater computational effort than the A/R scheme, which is easily explained in case $n=2$. In the first-to-default scheme, one starts by generating $T_{1}=\tau^{1} \wedge \tau^{2}$ by the inverse transform method from $P\left(T_{1}>t\right)=\phi\left(t, 1,0, X^{1}\right) \phi\left(t, 1,0, X^{2}\right)$. Given $T_{1}$, one then samples the identity of the first defaulter from the discrete distribution defined by $\gamma\left(T_{1}, i\right) /\left(\gamma\left(T_{1}, 1\right)+\gamma\left(T_{1}, 2\right)\right)$ for $i=1,2$. Here $\gamma(t, 1)=-\left.\phi\left(t, 1,0, X^{2}\right) \partial_{z} \phi\left(t, 1, z, X^{1}\right)\right|_{z=0}$; a similar expression holds for $\gamma(t, 2)$. This second step is similar to step 5 in Algorithm 4.1. In a third step, one draws the second default time $T_{2}$. To do this, note that, given $\mathcal{F}_{T_{1}}$, the time $T_{2}-T_{1}$ is equal in law to the first jump time of a doubly stochastic Poisson process started at $T_{1}$, with $\mathbb{F}$-intensity $\left(X_{t}^{2}+\beta^{21}\right)_{t \geq T_{1}}$ (this assumes $T_{1}=\tau^{1}$ ). Thus one needs to sample from $P\left(T_{2}-T_{1}>t \mid \mathcal{F}_{T_{1}}\right)=E\left(\exp \left(-\int_{T_{1}}^{T_{1}+t}\left(X_{s}^{2}+\beta^{21}\right) d s\right) \mid X_{T_{1}}^{2}, T_{1}\right)$. This can be done by the inverse method, once $X_{T_{1}}^{2}$ is drawn from the conditional distribution of $X_{t}^{2}$ given $T_{1}=\tau^{1}=t$. However, as shown by [24], this conditional distribution is known only in terms of its transform, making it relatively costly to sample from it. The A/R scheme avoids this third step and also tends to require less computational effort to generate the $T_{i}$.
6.4. Variance reduction. We demonstrate the effectiveness of the S/M Algorithm 4.2 for variance reduction under the self-exciting model (16)-(17). We estimate $P\left(C_{T}=k\right)$ for the test portfolio described in section 6.1. To measure the variance reduction offered by Algorithm 4.2 , we compute variance ratios for each of several values of $k$. Each variance ratio is calculated
by estimating the variance of the estimator generated by the plain $A / R$ Algorithm 4.1 and dividing it by the estimated variance of the estimator generated by Algorithm 4.2. The variance is estimated by the sample variance of the simulation output. ${ }^{9}$

To run Algorithm 4.2, we need to select the number of particles $R$, the number of selections $m$ performed during $[0, T]$, and the parameter $\delta$ of the selection probability (11). The number of particles $R$ is analogous to the number of trials in a standard simulation. In practice, this quantity is determined by the desired accuracy of the estimator, or the available computational budget, for fixed $m$. However, the theoretically optimal allocation of the computational budget between $m$ and $R$ has not yet been worked out to our knowledge. Therefore, we will illustrate the influence of $R$ and $m$ on the estimator through experiments, treating each variable separately.

Also the choice of $\delta$ is difficult. Intuitively, we want to pick $\delta$ so as to minimize the relative error of the estimator, given by the variance of the estimator divided by the product of the estimator and the square root of the number of simulation trials. Again, the theoretically optimal choice of $\delta$ has not yet been worked out, to our knowledge. Therefore, we approximate the optimal $\delta$ through experiments. Specifically, we discretize a range of values of $\delta$, run the simulation for each grid value using a small number of particles, and select the value of $\delta$ that produces the smallest relative error. The $\delta$ so chosen increases with the target event count $k$, because the selections must place greater weight on particles with transitions, the smaller the probability of interest.

Table 3 shows the results for $T=1, m=4$ selections, and $R=10,000$ particles. Table 4 reports the results when only $R=1,000$ particles are used. To provide a meaningful comparison between the two methods, the number of trials in the estimation using the plain $\mathrm{A} / \mathrm{R}$ Algorithm 4.1 is chosen such that the total time required to estimate $P\left(C_{T}=k\right)$ is approximately the same as that required by the $\mathrm{S} / \mathrm{M}$ Algorithm 4.2 for the given $R$ and $m$ (excluding the fixed time it takes to select $\delta$ ). We see that the smaller the probability, the larger the variance ratio. Further, for sufficiently small probabilities, the smaller the number of particles $R$, the higher the variance ratio. Figure 2 graphs the estimated probabilities $P\left(C_{1}=k\right)$ reported in Tables 3 and 4. It indicates the relative benefits of the $\mathrm{S} / \mathrm{M}$ scheme for each value of $R$.

Next we analyze the role of the number of selections $m$. Figure 3 graphs the estimated variance ratios for each of several values of $m$, fixing the number of particles $R=1,000$. The variance ratio increases with the number of selections if the probability of interest is only moderately small. For smaller event probabilities, this may not be the case anymore. The intuition is as follows. The computation time required by the $\mathrm{S} / \mathrm{M}$ scheme increases with $m$. The increase in computation time is relatively larger the larger $k$ becomes, i.e., the smaller the probability of interest. Now, by the design of our experiments, the number of trials that can be completed by the plain $\mathrm{A} / \mathrm{R}$ scheme increases with $m$ and $k$, so that the absolute variance of the corresponding estimator may decrease faster than the variance of the estimator generated by the $\mathrm{S} / \mathrm{M}$ scheme. However, note that, for a fixed number of particles, increasing $m$ always reduces the variance of the simulation estimator in absolute terms. This

[^94]Table 3
Variance reduction ratios for estimating $P\left(C_{1}=k\right)$ under the self-exciting model (16)-(17). In the $S / M$ Algorithm 4.2, $m=4$, and $\delta$ is chosen such that the relative error of the simulation estimator is minimized. The number of trials in the estimation using the plain $A / R$ Algorithm 4.1 is chosen such that the total time required to estimate $P\left(C_{T}=k\right)$ is approximately the same as that required by the $S / M$ scheme, excluding the fixed time it takes to select $\delta .\left(^{*}\right)$ indicates that the event of interest was not observed in any of the trials.

|  | S/M scheme |  |  |  |  | Plain A/R scheme |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $k$ | $\delta$ | Particles | Estimate | SE | Trials | Estimate | VarRatio |  |
| 8 | 0.55 | 10,000 | 0.02364596 | 0.00096165 | 15,806 | 0.02240 | 2.54 |  |
| 9 | 0.7 | 10,000 | 0.01253552 | 0.00065596 | 16,452 | 0.01349 | 3.08 |  |
| 10 | 0.75 | 10,000 | 0.00587429 | 0.00044357 | 16,452 | 0.00626 | 3.16 |  |
| 11 | 0.8 | 10,000 | 0.00337833 | 0.00020285 | 17,742 | 0.00293 | 7.10 |  |
| 12 | 0.8 | 10,000 | 0.00162340 | 0.00014706 | 17,742 | 0.00220 | 10.14 |  |
| 13 | 0.85 | 10,000 | 0.00068818 | 0.00007336 | 18,065 | 0.00066 | 12.33 |  |
| 14 | 0.85 | 10,000 | 0.00029433 | 0.00002062 | 18,387 | 0.00027 | 63.88 |  |
| 15 | 1.05 | 10,000 | 0.00016310 | 0.00001609 | 19,032 | 0.00011 | 121.80 |  |
| 16 | 1.05 | 10,000 | 0.00006790 | 0.00000471 | 19,032 | 0.00005 | 236.92 |  |
| 17 | 1.15 | 10,000 | 0.00002597 | 0.00000352 | 19,355 | $\left(^{*}\right)$ |  |  |
| 18 | 1.15 | 10,000 | 0.00000970 | 0.00000171 | 19,355 | $\left.*^{*}\right)$ |  |  |
| 19 | 1.15 | 10,000 | 0.00000500 | 0.00000054 | 19,355 | $\left.*^{*}\right)$ |  |  |
| 20 | 1.15 | 10,000 | 0.00000203 | 0.00000020 | 19,355 | $\left.*^{*}\right)$ |  |  |
| 21 | 1.15 | 10,000 | 0.00000106 | 0.00000008 | 19,355 | $\left.*^{*}\right)$ |  |  |
| 22 | 1.3 | 10,000 | 0.00000039 | 0.00000005 | 19,677 | $\left.*^{*}\right)$ |  |  |

## Table 4

Variance reduction ratios for estimating $P\left(C_{1}=k\right)$ under the self-exciting model (16)-(17). In the $S / M$ Algorithm 4.2, $m=4$, and $\delta$ is chosen such that the relative error of the simulation estimator is minimized. The number of trials in the estimation using the plain $A / R$ Algorithm 4.1 is chosen such that the total time required to estimate $P\left(C_{1}=k\right)$ is approximately the same as that required by the $S / M$ scheme, excluding the fixed time it takes to select $\delta .\left(^{*}\right)$ indicates that the event of interest was not observed in any of the trials.

|  | S/M scheme |  |  |  |  | Plain A/R scheme |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $k$ | $\delta$ | Particles | Estimate | SE | Trials | Estimate | VarRatio |  |
| 8 | 0.55 | 1,000 | 0.02433732 | 0.00310121 | 1,343 | 0.02010 | 2.05 |  |
| 9 | 0.7 | 1,000 | 0.01110821 | 0.00156666 | 1,407 | 0.00498 | 2.02 |  |
| 10 | 0.75 | 1,000 | 0.00738774 | 0.00108022 | 1,439 | 0.00625 | 5.33 |  |
| 11 | 0.8 | 1,000 | 0.00340511 | 0.00083321 | 1,600 | 0.00500 | 7.16 |  |
| 12 | 0.8 | 1,000 | 0.00156213 | 0.00029693 | 1,600 | 0.00125 | 14.16 |  |
| 13 | 0.85 | 1,000 | 0.00073482 | 0.00019565 | 1,600 | 0.00125 | 32.62 |  |
| 14 | 0.85 | 1,000 | 0.00024352 | 0.00005751 | 1,600 | 0.00188 | 565.20 |  |
| 15 | 1.05 | 1,000 | 0.00009061 | 0.00001921 | 1,726 | 0.00058 | 1562.75 |  |
| 16 | 1.05 | 1,000 | 0.00009384 | 0.00001385 | 1,759 | 0.00057 | 2951.99 |  |
| 17 | 1.15 | 1,000 | 0.00006344 | 0.00000372 | 1,790 | $\left(^{*}\right)$ |  |  |
| 18 | 1.15 | 1,000 | 0.00002131 | 0.00000152 | 1,823 | $\left.*^{*}\right)$ |  |  |
| 19 | 1.15 | 1,000 | 0.00000971 | 0.00000091 | 1,887 | $\left.*^{*}\right)$ |  |  |
| 20 | 1.15 | 1,000 | 0.00000041 | 0.00000014 | 1,887 | $\left.*^{*}\right)$ |  |  |
| 21 | 1.15 | 1,000 | 0.00000082 | 0.00000008 | 1,918 | $\left.*^{*}\right)$ |  |  |
| 22 | 1.3 | 1,000 | 0.00000031 | 0.00000004 | 1,983 | $\left.*^{*}\right)$ |  |  |

is indicated by Figure 4, which graphs the estimated probabilities $P\left(C_{1}=k\right)$ for each of several values of $m$, for fixed $R=1,000$.


Figure 2. Estimated probabilities $P\left(C_{1}=k\right)$ under the self-exciting model (16)-(17) for each of several values of $R$. The number of selections $m=4$.


Figure 3. Estimated variance reduction ratios under the self-exciting model (16)-(17) for each of several values of $m$. The number of particles $R=1,000$.


Figure 4. Estimated probabilities $P\left(C_{1}=k\right)$ under the self-exciting model (16)-(17) for each of several values of $m$. The number of particles $R=1,000$.
6.5. Potential extensions. There are several potential variations and extensions of the formulation (16)-(17). As explained in section 5.3, the additive specification of the feedback term in the intensity dynamics (16) could be replaced by a multiplicative specification. This would lead to different self-exciting dynamics and a greater degree of flexibility in designing the feedback behavior, without reducing the analyticity of $h^{i}$. The Feller diffusion risk factor dynamics (17) could be extended to include a compound Poisson jump term. This extension would allow for discontinuous movements of the intensity between defaults, while requiring only a minor modification of the coefficient functions (20) and (19) based on the results of [25]. The rate $h^{i}$ would still take a closed form. More generally, the Feller diffusion dynamics (17) could be replaced by more general affine jump-diffusion dynamics. The transform (18) would remain exponentially affine in the state, while the coefficient functions would be given as solutions to a system of ODEs.
7. Conclusion. This paper develops a simulation method for dynamic intensity-based models of correlated default risk. The method generates unbiased estimators of credit portfolio loss distributions, risk measures, and prices of derivative securities that are referenced on a portfolio of defaultable assets. It reduces the simulation problem to one of a simple Markov chain expectation. This problem can be treated with exact methods. An overarching selection/mutation scheme reduces variance in rare-event situations. The method is widely applicable to many intensity models in the literature. Numerical experiments demonstrate its effectiveness and highlight its advantages over alternative methods.

The simulation method has potential applications in other areas that deal with the arrival of correlated events. These include, in particular, applications in reliability, where intensitybased models have long been used to analyze the reliability of systems of interdependent components whose failure times are correlated.

## Appendix. Proofs.

Proof of Proposition 3.1. For all $B \in\{0,1\}^{n}$ and $t>0$, we have

$$
\begin{align*}
I\left(N_{t}=B\right) & =I\left(N_{0}=B\right)+\sum_{0<s \leq t}\left[I\left(N_{s}=B\right)-I\left(N_{s-}=B\right)\right] \\
& =I\left(N_{0}=B\right)+\sum_{i=1}^{n} \int_{0}^{t}\left[I\left(N_{s}=B\right)-I\left(N_{s-}=B\right)\right] d N_{s}^{i} \\
& =I\left(N_{0}=B\right)+\sum_{i=1}^{n} \int_{0}^{t}\left[I\left(N_{s-}+e^{i}=B\right) I\left(B^{i}=1\right)-I\left(N_{s-}=B\right)\right] d N_{s}^{i}, \tag{21}
\end{align*}
$$

where $e^{i}$ is an $n$-vector with $i$ th component equal to 1 and the rest equal to 0 . Next we recall the Doob-Meyer decomposition (1) of the submartingale $N^{i}$ into the sum of a martingale $M^{i}$ and a process $\int_{0}^{i} \lambda_{s}^{i}\left(1-N_{s}^{i}\right) d s$. Since the integrand in the integral

$$
\int_{0}^{t}\left[I\left(N_{s-}+e^{i}=B\right) I\left(B^{i}=1\right)-I\left(N_{s-}=B\right)\right] d M_{s}^{i}
$$

is bounded and predictable, the integral defines a martingale with initial value equal to zero. Thus, after we take expectation on both sides of (21) and apply Fubini's theorem, we obtain

$$
P\left(N_{t}=B\right)=P\left(N_{0}=B\right)+\sum_{i=1}^{n} \int_{0}^{t} E\left(\left[I\left(N_{s}=B-e^{i}\right) I\left(B^{i}=1\right)-I\left(N_{s}=B\right)\right] \lambda_{s}^{i}\left(1-N_{s}^{i}\right)\right) d s
$$

Now we differentiate both sides of this equation with respect to $t$. By the definition of the deterministic functions $h^{i}(t, B)$, we obtain

$$
\partial_{t} P\left(N_{t}=B\right)=\sum_{i=1}^{n} I\left(B^{i}=1\right) P\left(N_{t}=B-e^{i}\right) h^{i}\left(t, B-e^{i}\right)-\sum_{i=1}^{n} P\left(N_{t}=B\right) h^{i}(t, B) .
$$

But this equation coincides with the backward Kolmogorov equation,

$$
\partial_{t} P\left(M_{t}=B\right)=\sum_{i=1}^{n} I\left(B^{i}=1\right) P\left(M_{t}=B-e^{i}\right) h^{i}\left(t, B-e^{i}\right)-\sum_{i=1}^{n} P\left(M_{t}=B\right) h^{i}(t, B),
$$

which describes the time evolution of the distribution of the Markov chain $M$ with transition rates $h^{i}(t, B)$. Thus, the probabilities $P\left(M_{t}=B\right)$ and $P\left(N_{t}=B\right)$ satisfy the same ODE. Since the solution to this ODE is unique and $M_{0}=N_{0}=0_{n}$ by construction, we conclude that $M_{t}$ and $N_{t}$ must have the same distribution for all $t \geq 0$.

Proof of Proposition 5.1. By the independence of the random variables $\tau^{i}$, we get

$$
h^{i}(t, B)=E\left(X_{t}^{i}\left(1-N_{t}^{i}\right) \mid N_{t}^{i}=B^{i}\right)=\left(1-B^{i}\right) E\left(X_{t}^{i} \mid \tau^{i}>t\right) .
$$

By Bayes' formula and iterated expectations,

$$
\begin{equation*}
E\left(X_{t}^{i} \mid \tau^{i}>t\right)=\frac{E\left(X_{t}^{i} I\left(\tau^{i}>t\right)\right)}{P\left(\tau^{i}>t\right)}=\frac{E\left(X_{t}^{i} P\left(\tau^{i}>t \mid\left(X_{s}^{i}\right)_{s \leq t}\right)\right)}{E\left(P\left(\tau^{i}>t \mid\left(X_{s}^{i}\right)_{s \leq t}\right)\right)} . \tag{22}
\end{equation*}
$$

The doubly stochastic property of $\tau^{i}$ implies that

$$
P\left(\tau^{i}>t \mid\left(X_{s}^{i}\right)_{s \leq t}\right)=\exp \left(-\int_{0}^{t} X_{s}^{i} d s\right)
$$

whose expectation equals $\phi\left(t, 1,0, X^{i}\right)$. This gives the denominator of the right-hand side of (22). For the numerator,

$$
E\left(X_{t}^{i} P\left(\tau^{i}>t \mid\left(X_{s}^{i}\right)_{s \leq t}\right)\right)=E\left(X_{t}^{i} \exp \left(-\int_{0}^{t} X_{s}^{i} d s\right)\right)=-\left.\partial_{z} \phi\left(t, 1, z, X^{i}\right)\right|_{z=0}
$$

and this completes the proof.
Proof of Proposition 5.2. By the independence of the idiosyncratic factors $X^{i}$,

$$
h^{i}(t, B)=\left(1-B^{i}\right) E\left(X_{t}^{i} \mid \tau^{i}>t\right)+\left(1-B^{i}\right) \alpha^{i} E\left(Y_{t} \mid N_{t}=B\right),
$$

where the first summand is treated as in the proof of Proposition 5.1. It remains to calculate the second summand. By Bayes' formula and iterated expectations,

$$
E\left(Y_{t} \mid N_{t}=B\right)=\frac{E\left(Y_{t} I\left(N_{t}=B\right)\right)}{P\left(N_{t}=B\right)}=\frac{E\left(Y_{t} P\left(N_{t}=B \mid\left(Y_{s}\right)_{s \leq t}\right)\right)}{E\left(P\left(N_{t}=B \mid\left(Y_{s}\right)_{s \leq t}\right)\right)} .
$$

Given a path of the common factor $Y$, the intensities $\lambda^{i}$ are independent of one another, and so are the components $N^{i}$ of the process $N$. Thus,

$$
\begin{align*}
P\left(N_{t}=B \mid\left(Y_{s}\right)_{s \leq t}\right) & =\prod_{j=1}^{n} P\left(N_{t}^{j}=B^{j} \mid\left(Y_{s}\right)_{s \leq t}\right) \\
& =\prod_{j=1}^{n}\left[B^{j}-\left(2 B^{j}-1\right) P\left(\tau^{j}>t \mid\left(Y_{s}\right)_{s \leq t}\right)\right] . \tag{23}
\end{align*}
$$

By iterated expectations, the doubly stochastic property, and the independence of the processes $Y$ and $X^{j}$, we get

$$
\begin{align*}
P\left(\tau^{j}>t \mid\left(Y_{s}\right)_{s \leq t}\right) & =E\left[\exp \left(-\int_{0}^{t}\left(\alpha^{j} Y_{s}+X_{s}^{j}\right) d s\right) \mid\left(Y_{s}\right)_{s \leq t}\right] \\
& =\exp \left(-\alpha^{j} \int_{0}^{t} Y_{s} d s\right) E\left[\exp \left(-\int_{0}^{t} X_{s}^{j} d s\right)\right] \\
& =\exp \left(-\alpha^{j} \int_{0}^{t} Y_{s} d s\right) \phi\left(t, 1,0, X^{j}\right) \tag{24}
\end{align*}
$$

Now, expanding the $n$-fold product on the right-hand side of (23) and using (24), we see that there are deterministic functions $c_{k}(t)$ and constants $b_{k}$ satisfying (14) such that

$$
P\left(N_{t}=B \mid\left(Y_{s}\right)_{s \leq t}\right)=\sum_{k=0}^{2^{n}-1} c_{k}(t) \exp \left(-b_{k} \int_{0}^{t} Y_{s} d s\right) .
$$

Taking expectation on both sides of this equation leads to

$$
P\left(N_{t}=B\right)=\sum_{k=0}^{2^{n}-1} c_{k}(t) \phi\left(t, b_{k}, 0, Y\right) .
$$

A similar argument is applied to $E\left(Y_{t} P\left(N_{t}=B \mid\left(Y_{s}\right)_{s \leq t}\right)\right)$.
Proof of Proposition 5.3. Apply the argument used in the proof of Proposition 5.2.
Proof of Proposition 5.4. For clarity in the exposition, we consider the case $q=2$ and suppose that the observation filtration $\mathbb{F}$ is the right-continuous and complete filtration generated by the processes $X^{1}, \ldots, X^{n}, N$, and $Y_{1}$. Thus, the common factor $Y_{1}$ is observable (adapted to the filtration $\mathbb{F}$ ), while the common factor $Y_{2}$ is a frailty (not adapted to $\mathbb{F}$ ). An intensity $\lambda^{i}$ of $N^{i}$ relative to the observation filtration $\mathbb{F}$ is given by the optional projection (see [61, Chapter VI, p. 375]) of the complete information intensity $X^{i}+\left(\alpha_{1}^{i}, \alpha_{2}^{i}\right) \cdot\left(Y_{1}, Y_{2}\right)$ onto $\mathbb{F}$. Since $X^{i}$ and $Y_{1}$ are adapted to $\mathbb{F}$, we have

$$
\lambda_{t}^{i}=X_{t}^{i}+\alpha_{1}^{i} Y_{1 t}+\alpha_{2}^{i} E\left(Y_{2 t} \mid \mathcal{F}_{t}\right)
$$

almost surely, for each $t \geq 0$. Then

$$
h^{i}(t, B)=\left(1-B^{i}\right)\left\{E\left(X_{t}^{i} \mid N_{t}=B\right)+\alpha_{1}^{i} E\left(Y_{1 t} \mid N_{t}=B\right)+\alpha_{2}^{i} E\left(U_{t} \mid N_{t}=B\right)\right\}
$$

where $U_{t}=E\left(Y_{2 t} \mid \mathcal{F}_{t}\right)$. Since the $X^{i}$ are independent of one another, the first expectation on the right-hand side of this equation can be analyzed as in the proof of Proposition 5.1. The second expectation is treated as in the proof of Proposition 5.2, by conditioning on the path of $Y=\left(Y_{1}, Y_{2}\right)$ over $[0, t]$ and using the doubly stochastic property of $N$ in the complete information filtration. The third expectation can be calculated by an analogous conditioning argument by noting that

$$
E\left(U_{t} \mid N_{t}=B\right)=E\left(E\left(Y_{2 t} \mid \mathcal{F}_{t}\right) \mid N_{t}=B\right)=E\left(Y_{2 t} \mid N_{t}=B\right)
$$

since $\sigma\left(N_{t}\right) \subset \mathcal{F}_{t}$. The sum of these three expectations gives (15) for $q=2$.
Proof of Proposition 5.5. We have

$$
h^{i}(t, B)=\left(1-B^{i}\right)\left\{E\left(X_{t}^{i} \mid N_{t}=B\right)+c^{i}(t, B)\right\} .
$$

To calculate the conditional expectation, we apply a measure change argument developed by [17]. For clarity in the exposition, we consider the case $n=2$ and take $c^{i}\left(\cdot, 0_{2}\right)=0$. The general case can be treated by the same argument. We have

$$
E\left(X_{t}^{i} \mid N_{t}=0_{2}\right)=\frac{E\left(X_{t}^{i} I\left(N_{t}=0_{2}\right)\right)}{P\left(N_{t}=0_{2}\right)}=\frac{E^{*}\left(X_{t}^{i} \exp \left(-\int_{0}^{t}\left(\lambda_{s}^{1}+\lambda_{s}^{2}\right) d s\right)\right)}{E^{*}\left(\exp \left(-\int_{0}^{t}\left(\lambda_{s}^{1}+\lambda_{s}^{2}\right) d s\right)\right)}
$$

where $E^{*}$ denotes the expectation operator relative to the absolutely continuous probability measure $P^{*}$ on $\mathcal{F}_{t}$ defined by the density

$$
\left.\frac{d P^{*}}{d P}=I\left(N_{t}=0_{2}\right) \exp \left(\int_{0}^{t}\left(\lambda_{s}^{1}+\lambda_{s}^{2}\right) d s\right)\right) .
$$

The condition $\phi\left(t,-1,0, X^{i}\right)<\infty$ guarantees that $P^{*}$ is well-defined. Under $P^{*}$, the event $\left\{N_{t}=0_{2}\right\}$ has measure 1. Girsanov's theorem for absolutely continuous measure changes as in [61, Chapter III.8, Theorem 41] along with Lévy's theorem imply that the standard $P$ Brownian motion $W$ driving the $X^{i}$ 's remains a standard Brownian motion under $P^{*}$ on $[0, t]$, relative to the filtration $\mathbb{F}$ augmented by the $P^{*}$-null sets. The intensity of the point process $J$ remains $\gamma\left(t, X_{t}\right)$ relative to $P^{*}$ and the augmented filtration, because the components of $J$ do not have jumps in common with the components of $N$ almost surely. Therefore, the dynamics of the $X^{i}$ are invariant under the measure change. Taking $i=1$, it follows that

$$
\begin{aligned}
E^{*}\left(X_{t}^{1} \exp \left(-\int_{0}^{t}\left(\lambda_{s}^{1}+\lambda_{s}^{2}\right) d s\right)\right) & =E\left(X_{t}^{1} \exp \left(-\int_{0}^{t} X_{s}^{1} d s\right)\right) E\left(\exp \left(-\int_{0}^{t} X_{s}^{2} d s\right)\right) \\
& =-\left.\partial_{z} \phi\left(t, 1, z, X^{1}\right)\right|_{z=0} \phi\left(t, 1,0, X^{2}\right)
\end{aligned}
$$

An analogous expression holds for $i=2$. Similarly,

$$
E^{*}\left(\exp \left(-\int_{0}^{t}\left(\lambda_{s}^{1}+\lambda_{s}^{2}\right) d s\right)\right)=\phi\left(t, 1, z, X^{1}\right) \phi\left(t, 1,0, X^{2}\right)
$$

implying that

$$
\begin{equation*}
E\left(X_{t}^{i} \mid N_{t}=0_{2}\right)=-\frac{\left.\partial_{z} \phi\left(t, 1, z, X^{i}\right)\right|_{z=0}}{\phi\left(t, 1,0, X^{i}\right)} . \tag{25}
\end{equation*}
$$

Next, we use a similar argument to calculate

$$
E\left(X_{t}^{i} \mid N_{t}^{i}=0\right)=\frac{E\left(X_{t}^{i} I\left(N_{t}^{i}=0\right)\right)}{P\left(N_{t}^{i}=0\right)}=\frac{E^{i}\left(X_{t}^{i} \exp \left(-\int_{0}^{t} \lambda_{s}^{i} d s\right)\right)}{E^{i}\left(\exp \left(-\int_{0}^{t} \lambda_{s}^{i} d s\right)\right)}
$$

where $E^{i}$ denotes the expectation operator relative to the absolutely continuous probability measure $P^{i}$ on $\mathcal{F}_{t}$ defined by the density

$$
\frac{d P^{i}}{d P}=I\left(N_{t}^{i}=0\right) \exp \left(\int_{0}^{t} \lambda_{s}^{i} d s\right)
$$

Under $P^{i}$, the event $\left\{N_{t}^{i}=0\right\}$ has measure 1. As reasoned in the case of $P^{*}$, Girsanov's theorem implies that the dynamics of $\left(X^{1}, X^{2}\right)$ are invariant under the measure change. Taking $i=1$ and recalling that $c^{1}\left(s, N_{s}\right)$ takes the form $g^{1}\left(s, N_{s}^{2}\right)$,

$$
\begin{aligned}
E^{1}\left(X_{t}^{1} \exp \left(-\int_{0}^{t} \lambda_{s}^{1} d s\right)\right) & =E^{1}\left(X_{t}^{1} \exp \left(-\int_{0}^{t} X_{s}^{1} d s-\int_{0}^{t} c^{1}\left(s, N_{s}\right) d s\right)\right) \\
& =-\left.\partial_{z} \phi\left(t, 1, z, X^{1}\right)\right|_{z=0} E^{1}\left(\exp \left(-\int_{0}^{t} g^{1}\left(s, N_{s}^{2}\right) d s\right)\right) .
\end{aligned}
$$

An analogous expression holds for $i=2$. Applying a similar argument to the expectation $E^{i}\left(\exp \left(-\int_{0}^{t} \lambda_{s}^{i} d s\right)\right)$, we find that $E\left(X_{t}^{i} \mid N_{t}^{i}=0\right)=E\left(X_{t}^{i} \mid N_{t}=0_{2}\right)$. Bayes' rule, along with this relation, then shows that also $E\left(X_{t}^{1} \mid N_{t}=(0,1)\right)=E\left(X_{t}^{1} \mid N_{t}^{1}=0\right)$ and $E\left(X_{t}^{2} \mid N_{t}=\right.$ $(1,0))=E\left(X_{t}^{2} \mid N_{t}^{2}=0\right)$. Thus, we have shown that

$$
E\left(X_{t}^{i} \mid N_{t}=B\right)=-\frac{\left.\partial_{z} \phi\left(t, 1, z, X^{i}\right)\right|_{z=0}}{\phi\left(t, 1,0, X^{i}\right)}
$$

for all $B \in\{0,1\}^{2}$ with $B^{i}=0$.
Acknowledgments. We thank seminar participants for their comments, and Apaar Sadhwani for research assistance. We also thank Jose Blanchet, Rama Cont, Jean-Pierre Fouque, Josselin Garnier, Paul Glasserman, Martin Haugh, Baeho Kim, and Steve Kou for comments. We are grateful to two anonymous referees for their insightful reviews.

## REFERENCES

[1] L. Andersen, J. Sidenius, and S. Basu, All your hedges in one basket, Risk, 16 (2003), pp. 67-72.
[2] M. Arnsdorf and I. Halperin, BSLP: Markovian bivariate spread-loss model for portfolio credit derivatives, J. Comput. Finance, 12 (2008), pp. 77-100.
[3] S. Asmussen and P. Glynn, Stochastic Simulation: Algorithms and Analysis, Springer-Verlag, New York, Berlin, 2007.
[4] S. Azizpour, K. Giesecke, and G. Schwenkler, Exploring the Sources of Default Clustering, working paper, Department of Management Science and Engineering, Stanford University, Palo Alto, CA, 2008.
[5] A. Bassamboo and S. Jain, Efficient importance sampling for reduced form models in credit risk, in Proceedings of the 2006 Winter Simulation Conference, L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds., IEEE Press, Piscataway, NJ, 2006, pp. 741-748.
[6] A. Bassamboo, S. Juneja, and A. Zeevi, Portfolio credit risk with extremal dependence: Asymptotic analysis and efficient simulation, Oper. Res., 56 (2008), pp. 593-606.
[7] A. Bentata and R. Cont, Mimicking the Marginal Distributions of a Semimartingale, working paper, Columbia University, New York, 2009.
[8] P. Brémaud, Point Processes and Queues-Martingale Dynamics, Springer-Verlag, New York, 1980.
[9] M. Broadie and O. Kaya, Exact simulation of stochastic volatility and other affine jump diffusion processes, Oper. Res., 54 (2006), pp. 217-231.
[10] R. Carmona and S. Crepey, Particle methods for the estimation of Markovian credit portfolio loss distributions, Int. J. Theor. Appl. Finance, (2010), to appear.
[11] R. Carmona, J.-P. Fouque, and D. Vestal, Interacting particle systems for the computation of rare credit portfolio losses, Finance Stoch., 13 (2009), pp. 613-633.
[12] A. Chapovsky, A. Rennie, and P. Tavares, Stochastic intensity modeling for structured credit exotics, Int. J. Theor. Appl. Finance, 10 (2007), pp. 633-652.
[13] S. Chava and R. Jarrow, Bankruptcy prediction with industry effects, Rev. Finance, 8 (2004), pp. 537-569.
[14] Z. Chen and P. Glasserman, Fast pricing of basket default swaps, Oper. Res., 56 (2008), pp. 286-303.
[15] Z. Chen and P. Glasserman, Sensitivity estimates for portfolio credit derivatives using Monte Carlo, Finance Stoch., 12 (2008), pp. 507-540.
[16] P. Collin-Dufresne, R. Goldstein, and J. Helwege, How Large Can Jump-to-Default Risk Premia Be? Modeling Contagion via the Updating of Beliefs, working paper, Carnegie Mellon University, Pittsburgh, PA, 2009.
[17] P. Collin-Dufresne, R. Goldstein, and J. Hugonnier, A general formula for the valuation of defaultable securities, Econometrica, 72 (2004), pp. 1377-1407.
[18] R. Cont and A. Minca, Extracting Portfolio Default Rates from CDO Spreads, working paper, Columbia University, New York, 2008.
[19] S. Das, D. Duffie, N. Kapadia, and L. Saita, Common failings: How corporate defaults are correlated, J. Finance, 62 (2007), pp. 93-117.
[20] M. Davis and V. Lo, Modeling default correlation in bond portfolios, in Mastering Risk, Volume 2: Applications, C. Alexander, ed., Prentice-Hall, Englewood Cliffs, NJ, 2001, pp. 141-151.
[21] P. Del Moral and J. Garnier, Genealogical particle analysis of rare events, Ann. Appl. Probab., 15 (2005), pp. 2496-2534.
[22] M. Delloye, J.-D. Fermanian, and M. Sbai, Dynamic frailties and credit portfolio modeling, Risk, 19 (2006), pp. 101-109.
[23] G. di Graziano and C. Rogers, A dynamic approach to the modelling of correlation credit derivatives using Markov chains, Int. J. Theor. Appl. Finance, 12 (2009), pp. 45-62.
[24] D. Duffie, First-to-Default Valuation, working paper, Stanford University, Palo Alto, CA, 1998.
[25] D. Duffie and N. Garleanu, Risk and valuation of collateralized debt obligations, Fin. Anal. J., 57 (2001), pp. 41-59.
[26] D. Duffie and P. Glynn, Efficient Monte Carlo estimation of security prices, Ann. Appl. Probab., 4 (1995), pp. 897-905.
[27] D. Duffie and K. J. Singleton, Simulating Correlated Defaults, working paper, Graduate School of Business, Stanford University, Palo Alto, CA, 1999.
[28] D. Duffie, L. Saita, and K. Wang, Multi-period corporate default prediction with stochastic covariates, J. Fin. Econ., 83 (2006), pp. 635-665.
[29] D. Duffie, A. Eckner, G. Horel, and L. Saita, Frailty correlated default, J. Finance, 64 (2009), pp. 2089-2123.
[30] A. Eckner, Computational techniques for basic affine models of portfolio credit risk, J. Comput. Finance, 15 (2009), pp. 63-97.
[31] E. Errais, K. Giesecke, and L. R. Goldberg, Affine point processes and portfolio credit risk, SIAM J. Financial Math., 1 (2010), pp. 642-665.
[32] P. Feldhütter, An Empirical Investigation of an Intensity Based Model for Pricing CDO Tranches, working paper, Copenhagen Business School, Copenhagen, Denmark, 2007.
[33] R. Frey and J. Backhaus, Pricing and hedging of portfolio credit derivatives with interacting default intensities, Int. J. Theor. Appl. Finance, (2008), to appear.
[34] G. Giampieri, M. Davis, and M. Crowder, A hidden Markov model of default interaction, Quant. Finance, 5 (2005), pp. 27-34.
[35] K. Giesecke and L. Goldberg, Sequential defaults and incomplete information, J. Risk, 7 (2004), pp. 1-26.
[36] K. Giesecke and S. Weber, Credit contagion and aggregate loss, J. Econom. Dynam. Control, 30 (2006), pp. 741-761.
[37] K. Giesecke, L. Goldberg, and X. Ding, A top-down approach to multi-name credit, Oper. Res., (2009), to appear.
[38] K. Giesecke, H. Kakavand, and M. Mousavi, Exact simulation of point processes with stochastic intensities, Oper. Res., (2010), to appear.
[39] P. Glasserman, Monte Carlo Methods in Financial Engineering, Springer-Verlag, New York, Berlin, 2003.
[40] P. Glasserman and J. Li, Importance sampling for portfolio credit risk, Management Sci., 51 (2005), pp. 1643-1656.
[41] P. Glasserman and J. Li, Importance sampling for a mixed Poisson model of portfolio credit risk, in Proceedings of the 2003 Winter Simulation Conference, S. Chick, P. J. Sanchez, D. Ferrin, and D. J. Morrice, eds., IEEE Press, Piscataway, NJ, 2006, pp. 267-275.
[42] P. Glasserman, W. Kang, and P. Shahabuddin, Fast simulation of multifactor portfolio credit risk, Oper. Res., 56 (2008), pp. 1200-1217.
[43] P. Glynn and D. Iglehart, Importance sampling for stochastic simulations, Management Sci., 35 (1989), pp. 1367-1392.
[44] P. Heidelberger, V. Nicola, and P. Shahabuddin, Bounded relative error in estimating transient measures of highly dependable non-Markovian systems, ACM Trans. Model. Computer Simul., 4 (1994), pp. 137-164.
[45] A. Herbertsson and H. Rootzen, Pricing $k$-th-to-default-swaps under default contagion, J. Comput. Finance, 12 (2008), pp. 49-78.
[46] S. Inglis and A. Lipton, Factor models for credit correlation, Risk, 20 (2007), pp. 110-115.
[47] S. Inglis, A. Lipton, I. Savescu, and A. Sepp, Dynamic credit models, Stat. Interface, 1 (2008), pp. 211-227.
[48] R. A. Jarrow and F. Yu, Counterparty risk and the pricing of defaultable securities, J. Finance, 56 (2001), pp. 555-576.
[49] M. Joshi and D. Kainth, Rapid and accurate development of prices and greeks for nth to default credit swaps in the Li model, Quant. Finance, 4 (2004), pp. 266-275.
[50] S. J. Koopman, A. Lucas, and A. Monteiro, The multi-stage latent factor intensity model for credit rating transitions, J. Econometrics, 142 (2008), pp. 399-424.
[51] S. Kou and X. Peng, Default Clustering and Valuation of Collateralized Debt Obligations, working paper, Columbia University, New York, 2009.
[52] H. Kraft and M. Steffensen, Bankruptcy, counterparty risk, and contagion, Rev. Finance, 11 (2007), pp. 209-252.
[53] S. Kusuoka, A remark on default risk models, Adv. Math. Econ., 1 (1999), pp. 69-82.
[54] P. Lewis and G. Shedler, Simulation of nonhomogeneous Poisson processes by thinning, Naval Logistics Quart., 26 (1979), pp. 403-413.
[55] A. Lopatin, A Simple Dynamic Model for Pricing and Hedging Heterogenous CDOs, working paper, Numerix, New York, 2009.
[56] A. Lopatin and T. Misirpashaev, Two-dimensional Markovian model for dynamics of aggregate credit loss, Adv. Econometrics, 22 (2008), pp. 243-274.
[57] A. McNeil and J. Wendin, Bayesian inference for generalized linear mixed models of portfolio credit risk, J. Empirical Finance, 14 (2007), pp. 131-149.
[58] P.-A. Meyer, Démonstration simplifée d'un théorème de Knight, in Séminaire de Probabilités V, Lecture Notes in Math. 191, Springer-Verlag, Berlin, 1971, pp. 191-195.
[59] A. Mortensen, Semi-analytical valuation of basket credit derivatives in intensity-based models, J. Derivatives, 13 (2006), pp. 8-26.
[60] E. Papageorgiou and R. Sircar, Multiscale intensity models and name grouping for valuation of multiname credit derivatives, Appl. Math. Finance, 15 (2007), pp. 73-105.
[61] P. Protter, Stochastic Integration and Differential Equations, Springer-Verlag, New York, Berlin, 2004.
[62] P. Schönbucher, Information-Driven Default Contagion, working paper, ETH Zürich, Zürich, Switzerland, 2004.
[63] F. Yu, Correlated defaults in intensity based models, Math. Finance, 17 (2007), pp. 155-173.

# Optimal Portfolio Liquidation with Execution Cost and Risk* 

Idris Kharroubi ${ }^{\dagger}$ and Huyên Pham ${ }^{\ddagger}$

Abstract. We study the optimal portfolio liquidation problem over a finite horizon in a limit order book with bid-ask spread and temporary market price impact penalizing speedy execution trades. We use a continuous-time modeling framework, but in contrast with previous related papers (see, e.g., [L. C. G. Rogers and S. Singh, Math. Finance, 20 (2010), pp. 597-615] and [A. Schied and T. Schöneborn, Finance Stoch., 13 (2009), pp. 181-204]), we do not assume continuous-time trading strategies. We consider instead real trading that occur in discrete time, and this is formulated as an impulse control problem under a solvency constraint, including the lag variable tracking the time interval between trades. A first important result of our paper is to prove rigorously that nearly optimal execution strategies in this context actually lead to a finite number of trades with strictly increasing trading times, and this holds true without assuming ad hoc any fixed transaction fee. Next, we derive the dynamic programming quasi-variational inequality satisfied by the value function in the sense of constrained viscosity solutions. We also introduce a family of value functions which converges to our value function and is characterized as the unique constrained viscosity solutions of an approximation of our dynamic programming equation. This convergence result is useful for numerical purpose but is postponed until a companion paper [F. Guilbaud, M. Mnif, and H. Pham, Numerical Methods for an Optimal Order Execution Problem, preprint, 2010].

Key words. optimal portfolio liquidation, execution trade, liquidity effects, order book, impulse control, viscosity solutions

AMS subject classifications. 93E20, 91B28, 60H30, 49L25
DOI. 10.1137/09076372X

1. Introduction. Understanding trade execution strategies is a key issue for financial market practitioners and has attracted growing attention from the academic researchers. An important problem faced by stock traders is how to liquidate large block orders of shares. This is a challenge due to the following dilemma. By trading quickly, the investor is subject to higher costs due to market impact reflecting the depth of the limit order book. Thus, to minimize price impact, it is generally beneficial to break up a large order into smaller blocks. However, more gradual trading over time results in higher risks since the asset value can vary more during the investment horizon in an uncertain environment. There has recently been considerable interest in the literature in such liquidity effects, taking into account permanent and/or temporary price impacts, and problems of this type were studied in [7], [1], [3], [9],

[^95][22], [16], [29], [20], [28], and [10], to mention just a few.
There are essentially two popular formulation types for the optimal trading problem in the literature: discrete-time versus continuous-time. In the discrete-time formulation, we may distinguish papers considering that trading takes place at fixed deterministic times (see [7]), at exogenous random discrete times given, for example, by the jumps of a Poisson process (see [26], [5]), or at discrete times decided optimally by the investor through an impulse control formulation (see [16], [20]). In this last case, one usually assumes the existence of a fixed transaction cost paid at each trading in order to ensure that strategies do not accumulate in time and really occur at discrete points in time (see, e.g., [18] or [23]). The continuous-time trading formulation is not realistic in practice, but is commonly used (as in [9], [29], or [28]), due to the tractability and powerful theory of the stochastic calculus typically illustrated by Itô's formula. In a perfectly liquid market without transaction cost and market impact, continuous-time trading is often justified by arguing that it is a limit approximation of discretetime trading when the time step goes to zero. However, one may question the validity of such an assertion in the presence of liquidity effects.

In this paper, we propose a continuous-time framework taking into account the main liquidity features and risk/cost tradeoff of portfolio execution: there is a limit order book with bid-ask spread and temporary market price impact penalizing rapid execution trades. However, in contrast with previous related papers ([29] or [28]), we do not assume continuoustime trading strategies. We consider instead real trading that take place in discrete time, and without assuming ad hoc any fixed transaction cost, in accordance with the practitioner literature. Moreover, a key issue in line with the banking regulation and solvency constraints is to define in an economically meaningful way the portfolio value of a position in stock at any time, and this is addressed in our modeling. These issues are formulated conveniently through an impulse control problem including the lag variable tracking the time interval between trades. Thus, we combine the advantages of the stochastic calculus techniques and the realistic modeling of portfolio liquidation. In this context, we study the optimal portfolio liquidation problem over a finite horizon: the investor seeks to unwind an initial position in stock shares by maximizing his expected utility from terminal liquidation wealth and under a natural economic solvency constraint involving the liquidation value of a portfolio.

A first important result of our paper is to show that nearly optimal execution strategies in this modeling actually lead to a finite number of trading times. Actually, most models dealing with trading strategies via an impulse control formulation assumed a priori that admissible trades occur only finitely many times (see, e.g., [21]) or required fixed transaction cost in order to justify a posteriori the discrete nature of trading times. In this paper, we prove that discrete-time trading appears endogenously as a consequence of liquidity features represented by temporary price impact and bid-ask spread. Moreover, the optimal trading times are strictly increasing. To the best of our knowledge, the rigorous proof of these properties is new. Next, we derive the dynamic programming quasi-variational inequality (QVI) satisfied by the value function in the sense of constrained viscosity solutions in order to handle state constraints. There are some technical difficulties related to the nonlinearity of the impulse transaction function induced by the market price impact and the nonsmoothness of the solvency boundary. In particular, since we do not assume a fixed transaction fee, which precludes the existence of a strict supersolution to the QVI, we cannot prove directly a comparison prin-
ciple (hence a uniqueness result) for the QVI. However, by using a utility penalization method with small costs, we can prove that the value function is characterized as the minimal viscosity solution to its QVI. We next consider an approximation problem with fixed small transaction costs and whose associated value functions are characterized as unique constrained viscosity solutions to their dynamic programming equations. We then prove the convergence of these value functions to our original value function by relying on the finiteness of the number of trading strategies. This convergence result is new and useful for numerical purpose and is postponed until a further study.

The plan of this paper is organized as follows. Section 2 presents the details of the model and formulates the liquidation problem. In section 3, we show some interesting economical and mathematical properties of the model, in particular the finiteness of the number of trading strategies under illiquidity costs. Section 4 is devoted to the dynamic programming and viscosity properties of the value function of our impulse control problem. We prove in particular that our value function is characterized as the minimal constrained viscosity solution to its dynamic programming QVI. We propose in section 5 an approximation of the original problem by considering a small fixed transaction fee.
2. The model and liquidation problem. We consider a financial market where an investor has to liquidate an initial position of $y>0$ shares of risky asset (or stock) by time $T$. He faces with the following risk/cost tradeoff: if he trades rapidly, this results in higher costs for quickly executed orders and market price impact; he can then split the order into several smaller blocks but is then exposed to the risk of price depreciation during the trading horizon. These liquidity effects recently received considerable interest starting with the papers [7] and [1] in a discrete-time framework and further investigated, among others, in [22], [29], and [28] in a continuous-time model. These papers assume continuous trading with instantaneous trading rate inducing price impact. In a continuous time market framework, we propose here a more realistic modeling by considering that trading takes place at discrete points in time through an impulse control formulation, and with a temporary price impact depending on the time interval between trades, and including a bid-ask spread.

We present the details of the model. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space equipped with a filtration $\mathbb{F}=\left(\mathcal{F}_{t}\right)_{0 \leq t \leq T}$ satisfying the usual conditions and supporting a one-dimensional Brownian motion $W$ on a finite horizon $[0, T], T<\infty$. We denote by $P_{t}$ the market price of the risky asset, by $X_{t}$ the amount of money (or cash holdings), by $Y_{t}$ the number of shares in the stock held by the investor at time $t$, and by $\Theta_{t}$ the time interval between time $t$ and the last trade before $t$. We set $\mathbb{R}^{*}=\mathbb{R} \backslash\{0\}, \mathbb{R}_{+}^{*}=(0, \infty)$, and $\mathbb{R}_{-}^{*}=(-\infty, 0)$.

- Trading strategies. We assume that the investor can trade only discretely on $[0, T]$. This is modeled through an impulse control strategy $\alpha=\left(\tau_{n}, \zeta_{n}\right)_{n \geq 0}: \tau_{0} \leq \cdots \leq \tau_{n} \leq \cdots \leq T$ are nondecreasing stopping times representing the trading times of the investor, and $\zeta_{n}, n \geq 0$, are $\mathcal{F}_{\tau_{n}}$-measurable random variables valued in $\mathbb{R}$ and giving the number of stocks purchased if $\zeta_{n} \geq 0$ or sold if $\zeta_{n}<0$ at these times. We denote by $\mathcal{A}$ the set of trading strategies. The sequence $\left(\tau_{n}, \zeta_{n}\right)$ may be a priori finite or infinite. Notice also that we do not assume a priori that the sequence of trading times $\left(\tau_{n}\right)$ is strictly increasing. We introduce the lag variable tracking the time interval between trades:

$$
\Theta_{t}=\inf \left\{t-\tau_{n}: \tau_{n} \leq t\right\}, \quad t \in[0, T]
$$

which evolves according to

$$
\begin{equation*}
\Theta_{t}=t-\tau_{n}, \quad \tau_{n} \leq t<\tau_{n+1}, \Theta_{\tau_{n+1}}=0, \quad n \geq 0 \tag{2.1}
\end{equation*}
$$

The dynamics of the number of shares invested in stock is given by

$$
\begin{equation*}
Y_{t}=Y_{\tau_{n}}, \quad \tau_{n} \leq t<\tau_{n+1}, Y_{\tau_{n+1}}=Y_{\tau_{n+1}^{-}}+\zeta_{n+1}, \quad n \geq 0 \tag{2.2}
\end{equation*}
$$

- Cost of illiquidity. The market price of the risky asset process follows a geometric Brownian motion:

$$
\begin{equation*}
d P_{t}=P_{t}\left(b d t+\sigma d W_{t}\right) \tag{2.3}
\end{equation*}
$$

with constants $b$ and $\sigma>0$. We do not consider a permanent price impact on the price, i.e., the lasting effect of a large trader, but focus here on the effect of illiquidity, that is, the price at which an investor will trade the asset. Suppose now that the investor decides at time $t$ to make an order in stock shares of size $e$. If the current market price is $p$ and the time lag from the last order is $\theta$, then the price he actually get for the order $e$ is

$$
\begin{equation*}
Q(e, p, \theta)=p f(e, \theta), \tag{2.4}
\end{equation*}
$$

where $f$ is a temporary price impact function from $\mathbb{R} \times[0, T]$ into $\mathbb{R}_{+} \cup\{\infty\}$. The impact of liquidity modeled in (2.4) is like a transaction cost combining nonlinearity and proportionality effects. The nonlinear costs come from the dependence of the function $f$ on $e$ and $\theta$, and we assume the following natural condition:
(H1) $\quad f(0, \theta)=1$, and $f(., \theta)$ is nondecreasing for all $\theta \in[0, T]$.
Condition (H1) means that no trade incurs any impact on the market price, i.e., $Q(0, p, \theta)=p$, and a purchase (resp., a sale) of stock shares induces a cost (resp., gain) greater (resp., smaller) than the market price, which increases (resp., decreases) with the size of the order. In other words, we have $Q(e, p, \theta) \geq$ (resp., $\leq$ ) $p$ for $e \geq$ (resp., $\leq$ ) 0 , and $Q(., p, \theta)$ is nondecreasing. The proportional transaction cost effect is realized by considering a bid-ask spread, i.e., assuming the following condition:

$$
\begin{gather*}
\overline{\kappa_{b}}:=\sup _{\theta \in[0, T]} \kappa_{b}(\theta):=\sup _{\theta \in[0, T]} f\left(0^{-}, \theta\right)<1, \text { and }  \tag{H2}\\
\underline{\kappa_{a}}:=\inf _{\theta \in[0, T]} \kappa_{a}(\theta):=\inf _{\theta \in[0, T]} f\left(0^{+}, \theta\right)>1 .
\end{gather*}
$$

The term $\kappa_{b}\left(\overline{\theta)}\right.$ (resp., $\left.\kappa_{a}(\theta)\right)$ may be interpreted as the relative bid price (resp., ask price) given a time lag from last order $\theta$, and condition (H2) means that, given a current market or midprice $p$ at time $t, \overline{\kappa_{b}} p$ is the largest bid price, $\kappa_{a} p$ is the lowest ask price, and $\left(\underline{\kappa_{a}}-\overline{\kappa_{b}}\right) p$ is the bid-ask spread. In a typical example (see (2.6)), $\kappa_{a}(\theta)$ and $\kappa_{b}(\theta)$ do not depend on $\theta$, i.e., $\kappa_{a}(\theta)=\underline{\kappa_{a}}$, and $\kappa_{b}(\theta)=\overline{\kappa_{b}}$. On the other hand, this transaction cost function $f$ can be determined implicitly from the impact of a market order placed by a large trader in a limit order book (LOB), as explained in [22], [29], or [28]. Indeed, suppose that there is some midprice $p$ at current time $t$ and an order book of quotes posted on either side of the midprice. To fix the ideas, we consider the upper half of the LOB, and we denote by $\rho_{a}(k, \theta)$ the density of quotes to sell at relative price $k \geq \kappa_{a}(\theta)$ when the time lag from the last market order of the large trader is $\theta$. Similarly as in [22] or [12], we considered that the LOB may be affected by the past trades of the large investor through, e.g., its last trading time. If the large investor
places a buy market order for $e>0$ shares of the asset, this will consume all shares in the LOB located at relative prices between $\kappa_{a}(\theta)$ and $\hat{k}=\hat{k}(e, \theta)$ determined by

$$
\int_{\kappa_{a}(\theta)}^{\hat{k}} \rho_{a}(k, \theta) d k=e .
$$

Consequently, the cost paid by the large investor to acquire $e>0$ units of the asset through the LOB is

$$
\begin{equation*}
Q(p, e, \theta)=p \int_{\kappa_{a}(\theta)}^{\hat{k}} k \rho_{a}(k, \theta) d k=p f(e, \theta) . \tag{2.5}
\end{equation*}
$$

Therefore, the shape function $\rho_{a}$ of the LOB for sell quotes determines via the relation (2.5) the temporary market impact function $f$ for buy market order. Similarly, the shape function $\rho_{b}$ of the LOB for buy quotes determines the price impact function $f$ for sell market order, i.e., $f(e,$.$) for e<0$. Notice in particular that the dependence of $f$ in $\theta$ is induced by the dependence of $\rho_{a}$ and $\rho_{b}$ on $\theta$. Such an assumption is also made in the seminal paper [1], where the price impact function penalizes high trading volume per unit of time $e / \theta$. We assume that $f$ satisfies the following condition:
(i) $f(e, 0)=0$ for $e<0$, and (ii) $f(e, 0)=\infty$ for $e>0$.

Condition (H3) expresses the higher costs for immediacy in trading: indeed, the immediate market resiliency is limited, and the faster the investor wants to liquidate (resp., purchase) the asset, the deeper into the LOB he will have to go, and the lower (resp., higher) will be the price for the shares of the asset sold (resp., bought), with a zero (resp., infinite) limiting price for immediate block sale (resp., purchase). If the investor speeds up his buy trades, he will deplete the short-term supply and increase the immediate cost for additional trades. As more time is allowed between trades, supply will gradually recover. Moreover, the intervention of a large investor in an illiquid market has an important impact on the order book. His trading will execute the majority of orders on standby and so clear out the order book. Condition (H3) also prevents the investor from passing orders at immediate consecutive times, which is the case in practice. Notice that if we consider a market impact function $f(e)$, which does not depend on $\theta$, then the strict increasing monotonicity of trading times is not guaranteed a priori by a fixed transaction fee $\varepsilon>0$. Indeed, suppose, for example, that the investor wants to buy $e$ shares of stock, given the current market price $p$. Then, in the case where $e_{1} p f\left(e_{1}\right)+e_{2} p f\left(e_{2}\right)+\varepsilon<\operatorname{epf}(e)$, for some positive $e_{1}, e_{2}$ such that $e_{1}+e_{2}=e$, it is better to split the number of shares and trade separately the smaller quantities $e_{1}$ and $e_{2}$ at the same time.

We also assume some technical regularity conditions on the temporary price impact function, which shall be used later in Theorem 3.1.
(i) $f$ is continuous on $\mathbb{R}^{*} \times(0, T]$;
(ii) $f$ is $C^{1}$ on $\mathbb{R}_{-}^{*} \times[0, T]$, and $x \mapsto \frac{\partial f}{\partial \theta}$ is bounded on $\mathbb{R}_{-}^{*} \times[0, T]$.

A usual form (see, e.g., [19], [2]) of temporary price impact function $f$ (which includes here a transaction cost term as well) suggested by empirical studies is

$$
\begin{equation*}
f(e, \theta)=\exp \left(\lambda\left|\frac{e}{\theta}\right|^{\beta} \operatorname{sgn}(e)\right) \cdot\left(\underline{\kappa_{a}} \mathbf{1}_{e>0}+\mathbf{1}_{e=0}+\overline{\kappa_{b}} \mathbf{1}_{e<0}\right), \tag{2.6}
\end{equation*}
$$

with the convention $f(0,0)=1$. Here $0<\overline{\kappa_{b}}<1<\underline{\kappa_{a}}, \underline{\kappa_{a}}-\overline{\kappa_{b}}$ is the (relative) bid-ask spread parameter, $\lambda>0$ is the temporary price impact factor, and $\beta>0$ is the price impact exponent. The price impact function $f$ depends on $e$ and $\theta$ through the volume per unit of time $\vartheta=e / \theta$, and the penalization of quick trading, i.e., when $\theta$ goes to zero, is formulated by condition (H3), which is satisfied in (2.6). The power functional form in $e / \theta$ for the logarithm of the price impact function fits well with the statistical properties of order books (see [27]), and the parameters $\lambda, \beta$ can be determined by regressions on data. In particular, empirical observations suggest a value $\beta=1 / 2$. Notice that in the limiting case $\lambda=0$ the function $f$ is constant on $(0, \infty)$ and $(-\infty, 0)$, with a jump at 0 , which means that one ignores the nonlinear costs, keeping only the proportional costs.

In our illiquidity model, we focus on the cost of trading fast (that is, the temporary price impact) and ignore as in [9] and [28] the permanent price impact of a large trade. This last effect could be included in our model, by assuming a jump of the price process at the trading date, depending on the order size; see, e.g., [16] and [20].

- Cash holdings. We assume a zero risk-free return, so that the bank account is constant between two trading times:

$$
\begin{equation*}
X_{t}=X_{\tau_{n}}, \quad \tau_{n} \leq t<\tau_{n+1}, \quad n \geq 0 \tag{2.7}
\end{equation*}
$$

When a discrete trading $\Delta Y_{t}=\zeta_{n+1}$ occurs at time $t=\tau_{n+1}$, this results in a variation of the cash amount given by $\Delta X_{t}:=X_{t}-X_{t^{-}}=-\Delta Y_{t} \cdot Q\left(\Delta Y_{t}, P_{t}, \Theta_{t^{-}}\right)$due to the illiquidity effects. In other words, we have

$$
\begin{align*}
X_{\tau_{n+1}} & =X_{\tau_{n+1}^{-}}-\zeta_{n+1} Q\left(\zeta_{n+1}, P_{\tau_{n+1}}, \Theta_{\tau_{n+1}^{-}}\right) \\
& =X_{\tau_{n+1}^{-}}-\zeta_{n+1} P_{\tau_{n+1}} f\left(\zeta_{n+1}, \tau_{n+1}-\tau_{n}\right), \quad n \geq 0 \tag{2.8}
\end{align*}
$$

Notice that, similarly as in the above cited papers dealing with continuous-time trading, we do not assume fixed transaction fees to be paid at each trading. They are practically insignificant with respect to the price impact and bid-ask spread. We can then not exclude a priori trading strategies with immediate trading times; i.e., $\Theta_{\tau_{n+1}^{-}}=\tau_{n+1}-\tau_{n}=0$ for some $n$. However, notice that under condition (H3) an immediate sale does not increase the cash holdings, i.e., $X_{\tau_{n+1}}=X_{\tau_{n+1}^{-}}=X_{\tau_{n}}$, while an immediate purchase leads to a bankruptcy, i.e., $X_{\tau_{n+1}}=-\infty$.

Remark 2.1. Although assumption (H3) induces the worst gains for immediate successive trading, its does not prevent continuous-time trading at the limit. To see this, let us consider a market impact function $f(e, \theta)=\bar{f}(e / \theta)$ depending on $e, \theta$ through the volume per unit of time $e / \theta$, as in (2.6), and define the continuous-time (deterministic) strategy $\left(Y_{t}\right)_{t}$ with constant slope starting from $Y_{0}=y>0$ and ending at $Y_{T}=0$ at the liquidation date $T$. Consider now a uniform time discretization of the interval $[0, T]$ with time step $h=T / N$, and define the discrete-time strategy $\alpha^{h}=\left(\tau_{n}^{h}, \zeta_{n}^{h}\right)_{1 \leq n \leq N}$ by

$$
\tau_{n}^{h}=n h \quad \text { and } \quad \zeta_{n}^{h}=Y_{\tau_{n}^{h}}-Y_{\tau_{n-1}^{h}}=-\frac{y}{N}, \quad n \geq 1
$$

with $\tau_{0}^{h}=0$. This strategy $\alpha^{h}$ à la Almgren and Chriss is a uniform discrete-time approximation of $Y$. From (2.8), the cash holdings associated with this strategy, and starting from
an initial capital $x$, are then given at terminal date $T$ by

$$
\begin{aligned}
X_{T}^{h} & =x-\sum_{n=1}^{N} \zeta_{n}^{h} P_{\tau_{n}^{h}} \bar{f}\left(\frac{\zeta_{n}^{h}}{\tau_{n}^{h}-\tau_{n-1}^{h}}\right) \\
& =x+\sum_{n=1}^{N} \frac{y}{T} h P_{\tau_{n}^{h}} \bar{f}\left(-\frac{y}{T}\right) \longrightarrow x+\frac{y}{T} \bar{f}\left(-\frac{y}{T}\right) \int_{0}^{T} P_{t} d t
\end{aligned}
$$

when $h$ goes to zero. Thus, we see that the execution cost for the continuous-time limit strategy is also finite. Actually, the above argument shows more generally that any continuous-time finite variation strategy $\left(Y_{t}\right)$ with continuous instantaneous trading rate process $\eta_{t}=d Y_{t} / d t$ can be approximated by a discrete-time trading strategy $\alpha^{h}=\left(\tau_{n}^{h}, \zeta_{n}^{h}\right)_{1 \leq n \leq N}$, with $\tau_{n}^{h}=n h, \zeta_{n}^{h}$ $=Y_{\tau_{n}^{h}}-Y_{\tau_{n-1}^{h}}$, such that the corresponding execution cost $X_{T}^{h}$ converges to $x-\int_{0}^{T} P_{t} \bar{f}\left(\eta_{t}\right) d Y_{t}$ as $h=T / N$ goes to zero.

- Liquidation value and solvency constraint. A key issue in portfolio liquidation is to define in an economically meaningful way what the portfolio value of a position on cash and stocks is. In our framework, we impose a no-short selling constraint on the trading strategies, i.e.,

$$
Y_{t} \geq 0, \quad 0 \leq t \leq T .
$$

This constraint is consistent with the bank regulations following the financial crisis. We consider the liquidation function $L(x, y, p, \theta)$ representing the net wealth value that an investor with a cash amount $x$ would obtain by liquidating his stock position $y \geq 0$ by a single block trade when the market price is $p$ and given the time lag $\theta$ from the last trade. It is defined on $\mathbb{R} \times \mathbb{R}_{+} \times \mathbb{R}_{+}^{*} \times[0, T]$ by

$$
L(x, y, p, \theta)=x+y p f(-y, \theta),
$$

and we impose the following liquidation constraint on trading strategies:

$$
L\left(X_{t}, Y_{t}, P_{t}, \Theta_{t}\right) \geq 0, \quad 0 \leq t \leq T .
$$

We have $L(x, 0, p, \theta)=x$, and under condition (H3)(ii) we notice that $L(x, y, p, 0)=x$ for $y$ $\geq 0$. We naturally introduce the liquidation solvency region:

$$
\mathcal{S}=\left\{(z, \theta)=(x, y, p, \theta) \in \mathbb{R} \times \mathbb{R}_{+} \times \mathbb{R}_{+}^{*} \times[0, T]: y>0 \text { and } L(z, \theta)>0\right\} .
$$

We denote its boundary and its closure by

$$
\partial \mathcal{S}=\partial_{y} \mathcal{S} \cup \partial_{L} \mathcal{S} \quad \text { and } \quad \overline{\mathcal{S}}=\mathcal{S} \cup \partial \mathcal{S},
$$

where

$$
\begin{aligned}
& \partial_{y} \mathcal{S}=\left\{(z, \theta)=(x, y, p, \theta) \in \mathbb{R} \times \mathbb{R}_{+} \times \mathbb{R}_{+}^{*} \times[0, T]: y=0 \text { and } x=L(z, \theta) \geq 0\right\} \\
& \partial_{L} \mathcal{S}=\left\{(z, \theta)=(x, y, p, \theta) \in \mathbb{R} \times \mathbb{R}_{+} \times \mathbb{R}_{+}^{*} \times[0, T]: L(z, \theta)=0\right\}
\end{aligned}
$$



Figure 1. Domain $\mathcal{S}$ in the nonhatched zone for fixed $p=1$ and $\theta$ evolving from 1.5 to 0.1. Here $\overline{\kappa_{b}}=0.9$ and $f(e, \theta)=\overline{k_{b}} \exp \left(\frac{e}{\theta}\right)$ for $e<0$. Notice that when $\theta$ goes to 0 , the domain converges to the open orthant $\mathbb{R}_{+}^{*} \times \mathbb{R}_{+}^{*}$.

We also denote by $D_{0}$ the corner line in $\partial \mathcal{S}$ :

$$
D_{0}=\{0\} \times\{0\} \times \mathbb{R}_{+}^{*} \times[0, T]=\partial_{y} \mathcal{S} \cap \partial_{L} \mathcal{S}
$$

We represent in Figure 1 the graph of $\mathcal{S}$ in the plane $(x, y)$, in Figure 2 the graph of $\mathcal{S}$ in the space $(x, y, p)$, and in Figure 3 the graph of $\mathcal{S}$ in the space $(x, y, \theta)$.

- Admissible trading strategies. Given $(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}$, we say that the impulse control strategy $\alpha=\left(\tau_{n}, \zeta_{n}\right)_{n \geq 0}$ is admissible, denoted by $\alpha \in \mathcal{A}(t, z, \theta)$, if $\tau_{0}=t-\theta, \tau_{n} \geq t, n \geq$ 1, and the process $\left\{\left(Z_{s}, \Theta_{s}\right)=\left(X_{s}, Y_{s}, P_{s}, \Theta_{s}\right), t \leq s \leq T\right\}$ solution to (2.1)-(2.2)-(2.3)-(2.7)(2.8), with an initial state $\left(Z_{t^{-}}, \Theta_{t^{-}}\right)=(z, \theta)$ (and the convention that $\left(Z_{t}, \Theta_{t}\right)=(z, \theta)$ if $\tau_{1}$ $>t)$, satisfies $\left(Z_{s}, \Theta_{s}\right) \in[0, T] \times \overline{\mathcal{S}}$ for all $s \in[t, T]$. As usual, to alleviate notation, we omit the dependence of $(Z, \Theta)$ in $(t, z, \theta, \alpha)$ when there is no ambiguity.
- Portfolio liquidation problem. We consider a utility function $U$ from $\mathbb{R}_{+}$into $\mathbb{R}$ that is nondecreasing and concave, with $U(0)=0$, and such that there exist $K \geq 0$ and $\gamma \in[0,1)$ :

$$
\begin{equation*}
0 \leq U(x) \leq K x^{\gamma} \quad \forall x \in \mathbb{R}_{+} \tag{H5}
\end{equation*}
$$

The problem of optimal portfolio liquidation is formulated as

$$
\begin{equation*}
v(t, z, \theta)=\sup _{\alpha \in \mathcal{A}_{\ell}(t, z, \theta)} \mathbb{E}\left[U\left(X_{T}\right)\right], \quad(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}} \tag{2.9}
\end{equation*}
$$

theta $=1$

$y$ : stock amount

Figure 2. Lower bound of the domain $\mathcal{S}$ for fixed $\theta=1$. Here $\overline{\kappa_{b}}=0.9$ and $f(e, \theta)=\overline{\kappa_{b}} \exp \left(\frac{e}{\theta}\right)$ for $e<0$. Notice that when $p$ is fixed, we obtain Figure 1.


Figure 3. Lower bound of the domain $\mathcal{S}$ for fixed $p$ with $f(e, \theta)=\overline{\kappa_{b}} \exp \left(\frac{e}{\theta}\right)$ for $e<0$ and $\overline{\kappa_{b}}=0.9$. Notice that when $\theta$ is fixed, we obtain Figure 1.
where $\mathcal{A}_{\ell}(t, z, \theta)=\left\{\alpha \in \mathcal{A}(t, z, \theta): Y_{T}=0\right\}$. Notice that this set is nonempty. Indeed, let $(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}$, and consider the impulse control strategy $\tilde{\alpha}=\left(\tau_{n}, \zeta_{n}\right)_{n \geq 0}, \tau_{0}=$ $t-\theta$, consisting in liquidating all the stock shares immediately and then no longer doing transactions, i.e., $\left(\tau_{1}, \zeta_{1}\right)=(t,-y)$, and $\zeta_{n}=0, n \geq 2$. The associated state process $(Z=$ $(X, Y, P), \Theta)$ satisfies $X_{s}=L(z, \theta), Y_{s}=0$, which shows that $L\left(Z_{s}, \Theta_{s}\right)=X_{s}=L(z, \theta) \geq$ $0, t \leq s \leq T$, and thus $\tilde{\alpha} \in \mathcal{A}_{\ell}(t, z, \theta) \neq \emptyset$. Observe also that for $\alpha \in \mathcal{A}_{\ell}(t, z, \theta), X_{T}=$ $L\left(Z_{T}, \Theta_{T}\right) \geq 0$, so that the expectations in (2.9) and the value function $v$ are well defined in $[0, \infty]$. Moreover, by considering the particular strategy $\tilde{\alpha}$ described above, which leads to a final liquidation value $X_{T}=L(z, \theta)$, we obtain a lower bound for the value function:

$$
\begin{equation*}
v(t, z, \theta) \geq U(L(z, \theta)), \quad(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}} . \tag{2.10}
\end{equation*}
$$

Remark 2.2. We can shift the terminal liquidation constraint in $\mathcal{A}_{\ell}(t, z, \theta)$ to a terminal liquidation utility by considering the function $U_{L}$ defined on $\overline{\mathcal{S}}$ by

$$
U_{L}(z, \theta)=U(L(z, \theta)), \quad(z, \theta) \in \overline{\mathcal{S}}
$$

Then, problem (2.9) is written equivalently as

$$
\begin{equation*}
\bar{v}(t, z, \theta)=\sup _{\alpha \in \mathcal{A}(t, z, \theta)} \mathbb{E}\left[U_{L}\left(Z_{T}, \Theta_{T}\right)\right], \quad(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}} \tag{2.11}
\end{equation*}
$$

Indeed, by observing that for all $\alpha \in \mathcal{A}_{\ell}(t, z, \theta)$ we have $\mathbb{E}\left[U\left(X_{T}\right)\right]=\mathbb{E}\left[U_{L}\left(Z_{T}, \Theta_{T}\right)\right]$, and since $\mathcal{A}_{\ell}(t, z, \theta) \subset \mathcal{A}(t, z, \theta)$, it is clear that $v \leq \bar{v}$. Conversely, for any $\alpha \in \mathcal{A}(t, z, \theta)$ associated with the state controlled process $(Z, \Theta)$, consider the impulse control strategy $\tilde{\alpha}=\alpha \cup\left(T,-Y_{T}\right)$ consisting in liquidating all the stock shares $Y_{T}$ at time $T$. The corresponding state process $(\tilde{Z}, \tilde{\Theta})$ clearly satisfies $\left(\tilde{Z}_{s}, \tilde{\Theta}_{s}\right)=\left(Z_{s}, \Theta_{s}\right)$ for $t \leq s<T$, and $\tilde{X}_{T}=L\left(Z_{T}, \Theta_{T}\right), \tilde{Y}_{T}=0$, and so $\tilde{\alpha} \in \mathcal{A}_{\ell}(t, z, \theta)$. We deduce that $\mathbb{E}\left[U_{L}\left(Z_{T}, \Theta_{T}\right)\right]=\mathbb{E}\left[U\left(\tilde{X}_{T}\right)\right] \leq v(t, z, \theta)$, and so since $\alpha$ is arbitrary in $\mathcal{A}(t, z, \theta), \bar{v}(t, z, \theta) \leq v(t, z, \theta)$. This proves the equality $v=\bar{v}$. Actually, the above arguments also show that $\sup _{\alpha \in \mathcal{A}_{\ell}(t, z, \theta)} U\left(X_{T}\right)=\sup _{\alpha \in \mathcal{A}(t, z, \theta)} U_{L}\left(Z_{T}, \Theta_{T}\right)$.

Remark 2.3. Following Remark 2.1, we can formulate a continuous-time trading version of our illiquid market model with stock price $P$ and temporary price impact $\bar{f}$. The trading strategy is given by an $\mathbb{F}$-adapted process $\eta=\left(\eta_{t}\right)_{0 \leq t \leq T}$ representing the instantaneous trading rate, which means that the dynamics of the cumulative number of stock shares $Y$ is governed by $d Y_{t}=\eta_{t} d t$. The cash holdings $X$ are

$$
d X_{t}=-\eta_{t} P_{t} \bar{f}\left(\eta_{t}\right) d t
$$

Notice that in a continuous-time trading formulation the time interval between trades is $\Theta_{t}$ $=0$ at any time $t$. Under condition(H3), the liquidation value is then given at any time $t$ by

$$
L\left(X_{t}, Y_{t}, P_{t}, 0\right)=X_{t}, \quad 0 \leq t \leq T
$$

and does not take into account the position in stock shares, which is economically undesirable. On the contrary, by explicitly considering the time interval between trades in our discrete-time trading formulation, we take into account the position in stock.
3. Properties of the model. In this section, we show that the illiquid market model presented in the previous section displays some interesting and economically meaningful properties on the admissible trading strategies and the optimal performance, i.e., the value function. Let us consider the impulse transaction function $\Gamma$ defined on $\mathbb{R} \times \mathbb{R}_{+} \times \mathbb{R}_{+}^{*} \times[0, T] \times \mathbb{R}$ into $\mathbb{R} \cup\{-\infty\} \times \mathbb{R} \times \mathbb{R}_{+}^{*}$ by

$$
\Gamma(z, \theta, e)=(x-e p f(e, \theta), y+e, p)
$$

for $z=(x, y, p)$ and set $\bar{\Gamma}(z, \theta, e)=(\Gamma(z, \theta, e), 0)$. This corresponds to the value of the state variable $(Z, \Theta)$ immediately after a trading at time $t=\tau_{n+1}$ of $\zeta_{n+1}$ shares of stock, i.e., $\left(Z_{\tau_{n+1}}, \Theta_{\tau_{n+1}}\right)=\left(\Gamma\left(Z_{\tau_{n+1}^{-}}, \Theta_{\tau_{n+1}^{-}}, \zeta_{n+1}\right), 0\right)$. We then define the set of admissible transactions:

$$
\mathcal{C}(z, \theta)=\{e \in \mathbb{R}:(\Gamma(z, \theta, e), 0) \in \overline{\mathcal{S}}\}, \quad(z, \theta) \in \overline{\mathcal{S}}
$$

This means that for any $\alpha=\left(\tau_{n}, \zeta_{n}\right)_{n \geq 0} \in \mathcal{A}(t, z, \theta)$ with associated state process $(Z, \Theta)$ we have $\zeta_{n} \in \mathcal{C}\left(Z_{\tau_{n}^{-}}, \Theta_{\tau_{n}^{-}}\right), n \geq 1$. We define the impulse operator $\mathcal{H}$ by

$$
\mathcal{H} \varphi(t, z, \theta)=\sup _{e \in \mathcal{C}(z, \theta)} \varphi(t, \Gamma(z, \theta, e), 0), \quad(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}
$$

We also introduce the liquidation function corresponding to the classical Merton model without market impact:

$$
L_{M}(z)=x+p y \quad \forall z=(x, y, p) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_{+}^{*}
$$

For $(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}$, with $z=(x, y, p)$, we denote by $\left(Z^{0, t, z}, \Theta^{0, t, \theta}\right.$ ) the state process starting from $(z, \theta)$ at time $t$ and without any impulse control strategy: it is given by

$$
\left(Z_{s}^{0, t, z}, \Theta_{s}^{0, t, \theta}\right)=\left(x, y, P_{s}^{t, p}, \theta+s-t\right), \quad t \leq s \leq T
$$

where $P^{t, p}$ is the solution to (2.3) starting from $p$ at time $t$. Notice that $\left(Z^{0, t, z}, \Theta^{0, t, \theta}\right)$ is the continuous part of the state process $(Z, \Theta)$ controlled by $\alpha \in \mathcal{A}(t, z, \theta)$. The infinitesimal generator $\mathcal{L}$ associated with the process $\left(Z^{0, t, z}, \Theta^{0, t, \theta}\right)$ is

$$
\mathcal{L} \varphi+\frac{\partial \varphi}{\partial \theta}=b p \frac{\partial \varphi}{\partial p}+\frac{1}{2} \sigma^{2} p^{2} \frac{\partial^{2} \varphi}{\partial p^{2}}+\frac{\partial \varphi}{\partial \theta}
$$

We first prove a useful result on the set of admissible transactions.
Lemma 3.1. Assume that (H1), (H2), and (H3) hold. Then, for all $(z, \theta) \in \overline{\mathcal{S}}$, with $z=$ $(x, y, p)$, the set $\mathcal{C}(z, \theta)$ is compact in $\mathbb{R}$ and satisfies

$$
\begin{equation*}
\mathcal{C}(z, \theta) \subset[-y, \bar{e}(z, \theta)] \tag{3.1}
\end{equation*}
$$

where $-y \leq \bar{e}(z, \theta)<\infty$ is given by

$$
\bar{e}(z, \theta)=\left\{\begin{array}{cl}
\sup \{e \in \mathbb{R}: \operatorname{epf}(e, \theta) \leq x\} & \text { if } \theta>0 \\
0 & \text { if } \theta=0
\end{array}\right.
$$

For $\theta=0$, (3.1) becomes an equality: $\mathcal{C}(z, 0)=[-y, 0]$.
The set function $\mathcal{C}$ is continuous with respect to the Hausdorff metric; i.e., if $\left(z_{n}, \theta_{n}\right)$ converges to $(z, \theta)$ in $\overline{\mathcal{S}}$, and $\left(e_{n}\right)$ is a sequence in $\mathcal{C}\left(z_{n}, \theta_{n}\right)$ converging to $e$, then $e \in \mathcal{C}(z, \theta)$. Moreover, if $e \in \mathbb{R} \mapsto e f(e, \theta)$ is strictly increasing for $\theta \in(0, T]$, then for $(z, \theta) \in \partial_{L} \mathcal{S}$ with $\theta>0$ we have $\bar{e}(z, \theta)=-y$, i.e., $\mathcal{C}(z, \theta)=\{-y\}$.

Proof. By definition of the impulse transaction function $\Gamma$ and the liquidation function $L$, we immediately see that the set of admissible transactions is written as

$$
\begin{align*}
\mathcal{C}(z, \theta) & =\{e \in \mathbb{R}: x-\operatorname{epf}(e, \theta) \geq 0, \text { and } y+e \geq 0\} \\
& =\{e \in \mathbb{R}: \operatorname{epf}(e, \theta) \leq x\} \cap[-y, \infty)=: \mathcal{C}_{1}(z, \theta) \cap[-y, \infty) . \tag{3.2}
\end{align*}
$$

It is clear that $\mathcal{C}(z, \theta)$ is closed and bounded and thus a compact set. Under (H1) and (H2), we have $\lim _{e \rightarrow \infty} \operatorname{epf}(e, \theta)=\infty$. Hence we get $\bar{e}(z, \theta)<\infty$ and $\mathcal{C}_{1}(z, \theta) \subset(-\infty, \bar{e}(z, \theta)]$. From (3.2), we get (3.1). Suppose $\theta=0$. Under (H3), using $(z, \theta) \in \overline{\mathcal{S}}$, we have $\mathcal{C}_{1}(z, \theta)=\mathbb{R}_{-}$. From (3.2), we get $\mathcal{C}(z, \theta)=[-y, 0]$.

Let us now prove the continuity of the set of admissible transactions. Consider a sequence $\left(z_{n}, \theta_{n}\right)$ in $\overline{\mathcal{S}}$, with $z_{n}=\left(x_{n}, y_{n}, p_{n}\right)$, converging to $(z, \theta) \in \overline{\mathcal{S}}$ and a sequence $\left(e_{n}\right)$ in $\mathcal{C}\left(z_{n}, \theta_{n}\right)$ converging to $e$. Suppose first that $\theta>0$. Then, for $n$ large enough, $\theta_{n}>0$ and, by observing that $(z, \theta, e) \mapsto \bar{\Gamma}(z, \theta, e)$ is continuous on $\mathbb{R} \times \mathbb{R}_{+} \times \mathbb{R}_{+}^{*} \times \mathbb{R}_{+}^{*} \times \mathbb{R}$, we immediately deduce that $e \in \mathcal{C}(z, \theta)$. In the case $\theta=0$, writing $x_{n}-e_{n} f\left(e_{n}, \theta_{n}\right) \geq 0$, using (H3)(ii), and sending $n$ to infinity, we see that $e$ should necessarily be nonpositive. By writing also that $y_{n}+e_{n} \geq$ 0 , we get by sending $n$ to infinity that $y+e \geq 0$, and therefore $e \in \mathcal{C}(z, 0)=[-y, 0]$.

Suppose finally that $e \in \mathbb{R} \mapsto e f(e, \theta)$ is increasing, and fix $(z, \theta) \in \partial_{L} \mathcal{S}$, with $\theta>0$. Then, $L(z, \theta)=0$, i.e., $x=-y p f(-y, \theta)$. Set $\bar{e}=\bar{e}(z, \theta)$. By writing that $\bar{e} p f(\bar{e}, \theta) \leq x=$ $-y p f(-y, \theta)$, and $\bar{e} \geq-y$, we deduce from the increasing monotonicity of $e \mapsto \operatorname{epf}(e, \theta)$ that $\bar{e}=-y$.

Remark 3.1. The previous lemma implies in particular that $\mathcal{C}(z, 0) \subset \mathbb{R}_{-}$, which means that an admissible transaction after an immediate trading should necessarily be a sale. In other words, given $\alpha=\left(\tau_{n}, \zeta_{n}\right)_{n \geq 0} \in \mathcal{A}(t, z, \theta),(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}$, if $\Theta_{\tau_{n}^{-}}=0$, then $\zeta_{n} \leq$ 0 . The continuity property of $\mathcal{C}$ ensures that the operator $\mathcal{H}$ preserves the lower and upper semicontinuity (see (A.3) in Appendix A). This lemma also asserts that, under the assumption of increasing monotonicity of $e \mapsto e f(e, \theta)$, when the state is in the boundary $L=0$, the only admissible transaction is to liquidate all stock shares. This increasing monotonicity means that the amount traded is increasing with the size of the order. Such an assumption is satisfied in the example (2.6) of temporary price impact function $f$ for $\beta=2$, but it is not fulfilled for $\beta=1$. In this case, the presence of illiquidity cost implies that it may be more advantageous to split the order size.

We next state some useful bounds on the liquidation value associated with an admissible transaction.

Lemma 3.2. Assume that (H1) holds. Then, we have for all $(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}$

$$
\begin{align*}
0 \leq L(z, \theta) & \leq L_{M}(z),  \tag{3.3}\\
L_{M}(\Gamma(z, \theta, e)) & \leq L_{M}(z) \quad \forall e \in \mathbb{R},  \tag{3.4}\\
\sup _{\alpha \in \mathcal{A}(t, z, \theta)} L\left(Z_{s}, \Theta_{s}\right) & \leq L_{M}\left(Z_{s}^{0, t, z}\right), \quad t \leq s \leq T . \tag{3.5}
\end{align*}
$$

Furthermore, under (H2), we have for all $(z, \theta) \in \overline{\mathcal{S}}, z=(x, y, p)$

$$
\begin{equation*}
L_{M}(\Gamma(z, \theta, e)) \leq L_{M}(z)-\min \left(\underline{\kappa_{a}}-1,1-\overline{\kappa_{b}}\right)|e| p \quad \forall e \in \mathbb{R} . \tag{3.6}
\end{equation*}
$$

Proof. Under (H1), we have $f(e, \theta) \leq 1$ for all $e \leq 0$, which clearly shows (3.3). From the definition of $L_{M}$ and $\Gamma$, we see that for all $e \in \mathbb{R}$

$$
\begin{equation*}
L_{M}(\Gamma(z, \theta, e))-L_{M}(z)=e p(1-f(e, \theta)) \tag{3.7}
\end{equation*}
$$

which yields the inequality (3.4). Fix some arbitrary $\alpha=\left(\tau_{n}, \zeta_{n}\right)_{n \geq 0} \in \mathcal{A}(t, z, \theta)$ associated with the controlled state process $(Z, \Theta)$. When a transaction occurs at time $s=\tau_{n}, n \geq 1$, the jump of $L_{M}(Z)$ is nonpositive by (3.4):

$$
\Delta L_{M}\left(Z_{s}\right)=L_{M}\left(Z_{\tau_{n}}\right)-L_{M}\left(Z_{\tau_{n}^{-}}\right)=L_{M}\left(\Gamma\left(Z_{\tau_{n}^{-}}, \Theta_{\tau_{n}^{-}}, \zeta_{n}\right)\right)-L_{M}\left(Z_{\tau_{n}^{-}}\right) \leq 0
$$

We deduce that the process $L_{M}(Z)$ is smaller than its continuous part equal to $L_{M}\left(Z^{0, t, z}\right)$, and we then get (3.5) with (3.3). Finally, under the additional condition (H2), we easily obtain inequality (3.6) from relation (3.7).

We now check that our liquidation problem is well-posed by stating a natural upper bound on the optimal performance, namely, that the value function in our illiquid market model is bounded by the usual Merton bound in a perfectly liquid market.

Proposition 3.1. Assume that (H1) and (H5) hold. Then, for all $(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}$, the family $\left\{U_{L}\left(Z_{T}, \Theta_{T}\right), \alpha \in \mathcal{A}(t, z, \theta)\right\}$ is uniformly integrable, and we have

$$
\begin{align*}
v(t, z, \theta) \leq v_{0}(t, z) & :=\mathbb{E}\left[U\left(L_{M}\left(Z_{T}^{0, t, z}\right)\right)\right], \quad(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}} \\
& \leq K e^{\rho(T-t)} L_{M}(z)^{\gamma}, \tag{3.8}
\end{align*}
$$

where $\rho$ is a positive constant such that

$$
\begin{equation*}
\rho \geq \frac{\gamma}{1-\gamma} \frac{b^{2}}{2 \sigma^{2}} . \tag{3.9}
\end{equation*}
$$

Proof. From (3.5) and the nondecreasing monotonicity of $U$, we have for all $(t, z, \theta) \in$ $[0, T] \times \overline{\mathcal{S}}$

$$
\sup _{\alpha \in \mathcal{A}_{\ell}(t, z, \theta)} U\left(X_{T}\right)=\sup _{\alpha \in \mathcal{A}(t, z, \theta)} U_{L}\left(Z_{T}, \Theta_{T}\right) \leq U\left(L_{M}\left(Z_{T}^{0, t, z}\right)\right),
$$

and all the assertions of the proposition will follow once we prove the inequality (3.8). For this, consider the nonnegative function $\varphi$ defined on $[0, T] \times \overline{\mathcal{S}}$ by

$$
\varphi(t, z, \theta)=e^{\rho(T-t)} L_{M}(z)^{\gamma}=e^{\rho(T-t)}(x+p y)^{\gamma},
$$

and notice that $\varphi$ is smooth $C^{2}$ on $[0, T] \times\left(\overline{\mathcal{S}} \backslash D_{0}\right)$. We claim that, for $\rho>0$ large enough, the function $\varphi$ satisfies

$$
-\frac{\partial \varphi}{\partial t}-\frac{\partial \varphi}{\partial \theta}-\mathcal{L} \varphi \geq 0 \quad \text { on }[0, T] \times\left(\overline{\mathcal{S}} \backslash D_{0}\right) .
$$

Indeed, a straightforward calculation shows that for all $(t, z, \theta) \in[0, T] \times\left(\overline{\mathcal{S}} \backslash D_{0}\right)$

$$
\begin{align*}
& -\frac{\partial \varphi}{\partial t}(t, z, \theta)-\frac{\partial \varphi}{\partial \theta}(t, z, \theta)-\mathcal{L} \varphi(t, z, \theta) \\
= & e^{\rho(T-t)} L_{M}(z)^{\gamma-2}\left[\left(\sqrt{\rho} L_{M}(z)+\frac{b \gamma}{2 \sqrt{\rho}} y p\right)^{2}+\left(\frac{\gamma(1-\gamma) \sigma^{2}}{2}-\frac{b^{2} \gamma^{2}}{4 \rho}\right) y^{2} p^{2}\right] \tag{3.10}
\end{align*}
$$

which is nonnegative under condition (3.9).
Fix some $(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}$. If $(z, \theta)=(0,0, p, \theta) \in D_{0}$, then we clearly have $v_{0}(t, z, \theta)=$ $U(0)$, and inequality (3.8) is trivial. Otherwise, if $(z, \theta) \in \overline{\mathcal{S}} \backslash D_{0}$, then the process $\left(Z^{0, t, z}, \Theta^{0, t, \theta}\right)$ satisfies $L_{M}\left(Z^{0, t, z}, \Theta^{0, t, \theta}\right)>0$. Indeed, denote by $\left(\bar{Z}^{t, z}, \bar{\Theta}^{t, \theta}\right)$ the process starting from $(z, \theta)$ at $t$ and associated with the strategy consisting in liquidating all stock shares at $t$. Then we have $\left(\bar{Z}_{s}^{t, z}, \bar{\Theta}_{s}^{t, \theta}\right) \in \overline{\mathcal{S}} \backslash D_{0}$ for all $s \in[t, T]$ and hence $L_{M}\left(\bar{Z}_{s}^{t, z}, \bar{\Theta}_{s}^{t, \theta}\right)>0$ for all $s \in[t, T]$. Using (3.5) we get $L_{M}\left(Z_{s}^{0, t, z}, \Theta_{s}^{0, t, \theta}\right) \geq L_{M}\left(\bar{Z}_{s}^{t, z}, \bar{\Theta}_{s}^{t, \theta}\right)>0$.

We can then apply Itô's formula to $\varphi\left(s, Z_{s}^{0, t, z}, \Theta_{s}^{0, t, \theta}\right)$ between $t$ and $T_{R}=\inf \left\{s \geq t:\left|Z_{s}^{0, t, z}\right|\right.$ $\geq R\} \wedge T$ :

$$
\begin{aligned}
\mathbb{E}\left[\varphi\left(T_{R}, Z_{T_{R}}^{0, t, z}, \Theta_{T_{R}}^{0, t, \theta}\right)\right] & =\varphi(t, z)+\mathbb{E}\left[\int_{t}^{T_{R}}\left(\frac{\partial \varphi}{\partial t}+\frac{\partial \varphi}{\partial \theta}+\mathcal{L} \varphi\right)\left(s, Z_{s}^{0, t, z}, \Theta_{s}^{0, t, \theta}\right) d s\right] \\
& \leq \varphi(t, z)
\end{aligned}
$$

(The stochastic integral term vanishes in expectation since the integrand is bounded before $T_{R}$.) By sending $R$ to infinity, we get by Fatou's lemma and since $\varphi(T, z, \theta)=L_{M}(z)^{\gamma}$

$$
\mathbb{E}\left[L_{M}\left(Z_{T}^{0, t, z}\right)^{\gamma}\right] \leq \varphi(t, z, \theta)
$$

We conclude with the growth condition (H5).
As a direct consequence of the previous proposition, we obtain the continuity of the value function on the boundary $\partial_{y} \mathcal{S}$, i.e., when we start with no stock shares.

Corollary 3.1. Assume that (H1) and (H5) hold. Then, the value function $v$ is continuous on $[0, T] \times \partial_{y} \mathcal{S}$, and we have

$$
v(t, z, \theta)=U(x) \quad \forall t \in[0, T],(z, \theta)=(x, 0, p, \theta) \in \partial_{y} \mathcal{S}
$$

In particular, we have $v(t, z, \theta)=U(0)=0$ for all $(t, z, \theta) \in[0, T] \times D_{0}$.
Proof. From the lower bound (2.10) and the upper bound in Proposition 3.1, we have for all $(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}$

$$
U(x+y p f(-y, \theta)) \leq v(t, z, \theta) \leq \mathbb{E}\left[U\left(L_{M}\left(Z_{T}^{0, t, z}\right)\right)\right]=\mathbb{E}\left[U\left(x+y P_{T}^{t, p}\right)\right]
$$

These two inequalities imply the required result.
The following result states the finiteness of the total number of shares and amount traded.
Proposition 3.2. Assume that (H1) and (H2) hold. Then, for any $\alpha=\left(\tau_{n}, \zeta_{n}\right)_{n \geq 0} \in$ $\mathcal{A}(t, z, \theta),(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}$, we have

$$
\sum_{n \geq 1}\left|\zeta_{n}\right|<\infty, \quad \sum_{n \geq 1}\left|\zeta_{n}\right| P_{\tau_{n}}<\infty, \quad \text { and } \sum_{n \geq 1}\left|\zeta_{n}\right| P_{\tau_{n}} f\left(\zeta_{n}, \Theta_{\tau_{n}^{-}}\right)<\infty \quad \text { a.s. }
$$

Proof. Fix $(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}$, and $\alpha=\left(\tau_{n}, \zeta_{n}\right)_{n \geq 0} \in \mathcal{A}(t, z, \theta)$. Observe first that the continuous part of the process $L_{M}(Z)$ is $L_{M}\left(Z^{0, t, z}\right)$, and we denote its jump at time $\tau_{n}$ by $\Delta L_{M}\left(Z_{\tau_{n}}\right)=L_{M}\left(Z_{\tau_{n}}\right)-L_{M}\left(Z_{\tau_{n}^{-}}\right)$. From the estimates (3.3) and (3.6) in Lemma 3.2, we then have a.s. for all $n \geq 1$

$$
\begin{aligned}
0 \leq L_{M}\left(Z_{\tau_{n}}\right) & =L_{M}\left(Z_{\tau_{n}}^{0, t, z}\right)+\sum_{k=1}^{n} \Delta L_{M}\left(Z_{\tau_{k}}\right) \\
& \leq L_{M}\left(Z_{\tau_{n}}^{0, t, z}\right)-\bar{\kappa} \sum_{k=1}^{n}\left|\zeta_{k}\right| P_{\tau_{k}}
\end{aligned}
$$

where we set $\hat{\kappa}=\min \left(\underline{\kappa_{a}}-1,1-\overline{\kappa_{b}}\right)>0$. We deduce that for all $n \geq 1$

$$
\sum_{k=1}^{n}\left|\zeta_{k}\right| P_{\tau_{k}} \leq \frac{1}{\hat{\kappa}} \sup _{s \in[t, T]} L_{M}\left(Z_{s}^{0, t, z}\right)=\frac{1}{\hat{\kappa}}\left(x+y \sup _{s \in[t, T]} P_{s}^{t, p}\right)<\infty \quad \text { a.s. }
$$

This shows the almost sure convergence of the series $\sum_{n}\left|\zeta_{n}\right| P_{\tau_{n}}$. Moreover, since the price process $P$ is continuous and strictly positive, we also obtain the convergence of the series $\sum_{n}\left|\zeta_{n}\right|$. Recalling that $f(e, \theta) \leq 1$ for all $e \leq 0$ and $\theta \in[0, T]$, we have for all $n \geq 1$

$$
\begin{align*}
\sum_{k=1}^{n}\left|\zeta_{k}\right| P_{\tau_{k}} f\left(\zeta_{k}, \Theta_{\tau_{k}^{-}}\right) & =\sum_{k=1}^{n} \zeta_{k} P_{\tau_{k}} f\left(\zeta_{k}, \Theta_{\tau_{k}^{-}}\right)+2 \sum_{k=1}^{n}\left|\zeta_{k}\right| P_{\tau_{k}} f\left(\zeta_{k}, \Theta_{\tau_{k}^{-}}\right) \mathbf{1}_{\zeta_{k} \leq 0} \\
& \leq \sum_{k=1}^{n} \zeta_{k} P_{\tau_{k}} f\left(\zeta_{k}, \Theta_{\tau_{k}^{-}}\right)+2 \sum_{k=1}^{n}\left|\zeta_{k}\right| P_{\tau_{k}} \tag{3.11}
\end{align*}
$$

On the other hand, we have

$$
\begin{aligned}
0 \leq L_{M}\left(Z_{\tau_{n}}\right) & =X_{\tau_{n}}+Y_{\tau_{n}} P_{\tau_{n}} \\
& =x-\sum_{k=1}^{n} \zeta_{k} P_{\tau_{k}} f\left(\zeta_{k}, \Theta_{\tau_{k}^{-}}\right)+\left(y+\sum_{k=1}^{n} \zeta_{k}\right) P_{\tau_{n}} .
\end{aligned}
$$

Together with (3.11), this implies that for all $n \geq 1$

$$
\sum_{k=1}^{n}\left|\zeta_{k}\right| P_{\tau_{k}} f\left(\zeta_{k}, \Theta_{\tau_{k}^{-}}\right) \leq x+\left(y+\sum_{k=1}^{n}\left|\zeta_{k}\right|\right) \sup _{s \in[t, T]} P_{s}^{t, p}+2 \sum_{k=1}^{n}\left|\zeta_{k}\right| P_{\tau_{k}}
$$

The convergence of the series $\sum_{n}\left|\zeta_{n}\right| P_{\tau_{n}} f\left(\zeta_{n}, \Theta_{\tau_{n}^{-}}\right)$follows therefore from the convergence of the series $\sum_{n}\left|\zeta_{n}\right|$ and $\sum_{n}\left|\zeta_{n}\right| P_{\tau_{n}}$.

As a consequence of the above results, we can now prove that in the optimal portfolio liquidation it suffices to restrict to a finite number of trading times, which are strictly increasing. Given a trading strategy $\alpha=\left(\tau_{n}, \zeta_{n}\right)_{n \geq 0} \in \mathcal{A}$, let us denote by $N(\alpha)$ the process counting the number of intervention times:

$$
N_{t}(\alpha)=\sum_{n \geq 1} \mathbf{1}_{\tau_{n} \leq t}, \quad 0 \leq t \leq T .
$$

We denote by $\mathcal{A}_{\ell}^{b}(t, z, \theta)$ the set of admissible trading strategies in $\mathcal{A}_{\ell}(t, z, \theta)$ with a finite number of trading times such that these trading times are strictly increasing, namely,

$$
\begin{aligned}
\mathcal{A}_{\ell}^{b}(t, z, \theta)=\left\{\alpha=\left(\tau_{n}, \zeta_{n}\right)_{n \geq 0}\right. & \in \mathcal{A}_{\ell}(t, z, \theta): \quad N_{T}(\alpha)<\infty \quad \text { a.s. } \\
& \text { and } \left.\tau_{n}<\tau_{n+1} \text { a.s., } 0 \leq n \leq N_{T}(\alpha)-1\right\} .
\end{aligned}
$$

For any $\alpha=\left(\tau_{n}, \zeta_{n}\right)_{n} \in \mathcal{A}_{\ell}^{b}(t, z, \theta)$, the associated state process $(Z, \Theta)$ satisfies $\Theta_{\tau_{n+1}^{-}}>0$, i.e., $\left(Z_{\tau_{n+1}^{-}}, \Theta_{\tau_{n+1}^{-}}\right) \in \overline{\mathcal{S}}^{*}:=\{(z, \theta) \in \overline{\mathcal{S}}: \theta>0\}$. We also set $\partial_{L} \mathcal{S}^{*}=\partial_{L} \mathcal{S} \cap \overline{\mathcal{S}}^{*}$.

Theorem 3.1. Assume that (H1), (H2), (H3), (H4), and (H5) hold. Then, we have

$$
\begin{equation*}
v(t, z, \theta)=\sup _{\alpha \in \mathcal{A}_{\ell}^{b}(t, z, \theta)} \mathbb{E}\left[U\left(X_{T}\right)\right], \quad(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}} . \tag{3.12}
\end{equation*}
$$

Moreover, we have

$$
\begin{equation*}
v(t, z, \theta)=\sup _{\alpha \in \mathcal{A}_{\ell_{+}}^{b}(t, z, \theta)} \mathbb{E}\left[U\left(X_{T}\right)\right], \quad(t, z, \theta) \in[0, T] \times\left(\overline{\mathcal{S}} \backslash \partial_{L} \mathcal{S}\right), \tag{3.13}
\end{equation*}
$$

where $\mathcal{A}_{\ell_{+}}^{b}(t, z, \theta)=\left\{\alpha \in \mathcal{A}_{\ell}^{b}(t, z, \theta):\left(Z_{s}, \Theta_{s}\right) \in\left(\overline{\mathcal{S}} \backslash \partial_{L} \mathcal{S}\right), t \leq s<T\right\}$.
Proof. Step 1. Fix $(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}$, and denote by $\overline{\mathcal{A}}_{\ell}^{b}(t, z, \theta)$ the set of admissible trading strategies in $\mathcal{A}_{\ell}(t, z, \theta)$ with a finite number of trading times:

$$
\overline{\mathcal{A}}_{\ell}^{b}(t, z, \theta)=\left\{\alpha=\left(\tau_{k}, \zeta_{k}\right)_{k \geq 0} \in \mathcal{A}_{\ell}(t, z, \theta): N_{T}(\alpha) \text { is bounded a.s. }\right\} .
$$

Given an arbitrary $\alpha=\left(\tau_{k}, \zeta_{k}\right)_{k \geq 0} \in \mathcal{A}_{\ell}(t, z, \theta)$ associated with the state process $(Z, \Theta)=$ $(X, Y, P, \Theta)$, let us consider the truncated trading strategy $\alpha^{(n)}=\left(\tau_{k}, \zeta_{k}\right)_{k \leq n} \cup\left(\tau_{n+1},-Y_{\tau_{n+1}^{-}}\right)$, which consists in liquidating all stock shares at time $\tau_{n+1}$. This strategy $\alpha^{(n)}$ lies in $\overline{\mathcal{A}}_{\ell}(t, z, \theta)$ and is associated with the state process denoted by $\left(Z^{(n)}, \Theta^{(n)}\right)$. We then have

$$
X_{T}^{(n)}-X_{T}=\sum_{k \geq n+1} \zeta_{k} P_{\tau_{k}} f\left(\zeta_{k}, \Theta_{\tau_{k}^{-}}\right)+Y_{\tau_{n+1}^{-}} P_{\tau_{n+1}} f\left(-Y_{\tau_{n+1}^{-}}, \Theta_{\tau_{n+1}^{-}}\right)
$$

Now, from Proposition 3.2, we have

$$
\sum_{k \geq n+1} \zeta_{k} P_{\tau_{k}} f\left(\zeta_{k}, \Theta_{\tau_{k}^{-}}\right) \longrightarrow 0 \quad \text { a.s. } \quad \text { when } \quad n \rightarrow \infty
$$

Moreover, since $0 \leq Y_{\tau_{n+1}^{-}}=Y_{\tau_{n}}$ goes to $Y_{T}=0$ as $n$ goes to infinity, by definition of $\alpha \in$ $\mathcal{A}_{\ell}(t, z, \theta)$, and recalling that $f$ is smaller than 1 on $\mathbb{R}_{-} \times[0, T]$, we deduce that

$$
\begin{aligned}
0 \leq Y_{\tau_{n+1}^{-}} P_{\tau_{n+1}} f\left(-Y_{\tau_{n+1}^{-}} \Theta_{\tau_{n+1}^{-}}\right) & \leq Y_{\tau_{n+1}^{-}} \sup _{s \in[t, T]} P_{s}^{t, p} \\
& \longrightarrow 0 \quad \text { a.s. } \quad \text { when } \quad n \rightarrow \infty
\end{aligned}
$$

This proves that $X_{T}^{(n)} \longrightarrow X_{T}$ a.s. when $n$ goes to infinity. From Proposition 3.1, the sequence $\left(U\left(X_{T}^{(n)}\right)\right)_{n \geq 1}$ is uniformly integrable, and we can apply the dominated convergence theorem
to get $\mathbb{E}\left[U\left(X_{T}^{(n)}\right)\right] \longrightarrow \mathbb{E}\left[U\left(X_{T}\right)\right]$ when $n$ goes to infinity. Since $\alpha$ is arbitrary in $\mathcal{A}_{\ell}(t, z, \theta)$, this shows that

$$
v(t, z, \theta) \leq \bar{v}^{b}(t, z, \theta):=\sup _{\alpha \in \overline{\mathcal{A}}_{\ell}^{b}(t, z, \theta)} \mathbb{E}\left[U\left(X_{T}\right)\right],
$$

and actually the equality $v=\bar{v}^{b}$ since the other inequality $\bar{v}^{b} \leq v$ is trivial from the inclusion $\overline{\mathcal{A}}_{\ell}^{b}(t, z, \theta) \subset \mathcal{A}_{\ell}(t, z, \theta)$.

Step 2 . Denote by $v^{b}$ the value function in the right-hand side (r.h.s.) of (3.12). It is clear that $v^{b} \leq \bar{v}^{b}=v$ since $\mathcal{A}_{\ell}^{b}(t, z, \theta) \subset \overline{\mathcal{A}}_{\ell}^{b}(t, z, \theta)$. To prove the reverse inequality we need first to study the behavior of optimal strategies at time $T$. Introduce the set

$$
\tilde{\mathcal{A}}_{\ell}^{b}(t, z, \theta)=\left\{\alpha=\left(\tau_{k}, \zeta_{k}\right)_{k} \in \mathcal{A}_{\ell}^{b}(t, z, \theta): \#\left\{k: \tau_{k}=T\right\} \leq 1\right\},
$$

and denote by $\tilde{v}^{b}$ the associated value function. Then we have $\tilde{v}^{b} \leq \bar{v}^{b}$. Indeed, let $\alpha=\left(\tau_{k}, \zeta_{k}\right)_{k}$ be some arbitrary element in $\overline{\mathcal{A}}_{\ell}^{b}(t, z, \theta),(t, z=(x, y, p), \theta) \in[0, T] \times \overline{\mathcal{S}}$. If $\alpha \in \tilde{\mathcal{A}}_{\ell}^{b}(t, z, \theta)$, then we have $\tilde{v}^{b}(t, z, \theta) \geq \mathbb{E}\left[U_{L}\left(Z_{T}, \Theta_{T}\right)\right]$, where $(Z, \Theta)$ denotes the process associated with $\alpha$. Suppose now that $\alpha \notin \tilde{\mathcal{A}}_{\ell}^{b}(t, z, \theta)$. Set $m=\max \left\{k: \tau_{k}<T\right\}$. Then define the stopping time $\tau^{\prime}:=\frac{\tau_{m}+T}{2}$ and the $\mathcal{F}_{\tau^{\prime}}$-measurable random variable $\zeta^{\prime}:=\operatorname{argmax}\left\{e f\left(e, T-\tau_{m}\right): e \geq-Y_{\tau_{m}}\right\}$. Define the strategy $\alpha^{\prime}=\left(\tau_{k}, \zeta_{k}\right)_{k \leq m} \cup\left(\tau^{\prime}, Y_{\tau_{m}}-\zeta^{\prime}\right) \cup\left(T, \zeta^{\prime}\right)$. From the construction of $\alpha^{\prime}$, we easily check that $\alpha^{\prime} \in \tilde{\mathcal{A}}^{b}(t, z, \theta)$ and $\mathbb{E}\left[U_{L}\left(Z_{T}, \Theta_{T}\right)\right] \leq \mathbb{E}\left[U_{L}\left(Z_{T}^{\prime}, \Theta_{T}^{\prime}\right)\right]$, where $\left(Z^{\prime}, \Theta^{\prime}\right)$ denotes the process associated with $\alpha^{\prime}$. Thus, $\tilde{v}^{b} \geq \bar{v}^{b}$.

We now prove that $v^{b} \geq \tilde{v}^{b}$. Let $\alpha=\left(\tau_{k}, \zeta_{k}\right)_{k}$ be some arbitrary element in $\tilde{\mathcal{A}}_{\ell}^{b}(t, z, \theta)$, $(t, z=(x, y, p), \theta) \in[0, T] \times \overline{\mathcal{S}}$. Denote by $N=N_{T}(\alpha)$ the a.s. finite number of trading times in $\alpha$. We set $m=\inf \left\{0 \leq k \leq N-1: \tau_{k+1}=\tau_{k}\right\}$ and $M=\sup \left\{m+1 \leq k \leq N: \tau_{k}=\tau_{m}\right\}$ with the convention that $\inf \emptyset=\sup \emptyset=N+1$. We then define $\alpha^{\prime}=\left(\tau_{k}^{\prime}, \zeta_{k}^{\prime}\right)_{0 \leq k \leq N-(M-m)+1}$ $\in \mathcal{A}$ by

$$
\left(\tau_{k}^{\prime}, \zeta_{k}^{\prime}\right)=\left\{\begin{array}{cl}
\left(\tau_{k}, \zeta_{k}\right) & \text { for } 0 \leq k<m, \\
\left(\tau_{m}=\tau_{M}, \sum_{k=m} \zeta_{k}\right) & \text { for } k=m \text { and } m<N, \\
\left(\tau_{k+M-m}, \zeta_{k+M-m}\right) & \text { for } m+1 \leq k \leq N-(M-m) \text { and } m<N, \\
\left(\tau^{\prime}, \sum_{l=m+1}^{M} \zeta_{l}\right) & \text { for } k=N-(M-m)+1,
\end{array}\right.
$$

where $\tau^{\prime}=\frac{\hat{\tau}+T}{2}$ with $\hat{\tau}=\max \left\{\tau_{k}: \tau_{k}<T\right\}$, and we denote by $\left(Z^{\prime}=\left(X^{\prime}, Y^{\prime}, P\right), \Theta^{\prime}\right)$ the associated state process. It is clear that $\left(Z_{s}^{\prime}, \Theta_{s}^{\prime}\right)=\left(Z_{s}, \Theta_{s}\right)$ for $t \leq s<\tau_{m}$, and so $X_{(\tau)^{\prime-}}^{\prime}=$ $X_{\left(\tau^{\prime}\right)^{-}}, \Theta_{\left(\tau^{\prime}\right)^{-}}^{\prime}=\Theta_{\left(\tau^{\prime}\right)^{-}}$. Moreover, since $\tau_{m}=\tau_{M}$, we have $\Theta_{\tau_{k}^{-}}=0$ for $m+1 \leq k \leq M$. From Lemma 3.1 (or Remark 3.1), this implies that $\zeta_{k} \leq 0$ for $m+1 \leq k \leq M$, and so $\zeta_{N-(M-m)+1}^{\prime}$ $=\sum_{k=m+1}^{M} \zeta_{k} \leq 0$. We also recall that immediate sales does not increase the cash holdings, so that $X_{\tau_{k}}=X_{\tau_{m}}$ for $m+1 \leq k \leq M$. We then get

$$
X_{T}^{\prime}=X_{T}-\zeta_{N-(M-m)+1}^{\prime} P_{\tau^{\prime}} f\left(\zeta_{N-(M-m)+1}^{\prime}, \Theta_{\left(\tau^{\prime}\right)^{-}}^{\prime}\right) \geq X_{T}
$$

Moreover, we have $Y_{T}^{\prime}=y+\sum_{k=1}^{N} \zeta_{k}=Y_{T}=0$. By construction, notice that $\tau_{0}^{\prime}<\cdots<\tau_{m+1}^{\prime}$. Given an arbitrary $\alpha \in \overline{\mathcal{A}}_{\ell}^{b}(t, z, \theta)$, we can then construct by induction a trading strategy $\alpha^{\prime} \in$
$\mathcal{A}_{\ell}^{b}(t, z, \theta)$ such that $X_{T}^{\prime} \geq X_{T}$ a.s. By the nondecreasing monotonicity of the utility function $U$, this yields

$$
\mathbb{E}\left[U\left(X_{T}\right)\right] \leq \mathbb{E}\left[U\left(X_{T}^{\prime}\right)\right] \leq v^{b}(t, z, \theta)
$$

Since $\alpha$ is arbitrary in $\tilde{\mathcal{A}}_{\ell}^{b}(t, z, \theta)$, we conclude that $\tilde{v}^{b} \leq v^{b}$, and thus $v=\bar{v}^{b}=\tilde{v}^{b}=v^{b}$.
Step 3. Fix now an element $(t, z, \theta) \in[0, T] \times\left(\overline{\mathcal{S}} \backslash \partial_{L} \mathcal{S}\right)$, and denote by $v_{+}$the r.h.s of (3.13). It is clear that $v \geq v_{+}$. Conversely, take some arbitrary $\alpha=\left(\tau_{k}, \zeta_{k}\right)_{k} \in \mathcal{A}_{\ell}^{b}(t, z, \theta)$, associated with the state process $(Z, \Theta)$, and denote by $N=N_{T}(\alpha)$ the finite number of trading times in $\alpha$. Consider the first time before $T$ when the liquidation value reaches zero; i.e., $\tau^{\alpha}=\inf \left\{t \leq s \leq T: L\left(Z_{s}, \Theta_{s}\right)=0\right\} \wedge T$ with the convention $\inf \emptyset=\infty$. We claim that there exists $1 \leq m \leq N+1$ (depending on $\omega$ and $\alpha$ ) such that $\tau^{\alpha}=\tau_{m}$, with the convention that $m=N+1, \tau_{N+1}=T$ if $\tau^{\alpha}=T$. On the contrary, there would exist $1 \leq k \leq N$ such that $\tau_{k}<\tau^{\alpha}<\tau_{k+1}$, and $L\left(Z_{\tau^{\alpha}}, \Theta_{\tau^{\alpha}}\right)=0$. Between $\tau_{k}$ and $\tau_{k+1}$, there is no trading, and so $\left(X_{s}, Y_{s}\right)=\left(X_{\tau_{k}}, Y_{\tau_{k}}\right), \Theta_{s}=s-\tau_{k}$ for $\tau_{k} \leq s<\tau_{k+1}$. We then get

$$
\begin{equation*}
L\left(Z_{s}, \Theta_{s}\right)=X_{\tau_{k}}+Y_{\tau_{k}} P_{s} f\left(-Y_{\tau_{k}}, s-\tau_{k}\right), \quad \tau_{k} \leq s<\tau_{k+1} . \tag{3.14}
\end{equation*}
$$

Moreover, since $0<L\left(Z_{\tau_{k}}, \Theta_{\tau_{k}}\right)=X_{\tau_{k}}$, and $L\left(Z_{\tau^{\alpha}}, \Theta_{\tau^{\alpha}}\right)=0$, we see with (3.14) for $s=\tau^{\alpha}$ that $Y_{\tau_{k}} P_{\tau^{\alpha}} f\left(-Y_{\tau_{k}}, \tau^{\alpha}-\tau_{k}\right)$ should necessarily be strictly negative: $Y_{\tau_{k}} P_{\tau^{\alpha}} f\left(-Y_{\tau_{k}}, \tau^{\alpha}-\tau_{k}\right)$ $<0$, a contradiction with the admissibility conditions and the nonnegative property of $f$.

We then have $\tau^{\alpha}=\tau_{m}$ for some $1 \leq m \leq N+1$. Observe that if $m \leq N$, i.e., $L\left(Z_{\tau_{m}}, \Theta_{\tau_{m}}\right)=$ 0 , then $U\left(L\left(Z_{T}, \Theta_{T}\right)\right)=0$. Indeed, suppose that $Y_{\tau_{m}}>0$ and $m \leq N$. From the admissibility condition, and by Itô's formula to $L(Z, \Theta)$ in (3.14) between $\tau^{\alpha}$ and $\tau_{m+1}^{-}$, we get

$$
0 \leq L\left(Z_{\tau_{m+1}^{-}}, \Theta_{\tau_{k+1}^{-}}\right)=L\left(Z_{\tau_{m+1}^{-}}, \Theta_{\tau_{m+1}^{-}}\right)-L\left(Z_{\tau^{\alpha}}, \Theta_{\tau^{\alpha}}\right)
$$

$$
\begin{equation*}
=\int_{\tau^{\alpha}}^{\tau_{m+1}} Y_{\tau_{m}} P_{s}\left[\beta\left(Y_{\tau_{m}}, s-\tau_{m}\right) d s+\sigma f\left(-Y_{\tau_{k}}, s-\tau_{m}\right) d W_{s}\right] \tag{3.15}
\end{equation*}
$$

where $\beta(y, \theta)=b f(-y, \theta)+\frac{\partial f}{\partial \theta}(-y, \theta)$ is bounded on $\mathbb{R}_{+} \times[0, T]$ by (H4)(ii). Since the integrand in the above stochastic integral with respect to the Brownian motion $W$ is strictly positive, and thus nonzero, we must have $\tau^{\alpha}=\tau_{m+1}$. Otherwise, there is a nonzero probability that the r.h.s. of (3.15) becomes strictly negative, a contradiction with the inequality (3.15).

Hence we get $Y_{\tau_{m}}=0$, and thus $L\left(Z_{\tau_{m+1}^{-}}, \Theta_{\tau_{m+1}^{-}}\right)=X_{\tau_{m}}=0$. From the Markov feature of the model and Corollary 3.1, we then have

$$
\mathbb{E}\left[U\left(L\left(Z_{T}, \Theta_{T}\right)\right) \mid \mathcal{F}_{\tau_{m}}\right] \leq v\left(\tau_{m}, Z_{\tau_{m}}, \Theta_{\tau_{m}}\right)=U\left(X_{\tau_{m}}\right)=0
$$

Since $U$ is nonnegative, this implies that $U\left(L\left(Z_{T}, \Theta_{T}\right)\right)=0$. Let us next consider the trading strategy $\alpha^{\prime}=\left(\tau_{k}^{\prime}, \zeta_{k}^{\prime}\right)_{0 \leq k \leq(m-1)} \in \mathcal{A}$ consisting in following $\alpha$ until time $\tau^{\alpha}$ and liquidating all stock shares at time $\tau^{\bar{\alpha}}=\tau_{m-1}$ and defined by

$$
\left(\tau_{k}^{\prime}, \zeta_{k}^{\prime}\right)=\left\{\begin{array}{cl}
\left(\tau_{k}, \zeta_{k}\right) & \text { for } 0 \leq k<m-1 \\
\left(\tau_{m-1},-Y_{\tau_{(m-1)}^{-}}\right) & \text {for } k=m-1
\end{array}\right.
$$

and we denote by $\left(Z^{\prime}, \Theta^{\prime}\right)$ the associated state process. It is clear that $\left(Z_{s}^{\prime}, \Theta_{s}^{\prime}\right)=\left(Z_{s}, \Theta_{s}\right)$ for $t \leq s<\tau_{m-1}$, and so $L\left(Z_{s}^{\prime}, \Theta_{s}^{\prime}\right)=L\left(Z_{s}, \Theta_{s}\right)>0$ for $t \leq s \leq \tau_{m-1}$. The liquidation at time $\tau_{m-1}($ for $m \leq N)$ yields $X_{\tau_{m-1}}=L\left(Z_{\tau_{m-1}^{-}}, \Theta_{\tau_{m-1}^{-}}\right)>0$, and $Y_{\tau_{m-1}}=0$. Since there is no more trading after time $\tau_{m-1}$, the liquidation value for $\tau_{m-1} \leq s \leq T$ is given by $L\left(Z_{s}, \Theta_{s}\right)$ $=X_{\tau_{m-1}}>0$. This shows that $\alpha^{\prime} \in \mathcal{A}_{\ell_{+}}^{b}(t, z, \theta)$. When $m=N+1$, we have $\alpha=\alpha^{\prime}$, and so $X_{T}^{\prime}=L\left(Z_{T}^{\prime}, \Theta_{T}^{\prime}\right)=L\left(Z_{T}, \Theta_{T}\right)=X_{T}$. For $m \leq N$, we have $U\left(X_{T}^{\prime}\right)=U\left(L\left(Z_{T}^{\prime}, \Theta_{T}^{\prime}\right)\right) \geq 0=$ $U\left(L\left(Z_{T}, \Theta_{T}\right)\right)=U\left(X_{T}\right)$. We then get $U\left(X_{T}^{\prime}\right) \geq U\left(X_{T}\right)$ a.s., and so

$$
\mathbb{E}\left[U\left(X_{T}\right)\right] \leq \mathbb{E}\left[U\left(X_{T}^{\prime}\right)\right] \leq v_{+}(t, z, \theta)
$$

Since $\alpha$ is arbitrary in $\overline{\mathcal{A}}_{\ell}^{b}(t, z, \theta)$, we conclude that $v \leq v_{+}$, and thus $v=v_{+}$.
Remark 3.2. If we suppose that the function $e \in \mathbb{R} \mapsto e f(e, \theta)$ is increasing for $\theta \in(0, T]$, we get the value of $v$ on the bound $\partial_{L} \mathcal{S}^{*}: v(t, z, \theta)=U(0)=0$ for $(t, z=(x, y, p), \theta) \in$ $[0, T] \times \partial_{L} \mathcal{S}^{*}$. Indeed, fix some point $(t, z=(x, y, p), \theta) \in[0, T] \times \partial_{L} \mathcal{S}^{*}$, consider an arbitrary $\alpha=\left(\tau_{k}, \zeta_{k}\right)_{k} \in \mathcal{A}_{\ell}^{b}(t, z, \theta)$ with state process $(Z, \Theta)$, and denote by $N$ the number of trading times. We distinguish two cases: (i) If $\tau_{1}=t$, then by Lemma 3.1 the transaction $\zeta_{1}$ is equal to $-y$, which leads to $Y_{\tau_{1}}=0$, and a liquidation value $L\left(Z_{\tau_{1}}, \Theta_{\tau_{1}}\right)=X_{\tau_{1}}=L(z, \theta)=0$. At the next trading date $\tau_{2}$ (if it exists), we get $X_{\tau_{2}^{-}}=Y_{\tau_{2}^{-}}=0$ with liquidation value $L\left(Z_{\tau_{2}^{-}}, \Theta_{\tau_{2}^{-}}\right)$ $=0$, and by again using Lemma 3.1 we see that after the transaction at $\tau_{2}$ we shall also obtain $X_{\tau_{2}}=Y_{\tau_{2}}=0$. By induction, this leads at the final trading time to $X_{\tau_{N}}=Y_{\tau_{N}}=0$ and finally to $X_{T}=Y_{T}=0$. (ii) If $\tau_{1}>t$, we claim that $y=0$. On the contrary, by arguing similarly as in (3.15) between $t$ and $\tau_{1}^{-}$, we have then proved that any admissible trading strategy $\alpha \in$ $\mathcal{A}_{\ell}^{b}(t, z, \theta)$ provides a final liquidation value $X_{T}=0$, and so

$$
\begin{equation*}
v(t, z, \theta)=U(0)=0 \quad \forall(t, z, \theta) \in[0, T] \times \partial_{L} \mathcal{S}^{*} . \tag{3.16}
\end{equation*}
$$

Comments on Theorem 3.1. The representation (3.12) of the optimal portfolio liquidation reveals interesting economical and mathematical features. It shows that the liquidation problem in a continuous-time illiquid market model with discrete-time orders and temporary price impact with the presence of a bid-ask spread as considered in this paper leads to nearly optimal trading strategies with a finite number of orders and with strictly increasing trading times. While most models dealing with trading strategies via an impulse control formulation assumed fixed transaction fees in order to justify the discrete nature of trading times, we prove rigorously in this paper that discrete-time trading appears naturally as a consequence of temporary price impact and bid-ask spread. Although the result is quite intuitive, its proof uses technical arguments. In particular the separation of intervention times in Step 2 could not be done in a single step. Indeed, the natural idea consisting in replacing the sequence $\left(\tau_{k-1}, \tau_{k}, \tau_{k+1}\right)$ with $\tau_{k-1}=\tau_{k}<\tau_{k+1}$ by $\left(\tau_{k-1}, \tau_{k}^{\prime}=\frac{\tau_{k-1}+\tau_{k+1}}{2}, \tau_{k+1}\right)$ is not possible since $\frac{\tau_{k-1}+\tau_{k+1}}{2}$ is not necessarily a stopping time. We therefore gather all the cumulated orders at the terminal time and construct a stopping time of the form $\frac{\tau_{N}+T}{2}$, where $\tau_{N}$ is the last time intervention such that $\tau_{N}<T$. This allows us to obtain a strategy, which provides a better gain and for which the intervention times are separated. We mention a recent related paper [21], which considers an impulse control problem with subadditive transaction costs where the investor can trade only finitely many times during the trading horizon $[0, T]$. In this case, the authors prove that the number of trading times has finite expectation.

The representation (3.13) shows that when we are in an initial state with strictly positive liquidation value we can restrict in the optimal portfolio liquidation problem to admissible trading strategies with strictly positive liquidation value up to time $T^{-}$. The relation (3.16) means that when the initial state has a zero liquidation value, which is not a result of an immediate trading time, the liquidation value will stay at zero until the final horizon.
4. Dynamic programming and viscosity properties. In what follows, conditions (H1), (H2), (H3), (H4), and (H5) stand in force and are not recalled in the statement of theorems and propositions.

We use a dynamic programming approach to derive the equation satisfied by the value function of our optimal portfolio liquidation problem. The dynamic programming principle (DPP) for impulse controls was frequently used starting from the works in [6] and then considered, e.g., in [32], [24], [20], or [30]. In our context (recall the expression (2.11) of the value function), this is formulated as follows.
$D P P$. For all $(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}$, we have

$$
\begin{equation*}
v(t, z, \theta)=\sup _{\alpha \in \mathcal{A}(t, z, \theta)} \mathbb{E}\left[v\left(\tau, Z_{\tau}, \Theta_{\tau}\right)\right], \tag{4.1}
\end{equation*}
$$

where $\tau=\tau(\alpha)$ is any stopping time valued in $[t, T]$ eventually depending on the strategy $\alpha$ in (4.1).

The corresponding dynamic programming Hamilton-Jacobi-Bellman (HJB) equation is a quasi-variational inequality (QVI) written as

$$
\begin{equation*}
\min \left[-\frac{\partial v}{\partial t}-\frac{\partial v}{\partial \theta}-\mathcal{L} v, v-\mathcal{H} v\right]=0 \quad \text { in }[0, T) \times \overline{\mathcal{S}} \tag{4.2}
\end{equation*}
$$

together with the relaxed terminal condition

$$
\begin{equation*}
\min \left[v-U_{L}, v-\mathcal{H} v\right]=0 \quad \text { in }\{T\} \times \overline{\mathcal{S}} . \tag{4.3}
\end{equation*}
$$

The rigorous derivation of the HJB equation satisfied by the value function from the DPP is achieved by means of the notion of viscosity solutions and is by now rather classical in the modern approach of stochastic control (see, e.g., the books [13], [25]). There are some specific features here related to the impulse control and the liquidation state constraint, and we recall in Appendix A definitions of (discontinuous) constrained viscosity solutions for parabolic QVIs. The first result of this section is stated as follows.

Proposition 4.1. The value function $v$ is a constrained viscosity solution to (4.2)-(4.3).
The proof of Proposition 4.1 is quite routine, following, for example, arguments in [25] or [8], and is omitted here.

In order to have a complete characterization of the value function through its HJB equation, we usually need a uniqueness result and thus a comparison principle for the QVI (4.2)(4.3). A key argument originally due to [17] for getting a uniqueness result for variational inequalities with impulse parts is to produce a strict viscosity supersolution. However, in our model, this is not possible. Indeed, suppose we can find a strict viscosity l.s.c. supersolution $w$ to (4.2), so that $(w-\mathcal{H} w)(t, z, \theta)>0$ on $[0, T) \times \mathcal{S}$. But for $z=(x, y, p)$ and $\theta=0$, we
have $\Gamma(z, 0, e)=(x, y+e, p)$ for any $e \in \mathcal{C}(z, 0)$. Since $0 \in \mathcal{C}(z, 0)$ we have $\mathcal{H} w(t, z, 0)=$ $\sup _{e \in[-y, 0]} w(t, x, y+e, p, 0) \geq w(t, z, 0)>\mathcal{H} w(t, z, 0)$, a contradiction. Actually, the main reason why one cannot obtain a strict supersolution is the absence of fixed cost in the impulse function $\Gamma$ or in the objective functional.

However, we can prove a weaker characterization of the value function in terms of minimal solution to its dynamic programming equation. The argument is based on a small perturbation of the gain functional. The proof is postponed until Appendix B.

Proposition 4.2. The value function $v$ is the minimal constrained viscosity solution in $\mathcal{G}_{\gamma}([0, T] \times \overline{\mathcal{S}})$ to (4.2)-(4.3) satisfying the boundary condition

$$
\begin{equation*}
\lim _{\left(t^{\prime}, z^{\prime}, \theta^{\prime}\right) \rightarrow(t, z, \theta)} v\left(t^{\prime}, z^{\prime}, \theta^{\prime}\right)=v(t, z, \theta)=U(0) \quad \forall(t, z, \theta) \in[0, T] \times D_{0} \tag{4.4}
\end{equation*}
$$

5. An approximating problem with fixed transaction fee. In this section, we consider a small variation of our original model by adding a fixed transaction fee $\varepsilon>0$ at each trading. This means that given a trading strategy $\alpha=\left(\tau_{n}, \zeta_{n}\right)_{n \geq 0}$, the controlled state process ( $Z=$ $(X, Y, P), \Theta)$ jumps at time $\tau_{n+1}$ are defined now by

$$
\begin{equation*}
\left(Z_{\tau_{n+1}}, \Theta_{\tau_{n+1}}\right)=\left(\Gamma_{\varepsilon}\left(Z_{\tau_{n+1}^{-}}, \Theta_{\tau_{n+1}^{-}}, \zeta_{n+1}\right), 0\right) \tag{5.1}
\end{equation*}
$$

where $\Gamma_{\varepsilon}$ is the function defined on $\mathbb{R} \times \mathbb{R}_{+} \times \mathbb{R}_{+}^{*} \times[0, T] \times \mathbb{R}$ into $\mathbb{R} \cup\{-\infty\} \times \mathbb{R} \times \mathbb{R}_{+}^{*}$ by

$$
\Gamma_{\varepsilon}(z, \theta, e)=\Gamma(z, \theta, e)-(\varepsilon, 0,0)=(x-e p f(e, \theta)-\varepsilon, y+e, p)
$$

for $z=(x, y, p)$. The dynamics of $(Z, \Theta)$ between trading dates is given as before. We also introduce a modified liquidation function $L_{\varepsilon}$ defined by

$$
L_{\varepsilon}(z, \theta)=\max [x, L(z, \theta)-\varepsilon], \quad(z, \theta)=(x, y, p, \theta) \in \mathbb{R} \times \mathbb{R}_{+} \times \mathbb{R}_{+}^{*} \times[0, T]
$$

The interpretation of this modified liquidation function is the following. Due to the presence of the transaction fee at each trading, it may be advantageous for the investor not to liquidate his position in stock shares (which would give him $L(z, \theta)-\varepsilon$ ), and rather bin his stock shares, by keeping only his cash amount (which would give him $x$ ). Hence, the investor chooses the better of these two possibilities, which induces a liquidation value $L_{\varepsilon}(z, \theta)$.

We then introduce the corresponding solvency region $\mathcal{S}_{\varepsilon}$ with its closure $\overline{\mathcal{S}}_{\varepsilon}=\mathcal{S}_{\varepsilon} \cup \partial \mathcal{S}_{\varepsilon}$ and boundary $\partial \mathcal{S}_{\varepsilon}=\partial_{y} \mathcal{S}_{\varepsilon} \cup \partial_{L} \mathcal{S}_{\varepsilon}$ :

$$
\begin{aligned}
\mathcal{S}_{\varepsilon} & =\left\{(z, \theta)=(x, y, p, \theta) \in \mathbb{R} \times \mathbb{R}_{+} \times \mathbb{R}_{+}^{*} \times[0, T]: y>0 \text { and } L_{\varepsilon}(z, \theta)>0\right\}, \\
\partial_{y} \mathcal{S}_{\varepsilon} & =\left\{(z, \theta)=(x, y, p, \theta) \in \mathbb{R} \times \mathbb{R}_{+} \times \mathbb{R}_{+}^{*} \times[0, T]: y=0 \text { and } L_{\varepsilon}(z, \theta) \geq 0\right\}, \\
\partial_{L} \mathcal{S}_{\varepsilon} & =\left\{(z, \theta)=(x, y, p, \theta) \in \mathbb{R} \times \mathbb{R}_{+} \times \mathbb{R}_{+}^{*} \times \mathbb{R}_{+}: L_{\varepsilon}(z, \theta)=0\right\} .
\end{aligned}
$$

We also introduce the corner lines of $\partial \mathcal{S}_{\varepsilon}$. For simplicity of presentation, we consider a temporary price impact function $f$ in the form

$$
f(e, \theta)=\tilde{f}\left(\frac{e}{\theta}\right)=\exp \left(\lambda \frac{e}{\theta}\right)\left(\kappa_{a} \mathbf{1}_{e>0}+\mathbf{1}_{e=0}+\kappa_{b} \mathbf{1}_{e<0}\right) \mathbf{1}_{\theta>0}
$$

where $0<\overline{\kappa_{b}}<1<\underline{\kappa_{a}}$, and $\lambda>0$. A straightforward analysis of the function $L$ shows that $y \mapsto L(x, y, p, \theta)$ is increasing on $[0, \theta / \lambda]$, that it is decreasing on $[\theta / \lambda, \infty)$ with $L(x, 0, p, \theta)$ $=x=L(x, \infty, p, \theta)$, and that $\max _{y>0} L(x, y, p, \theta)=L(x, \theta / \lambda, p, \theta)=x+p \frac{\theta}{\lambda} \tilde{f}(-1 / \lambda)$. We first get the form of the $\operatorname{sets} \mathcal{C}(z, \theta): \mathcal{C}(z, \theta)=[-y, \bar{e}(z, \theta)]$, where the function $\bar{e}$ is defined in Lemma 3.1. We then distinguish two cases: (i) If $p \frac{\theta}{\lambda} \tilde{f}(-1 / \lambda)<\varepsilon$, then $L_{\varepsilon}(x, y, p, \theta)=$ $x$. (ii) If $p \frac{\theta}{\lambda} \tilde{f}(-1 / \lambda) \geq \varepsilon$, then there exist a unique $y_{1}(p, \theta) \in(0, \theta / \lambda]$ and a unique $y_{2}(p, \theta)$ $\in[\theta / \lambda, \infty)$ such that $L\left(x, y_{1}(p, \theta), p, \theta\right)=L\left(x, y_{2}(p, \theta), p, \theta\right)=x, L_{\varepsilon}(x, y, p, \theta)=x$ for $y \in$ $\left[0, y_{1}(p, \theta)\right) \cup\left(y_{2}(p, \theta), \infty\right)$, and $L_{\varepsilon}(x, y, p, \theta)=L(x, y, p, \theta)-\varepsilon$ for $y \in\left[y_{1}(p, \theta), y_{2}(p, \theta)\right]$. We then denote

$$
\begin{aligned}
D_{0} & =\{0\} \times\{0\} \times \mathbb{R}_{+}^{*} \times[0, T]=\partial_{y} \mathcal{S}_{\varepsilon} \cap \partial_{L} \mathcal{S}_{\varepsilon} \\
D_{1, \varepsilon} & =\left\{\left(0, y_{1}(p, \theta), p, \theta\right): p \frac{\theta}{\lambda} \tilde{f}\left(\frac{-1}{\lambda}\right) \geq \varepsilon, \theta \in[0, T]\right\} \\
D_{2, \varepsilon} & =\left\{\left(0, y_{2}(p, \theta), p, \theta\right): p \frac{\theta}{\lambda} \tilde{f}\left(\frac{-1}{\lambda}\right) \geq \varepsilon, \theta \in[0, T]\right\}
\end{aligned}
$$

Notice that the inner normal vectors at the corner lines $D_{1, \varepsilon}$ and $D_{2, \varepsilon}$ form an acute angle (positive scalar product), while we have a right angle at the corner $D_{0}$. We represent in Figure 4 the graph of $\mathcal{S}_{\varepsilon}$ in the plane $(x, y)$ for different values of $\varepsilon$, in Figure 5 the graph of $\mathcal{S}_{\varepsilon}$ in the space $(x, y, p)$, and in Figure 6 the graph of $\mathcal{S}_{\varepsilon}$ in the space $(x, y, \theta)$.

Next, we define the set of admissible trading strategies as follows. Given $(t, z, \theta) \in[0, T] \times$ $\overline{\mathcal{S}}_{\varepsilon}$, we say that the impulse control $\alpha$ is admissible, denoted by $\alpha \in \mathcal{A}^{\varepsilon}(t, z, \theta)$, if $\tau_{0}=t-\theta, \tau_{n}$ $\geq t, n \geq 1$, and the controlled state process $\left(Z^{\varepsilon}, \Theta\right)$ solution to (2.1)-(2.2)-(2.3)-(2.7)-(5.1), with an initial state $\left(Z_{t^{-}}^{\varepsilon}, \Theta_{t^{-}}\right)=(z, \theta)$ (and the convention that $\left(Z_{t}^{\varepsilon}, \Theta_{t}\right)=(z, \theta)$ if $\tau_{1}>$ $t)$, satisfies $\left(Z_{s}^{\varepsilon}, \Theta_{s}\right) \in[0, T] \times \overline{\mathcal{S}}_{\varepsilon}$ for all $s \in[t, T]$. Here, we stress the dependence of $Z^{\varepsilon}=$ $\left(X^{\varepsilon}, Y, P\right)$ in $\varepsilon$ appearing in the transaction function $\Gamma_{\varepsilon}$, and we notice that it affects only the cash component. Notice that $\mathcal{A}^{\varepsilon}(t, z, \theta)$ is nonempty for any $(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}_{\varepsilon}$. Indeed, for $(z, \theta) \in \overline{\mathcal{S}}_{\varepsilon}$ with $z=(x, y, p)$, i.e., $L_{\varepsilon}(z, \theta)=\max (x, L(z, \theta)-\varepsilon) \geq 0$, we distinguish two cases: (i) if $x \geq 0$, then by doing no transactions, the associated state process $\left(Z^{\varepsilon}=\left(X^{\varepsilon}, Y, P\right), \Theta\right)$ satisfies $X_{s}^{\varepsilon}=x \geq 0, t \leq s \leq T$, and thus this zero transaction is admissible; (ii) if $L(z, \theta)-\varepsilon \geq$ 0 , then by liquidating all the stock shares immediately, and doing nothing else, the associated state process satisfies $X_{s}^{\varepsilon}=L(z, \theta)-\varepsilon, Y_{s}=0$, and thus $L_{\varepsilon}\left(Z_{s}^{\varepsilon}, \Theta_{s}\right)=X_{s}^{\varepsilon} \geq 0, t \leq s \leq T$, which shows that this immediate transaction is admissible.

Given the utility function $U$ on $\mathbb{R}_{+}$and the liquidation utility function defined on $\overline{\mathcal{S}}_{\varepsilon}$ by $U_{L_{\varepsilon}}(z, \theta)=U\left(L_{\varepsilon}(z, \theta)\right)$, we then consider the associated optimal portfolio liquidation problem defined via its value function by

$$
\begin{equation*}
v_{\varepsilon}(t, z, \theta)=\sup _{\alpha \in \mathcal{A}^{\varepsilon}(t, z, \theta)} \mathbb{E}\left[U_{L_{\varepsilon}}\left(Z_{T}^{\varepsilon}, \Theta_{T}\right)\right], \quad(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}_{\varepsilon} \tag{5.2}
\end{equation*}
$$

Notice that when $\varepsilon=0$, the above problem reduces to the optimal portfolio liquidation problem described in section 2 , and in particular $v_{0}=v$. The main purpose of this section is to provide a unique PDE characterization of the value functions $v_{\varepsilon}, \varepsilon>0$, and to prove that the sequence $\left(v_{\varepsilon}\right)_{\varepsilon}$ converges to the original value function $v$ as $\varepsilon$ goes to zero.


Figure 4. Domain $\mathcal{S}_{\varepsilon}$ in the nonhatched zone for fixed $p=1$ and $\theta=1$ and $\varepsilon$ evolving from 0.1 to 0.4. Here $\kappa_{b}=0.9$ and $f(e, \theta)=\kappa_{b} \exp \left(\frac{e}{\theta}\right)$ for $e<0$. Notice that for $\varepsilon$ large enough, $\mathcal{S}_{\varepsilon}$ is equal to open orthant $\mathbb{R}_{+}^{*} \times \mathbb{R}_{+}^{*}$.

We define the set of admissible transactions in the model with fixed transaction fee by

$$
\mathcal{C}_{\varepsilon}(z, \theta)=\left\{e \in \mathbb{R}:\left(\Gamma_{\varepsilon}(z, \theta, e), 0\right) \in \overline{\mathcal{S}}_{\varepsilon}\right\}, \quad(z, \theta) \in \overline{\mathcal{S}}_{\varepsilon}
$$

A calculation similar to that in Lemma 3.1 shows that for $(z, \theta) \in \overline{\mathcal{S}}_{\varepsilon}, z=(x, y, p)$,

$$
\mathcal{C}_{\varepsilon}(z, \theta)=\left\{\begin{array}{cl}
{\left[-y, \bar{e}_{\varepsilon}(z, \theta)\right]} & \text { if } \theta>0 \text { or } x \geq \varepsilon \\
\emptyset & \text { if } \theta=0 \text { and } x<\varepsilon
\end{array}\right.
$$

where $\bar{e}(z, \theta)=\sup \{e \in \mathbb{R}: \operatorname{ep} \tilde{f}(e / \theta) \leq x-\varepsilon\}$ if $\theta>0$ and $\bar{e}(z, 0)=0$ if $x \geq \varepsilon$. Here, the set $\left[-y, \bar{e}_{\varepsilon}(z, \theta)\right]$ should be viewed as empty when $\bar{e}(z, \theta)<y$, i.e., $x+p y \tilde{f}(-y / \theta)-\varepsilon<0$. We also easily check that $\mathcal{C}_{\varepsilon}$ is continuous for the Hausdorff metric. We then consider the impulse operator $\mathcal{H}_{\varepsilon}$ defined by

$$
\mathcal{H}_{\varepsilon} w(t, z, \theta)=\sup _{e \in \mathcal{C}_{\varepsilon}(z, \theta)} w\left(t, \Gamma_{\varepsilon}(z, \theta, e), 0\right), \quad(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}_{\varepsilon}
$$

for any locally bounded function $w$ on $[0, T] \times \overline{\mathcal{S}}_{\varepsilon}$, with the convention that $\mathcal{H}_{\varepsilon} w(t, z, \theta)=-\infty$ when $\mathcal{C}_{\varepsilon}(z, \theta)=\emptyset$.


Figure 5. Lower bound of the domain $\mathcal{S}_{\varepsilon}$ for fixed $\theta=1$ and $f(e, \theta)=\kappa_{b} \exp \left(\frac{e}{\theta}\right)$ for $e<0$. Notice that when $p$ is fixed, we obtain Figure 4.


Figure 6. Lower bound of the domain $\mathcal{S}_{\varepsilon}$ for fixed $p=1$ and $\varepsilon=0.2$. Here $\kappa_{b}=0.9$ and $f(e, \theta)=\kappa_{b} \exp \left(\frac{e}{\theta}\right)$ for $e<0$. Notice that when $\theta$ is fixed, we obtain Figure 4.

Next, consider again the Merton liquidation function $L_{M}$, and observe similarly as in (3.7) that

$$
\begin{align*}
L_{M}\left(\Gamma_{\varepsilon}(z, \theta, e)\right)-L_{M}(z) & =e p(1-f(e, \theta))-\varepsilon \\
& \leq-\varepsilon \quad \forall(z, \theta) \in \overline{\mathcal{S}}_{\varepsilon}, \quad e \in \mathbb{R} \tag{5.3}
\end{align*}
$$

This implies, in particular, that

$$
\begin{equation*}
\mathcal{H}_{\varepsilon} L_{M}<L_{M} \quad \text { on } \overline{\mathcal{S}}_{\varepsilon} . \tag{5.4}
\end{equation*}
$$

Since $L_{\varepsilon} \leq L_{M}$, we observe from (5.3) that if $(z, \theta) \in \mathcal{N}_{\varepsilon}:=\left\{(z, \theta) \in \overline{\mathcal{S}}_{\varepsilon}: L_{M}(z)<\varepsilon\right\}$, then $\mathcal{C}_{\varepsilon}(z, \theta)=\emptyset$. Moreover, we deduce from (5.3) that for all $\alpha=\left(\tau_{n}, \zeta_{n}\right)_{n \geq 0} \in \mathcal{A}^{\varepsilon}(t, z, \theta)$ associated with the state process $(Z, \Theta),(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}_{\varepsilon}$

$$
\begin{aligned}
0 \leq L_{M}\left(Z_{T}\right) & =L_{M}\left(Z_{T}^{0, t, z}\right)+\sum_{n \geq 0} \Delta L_{M}\left(Z_{\tau_{n}}\right) \\
& \leq L_{M}\left(Z_{T}^{0, t, z}\right)-\varepsilon N_{T}(\alpha),
\end{aligned}
$$

where we recall that $N_{T}(\alpha)$ is the number of trading times over the whole horizon $T$. This shows that

$$
N_{T}(\alpha) \leq \frac{1}{\varepsilon} L_{M}\left(Z_{T}^{0, t, z}\right)<\infty \quad \text { a.s. }
$$

In other words, we see that, under the presence of fixed transaction fee, the number of intervention times over a finite interval for an admissible trading strategy is finite a.s.

The dynamic programming equation associated with the control problem (5.2) is

$$
\begin{align*}
\min \left[-\frac{\partial w}{\partial t}-\frac{\partial w}{\partial \theta}-\mathcal{L} w, w-\mathcal{H}_{\varepsilon} w\right]=0 & \text { in }[0, T) \times \overline{\mathcal{S}}_{\varepsilon}  \tag{5.5}\\
\min \left[w-U_{L_{\varepsilon}}, w-\mathcal{H}_{\varepsilon} w\right]=0 & \text { in }\{T\} \times \overline{\mathcal{S}}_{\varepsilon} \tag{5.6}
\end{align*}
$$

The main result of this section is stated as follows.
Theorem 5.1. (1) The sequence $\left(v_{\varepsilon}\right)_{\varepsilon}$ is nonincreasing and converges pointwise on $[0, T] \times$ $\left(\overline{\mathcal{S}} \backslash \partial_{L} \mathcal{S}\right)$ towards $v$ as $\varepsilon$ goes to zero.
(2) For any $\varepsilon>0$, the value function $v_{\varepsilon}$ is continuous on $[0, T) \times \mathcal{S}_{\varepsilon}$, and is the unique (in $[0, T) \times \mathcal{S}_{\varepsilon}$ ) constrained viscosity solution to (5.5)-(5.6), satisfying the growth condition

$$
\begin{equation*}
\left|v_{\varepsilon}(t, z, \theta)\right| \leq K\left(1+L_{M}(z)^{\gamma}\right) \quad \forall(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}_{\varepsilon} \tag{5.7}
\end{equation*}
$$

for some positive constant $K$ and the boundary condition

$$
\begin{align*}
\lim _{\left(t^{\prime}, z^{\prime}, \theta^{\prime}\right) \rightarrow(t, z, \theta)} v_{\varepsilon}\left(t^{\prime}, z^{\prime}, \theta^{\prime}\right) & =v(t, z, \theta) \\
& =U(0) \quad \forall(t, z=(0,0, p), \theta) \in[0, T] \times D_{0} . \tag{5.8}
\end{align*}
$$

We first prove rigorously the convergence of the sequence of value functions $\left(v_{\varepsilon}\right)$. The proof relies, in particular, on the discrete-time feature of nearly optimal trading strategies for
the original value function $v$; see Theorem 3.1. There are technical difficulties related to the dependence on $\varepsilon$ of the solvency constraint via the liquidation function $L_{\varepsilon}$ when passing to the limit $\varepsilon \rightarrow 0$.

Proof of Theorem 5.1. (1) Notice that for any $0<\varepsilon_{1} \leq \varepsilon_{2}$, we have $L_{\varepsilon_{2}} \leq L_{\varepsilon_{1}} \leq L$, $\mathcal{A}^{\varepsilon_{2}}(t, z, \theta) \subset \mathcal{A}^{\varepsilon_{1}}(t, z, \theta) \subset \mathcal{A}(t, z, \theta)$, for $t \in[0, T],(z, \theta) \in \overline{\mathcal{S}}_{\varepsilon_{2}} \subset \overline{\mathcal{S}}_{\varepsilon_{1}} \subset \overline{\mathcal{S}}$, and for $\alpha \in$ $\mathcal{A}^{\varepsilon_{2}}(t, z, \theta), L_{\varepsilon_{2}}\left(Z^{\varepsilon_{2}}, \Theta\right) \leq L_{\varepsilon_{2}}\left(Z^{\varepsilon_{1}}, \Theta\right) \leq L_{\varepsilon_{1}}\left(Z^{\varepsilon_{1}}, \Theta\right) \leq L(Z, \Theta)$. This shows that the sequence $\left(v_{\varepsilon}\right)$ is nonincreasing, and is upper bounded by the value function $v$ without transaction fee, so that

$$
\begin{equation*}
\lim _{\varepsilon \downarrow 0} v_{\varepsilon}(t, z, \theta) \leq v(t, z, \theta) \quad \forall(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}} . \tag{5.9}
\end{equation*}
$$

Fix now some point $(t, z, \theta) \in[0, T] \times\left(\overline{\mathcal{S}} \backslash \partial_{L} \mathcal{S}\right)$. From the representation (3.13) of $v(t, z, \theta)$, there exists, for any $n \geq 1$, an $1 / n$-optimal control $\alpha^{(n)}=\left(\tau_{k}^{(n)}, \zeta_{k}^{(n)}\right)_{k} \in \mathcal{A}_{\ell_{+}}^{b}(t, z, \theta)$ with associated state process $\left(Z^{(n)}=\left(X^{(n)}, Y^{(n)}, P\right), \Theta^{(n)}\right)$ and number of trading times $N^{(n)}$ :

$$
\begin{equation*}
\mathbb{E}\left[U\left(X_{T}^{(n)}\right)\right] \geq v(t, z, \theta)-\frac{1}{n} \tag{5.10}
\end{equation*}
$$

We denote by $\left.\left(Z^{\varepsilon,(n)}, \Theta^{(n)}\right)=\left(X^{\varepsilon,(n)}, Y^{(n)}, P\right), \Theta^{(n)}\right)$ the state process controlled by $\alpha^{(n)}$ in the model with transaction fee $\varepsilon$ (only the cash component is affected by $\varepsilon$ ), and we observe that for all $t \leq s \leq T$

$$
\begin{equation*}
X_{s}^{\varepsilon,(n)}=X_{s}^{(n)}-\varepsilon N_{s}^{(n)} \nearrow X_{s}^{(n)} \quad \text { as } \varepsilon \text { goes to zero. } \tag{5.11}
\end{equation*}
$$

Given $n$, we consider the family of stopping times

$$
\sigma_{\varepsilon}^{(n)}=\inf \left\{s \geq t: L\left(Z_{s}^{\varepsilon,(n)}, \Theta_{s}^{(n)}\right) \leq \varepsilon\right\} \wedge T, \quad \varepsilon>0 .
$$

Let us prove that

$$
\begin{equation*}
\lim _{\varepsilon \searrow 0} \sigma_{\varepsilon}^{(n)}=T \text { a.s. } \tag{5.12}
\end{equation*}
$$

Observe that for $0<\varepsilon_{1} \leq \varepsilon_{2}, X_{s}^{\varepsilon_{2},(n)} \leq X_{s}^{\varepsilon_{1},(n)}$, and so $L\left(Z_{s}^{\varepsilon_{2},(n)}, \Theta_{s}\right) \leq L\left(Z_{s}^{\varepsilon_{1},(n)}, \Theta_{s}\right)$ for $t \leq s \leq T$. This clearly implies that the sequence $\left(\sigma_{\varepsilon}^{(n)}\right)_{\varepsilon}$ is nonincreasing. Since this sequence is bounded by $T$, it admits a limit, denoted by $\sigma_{0}^{(n)}=\lim _{\varepsilon \downarrow 0} \uparrow \sigma_{\varepsilon}^{(n)}$. Now, by definition of $\sigma_{\varepsilon}^{(n)}$, we have $L\left(Z_{\sigma_{\varepsilon}^{(n)}}^{\varepsilon,(n)}, \Theta_{\sigma_{\varepsilon}^{(n)}}^{(n)}\right) \leq \varepsilon$ for all $\varepsilon>0$. By sending $\varepsilon$ to zero, we then get with (5.11)

$$
L\left(Z_{\sigma_{0}^{(n),-}}^{(n)}, \Theta_{\sigma_{0}^{(n),-}}^{(n)}\right)=0 \quad \text { a.s. }
$$

Recalling the definition of $\mathcal{A}_{\ell_{+}}^{b}(t, z, \theta)$, this implies that $\sigma_{0}^{(n)}=\tau_{k}^{(n)}$ for some $k \in\left\{1, \ldots, N^{(n)}+\right.$ $1\}$ with the convention $\tau_{N^{(n)}+1}^{(n)}=T$. If $k \leq N^{(n)}$, arguing as in (3.15), we get a contradiction with the solvency constraints. Hence we get $\sigma_{0}^{(n)}=T$.

Consider now the trading strategy $\tilde{\alpha}^{\varepsilon,(n)} \in \mathcal{A}$ consisting in following $\alpha^{(n)}$ until time $\sigma_{\varepsilon}^{(n)}$ and liquidating all the stock shares at time $\sigma_{\varepsilon}^{(n)}$, i.e.,

$$
\tilde{\alpha}^{\varepsilon,(n)}=\left(\tau_{k}^{(n)}, \zeta_{k}^{(n)}\right) \mathbf{1}_{\tau_{k}<\sigma_{\varepsilon}^{(n)}} \cup\left(\sigma_{\varepsilon}^{(n)},-Y_{\sigma_{\varepsilon}^{(n),-}}\right)
$$

We denote by ( $\left.\tilde{Z}^{\varepsilon,(n)}=\left(\tilde{X}^{\varepsilon,(n)}, \tilde{Y}^{\varepsilon,(n)}, P\right), \tilde{\Theta}^{\varepsilon,(n)}\right)$ the associated state process in the market with transaction fee $\varepsilon$. By construction, we have for all $t \leq s<\sigma_{\varepsilon}^{(n)}, L\left(\tilde{Z}_{s}^{\varepsilon,(n)}, \tilde{\Theta}_{s}^{\varepsilon,(n)}\right)=$ $L\left(Z_{s}^{\varepsilon,(n)}, \Theta_{s}^{(n)}\right) \geq \varepsilon$, and thus $L_{\varepsilon}\left(\tilde{Z}_{s}^{\varepsilon,(n)}, \tilde{\Theta}_{s}^{\varepsilon,(n)}\right) \geq 0$. At the transaction time $\sigma_{\varepsilon}^{(n)}$, we then have $\tilde{X}_{\sigma_{\varepsilon}^{(n)}}^{\varepsilon,(n)}=L\left(\tilde{Z}_{\sigma_{\varepsilon}^{(n),-}}^{\varepsilon,(n)}, \tilde{\Theta}_{\sigma_{\varepsilon}^{(n),-}}^{\varepsilon,(n)}\right)-\varepsilon=L\left(Z_{\sigma_{\varepsilon}^{\varepsilon,(n),-}}^{(n)}, \Theta_{\sigma_{\varepsilon}^{(n),-}}^{(n)}\right)-\varepsilon, \tilde{Y}_{\sigma_{\varepsilon}^{(n)}}^{\varepsilon,(n)}=0$. After time $\sigma_{\varepsilon}^{(n)}$, there are no more transactions in $\tilde{\alpha}^{\varepsilon,(n)}$, and so

$$
\begin{align*}
\tilde{X}_{s}^{\varepsilon,(n)} & =\tilde{X}_{\sigma_{\varepsilon}^{(n)}}^{\varepsilon,(n)}=L\left(Z_{\sigma_{\varepsilon}^{\varepsilon,(n),-}}^{(n)}, \Theta_{\sigma_{\varepsilon}^{(n),-}}^{(n)}\right)-\varepsilon \geq 0,  \tag{5.13}\\
\tilde{Y}_{s}^{\varepsilon,(n)} & =\tilde{Y}_{\sigma_{\varepsilon}^{\varepsilon,(n)}}^{\varepsilon,(n)}=0, \quad \sigma_{\varepsilon}^{(n)} \leq s \leq T, \tag{5.14}
\end{align*}
$$

and thus $L_{\varepsilon}\left(\tilde{Z}_{s}^{\varepsilon,(n)}, \tilde{\Theta}_{s}^{\varepsilon,(n)}\right)=\tilde{X}_{s}^{\varepsilon,(n)} \geq 0$ for $\sigma_{\varepsilon}^{(n)} \leq s \leq T$. This shows that $\tilde{\alpha}^{\varepsilon,(n)}$ lies in $\mathcal{A}^{\varepsilon}(t, z, \theta)$, and thus by definition of $v_{\varepsilon}$

$$
\begin{equation*}
v_{\varepsilon}(t, z) \geq \mathbb{E}\left[U_{L_{\varepsilon}}\left(\tilde{Z}_{T}^{\varepsilon,(n)}, \tilde{\Theta}_{T}^{\varepsilon,(n)}\right)\right] \tag{5.15}
\end{equation*}
$$

Let us check that, given $n$,

$$
\begin{equation*}
\lim _{\varepsilon \downarrow 0} L_{\varepsilon}\left(\tilde{Z}_{T}^{\varepsilon,(n)}, \tilde{\Theta}_{T}^{\varepsilon,(n)}\right)=X_{T}^{(n)} \quad \text { a.s. } \tag{5.16}
\end{equation*}
$$

To alleviate notation, we set $N=N_{T}^{(n)}$ as the total number of trading times of $\alpha^{(n)}$. If the last trading time of $\alpha^{(n)}$ occurs strictly before $T$, then we do not trade anymore until the final horizon $T$, and so

$$
\begin{equation*}
X_{T}^{(n)}=X_{\tau_{N}}^{(n)} \quad \text { and } \quad Y_{T}^{(n)}=Y_{\tau_{N}}^{(n)}=0 \quad \text { on }\left\{\tau_{N}<T\right\} \tag{5.17}
\end{equation*}
$$

By (5.12), we have for $\varepsilon$ small enough $\sigma_{\varepsilon}^{(n)}>\tau_{N}$, and so $\tilde{X}_{\sigma_{\varepsilon}^{(n),-}}^{\varepsilon,(n)}=X_{\tau_{N}}^{\varepsilon,(n)}, \tilde{Y}_{\sigma_{\varepsilon}^{(n),-}}^{\varepsilon,(n)}=Y_{\tau_{N}}^{(n)}=$ 0. The final liquidation at time $\sigma_{\varepsilon}^{(n)}$ yields $\tilde{X}_{T}^{\varepsilon,(n)}=\tilde{X}_{\sigma_{\varepsilon}^{(n)}}^{\varepsilon,(n)}=\tilde{X}_{\sigma_{\varepsilon}^{\varepsilon(n),-}}^{\varepsilon,(n)}-\varepsilon=X_{\tau_{N}}^{\varepsilon,(n)}-\varepsilon$, and $\tilde{Y}_{T}^{\varepsilon,(n)}=\tilde{Y}_{\sigma_{\varepsilon}^{(n)}}^{\varepsilon,(n)}=0$. We then obtain

$$
\begin{aligned}
L_{\varepsilon}\left(\tilde{Z}_{T}^{\varepsilon,(n)}, \tilde{\Theta}_{T}^{\varepsilon,(n)}\right) & =\max \left(\tilde{X}_{T}^{\varepsilon,(n)}, L\left(\tilde{Z}_{T}^{\varepsilon,(n)}, \tilde{\Theta}_{T}^{\varepsilon,(n)}\right)-\varepsilon\right) \\
& =\tilde{X}_{T}^{\varepsilon,(n)}=X_{\tau_{N}}^{\varepsilon,(n)}-\varepsilon \quad \text { on }\left\{\tau_{N}<T\right\} \\
& =X_{T}^{(n)}-(1+N) \varepsilon \quad \text { on }\left\{\tau_{N}<T\right\}
\end{aligned}
$$

by (5.11) and (5.17), which shows that the convergence in (5.16) holds on $\left\{\tau_{N}<T\right\}$. If the last trading of $\alpha^{(n)}$ occurs at time $T$, this means that we liquidate all stock shares at $T$, and so

$$
\begin{equation*}
X_{T}^{(n)}=L\left(Z_{T^{-}}^{(n)}, \Theta_{T^{-}}^{(n)}\right), \quad Y_{T}^{(n)}=0 \quad \text { on }\left\{\tau_{N}=T\right\} \tag{5.18}
\end{equation*}
$$

On the other hand, by (5.13)-(5.14), we have

$$
\begin{aligned}
L_{\varepsilon}\left(\tilde{Z}_{T}^{\varepsilon,(n)}, \tilde{\Theta}_{T}^{\varepsilon,(n)}\right)=\tilde{X}_{T}^{\varepsilon,(n)} & =L\left(Z_{\sigma_{\varepsilon}^{\varepsilon,(n),-}}^{(n)}, \Theta_{\sigma_{\varepsilon}^{(n),-}}^{(n)}\right)-\varepsilon \\
& \longrightarrow L\left(Z_{T^{-}}^{(n)}, \Theta_{T^{-}}^{(n)}\right) \quad \text { as } \varepsilon \text { goes to zero }
\end{aligned}
$$

by (5.12). Together with (5.18), this implies that the convergence in (5.16) also holds on $\left\{\tau_{N}=T\right\}$ and thus a.s. Since $0 \leq L_{\varepsilon} \leq L$, we immediately see by Proposition 3.1 that the sequence $\left\{U_{L_{\varepsilon}}\left(\tilde{Z}_{T}^{\varepsilon,(n)}, \tilde{\Theta}_{T}^{\varepsilon,(n)}\right), \varepsilon>0\right\}$ is uniformly integrable, so that by sending $\varepsilon$ to zero in (5.15) and using (5.16) we get

$$
\lim _{\varepsilon \downarrow 0} v_{\varepsilon}(t, z, \theta) \geq \mathbb{E}\left[U\left(X_{T}^{(n)}\right)\right] \geq v(t, z)-\frac{1}{n}
$$

from (5.10). By sending $n$ to infinity and recalling (5.9), this completes the proof of assertion (1) in Theorem 5.1.

We now turn to the viscosity characterization of $v_{\varepsilon}$. The viscosity property of $v_{\varepsilon}$ is proved similarly as for $v$ and is also omitted. From Proposition 3.1, and since $0 \leq v_{\varepsilon} \leq v$, we know that the value functions $v_{\varepsilon}$ lie in the set of functions satisfying the growth condition in (5.7), i.e.,

$$
\mathcal{G}_{\gamma}\left([0, T] \times \overline{\mathcal{S}}_{\varepsilon}\right)=\left\{w:[0, T] \times \overline{\mathcal{S}}_{\varepsilon} \rightarrow \mathbb{R}, \sup _{[0, T] \times \overline{\mathcal{S}}_{\varepsilon}} \frac{|w(t, z, \theta)|}{1+L_{M}(z)^{\gamma}}<\infty\right\} .
$$

The boundary property (5.8) is immediate. Indeed, fix $(t, z=(x, 0, p), \theta) \in[0, T] \times \partial_{y} \mathcal{S}_{\varepsilon}$, and consider an arbitrary sequence $\left(t_{n}, z_{n}=\left(x_{n}, y_{n}, p_{n}\right), \theta_{n}\right)_{n}$ in $[0, T] \times \overline{\mathcal{S}}_{\varepsilon}$ converging to $(t, z, \theta)$. Since $0 \leq L_{\varepsilon}\left(z_{n}, \theta_{n}\right)=\max \left(x_{n}, L\left(z_{n}, \theta_{n}\right)-\varepsilon\right)$, and $y_{n}$ goes to zero, this implies that for $n$ large enough, $x_{n}=L_{\varepsilon}\left(z_{n}, \theta_{n}\right) \geq 0$. By considering from $\left(t_{n}, z_{n}, \theta_{n}\right)$ the admissible strategy of doing no transactions, which leads to a final liquidation value $X_{T}=x_{n}$, we have $U\left(x_{n}\right)$ $\leq v_{\varepsilon}\left(t_{n}, z_{n}, \theta_{n}\right) \leq v\left(t_{n}, z_{n}, \theta_{n}\right)$. Recalling Corollary 3.1, we then obtain the continuity of $v_{\varepsilon}$ on $\partial_{y} \mathcal{S}_{\varepsilon}$ with $v_{\varepsilon}(t, z, \theta)=U(x)=v(t, z, \theta)$ for $(z, \theta)=(x, 0, p, \theta) \in \partial_{y} \mathcal{S}_{\varepsilon}$ and, in particular, (5.8). Finally, we address the uniqueness issue, which is a direct consequence of the following comparison principle for constrained (discontinuous) viscosity solution to (5.5)-(5.6).

Theorem 5.2 (comparison principle). Suppose $u \in \mathcal{G}_{\gamma}\left([0, T] \times \overline{\mathcal{S}}_{\varepsilon}\right)$ is a u.s.c. viscosity subsolution to (5.5)-(5.6) on $[0, T] \times \overline{\mathcal{S}}_{\varepsilon}$, and $w \in \mathcal{G}_{\gamma}\left([0, T] \times \overline{\mathcal{S}}_{\varepsilon}\right)$ is an l.s.c. viscosity supersolution to (5.5)-(5.6) on $[0, T] \times \mathcal{S}_{\varepsilon}$ such that

$$
\begin{equation*}
u(t, z, \theta) \leq \liminf _{\substack{\left(t^{\prime}, z^{\prime}, \theta^{\prime}\right) \rightarrow(t, z, \theta) \\\left(t^{\prime}, z^{\prime}, \theta^{\prime}\right) \in[0, T) \times \mathcal{S}_{\varepsilon}}} w\left(t^{\prime}, z^{\prime}, \theta^{\prime}\right) \quad \forall(t, z, \theta) \in[0, T] \times D_{0} . \tag{5.19}
\end{equation*}
$$

Then,

$$
\begin{equation*}
u \leq w \quad \text { on } \quad[0, T] \times \mathcal{S}_{\varepsilon} . \tag{5.20}
\end{equation*}
$$

Notice that with respect to usual comparison principles for parabolic PDEs where we compare a viscosity subsolution and a viscosity supersolution from the inequalities on the
domain and at the terminal date, we require here in addition a comparison on the boundary $D_{0}$ due to the nonsmoothness of the domain $\overline{\mathcal{S}}_{\varepsilon}$ on this right angle of the boundary. A similar feature appears also in [20], and we shall emphasize only the main arguments adapted from [4] for proving the comparison principle.

Proof of Theorem 5.2. Let $u$ and $w$ be as in Theorem 5.2, and (re)define $w$ on $[0, T] \times \partial \mathcal{S}_{\varepsilon}$ by

$$
\begin{equation*}
w(t, z, \theta)=\liminf _{\substack{\left(t^{\prime}, z^{\prime}, \theta^{\prime}\right) \rightarrow(t, z, \theta) \\\left(t^{\prime}, z^{\prime}, \theta^{\prime}\right) \in[0, T) \times \mathcal{S}_{\varepsilon}}} w\left(t^{\prime}, z^{\prime}, \theta^{\prime}\right), \quad(t, z, \theta) \in[0, T] \times \partial \mathcal{S}_{\varepsilon} . \tag{5.21}
\end{equation*}
$$

In order to obtain the comparison result (5.20), it suffices to prove that $\sup _{[0, T] \times \overline{\mathcal{S}}_{\varepsilon}}(u-w) \leq$ 0 , and we shall argue by contradiction by assuming that

$$
\begin{equation*}
\sup _{[0, T] \times \overline{\mathcal{S}}_{\varepsilon}}(u-w)>0 . \tag{5.22}
\end{equation*}
$$

- Step 1. Construction of a strict viscosity supersolution. Consider the function defined on $[0, T] \times \overline{\mathcal{S}}_{\varepsilon}$ by

$$
\psi(t, z, \theta)=e^{\rho^{\prime}(T-t)} L_{M}(z)^{\gamma^{\prime}}, \quad t \in[0, T],(z, \theta)=(x, y, p, \theta) \in \overline{\mathcal{S}}_{\varepsilon},
$$

where $\rho^{\prime}>0$, and $\gamma^{\prime} \in(0,1)$ will be chosen later. The function $\psi$ is smooth $C^{2}$ on $[0, T) \times$ $\left(\overline{\mathcal{S}}_{\mathcal{E}} \backslash D_{0}\right)$, and by the same calculations as in (3.10) we see that, by choosing $\rho^{\prime}>\frac{\gamma^{\prime}}{1-\gamma^{\prime}} \frac{b^{2}}{\sigma^{2}}$,

$$
\begin{equation*}
-\frac{\partial \psi}{\partial t}-\frac{\partial \psi}{\partial \theta}-\mathcal{L} \psi>0 \quad \text { on }[0, T) \times\left(\overline{\mathcal{S}}_{\varepsilon} \backslash D_{0}\right) . \tag{5.23}
\end{equation*}
$$

Moreover, from (5.4), we have

$$
\begin{align*}
\left(\psi-\mathcal{H}_{\varepsilon} \psi\right)(t, z, \theta) & =e^{\rho^{\prime}(T-t)}\left[L_{M}(z)^{\gamma^{\prime}}-\left(\mathcal{H}_{\varepsilon} L_{M}(z)\right)^{\gamma^{\prime}}\right]=: \Delta(t, z)  \tag{5.24}\\
& >0 \quad \text { on }[0, T] \times \overline{\mathcal{S}}_{\varepsilon} .
\end{align*}
$$

For $m \geq 1$, we denote

$$
\tilde{u}(t, z, \theta)=e^{t} u(t, z, \theta) \quad \text { and } \quad \tilde{w}_{m}(t, z, \theta)=e^{t}\left[w(t, z, \theta)+\frac{1}{m} \psi(t, z, \theta)\right] .
$$

From the viscosity subsolution property of $u$, we immediately see that $\tilde{u}$ is a viscosity subsolution to

$$
\begin{array}{rll}
\min \left[\tilde{u}-\frac{\partial \tilde{u}}{\partial t}-\frac{\partial \tilde{u}}{\partial \theta}-\mathcal{L} \tilde{u}, \tilde{u}-\mathcal{H}_{\varepsilon} \tilde{u}\right] \leq 0 & \text { on }[0, T) \times \overline{\mathcal{S}}_{\varepsilon}, \\
\min \left[\tilde{u}-\tilde{U}_{L_{\varepsilon}}, \tilde{u}-\mathcal{H}_{\varepsilon} \tilde{u}\right] \leq 0 & \text { on }\{T\} \times \overline{\mathcal{S}}_{\varepsilon}, \tag{5.26}
\end{array}
$$

where we set $\tilde{U}_{L_{\varepsilon}}(z, \theta)=e^{T} U_{L_{\varepsilon}}(z, \theta)$. From the viscosity supersolution property of $w$ and the relations (5.23)-(5.24), we also derive that $\tilde{w}_{m}$ is a viscosity supersolution to

$$
\begin{align*}
\tilde{w}_{m}-\frac{\partial \tilde{w}_{m}}{\partial t}-\frac{\partial \tilde{w}_{m}}{\partial \theta}-\mathcal{L} \tilde{w}_{m} \geq 0 & \text { on }[0, T) \times\left(\mathcal{S}_{\varepsilon} \backslash D_{0}\right),  \tag{5.27}\\
\tilde{w}_{m}-\mathcal{H}_{\varepsilon} \tilde{w}_{m} \geq \frac{1}{m} \Delta & \text { on }[0, T] \times \mathcal{S}_{\varepsilon}  \tag{5.28}\\
\tilde{w}_{m}-\tilde{U}_{L_{\varepsilon}} \geq 0 & \text { on }\{T\} \times \mathcal{S}_{\varepsilon} . \tag{5.29}
\end{align*}
$$

On the other hand, from the growth condition on $u$ and $w$ in $\mathcal{G}_{\gamma}\left([0, T] \times \overline{\mathcal{S}}_{\varepsilon}\right)$, and by choosing $\gamma^{\prime} \in(\gamma, 1)$, we have for all $(t, \theta) \in[0, T]^{2}$

$$
\lim _{|z| \rightarrow \infty}\left(u-w_{m}\right)(t, z, \theta)=-\infty
$$

Therefore, the u.s.c. function $\tilde{u}-\tilde{w}_{m}$ attains its supremum on $[0, T] \times \overline{\mathcal{S}}_{\varepsilon}$, and from (5.22) there exist $m$ large enough and $(\bar{t}, \bar{z}, \bar{\theta}) \in[0, T] \times \overline{\mathcal{S}}_{\varepsilon}$ such that

$$
\begin{equation*}
\tilde{M}=\sup _{[0, T] \times \overline{\mathcal{S}}_{\varepsilon}}\left(\tilde{u}-\tilde{w}_{m}\right)=\left(\tilde{u}-\tilde{w}_{m}\right)(\bar{t}, \bar{z}, \bar{\theta})>0 \tag{5.30}
\end{equation*}
$$

- Step 2. From the boundary condition (5.19), we know that $(\bar{z}, \bar{\theta})$ cannot lie in $D_{0}$, and we then have two possible cases:
(i) $(\bar{z}, \bar{\theta}) \in \mathcal{S}_{\varepsilon} \backslash D_{0}$.
(ii) $(\bar{z}, \bar{\theta}) \in \partial \mathcal{S}_{\varepsilon} \backslash D_{0}$.

Case (i), where $(\bar{z}, \bar{\theta})$ lies in $\mathcal{S}_{\varepsilon}$, is standard in the comparison principle for (nonconstrained) viscosity solutions, and we focus here on case (ii), which is specific to constrained viscosity solutions. From (5.21), there exists a sequence $\left(t_{n}, z_{n}, \theta_{n}\right)_{n \geq 1}$ in $[0, T) \times \mathcal{S}_{\varepsilon}$ such that

$$
\left(t_{n}, z_{n}, \theta_{n}, \tilde{w}_{m}\left(t_{n}, z_{n}, \theta_{n}\right)\right) \longrightarrow\left(\bar{t}, \bar{z}, \bar{\theta}, \tilde{w}_{m}(\bar{t}, \bar{z}, \bar{\theta})\right) \quad \text { as } n \rightarrow \infty
$$

We then set $\delta_{n}=\left|z_{n}-\bar{z}\right|+\left|\theta_{n}-\bar{\theta}\right|$ and consider the function $\Phi_{n}$ defined on $[0, T] \times\left(\overline{\mathcal{S}}_{\varepsilon}\right)^{2}$ by

$$
\begin{aligned}
\Phi_{n}\left(t, z, \theta, z^{\prime}, \theta^{\prime}\right)= & \tilde{u}(t, z, \theta)-\tilde{w}_{m}\left(t, z^{\prime}, \theta^{\prime}\right)-\varphi_{n}\left(t, z, \theta, z^{\prime}, \theta^{\prime}\right) \\
\varphi_{n}\left(t, z, \theta, z^{\prime}, \theta^{\prime}\right)= & |t-\bar{t}|^{2}+|z-\bar{z}|^{4}+|\theta-\bar{\theta}|^{4} \\
& +\frac{\left|z-z^{\prime}\right|^{2}+\left|\theta-\theta^{\prime}\right|^{2}}{2 \delta_{n}}+\left(\frac{d\left(z^{\prime}, \theta^{\prime}\right)}{d\left(z_{n}, \theta_{n}\right)}-1\right)^{4}
\end{aligned}
$$

Here, $d(z, \theta)$ denotes the distance from $(z, \theta)$ to $\partial \mathcal{S}_{\varepsilon}$. Since $(\bar{z}, \bar{\theta}) \notin D_{0}$, there exists an open neighborhood $\overline{\mathcal{V}}$ of $(\bar{z}, \bar{\theta})$ satisfying $\overline{\mathcal{V}} \cap D_{0}=\emptyset$ such that the function $d($.$) is twice continuously$ differentiable with bounded derivatives. This is well known (see, e.g., [14]) when $(\bar{z}, \bar{\theta})$ lies in the smooth parts of the boundary $\partial \mathcal{S}_{\varepsilon} \backslash\left(D_{1, \varepsilon} \cup D_{2, \varepsilon}\right)$. This is also true for $(\bar{z}, \bar{\theta}) \in D_{k, \varepsilon}$ for $k \in\{1,2\}$. Indeed, at these corner lines, the inner normal vectors form an acute angle (positive scalar product), and thus one can extend from $(\bar{z}, \bar{\theta})$ the boundary to a smooth boundary so that the distance $d$ is equal, locally on the neighborhood, to the distance to this smooth boundary. From the growth conditions on $u$ and $w$ in $\mathcal{G}_{\gamma}\left([0, T] \times \overline{\mathcal{S}}_{\varepsilon}\right)$, there exists a sequence $\left(\hat{t}_{n}, \hat{z}_{n}, \hat{\theta}_{n}, \hat{z}_{n}^{\prime}, \hat{\theta}_{n}^{\prime}\right)$ attaining the maximum of the u.s.c. $\Phi_{n}$ on $[0, T] \times\left(\overline{\mathcal{S}}_{\varepsilon}\right)^{2}$. By standard arguments (see, e.g., [4] or [20]), we have

$$
\begin{align*}
\quad\left(\hat{t}_{n}, \hat{z}_{n}, \hat{\theta}_{n}, \hat{z}_{n}^{\prime}, \hat{\theta}_{n}^{\prime}\right) & \longrightarrow(\bar{t}, \bar{z}, \bar{\theta}, \bar{z}, \bar{\theta})  \tag{5.31}\\
\frac{\left|\hat{z}_{n}-\hat{z}_{n}^{\prime}\right|^{2}+\left|\hat{\theta}_{n}-\hat{\theta}_{n}^{\prime}\right|^{2}}{2 \delta_{n}}+\left(\frac{d\left(\hat{z}_{n}^{\prime}, \hat{\theta}_{n}^{\prime}\right)}{d\left(z_{n}, \theta_{n}\right)}-1\right)^{4} & \longrightarrow 0  \tag{5.32}\\
\tilde{u}\left(\hat{t}_{n}, \hat{z}_{n}, \hat{\theta}_{n}\right)-\tilde{w}_{m}\left(\hat{t}_{n}, \hat{z}_{n}^{\prime}, \hat{\theta}_{n}^{\prime}\right) & \longrightarrow\left(\tilde{u}-\tilde{w}_{m}\right)(\bar{t}, \bar{z}, \bar{\theta}) \tag{5.33}
\end{align*}
$$

The convergence in (5.32) shows, in particular, that for $n$ large enough, $d\left(\hat{z}_{n}^{\prime}, \hat{\theta}_{n}^{\prime}\right) \geq d\left(z_{n}, \theta_{n}\right) / 2$ $>0$, and so $\left(\hat{z}_{n}^{\prime}, \hat{\theta}_{n}^{\prime}\right) \in \mathcal{S}_{\varepsilon}$. From the convergence in (5.31), we may also assume that for $n$ large enough, $\left(\hat{z}_{n}, \hat{\theta}_{n}\right),\left(\hat{z}_{n}^{\prime}, \hat{\theta}_{n}^{\prime}\right)$ lie in the neighborhood $\overline{\mathcal{V}}$ of $(\bar{z}, \bar{\theta})$ so that the derivatives upon order 2 of $d($.$) at \left(\hat{z}_{n}, \hat{\theta}_{n}\right)$ and $\left(\hat{z}_{n}^{\prime}, \hat{\theta}_{n}^{\prime}\right)$ exist and are bounded.

- Step 3. By arguments similar to those in [20], we show that for $n$ large enough, $\hat{t}_{n}<T$, and

$$
\begin{equation*}
\tilde{u}\left(\hat{t}_{n}, \hat{z}_{n}, \hat{\theta}_{n}\right)-\mathcal{H}_{\varepsilon} \tilde{u}\left(\hat{t}_{n}, \hat{z}_{n}\right)>0 . \tag{5.34}
\end{equation*}
$$

- Step 4. We use the viscosity subsolution property (5.25) of $\tilde{u}$ at $\left(\hat{t}_{n}, \hat{z}_{n}, \hat{\theta}_{n}\right) \in[0, T) \times \overline{\mathcal{S}}_{\varepsilon}$, which is written by (5.34) as

$$
\begin{equation*}
\left(\tilde{u}-\frac{\partial \tilde{u}}{\partial t}-\frac{\partial \tilde{u}}{\partial \theta}-\mathcal{L} \tilde{u}\right)\left(\hat{t}_{n}, \hat{z}_{n}, \hat{\theta}_{n}\right) \leq 0 . \tag{5.35}
\end{equation*}
$$

The above inequality is understood in the viscosity sense, and applied with the test function $(t, z, \theta) \mapsto \varphi_{n}\left(t, z, \theta, \hat{z}_{n}^{\prime}, \hat{\theta}_{n}^{\prime}\right)$, which is $C^{2}$ in the neighborhood $[0, T] \times \overline{\mathcal{V}}$ of $\left(\hat{t}_{n}, \hat{z}_{n}, \hat{\theta}_{n}\right)$. We also write the viscosity supersolution property (5.27) of $\tilde{w}_{m}$ at $\left(\hat{t}_{n}, \hat{z}_{n}^{\prime}, \hat{\theta}_{n}^{\prime}\right) \in[0, T) \times\left(\mathcal{S}_{\varepsilon} \backslash D_{0}\right)$ :

$$
\begin{equation*}
\left(\tilde{w}_{m}-\frac{\partial \tilde{w}_{m}}{\partial t}-\frac{\partial \tilde{w}_{m}}{\partial \theta}-\mathcal{L} \tilde{w}_{m}\right)\left(\hat{t}_{n}, \hat{z}_{n}^{\prime}, \hat{\theta}_{n}^{\prime}\right) \geq 0 \tag{5.36}
\end{equation*}
$$

The above inequality is again understood in the viscosity sense, and applied with the test function $\left(t, z^{\prime}, \theta^{\prime}\right) \mapsto-\varphi_{n}\left(t, \hat{z}_{n}, \hat{\theta}_{n}, z^{\prime}, \theta^{\prime}\right)$, which is $C^{2}$ in the neighborhood $[0, T] \times \overline{\mathcal{V}}$ of $\left(\hat{t}_{n}, \hat{z}_{n}^{\prime}, \hat{\theta}_{n}^{\prime}\right)$. The conclusion is achieved by arguments similar to those in [20]: we invoke Ishii's lemma, subtract the two inequalities (5.35)-(5.36), and finally get the required contradiction $\tilde{M} \leq 0$ by sending $n$ to infinity with (5.31)-(5.32)-(5.33).

Appendix A. Constrained viscosity solutions to parabolic QVIs. We consider a parabolic QV inequality in the form

$$
\begin{equation*}
\min \left[-\frac{\partial v}{\partial t}+F\left(t, x, v, D_{x} v, D_{x}^{2} v\right), v-\mathcal{H} v\right]=0 \quad \text { in }[0, T) \times \overline{\mathcal{O}}, \tag{A.1}
\end{equation*}
$$

together with a terminal condition

$$
\begin{equation*}
\min [v-g, v-\mathcal{H} v]=0 \quad \text { in }\{T\} \times \overline{\mathcal{O}} . \tag{A.2}
\end{equation*}
$$

Here, $\mathcal{O} \subset \mathbb{R}^{d}$ is an open domain, $F$ is a continuous function on $[0, T] \times \mathbb{R}^{d} \times \mathbb{R} \times \mathbb{R}^{d} \times \mathcal{S}^{d}$ ( $\mathcal{S}^{d}$ is the set of positive semidefinite symmetric matrices in $\mathbb{R}^{d \times d}$ ), nonincreasing in its last argument, $g$ is a continuous function on $\overline{\mathcal{O}}$, and $\mathcal{H}$ is a nonlocal operator defined on the set of locally bounded functions on $[0, T] \times \overline{\mathcal{O}}$ by

$$
\mathcal{H} v(t, x)=\sup _{e \in \mathcal{C}(t, x)}[v(t, \Gamma(t, x, e))+c(t, x, e)] .
$$

$\mathcal{C}(t, x)$ is a compact set of a metric space $E$, eventually empty for some values of $(t, x)$, in which case we set $\mathcal{H} v(t, x)=-\emptyset$, and is continuous for the Hausdorff metric; i.e., if $\left(t_{n}, x_{n}\right)$
converges to $(t, x)$ in $[0, T] \times \overline{\mathcal{O}}$, and $\left(e_{n}\right)$ is a sequence in $\mathcal{C}\left(t_{n}, x_{n}\right)$ converging to $e$, then $e \in$ $\mathcal{C}(t, x)$. The functions $\Gamma$ and $c$ are continuous and such that $\Gamma(t, x, e) \in \overline{\mathcal{O}}$ for all $e \in \mathcal{C}(t, x, e)$.

Given a locally bounded function $u$ on $[0, T] \times \overline{\mathcal{O}}$, we define its l.s.c. envelope $u_{*}$ and u.s.c. envelope $u^{*}$ on $[0, T] \times \overline{\mathcal{S}}$ by

$$
u_{*}(t, x)=\liminf _{\substack{\left(t^{\prime}, x^{\prime}\right) \rightarrow(t, x) \\\left(t^{\prime}, x^{\prime}\right) \in[0, T) \times \mathcal{O}}} u\left(t^{\prime}, x^{\prime}\right), u^{*}(t, x)=\operatorname{limisup}_{\substack{\left(t^{\prime}, x^{\prime}\right) \overrightarrow{(t, x)} \\\left(t^{\prime}, x^{\prime}\right) \in[0, T) \times \mathcal{O}}} u\left(t^{\prime}, x^{\prime}\right) .
$$

One can check (see, e.g., Lemma 5.1 in [20]) that the operator $\mathcal{H}$ preserves lower and upper semicontinuity:

$$
\begin{equation*}
\text { (i) } \mathcal{H} u_{*} \text { is lsc, and } \mathcal{H} u_{*} \leq(\mathcal{H} u)_{*}, \text { (ii) } \mathcal{H} u^{*} \text { is usc, and }(\mathcal{H} u)^{*} \leq \mathcal{H} u^{*} \text {. } \tag{A.3}
\end{equation*}
$$

We now give the definition of constrained viscosity solutions to (A.1)-(A.2). This notion, which extends the definition of viscosity solutions of Crandall, Ishii, and Lions (see [11]), was introduced in [31] for first-order equations for taking into account boundary conditions arising in state constraints and used in [33] for stochastic control problems in optimal investment.

Definition A.1. A locally bounded function $v$ on $[0, T] \times \overline{\mathcal{O}}$ is a constrained viscosity solution to (A.1)-(A.2) if the two following properties hold:
(i) Viscosity supersolution property on $[0, T] \times \mathcal{O}$ : for all $(\bar{t}, \bar{x}) \in[0, T] \times \mathcal{O}$, and $\varphi \in$ $C^{1,2}([0, T] \times \mathcal{O})$ with $0=\left(v_{*}-\varphi\right)(\bar{t}, \bar{x})=\min \left(v_{*}-\varphi\right)$, we have

$$
\begin{aligned}
\min \left[-\frac{\partial \varphi}{\partial t}(\bar{t}, \bar{x})+F\left(\bar{t}, \bar{x}, \varphi_{*}(\bar{t}, \bar{x}), D_{x} \varphi(\bar{t}, \bar{x}), D_{x}^{2} \varphi(\bar{t}, \bar{x})\right)\right. & \\
\left.v_{*}(\bar{t}, \bar{x})-\mathcal{H} v_{*}(\bar{t}, \bar{x})\right] & \geq 0, \quad(\bar{t}, \bar{x}) \in[0, T) \times \mathcal{O} \\
\min \left[v_{*}(\bar{t}, \bar{x})-g(\bar{x}), v_{*}(\bar{t}, \bar{x})-\mathcal{H} v_{*}(\bar{t}, \bar{x})\right] & \geq 0, \quad(\bar{t}, \bar{x}) \in\{T\} \times \mathcal{O}
\end{aligned}
$$

(ii) Viscosity subsolution property on $[0, T] \times \overline{\mathcal{O}}$ : for all $(\bar{t}, \bar{x}) \in[0, T] \times \overline{\mathcal{O}}$, and $\varphi \in$ $C^{1,2}([0, T] \times \overline{\mathcal{O}})$ with $0=\left(v^{*}-\varphi\right)(\bar{t}, \bar{x})=\max \left(v^{*}-\varphi\right)$, we have

$$
\begin{aligned}
& \min \left[-\frac{\partial \varphi}{\partial t}(\bar{t}, \bar{x})+F\left(\bar{t}, \bar{x}, \varphi_{*}(\bar{t}, \bar{x}), D_{x} \varphi(\bar{t}, \bar{x}), D_{x}^{2} \varphi(\bar{t}, \bar{x})\right)\right. \\
&\left.v^{*}(\bar{t}, \bar{x})-\mathcal{H} v^{*}(\bar{t}, \bar{x})\right] \leq 0, \quad(\bar{t}, \bar{x}) \in[0, T) \times \overline{\mathcal{O}} \\
& \min \left[v^{*}(\bar{t}, \bar{x})-g(\bar{x}), v^{*}(\bar{t}, \bar{x})-\mathcal{H} v^{*}(\bar{t}, \bar{x})\right] \leq 0, \quad(\bar{t}, \bar{x}) \in\{T\} \times \overline{\mathcal{O}}
\end{aligned}
$$

Appendix B. Proof of Proposition 4.2. We consider a small perturbation of our initial optimization problem by adding a cost $\varepsilon$ to the utility at each trading. We then define the value function $\bar{v}_{\varepsilon}$ on $[0, T] \times \overline{\mathcal{S}}$ by

$$
\begin{equation*}
\bar{v}_{\varepsilon}(t, z, \theta)=\sup _{\alpha \in \mathcal{A}_{\ell}^{b}(t, z, \theta)} \mathbb{E}\left[U_{L}\left(Z_{T}, \Theta_{T}\right)-\varepsilon N_{T}(\alpha)\right], \quad(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}} \tag{B.1}
\end{equation*}
$$

Step 1. We first prove that the sequence $\left(\bar{v}_{\varepsilon}\right)_{\varepsilon}$ converges pointwise on $[0, T] \times \overline{\mathcal{S}}$ towards $v$ as $\varepsilon$ goes to zero. It is clear that the sequence $\left(\bar{v}_{\varepsilon}\right)_{\varepsilon}$ is nondecreasing and that $\bar{v}_{\varepsilon} \leq v$ on $[0, T] \times \overline{\mathcal{S}}$ for any $\varepsilon>0$. Let us prove that $\lim _{\varepsilon} \searrow 0 \bar{v}_{\varepsilon}=v$. Fix $n \in \mathbb{N}^{*}$ and $(t, z, \theta) \in[0, T] \times \overline{\mathcal{S}}$, and consider some $\alpha^{(n)} \in \mathcal{A}_{\ell}^{b}(t, z, \theta)$ such that

$$
\mathbb{E}\left[U_{L}\left(Z_{T}^{(n)}, \Theta_{T}^{(n)}\right)\right] \geq v(t, z, \theta)-\frac{1}{n},
$$

where $\left(Z^{(n)}, \Theta^{(n)}\right)$ is the associated controlled process. From the monotone convergence theorem, we then get

$$
\lim _{\varepsilon \searrow 0} \bar{v}_{\varepsilon}(t, z, \theta) \geq \mathbb{E}\left[U_{L}\left(Z_{T}^{(n)}, \Theta_{T}^{(n)}\right)\right] \geq v(t, z, \theta)-\frac{1}{n}
$$

Sending $n$ to infinity, we conclude that $\lim _{\varepsilon} \bar{v}_{\varepsilon} \geq v$, which ends the proof since we already have $\bar{v}_{\varepsilon} \leq v$.

Step 2 . The nonlocal impulse operator $\overline{\mathcal{H}}^{\varepsilon}$ associated with (B.1) is given by

$$
\overline{\mathcal{H}}_{\varepsilon} \varphi(t, z, \theta)=\mathcal{H} \varphi(t, z, \theta)-\varepsilon,
$$

and we consider the corresponding dynamic programming equation:

$$
\begin{array}{rll}
\min \left[-\frac{\partial w}{\partial t}-\frac{\partial w}{\partial \theta}-\mathcal{L} w, w-\overline{\mathcal{H}}_{\varepsilon} w\right]=0 & \text { in }[0, T) \times \overline{\mathcal{S}} \\
\min \left[w-U_{L}, w-\overline{\mathcal{H}}_{\varepsilon} w\right]=0 & \text { in }\{T\} \times \overline{\mathcal{S}} \tag{B.3}
\end{array}
$$

One can show by routine arguments that $\bar{v}_{\varepsilon}$ is a constrained viscosity solution to (B.2)(B.3), and as in section 5 , the following comparison principle holds.

Suppose $u \in \mathcal{G}_{\gamma}([0, T] \times \overline{\mathcal{S}})$ is a u.s.c. viscosity subsolution to (B.2)-(B.3) on $[0, T] \times \overline{\mathcal{S}}$, and $w \in \mathcal{G}_{\gamma}([0, T] \times \overline{\mathcal{S}})$ is an l.s.c. viscosity supersolution to (B.2)-(B.3) on $[0, T] \times \mathcal{S}$, such that

$$
u(t, z, \theta) \leq \liminf _{\substack{\left(t^{\prime}, z^{\prime}, \theta^{\prime}\right) \rightarrow(t, z, \theta) \\\left(t^{\prime}, z^{\prime}, \theta^{\prime}\right) \in[0, T) \times \mathcal{S}}} w\left(t^{\prime}, z^{\prime}, \theta^{\prime}\right) \quad \forall(t, z, \theta) \in[0, T] \times D_{0} .
$$

Then,

$$
\begin{equation*}
u \leq w \quad \text { on } \quad[0, T] \times \mathcal{S} \tag{B.4}
\end{equation*}
$$

The proof follows the same lines of arguments as in the proof of Theorem 5.2 (the function $\psi$ is still a strict viscosity supersolution to (B.2)-(B.3) on $[0, T] \times \overline{\mathcal{S}}$ ), and so we omit it.

Step 3. Let $V \in \mathcal{G}_{\gamma}([0, T] \times \overline{\mathcal{S}})$ be a viscosity solution in $\mathcal{G}_{\gamma}([0, T] \times \overline{\mathcal{S}})$ to (4.2)-(4.3) satisfying the boundary condition (4.4). Since $\mathcal{H} \geq \overline{\mathcal{H}}_{\varepsilon}$, it is clear that $V_{*}$ is a viscosity supersolution to (B.2)-(B.3). Moreover, $\operatorname{since} \lim _{\left(t^{\prime}, z^{\prime}, \theta^{\prime}\right) \rightarrow(t, z, \theta)} V_{*}\left(t^{\prime}, z^{\prime}, \theta^{\prime}\right)=U(0)=v(t, z, \theta)$ $\geq \bar{v}_{\varepsilon}^{*}(t, z, \theta)$ for $(t, z, \theta) \in[0, T] \times D_{0}$, we deduce from the comparison principle (B.4) that $V$ $\geq V_{*} \geq \bar{v}_{\varepsilon}^{*} \geq \bar{v}_{\varepsilon}$ on $[0, T] \times \mathcal{S}$. By sending $\varepsilon$ to 0 , and from the convergence result in Step 1, we obtain $V \geq v$, which proves the required result.

Acknowledgments. We would like to thank Bruno Bouchard for useful comments. We also thank participants at the Istanbul Workshop on Mathematical Finance in May 2009 for relevant remarks. The comments of two anonymous referees were helpful in improving the first version of this paper.

## REFERENCES

[1] R. Almgren and N. Chriss, Optimal execution of portfolio transactions, J. Risk, 3 (2001), pp. 5-39.
[2] R. Almgren, C. Thum, E. Hauptmann, and H. Li, Equity market impact, Risk, July (2005), pp. 58-62.
[3] P. Bank and D. Baum, Hedging and portfolio optimization in illiquid financial markets with a large trader, Math. Finance, 14 (2004), pp. 1-18.
[4] G. Barles, Solutions de viscosité des équations d'Hamilton-Jacobi, Math. Appl. (Berlin) 17, SpringerVerlag, Paris, 1994.
[5] E. Bayraktar and M. Ludkovski, Optimal trade execution in illiquid markets, Math. Finance, to appear.
[6] A. Bensoussan and J. L. Lions, Impulse Control and Quasi-variational Inequalities, Gauthiers-Villars, Paris, 1984.
[7] D. Bertsimas and A. Lo, Optimal control of execution costs, J. Financial Markets, 1 (1998), pp. 1-50.
[8] B. Bouchard, A stochastic target formulation for optimal switching problems in finite horizon, Stochastics, 81 (2009), pp. 171-197.
[9] U. Cetin, R. Jarrow, and P. Protter, Liquidity risk and arbitrage pricing theory, Finance Stoch., 8 (2004), pp. 311-341.
[10] U. Cetin, M. Soner, And N. Touzi, Option hedging for small investors under liquidity costs, Finance Stoch., 14 (2010), pp. 317-341.
[11] M. Crandall, H. Ishir, and P. L. Lions, User's guide to viscosity solutions of second order partial differential equations, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1-67.
[12] R. Engle and R. Ferstenberg, Execution risk, J. Portfolio Management, 33 (2007), pp. 34-44.
[13] W. Fleming and M. Soner, Controlled Markov Processes and Viscosity Solutions, Appl. Math. (N. Y.) 25, Springer-Verlag, New York, 1993.
[14] D. Gilbarg and N. Trudinger, Elliptic Partial Differential Equations of Second Order, Springer-Verlag, Berlin, 1977.
[15] F. Guilbaud, M. Mnif, and H. Pham, Numerical Methods for an Optimal Order Execution Problem, preprint, 2010.
[16] H. He and H. Mamaysky, Dynamic trading policies with price impact, J. Econom. Dynam. Control, 29 (2005), pp. 891-930.
[17] K. IsHiI, Viscosity solutions of nonlinear second order elliptic PDEs associated with impulse control problems, Funkcial. Ekvac., 36 (1993), pp. 123-141.
[18] R. Korn, Portfolio optimization with strictly positive transaction costs and impulse control, Finance Stoch., 2 (1998), pp. 85-114.
[19] F. Lillo, J. Farmer, and R. Mantagna, Master curve for price impact function, Nature, 421 (2003), pp. 129-130.
[20] V. Ly Vath, M. Mnif, and H. Pham, A model of optimal portfolio selection under liquidity risk and price impact, Finance Stoch., 11 (2007), pp. 51-90.
[21] J. Ma, Q. Song, J. Xu, and J. Zhang, Impulse Control Problem and Optimal Portfolio Selection with General Transaction Costs, http://www-bcf.usc.edu/~jianfenz/Papers/MSXZ.pdf (2009).
[22] A. Obizhaeva and J. Wang, Optimal trading strategy and supply/demand dynamics, J. Financial Markets, to appear.
[23] B. Øksendal and A. Sulem, Optimal consumption and portfolio with both fixed and proportional transaction costs, SIAM J. Control. Optim., 40 (2002), pp. 1765-1790.
[24] B. Øksendal and A. Sulem, Applied Stochastic Control of Jump Diffusions, Universitext, SpringerVerlag, Berlin, 2007.
[25] H. Pham, Continuous-Time Stochastic Control and Optimization with Financial Applications, Stoch. Model. Appl. Probab. 61, Springer-Verlag, Berlin, 2009.
[26] H. Pham and P. Tankov, A model of optimal consumption under liquidity risk with random trading times, Math. Finance, 18 (2008), pp. 613-627.
[27] M. Potters and J. P. Bouchaud, More statistical properties of order books and price impact, Phys. A, 324 (2003), pp. 133-140.
[28] L. C. G. Rogers and S. Singh, The cost of illiquidity and its effects on hedging, Math. Finance, 20 (2010), pp. 597-615.
[29] A. Schied and T. Schöneborn, Risk aversion and the dynamics of optimal liquidation strategies in illiquid markets, Finance Stoch., 13 (2009), pp. 181-204.
[30] R. Seydel, Existence and uniqueness of viscosity solutions for QVI associated with impulse control of jump-diffusions, Stochastic Process. Appl., 119 (2009), pp. 3719-3748.
[31] H. M. Soner, Optimal control with state-space constraint I, SIAM J. Control Optim., 24 (1986), pp. 552561.
[32] S. Tang and J. Yong, Finite horizon stochastic optimal switching and impulse controls with a viscosity solution approach, Stochastics Stochastics Rep., 45 (1993), pp. 145-176.
[33] T. Zariphopoulou, Optimal Investment-Consumption Models with Constraints, Ph.D. thesis, Brown University, Providence, RI, 1988.

# Portfolio Selection Using Tikhonov Filtering to Estimate the Covariance Matrix* 

Sungwoo Park $^{\dagger}$ and Dianne P. O’Leary ${ }^{\ddagger}$


#### Abstract

Markowitz's portfolio selection problem chooses weights for stocks in a portfolio based on an estimated covariance matrix of stock returns. Our study proposes reducing noise in the estimation by using a Tikhonov filter function. In addition, we prevent rank deficiency of the estimated covariance matrix and propose a method for effectively choosing the Tikhonov parameter, which determines the filtering intensity. We put previous estimators into a common framework and compare their filtering functions for eigenvalues of the correlation matrix. We demonstrate the effectiveness of our estimator using stock return data from 1958 through 2007.


Key words. Tikhonov regularization, covariance matrix estimate, Markowitz portfolio selection, ridge regression

AMS subject classifications. 91G10, 91G60, 65F22, 62J07, 62J10
DOI. 10.1137/090749372

1. Introduction. A stock investor might want to construct a portfolio of stocks whose return has a small variance because large variance implies high risk. Given a target portfolio return $q$, a mean-variance (MV) problem [33] finds a stock weight vector $\boldsymbol{w}$ to determine a portfolio that minimizes the variance of the return. Let $\boldsymbol{\mu}$ be a vector of expected returns for each of $N$ stocks, and let $\boldsymbol{\Sigma}$ be an $N \times N$ covariance matrix for the returns. The problem can be written as

$$
\begin{equation*}
\min _{\boldsymbol{w}} \boldsymbol{w}^{T} \boldsymbol{\Sigma} \boldsymbol{w} \text { subject to } \quad \boldsymbol{w}^{T} \mathbf{1}=1, \quad \boldsymbol{w}^{T} \boldsymbol{\mu}=q \tag{1.1}
\end{equation*}
$$

where $\mathbf{1}$ is a vector of $N$ ones. On the other hand, a global minimum variance (GMV) problem finds a portfolio that minimizes the variances of the portfolio returns without the return constraint:

$$
\begin{equation*}
\min _{\boldsymbol{w}} \boldsymbol{w}^{T} \boldsymbol{\Sigma} \boldsymbol{w} \text { subject to } \quad \boldsymbol{w}^{T} \mathbf{1}=1 \tag{1.2}
\end{equation*}
$$

Even though these optimization problems play a central role in a modern portfolio theory, it has been observed that the solutions are very sensitive to their input parameters $[3,5,6,7]$. Thus, in order to construct a good portfolio using these formulations, the covariance matrix $\boldsymbol{\Sigma}$ must be well estimated. We let $\widetilde{\boldsymbol{\Sigma}}$ denote an estimate of $\boldsymbol{\Sigma}$ and $\widetilde{\boldsymbol{\Sigma}}_{\text {method }}$ denote a resulting estimate by a particular method.

[^96]Let $\boldsymbol{R}=[\boldsymbol{r}(1), \ldots, \boldsymbol{r}(T)]$ be an $N \times T$ matrix containing observations on $N$ stocks' returns for each of $T$ times. A conventional estimator-a sample covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$-can be computed from the stock return matrix $\boldsymbol{R}$ as

$$
\begin{equation*}
\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}=\frac{1}{T} \boldsymbol{R}\left(\boldsymbol{I}_{T}-\frac{1}{T} \mathbf{1 1}^{T}\right) \boldsymbol{R}^{T}, \tag{1.3}
\end{equation*}
$$

where $\boldsymbol{I}_{T}$ denotes a $T \times T$ identity matrix. From classical statistics, $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$ is a consistent estimate for fixed $N$; in our case, since $T$ is fixed and of the same order as $N$, this result is not so useful. Moreover, since the stock return matrix $\boldsymbol{R}$ contains noise, the sample covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$ might not estimate the true covariance matrix well. This paper uses principal component analysis and reduces the noise in the covariance matrix estimate by using a Tikhonov regularization method, as summarized in Table 1. We demonstrate experimentally that this improves the portfolio weight $\boldsymbol{w}$ obtained from (1.1) and (1.2).

Our study is closely related to factor analysis and principal component analysis, which were previously applied to explain the interdependency of stock returns and classify the securities into appropriate subgroups. Sharpe [43] first proposed a single-factor model in this context using market returns. King [26] analyzed stock behaviors with both multiple factors and multiple principal components. These factor models established a basis for the asset pricing models CAPM $[32,35,44,47]$ and APT [39, 40].

There have been previous efforts to improve the estimate of $\boldsymbol{\Sigma}$. Sharpe [43] proposed a market-index covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{\text {market }}$ derived from a single-factor model of market returns. Ledoit and Wolf [30] introduced a shrinkage method that averages $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$ and $\widetilde{\boldsymbol{\Sigma}}_{\text {market }}$. They in [31] also applied the shrinkage method with a different target, an identity matrix. Later, it was shown by DeMiguel et al. [10] that their shrinkage methods have the same effect as adding the constraint $\|\boldsymbol{w}\|_{\boldsymbol{A}} \leq \delta$ to the GMV problem (1.2), where $\boldsymbol{A}$ is the shrinkage target matrix ( $\widetilde{\boldsymbol{\Sigma}}_{\text {market }}$ or $\boldsymbol{I}_{N}$ ) and $\delta$ is a given threshold. Elton and Gruber [13] estimated $\boldsymbol{\Sigma}$ using a few principal components from a correlation matrix. More recently, Plerou et al. [38], Laloux et al. [28], Conlon, Ruskin, and Crane [8], and Kwapień, Drożdż, and Oświȩcimka [27] applied random matrix theory [34] to this problem. They found that most eigenvalues of correlation matrices from stock return data lie within the bound for a random correlation matrix and hypothesized that eigencomponents (principal components) outside this interval contain true information. Bengtsson and Holst [2] generalized the approach of Ledoit and Wolf [30] by damping all but the $k$ largest eigenvalues by a single rate. In summary, the estimator of Sharpe [43] uses $\widetilde{\boldsymbol{\Sigma}}_{\text {market }}$, the estimator of Ledoit and Wolf [30, 31] takes the weighted average of $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$ and different target matrices, the estimator of Elton and Gruber [13] truncates the smallest eigenvalues, the estimators of Plerou et al. [38], Laloux et al. [28], Conlon, Ruskin, and Crane [8], and Kwapień, Drożdż, and Oświȩcimka [27] adjust principal components in some interval, and the estimator of Bengtsson and Holst [2] attenuates the smallest eigenvalues by a single rate.

Jagannathan and Ma [25] showed that a short-sale constraint $(w \geq 0)$ is equivalent to shrinking the input covariance matrix $\boldsymbol{\Sigma}$ by subtracting $\left(\boldsymbol{\lambda} \mathbf{1}^{T}+\mathbf{1}^{T} \boldsymbol{\lambda}\right)$, where $\boldsymbol{\lambda}$ is a vector of Lagrange multipliers for the constraints. DeMiguel et al. [10] showed that adding the short-sale constraint to GMV is equivalent to adding a 1 -norm constraint $\|\boldsymbol{w}\|_{1} \leq 1$, and they
generalized this constraint to $\|\boldsymbol{w}\|_{1} \leq \delta$ for a certain threshold $\delta$ which determines a short-sale budget.

Our study focuses on estimating a good covariance matrix. We propose decreasing the contribution of the smaller eigenvalues of a correlation matrix gradually by using a Tikhonov filtering function. To derive the Tikhonov filtering, we construct a linear model based on principal component analysis and formulate an optimization problem that finds appropriately noise-filtered factors. Using the filtered factor data, we estimate a Tikhonov covariance matrix.

This paper is organized as follows. In section 2, we introduce Tikhonov regularization to reduce noise in the stock return data. In section 3, we show that applying Tikhonov regularization results in filtering the eigenvalues of the correlation matrix for the stock returns. In section 4, we discuss how we can choose a Tikhonov parameter that determines the intensity of Tikhonov filtering. In section 5, we put all of the factor-based estimators into a common framework and compare the characteristics of their filtering functions for the eigenvalues of the correlation matrix. In section 6 , we show the results of numerical experiments comparing the covariance estimators for portfolio construction using monthly return data of 100 randomly chosen stocks from the Center for Research in Security Prices. In section 7, we highlight the differences between Tikhonov filtering and the other methods.
2. Tikhonov filtering. To estimate the covariance matrix, we apply principal component analysis to find an orthogonal basis that maximizes the variance of the projected data into the basis. Based on the analysis, we use the Tikhonov regularization method to filter out the noise from the data. Next, we explain the feature of gradual downweighting, which is the key difference between Tikhonov filtering and other methods.
2.1. Principal component analysis. First, we establish some notation. We use a 2 -norm $\|\cdot\|$ for vectors and a Frobenius norm $\|\cdot\|_{F}$ for matrices, defined as

$$
\begin{equation*}
\|\boldsymbol{a}\|^{2}=\boldsymbol{a}^{T} \boldsymbol{a} \quad \text { and } \quad\|\boldsymbol{A}\|_{F}^{2}=\left(\sum_{i=1}^{M} \sum_{j=1}^{N} a_{i j}^{2}\right) \tag{2.1}
\end{equation*}
$$

for a given vector $\boldsymbol{a}$ and a given matrix $\boldsymbol{A}=\left(a_{i j}\right) \in \mathbb{R}^{M \times N}$.
For a random process $\boldsymbol{x}(t)$, let $\mathbb{E}[\boldsymbol{x}(t)] \in \mathbb{R}^{N \times 1}, \operatorname{Var}[\boldsymbol{x}(t)] \in \mathbb{R}^{N \times 1}, \operatorname{Cov}[\boldsymbol{x}(t)] \in \mathbb{R}^{N \times N}$, and $\operatorname{Corr}[\boldsymbol{x}(t)] \in \mathbb{R}^{N \times N}$ denote a mean, a variance, a covariance matrix, and a correlation matrix. For a given collection of observations $\boldsymbol{X}=[\boldsymbol{x}(1), \ldots, \boldsymbol{x}(T)]$ for $N$ objects during $T$ times, let $\mathbb{E}_{s}[\boldsymbol{x}(t)] \in \mathbb{R}^{N \times 1}, \operatorname{Var}_{s}[\boldsymbol{x}(t)] \in \mathbb{R}^{N \times 1}, \operatorname{Cov}_{s}[\boldsymbol{x}(t)] \in \mathbb{R}^{N \times N}$, and $\operatorname{Corr}_{s}[\boldsymbol{x}(t)] \in \mathbb{R}^{N \times N}$ denote the corresponding sample statistics, defined, for example, in [21, section 3.3]. An $N \times N$ identity matrix will be denoted by $\boldsymbol{I}_{N}$.

Now we apply principal component analysis (PCA) ${ }^{1}$ to the stock return data $\boldsymbol{R}$. Let $\boldsymbol{Z}=[\boldsymbol{z}(1), \ldots, \boldsymbol{z}(T)]$ be an $N \times T$ matrix of normalized stock returns derived from $\boldsymbol{R}$, defined so that

$$
\begin{equation*}
\mathbb{E}_{s}[\boldsymbol{z}(t)]=\mathbf{0}, \quad \operatorname{Var}_{s}[\boldsymbol{z}(t)]=\mathbf{1}, \tag{2.2}
\end{equation*}
$$

[^97]where $\mathbf{0}$ is a vector of $N$ zeros. We can compute $\boldsymbol{Z}$ as
\[

$$
\begin{equation*}
\boldsymbol{Z}=\boldsymbol{D}_{V}^{-\frac{1}{2}}\left(\boldsymbol{R}-\frac{1}{T} \boldsymbol{R} \mathbf{1 1} \mathbf{1}^{T}\right) \tag{2.3}
\end{equation*}
$$

\]

where $\boldsymbol{D}_{V}=\operatorname{diag}\left(\operatorname{Var}_{s}[\boldsymbol{r}(t)]\right) \in \mathbb{R}^{N \times N}$ is a diagonal matrix containing the $N$ sample variances for the $N$ stock returns. By using the normalized stock return matrix $\boldsymbol{Z}$ rather than $\boldsymbol{R}$, we can make the PCA independent of the different variance of each stock return [24, pp. 64-66].

PCA finds an orthogonal basis $\boldsymbol{U}=\left[\boldsymbol{u}_{1}, \ldots, \boldsymbol{u}_{k}\right] \in \mathbb{R}^{N \times k}$ for $\boldsymbol{Z}$, where $k=\operatorname{rank}(\boldsymbol{Z})$. Each basis vector $\boldsymbol{u}_{i}$ maximizes the variance of the projected data $\boldsymbol{u}_{i}^{T} \boldsymbol{Z}$, while maintaining orthogonality to all the preceding basis vectors $\boldsymbol{u}_{j}(j<i)$. By PCA, we can represent the given data $\boldsymbol{Z}=[\boldsymbol{z}(1), \ldots, \boldsymbol{z}(T)]$ as

$$
\begin{gather*}
\boldsymbol{Z}=\left[\boldsymbol{u}_{1}, \ldots, \boldsymbol{u}_{k}\right] \boldsymbol{F}=\boldsymbol{U} \boldsymbol{F},  \tag{2.4}\\
\boldsymbol{z}(t)=\boldsymbol{U} \boldsymbol{f}(t)=\left[\boldsymbol{u}_{1}, \ldots, \boldsymbol{u}_{k}\right] \boldsymbol{f}(t)=\sum_{i=1}^{k} f_{i}(t) \boldsymbol{u}_{i}, \tag{2.5}
\end{gather*}
$$

where $\boldsymbol{f}(t)=\left[f_{1}(t), \ldots, f_{k}(t)\right]^{T}$, a column of $\boldsymbol{F}$, is the projected data at time $t$, and $\operatorname{Var}_{s}\left[f_{1}(t)\right] \geq$ $\operatorname{Var}_{s}\left[f_{2}(t)\right] \geq \cdots \geq \operatorname{Var}_{s}\left[f_{k}(t)\right]$. The projected data $f_{i}(t)$ is called the $i$ th principal component in PCA or the $i$ th factor in the factor analysis. Larger $\operatorname{Var}_{s}\left[f_{i}(t)\right]$ implies that the corresponding $\boldsymbol{f}_{i}(t)$ plays a more important role in representing $\boldsymbol{Z}$. The orthogonal basis $\boldsymbol{U}$ and the projected data $\boldsymbol{F}$ can be obtained by the singular value decomposition (SVD) of $\boldsymbol{Z}$,

$$
\begin{equation*}
\boldsymbol{Z}=\boldsymbol{U}_{k} \boldsymbol{S}_{k} \boldsymbol{V}_{k}^{T}, \tag{2.6}
\end{equation*}
$$

where $k$ is the rank of $Z$,
$\boldsymbol{U}_{k}=\left[\boldsymbol{u}_{1}, \ldots, \boldsymbol{u}_{k}\right] \in \mathbb{R}^{N \times k}$ is a matrix of left singular vectors,
$\boldsymbol{S}_{k}=\operatorname{diag}\left(s_{1}, \ldots, s_{k}\right) \in \mathbb{R}^{k \times k}$ is a diagonal matrix of singular values $s_{i}$,
and $\boldsymbol{V}_{k}=\left[\boldsymbol{v}_{1}, \ldots, \boldsymbol{v}_{k}\right] \in \mathbb{R}^{T \times k}$ is a matrix of right singular vectors.
The left and right singular vector matrices $\boldsymbol{U}_{k}$ and $\boldsymbol{V}_{k}$ are orthonormal:

$$
\begin{equation*}
\boldsymbol{U}_{k}^{T} \boldsymbol{U}_{k}=\boldsymbol{I}_{k} \quad \text { and } \quad \boldsymbol{V}_{k}^{T} \boldsymbol{V}_{k}=\boldsymbol{I}_{k} . \tag{2.7}
\end{equation*}
$$

In PCA, the orthogonal basis matrix $\boldsymbol{U}$ corresponds to $\boldsymbol{U}_{k}$, and the projected data $\boldsymbol{F}$ corresponds to $\left(\boldsymbol{S}_{k} \boldsymbol{V}_{k}^{T}\right)$ [24, p. 193]. Moreover, the variance of the projected data $f_{i}(t)$ is proportional to the square of singular value $s_{i}^{2}$ as we now show. $\mathbb{E}_{s}[\boldsymbol{z}(t)]=\mathbf{0}$ means that $\boldsymbol{Z 1}=\mathbf{0}$. Therefore, since $\boldsymbol{z}(t)=\boldsymbol{U f}(t)$,

$$
\begin{equation*}
\mathbb{E}_{s}[\boldsymbol{f}(t)]=\boldsymbol{U}^{T} \boldsymbol{Z} \mathbf{1}=\mathbf{0}, \tag{2.8}
\end{equation*}
$$

so $\boldsymbol{f}(t)$ also has zero-mean. Therefore,

$$
\operatorname{Var}_{s}\left[f_{i}(t)\right]=\frac{1}{T} \sum_{t=1}^{T}\left(f_{i}(t)-\mathbb{E}_{s}\left[f_{i}(t)\right]\right)^{2}=\frac{1}{T} \sum_{t=1}^{T} f_{i}^{2}(t) .
$$

Since $\boldsymbol{F}$ is equal to $\boldsymbol{S}_{k} \boldsymbol{V}_{k}^{T}$,

$$
\begin{equation*}
f_{i}(t)=s_{i} v_{i}(t) \tag{2.9}
\end{equation*}
$$

where $v_{i}(t)$ is the $(t, i)$ element of $\boldsymbol{V}_{k}$. Thus,

$$
\begin{equation*}
\operatorname{Var}_{s}\left[f_{i}(t)\right]=\frac{1}{T} \sum_{t=1}^{T}\left(s_{i} v_{i}(t)\right)^{2}=\frac{1}{T} s_{i}^{2}\left(\boldsymbol{v}_{i}^{T} \boldsymbol{v}_{i}\right)=\frac{s_{i}^{2}}{T} \tag{2.10}
\end{equation*}
$$

by the orthonormality of $\boldsymbol{v}_{i}$. Thus, the singular value $s_{i}$ determines the magnitude of $\operatorname{Var}_{s}\left[f_{i}(t)\right]$, so it measures the contribution of the projected data $f_{i}(t)$ to $\boldsymbol{z}(t)$.
2.2. Tikhonov regularization. $\boldsymbol{U}$ and $\boldsymbol{f}(t)$ in (2.5) form a linear model with a $k$-dimensional orthogonal basis for the normalized stock return $\boldsymbol{Z}$, where $k=\operatorname{rank}(\boldsymbol{Z})$. As mentioned in the previous section, the singular value $s_{i}$ determines how much the principal component $f_{i}(t)$ contributes to $\boldsymbol{z}(t)$. However, since noise is included in $\boldsymbol{z}(t)$, the $k$-dimensional model is overfitted, containing unimportant principal components possibly corresponding to the noise. We use a Tikhonov regularization method [36, 46, 49], sometimes called ridge regression [22, 23], to reduce the contribution of unimportant principal components to the normalized stock return $\boldsymbol{Z}$. Eventually, we construct a filtered principal component $\widetilde{\boldsymbol{f}}(t)$ and a filtered market return $\widetilde{Z}$.

Originally, regularization methods were developed to reduce the influence of noise when solving a discrete ill-posed problem $\boldsymbol{b} \approx \boldsymbol{A} \boldsymbol{x}$, where the $M \times N$ matrix $\boldsymbol{A}$ has some singular values close to 0 [18, pp. 71-86]. If we write the SVD of $\boldsymbol{A}$ as

$$
\boldsymbol{A}=\boldsymbol{U} \boldsymbol{S} \boldsymbol{V}^{T}=\left[\boldsymbol{u}_{1}, \ldots, \boldsymbol{u}_{N}\right]\left[\begin{array}{lll}
s_{1} & & \\
& \ddots & \\
& & s_{N}
\end{array}\right]\left[\begin{array}{c}
\boldsymbol{v}_{1}^{T} \\
\vdots \\
\boldsymbol{v}_{N}^{T}
\end{array}\right],
$$

then the minimum norm least square solution $\boldsymbol{x}_{L S}$ to $\boldsymbol{b} \approx \boldsymbol{A} \boldsymbol{f}$ is

$$
\begin{equation*}
\boldsymbol{x}_{L S}=\boldsymbol{A}^{\dagger} \boldsymbol{b}=\boldsymbol{V} \boldsymbol{S}^{\dagger} \boldsymbol{U}^{T} \boldsymbol{b}=\sum_{i=1}^{\operatorname{rank}(\boldsymbol{A})} \frac{\boldsymbol{u}_{i}^{T} \boldsymbol{b}}{s_{i}} \boldsymbol{v}_{i} . \tag{2.11}
\end{equation*}
$$

If $\boldsymbol{A}$ has some small singular values, then $\boldsymbol{x}_{L S}$ is dominated by the corresponding singular vectors $\boldsymbol{v}_{i}$. Two popular methods are used for regularization to reduce the influence of components $\boldsymbol{v}_{i}$ corresponding to small singular values: a truncated SVD (TSVD) method [14, 20] and a Tikhonov method [46]. Briefly speaking, the TSVD simply truncates terms in (2.11) corresponding to singular values close to 0 . In contrast, Tikhonov regularization solves the least squares problem

$$
\begin{equation*}
\min _{f}\|\boldsymbol{b}-\boldsymbol{A} \boldsymbol{x}\|^{2}+\alpha^{2}\|\boldsymbol{P} \boldsymbol{x}\|^{2}, \tag{2.12}
\end{equation*}
$$

where $\alpha$ and $\boldsymbol{P}$ are predetermined. The penalty term $\|\boldsymbol{P} \boldsymbol{x}\|^{2}$ restricts the magnitude of the solution $\boldsymbol{x}$ so that the effects of small singular values are reduced.

Returning to our original problem, we use regularization in order to filter out the noise from the principal component $\boldsymbol{f}(t)$. We formulate the linear problem to find a filtered principal component $\widetilde{\boldsymbol{f}}(t)$ as

$$
\begin{gather*}
\widetilde{\boldsymbol{z}}(t)=\boldsymbol{U} \widetilde{\boldsymbol{f}}(t)  \tag{2.13}\\
\boldsymbol{z}(t)=\widetilde{\boldsymbol{z}}(t)+\boldsymbol{\epsilon}_{\boldsymbol{z}}(t)=\boldsymbol{U} \widetilde{\boldsymbol{f}}(t)+\boldsymbol{\epsilon}_{\boldsymbol{z}}(t), \tag{2.14}
\end{gather*}
$$

where $\widetilde{\boldsymbol{z}}(t)$ is the resulting filtered data and $\boldsymbol{\epsilon}_{z}(t)$ is the extracted noise. $\boldsymbol{f}(t)$ in (2.5) is the exact solution of (2.14) when $\boldsymbol{\epsilon}_{z}(t)=0$. By (2.9), we can express $\boldsymbol{f}(t)$ as

$$
\boldsymbol{f}(t)=\left[\begin{array}{c}
f_{1}(t) \\
\vdots \\
f_{k}(t)
\end{array}\right]=\left[\begin{array}{c}
s_{1} \\
v_{1}(t) \\
\vdots \\
s_{k} v_{k}(t)
\end{array}\right]=\sum_{i=1}^{k}\left(s_{i} v_{i}(t)\right) \boldsymbol{e}_{i},
$$

where $e_{i}$ is the $i$ th column of the identity matrix. Since we expect that the unimportant principal components $f_{i}(t)$ are more contaminated by the noise, we reduce the contribution of these principal components. We apply a filtering matrix $\boldsymbol{\Phi}=\operatorname{diag}\left(\phi_{1}, \ldots, \phi_{k}\right)$ to $\boldsymbol{f}(t)$ with each $\phi_{i} \in[0,1]$ so that

$$
\widetilde{\boldsymbol{f}}(t)=\boldsymbol{\Phi} \boldsymbol{f}(t)
$$

The element $\phi_{i}$ should be small when $s_{i}$ is small. The resulting filtered data are

$$
\begin{equation*}
\widetilde{\boldsymbol{z}}(t)=\boldsymbol{U} \boldsymbol{\Phi} \boldsymbol{f}(t), \tag{2.15}
\end{equation*}
$$

$$
\begin{equation*}
\widetilde{Z}=U \Phi F \tag{2.16}
\end{equation*}
$$

We introduce two different filtering matrices, $\boldsymbol{\Phi}_{\text {trun }}(\widehat{k})$ and $\boldsymbol{\Phi}_{t i k h}(\alpha)$, which correspond to TSVD and Tikhonov regularization. First, we can simply truncate all but $\widehat{k}$ most important components as Elton and Gruber [13] did by using a filtering matrix of $\boldsymbol{\Phi}_{\text {trun }}(\widehat{k})=$ $\operatorname{diag}(\underbrace{1, \ldots, 1}_{\widehat{k}}, \underbrace{0, \ldots, 0}_{k-\widehat{k}})$, so the truncated principal component $\widetilde{\boldsymbol{f}}_{\text {trun }}(t)$ is

$$
\widetilde{\boldsymbol{f}}_{\text {trun }}(t)=\boldsymbol{\Phi}_{\text {trun }}(\widehat{k}) \boldsymbol{f}(t) .
$$

By (2.15) and (2.16), the resulting filtered data are $\widetilde{\boldsymbol{z}}_{\text {trun }}(t)=\boldsymbol{U} \boldsymbol{\Phi}_{\text {trun }}(\widehat{k}) \boldsymbol{f}(t)$ and $\widetilde{\boldsymbol{Z}}_{\text {trun }}=$ $\boldsymbol{U} \boldsymbol{\Phi}_{\text {trun }}(\widehat{k}) \boldsymbol{F}$. Since $\boldsymbol{F}=\boldsymbol{S}_{k} \boldsymbol{V}_{k}^{T}$, we can rewrite $\widetilde{\boldsymbol{Z}}_{\text {trun }}$ as

$$
\begin{equation*}
\widetilde{\boldsymbol{Z}}_{t r u n}=\boldsymbol{U} \boldsymbol{\Phi}_{\text {trun }}(\widehat{k})\left(\boldsymbol{S}_{k} \boldsymbol{V}_{k}^{T}\right)=\sum_{i=1}^{\widehat{k}} s_{i} \boldsymbol{u}_{i} \boldsymbol{v}_{i}^{T} . \tag{2.17}
\end{equation*}
$$

From (2.17), we can see that this truncation method corresponds to the TSVD [14, 20].
Second, we can apply the Tikhonov method, and this is our approach to estimating the covariance matrix. We formulate the regularized least squares problem to solve (2.12) as

$$
\begin{equation*}
\min _{\tilde{\boldsymbol{f}}(t)} M(\widetilde{\boldsymbol{f}}(t)) \tag{2.18}
\end{equation*}
$$

with

$$
M(\tilde{\boldsymbol{f}}(t))=\|\boldsymbol{z}(t)-\tilde{\boldsymbol{U}} \tilde{\boldsymbol{f}}(t)\|^{2}+\alpha^{2}\|\tilde{\boldsymbol{P f}}(t)\|^{2}
$$

where $\alpha^{2}$ is a penalty parameter and $\boldsymbol{P}$ is a penalty matrix. The first term $\|\boldsymbol{z}(t)-\widetilde{\boldsymbol{U}} \widetilde{\boldsymbol{f}}(t)\|^{2}$ forces $\widetilde{\boldsymbol{f}}(t)$ to be close to the exact solution $\boldsymbol{f}(t)$. The second term $\|\widetilde{\boldsymbol{P}}(t)\|^{2}$ controls the size of $\widetilde{\boldsymbol{f}}(t)$. We can choose, for example,

$$
\boldsymbol{P}=\operatorname{diag}\left(s_{1}^{-1}, \ldots, s_{k}^{-1}\right)
$$

Let $\tilde{f}_{i}(t)$ denote the $i$ th element of $\tilde{\boldsymbol{f}}(t)$. The matrix $\boldsymbol{P}$ scales each $\tilde{f}_{i}(t)$ by $s_{i}^{-1}$, so the unimportant principal components corresponding to small $s_{i}$ are penalized more than the more important principal components since we expect that the unimportant principal components $f_{i}(t)$ are more contaminated by the noise. Thus, the penalty term prevents $\tilde{\boldsymbol{f}}(t)$ from containing large amounts of unimportant principal components. As we showed before, $s_{i}^{2}$ is proportional to the variance of the $i$ th principal component $f_{i}(t)$. Therefore, this penalty matrix $\boldsymbol{P}$ is statistically meaningful considering that the values of $\tilde{f}_{i}(t) / s_{i}$ in $\widetilde{\boldsymbol{P f}}(t)$ are in proportion to the normalized principal components $\widetilde{f}_{i}(t) / \sqrt{\operatorname{Var}_{s}\left[f_{i}(t)\right]}$.

The penalty parameter $\alpha$ balances the minimization between the error term $\| \boldsymbol{z}(t)-$ $\boldsymbol{U} \widetilde{\boldsymbol{f}}(t) \|^{2}$ and the penalty term $\|\widetilde{\boldsymbol{P}}(t)\|^{2}$. Therefore, as $\alpha$ increases, the regularized solution $\widetilde{f}(t)$ moves away from the exact solution $\boldsymbol{f}(t)$ but should discard more of $\boldsymbol{f}(t)$ as noise. We can quantify this property by determining the solution to (2.18). At the minimizer of (2.18), the gradient of $M(\widetilde{\boldsymbol{f}}(t))$ with respect to each $\widetilde{f}_{i}(t)$ becomes zero, so

$$
\nabla M(\tilde{\boldsymbol{f}}(t))=2 \boldsymbol{U}^{T} \boldsymbol{U} \widetilde{\boldsymbol{f}}(t)-2 \boldsymbol{U}^{T} \boldsymbol{z}(t)+2 \alpha^{2} \boldsymbol{P}^{T} \boldsymbol{P} \tilde{\boldsymbol{f}}(t)=0
$$

and thus

$$
\left(\boldsymbol{U}^{T} \boldsymbol{U}+\alpha^{2} \boldsymbol{P}^{T} \boldsymbol{P}\right) \tilde{\boldsymbol{f}}(t)=\boldsymbol{U}^{T} \boldsymbol{z}(t)
$$

Since $\boldsymbol{U}^{T} \boldsymbol{U}=\boldsymbol{I}_{k}, \boldsymbol{P}=\operatorname{diag}\left(s_{1}^{-1}, \ldots, s_{k}^{-1}\right)$, and $\boldsymbol{z}(t)=\boldsymbol{U} \boldsymbol{f}(t)$, this becomes

$$
\left(\boldsymbol{I}_{k}+\alpha^{2} \operatorname{diag}\left(s_{1}^{-2}, \ldots, s_{k}^{-2}\right)\right) \widetilde{\boldsymbol{f}}(t)=\boldsymbol{U}^{T}(\boldsymbol{U} \boldsymbol{f}(t))
$$

Therefore,

$$
\operatorname{diag}\left(\frac{s_{1}^{2}+\alpha^{2}}{s_{1}^{2}}, \ldots, \frac{s_{k}^{2}+\alpha^{2}}{s_{k}^{2}}\right) \widetilde{\boldsymbol{f}}(t)=\boldsymbol{f}(t)
$$

and

$$
\widetilde{\boldsymbol{f}}(t)=\operatorname{diag}\left(\frac{s_{1}^{2}}{s_{1}^{2}+\alpha^{2}}, \ldots, \frac{s_{k}^{2}}{s_{k}^{2}+\alpha^{2}}\right) \boldsymbol{f}(t)
$$

So, our Tikhonov estimate is

$$
\widetilde{\boldsymbol{f}}_{t i k h}(t)=\boldsymbol{\Phi}_{t i k h}(\alpha) \boldsymbol{f}(t)
$$

where $\boldsymbol{\Phi}_{t i k h}(\alpha)$, called the Tikhonov filtering matrix, denotes $\left(\boldsymbol{S}_{k}^{2}+\alpha^{2} \boldsymbol{I}_{k}\right)^{-1} \boldsymbol{S}_{k}^{2}$. Thus, we can see that the regularized principal component $\widetilde{\boldsymbol{f}}_{t i k h}(t)$ is the result after filtering the original principal component $\boldsymbol{f}(t)$ with the diagonal matrix $\boldsymbol{\Phi}_{t i k h}(\alpha)$, whose diagonal elements $\phi_{i}^{t i k h}(\alpha)=\frac{s_{i}^{2}}{s_{i}^{2}+\alpha^{2}}$ lie in $[0,1] . \quad$ By (2.15) and (2.16), the resulting filtered data become


Figure 1. Tikhonov filtering as a function of $s_{i}$ for various values of $\alpha$.
$\widetilde{\boldsymbol{z}}_{t i k h}(t)=\boldsymbol{U} \boldsymbol{\Phi}_{t i k h}(\alpha) \boldsymbol{f}(t)$ and $\widetilde{\boldsymbol{Z}}_{t i k h}=\boldsymbol{U} \boldsymbol{\Phi}_{t i k h}(\alpha) \boldsymbol{F}$. Let us see how $\phi_{i}^{t i k h}(\alpha)$ changes as $\alpha$ and $s_{i}$ vary. First, as $\alpha$ increases, $\phi_{i}^{t i k h}(\alpha)$ decreases, as illustrated in Figure 1. This is reasonable since $\alpha$ balances the error term and the penalty term. Later, in section 4 , we will propose how we can determine an appropriate parameter $\alpha$. Second, $\phi_{i}^{t i k h}(\alpha)$ monotonically increases as $s_{i}$ increases, so the Tikhonov filter matrix reduces the less important principal components more intensely. The main difference between the Tikhonov method and TSVD is that Tikhonov preserves some information from the least important principal components while TSVD discards all of it.
2.3. The relation between filtered PCA and a factor model. Some asset pricing models (see, e.g., $[40,44]$ ) model asset returns with a factor model:

$$
\begin{equation*}
\boldsymbol{r}(t)=\mathbb{E}[\boldsymbol{r}(t)]+\mathcal{B} \boldsymbol{\varphi}(t)+\boldsymbol{\epsilon}(t) \tag{2.19}
\end{equation*}
$$

The assumptions are that

$$
\begin{equation*}
\mathbb{E}[\boldsymbol{\varphi}(t)]=\mathbb{E}[\boldsymbol{\epsilon}(t)]=\mathbf{0} \tag{2.20}
\end{equation*}
$$

$$
\begin{equation*}
\mathbb{E}\left(\epsilon_{i}(t) \epsilon_{j}(t)\right)=\mathbb{E}\left(\epsilon_{i}(t) \varphi_{\ell}(t)\right)=\mathbb{E}\left(\varphi_{i}(t) \varphi_{j}(t)\right)=0 \quad \text { for all } i \neq j \tag{2.21}
\end{equation*}
$$

where $\boldsymbol{\varphi}(t)=\left[\varphi_{1}(t), \ldots, \varphi_{\ell}(t)\right]^{T}$ and $\boldsymbol{\epsilon}(t)=\left[\epsilon_{1}(t), \ldots, \epsilon_{N}(t)\right]^{T}$. The common factors $\varphi_{i}(t)$ are referred to as systematic factors, and $\epsilon_{i}(t)$ is called an unsystematic (idiosyncratic) factor. The matrix $\mathcal{B}=\left(\beta_{i k}\right)$ is called a factor-loading matrix, and $\beta_{i k}$ represents the sensitivity of the $i$ th asset to the $k$ th factor.

We can interpret our linear model (2.14) as a factor model. By (2.3) and (2.14), we have a linear equation for $\boldsymbol{r}(t)$ as

$$
\begin{align*}
\boldsymbol{r}(t) & =\mathbb{E}_{s}[\boldsymbol{r}(t)]+\boldsymbol{D}_{V}^{\frac{1}{2}}\left(\widetilde{\boldsymbol{U}}(t)+\boldsymbol{\epsilon}_{z}(t)\right)  \tag{2.22}\\
& =\mathbb{E}_{s}[\boldsymbol{r}(t)]+\boldsymbol{B} \tilde{\boldsymbol{f}}(t)+\boldsymbol{\epsilon}_{r}(t), \tag{2.23}
\end{align*}
$$

where

$$
\begin{equation*}
\boldsymbol{B}=\boldsymbol{D}_{V}^{\frac{1}{2}} \boldsymbol{U} \quad \text { and } \quad \boldsymbol{\epsilon}_{r}(t)=\boldsymbol{D}_{V}^{\frac{1}{2}} \boldsymbol{\epsilon}_{z}(t) \tag{2.24}
\end{equation*}
$$

Comparing (2.19) and (2.23), if we assume that $\widetilde{\boldsymbol{f}}(t)$ represents the systematic factors $\varphi(t)$ well, we can interpret $\boldsymbol{B}$ and $\boldsymbol{\epsilon}_{r}(t)$ as estimates of the loading matrix $\mathcal{B}$ and the unsystematic factor $\boldsymbol{\epsilon}(t)$ in (2.19). Since $\boldsymbol{\epsilon}_{z}(t)=\boldsymbol{z}(t)-\widetilde{\boldsymbol{U f}}(t), \boldsymbol{\epsilon}_{r}(t)$ becomes

$$
\begin{equation*}
\boldsymbol{\epsilon}_{r}(t)=\boldsymbol{D}_{V}^{\frac{1}{2}} \boldsymbol{\epsilon}_{z}(t)=\boldsymbol{D}_{V}^{\frac{1}{2}}(\boldsymbol{z}(t)-\widetilde{\boldsymbol{U}}(t)) \tag{2.25}
\end{equation*}
$$

Because $\boldsymbol{z}(t)=\boldsymbol{U f}(t)$ and $\widetilde{\boldsymbol{f}}(t)=\boldsymbol{\Phi} \boldsymbol{f}(t)$, the factor models result in the estimate

$$
\begin{equation*}
\boldsymbol{\epsilon}_{r}(t)=\boldsymbol{D}_{V}^{\frac{1}{2}}(\boldsymbol{U} \boldsymbol{f}(t)-\boldsymbol{U} \boldsymbol{\Phi} \boldsymbol{f}(t))=\left(\boldsymbol{D}_{V}^{\frac{1}{2}} \boldsymbol{U}\right)\left(\boldsymbol{I}_{k}-\boldsymbol{\Phi}\right) \boldsymbol{f}(t)=\boldsymbol{B}\left(\boldsymbol{I}_{k}-\boldsymbol{\Phi}\right) \boldsymbol{f}(t) \tag{2.26}
\end{equation*}
$$

3. Estimate of the covariance matrix $\boldsymbol{\Sigma}$. In this section we study how filtering changes the covariance and correlation estimates and the estimate of risk exposure, and how to ensure that the estimated covariance matrix has full rank.
3.1. A covariance estimate. Now we derive a covariance matrix estimate $\widetilde{\boldsymbol{\Sigma}}$ from (2.23), respecting the structure of the factor model (2.19). By (2.21), the covariance matrix $\boldsymbol{\Sigma}$ is

$$
\begin{equation*}
\boldsymbol{\Sigma}=\mathcal{B} \operatorname{Cov}[\boldsymbol{\varphi}(t)] \mathcal{B}^{T}+\operatorname{Cov}[\boldsymbol{\epsilon}(t)]=\boldsymbol{\Sigma}_{s}+\boldsymbol{D}_{\epsilon} \tag{3.1}
\end{equation*}
$$

where $\boldsymbol{\Sigma}_{s}$ denotes the systematic component $\mathcal{B} \operatorname{Cov}[\boldsymbol{\varphi}(t)] \mathcal{B}^{T}$ and $\boldsymbol{D}_{\epsilon}$ denotes the unsystematic component $\operatorname{Cov}[\boldsymbol{\epsilon}(t)]$. We estimate the systematic part $\boldsymbol{\Sigma}_{s}$ by $\widetilde{\boldsymbol{\Sigma}}_{s}=\boldsymbol{B C o v}_{s}[\widetilde{\boldsymbol{f}}(t)] \boldsymbol{B}^{T}$. Because $\boldsymbol{f}(t)$ has zero-mean, $\widetilde{\boldsymbol{f}}(t)=\boldsymbol{\Phi} \boldsymbol{f}(t)$ also has zero-mean, so

$$
\begin{equation*}
\operatorname{Cov}_{s}[\widetilde{\boldsymbol{f}}(t)]=\frac{1}{T}(\boldsymbol{\Phi} \boldsymbol{F})(\boldsymbol{\Phi} \boldsymbol{F})^{T}=\frac{1}{T}\left(\boldsymbol{\Phi}^{2} \boldsymbol{S}_{k}^{2}\right) \tag{3.2}
\end{equation*}
$$

Therefore, the estimate of $\boldsymbol{\Sigma}_{s}$ becomes

$$
\begin{equation*}
\widetilde{\boldsymbol{\Sigma}}_{s}=\boldsymbol{B C o v}_{s}[\widetilde{\boldsymbol{f}}(t)] \boldsymbol{B}^{T}=\frac{1}{T} \boldsymbol{B}\left(\boldsymbol{\Phi}^{2} \boldsymbol{S}_{k}^{2}\right) \boldsymbol{B}^{T} \tag{3.3}
\end{equation*}
$$

The unsystematic part $\boldsymbol{D}_{\epsilon}$ in (3.1) is diagonal since the unsystematic factors $\epsilon_{i}(t)$ are mutually uncorrelated. Thus, we estimate $\operatorname{Cov}[\boldsymbol{\epsilon}(t)]$ by the diagonal part of the difference $\widetilde{\boldsymbol{D}}_{\epsilon}$ between

$$
\begin{equation*}
\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}=\operatorname{Cov}_{s}[\boldsymbol{r}(t)]=\frac{1}{T} \boldsymbol{B} \boldsymbol{S}_{k}^{2} \boldsymbol{B}^{T} \tag{3.4}
\end{equation*}
$$

and $\widetilde{\boldsymbol{\Sigma}}_{s}$. Hence,

$$
\begin{equation*}
\widetilde{\boldsymbol{D}}_{\epsilon}=\operatorname{diag}\left(\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}-\widetilde{\boldsymbol{\Sigma}}_{s}\right)=\operatorname{diag}\left(\frac{1}{T}\left(\boldsymbol{B}\left(\boldsymbol{I}_{k}-\boldsymbol{\Phi}^{2}\right) \boldsymbol{S}_{k}^{2} \boldsymbol{B}^{T}\right)\right) \tag{3.5}
\end{equation*}
$$

Finally, the filtered covariance matrix $\widetilde{\boldsymbol{\Sigma}}$ will be

$$
\begin{equation*}
\tilde{\boldsymbol{\Sigma}}=\tilde{\boldsymbol{\Sigma}}_{s}+\widetilde{\boldsymbol{D}}_{\epsilon} \tag{3.6}
\end{equation*}
$$

where $\widetilde{\boldsymbol{\Sigma}}_{s}$ and $\widetilde{\boldsymbol{D}}_{\epsilon}$ are defined by (3.3) and (3.5). By the definition of $\widetilde{\boldsymbol{D}}_{\epsilon}$, the diagonal of $\widetilde{\boldsymbol{\Sigma}}$ equals $\operatorname{Var}_{s}[\boldsymbol{r}(t)]$.

Now we analyze how the filtering function $\boldsymbol{\Phi}$ affects the sample correlation matrix $\operatorname{Corr}_{s}[\boldsymbol{r}(t)]$. By (3.6), the filtered correlation matrix $\widetilde{\boldsymbol{\Omega}}$ can be calculated as

$$
\begin{equation*}
\widetilde{\boldsymbol{\Omega}}=\boldsymbol{D}_{V}^{-\frac{1}{2}} \widetilde{\boldsymbol{\Sigma}} \boldsymbol{D}_{V}^{-\frac{1}{2}}=\frac{1}{T} \boldsymbol{U} \boldsymbol{\Phi}^{2} \boldsymbol{S}_{k}^{2} \boldsymbol{U}^{T}+\boldsymbol{D}_{V}^{-\frac{1}{2}} \tilde{\boldsymbol{D}}_{\epsilon} \boldsymbol{D}_{V}^{-\frac{1}{2}} \tag{3.7}
\end{equation*}
$$

where the second term makes the diagonal elements of $\widetilde{\boldsymbol{\Omega}}$ equal one. On the other hand, the sample correlation matrix $\operatorname{Corr}_{s}[\boldsymbol{r}(t)]$ can be calculated as

$$
\operatorname{Corr}_{s}[\boldsymbol{r}(t)]=\boldsymbol{D}_{V}^{-\frac{1}{2}} \widetilde{\boldsymbol{\Sigma}}_{\text {sample }} \boldsymbol{D}_{V}^{-\frac{1}{2}}
$$

By (2.24) and (3.4), this becomes

$$
\begin{equation*}
\operatorname{Corr}_{s}[\boldsymbol{r}(t)]=\boldsymbol{D}_{V}^{-\frac{1}{2}}\left(\frac{1}{T} \boldsymbol{B} \boldsymbol{S}_{k}^{2} \boldsymbol{B}^{T}\right) \boldsymbol{D}_{V}^{-\frac{1}{2}}=\frac{1}{T} \boldsymbol{U} \boldsymbol{S}_{k}^{2} \boldsymbol{U}^{T} \tag{3.8}
\end{equation*}
$$

Comparing $\widetilde{\boldsymbol{\Omega}}$ in (3.7) and $\operatorname{Corr}_{s}[\boldsymbol{r}(t)]$ in (3.8), we can see that $\widetilde{\boldsymbol{\Omega}}$ is the result of applying the filtering matrix $\boldsymbol{\Phi}^{2}$ to $\boldsymbol{S}_{k}^{2}$ in $\operatorname{Corr}_{s}[\boldsymbol{r}(t)]$ and replacing the diagonal elements with one. Since each diagonal element of $\boldsymbol{S}_{k}^{2}$ corresponds to an eigenvalue of $\operatorname{Corr}_{s}[\boldsymbol{r}(t)]$, the filtering matrix $\boldsymbol{\Phi}^{2}$ attenuates the eigenvalues of $\operatorname{Corr}_{s}[\boldsymbol{r}(t)]$. In the previous section, we introduced two filtering matrices:

$$
\begin{equation*}
\boldsymbol{\Phi}_{\text {trun }}(\widehat{k})=\operatorname{diag}(\underbrace{1, \ldots, 1}_{\widehat{k}}, \underbrace{0, \ldots, 0}_{k-\widehat{k}}) \tag{3.9}
\end{equation*}
$$

and

$$
\begin{equation*}
\boldsymbol{\Phi}_{t i k h}(\alpha)=\operatorname{diag}\left(\frac{s_{1}^{2}}{s_{1}^{2}+\alpha^{2}}, \ldots, \frac{s_{k}^{2}}{s_{k}^{2}+\alpha^{2}}\right) \tag{3.10}
\end{equation*}
$$

Therefore, $\boldsymbol{\Phi}_{\text {trun }}^{2}(\widehat{k})$ truncates the eigencomponents corresponding to the $(k-\widehat{k})$ smallest eigenvalues, and $\boldsymbol{\Phi}_{\text {tikh }}^{2}(\alpha)$ downweights all the eigenvalues at a rate $\left(\frac{s_{i}^{2}}{s_{i}^{2}+\alpha^{2}}\right)^{2}=\left(\frac{\lambda_{i}}{\lambda_{i}+\alpha^{2}}\right)^{2}$, where $\lambda_{i}$ is the $i$ th largest eigenvalue of $\operatorname{Cov}_{s}[\boldsymbol{z}(t)]$. Hence, the TSVD filtering functions $\phi_{\text {trun }}^{2}\left(\lambda_{i}\right)$ for eigenvalues $\lambda_{i}$ become

$$
\phi_{t r u n}^{2}\left(\lambda_{i}\right)= \begin{cases}1 & \text { if } i \leq \widehat{k} \\ 0 & \text { otherwise }\end{cases}
$$

and the Tikhonov filtering functions $\phi_{t i k h}^{2}\left(\lambda_{i}\right)$ are

$$
\phi_{t i k h}^{2}\left(\lambda_{i}\right)=\left(\frac{\lambda_{i}}{\lambda_{i}+\alpha^{2}}\right)^{2}
$$

We let $\widetilde{\boldsymbol{\Sigma}}_{\text {trun }}$ and $\widetilde{\boldsymbol{\Sigma}}_{\text {tikh }}$ denote the estimates resulting from applying $\boldsymbol{\Phi}_{\text {trun }}^{2}(\widehat{k})$ and $\boldsymbol{\Phi}_{\text {tikh }}^{2}(\alpha)$ to (3.6). Finally, we can summarize the process of estimating the covariance matrix as Table 1.

Table 1
The algorithm to compute the covariance estimate $\widetilde{\boldsymbol{\Sigma}}$. For Tikhonov, the filter factors are $\boldsymbol{\Phi}_{\text {tikh }}=$ $\operatorname{diag}\left(\frac{s_{1}^{2}}{s_{1}^{2}+\alpha^{2}}, \ldots, \frac{s_{k}^{2}}{s_{k}^{2}+\alpha^{2}}\right)$.

Step 1. Estimate the systematic component of the covariance $\frac{1}{T} \boldsymbol{B}\left(\boldsymbol{\Phi}^{2} \boldsymbol{S}_{k}^{2}\right) \boldsymbol{B}^{T}$, where $\boldsymbol{\Phi}$ is the diagonal matrix of filter factors.
Step 2. Change the main diagonal to be the sample variances.
3.2. Risk exposure to factors. By (3.1), the variance of a portfolio return can be expressed as

$$
\begin{equation*}
\boldsymbol{w}^{T} \boldsymbol{\Sigma} \boldsymbol{w}=\boldsymbol{w}^{T}\left(\boldsymbol{\Sigma}_{s}+\boldsymbol{D}_{\epsilon}\right) \boldsymbol{w}=\boldsymbol{w}^{T} \boldsymbol{\Sigma}_{s} \boldsymbol{w}+\boldsymbol{w}^{T} \boldsymbol{D}_{\epsilon} \boldsymbol{w} \tag{3.11}
\end{equation*}
$$

The systematic risk is

$$
\begin{equation*}
\boldsymbol{w}^{T} \boldsymbol{\Sigma}_{s} \boldsymbol{w}=\boldsymbol{w}^{T}\left(\mathcal{B} \operatorname{Cov}[\boldsymbol{\varphi}(t)] \mathcal{B}^{T}\right) \boldsymbol{w}=\boldsymbol{w}^{T}\left(\mathcal{B} \operatorname{diag}(\operatorname{Var}[\boldsymbol{\varphi}(t)]) \mathcal{B}^{T}\right) \boldsymbol{w} \tag{3.12}
\end{equation*}
$$

because $\varphi_{i}(t)$ are mutually uncorrelated by (2.21). This can be expanded as

$$
\begin{equation*}
\boldsymbol{w}^{T} \boldsymbol{\Sigma}_{s} \boldsymbol{w}=\sum_{i=1}^{k} \operatorname{Var}\left[\varphi_{i}(t)\right]\left(\boldsymbol{w}^{T} \boldsymbol{\beta}_{i}\right)^{2} \tag{3.13}
\end{equation*}
$$

where $\boldsymbol{\beta}_{i}$ is the $i$ th column of $\mathcal{B}$. The $i$ th term in (3.13) represents the risk exposure of the portfolio to the $i$ th factor.

On the other hand, the estimated matrix $\widetilde{\boldsymbol{\Sigma}}_{s}$ in (3.3) can be rewritten as

$$
\begin{equation*}
\widetilde{\boldsymbol{\Sigma}}_{s}=\frac{1}{T} \boldsymbol{B} \boldsymbol{\Phi}^{2} \boldsymbol{S}_{k}^{2} \boldsymbol{B}^{T}=\boldsymbol{B} \boldsymbol{\Phi}^{2}\left(\frac{\boldsymbol{S}_{k}^{2}}{T}\right) \boldsymbol{B}^{T}=\boldsymbol{B} \boldsymbol{\Phi}^{2} \operatorname{diag}\left(\operatorname{Var}_{s}[\boldsymbol{f}(t)]\right) \boldsymbol{B}^{T} \tag{3.14}
\end{equation*}
$$

because $\operatorname{Var}_{s}[\boldsymbol{f}(t)]=\operatorname{diag}\left(\boldsymbol{S}_{k}^{2} / T\right)$ by (2.10). Hence, we can calculate the estimated systematic risk as

$$
\begin{equation*}
\boldsymbol{w}^{T} \widetilde{\boldsymbol{\Sigma}}_{s} \boldsymbol{w}=\sum_{i=1}^{k} \phi_{i}^{2}\left(\operatorname{Var}_{s}\left[\boldsymbol{f}_{i}(t)\right]\left(\boldsymbol{w}^{T} \boldsymbol{b}_{i}\right)^{2}\right), \tag{3.15}
\end{equation*}
$$

where $\boldsymbol{b}_{i}$ is the $i$ th column of $\boldsymbol{B}$. Therefore, we can see that our estimate of the risk exposure to the $i$ th factor is reduced by $\phi_{i}^{2}$. This equation explains how the estimated covariance matrix $\widetilde{\boldsymbol{\Sigma}}$ affects the estimated risk measure of a portfolio, downweighting risk factors corresponding to small values of $\phi_{i}(\alpha)$.
3.3. Rank deficiency of the covariance matrix. Since the covariance matrix is positive semidefinite, the MV problem (1.1) and the GMV problem (1.2) always have a minimizer $\boldsymbol{w}$. However, when the covariance matrix is rank deficient, the minimizer $w$ is not unique, which might not be desirable for investors who want to choose one portfolio. The sample covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$ from (1.3) has rank $(T-1)$ at most. Therefore, whenever the number of
observations $T$ is less than or equal to the number of stocks $N, \widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$ is rank deficient. To ensure full rank and a high quality estimate, we must have at least $(N+1)$ recent observations of returns, derived from at least $(N+1)$ recent trades, and this is not always possible.

Recall that the covariance matrix estimate $\widetilde{\boldsymbol{\Sigma}}$ is the sum of the systematic part $\widetilde{\boldsymbol{\Sigma}}_{s}$ and the unsystematic part $\widetilde{\boldsymbol{D}}_{\epsilon}$. By (3.3), we can see that $\widetilde{\boldsymbol{\Sigma}}_{s}$ has nonnegative eigenvalues. On the other hand, by (3.5),

$$
\begin{equation*}
\text { (The } \left.i \text { th diagonal element of } \widetilde{\boldsymbol{D}}_{\epsilon}\right)=\boldsymbol{e}_{i}^{T}\left(\frac{1}{T} \boldsymbol{B}\left(\boldsymbol{I}_{k}-\boldsymbol{\Phi}^{2}\right) \boldsymbol{S}_{k}^{2} \boldsymbol{B}^{T}\right) \boldsymbol{e}_{i} \tag{3.16}
\end{equation*}
$$

It is reasonable to assume that $\boldsymbol{e}_{i}^{T} \boldsymbol{B}$ is not zero for any $i$ since it becomes zero only when the $i$ th stock has zero variance of returns by (2.24). Thus, the diagonal matrix $\widetilde{\boldsymbol{D}}_{\epsilon}$ is positive definite whenever all $\phi_{i}<1$. In the case of Tikhonov filtering, whenever $\alpha>0$,

$$
\phi_{i}^{t i k h}(\alpha)=\frac{s_{i}^{2}}{s_{i}^{2}+\alpha^{2}}<1
$$

so $\widetilde{\boldsymbol{D}}_{\epsilon}$ is positive definite. Therefore, since $\widetilde{\boldsymbol{\Sigma}}_{s}$ is positive semidefinite, adding a positive definite matrix ensures that Tikhonov covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{t i k h}$ is positive definite and therefore full rank.

Sharpe [43], Ledoit and Wolf [30], Bengtsson and Holst [2], and Plerou et al. [38] also overcome the rank-deficiency problem by replacing the diagonals of their estimate with the sample variances as in Step 2 in Table 1. However, some of their filtering values $\phi_{i}$ could have a value of 1 as we will see in section 5 . This implies that the resulting estimate $\widetilde{\boldsymbol{\Sigma}}$ could be rank-deficient or very ill conditioned even after adding $\widetilde{\boldsymbol{D}}_{\epsilon}$ because $\widetilde{\boldsymbol{D}}_{\epsilon}$ is positive semidefinite. In the case that the estimate still has a large condition number even after Step 2, we can fix the problem by a small modification as follows:

$$
\begin{equation*}
\widetilde{\boldsymbol{\Sigma}}_{i i} \leftarrow \widetilde{\boldsymbol{\Sigma}}_{i i}+\delta_{i} \quad \text { for } i=1, \ldots, N, \tag{3.17}
\end{equation*}
$$

where $\delta_{i}$ is a small positive number.
Theorem 3.1 (condition number modification). Replacing the main diagonal of the covariance estimate $\widetilde{\boldsymbol{\Sigma}}$ as specified in (3.17) guarantees that

$$
\operatorname{cond}(\widetilde{\boldsymbol{\Sigma}}) \leq \frac{\lambda_{\max }(\tilde{\boldsymbol{\Sigma}})+\max \left(\delta_{i}\right)}{\min \left(\delta_{i}\right)}
$$

where $\lambda_{\max }(\cdot)$ is the maximum eigenvalue of the matrix.
Proof. This is a direct consequence of the eigenvalue interlacing theorem [45, p. 203] and the positive semidefiniteness of $\widetilde{\boldsymbol{\Sigma}}$.

This modification is useful especially for the sample covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$ when $T \leq$ $N$, and for the truncation-based estimators whose filtering factors $\phi_{i}$ equal 1 for some $i$.
4. Choice of Tikhonov parameter $\boldsymbol{\alpha}$. So far, we have seen how to filter noise from the covariance matrix using regularization and how to fix the rank deficiency of the resulting covariance matrix. In order to use Tikhonov regularization, we need to determine the Tikhonov


Figure 2. The difference $\left\|\operatorname{Corr}_{s}\left[\boldsymbol{\epsilon}_{r}(t)\right]-\mathbf{I}_{N}\right\|_{F}$ as a function of log-scaled $\alpha$ where $h=\max \left(s_{i}\right)$.
parameter $\alpha$. In regularization methods for discrete ill-posed problems, there are intensive studies about choosing $\alpha$ using methods such as generalized cross-validation [15], L-curves [17, 19], and residual periodograms [41, 42].

In factor analysis and PCA, there are analogous studies to determine the number of factors such as Bartlett's test [1], SCREE test [4], average root [16], partial correlation procedure [50], and cross-validation [51]. More recently, Plerou et al. [37, 38] applied random matrix theory, which will be described in section 5.6. In the context of arbitrage pricing theory, some different approaches were proposed to determine the number of factors: Trzcinka [48] studied the behavior of eigenvalues as the number of assets increases, and Connor and Korajczyk [9] studied the probabilistic behavior of noise factors.

The use of these methods requires various statistical properties for $\boldsymbol{\epsilon}_{r}(t)$ in the linear model (2.23). We note that since $\mathbb{E}_{s}[\boldsymbol{f}(t)]=\mathbf{0}$ by (2.8), the noise $\boldsymbol{\epsilon}_{r}(t)$ in (2.23) has zero-mean: By (2.26),

$$
\begin{equation*}
\mathbb{E}_{s}\left[\boldsymbol{\epsilon}_{r}(t)\right]=\boldsymbol{B}\left(\boldsymbol{I}_{k}-\boldsymbol{\Phi}\right) \mathbb{E}_{s}[\boldsymbol{f}(t)]=\mathbf{0} . \tag{4.1}
\end{equation*}
$$

For our Tikhonov estimation, we propose a new method adopting a mutually uncorrelated noise assumption in a factor model (2.21), so $\operatorname{Corr}_{s}\left[\boldsymbol{\epsilon}_{r}(t)\right] \simeq \boldsymbol{I}_{N}$. Hence, as a criterion to determine an appropriate parameter $\alpha$, we formulate an optimization problem minimizing the correlations among the noise,

$$
\begin{equation*}
\min _{\alpha \in\left[s_{k}, s_{1}\right]}\left\|\operatorname{Corr}_{s}\left[\boldsymbol{\epsilon}_{r}(t)\right]-\boldsymbol{I}_{N}\right\|_{F}, \tag{4.2}
\end{equation*}
$$

where $s_{1}$ and $s_{k}$ are the largest and the smallest singular values of $\boldsymbol{Z}$ as defined in (2.6). This is similar to Velicer's partial correlation procedure [50] to determine the number of principal components. Figure 2 illustrates an example of $\left\|\operatorname{Corr}_{s}\left[\boldsymbol{\epsilon}_{r}(t)\right]-\boldsymbol{I}_{N}\right\|_{F}$ as a function of $\alpha$ in the range $\left[s_{k}, s_{1}\right]$. The parameter might alternatively be determined by an asymptotic analysis proposed by Ledoit and Wolf [30, 31] or a cross-validation used by DeMiguel et al. [10].
5. Comparison to other estimators. In this section, we compare other covariance estimators to our Tikhonov estimator and put them all in a common framework. We summarize how they filter the eigenvalues of the sample correlation matrix with filtering functions $\phi^{2}\left(\lambda_{i}\right)$. Most of these methods use a two-step procedure as shown in Table 1: filter the eigenvalues and then adjust the main diagonal. We note any exceptions in our descriptions.
5.1. $\widetilde{\Sigma}_{\text {sample }}$ : Sample covariance matrix. A sample covariance matrix is the filtering target of most covariance estimators, including our Tikhonov estimator. Thus, the sample covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$ can be thought of as an unfiltered covariance matrix, so the filtering function $\phi_{s}^{2}\left(\lambda_{i}\right)$ for eigenvalues of $\operatorname{Cov}_{s}[\boldsymbol{z}(t)]$ is

$$
\phi_{s}^{2}\left(\lambda_{i}\right)=1 \quad \text { for } i=1, \ldots, \operatorname{rank}\left(\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}\right) .
$$

5.2. $\widetilde{\Sigma}_{\text {market }}$ from the single market index model [43]. Sharpe [43] proposed a single index market model

$$
\begin{equation*}
\boldsymbol{r}(t)=\mathbb{E}[\boldsymbol{r}(t)]+\boldsymbol{b} r_{m}(t)+\boldsymbol{\epsilon}(t), \tag{5.1}
\end{equation*}
$$

where $\boldsymbol{r}(t) \in \mathbb{R}^{N \times 1}$ is stock return at time $t$,
$r_{m}(t)$ is market return at time $t$,
$\boldsymbol{\epsilon}(t)$ is zero-mean uncorrelated error at time $t$,
and $\boldsymbol{b} \in \mathbb{R}^{N \times 1}$.
Unlike the factor model (2.19), this model assumes that the stock returns $\boldsymbol{r}(t)$ have only one common factor, the market return $r_{m}(t)$. Interestingly, Plerou et al. [38, p. 8] observed that the principal component corresponding to the largest eigenvalue of the correlation matrix $\operatorname{Corr}_{s}[\boldsymbol{r}(t)]\left(=\operatorname{Cov}_{s}[\boldsymbol{z}(t)]\right)$ is highly correlated with $r_{m}(t)$. This observation is natural in that most stocks are highly affected by the market situation. Based on their observation, we expect that the most important principal component $f_{1}(t)$ in (2.5) represents the market return $r_{m}(t)$. Thus, we can represent the relation between $\tilde{\boldsymbol{f}}(t)=\left[\tilde{f}_{1}(t), \ldots, \tilde{f}_{k}(t)\right]$ in $(2.23)$ and $r_{m}(t)$ as

$$
\tilde{f}_{i}(t) \simeq \begin{cases}C r_{m}(t) & \text { when } i=1  \tag{5.2}\\ 0 & \text { otherwise }\end{cases}
$$

for some constant $C$. Hence, the corresponding filtering function $\phi_{m}^{2}\left(\lambda_{i}\right)$ for $\widetilde{\boldsymbol{\Sigma}}_{\text {market }}$ becomes

$$
\phi_{m}^{2}\left(\lambda_{i}\right) \simeq \begin{cases}1 & \text { if } i=1  \tag{5.3}\\ 0 & \text { otherwise }\end{cases}
$$

Therefore, the filter function implicitly truncates all but the largest eigencomponent of $\operatorname{Corr}_{s}[\boldsymbol{r}(t)]$.
5.3. $\widetilde{\Sigma}_{s \rightarrow m}$ : Shrinkage toward $\widetilde{\Sigma}_{\text {market }}$ [30]. Ledoit and Wolf propose a shrinkage method from $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$ to $\widetilde{\boldsymbol{\Sigma}}_{\text {market }}$ as

$$
\begin{equation*}
\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow m}=\gamma \widetilde{\boldsymbol{\Sigma}}_{\text {market }}+(1-\gamma) \widetilde{\boldsymbol{\Sigma}}_{\text {sample }}, \tag{5.4}
\end{equation*}
$$

where $0 \leq \gamma \leq 1$. Thus, the shrinkage estimator is the weighed average of $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$ and $\widetilde{\boldsymbol{\Sigma}}_{\text {market }}$. In order to find an optimal weight $\gamma$, they minimize the distance between $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow m}$ and the true covariance matrix $\boldsymbol{\Sigma}$ :

$$
\min _{\gamma}\left\|\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow m}-\boldsymbol{\Sigma}\right\|_{F}^{2}
$$

Since the true covariance matrix $\boldsymbol{\Sigma}$ is unknown, they use an asymptotic variance to determine an optimal $\gamma$. (Refer to [30, sections 2.5-2.6] for a detailed description.) Considering that $\widetilde{\boldsymbol{\Sigma}}_{\text {market }}$ is the result of the implicit truncation method, we can think of this shrinkage method as implicitly downweighting all eigenvalues but the largest at a rate $(1-\gamma)$. Therefore, we can represent the filtering function $\phi_{s \rightarrow m}^{2}\left(\lambda_{i}\right)$ as

$$
\phi_{s \rightarrow m}^{2}\left(\lambda_{i}\right) \simeq \begin{cases}1 & \text { if } i=1  \tag{5.5}\\ 1-\gamma, \text { where } 0 \leq \gamma \leq 1 & \text { otherwise }\end{cases}
$$

5.4. Truncated covariance matrix $\widetilde{\Sigma}_{\text {trun }}$ [13]. As mentioned in section 3.1, the truncated covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{\text {trun }}$ has the filtering function $\phi_{\text {trun }}^{2}\left(\lambda_{i}\right)$ for the eigenvalues $\lambda_{i}$ of $\operatorname{Cov}_{s}[\boldsymbol{z}(t)]$, where

$$
\phi_{\text {trun }}^{2}\left(\lambda_{i}\right)= \begin{cases}1 & \text { if } i=1, \ldots, \widehat{k}  \tag{5.6}\\ 0 & \text { otherwise }\end{cases}
$$

Thus, the model of Elton and Gruber [13] truncates all but the $\widehat{k}$ largest eigencomponents of $\operatorname{Cov}_{s}[\boldsymbol{z}(t)]$.
5.5. $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow \text { trun }}$ : Shrinkage toward $\widetilde{\Sigma}_{\text {trun }}$ [2]. Bengtsson and Holst propose a shrinkage estimator from $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$ to $\widetilde{\boldsymbol{\Sigma}}_{\text {trun }}$ as

$$
\begin{equation*}
\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow \text { trun }}=\gamma \widetilde{\boldsymbol{\Sigma}}_{\text {trun }}+(1-\gamma) \widetilde{\boldsymbol{\Sigma}}_{\text {sample }}, \tag{5.7}
\end{equation*}
$$

where $0 \leq \gamma \leq 1$. They determine the parameter $\gamma$ in a way similar to [30]. (Refer to [2, sections 4.1-4.2] for a detailed description.) Therefore, $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow \text { trun }}$ is a variant of the shrinkage method toward $\widetilde{\boldsymbol{\Sigma}}_{\text {trun }}$. Because $\widetilde{\boldsymbol{\Sigma}}_{\text {trun }}$ is the truncated covariance matrix containing the $\widehat{k}$ most significant eigencomponents of $\operatorname{Cov}_{s}[\boldsymbol{z}(t)]$, we can regard $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow \text { trun }}$ as damping the smallest eigenvalues by $(1-\gamma)$. Thus, the filtering function corresponding to this approach is

$$
\phi_{s \rightarrow \text { trun }}^{2}\left(\lambda_{i}\right)= \begin{cases}1 & \text { if } i=1, \ldots, \widehat{k},  \tag{5.8}\\ 1-\gamma, \text { where } 0 \leq \gamma \leq 1 & \text { otherwise }\end{cases}
$$

Rather than removing all the least important principal components as Elton and Gruber did, Bengtsson and Holst try to preserve the potential information of unimportant principal components by this single-rate attenuation. Bengtsson and Holst conclude that their shrinkage matrix $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow \text { trun }}$ performed best in the Swedish stock market when the shrinkage target $\widetilde{\boldsymbol{\Sigma}}_{\text {trun }}$ took only the most significant principal component $(\widehat{k}=1)$. They also mention that the result is consistent with random matrix theory because only the largest eigenvalue deviates far from the range of $\left[\lambda_{\min }, \lambda_{\max }\right]$, explained in the next section.
5.6. $\widetilde{\Sigma}_{R M T: \text { trun }}$ truncation by random matrix theory [38]. Plerou et al. apply random matrix theory (RMT) [34], which shows that the eigenvalues of a random correlation matrix have a distribution within an interval determined by the ratio of $N$ and $T$. Let Corr $_{\text {random }}$ be a random correlation matrix,

$$
\begin{equation*}
\text { Corr }_{\text {random }}=\frac{1}{T} \boldsymbol{A} \boldsymbol{A}^{T}, \tag{5.9}
\end{equation*}
$$

where $\boldsymbol{A} \in \mathbb{R}^{N \times T}$ contains mutually independent random elements $a_{i, t}$ with zero-mean and unit variance. When $Q=T / N \geq 1$ is fixed, the eigenvalues $\lambda$ of $\operatorname{Corr}_{\text {random }}$ have a limiting distribution (as $N \rightarrow \infty$ ),

$$
f(\lambda)=\left\{\begin{array}{cl}
\frac{Q}{2 \pi \sigma^{2}} \frac{\sqrt{\left(\lambda_{\max }-\lambda\right)\left(\lambda_{\min }-\lambda\right)}}{\lambda} & \lambda_{\min } \leq \lambda \leq \lambda_{\max }  \tag{5.10}\\
0 & \text { otherwise }
\end{array}\right.
$$

where $\sigma^{2}$ is the variance of the elements of $\boldsymbol{A}, \lambda_{\min } \leq \lambda \leq \lambda_{\max }$, and $\lambda_{\min }^{\max }=\sigma^{2}\left(1+\frac{1}{Q} \pm\right.$ $\left.2 \sqrt{\frac{1}{Q}}\right)$. By comparing the eigenvalue distribution of $\operatorname{Corr}_{s}[\boldsymbol{r}(t)]$ with $f(\lambda)$, Plerou et al. show that most eigenvalues are within $\left[\lambda_{\min }, \lambda_{\max }\right.$ ]. They conclude that only a few large eigenvalues deviating from $\left[\lambda_{\min }, \lambda_{\max }\right]$ correspond to eigenvalues of the real correlation matrix, so the other eigencomponents should be removed from $\operatorname{Corr}_{s}[\boldsymbol{r}(t)]$. Thus, the filtering function $\phi_{R M T: \text { trun }}^{2}\left(\lambda_{i}\right)$ for the eigenvalue $\lambda_{i}$ of $\operatorname{Corr}_{s}[\boldsymbol{r}(t)]$ is

$$
\phi_{R M T: \text { trun }}^{2}\left(\lambda_{i}\right)= \begin{cases}1 & \text { if } \lambda_{i} \geq \lambda_{\max },  \tag{5.11}\\ 0 & \text { otherwise }\end{cases}
$$

5.7. $\widetilde{\Sigma}_{R M T: \text { repl }}$ replacing the RMT eigenvalues [28]. Laloux et al. apply RMT to this problem in a way somewhat different from Plerou et al. First, they find the best fitting $\sigma^{2}$ in (5.10) to the eigenvalue distribution of the observed correlation matrix rather than assuming that $\sigma^{2}=1$. Second, they replace each eigenvalue in the RMT interval with a constant value $C$, chosen so that the trace of the matrix is unchanged. Thus, the filtering function $\phi_{R M T: \text { :repl }}^{2}\left(\lambda_{i}\right)$ for eigenvalues is

$$
\phi_{R M T: \text { repl }}^{2}\left(\lambda_{i}\right)= \begin{cases}1 & \text { if } \lambda_{i} \geq \lambda_{\max }  \tag{5.12}\\ \frac{C}{\lambda_{i}} & \text { otherwise }\end{cases}
$$

This approach does not require the application of Step 2 in Table 1 since it replaces the smallest eigenvalues with a positive constant. The resulting covariance matrix does not preserve the original variances.
5.8. $\widetilde{\Sigma}_{s \rightarrow I}$ : Shrinkage toward I [31]. Ledoit and Wolf also introduced a shrinkage method from $\boldsymbol{\Sigma}_{\text {sample }}$ to the identity matrix $\boldsymbol{I}_{N}$ as

$$
\begin{equation*}
\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow I}=\gamma\left(m \boldsymbol{I}_{N}\right)+(1-\gamma) \widetilde{\boldsymbol{\Sigma}}_{\text {sample }}, \tag{5.13}
\end{equation*}
$$

where $m=\frac{\operatorname{trace}\left(\tilde{\boldsymbol{\Sigma}}_{\text {sample }}\right)}{N}$ and $0 \leq \gamma \leq 1$. They provide a method to estimate an optimal $\gamma$. (Refer to [31, section 3] for a detailed description.) There is no simple expression for the filter factors. In addition, this method does not use Step 2 in Table 1 since its shrinkage target $\boldsymbol{I}_{N}$ has full rank.

Table 2
Definition of the filter function $\phi^{2}\left(\lambda_{i}\right)$ for each covariance estimator, where $i=1, \ldots, \operatorname{rank}\left(\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}\right)$.

| Estimator | Filtering function $\phi^{2}\left(\lambda_{i}\right)$ |
| :--- | :--- |
| $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$ | $\phi_{s}^{2}\left(\lambda_{i}\right)=1$ |
| $\widetilde{\boldsymbol{\Sigma}}_{\text {market }}[43]$ | $\phi_{m}^{2}\left(\lambda_{i}\right) \simeq \begin{cases}1 & \text { if } i=1, \\ 0 & \text { otherwise. }\end{cases}$ |
| $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow m}[30]$ | $\phi_{s \rightarrow m}^{2}\left(\lambda_{i}\right) \simeq \begin{cases}1 & \text { if } i=1, \\ 1-\gamma & \text { otherwise. }\end{cases}$ |
| $\widetilde{\boldsymbol{\Sigma}}_{\text {trun }}[13]$ | $\phi_{\text {trun }}^{2}\left(\lambda_{i}\right)= \begin{cases}1 & \text { if } i=1, \ldots, k, \\ 0 & \text { otherwise. }\end{cases}$ |
| $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow \text { trun }}[2]$ | $\phi_{s \rightarrow \text { trun }}^{2}\left(\lambda_{i}\right)= \begin{cases}1 & \text { if } i=1, \ldots, \widehat{k}, \\ 1-\gamma & \text { otherwise. }\end{cases}$ |
| $\widetilde{\boldsymbol{\Sigma}}_{R M T: \text { trun }}[38]$ | $\phi_{R M T: \text { trun }}^{2}\left(\lambda_{i}\right)= \begin{cases}1 & \text { if } \lambda_{i} \geq \lambda_{\text {max }}, \\ 0 & \text { otherwise. }\end{cases}$ |
| $\widetilde{\boldsymbol{\Sigma}}_{R M T: \text { repl }}[28]$ | $\phi_{R M T: \text { repl }}^{2}\left(\lambda_{i}\right)= \begin{cases}1 & \text { if } \lambda_{i} \geq \lambda_{\text {max }}, \\ \frac{C}{\lambda_{i}} & \text { otherwise. }\end{cases}$ |
| $\widetilde{\boldsymbol{\Sigma}}_{\text {tikh }}$ | $\phi_{\text {tikh }}^{2}\left(\lambda_{i}\right)=\left(\frac{\lambda_{i}}{\lambda_{i}+\alpha^{2}}\right)^{2}$ |

5.9. Tikhonov covariance matrix $\widetilde{\Sigma}_{t i k h}$. As mentioned in section 3.1, the Tikhonov covariance matrix $\boldsymbol{\Sigma}_{t i k h}$ has the filtering function $\phi_{t i k h}^{2}\left(\lambda_{i}\right)$ for the eigenvalues $\lambda_{i}$ of $\operatorname{Cov}_{s}[\boldsymbol{z}(t)]$, where

$$
\begin{equation*}
\phi_{t i k h}^{2}\left(\lambda_{i}\right)=\left(\frac{\lambda_{i}}{\lambda_{i}+\alpha^{2}}\right)^{2} \tag{5.14}
\end{equation*}
$$

where the parameter $\alpha$ is determined as described in section 4 .
5.10. Comparison. The derivations in section 5 provide the proof of the following theorem.

Theorem 5.1 (filtering functions). The eight covariance estimators are characterized by the choice of filtering functions specified in Table 2.

Tikhonov filtering preserves potential information from less important principal components corresponding to small eigenvalues, rather than truncating them all like $\widetilde{\boldsymbol{\Sigma}}_{\text {market }}, \widetilde{\boldsymbol{\Sigma}}_{\text {trun }}$, and $\widetilde{\boldsymbol{\Sigma}}_{R M T: \text { trun }}$. In contrast to the single-rate attenuation of $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow m}$ and $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow \text { trun }}$ and the constant value replacement of $\widetilde{\boldsymbol{\Sigma}}_{R M T: r e p l}$, Tikhonov filtering reduces the effect of the smallest eigenvalues more intensely. This gradual downweighting with respect to the magnitude of eigenvalues is the key difference between the Tikhonov method and other estimators.

In addition, all the estimators except $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow I}$ and $\widetilde{\boldsymbol{\Sigma}}_{R M T: \text { repl }}$ overcome the rank deficiency of the covariance matrix by replacing the diagonal elements with the corresponding variances after filtering. This is what we did by preserving $\widetilde{\boldsymbol{D}}_{\epsilon}$ in Step 2 in Table 1. However, most
estimators have $\phi^{2}\left(\lambda_{i}\right)=1$ for the largest eigenvalues as Table 2 shows, so the resulting covariance matrix can be still rank deficient as we discussed in section 3.3. During experiments in section 6 , we actually observed the rank deficiency for some estimators even after preserving diagonal parts. This implies that an extra modification like (3.17) is necessary to overcome rank deficiency.
6. Experiments. In this section, we evaluate the covariance estimators using return data from the NYSE, AMEX, and NASDAQ. We collected the monthly data from January 1958 to December 2007 from the CRSP (Center for Research in Security Prices) database. There are 600 months over 50 years, and we randomly chose 100 stocks among those traded throughout this period.

Chopra and Ziemba [7] have noted that the MV problem is much more sensitive to errors in $\boldsymbol{\mu}$ than to errors in $\boldsymbol{\Sigma}$, and our experience confirms this observation. In fact, uncertainty in the estimates of $\boldsymbol{\mu}$ made the true return quite different from the target return. In addition, recently DeMiguel, Garlappi, and Uppal [11] showed that some common portfolio strategies do not yield consistently better Sharpe ratios, certainty-equivalent returns, or turnovers, compared to a naive $1 / N$ portfolio. The instability of the MV portfolio tends to increase turnover costs, so recent studies strengthen the stability by formulating new optimization problems [12]. However, since our study focuses on estimating the covariance matrix $\boldsymbol{\Sigma}$, we evaluated the estimators based on how well they minimize the risk variances in the MV and GMV portfolios.

First, in section 6.1, we evaluate the risk of the GMV portfolio using the covariance estimators of Table 2 with various in-sample periods. We then compare the stability and performance of the Tikhonov estimator to those of the shrinkage estimate $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow m}$. Next, in section 6.2, we perform similar experiments for the MV portfolio, varying the in-sample and out-of-sample periods as well as the required portfolio returns. We bypass the difficulties of estimating $\mu$ by assuming that it is known so that we can focus just on the effects of the different covariance estimators. Finally, in section 6.3, we compare the GMV and MV portfolio returns, and in section 6.4 we compare their predictions of risk.
6.1. GMV portfolio. We simulate portfolio construction under the following scenario. We solve the GMV problem to construct a portfolio to hold for 1 month, the out-of-sample period $T_{o}$. We repeat this process for every month until we reach December 2007. Finally, we evaluate the variance of the out-of-sample returns from the GMV portfolio for each covariance estimator.

When performing this experiment, the choice of in-sample window size $T_{w}$ is important. If $T_{w}$ is too long, the data may include out-of-date information. On the other hand, if $T_{w}$ is too short, the resulting covariance estimate could suffer from lack of information. We vary $T_{w}$ from 1 year to 10 years. Later, in section 6.2, we will consider the change of the out-of-sample period $T_{o}$ as well. We start each experiment at January 1968, giving 480 rebalancing steps for all values of $T_{w}$. For each covariance estimator, we perform the simulation for 20 different choices of 100 stocks.
6.1.1. Covariance estimators in experiments. We perform the experiment above for all the covariance estimators from section 5.1 to section 5.9 plus two diagonal matrices, $\widetilde{\boldsymbol{\Sigma}}_{V}$ and
$\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{I}}$, for a total of 11 estimators. $\widetilde{\boldsymbol{\Sigma}}_{V}$ has diagonal elements equal to $\operatorname{Var}_{s}[\boldsymbol{r}(t)]$, and any correlations between stocks are neglected. $\widetilde{\boldsymbol{\Sigma}}_{I}$ is an $N \times N$ identity matrix, which would yield an evenly distributed portfolio as the solution for the GMV problem (1.2); thus it is a good benchmark for a well-distributed portfolio. Since $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$ is rank deficient, we modify it by adding small positive constants $\delta_{i}$ to its diagonal elements as in (3.17). To compute $\widetilde{\boldsymbol{\Sigma}}_{\text {market }}$ and $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow m}$, we need the monthly market return data $r_{m}(t)$ in (5.1). In this experiment, we adopt equally weighted market portfolio returns including distributions from the CRSP database as $r_{m}(t)$. According to Ledoit and Wolf [30, p. 607], an equally weighted market portfolio is better than a value-weighted market portfolio for explaining stock market variances.

The parameter $\widehat{k}$ for $\widetilde{\boldsymbol{\Sigma}}_{\text {trun }}$ and $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow \text { trun }}$ is static, constant over all time periods. In our experiment, we perform the experiments with $\widehat{k}=1,5,9$ for $\widetilde{\boldsymbol{\Sigma}}_{\text {trun }}$ and $\widehat{k}=1,2,3$ for $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow \text { trun }}$. In contrast, the parameters of $\gamma$ for $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow m}$ and $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow \text { trun }}, \widehat{k}$ for $\widetilde{\boldsymbol{\Sigma}}_{R M T: \text { trun }}$ and $\widetilde{\boldsymbol{\Sigma}}_{R M T \text { :repl }}$, and $\alpha$ for $\widetilde{\boldsymbol{\Sigma}}_{t i k h}$ have their own parameter choice methods as described in section 5 , so we dynamically determine these parameters each time the portfolio is rebalanced.

Figure 3 shows singular value plots from each estimator, which illustrates the filtering characteristics for the first in-sample period of $T_{w}=4$ years with a particular set of 100 stocks.
6.1.2. Effect of in-sample period $\boldsymbol{T}_{\boldsymbol{w}}$. For each randomly chosen data set ( $i=1, \ldots, 20$ ), we calculate $\left(\sigma_{i}\right)_{\tilde{\Sigma}}$, the annualized standard deviation of the sample portfolio return, by multiplying the monthly standard deviation by $\sqrt{12}$. The subscript $\widetilde{\boldsymbol{\Sigma}}$ denotes the specific choice of covariance estimator. Figures $4(\mathrm{a})$ and $4(\mathrm{~b})$ show the means of $\left(\sigma_{i}\right)_{\tilde{\boldsymbol{\Sigma}}}$ for the static estimators and the dynamic estimators. The standard deviations of the $\left(\sigma_{i}\right)_{\tilde{\Sigma}}$ from each estimator were at most 0.56 for all time periods, except for the occurrence of values up to 3.38 for $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$ and up to 6.50 for $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow \operatorname{trun}(\widehat{k}=3)}$, so the results did not seem sensitive to the particular choice of 100 stocks.

For most estimators, the $\left(\sigma_{i}\right)_{\tilde{\boldsymbol{\Sigma}}}$ decrease until a particular $T_{w}$ and increase after that point, showing the advantage of using a sufficient amount of history but not too much out-of-date information. This is particularly evident for $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$ since it assumes that all of its data are reliable. At the opposite extreme, $\left(\sigma_{i}\right)_{\widetilde{\boldsymbol{\Sigma}}_{\text {market }}}$ from $\widetilde{\boldsymbol{\Sigma}}_{\text {market }}$ increases with $T_{w}$, which implies that the correlation among stocks cannot be fully explained by a single market index. For small values of $k, \widetilde{\boldsymbol{\Sigma}}_{\text {trun }}$ behaves like $\widetilde{\boldsymbol{\Sigma}}_{\text {market }}$, but performance can be improved by taking $k \approx 5$, making the estimator less sensitive to out-of-date information. The diagonal $\widetilde{\boldsymbol{\Sigma}}_{V}$ shows a better tolerance to out-of-date information than $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$, which may imply that the sample variance estimation is less sensitive to the choice of $T_{w}$ than the sample covariance estimation. The estimators that dynamically determine the filtering parameters ( $\widetilde{\boldsymbol{\Sigma}}_{t i k h}, \widetilde{\boldsymbol{\Sigma}}_{s \rightarrow m}, \widetilde{\boldsymbol{\Sigma}}_{s \rightarrow I}$, $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow \operatorname{trun}(\widehat{k}=1)}, \widetilde{\boldsymbol{\Sigma}}_{R M T: \text { repl }}$, and $\widetilde{\boldsymbol{\Sigma}}_{R M T: \text { trun }}$ ) also show good tolerance. Therefore, modestly filtered factor structures are better at filtering the out-of-date information than a singlefactor or full-factor structure, but all estimators benefit from an appropriate choice of window size.

Compared to the truncation-based estimators like $\widetilde{\boldsymbol{\Sigma}}_{R M T: \text { trun }}$ and $\widetilde{\boldsymbol{\Sigma}}_{\text {trun }}$, Tikhonov generally performs better when the in-sample period is shorter than its own optimal size, which


Figure 3. GMV portfolios: the singular values from each estimator when $T_{w}=4$ years.
is $T_{w}=4$. This result can be explained by the characteristics of their filtering functions. While $\phi_{t i k h}^{2}\left(\lambda_{i}\right)$ preserves the relative magnitudes of eigenvalues by gradual attenuation, $\phi_{R M T: \text { trun }}^{2}\left(\lambda_{i}\right)$ or $\phi_{\text {trun }}^{2}\left(\lambda_{i}\right)$ discard them all. Thus, when the smallest eigenvalues are still important, the Tikhonov filter shows superiority empirically. However, as noise level increases with longer $T_{w}$, the performance reverses.

Compared to the other shrinkage-based estimators, Tikhonov filtering $\phi_{t i k h}^{2}\left(\lambda_{i}\right)$ preserves


Figure 4. GMV portfolios: the mean of $\left(\sigma_{i}\right)_{\tilde{\Sigma}}$ over different choices of $T_{w}$.
the smallest but still informative factors better than a single rate reduction by $\phi_{s \rightarrow m}^{2}\left(\lambda_{i}\right)$ and $\phi_{s \rightarrow \text { trun }}^{2}\left(\lambda_{i}\right)$ or a replacement with a constant value by $\phi_{R M T: \text { :epl }}^{2}\left(\lambda_{i}\right)$ when $T_{w}$ is relatively short $\left(T_{w}<4\right)$. On the other hand, for $T_{w}>7$, it becomes evident that $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow m}, \widetilde{\boldsymbol{\Sigma}}_{s \rightarrow \operatorname{trun}(\widehat{k}=1)}$, and $\widetilde{\boldsymbol{\Sigma}}_{R M T: \text { repl }}$ show better performances than $\widetilde{\boldsymbol{\Sigma}}_{t i k h}$. This is because $\widetilde{\boldsymbol{\Sigma}}_{t i k h}$ has relatively weaker tolerance to the contamination by out-of-date information.
6.1.3. Stability of Tikhonov parameter choice. In this section, we evaluate the stability of our parameter choice method from section 4. For a particular choice of 100 stocks, we observe the change of the dynamic parameters $\alpha$ for $\widetilde{\boldsymbol{\Sigma}}_{t i k h}$ and $\gamma$ for $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow m}$. In this experiment,


Figure 5. GMV portfolios: the performance of static and dynamic choices of $\alpha$ and $\gamma$ in 20 experiments.
we set the window size as $T_{w}=48$ because both estimators have the smallest mean value of $\left(\sigma_{i}\right)_{\widetilde{\boldsymbol{\Sigma}}}$ for that window size.

Figure $5(\mathrm{a})$ illustrates the change of the ratio of the dynamically chosen Tikhonov parameter $\alpha_{D}$ to the largest singular value $s_{1}$ of $\operatorname{Corr}_{s}[\boldsymbol{r}(t)]$ and the change of $\gamma_{D}$ for $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow m}$. The results for 20 choices of the 100 stocks are shown, showing that both parameter choice methods for $\alpha_{D}$ and $\gamma_{D}$ are quite stable during the whole experiment. The resulting annualized standard deviations of $\left(\sigma_{i}\right)_{\tilde{\boldsymbol{\Sigma}}}$ range from $10.16 \%$ to $10.30 \%$ for $\widetilde{\boldsymbol{\Sigma}}_{t i k h}$ and $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow m}$, for both the static and dynamically determined parameters.

We repeated this numerical experiment keeping the ratio $\alpha / s_{1}$ and the parameter $\gamma$ constant over all time periods. (We use the notation $\alpha_{S}$ and $\gamma_{S}$ for this statically determined parameter.) This static parameter choice may not be practical in real market trading since we cannot access the future return information when we construct a portfolio. However, we can find a statically optimal ratio from this experiment for a comparison to $\alpha_{D} / s_{1}$ and $\gamma_{D}$. Figure $5(\mathrm{~b})$ shows how the standard deviation of portfolio returns changes as $\alpha_{S} / s_{1}$ and $\gamma_{S}$ increase. The optimal ratio $\alpha_{S}^{*} / s_{1}$ was 0.27 with resulting standard deviation of portfolio returns $10.16 \%$, and the optimal $\gamma_{S}^{*}$ was 0.59 with resulting standard deviation $10.27 \%$. These statically optimal values are represented by dashed lines in Figure 5(a). Therefore, we can see that both $\alpha_{D} / s_{1}$ and $\gamma_{D}$ remain near their statically optimal values $\alpha_{S}^{*} / s_{1}$ and $\gamma_{S}^{*}$. Moreover, the static and varying $\alpha$ values produce similar risk variance.
6.2. MV portfolio. Now, we observe the behavior of the MV portfolio resulting from each covariance estimator. In this experiment, we vary the out-of-sample period $T_{o}$ and the required portfolio return $q$ as well as the in-sample period $T_{w}$. We change $T_{o}$ from 2 months to 6 months, ${ }^{2} T_{w}$ from 1 year to 10 years, and $q$ from $0 \%$ to $20 \%$. As we mentioned before, the performance of the MV portfolio is quite sensitive to the estimation of stock returns $\boldsymbol{\mu}$.

[^98]Avg. std of portfolio returns as a function of $t_{w}$ and $t_{o}$ by $\widetilde{\Sigma}_{t i k h}$


Avg. std of portfolio returns as a function of $t_{w}$ and $t_{o}$ by $\widetilde{\Sigma}_{t i k h}$

Figure 6. MV portfolios: the average annualized standard deviations $\left(\sigma_{i}\right)_{\tilde{\Sigma}_{t i k h}}$ of portfolio returns as in-sample period $T_{w}$ and out-of-sample period $T_{o}$ change with different settings of required portfolio return $q$.

In order to evaluate covariance estimation with no influence of mean estimation, we assume a perfect prediction of stock returns $\boldsymbol{\mu}$, which means we estimate $\boldsymbol{\mu}$ by the average $\boldsymbol{r}(t)$ during the out-of-sample period.
6.2.1. Effect of out-of-sample period $\boldsymbol{T}_{\boldsymbol{o}}$. The out-of-sample period $T_{o}$ determines how fast we react to the changes in the market. Figure 6 shows how the average $\left(\sigma_{i}\right)_{\widetilde{\boldsymbol{\Sigma}}_{t i k h}}$ changes as $T_{o}$ and $T_{w}$ vary for $q=0 \%, 10 \%$, and $20 \%$. We can see that $\left(\sigma_{i}\right)_{\widetilde{\boldsymbol{\Sigma}}_{t i k h}}$ has a tendency to increase as we hold the portfolio for longer $T_{o}$. Similar results were obtained for all other covariance estimators.
6.2.2. Effect of in-sample period $\boldsymbol{T}_{\boldsymbol{w}}$. Similar to Figure 4 for the GMV experiment, we compared the mean of $\left(\sigma_{i}\right)_{\widetilde{\Sigma}}$ for different covariance estimators with varying $T_{w}$ and $q$ in Figure 7. Based on the result of section 6.2.1, we fixed $T_{o}$ as 2 months in order to compare


Figure 7. $M V$ portfolios: the mean of $\left(\sigma_{i}\right)_{\tilde{\Sigma}}$ over different choices of $T_{w}$ and $q$ when $T_{o}=2$ months.
the smallest standard deviations from the estimators. The behaviors of MV portfolios with respect to the change of $T_{w}$ are very similar to GMV portfolios for most covariance estimators. For example, as we observed for the previous GMV experiments, the MV portfolios in Figure 7 also suffered from lack of information when $T_{w}$ was too short and suffered from out-of-date


Figure 8. MV portfolios: the average annualized $\left(\sigma_{i}\right)_{\tilde{\Sigma}}$ versus required return $q$ for each estimator when $T_{o}=2$ months and $T_{w}=3$ years .
information when $T_{w}$ was too long. This implies that the choice of window size $T_{w}$ is very important for the MV portfolio as well as the GMV portfolio. Moreover, each estimator shows very similar shapes of curves for the GMV and the MV problems, except that the curves for the MV problems tend to shift upward as $q$ increases.

However, in contrast to the GMV problem, where most of the competitive estimators have optimal $T_{w}$ around 4 years, the optimal $T_{w}$ for most estimators was around 3 years for the MV problem (gray-colored vertical dot-dash lines indicate $T_{w}=3$ years in Figure 7). This may be because they have different out-of-sample periods: $T_{o}=1$ month for the GMV problem in Figure 4 and $T_{o}=2$ months for the MV problem in Figure 7.
6.2.3. Effect of required portfolio return $\boldsymbol{q}$. Figure $6(\mathrm{~d})$ summarizes the results from Figure 6(a) to Figure 6(c). As we can expect, the surfaces of $\left(\sigma_{i}\right)_{\widetilde{\boldsymbol{\Sigma}}_{t i k h}}$ move upward as $q$ increases. For all the estimators $\widetilde{\boldsymbol{\Sigma}}$ with particular choices of $T_{o}=2$ months and $T_{w}=3$ years, Figure 8 also shows that $\left(\sigma_{i}\right)_{\widetilde{\Sigma}}$ gradually increases as $q$ increases from $0 \%$ to $20 \%$, which explains a trade-off between risk and return from the MV portfolio.
6.2.4. Efficiency of portfolio. The mean-variance plot shows the efficiency of the MV portfolios. Let $\left(\mu_{i}\right)_{\widetilde{\Sigma}}$ denote the annualized mean of the realized portfolio returns in the $i$ th random choice of 100 stocks $(i=1, \ldots, 20)$. In order to evaluate the portfolio efficiency by each estimator, we compare the change of average $\left(\mu_{i}\right)_{\tilde{\boldsymbol{\Sigma}}}$ versus the change of average $\left(\sigma_{i}\right)_{\widetilde{\boldsymbol{\Sigma}}}$ with varying the required return $q$ from $0 \%$ to $20 \%$. Figure 9 presents the average of realized means and standard deviations of all the estimators for the cases of $T_{o}=2$ months and $T_{w}=1$ year or 3 years. Curves to the left of and above the others correspond to the more efficient portfolios.

When $T_{w}=1$ year, where we have insufficient historical data, $\widetilde{\boldsymbol{\Sigma}}_{t i k h}$ generates the most efficient portfolios (see Figure $9(\mathrm{~b})$ ). The shrinkage estimators with a target of a single factor like $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow m}$ and $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow \operatorname{trun}(k=1)}$ are also efficient compared to other dynamic estimators. When $T_{w}=3$ years, where we have near optimal historical data, $\widetilde{\boldsymbol{\Sigma}}_{t i k h}, \widetilde{\boldsymbol{\Sigma}}_{s \rightarrow m}, \widetilde{\boldsymbol{\Sigma}}_{R M T: r e p l}$, and $\widetilde{\boldsymbol{\Sigma}}_{s \rightarrow m}$ generate relatively efficient portfolios (see Figure $9(\mathrm{~d})$ ).


Figure 9. MV portfolios: the average annualized $\left(\mu_{i}\right)_{\tilde{\Sigma}}$ versus average annualized $\left(\sigma_{i}\right)_{\tilde{\boldsymbol{\Sigma}}}$.
6.3. Comparison of GMV and MV portfolios. Now we observe how the covariance estimators affect the realized portfolio returns at every rebalancing point for the GMV and the MV problems. For instance, Figure 10 shows the fluctuations of the portfolio returns by $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$ and $\widetilde{\boldsymbol{\Sigma}}_{\text {tikh }}$ at the first 100 rebalancing points when $T_{w}=3$ years and $T_{o}=2$ months. While the annualized returns of the GMV portfolios fluctuate around $11 \%$, the annualized returns of the MV portfolios fluctuate around their required return $q$. Note that the GMV mean return is greater than that for the MV portfolio with $q=0 \%$. Similarly, the standard deviations in Figure 4 are greater than the corresponding ones in Figures 7(a) and 7(b).

On the other hand, for both GMV and MV, the $\widetilde{\boldsymbol{\Sigma}}_{t i k h}$ portfolios have greater mean return and smaller variance than those from $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$, which implies more efficient portfolios. This result is consistent with the plots of means versus standard deviations in Figure 9.
6.4. Risk prediction. Laloux et al. [29] showed empirically that their estimator $\widetilde{\boldsymbol{\Sigma}}_{R M T: \text { repl }}$ predicts the risk more accurately than $\widetilde{\boldsymbol{\Sigma}}_{\text {sample }}$. They simply divided the dataset into two equal time periods for in-sample and out-of-sample periods and compared the estimated standard deviation $\left(\boldsymbol{w}^{T} \widetilde{\boldsymbol{\Sigma}} \boldsymbol{w}\right)^{\frac{1}{2}}$ from (1.1) to the realized standard deviation $\left(\sigma_{i}\right)_{\tilde{\boldsymbol{\Sigma}}}$ for the out-of-sample period. They assumed perfect prediction for means of stock returns as we did in section 6.2.

We evaluate the accuracy of the risk prediction of each covariance estimator in a similar way. However, rather than following their equal division of in-sample and out-of-sample periods, we varied $T_{w}$ with $T_{o}=2$ months, and we simulated the rebalancing scenario as


Figure 10. GMV and MV portfolios: the annualized portfolio returns at the rebalancing points for the $G M V$ and the MV problems with different required returns $q$.
in section 6.2. Finally, we compute the relative difference between the average estimated standard deviations from (1.1) and the average realized standard deviations for the most competitive estimators.

Figure 11 shows the relative difference for the cases of $T_{w}=1$ and 3 years, which correspond to the case of insufficient historical data and the minimizer of average $\left(\sigma_{i}\right)_{\tilde{\Sigma}}$. The realized standard deviations were greater than the estimated standard deviations for all estimators. However, it turns out that $\widetilde{\boldsymbol{\Sigma}}_{t i k h}$ has the smallest difference for both cases, giving us the best risk prediction.
7. Conclusion. In this study, we applied Tikhonov regularization to improve the covariance matrix estimate used in the Markowitz portfolio selection problem. We put the previous covariance estimators in a common framework based on the filtering function $\phi^{2}\left(\lambda_{i}\right)$ for the eigenvalues of $\operatorname{Corr}_{s}[\boldsymbol{r}(t)]$. The Tikhonov estimator $\widetilde{\boldsymbol{\Sigma}}_{t i k h}$ attenuates smaller eigenvalues more intensely, which is a key difference between it and the other filter functions.

In order to choose an appropriate Tikhonov parameter $\alpha$ that determines the intensity of attenuation, we formulated an optimization problem minimizing the difference between $\operatorname{Corr}_{s}\left[\boldsymbol{\epsilon}_{z}(t)\right]$ and $\boldsymbol{I}_{N}$ based on the assumption that the unsystematic factors are uncorrelated.

We performed empirical experiments to evaluate covariance estimators. For the GMV portfolio selection problem, the Tikhonov choice gave the smallest average standard deviation of the return when the out-of-sample period was 3 or 4 years, and it was not much worse than competitors for other periods. The choice of parameter was relatively stable. For the MV


Figure 11. MV portfolios: the relative differences between average estimated risks and average realized risks by each covariance matrix with varying different required returns $q$.
portfolio selection problem, the Tikhonov choice was among the most efficient portfolios and the best estimates of risk. Moreover, the Tikhonov estimator performs relatively well in the circumstance of insufficient historical data. We believe that this parameter selection method is quite promising relative to previously proposed methods.

Acknowledgments. We are grateful for the comments of the associate editor and the referees, who provided additional references and suggestions to make the presentation clearer and the experiments more convincing.

## REFERENCES

[1] M. S. Bartlett, Tests of significance in factor analysis, Br. J. Psych. Stat. Sec., 3 (1950), pp. 77-85.
[2] C. Bengtsson and J. Holst, On portfolio selection: Improved covariance matrix estimation for Swedish asset returns, in 31st Meeting of the Euro Working Group on Financial Modeling, Agia Napa, Cyprus, 2002; also available online from http://hermes-mba.cc.ucy.ac.cy/conferences/ewgfm/papers/bengtsson_holst_20021004.pdf.
[3] M. J. Best and R. R. Grauer, On the sensitivity of mean-variance-efficient portfolios to changes in asset means: Some analytical and computational results, Rev. Financ. Stud., 4 (1991), pp. 315-342.
[4] R. B. Cattell, The scree test for the number of factors, Mult. Behav. Res., 1 (1966), pp. 245-276.
[5] L. K. C. Chan, J. Karceski, and J. Lakonishok, On portfolio optimization: Forecasting covariances and choosing the risk model, Rev. Financ. Stud., 12 (1999), pp. 937-974.
[6] V. K. Chopra, C. R. Hensel, and A. L. Turner, Massaging mean-variance inputs: Returns from alternative global investment strategies in the 1980s, Management Sci., 39 (1993), pp. 845-855.
[7] V. K. Chopra and W. T. Ziemba, The effect of errors in means, variances, and covariances on optimal portfolio choice, J. Portfolio Management, 19:2 (1993), pp. 6-11.
[8] T. Conlon, H. J. Ruskin, and M. Crane, Random matrix theory and fund of funds portfolio optimisation, Phys. A, 382 (2007), pp. 565-576.
[9] G. Connor and R. A. Korajczyk, A test for the number of factors in an approximate factor model, J. Finance, 48 (1993), pp. 1263-1291.
[10] V. Demiguel, L. Garlappi, F. J. Nogales, and R. Uppal, A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms, Management Sci., 55 (2009), pp. 798-812.
[11] V. Demiguel, L. Garlappi, and R. Uppal, Optimal versus naive diversification: How inefficient is the $1 / N$ portfolio strategy?, Rev. Financ. Stud., 22 (2009), pp. 1915-1953.
[12] V. DeMiguel and F. J. Nogales, Portfolio selection with robust estimation, Oper. Res., 57 (2009), pp. 560-577.
[13] E. J. Elton and M. J. Gruber, Estimating the dependence structure of share prices-implications for portfolio selection, J. Finance, 28 (1973), pp. 1203-1232.
[14] G. Golub and W. Kahan, Calculating the singular values and pseudo-inverse of a matrix, J. Soc. Indust. Appl. Math. Ser. B Numer. Anal., 2 (1965), pp. 205-224.
[15] G. H. Golub, M. Heath, and G. Wahba, Generalized cross-validation as a method for choosing a good ridge parameter, Technometrics, 21 (1979), pp. 215-223.
[16] L. Guttman, Some necessary conditions for common-factor analysis, Psychometrika, 19 (1954), pp. 149161.
[17] P. C. Hansen, Analysis of discrete ill-posed problems by means of the L-curve, SIAM Rev., 34 (1992), pp. 561-580.
[18] P. C. Hansen, J. G. Nagy, and D. P. O'Leary, Deblurring Images: Matrices, Spectra, and Filtering, SIAM, Philadelphia, 2006.
[19] P. C. Hansen and D. P. O'Leary, The use of the L-curve in the regularization of discrete ill-posed problems, SIAM J. Sci. Comput., 14 (1993), pp. 1487-1503.
[20] R. J. Hanson, A numerical method for solving Fredholm integral equations of the first kind using singular values, SIAM J. Numer. Anal., 8 (1971), pp. 616-622.
[21] W. Härdle and L. Simar, Applied Multivariate Statistical Analysis, Springer, Berlin, 2003.
[22] A. E. Hoerl and R. W. Kennard, Ridge regression: Applications to nonorthogonal problems, Technometrics, 12 (1970), pp. 69-82.
[23] A. E. Hoerl and R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics, 12 (1970), pp. 55-67.
[24] J. E. Jackson, A User's Guide to Principal Components, Wiley-Interscience, New York, 2003.
[25] R. Jagannathan and T. Ma, Risk reduction in large portfolios: Why imposing the wrong constraints helps, J. Finance, 58 (2003), pp. 1651-1684.
[26] B. F. King, Market and industry factors in stock price behavior, J. Business, 39 (1966), pp. 139-190.
[27] J. Kwapień, S. Drożdż, and P. Oświȩcimka, The bulk of the stock market correlation matrix is not pure noise, Phys. A, 359 (2006), pp. 589-606.
[28] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, Noise dressing of financial correlation matrices, Phys. Rev. Lett., 83 (1999), pp. 1467-1470.
[29] L. Laloux, P. Cizeau, M. Potters, and J.-P. Bouchaud, Random matrix theory and financial correlations, Internat. J. Theoret. Appl. Finance, 3 (2000), pp. 391-397.
[30] O. Ledoit and M. Wolf, Improved estimation of the covariance matrix of stock returns with an application to portfolio selection, J. Empirical Finance, 10 (2003), pp. 603-621.
[31] O. Ledoit and M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivariate Anal., 88 (2004), pp. 365-411.
[32] J. Lintner, The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets, Rev. Econom. Statist., 47 (1965), pp. 13-37.
[33] H. Markowitz, Portfolio selection, J. Finance, 7 (1952), pp. 77-91.
[34] M. L. Mehta, Random Matrices, 3rd ed., Academic Press, New York, 2004. Cited by [38].
[35] J. Mossin, Equilibrium in a capital asset market, Econometrica, 34 (1966), pp. 768-783.
[36] D. L. Phillips, A technique for the numerical solution of certain integral equations of the first kind, J. Assoc. Comput. Mach., 9 (1962), pp. 84-97.
[37] V. Plerou, P. Gopikrishnan, L. A. Nunes Amaral, M. Meyer, and H. E. Stanley, Scaling of the distribution of price fluctuations of individual companies, Phys. Rev. E, 60 (1999), pp. 6519-6529.
[38] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, T. Guhr, and H. E. Stanley, Random matrix approach to cross-correlations in financial data, Phys. Rev. E, 65 (2002), article 066126.
[39] R. Roll and S. A. Ross, An empirical investigation of the arbitrage pricing theory, J. Finance, 35 (1980), pp. 1073-1103.
[40] S. A. Ross, The arbitrage theory of capital asset pricing, J. Econom. Theory, 13 (1976), pp. 341-360.
[41] B. W. Rust, Parameter selection for constrained solutions to ill-posed problems, in Proceedings of the 32nd Symposium on the Interface: Computing Science and Statistics, 2000, pp. 333-347; also available online from http://www.galaxy.gmu.edu/interface/I00/I2000Proceedings/BRust/BRust.pdf.
[42] B. W. Rust and D. P. O'Leary, Residual periodograms for choosing regularization parameters for ill-posed problems, Inverse Problems, 24 (2008), article 034005.
[43] W. F. Sharpe, A simplified model for portfolio analysis, Management Sci., 9 (1963), pp. 277-293.
[44] W. F. Sharpe, Capital asset prices: A theory of market equilibrium under conditions of risk, J. Finance, 19 (1964), pp. 425-442.
[45] G. W. Stewart and J. Sun, Matrix Perturbation Theory, Academic Press, Boston, 1990.
[46] A. N. Tikhonov and V. Y. Arsenin, Solution of Ill-Posed Problems, John Wiley \& Sons, New York, 1977.
[47] J. L. Treynor, Toward a theory of the market value of risky assets, 1962 manuscript published as Chapter 2 in Asset Pricing and Portfolio Performance: Models, Strategy and Performance Metrics, R. A. Korayczyk, ed., Risk Books, London, 1999.
[48] C. Trzcinka, On the number of factors in the arbitrage pricing model, J. Finance, 41 (1986), pp. 347-368.
[49] S. Twomey, On the numerical solution of Fredholm integral equations of the first kind by inversion of the linear system produced by quadrature, J. Assoc. Comput. Mach., 10 (1963), pp. 97-101.
[50] W. F. Velicer, Determining the number of components from the matrix of partial correlations, Psychometrika, 41 (1976), pp. 321-327.
[51] S. Wold, Cross-validatory estimation of the number of components in factor and principal component analysis, Technometrics, 20 (1978), pp. 397-405.


[^0]:    *Received by the editors February 24, 2009; accepted for publication (in revised form) August 10, 2009; published electronically January 21, 2010.
    http://www.siam.org/journals/sifin/1/75073.html
    ${ }^{\dagger}$ Bloomberg LP and Courant Institute, New York University, New York, NY 10022 (pcarr4@bloomberg.com).
    ${ }^{\ddagger}$ Robert H. Smith School of Business, University of Maryland, College Park, MD 20742 (dbm@rhsmith.umd.edu). This author's work was partially supported by the Financial Integrity Research Network, an ARC network, during his term in Australia as a FIRN visitor.

[^1]:    *Received by the editors November 17, 2008; accepted for publication (in revised form) August 10, 2009; published electronically January 21, 2010. This work was partially supported by the National Science Foundation under grant DMS-0908461.
    http://www.siam.org/journals/sifin/1/74112.html
    ${ }^{\dagger}$ Mathematics and Statistics Department, Boston University, 111 Cummington Street, Boston, MA 02215 (kardaras@bu.edu).
    ${ }^{\ddagger}$ School of Finance and Economics \& Department of Mathematical Sciences, University of Technology, Sydney, P.O. Box 123, Broadway, NSW 2007, Australia (eckhard.platen@uts.edu.au).

[^2]:    *Received by the editors January 9, 2009; accepted for publication (in revised form) August 19, 2009; published electronically January 21, 2010. This work was supported by National Basic Research Program of China (973 Program) 2007CB814903.
    http://www.siam.org/journals/sifin/1/74623.html
    ${ }^{\dagger}$ Department of Mathematics, Tongji University, Shanghai 200092, People's Republic of China (liang_jin@tongji. edu.cn, jiang_lishang@yahoo.com.cn).
    $\ddagger$ Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556 (b1hu@nd.edu).

[^3]:    *Received by the editors December 11, 2008; accepted for publication (in revised form) August 10, 2009; published electronically January 21, 2010.
    http://www.siam.org/journals/sifin/1/74349.html
    ${ }^{\dagger}$ Department of Operations Research and Financial Engineering (ORFE), Princeton University, Princeton, NJ 08544 (ahamel@princeton.edu).
    ${ }^{\ddagger}$ Institute of Mathematics, University Halle-Wittenberg, 06099 Halle, Germany (frank.heyde@mathematik. uni-halle.de).

[^4]:    *Received by the editors December 5, 2008; accepted for publication (in revised form) October 5, 2009; published electronically January 21, 2010.
    http://www.siam.org/journals/sifin/1/74288.html
    ${ }^{\dagger}$ Department of Mathematics, National University of Singapore (NUS), Singapore (matdm@nus.edu.sg). This author's research was partially supported by Singapore MOE AcRF grant R-146-000-096-112 and NUS RMI grant R-146-000-117-720/646. This author is also an affiliated member of Risk Management Institute of NUS.
    ${ }^{\ddagger}$ Mathematical Institute and Nomura Centre for Mathematical Finance, University of Oxford, 24-29 St Giles, Oxford OX1 3LB, UK (xuz@maths.ox.ac.uk). This author's research was partially supported by the Nomura Centre for Mathematical Finance and the Oxford-Man Institute of Quantitative Finance.
    ${ }^{\text {§ }}$ Mathematical Institute and Nomura Centre for Mathematical Finance, and Oxford-Man Institute of Quantitative Finance, University of Oxford, 24-29 St Giles, Oxford OX1 3LB, UK, and Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong (zhouxy@maths.ox.ac.uk). This author's research was partially supported by the Nomura Centre for Mathematical Finance and a start-up fund of the University of Oxford.

[^5]:    ${ }^{1}$ One should note that the trading strategies derived in Li and Ng [20] and Zhou and Li [35] are not time consistent.

[^6]:    ${ }^{2}$ This smoothness also plays a crucial role in studying the finite-horizon optimal investment and consumption with transaction costs within a utility framework; see section 3 of Dai et al. [7], where an integral over one free boundary is used.

[^7]:    ${ }^{3}$ In an EUM model-the Merton problem, for example - the optimal solution is to keep the bond-stock ratio exactly at a certain value. In the MV model the ratio must be "adjusted" in order to account for the constraint of meeting the terminal target.

[^8]:    ${ }^{4}$ The no-transaction cost frontier follows from the analytical solutions in Zhou and Li [35].

[^9]:    ${ }^{5}$ Strictly speaking, it is not entirely "fair" to compare the two models, because the MV model studied here allows bankruptcy, whereas the EUM model with the usual Inada condition inherently leads to a bankruptcyprohibited solution. It would be more meaningful to develop an MV model with transaction costs and no bankruptcy and then to compare the solution with its EUM counterpart.
    ${ }^{6}$ In Liu and Loewenstein [25] and Dai and Yi [8] only the case when $y>0$ is discussed, although we suspect that the methodology in our paper could be adapted to deal with the $y<0$ case in the EUM setting.

[^10]:    *Received by the editors January 2, 2009; accepted for publication (in revised form) November 4, 2009; published electronically February 3, 2010.
    http://www.siam.org/journals/sifin/1/74546.html
    ${ }^{\dagger}$ Department of Mathematics, University of Kansas, Lawrence, KS 66045 (jfeng@math.ku.edu). This author's research was partially supported by National Science Foundation grant DMS-0806434.
    ${ }^{\ddagger}$ Department of Mathematical Sciences, Dublin City University, Glasnevin, Dublin 9, Ireland (Martin.Forde@dcu. ie).
    ${ }^{\S}$ Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93016 (fouque@ pstat.ucsb.edu). This author's research was partially supported by National Science Foundation grant DMS-0806461.

[^11]:    *Received by the editors February 23, 2009; accepted for publication (in revised form) November 5, 2009; published electronically February 3, 2010. This research was supported by the Natural Sciences and Engineering Research Council of Canada.
    http://www.siam.org/journals/sifin/1/75042.html
    ${ }^{\dagger}$ Department of Mathematics and Statistics, McMaster University, Hamilton, ON, L8S 4K1, Canada (hurdt@ momaster.ca, zhouz3@univmail.cis.mcmaster.ca).

[^12]:    ${ }^{1}$ It came to our attention after the submission of our paper that the result of this theorem has been simultaneously and independently stated in another working paper by Antonov and Arneguy [1].
    ${ }^{2}$ Here and in rest of the paper, some variables such as $u, \epsilon, x$ are defined to be row vectors with components $u=\left(u_{1}, u_{2}\right)$, etc. We use implied matrix multiplication so that $u x^{\prime}=u_{1} x_{1}+u_{2} x_{2}$, where $x^{\prime}$ denotes the (unconjugated) transpose of $x$.

[^13]:    ${ }^{3}$ According to these authors, computing the gamma function becomes "not much more difficult than other built-in functions that we take for granted, such as $\sin x$ or $e^{x}$."

[^14]:    ${ }^{4}$ We thank a referee for this suggestion.

[^15]:    *Received by the editors March 30, 2009; accepted for publication (in revised form) November 5, 2009; published electronically February 17, 2010.
    http://www.siam.org/journals/sifin/1/75418.html
    ${ }^{\dagger}$ Department of Mathematics and Systems Analysis, Helsinki University of Technology, P.O. Box 1100, FI-02015 TKK, Espoo, Finland (teemu.pennanen@tkk.fi).
    ${ }^{\ddagger}$ Humboldt-Universität zu Berlin, Institut für Mathematik, Unter den Linden 6, 10099 Berlin, Germany (penner@ math.hu-berlin.de). This author's research was supported by the DFG Research Center Matheon "Mathematics for key technologies."

[^16]:    *Received by the editors November 3, 2008; accepted for publication (in revised form) November 9, 2009; published electronically February 17, 2010. This research has been supported by grants from both the Spanish and the Japanese governments.
    http://www.siam.org/journals/sifin/1/73976.html
    ${ }^{\dagger}$ Osaka University, Machikaneyama cho 1-3, Osaka 560-8531, Japan, and Japan Science and Technology Agency, Honcho, Kawaguchi-shi, Saitama 332-0012, Japan (kohatsu@sigmath.es.osaka-u.ac.jp).
    ${ }^{\ddagger}$ Facultat de Ciències UAB, Centre de Recerca Matemàtica, Bellaterra 08193, Barcelona, Spain (sortiz@crm.cat).

[^17]:    *Received by the editors February 24, 2009; accepted for publication (in revised form) November 18, 2009; published electronically February 19, 2010.
    http://www.siam.org/journals/sifin/1/75067.html
    ${ }^{\dagger}$ Mathematics Department, Indiana University, Bloomington, IN 47405 (goodmanv@indiana.edu).
    ${ }^{\ddagger}$ Mathematics Department, Florida State University, Tallahassee, FL 32306-4510 (kim@math.fsu.edu).

[^18]:    *Received by the editors January 29, 2009; accepted for publication (in revised form) December 6, 2009; published electronically February 19, 2010. This work was partially supported by the Italian M.I.U.R. project "Viscosity, metric, and control theoretic methods for nonlinear partial differential equations."
    http://www.siam.org/journals/sifin/1/74814.html
    ${ }^{\dagger}$ Dipartimento di Matematica P. e A., Università di Padova, via Trieste 63, 35121 Padova, Italy (bardi@math. unipd.it, acesar@math.unipd.it).
    ${ }^{\ddagger}$ Institut de Mathématiques de Toulon, Université du Sud Toulon-Var, Bâtiment U, BP 20132, 83957 La Garde Cedex, France (luigi.manca@univ-tln.fr). Most of this work was done while the author was holding a postdoc position at the University of Padua.

[^19]:    *Received by the editors November 3, 2008; accepted for publication (in revised form) December 9, 2009; published electronically March 24, 2010. This work was partially supported by the National Science Foundation (NSF grants DMS-0091946, NSF-RTG-0636586, DMS-FRG-0456118, and DMS-0706947).
    http://www.siam.org/journals/sifin/1/73973.html
    ${ }^{\dagger}$ Departments of Mathematics and Information, Risk and Operations Management, The University of Texas at Austin, 1 University Station, C1200, Austin, TX 78712 (zariphop@math.utexas.edu).
    ${ }^{\ddagger}$ Department of Mathematics, The University of Texas at Austin, 1 University Station, C1200, Austin, TX 78712 (gordanz@math.utexas.edu).

[^20]:    ${ }^{1}$ One could object to the above reasoning by pointing out that different maturities should give rise to different risk assessments due to the effect of time impatience. In response, we take the view that the market is efficient in the sense that all time impatience is already incorporated in the investment possibilities present in it. More specifically, we remind the reader that the assumption that $S^{0} \equiv 1$ effectively means that all contingent claims are quoted in terms of time-0 currency. As pointed out in Remark 2.1, one can easily extend the theory presented here to the more general case where the time value of money is modeled explicitly. We feel, however, that such a generalization would only obscure the central issue herein and render the present paper less accessible.

[^21]:    *Received by the editors March 25, 2009; accepted for publication (in revised form) February 1, 2010; published electronically April 21, 2010.
    http://www.siam.org/journals/sifin/1/75381.html
    ${ }^{\dagger}$ Pricing Partners, 6 rue Rougemont 75009 Paris, France (eric.benhamou@pricingpartners.com, mohammed.miri@ pricingpartners.com).
    ${ }^{\ddagger}$ Corresponding author. Laboratoire Jean Kuntzmann, Université de Grenoble and CNRS, BP 53, 38041 Grenoble cedex 9, France (emmanuel.gobet@imag.fr). This author's research was partially supported by University Joseph Fourier (MSTIC grant REFINE).
    ${ }^{1}$ Nice properties for the CIR process are derived by Dufresne [15], Göing-Jaeschke and Yor [20], Diop [14], Alfonsi [1], and Miri [32].

[^22]:    ${ }^{2}$ The approximation formula for the call price is obtained using the call/put parity relation: in (2.13), it consists of replacing on the left-hand side the put payoff by the call one, and on the right-hand side the put price function $P_{B S}$ by the similar call price function, while the coefficients remain the same.

[^23]:    ${ }^{3}$ In this approach, we leave the initial value $\bar{v}_{0}$ equal to $v_{0}$. Indeed, it is not natural to modify its value since it is not a parameter but rather an unobserved factor.

[^24]:    ${ }^{4}$ Error price bp $=\frac{\text { Price Approximation }- \text { True Price }}{\text { Spot }} \times 10000$.

[^25]:    ${ }^{5}$ For higher values of $\xi$, we compute the price taking in consideration the no-arbitrage interval.

[^26]:    ${ }^{6}$ Note that from the upper bound (4.11) in the proof we easily obtain that the continuity also holds a.s. and not only in $L_{p}$. Since only the latter is needed in what follows, we do not go into detail.

[^27]:    *Received by the editors January 5, 2009; accepted for publication (in revised form) November 15, 2009; published electronically May 26, 2010. This work was presented at the 4th World Congress of BFS (2006) and at meetings and workshops at, among others, U.C. Santa Barbara, U.M. College Park, Columbia University, MIT (2006), Oberwolfach, Stanford University, AMaMeF, and QMF (2007), and Oberwolfach, University of Kostanz, and RICAM (2008). The first version of this paper (September 2007) was titled "Implied preferences and bespoke portfolios."
    http://www.siam.org/journals/sifin/1/74525.html
    ${ }^{\dagger}$ BNP Paribas, NW1 6AA London, UK (marek.musiela@bnpparibas.com).
    ${ }^{\ddagger}$ Oxford-Man Institute and Mathematical Institute, University of Oxford, OX2 6ED Oxford, UK, and Department of Mathematics and Department of IROM, McCombs School of Business, The University of Texas at Austin, Austin, TX 78712 (Thaleia.Zariphopoulou@oxford-man.ox.ac.uk). The work of this author was partially supported by National Science Foundation grants DMS-FRG-0456118 and DMS-RTG-0636586.

[^28]:    ${ }^{1}$ See [13] and [7].

[^29]:    ${ }^{2}$ While revising this work, the authors came across the revised version of [1], where similar questions are studied for the nonnegative wealth case.

[^30]:    ${ }^{3}$ See, for example, the case $r(x, t)=\sqrt{a x^{2}+b e^{-a t}}$ analyzed in [18].

[^31]:    ${ }^{4}$ We remind the reader that this condition is sufficient for the finiteness of $h$ and $u$ but does not, in general, guarantee admissibility of the associated policies. For the latter, condition (22) is used.

[^32]:    ${ }^{5}$ One may alternatively represent $h$ as $h(x, t)=\int_{0}^{+\infty} e^{y x-\frac{1}{2} y^{2} t} \nu^{\prime}(d y)$ with $\nu^{\prime}(d y)=\frac{\nu(d y)}{y}$. Note that $\nu^{\prime} \in \mathcal{B}^{+}(\mathbb{R})$. Such a representation was used in [1].
    ${ }^{6}$ The authors would like to thank an anonymous referee for pointing out that this integrability condition is needed.

[^33]:    *Received by the editors December 1, 2008; accepted for publication (in revised form) March 8, 2010; published electronically May 26, 2010. This research was part of the Chair Financial Risks of the Risk Foundation sponsored by Société Générale, the Chair Derivatives of the Future sponsored by the Fédération Bancaire Française, and the Chair Finance and Sustainable Development sponsored by EDF and Calyon.
    http://www.siam.org/journals/sifin/1/74217.html
    ${ }^{\dagger}$ CEREMADE, Université Paris Dauphine, 75775 Cedex 16, Paris, France (imen@ceremade.dauphine.fr).
    ${ }^{\ddagger}$ Sabanci University, Istanbul 34956, Turkey (hmsoner@ethz.ch). This author's research was partly supported by the King Saud University D.S.F.P and by the Turkish Technical and Scientific Council, TÜBITAK.
    ${ }^{\text {§ }}$ CMAP, Ecole Polytechnique, 91128 Palaiseau Cedex, Paris, France (nizar.touzi@polytechnique.edu).

[^34]:    *Received by the editors March 30, 2009; accepted for publication (in revised form) March 8, 2010; published electronically May 26, 2010. This research was supported by the Swiss National Science Foundation under grant 200021-117606.
    http://www.siam.org/journals/sifin/1/75419.html
    ${ }^{\dagger}$ Department of Mathematical Statistics and Actuarial Science, University of Bern, Sidlerstrasse 5, 3012 Bern, Switzerland (ilya.molchanov@stat.unibe.ch, michael.schmutz@stat.unibe.ch).

[^35]:    *Received by the editors July 10, 2009; accepted for publication (in revised form) March 31, 2010; published electronically June 3, 2010.
    http://www.siam.org/journals/sifin/1/76457.html
    ${ }^{\dagger}$ CREST and ENSAE Paris Tech, Timbre J120, 3 Avenue Pierre Larousse, 92245 Malakoff Cedex, France (chrobert@ensae.fr).
    ${ }^{\ddagger}$ CMAP-École Polytechnique Paris, UMR CNRS 7641, Route de Saclay, 91128 Palaiseau Cedex, France (mathieu.rosenbaum@polytechnique.edu).

[^36]:    ${ }^{1}$ We do not assume that this theoretical strategy exactly replicates the payoff at maturity.
    ${ }^{2}$ Usual microstructure noise models mostly consider deterministic, exogenous sampling schemes and rarely allow for price discreteness; see, for example, [1], [13], [18], [27], and [28].

[^37]:    ${ }^{3}$ This is a simplifying, slightly incorrect framework since we should consider a payoff $F\left(P_{T}\right)$. However, the order of magnitude of the difference is clearly negligible in our context.

[^38]:    ${ }^{4}$ The agent is supposed to be a price follower, which means that the transaction price never moves because of its own trading.

[^39]:    ${ }^{5}$ Note that we could also express our results in terms of variation of the local volatility delta in the same way.

[^40]:    ${ }^{6}$ Note that a few typo errors seem to appear in the mentioned version of [14].

[^41]:    *Received by the editors October 1, 2009; accepted for publication (in revised form) April 18, 2010; published electronically July 1, 2010. This work was supported by the International Graduate School of Science and Engineering (IGSSE) of Technische Universität München.
    http://www.siam.org/journals/sifin/1/77271.html
    ${ }^{\dagger}$ Zentrum Mathematik, Technische Universität München, 85748 Garching bei München, Germany (hepperger@ ma.tum.de).

[^42]:    *Received by the editors June 23, 2009; accepted for publication (in revised form) April 18, 2010; published electronically July 1, 2010. This research was started at Cornell University and was supported by the School of Operations Research and Information Engineering.
    http://www.siam.org/journals/sifin/1/76278.html
    ${ }^{\dagger}$ Université Paris-Est, CERMICS, Project team MathFi ENPC-INRIA-UMLV, Ecole des Ponts, 6-8 avenue Blaise Pascal, 77455 Marne La Vallée, France (alfonsi@cermics.enpc.fr). This author's research was supported by the chair "Risques Financiers" of Fondation du Risque and the Credinext project from Finance Innovation.
    $\ddagger$ Institute of Mathematics, University of Mannheim, A 5, 6, 68159 Mannheim, Germany (schied@uni-mannheim. de).

[^43]:    ${ }^{1}$ This technique was already used in $[1,2]$.

[^44]:    ${ }^{2}$ These models are referred to as Models 1 and 2, respectively, in [1, 2].

[^45]:    *Received by the editors May 11, 2009; accepted for publication (in revised form) April 19, 2010; published electronically July 1, 2010.
    http://www.siam.org/journals/sifin/1/75859.html
    ${ }^{\dagger}$ Swiss Finance Institute, Ecole Polytechnique Fédérale de Lausanne, Quartier UNIL-Dorigny, Extranef 218, CH1015 Lausanne, Switzerland (damir.filipovic@epfl.ch). This author's research was supported by the WWTF (Vienna Science and Technology Fund) and Swissquote.
    ${ }^{\ddagger}$ Department of Mathematics, ETH Zürich, Rämistrasse 101, CH-8092 Zürich, Switzerland (stefan.tappe@math. ethz.ch, josef.teichmann@math.ethz.ch). The second author's research was supported by the WWTF. The third author's research was supported by FWF-grant Y 328 (START prize from the Austrian Science Fund).

[^46]:    *Received by the editors August 13, 2009; accepted for publication (in revised form) April 20, 2010; published electronically July 8, 2010.
    http://www.siam.org/journals/sifin/1/76800.html
    ${ }^{\dagger}$ IEOR Department, Columbia University, New York, NY 10027 (Rama.Cont@columbia.edu), and Laboratoire de Probabilités et Modèles Aléatoires, CNRS - Université Pierre et Marie Curie, Paris, France.
    ${ }^{\ddagger}$ IEOR Department, Columbia University, New York, NY 10027 (rd2304@columbia.edu, yk2246@columbia.edu).

[^47]:    ${ }^{1}$ Expression (9) is slightly different for the most senior tranche. See Appendix B for details.
    ${ }^{2}$ It is natural to assume that the number of tranches $I$ is fewer than the number of names $n$ in the reference portfolio.
    ${ }^{3}$ Our formulation can be easily extended to CDO tranches with multiple expirations, but it will not be further discussed in this paper.

[^48]:    *Received by the editors June 2, 2009; accepted for publication (in revised form) April 22, 2010; published electronically July 8, 2010.
    http://www.siam.org/journals/sifin/1/76080.html
    ${ }^{\dagger}$ Courant Institute of Mathematical Sciences, New York University, New York, NY 10012 and Finance Concepts, 490 Madison Avenue, New York, NY, 10022 (avellane@cims.nyu.edu, stanley@nyu.edu).
    ${ }^{1}$ To our knowledge, the first issuer of leveraged ETFs was Rydex in late 2006.
    ${ }^{2}$ The description of the hedging mechanism given here is not intended to be exact, but rather to illustrate the general approach used by ETF managers to achieve the targeted leveraged long and short exposures. For instance, managers can trade the stocks that compose the ETFs or indices, or enter into total-return swaps to synthetically replicate the returns of the index that they track. The fact that the returns are adjusted daily is important for our discussion. Recently Direxion Funds, an LETF manager, has announced the launch of products with monthly rebalancing.

[^49]:    ${ }^{3}$ Interestingly, similar considerations about volatility exposure apply to a class of over-the-counter (OTC) products known as Constant Proportion Portfolio Insurance (CPPI). These products are typically embedded in structured notes and other derivatives marketed by insurance companies and banks (see, for instance, Black and Perold [3] and Betrand and Prigent [2]). Although not directly relevant to this study, the parallel between the behavior of listed LETFs and OTC products may be of independent interest.

[^50]:    ${ }^{4}$ In the sense that this does not account for the costs of financing positions and management fees.
    ${ }^{5}$ We emphasize the cost of borrowing, since we are interested in LETFs which track financial indices. The latter have often been hard-to-borrow since July 2008. Moreover, broad market ETFs such as SPY have also been sporadically hard-to-borrow in the last quarter of 2008; see Avellaneda and Lipkin [1].

[^51]:    ${ }^{6}$ They are not assumed to be deterministic functions or constants.

[^52]:    ${ }^{7}$ For information about ProShares, see http://www.proshares.com.

[^53]:    ${ }^{8}$ See http://www.direxionfunds.com.

[^54]:    ${ }^{9}$ For investors subject to Regulation T margin requirements, the maximum allowed leverage for equities is 2. The use of a double-leveraged ETF would produce, accordingly, a leverage of 4 vis à vis the reference index. However, effective December 1, 2009, FINRA has introduced new margin requirements on LETFs to avoid such overleveraging by retail investors [9].

[^55]:    *Received by the editors September 25, 2009; accepted for publication (in revised form) May 1, 2010; published electronically July 15, 2010.
    http://www.siam.org/journals/sifin/1/77219.html
    ${ }^{\dagger}$ Statistical Laboratory, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK (L.C.G.Rogers@ statslab.cam.ac.uk).
    ${ }^{1}$ While we shall discuss only discrete-time problems in this paper, the methodology could be applied to the pricing of American options after discretizing the time.
    ${ }^{2}$ Some minor regularity condition needs to be imposed on $Z$; it is sufficient that $\sup _{0 \leq t \leq T}\left|Z_{t}\right| \in L^{p}$ for some $p>1$.

[^56]:    ${ }^{3}$ We know that $M^{(1)}$ has this form, and from (2.9) and the Markovian structure it is obvious by induction that this form persists.

[^57]:    *Received by the editors June 22, 2009; accepted for publication (in revised form) May 23, 2010; published electronically August 17, 2010.
    http://www.siam.org/journals/sifin/1/76271.html
    ${ }^{\dagger}$ Department of Mathematics, Ohio University, Athens, OH 45701 (guli@math.ohiou.edu).

[^58]:    ${ }^{1}$ Formula (82) was recently established for the stock price density in the correlated Heston model, and a better error estimate $O\left((\log x)^{-1 / 2}\right)$ was obtained in this formula (see [16]). Moreover, using the methods developed in the present paper, formula (105) (see section 7) was extended to the implied volatility in the correlated Heston model.

[^59]:    *Received by the editors September 16, 2009; accepted for publication (in revised form) June 12, 2010; published electronically September 16, 2010. This paper was previously circulated under the title "Pricing Credit From the Top-Down With Affine Point Processes."
    http://www.siam.org/journals/sifin/1/77127.html
    ${ }^{\dagger}$ CreditFlow Partners, 345 86th Street, Suite 12D, New York, NY 10028 (Eymen.Errais@thecreditflow.com).
    ${ }^{\ddagger}$ Corresponding author. Department of Management Science \& Engineering, Stanford University, Stanford, CA 94305-4026 (giesecke@stanford.edu, http://www.stanford.edu/~giesecke).
    ${ }^{\S} \mathrm{MSCI}$ Barra, 2100 Milvia St., Berkeley, CA 94704 (Lisa.Goldberg@mscibarra.com).

[^60]:    ${ }^{1}$ Exact valuation results can always be obtained at an additional computational expense. To this end, we replace $N$ by $N^{n}=N \wedge n$ and $L$ by $L^{n}=\sum_{k=1}^{N^{n}} \ell_{k}$ in the formulae. The distribution of the process $N^{n}$, which has intensity $\lambda 1_{\{N<n\}}$, is easily calculated from the distribution of $N$ : we have $P\left(N_{s}^{n}=k\right)=P\left(N_{s}=k\right)$ for all $k<n$ and $P\left(N_{s}^{n}=n\right)=P\left(N_{s} \geq n\right)$. The calculation of the distribution of $L^{n}$ requires further steps. If, for example, the losses $\ell_{k}$ are independent and identically distributed (i.i.d.) and independent of $N^{n}$, then this amounts to a convolution operation.

[^61]:    ${ }^{2}$ See section 3.3 for details on the method of computation.

[^62]:    ${ }^{3}$ That is, calibration required parameter values outside the set of admissible model parameter values. The copula model called for a correlation coefficient larger than 1 , which is not meaningful.

[^63]:    ${ }^{4}$ The tensor product $u \otimes v$ of vectors $u$ and $v$ is a matrix whose $i j$ th element is $u_{i} v_{j}$.

[^64]:    *Received by the editors August 13, 2009; accepted for publication (in revised form) July 6, 2010; published electronically September 29, 2010.
    http://www.siam.org/journals/sifin/1/76806.html
    ${ }^{\dagger}$ School of Management, University of Texas at Dallas, P.O. Box 830688, SM 30, Richardson, TX 75083-0688 (alain.bensoussan@utdallas.edu), and Graduate School of Business, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.
    ${ }^{\ddagger}$ Department of Finance and Real Estate, The University of Texas at Arlington, UTA Box 19449, Arlington, TX 76019 (diltz@uta.edu).
    ${ }^{\text {§ }}$ School of Management, University of Texas at Dallas, P.O. Box 830688, SM 30, Richardson, TX 75083-0688 (celinehoe@utdallas.edu).

[^65]:    ${ }^{1}$ See, for example, Dixit and Pindyck [10, Chap. 9], Grenadier [15], [16], [17], Smets [46], Lambrecht and Perraudin [32], Huisman [28], Smit and Trigeorgis [48], Smit and Ankum [47], etc.

[^66]:    ${ }^{2}$ While this statement holds in complete markets, the relationship between the two models is in general an open question.

[^67]:    ${ }^{3}$ This specification is for preparing the situation of two players. a represents market share by other active participants. For the single decision maker case, $a$ is zero. In the Stackelberg games that follow, $a$ represents the leader's market share after the follower's entry into the market. In the context of the leader-follower model, we assume $a \geq \frac{1}{2}$. Otherwise, the leader will surrender a majority interest in the project, providing a strong disincentive to enter at all.

[^68]:    ${ }^{4}$ It is straightforward that the uniqueness and existence of the solution, $U(v)$, guarantees the uniqueness and existence of the solution, $L(v)=U(v)+a v-K$.
    ${ }^{5} \varrho>\frac{3}{2}$.

[^69]:    ${ }^{6}$ Replace $\tilde{\Phi}$ with $\Phi$ in (3.25).

[^70]:    ${ }^{7}$ This formulation is similar to the specification that the investment cost grows at the risk-free rate.

[^71]:    ${ }^{8}$ Henderson [22] formulated the same investment problem, recognizing that this is a nonstandard situation. Wealth must be evaluated at a finite intermediate time. She also showed how this formulation eliminates bias that might influence the manager's choice of exercise/investment time in the infinite horizon setting. She also noted that the choice of $\frac{\lambda^{2}}{2}$ is a modeling choice, and it is not essential to solve the model in closed form. It is essential for unbiased investment timing.

[^72]:    ${ }^{9}$ By assuming investment cost grows at the risk-free rate in the complete market, we can see directly that when $|\rho| \rightarrow 1$ or $\gamma \rightarrow 0$ the optimal investment rule converges to the complete market/risk-neutral valuation case.

[^73]:    ${ }^{10}$ Note that this is a different $V(t)$ process from the original model. This $V(t)$ process arises from our redefined control problem $\Sigma(v)$. This is the process for which we define the optimal stopping rule for $\hat{\tau}_{\Sigma}$, not the original $V(t)$ process.

[^74]:    ${ }^{11}$ The condition of $\delta_{2}<\delta_{1}$ expresses the idea of the first-mover advantages; otherwise, the leader would have no desire to enter into the market prior to the follower.

[^75]:    ${ }^{12}$ The unique measure makes the market asset price a martingale while preserving the conditional of the nontraded factor.

[^76]:    *Received by the editors January 14, 2009; accepted for publication (in revised form) July 23, 2010; published electronically October 12, 2010. The first part of this research was supported by the Institute for Operations Research, ETH Zürich.
    http://www.siam.org/journals/sifin/1/74658.html
    ${ }^{\dagger}$ Department of Mathematics, National University of Singapore, 2 Science Drive, 117543 Singapore (mathj@nus. edu.sg).
    ${ }^{\ddagger}$ London School of Economics, Centre for the Analysis of Time Series, Houghton Street, London WC2A 2AE, UK (M.W.Fehr@lse.ac.uk). This author's research was partially supported by the Munich Re Programme within the Centre for Climate Change Economics and Policy, London School of Economics.

[^77]:    *Received by the editors June 15, 2009; accepted for publication (in revised form) July 3, 2010; published electronically October 21, 2010. This work was supported by the DFG reseach center Matheon "Mathematics for key technologies" in Berlin.
    http://www.siam.org/journals/sifin/1/75402.html
    ${ }^{\dagger}$ Corresponding author. Institute of Computational Science, University of Lugano, Via Giuseppe Buffi 13, CH6904 Lugano, Switzerland (lars.putzig@usi.ch).
    ${ }^{\ddagger}$ Institute of Mathematics, Humboldt-University Berlin, Rudower Chaussee 25, 12489 Berlin, Germany (becherer@ mathematik.hu-berlin.de).
    ${ }^{\S}$ Institute of Computational Science, University of Lugano, Via Giuseppe Buffi 13, CH-6904 Lugano, Switzerland (illia.horenko@usi.ch).

[^78]:    ${ }^{1}$ If the $\Theta_{i}$ were known or fixed to some value, the minimization problem (2.4) would be solved analytically w.r.t. $\Gamma$ by

    $$
    \gamma_{i}(t)=\left\{\begin{array}{lr}
    1, & i=\arg \min g\left(S_{t}, \Theta_{i}\right) \\
    0 & \text { else }
    \end{array}\right.
    $$

[^79]:    *Received by the editors September 9, 2009; accepted for publication (in revised form) August 13, 2010; published electronically October 21, 2010.
    http://www.siam.org/journals/sifin/1/77055.html
    ${ }^{\dagger}$ Department of Mathematics, National University of Singapore (NUS), 2, Science Drive 2, Singapore 117543, Singapore (matdm@nus.edu.sg). This author is also an affiliated member of Risk Management Institute, NUS. This author's research was supported by Singapore MOE AcRF grant R-146-000-096-112 and NUS RMI grant R-146-000-117/124-720/646.
    $\ddagger$ Department of Mathematics, University of Georgia, Athens, GA 30602 (qingz@math.uga.edu).
    ${ }^{\text {§ }}$ Department of Mathematics, Western Michigan University, Kalamazoo, MI 49008 (zhu@wmich.edu).

[^80]:    ${ }^{1}$ Intuitively, one should never buy stock if $\rho \geq \mu_{1}$ and never sell stock if $\rho \leq \mu_{2}$, which coincides with Lemma 2.4.

[^81]:    ${ }^{2}$ After finishing this paper, we found that the connection between a system of variational inequalities and a double obstacle problem was first revealed in Nakoulima [23].

[^82]:    ${ }^{3}$ On $p=0$ or 1 , no boundary conditions are required due to the degeneracy of the differential operator.

[^83]:    ${ }^{4}$ A rigorous proof needs the use of a penalization method (cf. [12]).

[^84]:    *Received by the editors November 3, 2009; accepted for publication (in revised form) August 13, 2010; published electronically October 21, 2010. This work was supported in part by the Deutsche Forschungsgemeinschaft through SFB 649 "Economic Risk" and DFG Research Center Matheon "Mathematics for Key Technologies" in Berlin.
    http://www.siam.org/journals/sifin/1/77584.html
    ${ }^{\dagger}$ Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany (kraetsch@ wias-berlin.de).
    ${ }^{\ddagger}$ Corresponding author. Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany (schoenma@wias-berlin.de).

[^85]:    *Received by the editors December 1, 2008; accepted for publication (in revised form) September 14, 2010; published electronically November 11, 2010.
    http://www.siam.org/journals/sifin/1/74233.html
    ${ }^{\dagger}$ Dept. of Finance \& Dept. of Decision Sciences, Bocconi University, Via Röntgen 1, 20136 Milan, Italy (Francesco.Corielli@uni-bocconi.it).
    ${ }^{\ddagger}$ Dipartimento di Scienze Statistiche "Paolo Fortunati," Università di Bologna, Via Belle Arti, 41 Bologna, Italy (paolo.foschi2@unibo.it).
     (pascucci@dm.unibo.it).

[^86]:    ${ }^{1}$ A comprehensive presentation of the classic parametrix method for uniformly parabolic PDEs can be found, for instance, in [17]. We also quote the papers [11] and [26], where the parametrix method is applied to a wider class of (possibly degenerate) equations that includes the pricing PDEs for Asian options. While well known in the classical theory of parabolic PDEs, the parametrix series is, as far as we know, rather unknown in the field of mathematical finance with the exception of [4] and of a quotation in [1].

[^87]:    ${ }^{2}$ Equivalently we may use $m:=\inf _{z \in \mathbb{R}^{N+1}} \mu(z)$ and $M:=\sup _{z \in \mathbb{R}^{N+1}} \mu(z)$, where $\mu(z)$ is the Euclidean norm of $A(z)$ in $\mathbb{R}^{N+1} \times \mathbb{R}^{N+1}$ (and also equal to the Euclidean norm of $\left(\lambda_{1}(z), \ldots, \lambda_{N}(z)\right)$, the vector of the eigenvalues of $A(z))$. This gives less precise, but more easily computable, estimates.

[^88]:    ${ }^{3}$ Recall that

    $$
    \frac{\Gamma_{E}\left(\frac{1}{2}\right)^{k}}{\Gamma_{E}\left(\frac{k}{2}\right)}=\frac{(2 \pi)^{\frac{k-1}{2}}}{(k-2)!!},
    $$

    where $n!!$ is the double factorial defined by $n!!=2 \cdot 4 \cdot 6 \cdots n$ if $n$ is even and $n!!=1 \cdot 3 \cdot 5 \cdots n$ if $n$ is odd.

[^89]:    *Received by the editors November 23, 2009; accepted for publication (in revised form) September 27, 2010; published electronically November 30, 2010. This work was supported by a Mizuho-DL Financial Technology grant. This paper was presented at seminars at the Chinese University of Hong Kong, Columbia University, Hong Kong University of Science and Technology, Mizuho-DL Financial Technology, the 2009 ICBI Global Derivatives and Risk Management Conference, and the 15th INFORMS Applied Probability Society Conference.
    http://www.siam.org/journals/sifin/1/77805.html
    ${ }^{\dagger}$ Department of Management Science \& Engineering, Stanford University, Stanford, CA 94305-4026 (giesecke@ stanford.edu, mousavi@stanford.edu, hossein@stanford.edu).
    $\ddagger$ Mizuho-DL Financial Technology, Tokyo (hideyuki-takada@fintec.co.jp).

[^90]:    ${ }^{1}$ Other specifications of the weight function can be envisioned. The formulation we adopted from [21] has computational (memory) advantages: it does not require us to keep track of the full path history of each particle, since only the most recent transitions are relevant for the selection.
    ${ }^{2}$ Note that $m$, unlike $\delta$, has a direct impact on the memory requirement for the $\mathrm{S} / \mathrm{M}$ scheme, which is $2 R n+m$ for our weight function specification.

[^91]:    ${ }^{3}$ Carmona and Crepey [10] discuss this analogy further.
    ${ }^{4}$ See [3] for an excellent discussion.
    ${ }^{5}$ However, see [43], [44] and the references in these papers for important results in this direction. See [5] for an optimal IS scheme for doubly stochastic intensity models $\lambda$.
    ${ }^{6}$ Saying that $N$ is doubly stochastic means that the intensity $\lambda$ is a function of an adapted process $Z$ and that, given a path of $Z$, the components $N^{i}$ are independent inhomogeneous Poisson processes, each stopped at its first jump time and having (conditionally deterministic) intensity $\lambda^{i}$.

[^92]:    ${ }^{7}$ Saying that $N$ is doubly stochastic relative to a complete information filtration $\mathbb{H}$ means that the doubly stochastic property holds if one artificially enlarges the reference observation filtration $\mathbb{F}$ to make all risk factor processes adapted. Note that $N$ is not $\mathbb{F}$-doubly stochastic.

[^93]:    ${ }^{8}$ It is unclear how to allocate the computational budget of the time-scaling method between the number of time steps and the number of replications. The square-root rule is adopted from [9] and others. It is motivated by the results in [26], which show that for first order methods it is asymptotically optimal to increase the number of time steps in a manner proportional to the square root of the number of replications. However, the optimal constant of proportionality is not known.

[^94]:    ${ }^{9}$ In the case of Algorithm 4.2, the sample variance is a biased estimator of the variance. This is because the samples are not independent due to the selections performed in the scheme.

[^95]:    *Received by the editors July 2, 2009; accepted for publication (in revised form) September 1, 2010; published electronically December 7, 2010.
    http://www.siam.org/journals/sifin/1/76372.html
    ${ }^{\dagger}$ Laboratoire de Probabilités et Modèles Aléatoires CNRS, UMR 7599, Université Paris 7, 75205 Paris Cedex 13, France, and CREST, 92245 Malakoff Cedex, France (kharroubi@ensae.fr). This author's research was partly supported by the European Research Council under grant 228053-FiRM.
    ₹Laboratoire de Probabilités et Modèles Aléatoires CNRS, UMR 7599, Université Paris 7, 75205 Paris Cedex 13, France, CREST, 92245 Malakoff Cedex, France, and Institut Universitaire de France, 75005 Paris, France (pham@math.jussieu.fr).

[^96]:    *Received by the editors February 11, 2009; accepted for publication (in revised form) September 27, 2010; published electronically December 14, 2010.
    http://www.siam.org/journals/sifin/1/74937.html
    ${ }^{\dagger}$ Department of Computer Science, University of Maryland, College Park, MD 20742 (swpark81@gmail.com).
    ${ }^{\ddagger}$ Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 (oleary@cs.umd.edu). This author's work was supported in part by NSF CCF 0514213.

[^97]:    ${ }^{1}$ In this paper, the term PCA always refers to applying PCA to the matrix $\boldsymbol{R}$ of sample stock returns. For convergence properties of the sample PCA toward its population PCA, refer to [24, Chapter 4].

[^98]:    ${ }^{2}$ We omit the case of $T_{o}=1$ month since it gives us a trivial result that the portfolio returns are equal to the required portfolio return $q$ making $\left(\sigma_{i}\right)_{\tilde{\Sigma}}$ zero for any covariance $\widetilde{\boldsymbol{\Sigma}}$ and any window size $T_{w}$. This is because $\boldsymbol{\mu}$ equals the realized stock returns $\boldsymbol{r}(t)$ in the out-of-sample period.

