SIAM J. FINANCIAL MATH.
Vol. 2, pp. 1–21

Good Deal Bounds Induced by Shortfall Risk*

Takuji Arai†

Abstract. We consider, throughout this paper, an incomplete financial market which is governed by a possibly nonlocally bounded right-continuous with left-limits (RCLL) special semimartingale. We shall provide good deal bounds for contingent claims induced by shortfall risk in the framework of the Orlicz heart setting. We prove that the upper and lower bounds of such a good deal bound are expressed by a convex risk measure on an Orlicz heart. In addition, we obtain representation results for three types of model, which are an unconstrained portfolio model, a W-admissible model, and a predictably convex model.

Key words. good deal bound, shortfall, convex risk measure, Orlicz space, predictably convex

AMS subject classifications. 91B28, 46N10, 60H05, 91B30

DOI. 10.1137/090769120

1. Introduction.

There exist many no-arbitrage prices for a contingent claim in an incomplete market, since there exist many equivalent martingale measures. Thus, the no-arbitrage principle provides merely a pricing bound. However, this bound is too wide to be useful from a practical point of view. It is then meaningful to obtain a sharper pricing bound. Good deal bounds are strong candidates for such a sharper pricing bound.

The study of good deal bounds has been undertaken by Cochrane and Saá-Requejo [11]. A claim (a payoff at the maturity) is called a good deal if a suitable index related to the claim is greater (or less) than a certain level at which an investor makes up her mind. Cochrane and Saá-Requejo, in the above paper, chose the Sharpe ratio as a criterion of good deals. Moreover, Bernardo and Ledoit [2] introduced the notion of good deals in terms of the gain-loss ratio. Klöppel and Schweizer [21] researched good deal bounds based on expected utilities and their dynamics. In addition, much literature on this topic has been published. See, for instance, Carr, Geman, and Madan [7], Jaschke and Küchler [18], Černý and Hodges [9], Černý [8], Larsen et al. [23], Staum [24], and Björk and Slinko [6]. In particular, [21] overviewed a history of the research of this topic.

In this paper, we select shortfall risk as such a criterion to obtain a sharper pricing bound. Although the good deal bounds introduced above are independent of hedging strategies, the bounds presented in this paper are influenced by the way to define hedging (or admissible) strategies. Now, we presume a seller who intends to sell a claim and who selects an increasing convex function as her loss function. Note that an investor's attitude toward risk is assumed

*Received by the editors August 26, 2009; accepted for publication (in revised form) October 16, 2010; published electronically January 11, 2011. This research was supported by FY2007 Researcher Exchange Program between JSPS and SNSF, and Scientific Research (C) 19540144 from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

http://www.siam.org/journals/sifin/2/76912.html

†Department of Economics, Keio University, 2-15-45 Mita, Minato-ku, Tokyo, 108-8345, Japan (arai@econ.keio.ac.jp).

to be described in terms of loss function; that is, loss functions look like utility functions with respect to losses. The convexity of loss functions represents the investor's risk-aversion. The case of lower partial moments $l(x) = x^p/p$ for $p \geq 1$ is well known as a common example. For more details, see Remark 2.2 of Föllmer and Leukert [14]. When she selects a hedging strategy, her shortfall is defined as the positive part of the claim minus the value of her hedging strategy. Her shortfall risk is given by the expectation of the shortfall weighted by her loss function. The starting point of the shortfall risk problem is given by [14]. Incidentally, if the seller succeeds in selling the claim for the upper bound of no-arbitrage prices, she could eliminate her shortfall risk, that is, attain a perfect hedge by selecting a suitable hedging strategy. Note that this upper bound is called the superhedging cost. However, it would be too expensive to trade in general. Thus, in order to get a deal, she should sell the claim for a price less than the superhedging cost; that is, she should accept some shortfall risk. A price for a claim is called a good deal price for a seller if there exists an admissible strategy whose corresponding cashflow at maturity suppresses its shortfall risk below a threshold level determined by the seller. A claim is called a good deal for a seller if the claim is priced for a good deal price for the seller. We can define good deal prices and good deals for a buyer, too. Note that loss functions may vary from investor to investor. The interval between the lowest good deal price for a seller and the upper good deal price for a buyer is called a good deal bound induced by shortfall risk. This bound becomes narrower than the no-arbitrage one in general.

Throughout the paper, we consider the shortfall risk problem under the Orlicz heart setting. Thus, we can take various functions as a loss function, such as x^p/p for $p \geq 1$, $e^x - 1$, $x - \log(x+1)$, $(x+1)\log(x+1) - x$, and so forth. Although it suffices to consider the L^p setting for the case of the lower partial moments, it is too wide to treat the exponential loss function $e^x - 1$. While we can apply the L^∞ setting to the exponential loss function, L^∞ is too restrictive to treat various claims. This is why we consider the Orlicz heart setting to treat various loss functions and various claims in a unified framework. We shall prove that, for continuous time models, a good deal bound induced by shortfall risk is given by a convex risk measure on an Orlicz heart. Furthermore, we shall obtain its robust representation theorems for various settings and introduce some examples.

Convex risk measures are introduced by Föllmer and Schied [15] and Frittelli and Rosazza-Gianin [17] for the first time. The article [17] treated a representation of convex risk measures having lower semicontinuity. On the other hand, [15] defined convex risk measures only for the L^∞-random variables and obtained representation results. In addition, they proved that the upper and lower bounds of a good deal bound induced by shortfall risk are described by a convex risk measure for bounded claims under the discrete time setting, although they did not use the terminology "good deal." Thus, our result may be regarded as a direct extension of [15] to the framework of Orlicz hearts and continuous time setting. Xu [25] and Klöppel and Schweizer [20] considered problems similar to ours. Recently, Kaina and Rüschendorf [19] studied some properties of convex risk measures on L^p-spaces. In addition, they treated a convex risk measure related to shortfall risk, although theirs is somewhat different from ours. Cheridito and Li [10] extended the concept of convex risk measures to Orlicz hearts.

The paper is organized as follows: After some preparations in section 2, we shall prove in section 3 that a good deal bound induced by shortfall risk is given by a convex risk measure on an Orlicz heart. We obtain in section 4 representation results of this convex risk measure when the space of admissible strategies is linear. Moreover, we consider the case where the gain process is bounded from below by a random variable W in section 5. The σ-martingale measures will play a vital role. In addition, we consider the case where the admissible strategies form a predictable convex set in section 6. As an extension of a result of Föllmer and Kramkov [13], we introduce upper variation processes. We prove that the minimal penalty function of the convex risk measure on this setting is represented by using an upper variation process. We shall introduce some examples in subsection 6.3.

2. Preliminaries. Consider an incomplete financial market being composed of one riskless asset and d risky assets. The fluctuation of the risky assets is described by an $\mathbf{R}^d$-valued RCLL special semimartingale S, which is possibly nonlocally bounded, defined on a completed probability space $(\Omega, \mathcal{F}, P; \mathbf{F} = \{\mathcal{F}_t\}_{t\in[0,T]})$, where $T > 0$ is the maturity of our market and $\mathbf{F}$ is a filtration satisfying the so-called usual condition, that is, $\mathbf{F}$ is right-continuous, $\mathcal{F}_T = \mathcal{F}$, and $\mathcal{F}_0$ contains all null sets of $\mathcal{F}$. Suppose that the interest rate is given by 0. Let $\mathcal{X}$ be a suitable subset of L^0, the set of all random variables defined on $(\Omega, \mathcal{F}_T)$. We assume that any contingent claim belongs to the set $\mathcal{X}$.

Next, we presume a seller who intends to sell a claim $X \in \mathcal{X}$. She selects a loss function l, which is an $\mathbf{R}_+$-valued continuous nondecreasing convex function defined on $\mathbf{R}$, where $\mathbf{R}_+ = [0, \infty)$. We assume that $l(x) = 0$ if $x \le 0$, and $l(x) > 0$ if $x > 0$. Let Θ be a convex subset of S-integrable predictable processes including 0, and let $G_t(\vartheta) := \int_0^t \vartheta_s dS_s$ for any $t \in [0, T]$ and any $\vartheta \in \Theta$. Throughout the paper, Θ represents the set of all admissible strategies. Note that any admissible strategy in this paper is self-financing and represented by a predictable process. The process $G(\vartheta)$ represents the gain process induced by an admissible strategy $\vartheta \in \Theta$. In sections 4–6, we shall take Θ^M, Θ^W, and $\Theta^{\mathcal{S}}$ as Θ, respectively, where Θ^M is the linear space of S-integrable predictable processes ϑ such that $G_T(\vartheta)$ belongs to the Orlicz heart generated by l, Θ^W is defined, for a random variable W satisfying some conditions, as the conic set of ϑ such that there exists a constant $c > 0$ satisfying $G_t(\vartheta) \ge -cW$ for any $t \in [0, T]$, and $\Theta^{\mathcal{S}}$ is defined as the corresponding subset of Θ^W to $\mathcal{S}$ being a predictably convex subset of $\{G_T(\vartheta) | \vartheta \in \Theta^W\}$. When X is priced for $x \in \mathbf{R}$ and the seller selects a hedging strategy $\vartheta \in \Theta$, her shortfall and shortfall risk are defined by $(-x - G_T(\vartheta) + X) \vee 0$ and $E[l(-x - G_T(\vartheta) + X)]$, respectively.

In order to determine whether a price of X is a good deal price or not, we presume that the seller selects a constant $\delta > 0$, which is called a threshold level in this paper. That is, a price $x \in \mathbf{R}$ is a good deal price of X for her if there exists a $\vartheta \in \Theta$ such that $x + G_T(\vartheta) - X \in \mathcal{A}^0$, where $\mathcal{A}^0 := \{Y \in \mathcal{X} | E[l(-Y)] \le \delta\}$. Note that $x + G_T(\vartheta) - X$ is required to be in $\mathcal{X}$. Hence, the lower bound B^s of good deal prices is given by

$$B^s = \inf\{x \in \mathbf{R} |\text{ there exists a } \vartheta \in \Theta \text{ such that } x + G_T(\vartheta) - X \in \mathcal{A}^0\}.$$

In the case where $\delta = 0$, no matter which l is selected, B^s is given by $\inf\{x \in \mathbf{R} |$ there exists a $\vartheta \in \Theta$ such that $x + G_T(\vartheta) \ge X\}$. This is why B^s coincides with the superhedging cost. In addition, B^s may be regarded as a decreasing function on δ. Similarly, we can consider the upper bound B^b of good deal prices for a buyer. Now, we assume a buyer with a loss function l^b and a threshold level δ^b. Note that any buyer or seller would select a different loss function and a different threshold level from others. Then, we have

$$B^b = \sup\{x \in \mathbf{R} |\ \text{there exists a } \vartheta \in \Theta \text{ such that } -x + G_T(\vartheta) + X \in \mathcal{A}^{0,b}\},$$

where $\mathcal{A}^{0,b} = \{Y \in \mathcal{X} | E[l^b(-Y)] \leq \delta^b\}$. Thus, the interval (B^b, B^s) must be called the good deal bound of X induced by shortfall risk between the seller and the buyer.

In order to get a representation of B^s as well as B^b, we define a functional ρ_l on $\mathcal{X}$ as

$$\begin{aligned}\rho_l(X) := \inf\{x \in \mathbf{R} |\ & \text{there exists a } \vartheta \in \Theta \\ & \text{such that } x + G_T(\vartheta) + X \in \mathcal{A}^0\}.\end{aligned} \tag{2.1}$$

Note that $B^s = \rho_l(-X)$, and $B^b = -\rho_{l^b}(X)$ with the threshold level δ^b. Hence, we have only to obtain a robust representation of ρ_l in order to represent the lower and the upper bounds of good deal bounds induced by shortfall risk. Thus, the main aim of the paper is to obtain representations of ρ_l for various settings. Actually, the functional ρ_l is a convex risk measure on $\mathcal{X}$. This fact will be proved in the next section under the Orlicz heart setting. Moreover, the set $\mathcal{A}^0$ is closely related to the acceptance set or the penalty function of ρ_l. Thus, we define it as a subset of $\mathcal{X}$.

Example 2.1 (one period model). Consider a one period model with trading times $t = 0$ and 1. In particular, we consider the case where the asset price process S is nonlocally bounded. Let S be given by $S_0 = 1$ and $S_1 = |Y|$, where Y is a random variable following a normal distribution $N(\mu, \sigma^2)$. We consider a European call option $X := (S_1 - K)^+$, where $K > 0$. The set Θ of all admissible strategies is given by $\mathbf{R}$. Then, the superhedging costs of X for sellers and for buyers are given by 1 and $0 \vee (1 - K)$, respectively. Let the loss function l be the exponential type $e^x - 1$ and the threshold level $\delta > 0$. Noting that $l(x) = 0$ for any $x \leq 0$, we have

$$\begin{aligned}
&\inf_{\vartheta \in \mathbf{R}} E[l(-1 + \varepsilon - \vartheta(S_1 - 1) + X)] \\
&\leq E[l(-1 + \varepsilon - (S_1 - 1) + X)] = E[l(\varepsilon - S_1 + (S_1 - K)^+)] \\
&= E[l(\varepsilon - S_1) 1_{\{S_1 < K\}}] + E[l(\varepsilon - K) 1_{\{S_1 \geq K\}}] \\
&= \int_{-\varepsilon}^{\varepsilon} \left(e^{\varepsilon - |y|} - 1 \right) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y - \mu)^2}{2\sigma^2} \right\} dy \leq \frac{2\varepsilon(e^{\varepsilon} - 1)}{\sqrt{2\pi\sigma^2}} \leq \delta
\end{aligned}$$

for any sufficiently small $\varepsilon \in (0, K)$, Hence, $\rho_l(-X)$, that is, the lowest good deal price for a seller, is less than the superhedging cost 1.

Remark 2.2. In general, good deal bounds are induced by a performance measure, such as the Sharpe ratio, the gain-loss ratio, or expected utilities. These bounds have nothing to do with how to take the set Θ of all admissible strategies. On the other hand, the pricing bounds treated in this paper take the structure of Θ into account. In other words, comparing Theorems 4.1, 5.2, and 6.9, we see that representations of the functional ρ_l change according to the definitions of admissible strategies.

Remark 2.3. Xu [25] proposed the concept of the risk indifference valuation, which is a valuation method for claims by using a convex risk measure. Given a convex risk measure ρ

and an initial liability L which is an $\mathcal{F}_T$-measurable random variable, the selling price of a claim X is defined as

$$\inf\left\{x \in \mathbf{R} \middle| \inf_{\vartheta\in\Theta} \rho(x + G_T(\vartheta) - L - X) \leq \inf_{\vartheta\in\Theta} \rho(G_T(\vartheta) - L)\right\}.$$

We can define the buying price in a similar way. It was pointed out that these prices form a good deal bound. Although the mathematical setting in [25] is slightly different from ours, this problem is closely related. Actually, the risk indifference valuation based on ρ_l was mentioned there. Its selling price of X is given by $\rho_l(-X) - \rho_l(0)$, which is subtly different from $B^s = \rho_l(-X)$ if L is given by 0. However, [25] did not discuss any robust representation result on convex risk measures.

Throughout the paper, we consider ρ_l on an Orlicz heart generated by a loss function. That is, we take an Orlicz heart as $\mathcal{X}$. Thus, we have to introduce terminologies and concepts on Orlicz spaces. A left-continuous nondecreasing convex nontrivial function $\Phi : \mathbf{R}_+ \to [0, \infty]$ with $\Phi(0) = 0$ is called an Orlicz function, where Φ is nontrivial if $\Phi(x) > 0$ for some $x > 0$ and $\Phi(x) < \infty$ for some $x > 0$. When Φ is an $\mathbf{R}_+$-valued continuous, strictly increasing Orlicz function, we call it a strict Orlicz function in this paper. Note that, for any strict Orlicz function Φ, we have $\Phi(x) \in (0, \infty)$ for any $x > 0$ and $\lim_{x\to\infty} \Phi(x) = \infty$. Moreover, a strict Orlicz function Φ is differentiable a.e. and its left derivative Φ' satisfies

$$\Phi(x) = \int_0^x \Phi'(u)du.$$

Note that Φ' is left-continuous and may have at most countably many jumps. Define $I(y) := \inf\{x \in (0, \infty)|\Phi'(x) \geq y\}$, which is called the generalized left-continuous inverse of Φ'. We define $\Psi(y) := \int_0^y I(v)dv$ for $y \geq 0$, which is an Orlicz function called the conjugate function of Φ.

Remark 2.4. Any polynomial function starting at 0 whose minimal degree is equal to or greater than 1 and whose coefficients are all positive is a strict Orlicz function. For example, cx^p for $c > 0$, $p \geq 1$, $x^2 + 3x^5$, and so forth. Moreover, $e^x - 1$, $e^x - x - 1$, $(x+1)\log(x+1) - x$, and $x - \log(x + 1)$ are strict Orlicz functions.

Now, we need the following definitions.

Definition 2.5. *For an Orlicz function Φ, we define three spaces of random variables:*
Orlicz space: $L^\Phi := \{X \in L^0|E[\Phi(c|X|)] < \infty$ *for some* $c > 0\}$,
Orlicz heart: $M^\Phi := \{X \in L^0|E[\Phi(c|X|)] < \infty$ *for any* $c > 0\}$.
In addition, we define two norms:
Luxemburg norm: $\|X\|_\Phi := \inf\left\{\lambda > 0 \middle| E\left[\Phi\left(\left|\frac{X}{\lambda}\right|\right)\right] \leq 1\right\}$,
Orlicz norm: $\|X\|^*_\Phi := \sup\{E[XY]| \, \|Y\|_\Phi \leq 1\}$.

Note that $M^\Phi \subset L^\Phi$ and both spaces L^Φ and M^Φ are linear. Moreover, if Φ is a strict Orlicz function, the norm dual of $(M^\Phi, \|\cdot\|_\Phi)$ is given by $(L^\Psi, \|\cdot\|^*_\Phi)$.

Remark 2.6. In the case of the lower partial moments $\Phi(x) = x^p/p$ for $p > 1$, the Orlicz space L^Φ and the Orlicz heart M^Φ are both identical to L^p. In this case, the conjugate function is given by x^q/q, where $q = p/(p - 1)$, and $M^\Psi = L^\Psi = L^q$. See Example 4.5.

In general, if $\limsup_{x\to\infty} \frac{x\Phi'(x)}{\Phi(x)} < \infty$, then M^Φ is identical to L^Φ. For instance, $\Phi(x) = x - \log(x + 1)$. Otherwise, M^Φ must be a subset of L^Φ—for example, $\Phi(x) = e^x - 1$.

Hereafter, a strict Orlicz function Φ is fixed arbitrarily, and l is taken as the loss function associated with Φ as follows:

$$l(x) := \begin{cases} \Phi(x) & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

Moreover, we fix a constant $\delta > 0$ and treat ρ_l as a functional defined on M^Φ. Thus, we take M^Φ as $\mathcal{X}$, and $\mathcal{A}^0$ is denoted by $\mathcal{A}^0 := \{Y \in M^\Phi | E[l(-Y)] \leq \delta\}$.

3. Convex risk measure property of ρ_l. Föllmer and Schied [15] have proved that, roughly speaking, ρ_l defined by (2.1) becomes a convex risk measure in the framework of bounded claims and discrete time trading. In this section, we try to extend their result to the framework of Orlicz hearts and continuous time trading. Recall that the set Θ of all admissible strategies is a convex subset of S-integrable predictable processes including 0. Let $\mathcal{P}^\Psi$ be the set of all probability measures absolutely continuous with respect to P and having L^Ψ-density with respect to P, that is,

$$\mathcal{P}^\Psi := \{Q \ll P | dQ/dP \in L^\Psi\}.$$

Next, we state the standing assumption of this paper.

Assumption 3.1. $\rho_l(0) > -\infty$.

Assumption 3.1 is used to prove only the properness of ρ_l in Proposition 3.3. There are many examples satisfying this condition. Now, we introduce a sufficient condition for Assumption 3.1.

Example 3.2. We assume that there exists an $m > 0$ such that $\inf_{\vartheta\in\Theta} E[l(m - G_T(\vartheta))] > 0$. This condition would be very similar to the no-arbitrage one. Under this assumption, we can find a $c \geq 1$ satisfying $\inf_{\vartheta\in\Theta} E[l(cm - G_T(\vartheta))] > \delta$ by the convexity of Θ. Thus, $\rho_l(0) > -\infty$.

In Cheridito and Li [10], a convex risk measure on an Orlicz heart $\mathcal{Y}$ is defined as a $(-\infty, +\infty]$-valued functional ρ on $\mathcal{Y}$ satisfying the following:

(1) *Properness*: $\rho(0) \in \mathbf{R}$ and $\rho > -\infty$.
(2) *Monotonicity*: $\rho(X) \geq \rho(Y)$ for any $X, Y \in \mathcal{Y}$ such that $X \leq Y$.
(3) *Translation invariance*: $\rho(X + m) = \rho(X) - m$ for $X \in \mathcal{Y}$ and $m \in \mathbf{R}$.
(4) *Convexity*: $\rho(\lambda X + (1-\lambda)Y) \leq \lambda\rho(X) + (1-\lambda)\rho(Y)$ for any $X, Y \in \mathcal{Y}$ and $\lambda \in [0, 1]$.

First, we prove the following proposition.

Proposition 3.3. *ρ_l is an $\mathbf{R}$-valued convex risk measure on M^Φ.*

Proof. It is clear that ρ_l is translation invariant and monotone. We prove the convexity. Denoting $\mathcal{A}^0 - G_T(\vartheta) := \{Y - G_T(\vartheta) | Y \in \mathcal{A}^0\}$ for any $\vartheta \in \Theta$, we have, for any $X \in M^\Phi$,

$$\begin{aligned}
\rho_l(X) &= \inf\left\{x \in \mathbf{R} \middle| \text{ there exists a } \vartheta \in \Theta \text{ such that } x + G_T(\vartheta) + X \in \mathcal{A}^0\right\} \\
&= \inf\left\{x \in \mathbf{R} \middle| \text{ there exists a } \vartheta \in \Theta \text{ such that } x + X \in (\mathcal{A}^0 - G_T(\vartheta))\right\} \\
&= \inf\left\{x \in \mathbf{R} \middle| x + X \in \bigcup_{\vartheta\in\Theta} (\mathcal{A}^0 - G_T(\vartheta))\right\}.
\end{aligned}$$

Note that, although a union of convex sets is in general not convex, the set $\bigcup_{\vartheta\in\Theta}(\mathcal{A}^0 - G_T(\vartheta))$ becomes convex by the convexity of Θ and the linearity of G_T. Let X_1 and X_2 be in M^Φ.

For any $x_1 > \rho_l(X_1)$, $x_2 > \rho_l(X_2)$, and $\lambda \in [0,1]$, we can easily prove $\lambda x_1 + (1-\lambda)x_2 \geq \rho_l(\lambda X_1 + (1-\lambda)X_2)$. By the arbitrariness of x_1 and x_2, the convexity of ρ_l follows.

Next, we prove the properness of ρ_l, that is, $\rho_l(0) \in \mathbf{R}$ and $\rho_l > -\infty$. It is clear that $\rho_l(0) \leq 0$. The rest is to prove $\rho_l > -\infty$. Suppose that there exists an $X \in M^\Phi$ such that $\rho_l(X) = -\infty$. The convexity of ρ_l yields that

$$\rho_l(\lambda X) \leq \lambda \rho_l(X) + (1-\lambda)\rho_l(0) = -\infty$$

for any $\lambda \in (0,1]$. Thus, we may assume that $E[l(X)] \leq \delta/2$. Since $\rho_l(0) > -\infty$, there exists an $m > 0$ such that $E[l(m - G_T(\vartheta))] > \delta$ for any $\vartheta \in \Theta$. Then, since $\rho_l(X) = -\infty$, there exists a $\vartheta^X \in \Theta$ such that

$$\delta \leq \inf_{\vartheta \in \Theta} E[l(m - G_T(\vartheta))] \leq \frac{1}{2} E[l(2m - G_T(\vartheta^X) - X)] + \frac{1}{2} E[l(X)] \leq \frac{3}{4}\delta,$$

which is a contradiction. Hence, the properness is obtained.

Finally, we prove that $\rho_l < \infty$. Since $\vartheta = 0$ belongs to Θ, we have, for any $X \in M^\Phi$, $\rho_l(X) \leq \inf\{m \in \mathbf{R} | E[l(-m - X)] \leq \delta\} < +\infty$. The last inequality is implied by the fact $E[l(-m-X)] \to 0$ as m tends to ∞ by the dominated convergence theorem. ■

Note that $(M^\Phi, \|\cdot\|_\Phi)$ is a locally convex Fréchet lattice with order continuous topology (see [5] or Aliprantis and Border [1]). Corollary 1 of [5], together with Proposition 3.3, implies that ρ_l is represented as

$$\rho_l(X) = \max_{Q \in \mathcal{P}^\Psi} \{E_Q[-X] - a_l(Q)\}, \tag{3.1}$$

where $a_l : \mathcal{P}^\Psi \to \mathbf{R}$ is the convex conjugate of ρ_l, which is called the minimal penalty function. Note that a_l is given by

$$a_l(Q) := \sup_{X \in M^\Phi} \{E_Q[-X] - \rho_l(X)\}. \tag{3.2}$$

In order to obtain a representation of a_l, we need some preparations. First, we define two acceptance sets before stating a significant lemma:

$$\mathcal{A}_{\rho_l} := \{X \in M^\Phi | \rho_l(X) \leq 0\} \tag{3.3}$$

and

$$\begin{aligned} \widetilde{\mathcal{A}} := \{X \in M^\Phi | & \text{ there exists a } \vartheta \in \Theta \\ & \text{such that } X + G_T(\vartheta) \geq Y \text{ for some } Y \in \mathcal{A}^0\}. \end{aligned}$$

Then, we can prove that $\rho_l(X) = \inf\{m \in \mathbf{R} | m + X \in \widetilde{\mathcal{A}}\} (=: \rho_{\widetilde{\mathcal{A}}}(X))$ by the very definition of $\widetilde{\mathcal{A}}$. Thus, by Remark 4.16(c) of Föllmer and Schied [16], we have $a_l(Q) = \sup_{X \in \widetilde{\mathcal{A}}} E_Q[-X]$. Since only the L^∞-case was discussed in [16], we try to give another proof of this fact just to be sure.

Lemma 3.4. *$a_l(Q) = \sup_{X \in \widetilde{\mathcal{A}}} E_Q[-X]$ for any $Q \in \mathcal{P}^\Psi$.*

Proof. By (3.2) and (3.3), we have

$$\sup_{X\in\mathcal{A}_{\rho_l}} E_Q[-X] \le \sup_{X\in\mathcal{A}_{\rho_l}} \{E_Q[-X]-\rho_l(X)\} \le \sup_{X\in M^\Phi} \{E_Q[-X]-\rho_l(X)\} = a_l(Q). \tag{3.4}$$

Since $\rho_l=\rho_{\widetilde{\mathcal{A}}}$, we have $\widetilde{\mathcal{A}}\subset\mathcal{A}_{\rho_l}$. Hence, (3.4) implies

$$a_l(Q) \ge \sup_{X\in\mathcal{A}_{\rho_l}} E_Q[-X] \ge \sup_{X\in\widetilde{\mathcal{A}}} E_Q[-X]. \tag{3.5}$$

Note that the translation invariance of ρ_l implies that

$$E_Q[-(X+m)]-\rho_l(X+m) = E_Q[-X]-\rho_l(X) \tag{3.6}$$

for any $m\in\mathbf{R}$, $X\in M^\Phi$, and $Q\in\mathcal{P}^\Psi$. Since $\{X\in M^\Phi|\rho_l(X)=-\varepsilon\}\subset\widetilde{\mathcal{A}}$ for any $\varepsilon>0$, we have

$$\begin{aligned} a_l(Q) &= \sup_{X\in M^\Phi}\{E_Q[-X]-\rho_l(X)\} = \sup_{X\in M^\Phi,\rho_l(X)=-\varepsilon}\{E_Q[-X]-\rho_l(X)\} \\ &= \sup_{X\in M^\Phi,\rho_l(X)=-\varepsilon} E_Q[-X]+\varepsilon \le \sup_{X\in\widetilde{\mathcal{A}}} E_Q[-X]+\varepsilon \end{aligned}$$

for any $\varepsilon>0$. The second equality holds due to (3.6). Thus, we have $a_l(Q)\le\sup_{X\in\widetilde{\mathcal{A}}}E_Q[-X]$. Hence, we can conclude, together with (3.5), that $a_l(Q)=\sup_{X\in\widetilde{\mathcal{A}}}E_Q[-X]$. ∎

Based on the above preparations, we prove the following representation result of the convex risk measure ρ_l.

Proposition 3.5. *The convex risk measure ρ_l is represented as, for any $X\in M^\Phi$,*

$$\rho_l(X) = \max_{Q\in\mathcal{P}^\Psi}\left\{E_Q[-X]-\sup_{X^1\in\mathcal{A}^1}E_Q[-X^1]-\inf_{\lambda>0}\frac{1}{\lambda}\left\{\delta+E\left[\Psi\left(\lambda\frac{dQ}{dP}\right)\right]\right\}\right\}, \tag{3.7}$$

where

$$\mathcal{A}^1 := \{X^1\in M^\Phi|\ \textit{there exists a } \vartheta\in\Theta \textit{ such that } X^1+G_T(\vartheta)\ge 0\ P\text{-a.s.}\}. \tag{3.8}$$

Proof. For any $X\in\widetilde{\mathcal{A}}$, there exist a $\vartheta\in\Theta$ and a $Y\in\mathcal{A}^0$ such that $X+G_T(\vartheta)\ge Y$, that is, $X-Y+G_T(\vartheta)\ge 0$. Hence, $X-Y\in\mathcal{A}^1$, that is, $\widetilde{\mathcal{A}}=\{X^1+X^0|X^1\in\mathcal{A}^1,X^0\in\mathcal{A}^0\}$. Thus, Lemma 3.4 implies that, for any $Q\in\mathcal{P}^\Psi$,

$$\begin{aligned} a_l(Q) &= \sup_{X\in\widetilde{\mathcal{A}}}E_Q[-X] = \sup_{X^1\in\mathcal{A}^1}\sup_{X^0\in\mathcal{A}^0}E_Q[-X^1-X^0] \\ &= \sup_{X^1\in\mathcal{A}^1}\left\{E_Q[-X^1]+\sup_{X^0\in\mathcal{A}^0}E_Q[-X^0]\right\} \\ &= \sup_{X^1\in\mathcal{A}^1}E_Q[-X^1]+\sup_{X^0\in\mathcal{A}^0}E_Q[-X^0]. \end{aligned} \tag{3.9}$$

Consequently, (3.9) implies that

$$\rho_l(X) = \max_{Q\in\mathcal{P}^\Psi}\{E_Q[-X] - a_l(Q)\}$$

$$= \max_{Q\in\mathcal{P}^\Psi}\left\{E_Q[-X] - \sup_{X^1\in\mathcal{A}^1} E_Q[-X^1] - \sup_{X^0\in\mathcal{A}^0} E_Q[-X^0]\right\}. \tag{3.10}$$

Moreover, by the same sort argument as in Theorem 10 of [15], we can obtain that

$$\sup_{X^0\in\mathcal{A}^0} E_Q[-X^0] = \inf_{\lambda>0}\frac{1}{\lambda}\left\{\delta + E\left[\Psi\left(\lambda\frac{dQ}{dP}\right)\right]\right\}.$$

This, together with (3.10), completes the proof of Proposition 3.5. ■

Remark 3.6. We can extend the above result to more general cases. Let G be a convex set of $\mathcal{F}_T$-measurable random variables including 0. We can rewrite the definition of ρ_l as follows:

$$\rho_l(X) := \inf\{x\in\mathbf{R}|\text{ there exists a } g\in G \text{ such that } x+g+X\in\mathcal{A}^0\}.$$

Under this definition, ρ_l must be represented in the same way as in Proposition 3.5.

4. Unconstrained portfolio. We consider the case where Θ, which is the space of all admissible strategies, forms a linear subspace of M^Φ. Let $L(S)$ be the set of all S-integrable predictable processes. Denoting $\Theta^M := \{\vartheta\in L(S)|G_T(\vartheta)\in M^\Phi\}$, we regard Θ^M as the collection of admissible strategies. Note that Θ^M is linear. We shall introduce a representation theorem for ρ_l under this setting and some examples.

Now, we define the set of martingale measures as follows:

$$\mathcal{M}^\Psi := \left\{Q\in\mathcal{P}^\Psi|E_Q[G_T(\vartheta)] = 0 \text{ for any } \vartheta\in\Theta^M\right\}. \tag{4.1}$$

The following theorem is the main result of this section. The proof is strongly dependent on the linearity of Θ^M.

Theorem 4.1. *Let Θ be given by Θ^M. Suppose that $\mathcal{M}^\Psi\neq\emptyset$. The convex risk measure ρ_l is represented as*

$$\rho_l(X) = \max_{Q\in\mathcal{M}^\Psi}\left\{E_Q[-X] - \inf_{\lambda>0}\frac{1}{\lambda}\left\{\delta + E\left[\Psi\left(\lambda\frac{dQ}{dP}\right)\right]\right\}\right\}$$

for any $X\in M^\Phi$.

Proof. By Proposition 3.5, we have only to calculate $\sup_{X^1\in\mathcal{A}^1} E_Q[-X^1]$. For any $X^1\in\mathcal{A}^1$, there exists a $\vartheta\in\Theta^M$ such that $X^1\geq G_T(-\vartheta)$. Thus, $\sup_{X^1\in\mathcal{A}^1} E_Q[-X^1]\leq\sup_{\vartheta\in\Theta^M} E_Q[G_T(\vartheta)]$. On the other hand, since $G_T(\vartheta)\in\mathcal{A}^1$ for any $\vartheta\in\Theta^M$, we have $\sup_{X^1\in\mathcal{A}^1} E_Q[-X^1]\geq\sup_{\vartheta\in\Theta^M} E_Q[-G_T(\vartheta)] = \sup_{\vartheta\in\Theta^M} E_Q[G_T(\vartheta)]$. The linearity of Θ^M and (4.1) imply that

$$\sup_{X^1\in\mathcal{A}^1} E_Q[-X^1] = \sup_{\vartheta\in\Theta^M} E_Q[G_T(\vartheta)] = \begin{cases} 0 & \text{if } Q\in\mathcal{M}^\Psi, \\ \infty & \text{if } Q\notin\mathcal{M}^\Psi, \end{cases}$$

which completes the proof of Theorem 4.1. ■

Remark 4.2. We have to remark that the assumption $\mathcal{M}^\Psi \neq \emptyset$ does not ensure the no-arbitrage condition, which is exemplified as follows: Consider a one period model. The price process S is defined as follows: $S_0 = 0$ and S_1 follows the exponential distribution with parameter 1. Thus, this model includes arbitrage opportunities. Let $\Phi(x) = e^x - 1$. Thus, $\Theta^M = \{0\}$. Hence, $\mathcal{M}^\Psi$ is given by $\{Q \ll P | dQ/dP \in L^\Psi\}$. Note that each element of $\mathcal{M}^\Psi$ is not a martingale measure.

Corollary 4.3. *For any $Q \in \mathcal{M}^\Psi$, if we find a $\widehat{\lambda}_Q > 0$ satisfying $\delta = E[\Phi(I(\widehat{\lambda}_Q \frac{dQ}{dP}))]$, then we have*

$$a_l(Q) = E_Q\left[I\left(\widehat{\lambda}_Q \frac{dQ}{dP}\right)\right]. \tag{4.2}$$

Recall that I is the generalized left-continuous inverse of the left derivative Φ'. Note that we can find such a $\widehat{\lambda}_Q$ at least when I is continuous.

Proof. For any $\lambda > 0$, we have

$$\frac{1}{\lambda}\left\{\delta + E\left[\Psi\left(\lambda\frac{dQ}{dP}\right)\right]\right\} \geq E_Q\left[I\left(\widehat{\lambda}_Q \frac{dQ}{dP}\right)\right]$$

by Young's inequality. Moreover, we have $\frac{1}{\widehat{\lambda}_Q}\{\delta + E[\Psi(\widehat{\lambda}_Q \frac{dQ}{dP})]\} = E_Q[I(\widehat{\lambda}_Q \frac{dQ}{dP})]$. Therefore, (4.2) holds. ∎

We introduce three typical examples.

Example 4.4 (expected shortfall). Let Φ be $\Phi(x) = x$. Its conjugate is given by

$$\Psi(y) = \begin{cases} 0 & \text{if } y \leq 1, \\ +\infty & \text{if } y > 1. \end{cases}$$

Denoting $\mathcal{M}^{bdd} := \{Q \ll P | dQ/dP \in L^\infty, E_Q[G_T(\vartheta)] = 0 \text{ for any } \vartheta \in \Theta^M\}$, we have then

$$\rho_l(X) = \max_{Q \in \mathcal{M}^{bdd}} \left\{E_Q[-X] - \delta\left\|\frac{dQ}{dP}\right\|_\infty\right\}.$$

Example 4.5 (lower partial moments). Let Φ be given by $\Phi(x) = x^p/p$ for $p > 1$. Note that $M^\Phi = L^\Phi = L^p$ and $M^\Psi = L^\Psi = L^q$, where q is the conjugate index of p. Example 13 of [15] implies that

$$a_l(Q) = (p\delta)^{\frac{1}{p}}\left\|\frac{dQ}{dP}\right\|_q.$$

Thus, ρ_l is represented as

$$\rho_l(X) = \max_{Q \in \mathcal{M}^q}\left\{E_Q[-X] - (p\delta)^{\frac{1}{p}}\left\|\frac{dQ}{dP}\right\|_q\right\},$$

where $\mathcal{M}^q$ is the set of all absolutely continuous martingale measures whose density is in the space L^q.

Example 4.6 (exponential loss functions). If Φ is given by $\Phi(x)=C(e^{\alpha x}-1)$, where $C>0$ and $\alpha>0$, then $\Phi'(x)=C\alpha e^{\alpha x}$, and $I(y)=\frac{1}{\alpha}\log\frac{y}{C\alpha}$ when $y\geq C\alpha$. Hence, for $y\geq C\alpha$, we have

$$\Psi(y)=\frac{y}{\alpha}\left(\log\frac{y}{C\alpha}-1\right)+C.$$

Note that $L^\infty\subset M^\Phi\subset L^p\subset L^\Psi\subset L^1$ for any $1<p<\infty$. Define $\mathcal{M}^f:=\{Q\ll P|H(Q|P)<\infty, E_Q[G_T(\vartheta)]=0 \text{ for any } \vartheta\in\Theta^M\}$, where $H(Q|P):=E[\frac{dQ}{dP}\log\frac{dQ}{dP}]$. For any $Q\in\mathcal{M}^f$, we denote by $\widehat{\lambda}_Q$ a real number satisfying $E[\Phi(I(\widehat{\lambda}_Q\frac{dQ}{dP}))]=\delta$. Then, Corollary 4.3 implies that

$$\rho_l(X)=\max_{Q\in\mathcal{M}^f}\left\{E_Q[-X]-\frac{1}{\alpha}E_Q\left[\log\left(\frac{\widehat{\lambda}_Q\frac{dQ}{dP}}{C\alpha}\vee 1\right)\right]\right\}.$$

5. W-admissible case. In this section and the next, we shall investigate models with portfolio constraints. That is, we shall obtain representation results of ρ_l when Θ is not a linear space. In this section, we consider a model with cone constraint, which is so-called W-admissibility. As for a financial interpretation of W-admissibility, see Biagini and Frittelli [3]. From the viewpoint of Proposition 3.5, all we have to do is to calculate the term $\sup_{X^1\in\mathcal{A}^1}E_Q[-X^1]$ in (3.7), where the definition of $\mathcal{A}^1$ is given in (3.8). Hereafter, we fix an $\mathcal{F}_T$-measurable random variable W such that $W\geq 1$ and $W\in M^\Phi$. Moreover, we define

$$\Theta^W:=\{\vartheta\in L(S)|\text{ there exists a } c>0 \text{ such that } G_t(\vartheta)\geq -cW\ P\text{- a.s. for any } t\in[0,T]\}. \tag{5.1}$$

We will use the following lemma not only in this section but also throughout the paper.

Lemma 5.1. *Suppose that $\Theta\subset\Theta^W$. For any $Q\in\mathcal{P}^\Psi$, we have*

$$\sup_{X^1\in\mathcal{A}^1}E_Q[-X^1]=\sup_{\vartheta\in\Theta}E_Q[G_T(\vartheta)].$$

Proof. For any $\widetilde{c}>0$ and any $\vartheta\in\Theta$, there exists a $c>0$ such that $-cW\leq G_T(\vartheta)\wedge\widetilde{c}W\leq\widetilde{c}W$. Thus, we have $G_T(\vartheta)\wedge\widetilde{c}W\in M^\Phi$ and $-(G_T(\vartheta)\wedge\widetilde{c}W)\in\mathcal{A}^1$ for any $\widetilde{c}>0$. The monotone convergence theorem implies that

$$\sup_{\vartheta\in\Theta}E_Q[G_T(\vartheta)]=\sup_{\vartheta\in\Theta}\lim_{\widetilde{c}\to\infty}E_Q[G_T(\vartheta)\wedge\widetilde{c}W]\leq\sup_{X^1\in\mathcal{A}^1}E_Q[-X^1]. \tag{5.2}$$

On the other hand, for any $X^1\in\mathcal{A}^1$, there exists a $\vartheta\in\Theta$ such that $G_T(\vartheta)\geq -X^1$. Hence, we have $\sup_{X^1\in\mathcal{A}^1}E_Q[-X^1]\leq\sup_{\vartheta\in\Theta}E_Q[G_T(\vartheta)]$. Together with (5.2), this completes the proof of Lemma 5.1. ■

We regard Θ^W as the set of all admissible strategies. In other words, Θ coincides with Θ^W. We define the set of all absolutely continuous σ-martingale measures as

$$\mathbf{M}^\sigma:=\{Q\ll P|S\text{ is a }\sigma\text{-martingale under }Q\}.$$

Moreover, a nonnegative random variable Y is said to be *suitable* if

1. $Y\geq 1$, and

2. for each $i=1,\dots,d$, there exists a $\vartheta^i \in L(S^i)$ such that $P(\{\omega|$ there exists a $t\in[0,T]$ such that $\vartheta^i_t(\omega)=0\})=0$ and $|\int_0^t \vartheta^i_s dS^i_s| \leq Y$ for any $t\in[0,T]$, P-a.s.

We state the main theorem of this section. In the W-admissible case, ρ_l has the same representation as in the unconstrained case discussed in the previous section, except for replacing $\mathcal{M}^\Psi$ by $\mathbf{M}^\sigma$.

Theorem 5.2. *Let Θ be given by Θ^W. Suppose that $\mathbf{M}^\sigma \cap \mathcal{P}^\Psi \neq \emptyset$ and W is suitable. For any $Q\in\mathcal{P}^\Psi$, we have then*

$$\sup_{X^1\in\mathcal{A}^1} E_Q[-X^1] = \begin{cases} 0 & \text{if } Q\in\mathbf{M}^\sigma, \\ +\infty & \text{if } Q\notin\mathbf{M}^\sigma. \end{cases}$$

Hence, we have

$$\rho_l(X) = \max_{Q\in\mathbf{M}^\sigma\cap\mathcal{P}^\Psi} \left\{ E_Q[-X] - \inf_{\lambda>0}\frac{1}{\lambda}\left\{\delta + E\left[\Psi\left(\lambda\frac{dQ}{dP}\right)\right]\right\}\right\}$$

for any $X\in M^\Phi$.

Proof. Step 1. By Proposition 19(d) of Biagini and Frittelli [4], we have

$$\mathbf{M}^\sigma\cap\mathcal{P}^\Psi = \{Q\in\mathcal{P}^\Psi | G(\vartheta) \text{ is a } Q\text{-supermartingale for any } \vartheta\in\Theta^W\}. \tag{5.3}$$

Thus, when $Q\in\mathbf{M}^\sigma\cap\mathcal{P}^\Psi$, $E_Q[G_T(\vartheta)]\leq 0$ for any $\vartheta\in\Theta^W$. The fact $0\in\Theta^W$ implies $\sup_{\vartheta\in\Theta^W} E_Q[G_T(\vartheta)]=0$. Together with Lemma 5.1, $\sup_{X^1\in\mathcal{A}^1} E_Q[-X^1] = \sup_{\vartheta\in\Theta^W} E_Q[G_T(\vartheta)]$ $=0$ for any $Q\in\mathbf{M}^\sigma\cap\mathcal{P}^\Psi$.

Step 2. Suppose that $Q\in\mathcal{P}^\Psi$ and $Q\notin\mathbf{M}^\sigma$. Then, there exists a $\vartheta\in\Theta^W$ such that $G(\vartheta)$ is not a Q-supermartingale from the viewpoint of (5.3). Now, we define, for $0\leq t_1<t_2\leq T$ and $\vartheta\in\Theta^W$,

$$U(t_1,t_2;\vartheta) := \{E_Q[G_{t_2}(\vartheta)|\mathcal{F}_{t_1}] > G_{t_1}(\vartheta)\}. \tag{5.4}$$

Then, there exist $t_1,t_2\in[0,T]$ with $t_1<t_2$ and a $\overline{\vartheta}\in\Theta^W$ such that $Q(U(t_1,t_2;\overline{\vartheta}))>0$. Otherwise, we have $E_Q[G_{t_2}(\vartheta)|\mathcal{F}_{t_1}]\leq G_{t_1}(\vartheta)$ for any $t_1,t_2\in[0,T]$ with $t_1<t_2$ and any $\vartheta\in\Theta^W$. Substituting $t_1=0$, we have $E_Q[G_{t_2}(\vartheta)]\leq 0$ for any $t_2\in[0,T]$ and any $\vartheta\in\Theta^W$. Incidentally, for any $\vartheta\in\Theta^W$, there exists a $c>0$ such that $G_t(\vartheta)\geq -cW$ for any $t\in[0,T]$ by (5.1). Thus, $G_t(\vartheta)\in L^1(Q)$ for any $t\in[0,T]$ and any $\vartheta\in\Theta^W$. Hence, $G(\vartheta)$ is a Q-supermartingale for any $\vartheta\in\Theta^W$. This is a contradiction.

Now, for $t\in[0,T]$, $k\geq 1$, and $\vartheta\in\Theta^W$, we define

$$B(t,k;\vartheta) := \{G_t(\vartheta)\leq k\}.$$

Then, for any $t\in[0,T]$ and any $\vartheta\in\Theta^W$, $Q(\cup_{k=1}^\infty B(t,k;\vartheta))=1$. Taking a sufficiently large $\overline{k}\geq 1$, we have $Q(U(t_1,t_2;\overline{\vartheta})\cap B(t_1,\overline{k};\overline{\vartheta}))>0$. Note that, denoting $\overline{U}=U(t_1,t_2;\overline{\vartheta})\cap B(t_1,\overline{k};\overline{\vartheta})$, we have $\overline{U}\in\mathcal{F}_{t_1}$. We construct $\widetilde{\vartheta}$ by using $\overline{\vartheta}$ as follows:

$$\widetilde{\vartheta}_t := \begin{cases} 0 & \text{if } t\leq t_1 \text{ or } t_2<t, \\ 1_{\overline{U}}\overline{\vartheta}_t & \text{if } t_1<t\leq t_2. \end{cases} \tag{5.5}$$

We have $G_{t_1}(\overline{\vartheta}) \leq \overline{k} \leq \overline{k}W$ on $\overline{U}$. Moreover, since $\overline{\vartheta} \in \Theta^W$, there exists a $\overline{c} > 0$ such that $G_t(\overline{\vartheta}) \geq -\overline{c}W$ for any $t \in [0,T]$. By (5.5), we obtain, for any $t \in [0,T]$,

$$G_t(\widetilde{\vartheta}) \geq -(\overline{c}+\overline{k})W,$$

that is, $\widetilde{\vartheta} \in \Theta^W$. As a result, (5.4) and the definition of $\overline{U}$ yield

$$\begin{aligned} E_Q[G_T(\widetilde{\vartheta})] &= E_Q[1_{\overline{U}}(G_{t_2}(\overline{\vartheta}) - G_{t_1}(\overline{\vartheta}))] \\ &= E_Q[1_{\overline{U}}(E_Q[G_{t_2}(\overline{\vartheta})|\mathcal{F}_{t_1}] - G_{t_1}(\overline{\vartheta}))] > 0. \end{aligned}$$

Since $c\widetilde{\vartheta} \in \Theta^W$ for any $c > 0$, we can conclude that $\sup_{\vartheta\in\Theta^W} E_Q[G_T(\vartheta)] = +\infty$. ■

6. Predictably convex case.

A family of semimartingales $\mathcal{S}$ is said to be *predictably convex* if, for any $S^1, S^2 \in \mathcal{S}$ and any $[0,1]$-valued predictable process h, $\int_0^\cdot h dS^1 + \int_0^\cdot (1-h) dS^2$ belongs to $\mathcal{S}$. In this section, let $\mathcal{S}$ be a predictable convex subset of $G(\Theta^W) := \{G(\vartheta)|\vartheta \in \Theta^W\}$. We denote by $\Theta^{\mathcal{S}}$ the corresponding subset of Θ^W to the fixed $\mathcal{S}$. That is, we can describe $\mathcal{S} = \{G(\vartheta)|\vartheta \in \Theta^{\mathcal{S}}\}$. Throughout this section, we regard $\Theta^{\mathcal{S}}$ as the set of all admissible strategies. Examples of $\mathcal{S}$ satisfying the above conditions shall be discussed in subsection 6.3. Now, we introduce an example of the case where Θ^W is not predictably convex.

Example 6.1 (Θ^W is not necessarily predictably convex). Let $d = 1$. We consider a discrete time model with maturity 3. Let Z_1, Z_2, Y_1, and Y_2 be four independent random variables. Suppose that $Z_i, i = 1,2$, takes values in $\{1, 2, \dots\}$, and $Y_i, i = 1,2$, in $\{-1, 1\}$, respectively. Let $\mathcal{F}_0 = \{\emptyset, \Omega\}$, $\mathcal{F}_1 = \sigma(Z_1)$, $\mathcal{F}_2 = \sigma(Y_1, Z_1)$, and $\mathcal{F}_3 = \sigma(Y_1, Y_2, Z_1, Z_2)$. Moreover, the asset price process S is given by $S_0 = S_1 = 0$, $S_2 = Z_1 1_{\{Y_1=1\}} - 1_{\{Y_1=-1\}}$, and

$$S_3 = (S_2+1)1_{\{Y_1=1,Y_2=1\}} - Z_2 1_{\{Y_1=1,Y_2=-1\}} - 1_{\{Y_1=-1\}}.$$

Suppose that $Z_2 \in M^\Phi$. Put $W = Z_2 + 1$.

For instance, letting $\vartheta_1 = 1$ and $\vartheta_2 = \vartheta_3 = Z_1^{-1}$, we have $|G_t(\vartheta)| \leq W$ for $t = 1,2,3$. Thus, W is suitable. Next, letting $\vartheta^1 \equiv 1$, we have $G_t(\vartheta^1) \geq -Z_2$ for $t = 1,2,3$, which means $\vartheta^1 \in \Theta^W$. However, letting $\vartheta^2_1 = \vartheta^2_2 = 0$ and $\vartheta^2_3 = 1$, we have $G_3(\vartheta^2) = -Z_1 - Z_2$ on $\{Y_1 = 1, Y_2 = -1\}$. Since Z_1 is independent of Z_2, we get $\vartheta^2 \notin \Theta^W$. Hence, Θ^W is not predictably convex.

6.1. Upper variation process.

Before stating the main theorem of this section, we need some mathematical preparations. We first define

$$\begin{aligned} \mathcal{P}(\mathcal{S}) := \{Q \ll P|\ &\text{there exists an increasing predictable process } A \\ &\text{such that } G(\vartheta) - A \text{ is a } Q\text{-supermartingale for any } \vartheta \in \Theta^{\mathcal{S}}\}. \end{aligned} \tag{6.1}$$

We can now prove the following lemma.

Lemma 6.2. *If $Q \in \mathcal{P}(\mathcal{S})$, then, for any $\vartheta \in \Theta^{\mathcal{S}}$, $G(\vartheta)$ is a special semimartingale under Q.*

Proof. By (6.1), for any $Q \in \mathcal{P}(\mathcal{S})$, there exists an increasing predictable process A^Q satisfying that $G(\vartheta) - A^Q$ is a Q-supermartingale for any $\vartheta \in \Theta^{\mathcal{S}}$. Now, we denote by $\overline{M^\vartheta} - \overline{A^\vartheta}$ the Doob decomposition of $G(\vartheta) - A^Q$, where $\overline{M^\vartheta}$ is a Q-local martingale, and $\overline{A^\vartheta}$

is an increasing predictable process null at 0. See Theorem VII.12 of Dellacherie and Meyer [12]. Then, we have $G(\vartheta) = \overline{M^\vartheta} - \overline{A^\vartheta} + A^Q$, which gives a canonical decomposition of $G(\vartheta)$ under Q. ■

Unless otherwise stated explicitly, for $Q \in \mathcal{P}(\mathcal{S})$, we denote by $M^\vartheta + A^\vartheta$ the canonical decomposition of $G(\vartheta)$ under Q. Note that this decomposition $M^\vartheta + A^\vartheta$ depends on Q. Now, for $Q \in \mathcal{P}(\mathcal{S})$, we define

$$\mathcal{A} := \{A^\vartheta | \vartheta \in \Theta^{\mathcal{S}}\}.$$

Note that the set $\mathcal{A}$ depends on Q. In particular, when $Q \notin \mathcal{P}(\mathcal{S})$, we are not necessarily able to define $\mathcal{A}$. In addition, for two stochastic processes X and Y, we define an order $\preceq$ as follows:

$$X \preceq Y \Longleftrightarrow Y - X \text{ is an increasing process.}$$

Remark 6.3. For any $A \in \mathcal{A}$, we can construct an increasing process $\widetilde{A} \in \mathcal{A}$ by modifying A. Denote

$$h_t := \begin{cases} 1, & dA_t \geq 0, \\ 0, & dA_t < 0. \end{cases}$$

Note that h is a $[0,1]$-valued predictable process. Defining $\widetilde{A}_t := \int_0^t h_s dA_s$, $\widetilde{A}$ is an increasing predictable process satisfying $A \preceq \widetilde{A}$. The predictable convexity of $\mathcal{S}$ implies that $\widetilde{A} \in \mathcal{A}$. By taking $\widetilde{A}$ instead of A if need be, we can regard every element of $\mathcal{A}$ as an increasing predictable process.

Remark 6.3 will be used in many places. Now, we need to prepare one more lemma.

Lemma 6.4. *For $Q \in \mathcal{P}(\mathcal{S})$, the ordered set $(\mathcal{A}, \preceq)$ is directed upward.*

Proof. Let $A^1, A^2 \in \mathcal{A}$. Define $B_t := \frac{1}{2}(A^1_t + A^2_t + \int_0^t |dA^1_s - dA^2_s|)$. By the proof of Lemma 2.2 of Kramkov [22], there exists a $\{-1,+1\}$-valued predictable process h satisfying $\int_0^t |dA^1_s - dA^2_s| = \int_0^t h_s(dA^1_s - dA^2_s)$. Hence, we can denote $B_t = \int_0^t \frac{1+h_s}{2} dA^1_s + \int_0^t \frac{1-h_s}{2} dA^2_s$, and the predictable convexity of $\mathcal{S}$ yields $B \in \mathcal{A}$. Moreover, $A^1 \preceq B$ holds, since $B_t - A^1_t = \int_0^t (dA^1_s - dA^2_s)^-$. We obtain $A^2 \preceq B$ in the same manner. As a result, $(\mathcal{A}, \preceq)$ is directed upward. ■

For any $A^1, A^2 \in \mathcal{A}$, we denote the above B by $A^1 \vee A^2$. An increasing predictable process $A^{\mathcal{S}}$ is called an *upper variation process* of the ordered set $(\mathcal{A}, \preceq)$ if $A^{\mathcal{S}}$ satisfies the following two conditions:

1. $A \preceq A^{\mathcal{S}}$ for any $A \in \mathcal{A}$;
2. if an increasing predictable process $\widehat{A}$ satisfies $A \preceq \widehat{A}$ for any $A \in \mathcal{A}$, then $A^{\mathcal{S}} \preceq \widehat{A}$ holds.

We can prove the existence of an upper variation process and its integrability. The following lemma will be essential in proving the main theorem in this section. Note that Lemma 2.1 of Föllmer and Kramkov [13] proved a result very similar to the following lemma.

Lemma 6.5. *Let $Q \in \mathcal{P}(\mathcal{S})$. We have the following three assertions:*

(a) *An upper variation process $A^{\mathcal{S}}$ of $(\mathcal{A}, \preceq)$ exists.*

(b) *There exists an increasing sequence $(A^n)_{n\geq 1}$ of $(\mathcal{A}, \preceq)$ such that $A^n_t \uparrow A^{\mathcal{S}}_t$ Q-a.s. for any $t \in [0,T]$.*

(c) $E_Q[A^{\mathcal{S}}_T] < \infty$.

Proof. We prove this lemma in a similar way to Theorem A.32 of Föllmer and Schied [16].

Step 1. Since $Q \in \mathcal{P}(\mathcal{S})$, there exists an increasing predictable process A^Q such that $G(\vartheta) - A^Q$ is a Q-supermartingale for any $\vartheta \in \Theta^{\mathcal{S}}$. Taking $\vartheta \equiv 0$, $-A^Q$ itself is a Q-supermartingale. That is, $E_Q[A^Q_T] < \infty$. Since $A^Q - A^\vartheta$ is increasing by the proof of Lemma 6.2, we have $E_Q[A^\vartheta_T] \le E_Q[A^Q_T] < \infty$ for any $\vartheta \in \Theta^{\mathcal{S}}$.

Step 2. Let $\mathcal{B} \subset \mathcal{A}$ be a countable set. Thus, we can denote $\mathcal{B} = \{A^1, A^2, \dots\}$. We then define inductively $\widetilde{A}^1 := A^1$ and $\widetilde{A}^k := A^k \vee \widetilde{A}^{k-1}$ for $k \ge 2$. Clearly, we have $\widetilde{A}^1 \preceq \widetilde{A}^2 \preceq \cdots$, and $\widetilde{A}^k \in \mathcal{A}$ for any $k \ge 1$. Let $\widetilde{\mathcal{B}} := \{\widetilde{A}^1, \widetilde{A}^2, \dots\}$. Considering $\widetilde{\mathcal{B}}$ instead of $\mathcal{B}$, we may assume that every countable subset $\mathcal{B} = \{A^1, A^2, \dots\}$ has the monotonicity in the sense of that $A^1 \preceq A^2 \preceq \cdots$.

Now, we define, for any countable set $\mathcal{B}$, $A^{\mathcal{B}}_t := \lim_{n\to\infty} A^n_t$ for any $t \in [0,T]$. The monotone convergence theorem implies that $E_Q[A^{\mathcal{B}}_t] = \lim_{n\to\infty} E_Q[A^n_t]$ for any $t \in [0,T]$. Note that the stochastic process $A^{\mathcal{B}}$ is an increasing predictable process. Denoting

$$\alpha := \sup\{E_Q[A^{\mathcal{B}}_T] \mid \mathcal{B} \subset \mathcal{A} \text{ countable}\}, \tag{6.2}$$

α is finite by Step 1. We take a sequence $(\mathcal{B}^n)_{n\ge1}$ of countable subsets of $\mathcal{A}$ so that $E_Q[A^{\mathcal{B}^n}_T] \uparrow \alpha$ as $n \to \infty$. Letting $\mathcal{B}^* := \cup_{n=1}^\infty \mathcal{B}^n$, $\mathcal{B}^*$ is countable and satisfies $E_Q[A^{\mathcal{B}^*}_T] = \alpha$. Without loss of generality, $\mathcal{B}^*$ is denoted by $\{A^{*,1}, A^{*,2}, \dots\}$ such that $A^{*,1} \preceq A^{*,2} \preceq \cdots$ and $A^{*,n}_t \to A^{\mathcal{B}^*}_t$ Q-a.s. as $n \to \infty$ for any $t \in [0,T]$.

Step 3. In order to complete the proof of Lemma 6.5, we have only to prove that $A^{\mathcal{B}^*}$ is an upper variation process of $(\mathcal{A}, \preceq)$. First, we shall prove that $A \preceq A^{\mathcal{B}^*}$ for any $A \in \mathcal{A}$.

For $\varepsilon > 0$, $A \in \mathcal{A}$, and $0 \le t_1 < t_2 \le T$, we define

$$D(t_1, t_2; \varepsilon, A) := \{(A_{t_2} - A_{t_1}) - (A^{\mathcal{B}^*}_{t_2} - A^{\mathcal{B}^*}_{t_1}) > \varepsilon\}.$$

Assume that there exist $\varepsilon > 0$, $\widehat{A} \in \mathcal{A}$, and $0 \le t_1 < t_2 \le T$ such that $Q(D(t_1, t_2; \varepsilon, \widehat{A})) > 0$. Thus, we have $Q(\{A^{\mathcal{B}^*}_{t_1} - A^{*,N}_{t_1} < \varepsilon/2\} \cap D(t_1, t_2; \varepsilon, \widehat{A})) > 0$ for any sufficiently large number N, where $(A^{*,n})_{n\ge1}$ is the sequence of increasing processes that appeared in Step 2; that is, $A^{*,n}$ converges increasingly to $A^{\mathcal{B}^*}$ as n tends to ∞. For such an N, denote $\overline{A} := \widehat{A} \vee A^{*,N}$. Then, we have $Q(\overline{A}_{t_2} > A^{\mathcal{B}^*}_{t_2}) > 0$. Defining $\overline{A}^k := \overline{A} \vee A^{*,N+k}$ for $k \ge 1$, we have $E_Q[A^{\mathcal{B}^*}_T] < \lim_{k\to\infty} E_Q[\overline{A}^k_T]$. That is, there exists an $A' \in \mathcal{A}$ such that $E_Q[A^{\mathcal{B}^*}_T] < E_Q[A'_T]$. Letting $\mathcal{B}' := \mathcal{B}^* \cup \{A'\}$, we have $E_Q[A^{\mathcal{B}'}_T] > E_Q[A^{\mathcal{B}^*}_T] = \alpha$, which contradicts the definition of α, since $\mathcal{B}'$ is countable. Thus, $Q(D(t_1, t_2; \varepsilon, A)) = 0$ for any $\varepsilon > 0$, any $A \in \mathcal{A}$, and any $0 \le t_1 < t_2 \le T$, that is, $A \preceq A^{\mathcal{B}^*}$ for any $A \in \mathcal{A}$.

Step 4. Let B be an increasing predictable process such that $A \preceq B$ for any $A \in \mathcal{A}$. We prove that $A^{\mathcal{B}^*} \preceq B$ as the last step of the proof. Defining

$$D_{t_1,t_2} := \{A^{\mathcal{B}^*}_{t_2} - A^{\mathcal{B}^*}_{t_1} > B_{t_2} - B_{t_1}\}, \quad 0 \le t_1 < t_2 \le T,$$

we assume that $Q(D_{t_1,t_2}) > 0$ for some $0 \le t_1 < t_2 \le T$.

Denoting

$$\widetilde{B}_t := \begin{cases} A^{\mathcal{B}^*}_t & \text{if } t \le t_1, \\ A^{\mathcal{B}^*}_{t_1} + B_t - B_{t_1} & \text{if } t_1 < t \le t_2, \\ A^{\mathcal{B}^*}_{t_1} + B_{t_2} - B_{t_1} + A^{\mathcal{B}^*}_t - A^{\mathcal{B}^*}_{t_2} & \text{if } t > t_2, \end{cases}$$

we have

$$Q(A^{\mathcal{B}^*}_{t_2} > \widetilde{B}_{t_2}) = Q(D_{t_1,t_2}) > 0. \tag{6.3}$$

Note that $A \preceq A^{\mathcal{B}^*}$ and $A \preceq B$ for any $A \in \mathcal{A}$. Thus, $A \preceq \widetilde{B}$ for any $A \in \mathcal{A}$. This fact, together with (6.3), contradicts that $A^{\mathcal{B}^*}_t = \lim_{n\to\infty} A^{*,n}_t$ for any $t \in [0,T]$, and each $A^{*,n}$ belongs to $\mathcal{A}$. Hence, $Q(D_{t_1,t_2}) = 0$ for any $0 \le t_1 < t_2 \le T$, that is, $A^{\mathcal{B}^*} \preceq B$. ■

From the viewpoint of Lemma 6.5, we obtain the following proposition. In addition to this, we introduce one corollary, which is proved immediately by the proof of Proposition 6.6.

Proposition 6.6. *Let $Q \in \mathcal{P}^\Psi$. Then, $Q \in \mathcal{P}(\mathcal{S})$ if and only if*

(1) *$G(\vartheta)$ is a special semimartingale under Q for any $\vartheta \in \Theta^{\mathcal{S}}$, and*

(2) *an upper variation process $A^{\mathcal{S}}$ exists with $E_Q[A^{\mathcal{S}}_T] < \infty$.*

Proof. The "only if" part has been proved by Lemmas 6.2 and 6.5(c). We prove the "if" part. Note that, when condition (1) holds, the ordered set $(\mathcal{A}, \preceq)$ is well defined; that is, condition (2) becomes meaningful.

Fix $\vartheta \in \Theta^{\mathcal{S}}$ arbitrarily. There exists then a constant $c > 0$ such that, for any $t \in [0,T]$, $G_t(\vartheta) \ge -cW$ and $cW \in L^1(Q)$, since $Q \in \mathcal{P}^\Psi$. Let $(\tau_n)_{n\ge1}$ be a localizing sequence of M^ϑ. For any $0 \le t_1 < t_2 \le T$, we have

$$\begin{aligned}
E_Q[G_{t_2}(\vartheta) - A^{\mathcal{S}}_{t_2}|\mathcal{F}_{t_1}] &= E_Q[\liminf_{n\to\infty}(M^\vartheta_{t_2\wedge\tau^n} + A^\vartheta_{t_2\wedge\tau^n}) - A^{\mathcal{S}}_{t_2}|\mathcal{F}_{t_1}] \\
&\le \liminf_{n\to\infty} E_Q[(M^\vartheta_{t_2\wedge\tau^n} + A^\vartheta_{t_2\wedge\tau^n}) - A^{\mathcal{S}}_{t_2}|\mathcal{F}_{t_1}] \\
&= \liminf_{n\to\infty}\left(M^\vartheta_{t_1\wedge\tau^n} + E_Q[A^\vartheta_{t_2\wedge\tau^n} - A^{\mathcal{S}}_{t_2}|\mathcal{F}_{t_1}]\right) \\
&\le M^\vartheta_{t_1} + E_Q[A^\vartheta_{t_2} - A^{\mathcal{S}}_{t_2}|\mathcal{F}_{t_1}] \\
&\le M^\vartheta_{t_1} + A^\vartheta_{t_1} - A^{\mathcal{S}}_{t_1} = G_{t_1}(\vartheta) - A^{\mathcal{S}}_{t_1}.
\end{aligned}$$

The first inequality is given by the Fatou lemma, the second equality the martingale property of the stopped process $(M^\vartheta)^{\tau_n}$, the second inequality the increasing property of A^ϑ, and the last inequality the fact that $A^\vartheta \preceq A^{\mathcal{S}}$, respectively. Hence, $Q \in \mathcal{P}(\mathcal{S})$. ■

Remark 6.7. Although we treat A^ϑ as an increasing process in the above proof, we can extend it to the case where A^ϑ is not increasing.

Corollary 6.8. *Let $Q \in \mathcal{P}^\Psi$. Then, $Q \in \mathcal{P}(\mathcal{S})$ if and only if $A^{\mathcal{S}}$ exists and $G(\vartheta) - A^{\mathcal{S}}$ is a Q-supermartingale for any $\vartheta \in \Theta^{\mathcal{S}}$.*

6.2. Main theorem.

At last, we state the main theorem of this section. By using an upper-variation process, the representation of the first term of the minimal penalty function will be given.

Theorem 6.9. *Let Θ be given by $\Theta^{\mathcal{S}}$. Suppose that $\mathcal{P}(\mathcal{S}) \cap \mathcal{P}^\Psi \ne \emptyset$. For any $Q \in \mathcal{P}^\Psi$, we have then*

$$\sup_{X^1\in\mathcal{A}^1} E_Q[-X^1] = \begin{cases} E_Q[A^{\mathcal{S}}_T] & \text{if } Q \in \mathcal{P}(\mathcal{S}), \\ +\infty & \text{if } Q \notin \mathcal{P}(\mathcal{S}). \end{cases} \tag{6.4}$$

Hence, we have

$$\rho_l(X) = \max_{Q\in\mathcal{P}(\mathcal{S})\cap\mathcal{P}^\Psi}\left\{E_Q[-X] - E_Q[A^{\mathcal{S}}_T] - \inf_{\lambda>0}\frac{1}{\lambda}\left\{\delta + E\left[\Psi\left(\lambda\frac{dQ}{dP}\right)\right]\right\}\right\}$$

for any $X \in M^\Phi$.

Proof. *Step* 1. Let $Q \in \mathcal{P}(\mathcal{S})$. Corollary 6.8 implies that $E_Q[G_T(\vartheta)] \le E_Q[A^{\mathcal{S}}_T]$ for any $\vartheta \in \Theta^{\mathcal{S}}$. Thus, we have

$$\sup_{\vartheta \in \Theta^{\mathcal{S}}} E_Q[G_T(\vartheta)] \le E_Q[A^{\mathcal{S}}_T]. \tag{6.5}$$

Let $(\vartheta^n)_{n \ge 1} \subset \Theta^{\mathcal{S}}$ be a sequence such that the corresponding sequence A^{ϑ^n} is increasing in the order $\preceq$, and $A^{\vartheta^n}_t \uparrow A^{\mathcal{S}}_t$ Q-a.s. for any $t \in [0,T]$. The existence of such a sequence is ensured by Lemma 6.5(b). Let $(\tau^n_k)_{k \ge 1}$ be a localizing sequence of M^{ϑ^n}. Then, $A^{\vartheta^n}_{T \wedge \tau^n_k} \uparrow A^{\vartheta^n}_T$ Q-a.s. as $k \to \infty$. Indeed, this convergence holds in $L^1(Q)$, too, because $A^{\mathcal{S}}_T \in L^1(Q)$. Take $\varepsilon > 0$ arbitrarily. Then, for each $n \ge 1$, we can take a sufficiently large number $k(n)$ to satisfy that $E_Q[A^{\vartheta^n}_T - A^{\vartheta^n}_{T \wedge \tau^n_{k(n)}}] < \varepsilon/2$. Denoting $\sigma^n := \tau^n_{k(n)}$, we have $E_Q[A^{\mathcal{S}}_T - A^{\vartheta^n}_{T\wedge\sigma^n}] < \varepsilon$ for any sufficiently large number n, since $E_Q[A^{\mathcal{S}}_T - A^{\vartheta^n}_T] < \varepsilon/2$ holds for any sufficiently large number n. Noting that $G(\vartheta)^{\sigma^n} \in \mathcal{S}$ because the predictable convexity of $\mathcal{S}$, we obtain, for such a sufficiently large number n,

$$\sup_{\vartheta \in \Theta^{\mathcal{S}}} E_Q[G_T(\vartheta)] \ge E_Q[G_T(\vartheta^n)^{\sigma^n}] = E_Q[A^{\vartheta^n}_{T\wedge\sigma^n}] > E_Q[A^{\mathcal{S}}_T] - \varepsilon.$$

By the arbitrariness of ε, we can conclude that $\sup_{\vartheta \in \Theta^{\mathcal{S}}} E_Q[G_T(\vartheta)] \ge E_Q[A^{\mathcal{S}}_T]$. This, together with (6.5) and Lemma 5.1, implies that (6.4) holds for any $Q \in \mathcal{P}(\mathcal{S})$.

Step 2. We consider the case of $Q \notin \mathcal{P}(\mathcal{S})$. We divide it into three cases from the viewpoint of Proposition 6.6.

Case 1. Consider the case where condition (1) in Proposition 6.6 holds, $A^{\mathcal{S}}$ exists, and $E_Q[A^{\mathcal{S}}_T] = +\infty$. Assuming that α in (6.2) is finite, Steps 2–4 in the proof of Lemma 6.5 hold. We have then $E_Q[A^{\mathcal{S}}_T] = \alpha < \infty$, which is a contradiction. Hence, α must be $+\infty$. Then, for any $M > 0$, there exists a countable subset $\mathcal{B}^M$ of $\mathcal{A}$ such that $E_Q[A^{\mathcal{B}^M}_T] > M$. Thus, there exists an $A^M \in \mathcal{B}^M$ such that $E_Q[A^M_T] > M$. Denoting by ϑ^M the corresponding element of $\Theta^{\mathcal{S}}$ to A^M, and by $(\tau^M_k)_{k\ge1}$ a localizing sequence of M^{ϑ^M}, we have $G(\vartheta^M)^{\tau^M_k} \in \mathcal{S}$ for any $k \ge 1$. Thus, for any $k \ge 1$, the following holds:

$$\sup_{X^1 \in \mathcal{A}^1} E_Q[-X^1] = \sup_{\vartheta \in \Theta^{\mathcal{S}}} E_Q[G_T(\vartheta)] \ge E_Q[G_T(\vartheta^M)^{\tau^M_k}] = E_Q[A^M_{T\wedge\tau^M_k}].$$

Note that $A^{\vartheta^M} = A^M$. As $k \to \infty$,

$$\sup_{X^1 \in \mathcal{A}^1} E_Q[-X^1] \ge \lim_{k\to\infty} E_Q[A^M_{T\wedge\tau^M_k}] = E_Q[A^M_T] > M.$$

By the arbitrariness of M, we obtain $\sup_{X^1 \in \mathcal{A}^1} E_Q[-X^1] = +\infty$.

Case 2. We treat the case where condition (1) in Proposition 6.6 holds and $A^{\mathcal{S}}$ does not exist. Assuming that α in (6.2) is finite, Steps 2–4 in Lemma 6.5 hold. Thus, $A^{\mathcal{S}}$ exists, which is a contradiction. Letting $\alpha = +\infty$, the same argument as Case 1 provides that $\sup_{X^1 \in \mathcal{A}^1} E_Q[-X^1] = +\infty$.

Case 3. Consider the case where condition (1) in Proposition 6.6 does not hold. We can take some $\vartheta \in \Theta^{\mathcal{S}}$ such that $G(\vartheta)$ is not a special semimartingale under Q. Remark that $G(\vartheta)$ is a semimartingale under Q by Theorem VII.45 of [12].

Denote $R^\vartheta_t := \sup_{s\in[0,t]}|G_s(\vartheta)|$. Theorem VII.25 of [12] says that R^ϑ is not locally Q-integrable. Now, for any $n\geq 1$, we define $\tau^n := \inf\{t\geq 0 | |G_t(\vartheta)|\geq n\}\wedge T$. Then, there exists an $N\geq 1$ such that $E_Q[R^\vartheta_{\tau^N}] = +\infty$. Let $A^N := \{|G_t(\vartheta)| < N \text{ for any } t\in[0,T]\}$. Then, $R^\vartheta_{\tau^N}\leq N$ on A^N. On the other hand, $R^\vartheta_{\tau^N} = |G_{\tau^N}(\vartheta)|$ on $(A^N)^c$.

Hence, we have

$$\begin{aligned}+\infty = E_Q[R^\vartheta_{\tau^N}] &= E_Q[1_{A^N}R^\vartheta_{\tau^N} + 1_{(A^N)^c}R^\vartheta_{\tau^N}]\\ &\leq NQ(A^N) + E_Q[|G_{\tau^N}(\vartheta)|],\end{aligned}$$

which means $E_Q[|G_{\tau^N}(\vartheta)|] = +\infty$. Since $G(\vartheta)^{\tau^N}\in\mathcal{S}$, there exists a $c>0$ such that $G_t(\vartheta)^{\tau^N}\geq -cW$ for any $t\in[0,T]$. Thus, since $W\in L^1(Q)$ and $|G_T(\vartheta)^{\tau^N}|1_{\{G_T(\vartheta)^{\tau^N}\leq 0\}}\leq cW$, we have

$$\begin{aligned}\sup_{X^1\in\mathcal{A}^1} E_Q[-X^1] &\geq E_Q[G_T(\vartheta)^{\tau^N}] \geq E_Q[|G_T(\vartheta)^{\tau^N}|1_{\{G_T(\vartheta)^{\tau^N}>0\}} - cW]\\ &\geq E_Q[|G_T(\vartheta)^{\tau^N}| - 2cW] = +\infty.\end{aligned}$$

Hence, (6.4) holds. This completes the proof of Theorem 6.9. ■

6.3. Examples. In this subsection, we shall introduce two examples of $\Theta^{\mathcal{S}}$. We aim to illustrate examples of nonlocally bounded asset price processes. Thus, we treat exponential compound Poisson processes as typical cases of continuous time asset price processes with jumps. Furthermore, we shall calculate upper-variation processes for the first example. For simplicity, we consider only the case of one-dimensional pure jump type models.

Let N be a Poisson process with parameter $\lambda>0$. Set $T_j = 0$ for $j=0,-1$. For $j\geq 1$, the jth jump time of N is denoted by T_j. Let $(Y_j)_{j\geq 1}$ be a sequence of i.i.d. $(-1,\infty)$-valued random variables which are independent of N, and $Y_0 = 0$. Define a compound Poisson process Z as $Z_t := \sum_{i=0}^{N_t} Y_i$. The asset price process S is given by $S_t = S_0\mathcal{E}(Z)_t$, where $\mathcal{E}$ means the stochastic exponential, and $S_0>0$. In other words, S is represented as $S_t = S_0\prod_{i=0}^{N_t}(1+Y_i)$, which is a positive semimartingale.

Before stating the first example, we need some preparations.

Assumption 6.10. We assume the existence of W satisfying the following conditions:

1. $W\in M^\Phi$;
2. $W\geq 1+\sup\{|G_t(h)| : h \text{ is a } [0,1]\text{-valued predictable process, and } t\in[0,T]\}$.

Note that such a W is suitable.

Proposition 6.11. *Under Assumption 6.10, we fix two nonnegative constants k_1 and k_2 and define*

$$\Theta_{k_1,k_2} := \{\vartheta\in L(S) | -k_1\leq\vartheta\leq k_2\}.$$

Then, Θ_{k_1,k_2} is predictably convex and a subset of Θ^W.

Proof. We have only to prove $\Theta_{k_1,k_2}\subset\Theta^W$. Taking a $\vartheta\in\Theta_{k_1,k_2}$, there exists a $[0,1]$-valued predictable process h such that $\vartheta = -k_1h + k_2(1-h)$. Then, for any $t\in[0,T]$, we have $G_t(\vartheta) = -k_1G_t(h) + k_2G_t(1-h)\geq -k_1W - k_2W$. This completes the proof of Proposition 6.11. ■

This proposition implies that, under Assumption 6.10, Θ_{k_1,k_2} becomes an example of $\Theta^{\mathcal{S}}$.

Example 6.12. We consider the case where $\Phi(x) = x^2/2$. Suppose that the random variable Y_1 satisfies $E[Y_1^2] < \infty$. Taking

$$W := 1 + \sup\{|G_t(h)| : h \text{ is a } [0,1]\text{-valued predictable process, and } t \in [0,T]\},$$

we prove $W \in M^\Phi$. Since we have, for any $[0,1]$-valued predictable process h,

$$G_t^2(h) \le N_t \sum_{i=0}^{N_t} h_{T_i}^2 (S_{T_i} - S_{T_{i-1}})^2 \le N_t \sum_{i=0}^{N_t} (S_{T_i} - S_{T_{i-1}})^2 \le 2N_t \sum_{i=0}^{N_t} S_{T_i}^2,$$

it suffices to prove that $E[N_T \sum_{i=0}^{N_T} S_{T_i}^2] < \infty$. Denoting $\sigma := E[(1+Y_1)^2] > 0$, we have

$$\begin{aligned}
E\left[N_T \sum_{i=0}^{N_T} S_{T_i}^2\right] &= \sum_{k=0}^{\infty} E\left[k \sum_{i=0}^{k} S_{T_i}^2 \middle| N_T = k\right] P(N_T = k) \\
&= \sum_{k=0}^{\infty} k \sum_{i=0}^{k} E\left[S_0^2 \prod_{j=0}^{i} (1+Y_j)^2\right] P(N_T = k) \\
&= S_0^2 \sum_{k=0}^{\infty} k \sum_{i=0}^{k} \sigma^i \frac{(\lambda T)^k}{k!} e^{-\lambda T} \le S_0^2 \sum_{k=0}^{\infty} k \sum_{i=0}^{k} (\sigma+1)^i \frac{(\lambda T)^k}{k!} e^{-\lambda T} \\
&= S_0^2 \sum_{k=1}^{\infty} \frac{(\sigma+1)^{k+1} - 1}{\sigma} \frac{(\lambda T)^k}{(k-1)!} e^{-\lambda T} \le S_0^2 \frac{(\sigma+1)^2 \lambda T}{\sigma} e^{\sigma \lambda T} < \infty.
\end{aligned}$$

By Proposition 6.11, Θ_{k_1,k_2} for any $k_1, k_2 \ge 0$ is predictably convex. Denoting $\widetilde{Z}_t := Z_t - E[Y_1]\lambda t$, we have that $\widetilde{Z}$ is a P-martingale. For any $\vartheta \in \Theta_{k_1,k_2}$, we have $G_t(\vartheta) = \int_0^t \vartheta_s S_{s-} dZ_s = \int_0^t \vartheta_s S_{s-} d\widetilde{Z}_s + E[Y_1]\lambda \int_0^t \vartheta_s S_{s-} ds$. Note that the process S is positive. Thus, by the definition of upper variation processes, we can see that P belongs to $\mathcal{P}(\mathcal{S})$ and its upper-variation process $A_t^{\mathcal{S}}$ is given by

$$A_t^{\mathcal{S}} = \begin{cases} E[Y_1]\lambda k_2 \int_0^t S_{s-} ds & \text{if } E[Y_1] \ge 0, \\ -E[Y_1]\lambda k_1 \int_0^t S_{s-} ds & \text{if } E[Y_1] < 0. \end{cases}$$

In the next example, we consider the exponential loss function case.

Example 6.13. Suppose that Y_1 belongs to the Orlicz heart M^Φ, where $\Phi(x) = e^x - 1$. For instance, $1 + Y_1$ has the same distribution as $|Y|$, where Y is a normal distributed random variable. For any predictable process h, we have $G_t(\vartheta) = \sum_{j \ge 0, T_j \le t} \vartheta_{T_j} S_{T_{j-1}} Y_j$. If ϑ belongs to the set

$$F := \{\vartheta \in L(S) | \ \vartheta_{T_j} S_{T_{j-1}} \text{ is } [0,1]\text{-valued for any } j \ge 0\},$$

then $|G_t(\vartheta)| \le \sum_{T_j \le T} |Y_j|$ for any $t \in [0,T]$. Let W be defined by

$$W := 1 + \sup\{|G_t(\vartheta)| : \vartheta \in F, t \in [0,T]\}. \tag{6.6}$$

We have

$$E\left[\exp\left\{c\sum_{T_j\le T}|Y_j|\right\}\right]=\sum_{k=0}^{\infty}E\left[\prod_{j=0}^{k}e^{c|Y_j|}\middle|N_T=k\right]P(N_T=k)$$
$$=\sum_{k=0}^{\infty}E^k\left[e^{c|Y_1|}\right]\frac{(\lambda T)^k}{k!}e^{-\lambda T}<\infty,$$

from which $W\in M^\Phi$ follows. We can define the set Θ^W associated with (6.6). Defining, for two nonnegative constants k_1 and k_2,

$$\Theta^{\mathcal{S}}:=\{\vartheta\in L(S)|-k_1\le\vartheta_{T_j}S_{T_{j-1}}\le k_2\text{ for any }j\ge 1\},$$

we can see that $\Theta^{\mathcal{S}}$ is a predictably convex subset of Θ^W by the same argument as Proposition 6.11.

Acknowledgments. The author would like to thank Martin Schweizer for his valuable comments and his hospitality during the FY2007 Researcher Exchange Program between JSPS and SNSF. Moreover, the author also thanks Marco Frittelli and two anonymous referees for their helpful comments and suggestions.

REFERENCES

[1] C. D. Aliprantis and K. C. Border, *Infinite Dimensional Analysis*, 3rd ed., Springer, New York, 2005.

[2] A. E. Bernardo and O. Ledoit, *Gain, loss, and asset pricing*, J. Political Economy, 108 (2000), pp. 144–172.

[3] S. Biagini and M. Frittelli, *Utility maximization in incomplete markets for unbounded processes*, Finance Stoch., 9 (2005), pp. 493–517.

[4] S. Biagini and M. Frittelli, *A unified framework for utility maximization problems: An Orlicz space approach*, Ann. Appl. Probab., 18 (2008), pp. 929–966.

[5] S. Biagini and M. Frittelli, *On the extension of the Namioka-Klee theorem and on the Fatou property for risk measures*, in Optimality and Risk: Modern Trends in Mathematical Finance. The Kabanov Festschrift, Delbaen, Rasonyi, and Stricker, eds., Springer, New York, 2009, pp. 1–28.

[6] T. Björk and I. Slinko, *Towards a general theory of good-deal bounds*, Rev. Finance, 10 (2006), pp. 221–260.

[7] P. Carr, H. Geman, and D. B. Madan, *Pricing and hedging in incomplete markets*, J. Financial Economics, 62 (2001), pp. 131–167.

[8] A. Černý, *Generalized Sharpe ratios and asset pricing in incomplete markets*, European Finance Review, 7 (2003), pp. 191–233.

[9] A. Černý and S. Hodges, *The theory of good-deal pricing in financial markets*, in Mathematical Finance, Bachelier Congress 2000, Geman, Madan, Pliska, and Vorst, eds., Springer-Verlag, New York, 2002, pp. 175–202.

[10] P. Cheridito and T. Li, *Risk measures on Orlicz hearts*, Math. Finance, 19 (2009), pp. 189–214.

[11] J. H. Cochrane and J. Saá-Requejo, *Beyond arbitrage: Good-deal asset price bounds in incomplete markets*, J. Political Economy, 108 (2000), pp. 79–119.

[12] C. Dellacherie and P. A. Meyer, *Probabilities and Potential B*, North–Holland, Amsterdam, 1982.

[13] H. Föllmer and D. O. Kramkov, *Optional decomposition under constraints*, Probab. Theory Related Fields, 109 (1997), pp. 1–25.

[14] H. Föllmer and P. Leukert, *Efficient hedging: Cost versus shortfall risk*, Finance Stoch., 4 (2000), pp. 117–146.

[15] H. Föllmer and A. Schied, *Convex measures of risk and trading constraints*, Finance Stoch., 6 (2002), pp. 429–447.

[16] H. Föllmer and A. Schied, *Stochastic Finance: An Introduction in Discrete Time*, 2nd ed., de Gruyter Stud. Math., Walter De Gruyter, Berlin, 2004.

[17] M. Frittelli and E. Rosazza-Gianin, *Putting order in risk measures*, J. Banking Finance, 26 (2002), pp. 1473–1486.

[18] S. Jaschke and U. Küchler, *Coherent risk measures and good-deal bounds*, Finance Stoch., 5 (2001), pp. 181–200.

[19] M. Kaina and L. Rüschendorf, *On convex risk measures on L^p-spaces*, Math. Methods Oper. Res., 69 (2009), pp. 475–495.

[20] S. Klöppel and M. Schweizer, *Dynamic indifference valuation via convex risk measures*, Math. Finance, 17 (2007), pp. 599–627.

[21] S. Klöppel and M. Schweizer, *Dynamic utility-based good deal bounds*, Statist. Decisions, 25 (2007), pp. 285–309.

[22] D. O. Kramkov, *Optional decomposition of supermartingales and hedging contingent claims in incomplete security markets*, Probab. Theory Related Fields, 105 (1996), pp. 459–479.

[23] K. Larsen, T. Pirvu, S. Shreve, and R. Tütüncü, *Satisfying convex risk limits by trading*, Finance Stoch., 9 (2004), pp. 177–195.

[24] J. Staum, *Fundamental theorems of asset pricing for good deal bounds*, Math. Finance, 14 (2004), pp. 141–161.

[25] M. Xu, *Risk measure pricing and hedging in incomplete markets*, Ann. Finance, 2 (2006), pp. 51–71.

SIAM J. FINANCIAL MATH.
Vol. 2, pp. 22–54

Jump-Diffusion Risk-Sensitive Asset Management I: Diffusion Factor Model*

Mark Davis† and Sébastien Lleo‡

Abstract. This paper considers a portfolio optimization problem in which asset prices are represented by SDEs driven by Brownian motion and a Poisson random measure, with drifts that are functions of an auxiliary diffusion factor process. The criterion, following earlier work by Bielecki, Pliska, Nagai, and others, is risk-sensitive optimization (equivalent to maximizing the expected growth rate subject to a constraint on variance). By using a change of measure technique introduced by Kuroda and Nagai we show that the problem reduces to solving a certain stochastic control problem in the factor process, which has no jumps. The main result of this paper is to show that the risk-sensitive jump-diffusion problem can be fully characterized in terms of a parabolic Hamilton–Jacobi–Bellman PDE rather than a partial integro-differential equation, and that this PDE admits a classical ($C^{1,2}$) solution.

Key words. stochastic control, risk-sensitive control, jump-diffusion processes, Lévy processes, Hamilton–Jacobi–Bellman partial differential equation, linear parabolic partial differential equation, partial integro-differential equation, policy improvement

AMS subject classifications. 91G10, 91G80, 93E20, 93E11

DOI. 10.1137/090760180

1. Introduction.

In this article, we consider a finite time jump-diffusion version of the risk-sensitive asset management problem of Bielecki and Pliska [5]. Fundamentally, our main result is to show that the resulting stochastic control problem can be fully characterized by a parabolic Hamilton–Jacobi–Bellman (HJB) PDE rather than a partial integro-differential equation (PIDE) and that this PDE admits a classical ($C^{1,2}$) solution.

Risk-sensitive control is a generalization of classical stochastic control in which the degree of risk aversion or risk tolerance of the optimizing agent is explicitly parameterized in the objective criterion and influences the outcome of the optimization directly. In risk-sensitive control, the decision maker's objective is to select a control policy $h(t)$ to maximize the criterion

$$J(x,t,h;\theta) := -\frac{1}{\theta}\ln \mathbf{E}\left[e^{-\theta F(x,h)}\right], \tag{1.1}$$

where t is the time, x is the state variable, F is a given reward function, and the risk sensitivity $\theta > 0$ is an exogenous parameter representing the decision maker's degree of risk aversion. A

*Received by the editors May 27, 2009; accepted for publication (in revised form) October 24, 2010; published electronically January 13, 2011. This research was supported by the EPSRC under Mathematical and Industry grant EP/F035578/1.
http://www.siam.org/journals/sifin/2/76018.html
†Department of Mathematics, Imperial College London, London SW7 2AZ, England (mark.davis@imperial.ac.uk).
‡Finance Department, Reims Management School, 59 rue Pierre Taittinger, 51100 Reims, France (sebastien.lleo@reims-ms.fr).

Taylor expansion of the previous expression around $\theta = 0$ evidences the vital role played by the risk sensitivity parameter:

$$J(x,t,h;\theta) = \mathbf{E}\left[F(x,t,h)\right] - \frac{\theta}{2}\mathbf{Var}\left[F(x,t,h)\right] + O(\theta^2). \tag{1.2}$$

This criterion amounts to maximizing $\mathbf{E}\left[F(x,t,h)\right]$ subject to a penalty for variance. For a general reference, see Whittle [40]. Much of the recent literature concerns the infinite time horizon problem:

$$J_\infty(x,h;\theta) := \liminf_{t\to\infty} -\frac{1}{\theta}t^{-1}\ln \mathbf{E}\left[e^{-\theta f(x,h)}\right]. \tag{1.3}$$

This is interesting from a theoretical perspective but is not applicable to practical asset management because of the nonuniqueness of controls. Optimality in this sense is a "tail property": if $h^*(t)$ is optimal, then so is $\tilde{h}(t) = h^*(t)1_{t>T} + h(t)1_{t\le T}$ for any arbitrary process $h(t)$ and time $T > 0$. Of course, near-term decisions are the ones that are of primary importance to investment managers.

In the past decade, the applications of risk-sensitive control to asset management have flourished. Risk-sensitive control was first applied to solve financial problems by Lefebvre and Montulet [29] in a corporate finance context and by Fleming [15] in a portfolio selection context. However, Bielecki and Pliska [5] were the first to apply the continuous time risk-sensitive control as a practical tool that could be used to solve "real world" portfolio selection problems. They considered a long-term asset allocation problem and proposed the logarithm of the investor's wealth as a reward function, so that the investor's objective is to maximize the risk-sensitive (log) return of his/her portfolio or alternatively to maximize a function of the power utility (HARA) of terminal wealth. They derived the optimal control and solved the associated HJB PDE under the restrictive assumption that the asset and factor noise are uncorrelated. This assumption is unrealistic, and it was later relaxed (see [8]). The contribution of Bielecki and Pliska to the field is immense: they studied the economic properties of the risk-sensitive asset management criterion (see [7]), extended the asset management model into an intertemporal CAPM [8], worked on transaction costs [6] and numerical methods [4], and considered factors driven by a CIR model [9]. A major contribution to the mathematical theory was made by Kuroda and Nagai [26], who introduced an elegant solution method based on a change of measure argument which transforms the risk-sensitive control problem into a linear exponential of quadratic regulator. They solved the associated HJB PDE over a finite time horizon and then studied the properties of the ergodic HJB PDE. Recently, Davis and Lleo [12] applied this change of measure technique to solve, at a finite and an infinite horizon, a benchmarked investment problem in which an investor selects an asset allocation to outperform a given financial benchmark. The problem we consider is also related to the vast literature on HARA utility maximization that has flourished in the past 50 years. This literature includes a number of references related to risk-sensitive control, such as works by Fleming and Sheu [17], [18], [19] or Hansen and Sargent [22] in the context of robust control.

Risk-sensitive asset management theory was originally set in a world of diffusion dynamics where randomness is modelled using correlated Brownian motions. To our knowledge, the only attempt to extend the risk-sensitive asset management theory from a diffusion to a jump-diffusion setting was made by Wan [39], who briefly sketched a jump-diffusion extension of

Bielecki and Pliska's original infinite horizon risk-sensitive asset management model [5]. Wan's treatment is, however, restrictive, as it considers only a single Poisson process-driven jump per asset and assumes that the underlying valuation factor risks and asset risks are uncorrelated. Our paper addresses these two limitations. The setting of our control problem, which takes place within a finite time horizon, allows for both infinite activity jumps in asset prices and for a correlation structure between factor risks and asset risks. To solve this control problem we extend Kuroda and Nagai's powerful change of measure technique to account for the jumps. One of the difficulties we face in extending this technique is proving that the optimal control is admissible, as this requires showing that the Doléans exponential (2.9) associated with this control is a martingale. In a pure diffusion setting, this would follow easily from the Kamazaki condition or the Novikov condition. However, when the Doléans exponential does not have continuous paths, as is the case in a jump-diffusion setting, proving that it is indeed a martingale is more difficult, as only weaker partial results exist. This question is addressed in the appendix of the present paper.

In this paper the asset price processes are modelled as jump diffusions whose growth rates are functions of an auxiliary "factor" process $X(t)$ which satisfies a linear diffusion SDE. Our main result is that the risk-sensitive jump-diffusion asset management problem is equivalent to an optimal control problem for a *diffusion* process (no jumps) and that the HJB equation for the latter admits a unique classical $C^{1,2}([0,T)\times\mathbb{R}^n)$ solution. Showing the existence and uniqueness of a solution to a risk-sensitive control problem can prove difficult even in a pure diffusion setting. For example, Bensoussan, Freshe, and Nagai [3] had to constrain the behavior of the Hamiltonian in order to prove the existence of a classical solution. Still in a pure diffusion setting, Fleming and Soner (see section V.9 in [20]) proved that the value function is a continuous viscosity solution of the associated HJB PDE but had to assume boundedness of all coefficients and of the derivatives of the reward function. No such strong condition is required to solve the jump-diffusion problem considered in this article. In fact all our assumptions arise naturally from the structure of the risk-sensitive asset management problem. Uniqueness follows from a classical verification argument, while the proof of existence relies on a policy improvement algorithm and on the properties of linear parabolic PDEs.

This paper is organized as follows. We first introduce the general setting of the model in section 2 and define the class of random Poisson measures which will be used to model the jump component of the asset dynamics. In section 3, we formulate the jump-diffusion control problem and introduce the change of measure argument of Kuroda and Nagai [26]. In a pure diffusion case, this is enough to transform the problem into a standard linear exponential of quadratic regulator (LEQR) problem. In our jump-diffusion setting, the change of measure simplifies the problem by associating the HJB PDE given in section 3.3, rather than the expected PIDE containing nonlocal terms, to the value function. It is striking that an optimal control problem for a jump-diffusion model has a solution that is characterized in terms of an HJB PDE and not an HJB PIDE.[1]

Our main result is Theorem 4.3 in section 4. The proof depends on various technical arguments which are given in sections 5 to 7. In section 5, we show the existence of a unique optimal control before addressing two key questions in section 6. First, the admissibility of

[1]See Øksendal and Sulem [34] for a treatment of jump-diffusion control problems.

the optimal control is no longer a priori guaranteed because the Doléans exponential defining the Radon–Nikodym derivative does not have continuous paths. This point is addressed in Propositions 6.3 and 6.4. Second, the risk-sensitive (RS) HJB PDE contains a jump-induced control-dependent integral term: it is no longer possible to find an analytical solution, and the existence of a strong, classical solution is no longer guaranteed. However, should we be able to prove the existence of a classical $C^{1,2}$ solution to the RS HJB PDE, then we can prove uniqueness and resolve the control problem using a straightforward verification theorem, presented in Theorem 6.1 and Corollary 6.2 in section 6.

In section 7, we address the existence and regularity of solutions to the RS HJB PDE. We show, in Theorem 7.2 and Corollary 7.3, that the risk-sensitive jump-diffusion control problem we consider admits a unique classical $C^{1,2}([0,T) \times \mathbb{R}^n)$ solution. Showing the existence and uniqueness of a solution to a risk-sensitive control problem can prove difficult even in a pure diffusion setting. For example, Bensoussan, Frehse, and Nagai [3] had to constrain the behavior of the Hamiltonian in their finite time horizon problem to prove the existence of a classical solution. Still in a pure diffusion setting and over a finite time horizon, Fleming and Soner (see section V.9 in [20]) proved that the value function is a continuous viscosity solution of the associated HJB PDE but had to assume boundedness of all coefficients and of the derivatives of the reward function. No such strong condition is required to solve the jump-diffusion problem considered in this article. In fact all our assumptions arise naturally from the structure of the risk-sensitive asset management problem. We obtain our result by applying an approximation in policy space in a two-step process: first, we show existence on a bounded region and then extend to the unbounded state space. With this result in hand, we have all the ingredients needed for the proof of Theorem 4.3.

Up to this point, we have assumed that the factor process $X(t)$ is directly observed by the controller and therefore represents real economic factors: GDP growth, inflation, the S&P500 index, etc. We may, however, wish to use $X(t)$ as an abstract latent factor, introduced to model volatility of returns, in which case only the prices, and not $X(t)$, will be observed. In section 8, we note that this problem, once adequately reformulated, can be solved using a classical Kalman filter, as in [32], as the jump noise is absent from the dynamics of $X(t)$. While this is from a technical point of view a simple observation, it greatly enhances the applicability of our results.

In a companion paper [14] we consider the case in which there are jumps in both the price and factor processes. There the measure change technique does not remove the jumps and the argument is substantially different.

2. Analytical setting.

2.1. Overview. The growth rates of the assets are assumed to depend on the n factors $X_1(t), \dots, X_n(t)$ which follow the dynamics given in (2.3). As in Kuroda and Nagai's asset-only model, the assets market comprises m risky securities S_i, $i = 1, \dots, m$. In contrast to Kuroda and Nagai, we assume that the money market account process, S_0, is an affine function of the valuation factors, which enables us to easily model a stochastic short-term rate. Let $M := n + m$. Throughout, we will assume that $m > n$; this is needed in connection with the "zero beta" policies introduced in section 7.1.

Let $(\Omega, \{\mathcal{F}_t\}, \mathcal{F}, \mathbb{P})$ be the underlying probability space. On this space is defined an $\mathbb{R}^M$-

valued $(\mathcal{F}_t)$-Brownian motion $W(t)$ with components $W_k(t)$, $k = 1, \dots, M$. Moreover, let N be an $(\mathcal{F}_t)$-Poisson point process on $(0, \infty) \times \mathbf{Z}$, independent of $W(t)$, where $(\mathbf{Z}, \mathcal{B}_{\mathbf{Z}})$ is a given Borel space.[2] Define

$$\mathfrak{Z} := \{U \in \mathcal{B}(\mathbf{Z}), \mathbb{E}[N(t,U)] < \infty \quad \forall t\}. \tag{2.1}$$

Finally, for notational convenience, we fix throughout the paper a set $\mathbf{Z}_0 \in \mathcal{B}_{\mathbf{Z}}$ such that $\nu(\mathbf{Z} \backslash \mathbf{Z}_0) < \infty$ and define

$$\begin{aligned} &\bar{N}(dt,dz) \\ &= \begin{cases} N(dt,dz) - \hat{N}(dt,dz) = N(dt,dz) - \nu(dz)dt =: \tilde{N}(dt,dz) & \text{if } z \in \mathbf{Z}_0, \\ N(dt,dz) & \text{if } z \in \mathbf{Z}\backslash\mathbf{Z}_0. \end{cases} \end{aligned} \tag{2.2}$$

2.2. Factor dynamics. The dynamics of the n factors are expressed by the affine diffusion equation

$$dX(t) = (b + BX(t))dt + \Lambda dW(t), \qquad X(0) = x, \tag{2.3}$$

where $X(t)$ is the $\mathbb{R}^n$-valued factor process with components $X_j(t)$ and $b \in \mathbb{R}^n$, $B \in \mathbb{R}^{n \times n}$, and $\Lambda \in \mathbb{R}^{n \times M}$.

2.3. Asset market dynamics. Let S_0 denote the wealth invested in the money market account with dynamics given by the equation

$$\frac{dS_0(t)}{S_0(t)} = \left(a_0 + A_0^T X(t)\right) dt, \qquad S_0(0) = s_0, \tag{2.4}$$

where $a_0 \in \mathbb{R}$ is a scalar constant, $A_0 \in \mathbb{R}^n$ is an n-element column vector, and throughout the paper x^T denotes the transpose of the matrix or vector x.

Let $S_i(t)$ denote the price at time t of the ith security, with $i = 1, \dots, m$. The dynamics of risky security i can be expressed as

$$\begin{aligned} \frac{dS_i(t)}{S_i(t^-)} &= (a + AX(t))_i dt + \sum_{k=1}^{N} \sigma_{ik} dW_k(t) + \int_{\mathbf{Z}} \gamma_i(z) \bar{N}(dt,dz), \\ S_i(0) &= s_i, \quad i = 1, \dots, m, \end{aligned} \tag{2.5}$$

where $a \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$, $\Sigma := [\sigma_{ij}]$, $i = 1, \dots, m$, $j = 1, \dots, M$, and $\gamma(z) \in \mathbb{R}^m$ satisfies Assumption 1.

Assumption 1. $\gamma(z) \in \mathbb{R}^m$ satisfies

$$-1 \leq \gamma_i^{min} \leq \gamma_i(z) \leq \gamma_i^{max} < +\infty, \qquad i = 1, \dots, m,$$

and

$$-1 \leq \gamma_i^{min} < 0 < \gamma_i^{max} < +\infty, \qquad i = 1, \dots, m,$$

[2] $\mathbf{Z}$ is a Polish space and $\mathcal{B}_{\mathbf{Z}}$ is the Borel σ-field. See Ikeda and Watanabe [23] for a formal definition of the Poisson point process.

for $i = 1, \ldots, m$. Furthermore, define

$$\mathbf{S} := \mathrm{supp}(\nu) \in \mathcal{B}_{\mathbf{Z}}$$

and

$$\tilde{\mathbf{S}} := \mathrm{supp}(\nu \circ \gamma^{-1}) \in \mathcal{B}(\mathbb{R}^m),$$

where $\mathrm{supp}(\cdot)$ denotes the measure's support; then we assume that $\prod_{i=1}^{m}[\gamma_i^{min}, \gamma_i^{max}]$ is the smallest closed hypercube containing $\tilde{\mathbf{S}}$.

In addition, the vector-valued function $\gamma(z)$ satisfies

$$\int_{\mathbf{Z}_0} |\gamma(z)|^2 \nu(dz) < \infty. \tag{2.6}$$

Note that Assumption 1 implies that each asset has, with positive probability, both upward and downward jumps. As will become evident in section 3.2, the effect of this assumption is to bound the space of controls. Relation (2.6) is a standard condition.

Define the set $\mathcal{J}$ as

$$\mathcal{J} := \left\{ h \in \mathbb{R}^m : -1 - h^T \psi < 0 \quad \forall \psi \in \tilde{\mathbf{S}} \right\}, \tag{2.7}$$

and let $\overline{\mathcal{J}}$ be the closure of $\mathcal{J}$. For a given z, the equation $h^T\gamma(z) = -1$ describes a hyperplane in $\mathbb{R}^m$. $\mathcal{J}$ is a bounded open convex subset of $\mathbb{R}^m$.

2.4. Portfolio dynamics. We will need the following assumptions.

Assumption 2. $\Sigma\Sigma^T > 0$.

The effect of this assumption is to prevent redundant assets. For example, we will not able to model in our investment market a share and an option or futures on that share. However, this assumption leaves us free to model a wide range of assets such as shares, bonds, and commodity products as well as related indices.

Assumption 3. $\Lambda\Lambda^T > 0$.

Let $\mathcal{G}_t := \sigma((S(s), X(s)), 0 \leq s \leq t)$ be the sigma-field generated by the security and factor processes up to time t. An *investment strategy* or *control process* is an $\mathbb{R}^m$-valued process with the interpretation that $h_i(t)$ is the fraction of current portfolio value invested in the ith asset, $i = 1, \ldots, m$. The fraction invested in the money market account is then $h_0(t) = 1 - \sum_{i=1}^m h_i(t)$.

Definition 2.1. *An $\mathbb{R}^m$-valued control process $h(t)$ is in class $\mathcal{H}$ if the following conditions are satisfied:*

1. *$h(t)$ is progressively measurable with respect to $\{\mathcal{B}([0,t]) \otimes \mathcal{G}_t\}_{t\geq 0}$ and is càdlàg;*
2. *$\mathbb{P}(\int_0^t |h(s)|^2\, ds < +\infty) = 1\ \forall t > 0$;*
3. *$h^T(t)\gamma(z) > -1\ \forall t > 0, z \in \mathbf{Z}$, a.s. $d\nu$.*

Define the set $\mathcal{K}$ as

$$\mathcal{K} := \left\{ h \in \mathcal{H} : h(t,\omega) \in \mathcal{J} \quad \text{a.e. } dt \times d\mathbb{P} \right\}. \tag{2.8}$$

Lemma 2.2. *Under Assumption 1, a control process $h(t)$ satisfying condition 3 in Definition 2.1 is bounded.*

Proof. The proof of this result is immediate. ■

Definition 2.3. *A control process h is in class $\mathcal{A}(T)$ if the following conditions are satisfied:*

1. $h \in \mathcal{H}$;
2. $\mathbf{E}\chi_T^h = 1$, *where* χ_t^h, $t \in (0,T]$, *is the Doléans exponential defined as*

$$\chi_t^h := \exp\left\{-\theta \int_0^t h(s)^T \Sigma dW_s - \frac{1}{2}\theta^2 \int_0^t h(s)^T \Sigma\Sigma^T h(s) ds \right.$$
$$+ \int_0^t \int_{\mathbf{Z}} \ln\left(1 - G(z, h(s))\right) \tilde{N}(ds, dz)$$
$$\left. + \int_0^t \int_{\mathbf{Z}} \left\{\ln\left(1 - G(z, h(s))\right) + G(z, h(s))\right\} \nu(dz) ds \right\} \tag{2.9}$$

and

$$G(z,h) = 1 - \left(1 + h^T \gamma(z)\right)^{-\theta}. \tag{2.10}$$

We say that a control process h is admissible *if $h \in \mathcal{A}(T)$.*

The proportion invested in the money market account is $h_0(t) = 1 - \sum_{i=1}^m h_i(t)$. Taking this budget equation into consideration, the wealth, $V(t)$, of the investor in response to an investment strategy $h(t) \in \mathcal{H}$ follows the dynamics

$$\frac{dV(t)}{V(t^-)} = \left(a_0 + A_0^T X(t)\right) dt + h^T(t)\left(a - a_0\mathbf{1} + \left(A - \mathbf{1}A_0^T\right) X(t)\right) dt$$
$$+ h^T(t)\Sigma dW_t + \int_{\mathbf{Z}} h^T(t)\gamma(z)\bar{N}(dt, dz),$$

with $\mathbf{1} \in \mathbf{R}^m$ denoting the m-element unit column vector and with $V(0) = v$. Defining $\hat{a} := a - a_0\mathbf{1}$ and $\hat{A} := A - \mathbf{1}A_0^T$, we can express the portfolio dynamics as

$$\frac{dV(t)}{V(t^-)} = \left(a_0 + A_0^T X(t)\right) dt + h^T(t)\left(\hat{a} + \hat{A}X(t)\right) dt + h^T(t)\Sigma dW_t + \int_{\mathbf{Z}} h^T(t)\gamma(z)\bar{N}(dt, dz) \tag{2.11}$$

with initial endowment $V(0) = 0$.

3. Problem setup.

3.1. Optimization criterion.

We will assume that the objective of the investor is to maximize the risk adjusted growth of his/her portfolio of assets over a finite time horizon. In this context, the objective of the risk-sensitive management problem is to find $h^* \in \mathcal{A}(T)$, which maximizes the control criterion

$$J(x,t,h;\theta) := -\frac{1}{\theta} \ln \mathbf{E}\left[e^{-\theta \ln V(t,x,h)}\right]. \tag{3.1}$$

By Itô's lemma, the log of the portfolio value in response to a strategy h is

$$\begin{aligned}\ln V(t) = \ln v &+ \int_0^t \left(a_0 + A_0^T X(s)\right) + h(s)^T \left(\hat{a} + \hat{A}X(s)\right) ds - \frac{1}{2}\int_0^t h(s)^T \Sigma\Sigma^T h(s) ds \\ &+ \int_0^t h(s)^T \Sigma dW(s) \\ &+ \int_0^t \int_{\mathbf{Z}_0} \left\{\ln\left(1 + h(s)^T \gamma(z)\right) - h(s)^T \gamma(z)\right\} \nu(dz) ds \\ &+ \int_0^t \int_{\mathbf{Z}} \ln\left(1 + h(s)^T \gamma(z)\right) \bar{N}(ds, dz). \end{aligned} \tag{3.2}$$

Hence,

$$e^{-\theta \ln V(t)} = v^{-\theta} \exp\left\{\theta \int_0^t g(X_s, h(s)) ds\right\} \chi_t^h, \tag{3.3}$$

where

$$\begin{aligned} g(x,h) = &\frac{1}{2}(\theta+1)\, h^T \Sigma\Sigma^T h - a_0 - A_0^T x - h^T(\hat{a} + \hat{A}x) \\ &+ \int_{\mathbf{Z}} \left\{\frac{1}{\theta}\left[\left(1 + h^T \gamma(z)\right)^{-\theta} - 1\right] + h^T \gamma(z) \mathit{1}_{\mathbf{Z}_0}(z)\right\} \nu(dz) \end{aligned} \tag{3.4}$$

and the Doléans exponential χ_t^h is given by (2.9).

3.2. Change of measure. Let $\mathbb{P}_h$ be the measure on $(\Omega, \mathcal{F}_T)$ defined via the Radon–Nikodym derivative

$$\frac{d\mathbb{P}_h}{d\mathbb{P}} := \chi_T^h. \tag{3.5}$$

For a change of measure to be possible, we must ensure that the following technical condition holds:

$$G(z, h(s)) < 1.$$

This condition is satisfied iff

$$h^T(s)\gamma(z) > -1 \tag{3.6}$$

a.s. $d\nu$, which was already required for $h(t)$ to be in class $\mathcal{H}$ (condition 3 in Definition 2.1). Condition (3.6) is endogenous to the control problem and can be interpreted as a risk management safeguard preventing the investor from investing in some of the portfolios if the jump component of these portfolios could result in the investor's bankruptcy.

Observe that $\mathbb{P}_h$ is a probability measure for $h \in \mathcal{A}(T)$. Then

$$W_t^h = W_t + \theta \int_0^t \Sigma^T h(s) ds$$

is a standard Brownian motion under the measure $\mathbb{P}_h$ and we have (recall the notation defined at (2.2))

$$\begin{aligned}\int_0^t \int_{\mathbf{Z}_0} \tilde{N}^h(ds,dz) &= \int_0^t \int_{\mathbf{Z}_0} N(ds,dz) - \int_0^t \int_{\mathbf{Z}_0} \{1 - G(z,h(s))\}\, \nu(dz)ds \\ &= \int_0^t \int_{\mathbf{Z}_0} N(ds,dz) - \int_0^t \int_{\mathbf{Z}_0} \left\{\left(1 + h^T\gamma(z)\right)^{-\theta}\right\} \nu(dz)ds.\end{aligned}$$

As a result, $X(t)$ satisfies the SDE

$$dX(t) = f\left(X(t), h(t)\right) dt + \Lambda dW_t^h, \qquad t \in [0,T], \tag{3.7}$$

where

$$f(x,h) := \left(b + Bx - \theta\Lambda\Sigma^T h\right). \tag{3.8}$$

We will now introduce the following two auxiliary criterion functions under the measure $\mathbb{P}_h$:

- the auxiliary function directly associated with the risk-sensitive control problem:

$$I(v,x;h;t,T) = -\frac{1}{\theta} \ln \mathbf{E}_{t,x}^h \left[\exp\left\{\theta \int_t^T g(X_s,h(s))ds - \theta \ln v\right\}\right], \tag{3.9}$$

  with $\mathbf{E}_{t,x}[\cdot]$ denoting the expectation taken with respect to the measure $\mathbb{P}_h$ and with initial conditions (t,x);
- the exponentially transformed criterion

$$\tilde{I}(v,x,h;t,T) := \mathbf{E}_{t,x}^h \left[\exp\left\{\theta \int_t^T g(s,X_s,h(s))ds - \theta \ln v\right\}\right], \tag{3.10}$$

  which we will find convenient to use in our derivations.

We have completed our reformulation of the problem under the measure $\mathbb{P}_h$. It is striking that the asset allocation problem with jump-diffusion asset prices reduces to a stochastic control problem for a *diffusion* process, with dynamics (3.7) and reward function (3.9) or (3.10).

3.3. The risk-sensitive control problems under $\mathbb{P}_h$. Let Φ be the value function for the auxiliary criterion function $I(v,x;h;t,T)$. Then Φ is defined as

$$\Phi(t,x) = \sup_{h\in\mathcal{A}(T)} I(v,x;h;t,T). \tag{3.11}$$

We will show that Φ satisfies the HJB PDE

$$\frac{\partial\Phi}{\partial t}(t,x) + \sup_{h\in\mathcal{J}} L_t^h\Phi(t,x) = 0, \qquad (t,x) \in (0,T)\times\mathbb{R}^n, \tag{3.12}$$

where

$$L_t^h\Phi(t,x) = \left(b + Bx - \theta\Lambda\Sigma^T h(s)\right)^T D\Phi \\ + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T D^2\Phi\right) - \frac{\theta}{2}(D\Phi)^T\Lambda\Lambda^T D\Phi - g(x,h), \tag{3.13}$$

and is subject to terminal condition

$$\Phi(T,x) = \ln v, \qquad x \in \mathbb{R}^n. \tag{3.14}$$

Similarly, let $\tilde{\Phi}$ be the value function for the auxiliary criterion function $\tilde{I}(v,x;h;t,T)$. Then $\tilde{\Phi}$ is defined as

$$\tilde{\Phi}(t,x) = \inf_{h\in\mathcal{A}(T)} \tilde{I}(v,x;h;t,T). \tag{3.15}$$

The corresponding HJB PDE is

$$\frac{\partial\tilde{\Phi}}{\partial t}(t,x) + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T D^2\tilde{\Phi}(t,x)\right) + H(t,x,\tilde{\Phi},D\tilde{\Phi}) = 0, \tag{3.16}$$

which is subject to terminal condition

$$\tilde{\Phi}(T,x) = v^{-\theta} \tag{3.17}$$

and where

$$H(x,r,p) = \inf_{h\in\mathcal{J}} \left\{f(x,h)^T p + \theta g(x,h) r\right\} \tag{3.18}$$

for $r \in \mathbb{R}$, $p \in \mathbb{R}^n$ and, in particular,

$$\tilde{\Phi}(t,x) = \exp\left\{-\theta\Phi(t,x)\right\}. \tag{3.19}$$

Note that since Φ and $\tilde{\Phi}$ are related through a strictly monotone continuous transformation, an admissible (optimal) strategy for the exponentially transformed problem is also admissible (optimal) for the risk-sensitive problem.

4. Main result. In this section, we present the main result of this article, namely, that the risk-sensitive jump diffusion problem admits a classical ($C^{1,2}$) solution, and show that the value function Φ is convex in x.

Proposition 4.1. *The value function $\Phi(t,x)$ is convex in x.*

Proof. To prove that the value function $\Phi(t,x)$ is convex in x, it is necessary and sufficient to show that $\forall(x_1,x_2) \in \mathbb{R}^n$ and for any $\kappa \in (0,1)$,

$$\Phi(t,\kappa x_1 + (1-\kappa)x_2) \leq \kappa\Phi(t,x_1) + (1-\kappa)\Phi(t,x_2). \tag{4.1}$$

Start from the left-hand side:

$$
\begin{aligned}
&\Phi(t, \kappa x_1 + (1-\kappa)x_2) \\
&= \sup_{h\in\mathcal{A}(T)} -\frac{1}{\theta} \ln \mathbf{E}^h_{t,\kappa x_1+(1-\kappa)x_2} \left[\exp\left\{ \theta \int_t^T g(X_s, h(s))ds - \theta \ln v \right\} \chi(t) \right] \\
&= \sup_{h\in\mathcal{A}(T)} -\frac{1}{\theta} \ln \mathbf{E}^h_{t,(x_1,x_2)} \left[\exp\left\{ \theta \int_t^T g(\kappa X_1(s) + (1-\kappa)X_2(s), h(s))ds - \theta \ln v \right\} \chi(t) \right] \\
&= \sup_{h\in\mathcal{A}(T)} -\frac{1}{\theta} \ln \mathbf{E}^h_{t,(x_1,x_2)} \left[\exp\left\{ \kappa\theta \int_t^T g(X_1(s), h(s))ds \right.\right. \\
&\quad \left.\left. + (1-\kappa)\theta \int_t^T g(X_2(s), h(s))ds - \theta \ln v \right\} \chi(t) \right] \\
&= \sup_{h\in\mathcal{A}(T)} -\frac{1}{\theta} \ln \mathbf{E}^h_{t,(x_1,x_2)} \left[\left(\exp\left\{ \theta \int_t^T g(X_1(s), h(s))ds - \theta \ln v \right\} \chi(t) \right)^\kappa \right. \\
&\quad \left. \times \left(\exp\left\{ \theta \int_t^T g(X_2(s), h(s))ds - \theta \ln v \right\} \chi(t) \right)^{1-\kappa} \right] \\
&\le \sup_{h\in\mathcal{A}(T)} -\frac{1}{\theta} \ln \left\{ \mathbf{E}^h_{t,x_1} \left[\left(\exp\left\{ \theta \int_t^T g(X_1(s), h(s))ds - \theta \ln v \right\} \chi(t) \right)^\kappa \right] \right. \\
&\quad \left. \times \mathbf{E}^h_{t,x_2} \left[\left(\exp\left\{ \theta \int_t^T g(X_2(s), h(s))ds - \theta \ln v \right\} \chi(t) \right)^{1-\kappa} \right] \right\} \\
&\quad = \sup_{h\in\mathcal{A}(T)} \left\{ -\frac{1}{\theta} \ln \mathbf{E}^h_{t,x_1} \left[\left(\exp\left\{ \theta \int_t^T g(X_1(s), h(s))ds - \theta \ln v \right\} \chi(t) \right)^\kappa \right] \right. \\
&\quad\quad \left. -\frac{1}{\theta} \ln \mathbf{E}^h_{t,x_2} \left[\left(\exp\left\{ \theta \int_t^T g(X_2(s), h(s))ds - \theta \ln v \right\} \chi(t) \right)^{1-\kappa} \right] \right\} \\
&\quad \le \sup_{h\in\mathcal{A}(T)} -\frac{1}{\theta} \ln \mathbf{E}^h_{t,x_1} \left[\left(\exp\left\{ \theta \int_t^T g(X_1(s), h(s))ds - \theta \ln v \right\} \chi(t) \right)^\kappa \right] \\
&\quad\quad + \sup_{h\in\mathcal{A}(T)} -\frac{1}{\theta} \ln \mathbf{E}^h_{t,x_2} \left[\left(\exp\left\{ \theta \int_t^T g(X_2(s), h(s))ds - \theta \ln v \right\} \chi(t) \right)^{1-\kappa} \right] \\
&\quad \le \sup_{h\in\mathcal{A}(T)} -\frac{\kappa}{\theta} \ln \mathbf{E}^h_{t,x_1} \left[\exp\left\{ \theta \int_t^T g(X_1(s), h(s))ds - \theta \ln v \right\} \chi(t) \right] \\
&\quad\quad + \sup_{h\in\mathcal{A}(T)} -\frac{1-\kappa}{\theta} \ln \mathbf{E}^h_{t,x_2} \left[\exp\left\{ \theta \int_t^T g(X_2(s), h(s))ds - \theta \ln v \right\} \chi(t) \right] \\
&\quad = \kappa\Phi(t, x_1) + (1-\kappa)\Phi(t, x_2),
\end{aligned}
$$

where the fourth equality follows from the fact that the covariance of two random variables inside the expectations is positive and the third inequality is due to the fact that the function $x \mapsto x^\alpha$ for $x > 0$ and $\alpha \in (0, 1)$ is concave. ■

Corollary 4.2. *The exponentially transformed value function $\tilde{\Phi}$ has the following property:* $\forall(x_1, x_2) \in \mathbb{R}^2,\ \kappa \in (0, 1,),$

$$\tilde{\Phi}(t, \kappa x_1 + (1-\kappa)x_2) \geq \tilde{\Phi}^{\kappa}(t, x_1)\tilde{\Phi}^{1-\kappa}(t, x_2). \tag{4.2}$$

Proof. The property follows immediately from the definition of $\Phi = -\frac{1}{\theta}\ln\tilde{\Phi}$. ■

We now come to the main result of this article. Recall the standing assumptions: in section 2, Assumption 1 is a condition on the support of the jump measure, and Assumptions 2 and 3 are the nondegeneracy conditions $\Sigma\Sigma^T > 0$, $\Lambda\Lambda^T > 0$, while Assumption 4, introduced in section 7.1, is a full-rank condition on the matrix $\hat{A}$ defined at (2.11).

Theorem 4.3. *Under Assumptions 1–4 the following hold:*

1. *The optimal asset allocation is the unique maximizer of the supremum* (3.12)*.*
2. $\tilde{\Phi}$ *is the unique* $C^{1,2}([0,T]\times\mathbb{R}^n)$ *solution of the RS HJB PIDE* (3.16)*–*(3.17)*. Moreover,* $\tilde{\Phi}$ *satisfies the property* (4.2)*.*
3. Φ *is the unique* $C^{1,2}([0,T]\times\mathbb{R}^n)$ *solution of the RS HJB PIDE* (3.12)*–*(3.14)*. Moreover,* Φ *is convex in its argument* x*.*

Proof. The proof is based on a series of results proved in sections 5–7. These combine to give us the following arguments.

Existence of an optimal control. By Proposition 5.1, the supremum in (3.12) admits a unique Borel measurable maximizer. Moreover, by Proposition 6.3, this maximizer is admissible, and by Proposition 6.4 it is also a maximizer with respect to the $\mathbb{P}$-measure criterion J defined in (3.1). Thus, we can take this maximizer as our optimal asset allocation.

Existence of a classical ($C^{1,2}$) solution. By Corollary 7.3, $\tilde{\Phi}$ is a $C^{1,2}([0,T]\times\mathbb{R}^n)$ solution of the RS HJB PDE (3.16)–(3.17).

Uniqueness of the classical solution. The existence of zero beta policies enable us to deduce (as in Step 1 of the proof of Theorem 7.2) that $\tilde{\Phi}$ is bounded. Part (i) of Theorem 6.1 therefore applies. Choosing as optimal control the unique maximizer of the supremum in (3.12), part (ii) of Theorem 6.1 also applies: $\tilde{\Phi}$ is the unique solution to the HJB PIDE. Property (4.2) is proved in Corollary 4.2.

By Corollaries 7.3 and 6.2, it then follows that Φ is the unique classical solution to the HJB PIDE (3.12) with terminal condition (3.14). Moreover, Φ is convex in its second argument x. ■

This result proves that we have solved our original control problem in the context of strong, classical solutions. What would this imply in terms of weaker viscosity solutions? As a classical solution is also a viscosity solution, our result implies that the value function is indeed a viscosity solution of the HJB PDE. However, uniqueness of classical solutions does not necessarily imply uniqueness of viscosity solutions. To prove uniqueness in the viscosity sense, we would need a comparison result such as Theorem 33 in Davis and Lleo [13].

In the remainder of this paper, we develop the various technical arguments required in the proof of Theorem 4.3.

5. Existence of a maximizing control.

Proposition 5.1. *Under Assumption 2, the supremum in* (3.12) *admits a unique Borel measurable maximizer* $\hat{h}(t,x,p)$ *for* $(t,x,p)\in[0,T]\times\mathbb{R}^n\times\mathbb{R}^n$*. Moreover, the maximizer* $\hat{h}(t,x,p)$ *is an interior point of the set* $\overline{\mathcal{J}}$*.*

Proof. The supremum in (3.12) can be expressed as

$$\begin{aligned}
&\sup_{h\in\mathcal{J}} L_t^h\Phi \\
&= (b+Bx)^T D\Phi + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T D^2\Phi\right) - \frac{\theta}{2}(D\Phi)^T\Lambda\Lambda^T D\Phi + a_0 + A_0^T x \\
&\quad + \sup_{h\in\mathcal{J}}\Big\{ -\frac{1}{2}(\theta+1)h^T\Sigma\Sigma^T h - \theta h^T\Sigma\Lambda^T D\Phi + h^T(\hat{a}+\hat{A}x) \\
&\quad - \frac{1}{\theta}\int_{\mathbf{Z}}\left\{\left[\left(1+h^T\gamma(z)\right)^{-\theta} - 1\right] + \theta h^T\gamma(z)\mathit{1}_{\mathbf{Z}_0}(z)\right\}\nu(dz)\Big\}. \qquad (5.1)
\end{aligned}$$

Under Assumption 2, for any $p \in \mathbb{R}^n$, the terms

$$-\frac{1}{2}(\theta+1)h^T\Sigma\Sigma^T h - \theta h^T\Sigma\Lambda^T p + h^T(\hat{a}+\hat{A}x) - \int_{\mathbf{Z}} h^T\gamma(z)\mathit{1}_{\mathbf{Z}_0}(z)\nu(dz)$$

and

$$-\frac{1}{\theta}\int_{\mathbf{Z}}\left\{\left[\left(1+h^T\gamma(z)\right)^{-\theta} - 1\right]\right\}\nu(dz)$$

are both strictly concave in h $\forall z \in \mathbb{Z}$ a.s. $d\nu$. Therefore, the supremum is reached for a unique maximizer $\hat{h}(t,x,p)$, which is an interior point of the set $\overline{\mathcal{J}}$, which is the closure of the set $\mathcal{J}$ defined in (2.7), and the supremum, evaluated at $\hat{h}(t,x,p) \in \mathbb{R}^n$, is finite. By measurable selection, $\hat{h}$ can be taken as a Borel measurable function on $[0,T]\times\mathbb{R}^n\times\mathbb{R}^n$. ■

6. Verification theorems. In this section, we prove a verification theorem to the effect that if (3.12) has a $C^{1,2}$ solution, then that solution is equal to Φ defined by (3.11) and the control $h^*(t) = \hat{h}(t,x,D\Phi)$ is optimal. We will first prove a verification theorem for the exponentially transformed problem (3.15) with HJB PDE (3.16) and value function $\tilde{\Phi}(t,x)$. As a corollary, we will obtain a verification theorem for the risk-sensitive control problem with (3.11), HJB PDE (3.12), and value function $\Phi(t,x)$. Define the first order operator

$$\tilde{L}^h\varphi(t,x) = \left(b + Bx - \theta\Lambda\Sigma^T h\right)^T D\varphi(t,x) + \theta g(x,h)\varphi(t,x). \qquad (6.1)$$

Theorem 6.1 (verification theorem for the exponentially transformed control problem). *Let $\tilde{\phi}$ be a $C^{1,2}([0,T]\times\mathbb{R}^n)$ bounded function.*

(i) *Assume that $\tilde{\phi}(T,x) \le e^{-\theta\ln v}$ $\forall x \in \mathbb{R}^n$ and*

$$\frac{\partial\tilde{\phi}}{\partial t}(t,x) + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T D^2\tilde{\phi}(t,x)\right) + H(t,x,\tilde{\phi},D\tilde{\phi}) \ge 0$$

on $[0,T]\times\mathbb{R}^n$; then $\tilde{\phi}(t,x) \le \tilde{\Phi}(t,x)$ $\forall (t,x) \in [0,T]\times\mathbb{R}^n$.

(ii) *Further assume that $\tilde{\phi}(T,x) = e^{-\theta\ln v}$ $\forall x \in \mathbb{R}^n$ and that there exists a Borel-measurable minimizer $\tilde{h}^*(t,x)$ of $\tilde{h} \mapsto \tilde{L}^{\tilde{h}}\tilde{\phi}$ defined in (6.1) such that*

$$\begin{aligned}
&\frac{\partial\tilde{\phi}}{\partial t}(t,x) + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T D^2\tilde{\phi}(t,x)\right) + H(t,x,\tilde{\phi},D\tilde{\phi}) \\
&= \frac{\partial\tilde{\phi}}{\partial t}(t,x) + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T D^2\tilde{\phi}(t,x)\right) + \tilde{L}^{\tilde{h}^*}\tilde{\phi} \\
&= 0
\end{aligned}$$

and the stochastic differential equation

$$dX(t) = \left(b + BX(t) - \theta\Lambda\Sigma^T h(t)\right) dt + \Lambda dW_t^\theta$$

defines a unique solution $X(s)$ *for each given initial data* $X_t = x$ *and the process* $\pi^*(s) := \tilde{h}^*(s, X(s))$ *is a well-defined control process in* $\tilde{\mathcal{A}}(T)$. *Then* $\tilde{\phi} = \tilde{\Phi}$, *and* $\pi^*(s)$ *is an optimal Markov control process.*

Proof. The following proof is based on an argument used by Touzi [38].

(i) Let $\tilde{h} \in \tilde{\mathcal{A}}(T)$ be an arbitrary control, with $X(t)$ the state process with initial data $X(t) = x$. Define the stopping time

$$\tau_N := T \wedge \inf\left\{s > t : |X_s - x| \geq N\right\}.$$

Define $Z(s) = \theta \int_t^s g(s, X_s, \hat{h}_s) ds$; then

$$d\left(e^{Z_s}\right) := \theta g(s, X_s, \hat{h}_s) e^{Z_s}.$$

Also, by the Itô formula, for $s \in [t, \tau_\delta]$,

$$d\tilde{\phi}_s = \left\{\frac{\partial\tilde{\phi}}{\partial s} + \mathcal{L}\tilde{\phi}\right\} ds + D\tilde{\phi}^T \Lambda dW_s^\theta,$$

where $\mathcal{L}$ is the generator of the state process $X(t)$ defined as

$$\mathcal{L}\tilde{\phi}(t,x) := \left(b + Bx - \theta\Lambda\Sigma^T h(s)\right)^T D\tilde{\phi} + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T D^2\tilde{\phi}\right).$$

By the Itô product rule, and since $dZ_s \cdot \tilde{\phi}_s = 0$, we get

$$d\left(\tilde{\phi}_s e^{Z_s}\right) = \tilde{\phi}_s d\left(e^{Z_s}\right) + e^{Z_s} d\tilde{\phi}_s,$$

and hence for $s \in [t, \tau_N]$

$$\begin{aligned}
\tilde{\phi}(s, X_s) e^{Z_s} &= \tilde{\phi}(t,x) e^{Z_t} + \theta \int_t^s \tilde{\phi}(u, X_u) g(u, X_u, \hat{h}_u) e^{Z_u} du \\
&\quad + \int_t^s \left(\frac{\partial\tilde{\phi}}{\partial u}(u, X_u) + \mathcal{L}\tilde{\phi}(u, X_u) e^{Z_u}\right) du + \int_t^s D\tilde{\phi}^T \Lambda dW_u^\theta.
\end{aligned}$$

Because, for an arbitrary control h,

$$\begin{aligned}
&\frac{\partial\tilde{\phi}}{\partial t}(t,x) + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T D^2\tilde{\phi}(t,x)\right) + \mathcal{L}^{\tilde{h}}\tilde{\phi}(t, X_t) \\
&\geq \frac{\partial\tilde{\phi}}{\partial t}(t,x) + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T D^2\tilde{\phi}(t,x)\right) + H(t, x, \tilde{\phi}, D\tilde{\phi}) \\
&\geq 0
\end{aligned}$$

and $e^{Z_s} \geq 0 \; \forall s \in [t, \tau_N]$, we have

$$\tilde{\phi}(t,x)e^{Z_t} \leq \tilde{\phi}(s, X_s)e^{Z_s} + \int_t^s D\tilde{\phi}^T \Lambda dW_u^\theta.$$

Taking the expectation, we obtain

$$\tilde{\phi}(t,x)e^{Z_t} \leq \mathbf{E}_{t,x}^{\tilde{h},\theta}\left[\tilde{\phi}(s,X_s)e^{Z_s}\right] = \mathbf{E}_{t,x}^{\tilde{h},\theta}\left[\tilde{\phi}(s,X_s)e^{\theta\int_t^s g(u,X_u,\hat{h}_u)du}\right].$$

In particular, take $s = \tau_N$, and note that $e^{Z_t} = 1$; then

$$\tilde{\phi}(t,x)e^{Z_t} \leq \mathbf{E}_{t,x}^{\tilde{h},\theta}\left[\tilde{\phi}(\tau_N, X_{\tau_N})e^{\theta\int_t^{\tau_N} g(u,X_u,\hat{h}_u)du}\right].$$

Since $\tilde{\phi}$ is assumed to be bounded, there exists a constant $C_1 > 0$ such that

$$\left|\tilde{\phi}(s,X_s)e^{\theta\int_s^{\tau_N} g(s,X_s,\hat{h}_s)ds}\right| \leq C_1 e^{\theta\int_t^{\tau_N} g(u,X_u,\hat{h}_u)du}.$$

Since for an arbitrary admissible control $\tilde{h} \in \mathcal{A}(T)$ and fixed $s \in [t,T]$ there exists some constant $C_2 > 0$ such that

$$\left|g(s,X_s,\hat{h}_s)\right| \leq C_2\left|1 + X(s)\right|,$$

then

$$\begin{aligned}\left|\tilde{\phi}(s,X_s)e^{\theta\int_s^{\tau_N} g(s,X_s,\hat{h}_s)ds}\right| &\leq C_3 e^{\theta\int_t^{\tau_N}|1+X(s)|du} \\ &\leq C_3 e^{\theta(\tau_N - t) + \theta\int_t^{\tau_N}|X(s)|du} \\ &\leq C_4 e^{\theta\int_t^{\tau_N}|X(s)|du} \\ &\leq C_4 e^{\theta(T-t)\sup_{t\leq s\leq T}|X(s)|}\end{aligned}$$

for $C_3 = C_1 e^{C_2}$ and $C_4 = C_3 e^{\theta(T-t)}$.
By the dominated convergence theorem and the assumption that $\tilde{\phi}(T, X_t) \leq e^{-\theta \ln v}$,

$$\begin{aligned}\tilde{\phi}(t,x) &\leq \mathbf{E}_{t,x}^{\tilde{h},\theta}\left[\tilde{\phi}(T,X_T)e^{\theta\int_t^T g(u,X_u,\hat{h}_u)du}\right] \\ &\leq \mathbf{E}_{t,x}^{\tilde{h},\theta}\left[e^{\theta\int_t^T g(u,X_u,\hat{h}_u)du} - \theta\ln v\right].\end{aligned}$$

We have now proved the first part of the theorem.

(ii) To prove the second part, we can simply apply the same reasoning for the optimal control $\tilde{h}^*$. Note, however, that with this choice of control we would have

$$\begin{aligned}&\frac{\partial\tilde{\phi}}{\partial t}(t,x) + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T D^2\tilde{\phi}(t,x)\right) + H(t,x,\tilde{\phi},D\tilde{\phi}) \\ &= \frac{\partial\tilde{\phi}}{\partial t}(t,x) + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T D^2\tilde{\phi}(t,x)\right) + \tilde{L}^{\tilde{h}^*}\tilde{\phi} \\ &= 0,\end{aligned}$$

which would lead us to equality in the last equation, i.e.,

$$\tilde{\phi}(t,x) = \mathbf{E}_{t,x}^{\tilde{h},\theta}\left[e^{\theta\int_t^T g(u,X_u,\hat{h}_u)du} - \theta\ln v\right]. \qquad \blacksquare$$

Corollary 6.2 (verification theorem for the risk-sensitive control problem). *Letting* ϕ *be a* $C^{1,2}([0,T]\times\mathbb{R}^n)\cap C([0,T]\times\mathbb{R}^n)$ *bounded function, we assume the following:*

(i) $\phi(T,x)\leq e^{-\theta\ln v}$ $\forall x\in\mathbb{R}^n$ *and*

$$\frac{\partial\phi}{\partial t}+\sup_{h\in\mathcal{J}}L_t^h\phi(t,X(t))\geq 0$$

on $[0,T]\times\mathbb{R}^n$*; then* $\phi(t,x)\leq\tilde{\Phi}(t,x)$ $\forall(t,x)\in[0,T]\times\mathbb{R}^n$.

(ii) *Further,* $\phi(T,x)=e^{-\theta\ln v}$ $\forall x\in\mathbb{R}^n$ *and there exists a minimizer* $h^*(t,x)$ *of* $h\mapsto L^h\phi$ *defined in* (3.13) *such that*

$$\frac{\partial\phi}{\partial t}+\sup_{h\in\mathcal{J}}L_t^h\phi(t,X(t))=\frac{\partial\phi}{\partial t}+L_t^{h^*}\phi(t,X(t))=0$$

and the stochastic differential equation

$$dX(s)=\left(b+BX(s^-)-\theta\Lambda\Sigma^T h(s)\right)ds+\Lambda dW_s^\theta$$

defines a unique solution X *for each given initial data* $X_t=x$ *and the process* $\pi^*(s):=\tilde{h}^*(s,X(s))$ *is a well-defined control process in* $\tilde{\mathcal{A}}(T)$*. Then* $\phi=\Phi$*, and* $\pi^*(s)$ *is an optimal Markov control process.*

Proof. This corollary follows from (3.19) and from the fact that an admissible (optimal) strategy for the exponentially transformed problem is also admissible (optimal) for the risk-sensitive problem. ■

Proposition 6.3. *The process* $h^*(t)$ *is admissible:* $h^*(t)\in\mathcal{A}(T)$.

Proof. Refer to the appendix for a full discussion and a proof of this proposition. ■

Applying Proposition 6.3 we deduce that the control $h^*(t)$ is optimal for the auxiliary problems (3.9) and (3.10) resulting from the change of measure. However, this proposition is not sufficient to conclude that $h^*(t)$ is optimal for the original problem (3.1) set under the $\mathbb{P}$-measure. The next result shows that this is indeed the case.

Proposition 6.4. *The optimal control* $h^*(t)$ *for the auxiliary problem*

$$\sup_{h\in\mathcal{A}(T)} I(v,x;h;t,T),$$

where I *is defined in* (3.9)*, is also optimal for the initial problem* $\sup_{h\in\mathcal{A}(T)}J(x,t,h)$*, where* J *is defined in* (3.1).

Proof. See the appendix. ■

7. Existence of a classical solution. Historically, proving the existence of a strong, analytical solution to the HJB PDE was both the main difficulty and the main objective when solving a control problem. Fleming and Rishel [16] as well as Krylov [24], [25] have been the main contributors, proposing techniques based either on PDE arguments or on probability theory. Recently, however, the emphasis has switched from strong solutions to weaker types of solutions. Viscosity solutions have proved particularly useful and successful, gaining many applications in stochastic control theory (see, for example, the classic article by Crandall, Ishii, and Lions [11] as well as Fleming and Soner [20] for their applications to stochastic control). The reason for this appeal is twofold. First, it is significantly easier to prove the existence of a viscosity solution than a classical solution. In the viscosity world, the difficulty is shifted from proof of existence to proof of uniqueness, and even then it is generally easier to prove uniqueness of a viscosity solution via a comparison theorem than the existence of a classical solution. Second, the stability result due to Barles and Souganidis [1] connects viscosity solutions to numerical methods directly, making it easy to solve "real world" control problems.

This section follows similar arguments to those developed by Fleming and Rishel [16] (Theorem 6.2 and Appendix E). Namely, we use an approximation in policy space alongside results on linear parabolic PDEs to prove that the exponentially transformed value function $\tilde{\Phi}$ is of class $C^{1,2}((0,T) \times \mathbb{R}^n)$. Then it follows that the value function Φ is also of class $C^{1,2}((0,T) \times \mathbb{R}^n)$. The approximation in policy space algorithm was originally proposed by Bellman in the 1950s (see Bellman [2] for details) as a numerical method to compute the value function. Our approach has two steps. First, we use the approximation in policy space algorithm to show the existence of a classical solution in a bounded region. Then we extend our argument to unbounded state space. To derive this second result we follow an argument different from that of Fleming and Rishel [16] which makes more use of the actual structure of the control problem.

Our interest in classical solutions is as much mathematical as practical. First, since a smooth solution is a viscosity solution but the converse is not necessarily true, we are proving a stronger result. Second, this stronger result immediately translates a better grasp of the analytical properties of the value function. While viscosity solutions provide continuity, they do not generally give information about higher order derivatives. By contrast, classical solutions are smooth in the state, implying that they are (at least) C^1 in time and C^2 in the state. Third, viscosity solutions are purely about solving the PDE, and although they show that the value function is the unique solution of the HJB PDE, they do not prove directly that the control problem has a solution, that is, a pair of a value function and an admissible optimal control. Fourth and finally, in our case seeking a strong solution does not impair our search for numerical results. Because our state process $X(t)$ can clearly be interpreted as the continuous time limit of a Markov chain, we can apply well-known results by Kushner and Dupuis [27] to prove convergence of a finite approximation scheme to the value function. We can therefore solve concrete portfolio selection problems quite directly.

7.1. "Zero beta" policies. In this section, we introduce a new class of control policies: the zero beta (0β) policies.

Definition 7.1 (0β policy). *By reference to the definition of the function g in (3.4), a 0β* control policy $\check{h}(t)$ *is an admissible control policy for which the function g is independent of the state variable x.*

The term "zero beta" is borrowed from financial economics (see, for instance, Black [10]). To avoid assuming the existence of a globally risk-free rate in factor models such as the CAPM and the APT or in ad hoc valuation models, it is customary to build portfolios without any exposure to the factor(s) as a substitute for the risk-free rate. These special portfolios are referred to as 0β portfolios by reference to the slope coefficient β used to measure the sensitivity of asset returns to the valuation factor(s).

In the risk-sensitive asset management model, if $A_0 = 0$, then the policy $h^0 = 0$, i.e., invest all the wealth in the risk-free asset, is a 0β policy. When $A_0 \neq 0$, we see from (2.11) that a 0β policy can exist only if there is a vector $\check{h}$ satisfying

$$\check{h}^T \hat{A} = -A_0. \tag{7.1}$$

We introduce the following standing assumption.

Assumption 4. The matrix $\hat{A}$ has rank n.

Under this assumption there are always 0β policies: we need only take a vector $\check{h}$ satisfying (7.1) and scale it if necessary so that $\check{h} \in \mathcal{J}$ (see (2.7)), and then $h(t,\omega) = \check{h}$ is a 0β policy. We have no reason to consider anything other than the set $\mathcal{Z}$ of constant policies of this kind. Note that when $\check{h} \in \mathcal{Z}$ the function g of (3.4) is a constant: $g(x,\check{h}) \equiv \check{g}$.

7.2. The $L^\eta(K)$ and $\mathscr{L}^\eta(K), 1 < \eta < \infty$ spaces. The following ideas and notation relate to the treatment of linear parabolic PDEs found in Ladyženskaja, Solonnikov, and Uralceva [28]. The relevant results are summarized in Appendix E of Fleming and Rishel [16]. They concern PDEs of the form

$$\frac{\partial \psi}{\partial t} + \frac{1}{2}\mathrm{tr}\left(a(t,x)D^2\psi\right) + b(t,x)^T D\psi + \theta c(t,x)\psi + d(t,x) = 0 \tag{7.2}$$

on a set $Q = (0,T) \times G$ and with boundary conditions

$$\begin{aligned} \psi(t,x) &= \Psi_T(x), \quad x \in G, \\ \psi(t,x) &= \Psi(t,x), \quad (t,x) \in (0,T) \times \partial G. \end{aligned}$$

The set G is open and is such that ∂G is a compact manifold of class C^2. Denote the following:

- $\partial^* Q$ is the boundary of Q, i.e.,

$$\partial^* Q := (\{T\} \times G) \cup ((0,T) \times \partial G).$$

- $L^\eta(K)$ is the space of ηth power integrable functions on $K \subset Q$.
- $\|\cdot\|_{\eta,K}$ is the norm in $L^\eta(K)$.

Also, denote by $\mathscr{L}^\eta(Q), 1 < \eta < \infty$, the space of all functions ψ such that ψ and all its generalized partial derivatives are in $L^\eta(K)$. We associate with this space the Sobolev-type norm

$$\|\psi\|^{(2)}_{\eta,K} := \|\psi\|_{\eta,K} + \left\|\frac{\partial \psi}{\partial t}\right\|_{\eta,K} + \sum_{i=1}^n \left\|\frac{\partial \psi}{\partial x_i}\right\|_{\eta,K} + \sum_{i,j=1}^n \left\|\frac{\partial^2 \psi}{\partial x_i x_j}\right\|_{\eta,K}. \tag{7.3}$$

We will also introduce additional notation and concepts as required in the proofs.

7.3. Existence of a classical solution. In this section, we use an approximation in policy space to show the existence of a $C^{1,2}$ solution to the RS HJB PDE (3.12).

Theorem 7.2 (existence of a classical solution for the exponentially transformed control problem). *The RS HJB PDE (3.16) with terminal condition $\tilde{\Phi}(T,x) = e^{-\theta \ln v}$ has a solution $\tilde{\Phi} \in C^{1,2}((0,T) \times \mathbb{R}^n)$ with $\tilde{\Phi}$ continuous in $[0,T] \times \mathbb{R}^n$.*

Proof.

Step 1: Approximation in policy space—bounded space. Consider the following auxiliary problem: fix $R > 0$, and let $\mathscr{B}_R$ be the open n-dimensional ball of radius $R > 0$ centered at 0 defined as $\mathscr{B}_R := \{x \in \mathbb{R}^n : |x| < R\}$. We construct an investment portfolio by solving the optimal risk-sensitive asset allocation problem as long as $X(t) \in \mathscr{B}_R$ for $R > 0$. Then, as soon as $X(t) \notin \mathscr{B}_R$, we switch all of the wealth into the 0β policy $\check{h}$ from the exit time t until the end of the investment horizon at time T. The HJB PDE for this auxiliary problem can be expressed as

$$\frac{\partial \tilde{\Phi}}{\partial t} + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T(t)D^2\tilde{\Phi}\right) + H(t,x,\tilde{\Phi},D\tilde{\Phi}) = 0 \qquad \forall (t,x) \in Q_R := (0,T) \times \mathscr{B}_R, \tag{7.4}$$

where, as in (3.18),

$$H(x,r,p) = \inf_{h \in \overline{\mathcal{J}}} \left\{ f(x,h)^T p + \theta g(t,x,h) r \right\}$$

for $p \in \mathbb{R}^n$ and for g and f defined in (3.4) and (3.8), respectively. $\overline{\mathcal{J}}$, defined as the closure of $\mathcal{J}$, is a compact set. The boundary conditions are

$$\tilde{\Phi}(t,x) = \Psi(t,x) \qquad \forall (t,x) \in \partial^* Q_R := ((0,T) \times \partial\mathscr{B}_R) \cup (\{T\} \times \mathscr{B}_R),$$

where the following hold:

- $\Psi(T,x) = e^{-\theta \ln v}$ $\forall x \in \mathscr{B}_R$.
- $\Psi(t,x) := \psi(t,x) := e^{\theta \check{g}(T-t)}$ $\forall (t,x) \in (0,T) \times \partial\mathscr{B}_R$, and $\check{h}$ is a fixed arbitrary 0β policy which is constant as a function of time. Note that ψ is obviously of class $C^{1,2}(\overline{Q_R})$ and that the Sobolev-type norm

$$\|\Psi\|^{(2)}_{\eta,\partial^* Q_R} = \|\tilde{\Psi}\|^{(2)}_{\eta,Q_R} \tag{7.5}$$

is finite.

Define a sequence of functions $\tilde{\Phi}^1, \tilde{\Phi}^2, \ldots, \tilde{\Phi}^k, \ldots$ on $\overline{Q_R} = [0,T] \times \overline{\mathscr{B}_R}$ and of bounded measurable feedback control laws $h^0, h^1, \ldots, h^k, \ldots$, where h^0 is an arbitrary control. Assuming h^{k-1} is known, we define the function $\tilde{\Phi}^{k+1}$ as the solution to the boundary value problem

$$\frac{\partial \tilde{\Phi}^k}{\partial t} + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T(t)D^2\tilde{\Phi}^k\right) + f(x,h^{k-1})^T D\tilde{\Phi}^k + \theta g(t,x,h^{k-1})\tilde{\Phi}^k = 0 \tag{7.6}$$

subject to boundary conditions

$$\tilde{\Phi}(t,x) = \Psi(t,x) \qquad \forall (t,x) \in \partial^* Q_R := ((0,T) \times \partial\mathscr{B}_R) \cup (\{T\} \times \mathscr{B}_R).$$

Note that the boundary value problem (7.6) is a special case of the generic problem introduced earlier in (7.2) with

$$\begin{aligned} a(t,x) &= \Lambda\Lambda^T(t), \\ b(t,x) &= f(x,h^k), \\ c(t,x) &= g(t,x,h^k), \\ d(t,x) &= 0. \end{aligned}$$

Moreover, since $\mathscr{B}_R$ is bounded and $\overline{\mathcal{J}}$ is compact, all of these functions are also bounded. Thus, based on standard results on parabolic PDEs (see, for example, Appendix E in Fleming and Rishel [16] and Chapter IV in Ladyženskaja, Solonnikov, and Uralceva [28]), the boundary value problem (7.6) admits a unique solution in $\mathscr{L}^\eta(Q_R)$.

Next, for almost all $(t,x) \in Q_R$, $k = 1, 2, \ldots$, we define h^k by the prescription

$$h^k = \operatorname*{Argmin}_{h\in\overline{\mathcal{J}}} \left\{ f(x,h)^T D\tilde{\Phi}^k + \theta g(t,x,h)\tilde{\Phi}^k \right\} \tag{7.7}$$

so that

$$\begin{aligned} f(x,h^k)^T D\tilde{\Phi}^k + \theta g(t,x,h^k)\tilde{\Phi}^k &= \inf_{h\in\overline{\mathcal{J}}} \left\{ f(x,h)^T D\tilde{\Phi}^k + \theta g(t,x,h)\tilde{\Phi}^k \right\} \\ &= H(t,x,\tilde{\Phi}^k, D\tilde{\Phi}^k). \end{aligned} \tag{7.8}$$

Note, in view of the definition of g in (3.4), that the minimum is never achieved on the boundary of $\overline{\mathcal{J}}$; i.e., h^k takes values in $\mathcal{J}$.

Observe that the sequence $(\tilde{\Phi}^k)_{k\in\mathbb{N}}$ is globally bounded. Indeed, by the Feynman–Kac formula, the sequence $(\tilde{\Phi}^k)_{k\in\mathbb{N}}$ is bounded from below by 0. By the optimality principle, it is also bounded from above by $e^{\theta\int_t^T g(X(s),\check{h})ds} = e^{\theta\check{g}(T-t)}$. Moreover, these bounds do not depend on the radius R and are therefore valid over the entire space $(0,T)\times\mathbb{R}^n$.

Step 2: Convergence inside the cylinder $(0,T)\times\mathscr{B}_R$.

Step 2.1: Monotonicity of the sequence. Take $k \geq 1$. Subtracting the PDE for $\tilde{\Phi}^{k+1}$ from the PDE for $\tilde{\Phi}^k$, we see that

$$\begin{aligned} &\left(\frac{\partial\tilde{\Phi}^{k+1}}{\partial t} - \frac{\partial\tilde{\Phi}^k}{\partial t}\right) + \left(\frac{1}{2}\mathrm{tr}\left[\left(\Lambda\Lambda^T(t)D^2\tilde{\Phi}^{k+1}\right) - \left(\Lambda\Lambda^T(t)D^2\tilde{\Phi}^k\right)\right]\right) \\ &\quad + \left(f(x,h^k)^T D\tilde{\Phi}^{k+1} - f(x,h^{k-1})^T D\tilde{\Phi}^k\right) + \theta\left(g(t,x,h^k)\tilde{\Phi}^{k+1} - g(t,x,h^{k-1})\tilde{\Phi}^k\right) \\ &= 0 \qquad \text{in } (0,T)\times\mathbb{R}^n \end{aligned}$$

with $\tilde{\Phi}^{k+1} - \tilde{\Phi}^k = 0$ on $\mathbb{R}^n$.

Add and subtract $f(x,h^k)^T D\tilde{\Phi}^k + \theta g(t,x,h^k)\tilde{\Phi}^k$:

$$\begin{aligned} &\left(\frac{\partial\tilde{\Phi}^{k+1}}{\partial t} - \frac{\partial\tilde{\Phi}^k}{\partial t}\right) + \left(\frac{1}{2}\mathrm{tr}\left[\left(\Lambda\Lambda^T(t)D^2\tilde{\Phi}^{k+1}\right) - \left(\Lambda\Lambda^T(t)D^2\tilde{\Phi}^k\right)\right]\right) \\ &\quad + \left(f(x,h^k)^T D\tilde{\Phi}^{k+1} - f(x,h^{k-1})^T D\tilde{\Phi}^k\right) + \theta\left(g(t,x,h^k)\tilde{\Phi}^{k+1} - g(t,x,h^{k-1})\tilde{\Phi}^k\right) \\ &\quad + \left(f(x,h^k)^T D\tilde{\Phi}^k + \theta g(t,x,h^k)\tilde{\Phi}^k\right) - \left(f(x,h^k)^T D\tilde{\Phi}^k + \theta g(t,x,h^k)\tilde{\Phi}^k\right) \\ &= 0 \qquad \text{in } (0,T)\times\mathbb{R}^n. \end{aligned}$$

Rearranging,

$$\begin{aligned}
&\left(\frac{\partial\tilde{\Phi}^{k+1}}{\partial t} - \frac{\partial\tilde{\Phi}^{k}}{\partial t}\right) + \left(\frac{1}{2}\mathrm{tr}\left[\left(\Lambda\Lambda^T(t)D^2\tilde{\Phi}^{k+1}\right) - \left(\Lambda\Lambda^T(t)D^2\tilde{\Phi}^{k}\right)\right]\right)\\
&+ f(x,h^k)^T\left(D\tilde{\Phi}^{k+1} - D\tilde{\Phi}^{k}\right) + \theta g(t,x,h^k)\left(\tilde{\Phi}^{k+1} - \tilde{\Phi}^{k}\right)\\
&+ \left(f(x,h^k)^T D\tilde{\Phi}^{k} + \theta g(t,x,h^k)\tilde{\Phi}^{k}\right) - \left(f(x,h^{k-1})^T D\tilde{\Phi}^{k} + \theta g(t,x,h^{k-1})\tilde{\Phi}^{k}\right)\\
=&\ 0 \qquad \text{in } (0,T)\times\mathbb{R}^n.
\end{aligned}$$

Define the function $\ell^k(t,x)$ as

$$\ell^k(t,x) := \left(f(x,h^k)^T D\tilde{\Phi}^{k} + \theta g(t,x,h^k)\tilde{\Phi}^{k}\right) - \left(f(x,h^{k-1})^T D\tilde{\Phi}^{k} + \theta g(t,x,h^{k-1})\tilde{\Phi}^{k}\right).$$

By the definition of h^k given in (7.7), $\ell^k(t,x) \leq 0\ \forall (t,x) \in [0,T]\times\mathbb{R}^n, \forall k \in \mathbb{N}$. Define the sequence of functions $(W^k)_{k\in\mathbb{N}}$ as

$$W^k := \tilde{\Phi}^{k+1} - \tilde{\Phi}^{k};$$

then W^k satisfies the PDE

$$\frac{\partial W^k}{\partial t} + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T(t)D^2W^k\right) + f(x,h^k)^T DW^k + \theta g(t,x,h^k)W^k + \ell^k(t,x) = 0 \tag{7.9}$$

in $(0,T)\times\mathscr{B}_R$ and has boundary condition $W^k(T,x) = 0$ on $\partial^* Q_R = ((0,T)\times\partial\mathscr{B}_R)\cup(\{T\}\times\mathscr{B}_R)$.

Define the stopping time τ_G as the first exit time from $\mathscr{B}_R$:

$$\tau_G := \inf\{t : X(t) \notin G\}.$$

By a standard Feynman–Kac representation, $W^k(t,x)$ can be represented by the expectation

$$W^k(t,x) = \mathbf{E}\left[\int_t^{T\wedge\tau_G} \ell^k(s,X_s)e^{\theta\int_0^s g(r,X_r)dr}ds\right]. \tag{7.10}$$

Because $\ell(t,x) \leq 0$, $W^k(t,x) \leq 0$ for $k \geq 1$, and hence, by definition of W^k,

$$\tilde{\Phi}^k \geq \tilde{\Phi}^{k+1} \qquad \forall k \in \mathbb{N},$$

which implies that the sequence $\{\tilde{\Phi}^k\}_{k\in\mathbb{N}}$ is nonincreasing.

Step 2.2: Convergence of the sequence. Since the sequence $(\tilde{\Phi}^k)_{k\in\mathbb{N}}$ is nonincreasing and is also bounded, it converges. Denote by $\tilde{\Phi}$ its limit as $k\to\infty$. Now, since the Sobolev-type norm $\|\tilde{\Phi}^{k+1}\|^{(2)}_{\eta,Q_R}$ is bounded for $1<\eta<\infty$, we can apply the following estimate given by equation (E.9) in Appendix E of Fleming and Rishel [16]:

$$|\tilde{\Phi}^k|^{1+\mu}_{Q_R} \leq M_R\|\tilde{\Phi}^k\|^{(2)}_{\eta,Q_R} \tag{7.11}$$

for some constant M_R (depending on R) and where

$$\mu = 1 - \frac{n+2}{\eta},$$

$$|\tilde{\Phi}^k|^{1+\mu}_{Q_R} = |\tilde{\Phi}^k|^{\mu}_{Q_R} + \sum_{i=1}^{n} |\tilde{\Phi}^k_{x_i}|^{\mu}_{Q_R},$$

and

$$\begin{aligned} |\tilde{\Phi}^k|^{\mu}_{Q_R} = & \sup_{(t,x)\in Q_R} |\tilde{\Phi}^k(t,x)| + \sup_{\substack{(x,y)\in\overline{G} \\ 0\le t\le T}} \frac{|\tilde{\Phi}^k(t,x) - \tilde{\Phi}^k(t,y)|}{|x-y|^{\mu}} \\ & + \sup_{\substack{x\in\overline{G} \\ 0\le s,t\le T}} \frac{|\tilde{\Phi}^k(s,x) - \tilde{\Phi}^k(t,x)|}{|s-t|^{\mu/2}} \end{aligned}$$

to show that the Hölder-type norm $|\tilde{\Phi}^k|^{1+\mu}_{Q_R}$ is bounded. As $k \to \infty$ we conclude the following:

- $D\tilde{\Phi}^k$ converges to $D\tilde{\Phi}$ uniformly in $L^{\eta}(Q_R)$;
- $D^2\tilde{\Phi}^k$ converges to $D^2\tilde{\Phi}$ weakly in $L^{\eta}(Q_R)$; and
- $\frac{\partial \tilde{\Phi}^k}{\partial t}$ converges to $\frac{\partial \tilde{\Phi}}{\partial t}$ weakly in $L^{\eta}(Q_R)$.

Step 2.3: Proving that $\tilde{\Phi} \in C^{1,2}(Q_R)$. Using estimate (7.11), we see that $|\tilde{\Phi}^k|^{1+\mu}_{Q_R}$ is bounded for $\mu > 0$, which implies that $\eta > n+2$. Using relationship (7.8) and then (7.6), we get

$$\begin{aligned} & \frac{\partial \tilde{\Phi}^k}{\partial t} + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T(t) D^2\tilde{\Phi}^k\right) + f(x,h)^T D\tilde{\Phi}^k + \theta g(t,x,h)\tilde{\Phi}^k \\ \ge & \frac{\partial \tilde{\Phi}^k}{\partial t} + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T(t) D^2\tilde{\Phi}^k\right) + f(x,h^k)^T D\tilde{\Phi}^k + \theta g(t,x,h^k)\tilde{\Phi}^k \\ = & \left(\frac{\partial \tilde{\Phi}^k}{\partial t} - \frac{\partial \tilde{\Phi}^{k+1}}{\partial t}\right) + \left(\frac{1}{2}\mathrm{tr}\left[\left(\Lambda\Lambda^T(t) D^2\tilde{\Phi}^k\right) - \left(\Lambda\Lambda^T(t) D^2\tilde{\Phi}^{k+1}\right)\right]\right) \\ & + f(x,h^k)^T \left(D\tilde{\Phi}^k - D\tilde{\Phi}^{k+1}\right) + \theta g(t,x,h^k)\left(\tilde{\Phi}^k - \tilde{\Phi}^{k+1}\right) \end{aligned} \tag{7.12}$$

for any admissible control h.

Since the left-hand side of (7.12) tends weakly in $L^{\eta}(Q_R)$ to

$$\frac{\partial \tilde{\Phi}}{\partial t} + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T(t) D^2\tilde{\Phi}\right) + f(x,h)^T D\tilde{\Phi} + \theta g(t,x,h)\tilde{\Phi} \tag{7.13}$$

as $k \to \infty$ and the right-hand side tends tends weakly to 0, we obtain the following inequality:

$$\frac{\partial \tilde{\Phi}}{\partial t} + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T(t) D^2\tilde{\Phi}\right) + f(x,h)^T D\tilde{\Phi} + \theta g(t,x,h)\tilde{\Phi} \ge 0$$

a.e. in Q_R.

Using a measurable selection theorem and following an argument similar to that of Lemma VI.6.1 of Fleming and Rishel [16], we see that there exists a Borel measurable function h^* from $(0,T)\times\mathscr{B}_R$ into $\overline{\mathcal{J}}$ (in fact $\mathcal{J}$) such that

$$f(x,h^*)^T D\tilde{\Phi} + \theta g(t,x,h^*)\tilde{\Phi} = \inf_{h\in\overline{\mathcal{J}}}\left\{f(x,h)^T D\tilde{\Phi} + \theta g(t,x,h)\tilde{\Phi}\right\}$$

holds for almost all $(t,x)\in(0,T)\times\mathscr{B}_R$. Then

$$\begin{aligned}
&\frac{\partial\tilde{\Phi}^k}{\partial t} + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T(t)D^2\tilde{\Phi}^k\right) + f(x,h^*)^T D\tilde{\Phi}^k + \theta g(t,x,h^*)\tilde{\Phi}^k\\
&\leq \frac{\partial\tilde{\Phi}^k}{\partial t} + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T(t)D^2\tilde{\Phi}^k\right) + f(x,h^k)^T D\tilde{\Phi}^k + \theta g(t,x,h^k)\tilde{\Phi}^k\\
&= \left(\frac{\partial\tilde{\Phi}^k}{\partial t} - \frac{\partial\tilde{\Phi}^{k+1}}{\partial t}\right) + \left(\frac{1}{2}\mathrm{tr}\left[\left(\Lambda\Lambda^T(t)D^2\tilde{\Phi}^k\right) - \left(\Lambda\Lambda^T(t)D^2\tilde{\Phi}^{k+1}\right)\right]\right)\\
&\quad + f(x,h^k)^T\left(D\tilde{\Phi}^k - D\tilde{\Phi}^{k+1}\right) + \theta g(t,x,h^k)\left(\tilde{\Phi}^k - D\tilde{\Phi}^{k+1}\right).
\end{aligned}\tag{7.14}$$

Since the left-hand side of (7.14) tends weakly in $L^\eta(Q_R)$ to

$$\frac{\partial\tilde{\Phi}}{\partial t} + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T(t)D^2\tilde{\Phi}\right) + f(x,h^*)^T D\tilde{\Phi} + \theta g(t,x,h^*)\tilde{\Phi}$$

as $k\to\infty$ and the right-hand side tends weakly to 0, we obtain the inequality

$$\frac{\partial\tilde{\Phi}}{\partial t} + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T(t)D^2\tilde{\Phi}\right) + f(x,h^*)^T D\tilde{\Phi} + \theta g(t,x,h^*)\tilde{\Phi} \leq 0 \tag{7.15}$$

a.e. in Q_R.

Combining (7.13) and (7.15), we have shown that

$$\frac{\partial\tilde{\Phi}}{\partial t} + \frac{1}{2}\mathrm{tr}\left(\Lambda\Lambda^T(t)D^2\tilde{\Phi}\right) + f(x,h^*)^T D\tilde{\Phi} + \theta g(t,x,h^*)\tilde{\Phi} = 0$$

a.e. in Q_R.

Hence, $\tilde{\Phi}$ is a solution of (3.16) on a bounded domain. Moreover, $\tilde{\Phi}\in\mathcal{L}^\eta(Q_R)$. Also, since H is locally Lipschitz, $|\tilde{\Phi}^k|^\mu_{Q_R}<\infty$ for $\mu>0$, and $|D\tilde{\Phi}^k|^\mu_{Q_R}<\infty$ for $\mu>0$, $|H(t,x,\tilde{\Phi}^k,D\tilde{\Phi}^k)|^\mu_{Q_R}$ $<\infty$.

We can now show that $\tilde{\Phi}\in C^{1,2}(Q_R)$. Define

$$|\tilde{\Phi}^k|^{2+\mu}_{Q_R} := |\tilde{\Phi}^k|^{1+\mu}_{Q_R} + \left|\frac{\partial\tilde{\Phi}^k}{\partial t}\right|^\mu_{Q_R} + \sum_{i,j=1}^n|\tilde{\Phi}^k_{x_ix_j}|^\mu_{Q_R}.$$

Consider the following estimate given by equation (E.10) in Appendix E of Fleming and Rishel [16]:

$$|\tilde{\Phi}|^{2+\mu}_{Q'} \leq M_2\|\tilde{\Phi}\|_{Q''} \tag{7.16}$$

for some constant M_2 and two open subsets Q' and Q'' of Q such that $\bar{Q}' \subset \bar{Q}''$. In this estimate, set $Q'' = Q_R$, and take Q' to be any subset of Q such that $\bar{Q}' \subset Q$. Thus

$$|\tilde{\Phi}|_{Q'}^{2+\mu} < \infty. \tag{7.17}$$

When interpreted in light of estimate (7.11) (stemming from (E9)), we see that the derivatives $\frac{\partial \tilde{\Phi}}{\partial t}$, $\frac{\partial \tilde{\Phi}}{\partial x_i}$, and $\frac{\partial^2 \tilde{\Phi}}{\partial x_i x_j}$ satisfy a uniform Hölder condition on any compact subset Q' of Q_R. By Theorem 10.1 in Chapter IV of Ladyženskaja, Solonnikov, and Uralceva [28], we can therefore conclude that $\tilde{\Phi} \in C^{1,2}(Q_R)$.

Step 3: Convergence from the cylinder $[0,T) \times \mathscr{B}_R$ to the state space $[0,T) \times \mathbb{R}^n$.

Step 3.1: Setting. Let $\{R_i\}_{i\in\mathbb{N}} > 0$ be a nondecreasing sequence with $\lim_{i\to\infty} R_i \to \infty$, and let $\{\tau_i\}_{i\in\mathbb{N}}$ be the sequence of stopping times defined as

$$\tau_i := \inf \{t : X(t) \notin \mathscr{B}_{R_i}\}.$$

Note that $\{\tau_i\}_{i\in\mathbb{N}}$ is nondecreasing and $\lim_{i\to\infty} \tau_i = \infty$.

Denote by $\tilde{\Phi}^{(i)}$ the limit of the sequence $(\tilde{\Phi}^k)_{k\in\mathbb{N}}$ on $(0,T) \times \mathscr{B}_{R_i}$, i.e.,

$$\tilde{\Phi}^{(i)}(t,x) = \lim_{k\to\infty} \tilde{\Phi}^k(t,x) \qquad \forall (t,x) \in (0,T) \times \mathscr{B}_{R_i}. \tag{7.18}$$

Step 3.2: Convergence of the sequence $(\tilde{\Phi}^{(i)})_{i\in\mathbb{N}}$. First, observe that the sequence $(\tilde{\Phi}^{(i)})_{i\in\mathbb{N}}$ is nonincreasing. Indeed, for $i < j$, the stochastic control problem defined over $(0,T) \times \mathscr{B}_{R_i}$ is nested into the stochastic control problem defined over $(0,T) \times \mathscr{B}_{R_j}$. In particular, a suboptimal strategy for the stochastic control problem defined over $(0,T) \times \mathscr{B}_{R_j}$ would be to invest optimally while $x \in \mathscr{B}_{R_i}$ and then switch to the 0β policy $\check{h}$ when $x \in \mathscr{B}_{R_j} \backslash \mathscr{B}_{R_i}$. By the optimality principle, the expected total cost of such a strategy is greater than the value function $\tilde{\Phi}^{(j)}$. But this suboptimal strategy also corresponds to the optimal strategy for the stochastic control problem defined over $(0,T) \times \mathscr{B}_{R_i}$. Hence

$$\tilde{\Phi}^{(i)}(t,x) \geq \tilde{\Phi}^{(j)}(t,x) \qquad \forall i,j \in \mathbb{N},\ \forall (t,x) \in (0,T) \times \mathscr{B}_{R_i}.$$

By the argument in Step 1, the sequence $(\tilde{\Phi}^{(i)})_{i\in\mathbb{N}}$ is also bounded. As a result, it converges to a limit $\tilde{\Phi}$. This limit satisfies the boundary condition (3.17). We now show that $\tilde{\Phi}$ is $C^{1,2}$ and satisfies the HJB PDE. These statements are local properties, so we can restrict ourselves to a finite ball Q_R.

Step 3.3: Proving that $\tilde{\Phi} \in C^{1,2}(Q_R)$. Now that we have shown the convergence of the sequence $(\tilde{\Phi}^{(i)})_{i\in\mathbb{N}}$ through a simple control-based argument, we can conclude the proof using the same arguments based on Ascoli's theorem as those of Fleming and Rishel (see [16], proof of Theorem 6.2 in Appendix E). ■

Corollary 7.3 (existence of a classical solution for the risk-sensitive control problem). *The RS HJB PDE (3.12) with terminal condition $\Phi(T,x) = \ln v$ has a solution $\Phi \in C^{1,2}([0,T] \times \mathbb{R}^n)$ with Φ continuous in $\overline{[0,T] \times \mathbb{R}^n}$.*

8. Partial observation. In this section we show how the results of this paper can be extended to the case where the factor process $X(t)$ is not directly observed and the asset allocation strategy h_t must be adapted to the filtration $\mathcal{F}_t^S = \sigma\{S_i(u), 0 \leq u \leq t, j = 0, \dots, m\}$ generated by the asset price processes alone. In the linear diffusion case studied by Nagai [31] and Nagai and Peng [32], the authors noted that the pair of processes $(X(t), Y(t))$, where $Y_i(t) = \log S_i(t)$, takes the form of the "signal" and "observation" processes in a Kalman filter system, and consequently the conditional distribution of $X(t)$ is normal $N(\hat{X}(t), P(t))$, where $\hat{X}(t) = \mathbf{E}[X(t)|\mathcal{F}_t^S]$ satisfies the Kalman filter equation and $P(t)$ is a deterministic matrix-valued function. By using this idea they obtain an equivalent form of the problem in which $X(t)$ is replaced by $\hat{X}(t)$ and the dynamic equation (2.3) by the Kalman filter. Optimal strategies take the form $h(t, \hat{X}(t))$. This is in fact a very old idea in stochastic control, going back at least to Wonham [41].

8.1. Decomposition. At first sight it does not seem apparent that the same approach can be used here, as the price processes contain jumps, but a simple observation shows that the jumps play no role in the estimation process, which is still, at the base, the Kalman filter; see Proposition 8.1. A further complication is that the money market interest rate $r(t) = a_0 + A_0^T X(t)$ (see (2.4)) is observed directly and contains information about $X(t)$. This was not the case in [31] and [32], where, in our notation, $A_0 = 0$. We start by assuming that $A_0 = 0$ and briefly discuss the extension to $A_0 \neq 0$ at the end of the section.

Recall first that $X(t)$ satisfies

$$dX(t) = (b + BX(t))dt + \Lambda dW(t), \qquad X(0) = X_0. \tag{8.1}$$

When X_t is observed, the initial value X_0 is just a constant. In the present case we need to assume that X_0 is a normal random vector $N(m_0, P_0)$ with known mean m_0 and covariance P_0 and that X_0 is independent of the processes $W, N_{\mathbf{p}}$.

An application of the general Itô formula[3] shows that for $i = 1, \dots, m$ the log-prices $Y_i(t)$ satisfy $Y_i(0) = \log s_i$ and

$$\begin{aligned} dY_i(t) = {} & \left[(\hat{a} + \hat{A}X(t))_i - \frac{1}{2}\Sigma\Sigma_{ii}^T \right] dt + \sum_{k=1}^N \sigma_{ik} dW_k(t) \\ & + \int_{\mathbf{Z}_0} \left\{ \ln\left(1 + \gamma_i(z)\right) - \gamma_i(z) \right\} \nu(dz)dt + \int_{\mathbf{Z}} \ln\left(1 + \gamma_i(z)\right) \bar{N}(dt, dz). \end{aligned} \tag{8.2}$$

Proposition 8.1. *Define processes $Y^1(t), Y^2(t) \in \mathbb{R}^m$ as follows:*

$$\begin{aligned} dY^1(t) &= \hat{A}X(t) + \Sigma dW(t), & & Y_i^1(0) = 0, \\ dY_i^2(t) &= c_i dt + \int_{\mathbf{Z}} \ln\left(1 + \gamma(z)\right)_i \bar{N}(dt, dz), & i = 1, \dots, m, \qquad & Y_i^2(0) = \log s_i \end{aligned} \tag{8.3}$$

with $c \in \mathbb{R}^m$ defined by

$$c_i := \hat{a}_i - \frac{1}{2}\Sigma\Sigma_{ii}^T + \int_{\mathbf{Z}_0} \left\{ \ln\left(1 + \gamma_i(z)\right) - \gamma_i(z) \right\} \nu(dz)$$

[3] See Øksendal and Sulem [34] for this calculation.

so that $Y(t) = Y^1(t) + Y^2(t)$*. Also, define* $\mathcal{Y}_{it} = \sigma\{Y^i(u), 0 \leq u \leq t\}$, $i = 1, 2$*. Then the following hold:*

(i) *The processes* Y^1, Y^2 *are each adapted to the filtration* $\mathcal{F}_t^S$*.*

(ii) *For any bounded measurable function* f *and* $t \geq 0$,

$$\mathbf{E}[f(X(t))|\mathcal{F}_t^S] = \mathbf{E}[f(X(t))|\mathcal{Y}_{1t}].$$

Proof. (i) $S(t)$ and $Y(t)$ are in 1-1 correspondence and therefore generate the same filtration $\mathcal{F}_t^S$. Apart from rearrangement of deterministic terms, the decomposition $Y = Y^1 + Y^2$ is the same as the standard decomposition $Y = Y^c + Y^d$ of a semimartingale into its continuous and discontinuous components; see paragraph VI. 37 of Rogers and Williams [37].

(ii) N and $W(t)$ are independent; as a result $\mathcal{Y}_{1t}$ and $\mathcal{Y}_{2t}$ are independent, and clearly $\mathcal{F}_t^S = \mathcal{Y}_{1t} \vee \mathcal{Y}_{2t}$. The result follows, since $X(t)$ is independent of $\mathcal{Y}_{2t}$. ■

8.2. Kalman filter. The processes $(X(t), Y^1(t))$ satisfying (8.1) and (8.3) and the filtering equations, which are standard, are stated in the following proposition.

Proposition 8.2 (Kalman filter). *The conditional distribution of* $X(t)$ *given* $\mathcal{Y}_{1t}$ *is* $N(\hat{X}(t), P(t))$*, calculated as follows:*

(i) *The* innovation process $U(t) \in \mathbb{R}^m$ *defined by*

$$dU(t) = \left(\Sigma\Sigma^T\right)^{-1/2}(dY^1(t) - \hat{A}\hat{X}(t)dt), \qquad U(0) = 0 \tag{8.4}$$

is a vector Brownian motion.

(ii) $\hat{X}(t)$ *is the unique solution of the SDE*

$$d\hat{X}(t) = (b + B\hat{X}(t))dt + \left(\Lambda\Sigma^T + P(t)\hat{A}^T\right)\left(\Sigma\Sigma^T\right)^{-1/2} dU(t), \qquad \hat{X}(0) = m_0. \tag{8.5}$$

(iii) $P(t)$ *is the unique nonnegative definite symmetric solution of the matrix Riccati equation*

$$\begin{aligned}\dot{P}(t) = \Lambda\Xi\Xi^T\Lambda^T - P(t)\hat{A}^T\left(\Sigma\Sigma^T\right)^{-1}\hat{A}P(t) + \left(B - \Lambda\Sigma^T\left(\Sigma\Sigma^T\right)^{-1}\hat{A}\right)P(t)\\ + P(t)\left(B^T - \hat{A}^T\left(\Sigma\Sigma^T\right)^{-1}\Sigma\Lambda^T\right), \qquad P(0) = P_0,\end{aligned}$$

where $\Xi := I - \Sigma^T\left(\Sigma^T\Sigma\right)^{-1}\Sigma$.

To conclude, the Kalman filter has replaced our initial state process $X(t)$ by an estimate $\hat{X}(t)$ with dynamics given in (8.5). To recover the asset price process, we use (8.3) together with (8.4) to obtain the dynamics of $Y(t)$:

$$\begin{aligned}dY_i(t) &= dY_i^1(t) + dY_i^2(t)\\ &= \hat{a}_i + \hat{A}\hat{X}(t)dt - \frac{1}{2}\Sigma\Sigma_{ii}^T dt + \left(\Sigma\Sigma^T\right)^{1/2} dU(t)\\ &\quad + \int_{\mathbf{Z}_0}\{\ln(1 + \gamma_i(z)) - \gamma_i(z)\}\,\nu(dz) + \int_{\mathbf{Z}}\ln(1 + \gamma(z))_i\,\bar{N}(dt, dz).\end{aligned} \tag{8.6}$$

We then apply Itô's lemma to $S_i(t) = \exp Y_i(t)$ to get

$$\begin{aligned}\frac{dS_i(t)}{S_i(t^-)} &= (a + A\hat{X}(t))_i dt + \sum_{k=1}^N\left[\left(\Sigma\Sigma^T\right)^{1/2}\right]_{ik} dU_k(t) + \int_{\mathbf{Z}}\gamma_i(z)\bar{N}(dt, dz),\\ S_i(0) &= s_i, \quad i = 1, \ldots, m.\end{aligned} \tag{8.7}$$

We now solve the stochastic control problem with partial observation simply by replacing the original asset price description (2.5) by (8.7), and the factor process description (2.3) by the Kalman filter equation (8.5), in our solution of the full observation case. The Kalman filter has time-varying coefficients, but this does not affect the preceding arguments.

Finally, we briefly sketch what to do if $A_0 \neq 0$. We observe the short rate $r(t) = a_0 + A_0^T X(t)$, and hence the 1-dimensional statistic $Y_0(t) \equiv A_0^T X(t)$, exactly. We need to assume that this observation contains positive "noise," i.e., $A_0^T \Lambda\Lambda^T A_0 > 0$. Changing coordinates if necessary, we can assume that $A_0^T = (0, 0, \ldots, 1)$ and hence $Y_0(t) = X_n(t)$. Our "observation" is now the $(m+1)$-dimensional process $\bar{Y} = (Y_0, \ldots, Y_m)$, and we can set up a Kalman filter system to estimate the unobserved states $\bar{X} = (X_1, \ldots, X_{n-1})^T \in \mathbb{R}^{n-1}$. Ultimately, our optimal strategy will take the form $h(t, X_1(t), \hat{\bar{X}}(t))$, where $\hat{\bar{X}}(t)$ is the Kalman filter estimate for $\bar{X}(t)$ given $\{\bar{Y}(u), u \leq t\}$. The details are left to the reader.

9. Conclusion. In this article, we extended the classical risk-sensitive asset management setting to include the possibility of infinite activity jumps in asset prices. We applied the change of measure technique proposed by Kuroda and Nagai [26] to derive the HJB PDE associated with the control problem and then proved the existence and uniqueness of an admissible optimal control policy. Using an approximation in policy space algorithm, we established the existence of a classical $C^{1,2}\left((0,T) \times \mathbb{R}^n\right)$ solution and obtained the uniqueness of this solution through a verification result. This approach also extends naturally and with similar results to a jump-diffusion version of the risk-sensitive benchmarked asset management problem.

Finally, we have observed that an attractive, if somewhat surprising, feature of the jump-diffusion risk-sensitive asset management is that it naturally prohibits any investment policy which may result in the investor's bankruptcy. In particular, in the risk-sensitive setting presented in this article, an investor who implements the optimal asset allocation is certain to remain solvent over the investment horizon. This contrasts with the Merton type of approach in which the threat of bankruptcy remains present and has to be accounted for using a stopping time.

The approach presented in this article extends naturally to a jump-diffusion version of the risk-sensitive benchmarked asset management problem introduced by Davis and Lleo [12] and would yield similar results, namely, the existence of a unique admissible control policy and of a classical $C^{1,2}$ solution to the associated RS HJB PDE.

Appendix. Admissibility of the optimal control policy. The admissibility of the optimal control process $h^*(t)$ solving (5.1) is linked to the existence of a probability measure $\mathbb{P}_{h^*}^{\theta}$, which itself hinges on the characterization as an exponential martingale of the Radon–Nikodym derivative $\frac{d\mathbb{P}_{h^*}^{\theta}}{d\mathbb{P}} = \chi_T^*$ defined in (3.5) via the Doléans exponential introduced in (2.9). In the setting of Kuroda and Nagai [26], the admissibility of the control follows easily from an argument in Gihman and Skorohod [21] which proves that the Doléans exponential (here a Girsanov exponential with Gaussian integrand) is an exponential martingale. However, when the Doléans exponential does not have continuous paths, as is the case in a jump-diffusion setting, proving that it is indeed a martingale is more difficult. As noted by Protter [36], some partial results exist in this case (see, for example, Mémin [30] and more recently Protter and

Shimbo [35]), but none is as powerful as their counterparts in the continuous case, namely, the Kamazaki or the Novikov conditions.

To show that the Doléans exponential introduced in (2.9) is a martingale we will apply results derived by Mémin [30]. We recall here the definition of the Doléans-Dade exponential as well as results from Mémin [30] (see also Exercise 13 in Chapter V of [36]) on the multiplicative decomposition of local martingales that we will use to prove our point.

Definition A.1 (Doléans-Dade exponential). *The Doléans-Dade exponential $\mathcal{E}(X)(t)$ of a semimartingale $X(t)$ is defined as*

$$\mathcal{E}(X)(t) = \exp\left\{X(t) - \frac{1}{2}\left[X^c, X^c\right]_t\right\} \prod_{0<s\leq t} (1+\Delta X_t)e^{-\Delta X_s}. \tag{A.1}$$

Definition A.2 (Mémin's additive decomposition of local martingales). *Let $M(t)$ be a local martingale. We define an additive decomposition of M into two processes $M_1(t)$ and $M_2(t)$, i.e., such that $M(t) = M_1(t) + M_2(t)$.*

In this decomposition, the process $M_1(t)$ is defined as $M_1(t) = L(t) - \tilde{L}(t)$, where

$$L(t) = \sum_{0<s\leq t} \Delta M_s 1_{\left\{|\Delta M_s| \geq \frac{1}{2}\right\}}$$

and $\tilde{L}(t)$ is the compensator of $L(t)$.

Proposition A.3 (Mémin's Proposition III-1). *Let $M(t)$ be a local martingale with additive decomposition as per Definition A.2 and such that $M_0 = 0$. Then the following hold:*

(i) *$\mathcal{E}(M)$ has the decomposition*

$$\mathcal{E}(M) = \mathcal{E}(M_2)\mathcal{E}(\tilde{M}_1),$$

where

$$\tilde{M}_1(t) = M_1(t) - \sum_{0<s\leq t} \frac{\Delta M_1(s)\Delta M_2(s)}{1+\Delta M_2(s)}, \qquad t < \infty.$$

(ii) *$\mathcal{E}(M_2)\tilde{M}_1$ is a local martingale.*

(iii) *If $\Delta M(s) > -1$, then $\Delta \tilde{M}_1(s) > -1$ for all finite s.*

Corollary A.4 (Mémin's Corollary III-2). *Let N be a local martingale such that $\Delta N(s) > -1$ for all finite s and such that $\mathcal{E}(N)(\infty)$ is uniformly integrable. Let $\mathbb{P}'$ be the probability defined as*

$$\frac{d\mathbb{P}'}{d\mathbb{P}} = \mathcal{E}(N)(\infty).$$

Let N_1 be a local martingale with locally integrable variations, and denote by $\tilde{N}_1$ the $\mathbb{P}$-semimartingale defined as

$$\tilde{N}_1(t) = N_1(t) - \sum_{0<s\leq t} \frac{\Delta N_1(s)\Delta N(s)}{1+\Delta N(s)}, \qquad t < \infty;$$

then $\tilde{N}_1$ is a $\mathbb{P}'$ local martingale, with locally integrable variations. Moreover, the $\mathbb{P}'$ predictable compensator of $\sum_{0<s\leq t} |\Delta \tilde{N}_1(s)|$ is equal to the $\mathbb{P}$ predictable compensator of $\sum_{0<s\leq t} |\Delta N_1(s)|$.

Theorem A.5 (Mémin's Theorem III-3). *Let $M(t)$ be a local martingale with additive decomposition as per Definition A.2. If the predictable compensator of the process*

$$Y(t) = [M^c, M^c]_t + \sum_{0<s\leq t} |\Delta M_1(s)| + \sum_{0<s\leq t} (\Delta M_2(s))^2 \tag{A.2}$$

is bounded, then $\mathcal{E}(M)(t)$ is uniformly integrable.

Proof of Proposition 6.3. To prove that the control $h^*(t)$ is admissible, we need to show that the local martingale $M^*(t)$ defined as

$$M^*(t) := -\theta \int_0^t (h^*(s))^T \Sigma dW_s - \int_0^t \int_{\mathbf{Z}} \ln\left(1 - G(z, h^*(s))\right) \tilde{N}(ds, dz) \tag{A.3}$$

and such that

$$\mathcal{E}(M)(t) = \chi_t^*$$

is an exponential martingale.

To achieve this objective, we will define a new class of control processes to which the optimal control belongs. We will start from the definition of a control h as a function:

$$\begin{aligned} h : [0,T] \times \mathbb{R}^n &\to \mathcal{J}, \\ (t,x) &\mapsto h(t,x), \end{aligned}$$

where the set $\mathcal{J}$ was defined in (2.7). Based on this definition, the control space can be viewed as a functional space.

Define the functional $\mathcal{L}(x,p,h)$ as

$$\begin{aligned} \mathcal{L}(x,p,h) :=& -\frac{1}{2}(\theta+1) h^T \Sigma\Sigma^T h - \theta h^T \Sigma \Lambda^T p + h^T(\hat{a} + \hat{A}x) \\ & - \frac{1}{\theta} \int_{\mathbf{Z}} \left\{ \left[\left(1 + h^T \gamma(z)\right)^{-\theta} - 1 \right] + \theta h^T \gamma(z) \mathit{1}_{\mathbf{Z}_0}(z) \right\} \nu(dz), \end{aligned} \tag{A.4}$$

where $p \in \mathbb{R}^n$ so that

$$\begin{aligned} \sup_{h\in\mathcal{J}} L_t^h \Phi =& (b + Bx)^T D\Phi + \frac{1}{2} \mathrm{tr}\left(\Lambda\Lambda^T D^2 \Phi\right) - \frac{\theta}{2} (D\Phi)^T \Lambda\Lambda^T D\Phi + a_0 + A_0^T x \\ & + \sup_{h\in J} \mathcal{L}(x, D\Phi, h) \end{aligned} \tag{A.5}$$

and the unique maximizer of $L_t^h\Phi(t,x)$, $\hat{h}(t,x)$, is also the unique maximizer of $\mathcal{L}(x, D\Phi, h)$.

Observe that with the choice of control function $h^0(t,x) := 0 \; \forall (t,x) \in [0,T] \times \mathbb{R}^n$ the functional $\mathcal{L}(x,p,h^0) = 0 \; \forall (t,x,p) \in [0,T] \times \mathbb{R}^n \times \mathbb{R}^n$. Invoking the optimality principle, we deduce that $\mathcal{L}(x, D\Phi, h^*(t,x)) \geq 0$.

Denote by $\hat{\mathcal{J}}$ the range of the control functions $\tilde{h}(t,x)$ such that $\mathcal{L}(x,p,\tilde{h}) \geq 0$. Under Assumption 1, the set $\mathcal{J}$, defined by (2.7), is in the interior of a hypercube, and since the functional $\mathcal{L}(x,p,h)$ is smooth and strictly concave in h and $\lim_{h\to\partial\mathcal{J}} \mathcal{L}(x,p,h) = -\infty$, we deduce that the set $\hat{\mathcal{J}}$ is a closed convex subset of $\mathcal{J}$ $\forall(t,x) \in [0,T]\times\mathbb{R}^n$. The control functions $\tilde{h}$ take the form

$$\begin{aligned}\tilde{h} : [0,T] \times \mathbb{R}^n &\to \hat{\mathcal{J}} \subset \mathcal{J},\\ (t,x) &\mapsto \tilde{h}(t,x).\end{aligned}$$

More formally, we can define a class $\hat{\mathcal{H}}(T)$ of Markov control processes as follows.

Definition A.6. *A control process $\tilde{h}(t)$ is in class $\hat{\mathcal{H}}(T)$ if the following conditions are satisfied:*

1. *$\tilde{h}(t)$ is in class $\mathcal{H}$ introduced in Definition 2.1;*
2. *$\tilde{h}(t,x) \in \hat{\mathcal{J}}$ $\forall(t,x) \in [0,T] \times \mathbb{R}^n$.*

In particular, we note that the optimal control process $h^*(t) \in \hat{\mathcal{H}}(T)$ $\forall t \in [0,T]$ and $\forall\omega \in \Omega$.

For any control policy $\tilde{h}(t) \in \hat{\mathcal{H}}(T)$, define the local martingale $\hat{M}(t)$ as

$$\hat{M}(t) := -\theta \int_0^t \tilde{h}(s)^T \Sigma dW_s - \int_0^t \int_{\mathbf{Z}} \ln\left(1 - G(z,\tilde{h}(s))\right) \tilde{N}(ds,dz). \tag{A.6}$$

Also, let $L(t)$ be the process defined as

$$\begin{aligned}L(t) &= \sum_{0<s\leq t} \Delta Y_s 1_{\{|\Delta Y_s| \geq \frac{1}{2}\}}\\ &= -\int_0^t \int_{\mathbf{Z}\setminus\mathbf{Z}_1} \ln\left(1 - G(z,\tilde{h}(s))\right) N(ds,dz),\end{aligned}$$

where $\mathbf{Z}_1 = \left\{z \in \mathbf{Z} : |\Delta Y_s| < \frac{1}{2}, 0 \leq s \leq t\right\}$. Then the process $M_1(t) := L(t) - \tilde{L}(t)$ can be expressed as

$$M_1(t) = -\int_0^t \int_{\mathbf{Z}\setminus\mathbf{Z}_1} \ln\left(1 - G(z,\tilde{h}(s))\right) \tilde{N}(ds,dz).$$

To complete our decomposition of the local martingale $M(t)$, we define the process $M_2(t)$ as

$$\begin{aligned}M_2(t) &= M(t) - M_1(t)\\ &= -\int_0^t \int_{\mathbf{Z}_1} \ln\left(1 - G(z,\tilde{h}(s))\right) \tilde{N}(ds,dz).\end{aligned}$$

The next step is to study each component of the process $Y(t)$ defined in (A.2):

- The process
$$[M^c, M^c]_t = \exp\left\{\theta^2 \int_0^t \tilde{h}(s)^T \Sigma\Sigma^T \tilde{h}(s) ds\right\}$$
is clearly bounded because $\tilde{h}(s) \in \hat{\mathcal{H}}(T)$ $\forall s \in [0,t]$.
- The process
$$\sum_{0<s\leq t} |\Delta M_1(s)| = \int_0^t \int_{\mathbf{Z}\setminus\mathbf{Z}_1} \left|\ln\left(1 - G(z,\tilde{h}(s))\right)\right| N(ds,dz)$$

is bounded because $\tilde{h}(s) \in \hat{\mathcal{H}}(T)$ $\forall s \in [0,t]$. In addition, the number of jumps greater than $\frac{1}{2}$ is finite:

$$\begin{aligned}\#\left\{0 \leq s \leq t; |\Delta M_1(s)|\right\} &= \#\left\{0 \leq s \leq t; |\Delta M(s)| \mathit{1}_{\{|\Delta M_s| \geq \frac{1}{2}\}}\right\} \\ &= N\left(t, \left]-\infty, -\frac{1}{2}\right[\cup \left]\frac{1}{2}, \infty\right[\right) \\ &< \infty.\end{aligned}$$

- Finally, we turn our attention to the process

$$\sum_{0<s\leq t} (\Delta M_2(s))^2 = \int_0^t \int_{\mathbf{Z}_1} \left| \ln\left(1 - G(z, \tilde{h}(s))\right)\right|^2 N(ds, dz).$$

Recalling that we assumed that in our setting

$$\int_{\mathbf{Z}_0} |\gamma(z)|^2 \nu(dz) < \infty$$

and taking into consideration the fact that $\tilde{h}(s) \in \hat{\mathcal{H}}(T) \forall s \in [0,t]$, then we deduce that

$$\int_{\mathbf{Z}_0} \left| \ln\left(1 - G(z, \tilde{h}(s))\right)\right|^2 \nu(dz) < \infty$$

for any $\omega \in \Omega$, which proves that

$$\int_0^t \int_{\mathbf{Z}_1} \left| \ln\left(1 - G(z, \tilde{h}(s))\right)\right|^2 N(ds, dz) < \infty.$$

By Theorem A.5, the Doléans-Dade exponential

$$\mathcal{E}(\hat{M})(t) = \chi_t^*$$

is uniformly integrable for all $\tilde{h} \in \hat{\mathcal{H}}(T)$. We can now apply Corollary A.4 to formally define the measure $\mathbb{P}_{\tilde{h}}^\theta$. In particular, the measure $\mathbb{P}_{h^*}^\theta$ characterized via the Radon–Nikodym derivative χ_t^* is well defined because $h^*(t) \in \hat{\mathcal{H}}(T)$ $\forall \omega \in \Omega$. This proves that the control $h^*(t)$ is admissible $\forall t \in [0,T]$ and $\omega \in \Omega$. ■

Note that the control policy $h^0(t) = 0$ corresponds to investing the entire wealth into the money market asset for the duration of the investment period. The associated measure $\mathbb{P}_{h^0}^\theta$ is well defined, and it is equal to the physical measure $\mathbb{P}$. In fact, this proof not only shows that the optimal control process $h^*(t)$ is admissible but also that a large class of "reasonable" control processes $\tilde{h}(t)$ is also admissible and is associated with a well-defined probability measure.

Proof of Proposition 6.4. Consider the exponentially transformed problem $\inf_{h\in\mathcal{A}(T)} \tilde{J}(x,t,h)$, where

$$\tilde{J}(x,t,h) := \ln \mathbf{E}\left[e^{-\theta \ln V(t,x,h)}\right]. \tag{A.7}$$

Note that because the term $e^{-\theta \ln V(t,x,h)}$ is bounded from below by 0, $\inf_{h\in\mathcal{A}(T)} \tilde{J}(x,t,h)$ is well defined, which implies that there exists at least one minimizer $\tilde{h}$:

$$\mathbf{E}\left[e^{-\theta \ln V(t,x,h)}\right] = \mathbf{E}^h_{t,x}\left[\exp\left\{\theta \int_t^T g(s,X_s,h(s))ds - \theta \ln v\right\}\right]$$

(see, for example, Lemma 8.6.2 in [33]) and hence

$$\begin{aligned}\inf_{h\in\mathcal{A}(T)} \mathbf{E}\left[e^{-\theta \ln V(t,x,h)}\right] &= \inf_{h\in\mathcal{A}(T)} \mathbf{E}^h_{t,x}\left[\exp\left\{\theta \int_t^T g(s,X_s,h(s))ds - \theta \ln v\right\}\right] \\ &= I(v,x;h^*(t);t,T),\end{aligned}$$

which proves that the optimal control $h^*(t)$ for the auxiliary problem $\sup_{h\in\mathcal{A}(T)} I(v,x;h;t,T)$ derived in section 3.3 is indeed optimal for the problem $\sup_{h\in\mathcal{A}(T)} J(x,t,h)$. ∎

REFERENCES

[1] G. Barles and P. E. Souganidis, *Convergence of approximation schemes for fully nonlinear second order equations*, J. Asymptotic Anal., 4 (1991), pp. 271–283.

[2] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.

[3] A. Bensoussan, J. Frehse, and H. Nagai, *Some results on risk-sensitive control with full observation*, Appl. Math. Optim., 37 (1998), pp. 1–41.

[4] T. R. Bielecki, D. Hernandez-Hernandez, and S. R. Pliska, *Risk sensitive asset management with constrained trading strategies*, in Recent Developments in Mathematical Finance, World Scientific, Singapore, 2002, pp. 127–138.

[5] T. R. Bielecki and S. R. Pliska, *Risk-sensitive dynamic asset management*, Appl. Math. Optim., 39 (1999), pp. 337–360.

[6] T. R. Bielecki and S. R. Pliska, *Risk sensitive asset management with transaction costs*, Finance Stoch., 4 (2000), pp. 1–33.

[7] T. R. Bielecki and S. R. Pliska, *Economic properties of the risk sensitive criterion for portfolio management*, The Review of Accounting and Finance, 2 (2003), pp. 3–17.

[8] T. R. Bielecki and S. R. Pliska, *Risk sensitive intertemporal CAPM with applications to fixed-income management*, IEEE Trans. Automat. Control, 49 (2004), pp. 420–432.

[9] T. R. Bielecki, S. R. Pliska, and S.-J. Sheu, *Risk sensitive portfolio management with Cox–Ingersoll–Ross interest rates: The HJB equation*, SIAM J. Control Optim., 44 (2005), pp. 1811–1843.

[10] F. Black, *Capital market equilibrium with restricted borrowing*, J. Business, 45 (1972), pp. 445–454.

[11] M. Crandall, H. Ishii, 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] M. H. A. Davis and S. Lleo, *Risk-sensitive benchmarked asset management*, Quant. Finance, 8 (2008), pp. 415–426.

[13] M. H. A. Davis and S. Lleo, *Risk-sensitive asset management and affine processes*, in Recent Advances in Financial Engineering 2009: Proceedings of the KIER-TMU International Workshop on Financial Engineering 2009, M. Kijima, C. Hara, and K. Tanaka, eds., World Scientific, Singapore, 2010, pp. 1–41.

[14] M. H. A. Davis and S. Lleo, *Jump-Diffusion Risk-Sensitive Asset Management* II: *Jump-Diffusion Factors*, preprint, Imperial College London, London, England, 2010.

[15] W. H. Fleming, *Optimal investment models and risk-sensitive stochastic control*, in Mathematical Finance, IMA Vol. Math. Appl. 65, Springer, Berlin, 1995, pp. 75–88.

[16] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer, Berlin, New York, 1975.

[17] W. H. Fleming and S. J. Sheu, *Optimal long term growth rate of expected utility of wealth*, Ann. Appl. Probab., 9 (1999), pp. 871–903.

[18] W. H. Fleming and S. J. Sheu, *Risk-sensitive control and an optimal investment model*, Math. Finance, 10 (2000), pp. 197–213.
[19] W. H. Fleming and S. J. Sheu, *Risk-sensitive control and an optimal investment model* II, Ann. Appl. Probab., 12 (2000), pp. 730–767.
[20] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, 2nd ed., Stoch. Model. Appl. Probab. 25, Springer, New York, 2006.
[21] I. I. Gihman and A. Skorohod, *Stochastic Differential Equations*, Springer, New York, 1972.
[22] L. P. Hansen and T. J. Sargent, *Robustness*, Princeton University Press, Princeton, NJ, 2008.
[23] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, North–Holland, Amsterdam, 1981.
[24] N. V. Krylov, *Controlled Diffusion Processes*, Springer, Berlin, New York, 1980.
[25] N. V. Krylov, *Nonlinear Elliptic and Parabolic Equations of the Second Order*, D. Reidel, Dordrecht, The Netherlands, 1987.
[26] K. Kuroda and H. Nagai, *Risk-sensitive portfolio optimization on infinite time horizon*, Stoch. Stoch. Rep., 73 (2002), pp. 309–331.
[27] H. J. Kushner and P. G. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, 2nd ed., Appl. Math. (N. Y.) 24, Springer, New York, 2001.
[28] O. A. Ladyženskaja, V. A. Solonnikov, and O. O. Uralceva, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.
[29] M. Lefebvre and P. Montulet, *Risk-sensitive optimal investment policy*, Internat. J. Systems Sci., 25 (1994), pp. 183–192.
[30] J. Mémin, *Décomposition in multiplicatives de semimartingales exponentielles et applications*, Séminaires de Probabilités XII, Lecture Notes in Math. 649, Springer, Berlin, 1978, pp. 35–46.
[31] H. Nagai, *Risk-sensitive dynamic asset management with partial information*, in Stochastics in Finite and Infinite Dimensions, in Honor of Gopinath Kallianpur, T. Hida, R. L. Karandikar, H. Kunita, B. S. Rajput, and S. Watanabe, eds., Birkhäuser, Boston, 2001, pp. 321–339.
[32] H. Nagai and S. Peng, *Risk-sensitive dynamic portfolio optimization with partial information on infinite time horizon*, Ann. Appl. Probab., 12 (2002), pp. 173–195.
[33] B. Øksendal, *Stochastic Differential Equations*, 6th ed., Universitext, Springer, Berlin, 2003.
[34] B. Øksendal and A. Sulem, *Applied Stochastic Control of Jump Diffusions*, Universitext, Springer, Berlin, 2005.
[35] P. Protter and K. Shimbo, *No arbitrage and general semimartingales*, in Markov Processes and Related Topics: A Festschrift for Thomas G. Kurtz, Inst. Math. Stat. Collect. 4, Institute of Mathematical Statistics, Beachwood, OH, 2008, pp. 267–283.
[36] P. E. Protter, *Stochastic Integration and Differential Equations*, 2nd ed., Springer, Berlin, 2005.
[37] L. C. G. Rogers and D. Williams, *Diffusions, Markov Processes and Martingales: Volume* II, *Ito Calculus*, 2nd ed., Cambridge University Press, Cambridge, UK, 2000.
[38] N. Touzi, *Stochastic Control Problems, Viscosity Solutions, and Application to Finance*, http://www.cmap.polytechnique.fr/~touzi/pise02.pdf (2002).
[39] S. Wan, *Risk sensitive optimal portfolio model under jump processes*, in Proceedings of the IEEE Chinese Control Conference, 2006, pp. 607–610.
[40] P. Whittle, *Risk-Sensitive Optimal Control*, Wiley, Chichester, 1990.
[41] W. M. Wonham, *On the separation theorem of stochastic control*, SIAM J. Control, 6 (1968), pp. 312–326.

SIAM J. FINANCIAL MATH.
Vol. 2, pp. 55–78

On the Convergence of Higher Order Hedging Schemes: The Delta-Gamma Case*

Mats Brodén† and Magnus Wiktorsson†

Abstract. Hedging errors induced by discrete rebalancing of the hedge portfolio of a delta-gamma hedging strategy are investigated. The rate of convergence of the expected squared hedging error as the number of adjustments of the hedge portfolio goes to infinity is analyzed. It is found that the delta-gamma strategy produces higher convergence rates than the usual delta strategy.

Key words. discrete time hedging, rate of convergence, delta-gamma hedging

AMS subject classifications. 45M05, 60H05, 91B28

DOI. 10.1137/090779905

1. Introduction.

The aim of this paper is to investigate the convergence of the expected squared hedging error of a discretely rebalanced hedge portfolio containing two hedge instruments with respect to the number of rebalancings. We consider a complete diffusion market containing a number of instruments with sufficiently smooth payoff functions (i.e., European call options), where the hedge portfolio is rebalanced on an equidistant time grid.

The hedging strategy considered is the well-known delta-gamma hedging strategy (see, for example, Björk [2] or Hull [12]), a strategy where the hedge portfolio is made both delta and gamma neutral using the underlying and some other derivative as hedge instruments.

The hedging error at maturity T using n rebalancing points is defined by

$$\mathcal{R}(n) = \Phi(X_T) - V_T,$$

where Φ denotes the payoff function, X_T denotes the stock price, and V_T is the value of the hedge portfolio. In this paper we will measure the error using an L^2 criterion and study the rate at which $\mathbb{E}[\mathcal{R}^2(n)]$ goes to zero as n approaches infinity.

The question of hedging errors due to discrete trading has already been raised in the paper by Black and Scholes [3], where it was claimed that the risk induced from discrete trading is nonsystematic and to a large extent can be diversified away. This claim was supported by Boyle and Emanuel [4], who utilized a Taylor expansion approach. Moreover, they also investigated the actual distribution of the hedging error of a discretely rebalanced hedge portfolio. In both of these papers the delta of the hedge portfolio was set to equal delta of the derivative when rebalancing the portfolio. Robins and Schachter [14] depart from this idea and instead minimize the variance over each rebalancing period with respect to the hedge portfolio weights. They show that this strategy yields a significant reduction in variance of the hedging error compared with the delta neutral strategy.

*Received by the editors December 11, 2009; accepted for publication (in revised form) October 17, 2010; published electronically January 20, 2011.
http://www.siam.org/journals/sifin/2/77990.html

†Centre for Mathematical Sciences, Lund University, Box 118, 221 00 LUND, Sweden (matsb@maths.lth.se, magnusw@maths.lth.se).

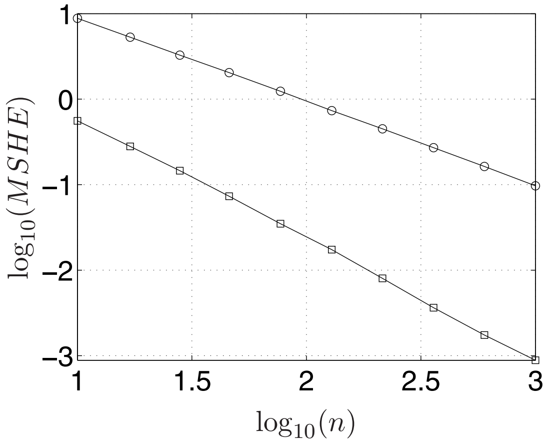


Figure 1. *Log mean squared hedging error (MSHE) as a function of the log number of rebalancings for the delta hedge (circles) and the delta-gamma hedge (squares).*

Later accounts on the topic of discrete time hedging have often focused on the order of convergence to zero of the hedging error as the number of rebalancing dates approaches infinity. In Zhang [15] and Bertsimas, Kogan, and Lo [1], discrete hedging of European options in complete diffusion markets on an equidistant time grid was considered. Whereas Zhang considers an L^2 criterion for the order of convergence, Bertsimas, Kogan, and Lo [1] analyze both the L^2 convergence of the hedging error and convergence in the weak sense. Among other things it was shown that, when delta hedging a European option, $n\mathbb{E}[\mathcal{R}^2(n)]$ converges to a nonzero finite limit as n approaches infinity. The results by Bertsimas, Kogan, and Lo [1] were extended in different directions in the paper by Hayashi and Mykland [11]. In Gobet and Temam [10] it was shown that, when the derivative has a more irregular payoff, the order of convergence may decrease. In particular, for a digital option it was found that $n^{1/2}\mathbb{E}[\mathcal{R}^2(n)]$ converges to a nondegenerate limit as $n \to \infty$. In Geiss [8] it was shown that the order of convergence in the case of digital options may be improved by using a nonequidistant net of adjustment dates. For a particular choice of nonequidistant rebalancing dates it was shown that $n\mathbb{E}[\mathcal{R}^2(n)]$ converges to a nonzero finite limit also in the case of digital options.

In the work presented here we extend previous results to cover the case of contracts that are hedged using two hedge instruments (the underlying and some other derivative) on an equidistant time grid. We find that $n^{3/2}\mathbb{E}[\mathcal{R}^2(n)]$ converges to a nondegenerate limit as n approaches infinity. Thus the order of convergence increases substantially when adding one more hedge instrument to the hedge portfolio.

To illustrate the improvement in order of convergence for the delta-gamma hedge, we performed a simulation of the delta hedging strategy and of the delta-gamma hedging strategy in a Black and Scholes market. Figure 1 shows the log mean squared hedging error (MSHE) as a function of the log number of adjustments of the hedge portfolio. The advantage of the delta-gamma strategy over the delta strategy is clearly visible in the picture.

In the case of a market with normally distributed log returns (i.e., the Black and Scholes market) we investigate by simulation to what extent the asymptotic expression approximates the true mean squared error for finite n, and we analyze the MSHE with respect to the choice of the additional hedge instrument.

This paper is mainly an extension of the paper by Zhang [15], but the techniques used in the proof are more related to those used in Gobet and Temam [10]. In Gobet and Makhlouf [9] the convergence of the tracking error for the delta-gamma hedging strategy is analyzed in a multiasset setting where log returns are assumed to be normally distributed. They derive upper bounds of the expected squared hedging error for a class of payoff functions with variable regularity properties. In particular it is seen that, in the case of digital options, delta-gamma hedging does not increase the order of convergence compared to delta hedging. In our paper we analyze a wider class of prices processes, i.e., time homogeneous local volatility models, and prove the exact order of L^2 convergence in the case of European call options. We also derive an explicit expression of the constant $\lim_{n\to\infty} n^{3/2}\mathbb{E}[\mathcal{R}^2(n)]$.

Discrete trading may arise because of a number of reasons, e.g., due to the presence of transaction costs or if the stock price process can be observed only at discrete points in time. Since options in many cases are subjected to higher transaction costs than the underlying stock, it is clear that transaction costs might have an impact on the practical applicability of delta-gamma hedging strategies. However, in this paper we assume that there are no transaction costs.

The next section introduces some notation, basic facts, and previous results together with our results. In section 3 we perform a simulation study investigating how the results from section 2 relate to a simulated example. We conclude our findings in section 4. The proofs of the lemmas as well as some technical results are presented in Appendix A.

2. Preliminaries and results.

Setting and assumptions. Let the diffusion process X, defined on a filtered probability space $\{\Omega,\ \mathcal{F}_t,\ \{\mathcal{F}_t\}_{t\geq 0},\ \mathbb{P}\}$, describe the price of a risky asset. Assuming that the risk free interest rate equals r, under the risk neutral probability measure $\mathbb{Q}$ the process X has the dynamics

$$\mathrm{d}X_t = rX_t\mathrm{d}t + \sigma(X_t)X_t\mathrm{d}W_t, \tag{1}$$

with $X_0 = x_0$ and where $\{W_t\}_{t\geq 0}$ is a one-dimensional Wiener process.

The above market model may be recognized as a local volatility model with a time invariant diffusion coefficient. In this setting it is often convenient to work with the transformed process $Y_t = \log(X_t)$. In connection to this define $\tilde{\sigma}(y) = \sigma(e^y)$, which is simply the diffusion coefficient of the process Y. Let $p_X(t, x, x')$ and $p_Y(t, y, y')$ denote the transition densities of the processes X and Y, respectively (in the case they exist). One typically wants the processes to have sufficiently smooth transition densities, which translates to smoothness properties of the pricing functions of derivatives in this market. The smoothness of the transition densities is dependent on properties of the diffusion coefficient $\tilde{\sigma}$. We introduce the following set of assumptions:

H1.

(i) There is a positive constant σ_0 such that $\tilde{\sigma}(y) \geq \sigma_0$ for all $y \in \mathbb{R}$.

(ii) The function $\tilde{\sigma}$ is bounded, uniformly Lipschitz continuous in compact subsets of $\mathbb{R}$, and uniformly Hölder continuous.

H2. The functions $(\partial^k/\partial y^k)\tilde{\sigma}(y)$, $i \in \{1,2,3,4\}$, are bounded.

Under assumption H1 the transition probability function of Y has a transition density, i.e., p_Y (see Theorem 6.5.4 of [7]). The densities p_Y and p_X are related through

$$p_X(t,x,x') = \frac{p_Y(t,\log(x),\log(x'))}{x'},$$

and thus under assumption H1 X also has a density.

Now consider a T_1 claim denoted F_1 with payoff function $\Phi_1(x) = (x-K_1)^+$. We would like to replicate this contract using the underlying and some other derivative F_2 with payoff function $\Phi_2(x) = (x-K_2)^+$ and which expires at $T_2 > T_1$. Thus the contracts should satisfy the following assumption:

H3. $\Phi_1(x) = (x-K_1)^+$, $\Phi_2(x) = (x-K_2)^+$, and $T_2 > T_1$.

Pricing and delta hedging. The price of the contracts at time t will be denoted $F_1(t,X_t)$ and $F_2(t,X_t)$, respectively. We will let $F_{1,x}(t,X_t)$ and $F_{2,x}(t,X_t)$ denote their respective derivatives with respect to the underlying asset and equivalently for higher order derivatives and derivatives with respect to the variable t.

It can be shown using replication arguments (see, e.g., [2]) that the price of some derivative F_i at time t with payoff function Φ_i solves the following Cauchy problem:

$$\begin{aligned} F_{i,t}(t,x) + rxF_{i,x}(t,x) + \frac{1}{2}\sigma^2(x)x^2F_{i,xx}(t,x) &= rF_i(t,x), \\ F_i(T,x) &= \Phi_i(x). \end{aligned} \tag{2}$$

Under assumption H1 the Cauchy problem (2) has a solution given by the following stochastic representation:

$$F_i(t,x) = e^{-r(T_i-t)}\mathbb{E}_{t,x}[\Phi_i(X_{T_i})], \tag{3}$$

where $\mathbb{E}_{t,x}$ denotes the expectation under the risk neutral measure $\mathbb{Q}$ conditioned on $\mathcal{F}_t$ with the process in question starting in the point x. At times we will suppress this dependence and simply write $\mathbb{E}$. It is also possible to deduce the well-known risk neutral valuation formula (3) using arbitrage arguments (see [5] for a full account on this topic).

Introduce the processes $\tilde{X}_t = e^{-rt}X_t$ and $\tilde{F}_i(t,X_t) = e^{-rt}F_i(t,X_t)$. Using (2) and Itô's formula, it is seen that both $\tilde{X}_t$ and $\tilde{F}_i(t,X_t)$ are $\mathbb{Q}$-martingale adapted to $\mathcal{F}_t^W$. By the martingale representation theorem for Wiener processes there exists a uniquely determined $\mathcal{F}_t^W$-adapted process h_t such that

$$e^{-rT_i}\Phi_i(X_{T_i}) = \tilde{F}_i(0,X_0) + \int_0^{T_i} h_s \mathrm{d}W_s.$$

Using Itô and the fact that $d\tilde{X}_t = e^{-rt}\sigma(X_t)X_t\mathrm{d}W_t$ yields

$$h_t = \sigma(X_t)X_t\tilde{F}_{i,x}(t),$$

and thus

$$e^{-rT_i}\Phi_i(X_{T_i}) = F_i(0,X_0) + \int_0^{T_i} F_{i,x}(s)\mathrm{d}\tilde{X}_s.$$

Since the market considered is complete, the derivatives can be perfectly replicated using the underlying asset as a hedge instrument. This is what is usually called delta hedging.

One major concern regarding hedging is the possibility of exploding hedge quotes. This problem is an issue when, for example, one is delta hedging an up and out barrier call option. When the option is close to maturity and at the same time approaches the barrier, the delta of the option may grow infinitely high. This is not a problem in our case, i.e., delta hedging of a European call option, which is seen in Lemma 2.1. The gamma of the option, on the other hand, may grow like $1/\sqrt{T_i - t}$ as the option approaches its maturity T_i.

Lemma 2.1. *Assume that* H1 *and* H2 *hold; then there exists a bounded constant C such that*

$$|F_{i,x}(t,x)| \leq C \quad \forall\,(t,x) \in [0,T_i] \times \mathbb{R}_+. \tag{4}$$

Assume in addition that p is a bounded real number; then there exists a bounded constant C such that

$$|x^p F_{i,xx}(t,x)| \leq \frac{C}{(T_i - t)^{\frac{1}{2}}} \quad \forall\,(t,x) \in [0,T_i] \times \mathbb{R}_+. \tag{5}$$

The proof of Lemma 2.1 is presented in Appendix A.

Delta-gamma hedging. We now turn our attention to the case where we use both the underlying asset and F_2 to hedge F_1. We will let h_t^B and h_t^X and $h_t^{F_2}$ denote the amount held in the bank account and the number of shares of the underlying and of the hedge derivative, respectively. We will let V_t denote the value of the hedge portfolio at time t:

$$V_t = h_t^X X_t + h_t^{F_2} F_2(t, X_t) + h_t^B.$$

Clearly we want the value of this portfolio to equal the value of F_1. In order to hedge the issued contract, one way is to make the hedge portfolio both delta and gamma neutral. Thus (see [2])

$$\begin{aligned}
F_1(t, X_t) &= h_t^X X_t + h_t^{F_2} F_2(t, X_t) + h_t^B\,,\\
F_{1,x}(t, X_t) &= h_t^X + h_t^{F_2} F_{2,x}(t, X_t)\,,\\
F_{1,xx}(t, X_t) &= h_t^{F_2} F_{2,xx}(t, X_t)\,.
\end{aligned}$$

By solving this we get the portfolio

$$\begin{aligned}
h_t^X &= F_{1,x}(t, X_t) - \frac{F_{2,x}(t, X_t) F_{1,xx}(t, X_t)}{F_{2,xx}(t, X_t)}\,,\\
h_t^{F_2} &= \frac{F_{1,xx}(t, X_t)}{F_{2,xx}(t, X_t)}\,,\\
h_t^B &= F_1(t, X_t) - h_t^X X_t - h_t^{F_2} F_2(t, X_t)\,.
\end{aligned}$$

We will refer to this type of hedging as delta-gamma hedging. As in the delta hedging case, the value of our hedge portfolio at time T_1 can be written as

$$\begin{aligned}
V_{T_1} = F_1(0, X_0) &+ \int_0^{T_1} \left(F_{1,x}(t, X_t) - \frac{F_{1,xx}(t, X_t)}{F_{2,xx}(t, X_t)} F_{2,x}(t, X_t) \right) \mathrm{d}\tilde{X}_t\\
&+ \int_0^{T_1} \frac{F_{1,xx}(t, X_t)}{F_{2,xx}(t, X_t)} \mathrm{d}\tilde{F}_2(t, X_t)\,.
\end{aligned}$$

Using that $d\tilde{F}_2(t,X_t) = F_{2,x}(t,X_t)\mathrm{d}\tilde{X}_t$ yields

$$\begin{aligned}V_{T_1} &= F_1(0,X_0) + \int_0^{T_1} \left(F_{1,x}(t,X_t) - \frac{F_{1,xx}(t,X_t)}{F_{2,xx}(t,X_t)} F_{2,x}(t,X_t)\right) \mathrm{d}\tilde{X}_t \\ &\quad + \int_0^{T_1} \frac{F_{1,xx}(t,X_t)}{F_{2,xx}(t,X_t)} F_{2,x}(t,X_t)\mathrm{d}\tilde{X}_t \\ &= F_1(0,X_0) + \int_0^{T_1} (F_{1,x}(t,X_t))\,\mathrm{d}\tilde{X}_t = \Phi_1(X_{T_1}),\end{aligned}$$

and thus this portfolio also replicates the contract.

It has already been pointed out that the delta of our option is bounded (see Lemma 2.1). However, the hedge portfolio is now also dependent on the fraction $F_{1,xx}/F_{2,xx}$, that is, the position held in F_2. It is seen that in order to guarantee boundedness one needs to make sure that gamma of F_2 is bounded away from zero, or at least that $F_{1,xx}$ goes to zero faster than $F_{2,xx}$ when $F_{2,xx}$ approaches zero. For practical considerations we point out that the only point at which h^{F_2} may explode is when X_t equals K_1 and t approaches T_1 (this is apparent from the proof of Lemma 2.2). Later in the proofs we will need the following bound (which is sufficient for our needs) presented in the following lemma.

Lemma 2.2. *Assume that* H1–H3 *hold; then there exists a bounded constant C such that*

$$\left|\frac{F_{1,xx}(t,x)}{F_{2,xx}(t,x)}\right| \le \frac{C}{(T_1-t)^{\frac{1}{2}}} \quad \forall\,(t,x)\in[0,T_1]\times\mathbb{R}_+\,. \tag{6}$$

The proof of Lemma 2.2 is presented in Appendix A.

Discrete time hedging. In our setting we will adjust the hedge portfolio at a prespecified time grid t_i^n, where n denotes the number of rebalancing points. In the equidistant case we have that $t_i^n = iT_1/n$, where $i = \{0,\dots,n-1\}$. The difference between the hedge portfolio and the derivative at time $t = T_1$ with n the number of adjustments will be denoted by $\mathcal{R}_\Delta(n)$ in the delta hedging case and $\mathcal{R}_\Gamma(n)$ in the delta-gamma hedging case. In the delta hedging case the discounted hedging error, $\mathcal{R}_\Delta(n)$, can be written as

$$\mathcal{R}_\Delta(n) = \int_0^{T_1} \big(F_{1,x}(t,X_t) - F_{1,x}(\varphi_n(t),X_{\varphi_n(t)})\big)\,\mathrm{d}\tilde{X}_t\,,$$

where $\varphi_n(t) = \sup\{t_i^n \mid t_i^n < t\}$. In the delta-gamma hedging case we have that

$$\mathcal{R}_\Gamma(n) = \int_0^{T_1} \left(h_t^X - h_{\varphi_n(t)}^X\right)\mathrm{d}\tilde{X}_t + \int_0^{T_1} \left(h_t^{F_2} - h_{\varphi_n(t)}^{F_2}\right)\mathrm{d}\tilde{F}_2(t,X_t)\,.$$

In [15] the discretely rebalanced delta hedge is investigated. The expected squared hedging error is found to approach zero as n^{-1} when $n\to\infty$.

$$\mathbb{E}[\mathcal{R}_\Delta^2(n)] = \frac{T_1}{2n}\mathbb{E}\left[\int_0^{T_1} e^{-2rt}\sigma^4(X_t)X_t^4\,(F_{1,xx}(t,X_t))^2\,\mathrm{d}t\right] + o\left(\frac{1}{n}\right).$$

In Proposition 2.3 our main result is presented. It states that the expected squared hedging error, induced by discretely rebalancing the delta-gamma hedge portfolio, goes to zero as $n^{-\frac{3}{2}}$ when $n\to\infty$.

Proposition 2.3. *Assume that* H1–H3 *hold; then*

$$\begin{aligned}\mathbb{E}[\mathcal{R}_\Gamma^2(n)] &= \left(\frac{T_1}{n}\right)^{\frac{3}{2}} C_{\frac{3}{2}} \lim_{t\to T_1} g(t) + o\left(\frac{1}{n^{3/2}}\right)\\ &= \left(\frac{T_1}{n}\right)^{\frac{3}{2}} C_{\frac{3}{2}} e^{-2rT_1}\frac{K_1^3\sigma^3(K_1)}{4\sqrt{\pi}} p_X(T_1,x_0,K_1) + o\left(\frac{1}{n^{3/2}}\right),\end{aligned} \tag{7}$$

where

$$g(t) = (T_1-t)^{3/2}\mathbb{E}\left[e^{-2rt}F_{1,xxx}^2(t,X_t)X_t^6\sigma^6(X_t)\right], \tag{8}$$

and

$$C_a = \sum_{k=1}^\infty \int_0^1\int_0^x\int_0^w \frac{1}{(k-v)^a}\mathrm{d}v\,\mathrm{d}w\,\mathrm{d}x.$$

Proof. In Lemma A.3 the error is decomposed into seven terms A_i $(1\le i\le 7)$, where A_1 is given by

$$A_1(n) = \mathbb{E}\left[\int_0^{T_1}\int_{\varphi_n(t)}^t\int_{\varphi_n(s)}^s e^{-2ru}F_{1,xxx}^2(u,X_u)X_u^6\sigma^6(X_u)\mathrm{d}u\mathrm{d}s\mathrm{d}t\right].$$

The rest of the terms, A_i $(2\le i\le 7)$, go to zero as $O(1/n^{13/8})$ or faster (see Lemma A.5). Lemma A.2, which states that

$$\lim_{t\to T_1}(T_1-t)^{\frac{3}{2}}\mathbb{E}[e^{-2rt}F_{1,xxx}^2(t,X_t)X_t^6\sigma^6(X_t)] = e^{-2rT_1}\frac{K_1^3\sigma^3(K_1)}{4\sqrt{\pi}}p_X(T_1,x_0,K_1),$$

together with Lemma A.4 yields that

$$\mathbb{E}[\mathcal{R}_\Gamma^2(n)] = \left(\frac{T_1}{n}\right)^{\frac{3}{2}} C_{\frac{3}{2}} e^{-2rT_1}\frac{K_1^3\sigma^3(K_1)}{4\sqrt{\pi}}p_X(T_1,x_0,K_1) + o\left(\frac{1}{n^{3/2}}\right),$$

which is the desired result, so the proof is complete. ■

3. Simulation study. In this section we conduct a simulation study to examine the behavior of the expected squared hedging error as we let n approach infinity as well for some fixed choices of n. First, we investigate the transition of the expected squared hedging error to its limiting value as we let n go from some low value to some high value for some different choices of K_2 and T_2. Second, we investigate the expected squared error as a function of K_2 for some different values of n and T_2.

The setup for our simulation study is that of the Black and Scholes market with $r=0$. More specifically, we let the risky asset X be defined by the SDE

$$dX_t = \sigma X_t\mathrm{d}W_t,$$

where $X_0 = x_0$ and σ is a constant. The process X is log-normally distributed with transition density given by

$$p_X(t,x,x') = \frac{1}{x'\sqrt{2\pi\sigma^2 t}}e^{-\frac{(\log(x')-\log(x)+\sigma^2t/2)^2}{2\sigma^2 t}}.$$

According to Proposition 2.3, for large values of n

$$\mathbb{E}[\mathcal{R}_\Gamma^2(n)] \approx \left(\frac{T_1}{n}\right)^{\frac{3}{2}} C_{\frac{3}{2}} \frac{K_1^3\sigma^3(K_1)}{4\sqrt{\pi}} p_X(T_1, x_0, K_1). \tag{9}$$

Our results in the simulation study will be compared with the above approximate value of the expected squared hedging error.

A number of trajectories, N_{MC}, of the process X, where $\sigma = 0.4$ and $x_0 = 100$, together with option prices and hedge ratios were simulated. For the option to be hedged the strike price was set to $K_1 = 100$, and the time to maturity was set to $T_1 = 0.5$.

In the first part we investigate the expected squared hedging error as we let n approach infinity. This is done for three values of $K_2 \in \{60, 100, 140\}$ and two values of $T_2 \in \{0.6, 1.0\}$. The measured MSHEs from the simulations are depicted in Figure 2 together with the value of expression (9). In this simulation N_{MC} was set to $N_{MC} = 100000$.

It is seen that the graphs for the different values of K_2 are more spread out around the value of (9) in the case when $T_2 = 0.6$ than when $T_2 = 1.0$, and that the mean squared error attains the limiting value faster for high values of T_2 than for low values of T_2 when n approaches infinity. As expected, the hedge derivative most similar to F_1, that is, when $T_2 = 0.6$ and $K_2 = 100$, is the one giving the lowest mean squared error. As can be seen, all Monte Carlo estimates are relatively close to the value of (9), especially for high values of T_2.

In Figure 3 the MSHE as a function of K_2 for different values of $T_2 \in \{0.6, 0.8, 1.0\}$ and $n \in \{10, 100\}$ is depicted. In this case the numbers of simulations was set to $N_{MC} = 1000000$.

As can be seen, for every choice of T_2 there is a clear smile pattern with a distinct optimal value of K_2, and this smile is more pronounced for low values of T_2 and n. It is also seen that even though hedge contracts with short maturities might produce very feasible hedges there is also the possibility of ending up with very high values of the expected squared hedging error if one is not able to find a contact with the right strike price. Hence, with hedge contracts with longer time to maturity the expected squared error is more robust with respect to the choice of K_2.

4. Conclusions.

We have shown that the expected squared hedging error using two hedge instruments converges to zero as $n^{-\frac{3}{2}}$ when the number of adjustments of the hedge portfolio n goes to infinity. An expression of the leading $n^{-\frac{3}{2}}$-order term has also been derived.

By simulation it is shown that the derived expression of the leading term approximates the true MSHE quite well (see Figure 2), especially for high values of T_2. It can also be seen that, as expected, using hedge instruments that are similar to the instrument to be hedged, that is, values of K_2 close to K_1 and values of T_2 close to T_1, gives a lower mean squared error. However, in the case when T_2 is chosen to be close to T_1, the value of the MSHE is very sensitive with respect to the value of K_2 (see Figure 3).

Further research could be directed to the investigation of the higher order terms in the expansion of the MSHE in order to find an optimal choice of hedge instrument in a collection of possible hedge instruments.

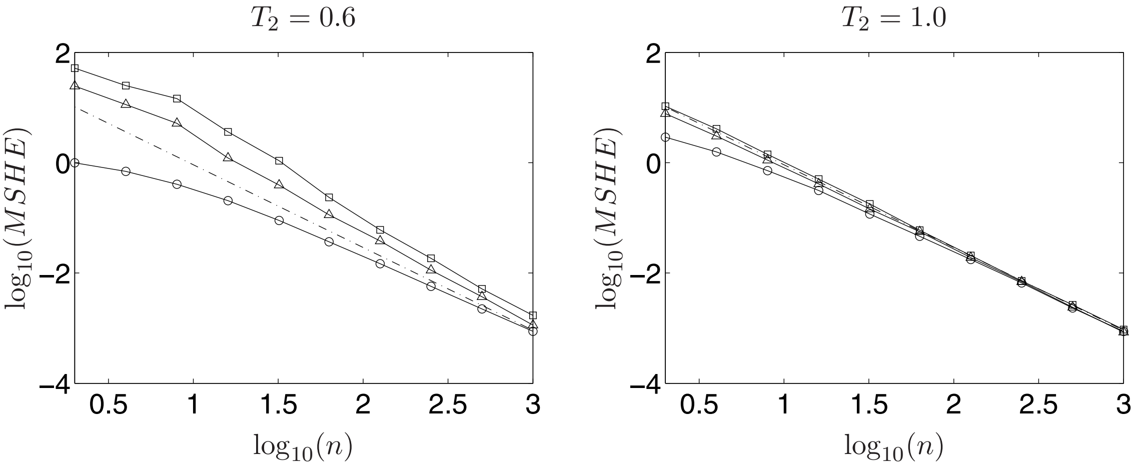


Figure 2. *Log MSHE as a function of the log number of rebalancings, where $K_2 = 60$ (squares), $K_2 = 100$ (circles), $K_2 = 140$ (triangles), and expression* (9) *(dash-dotted line).*

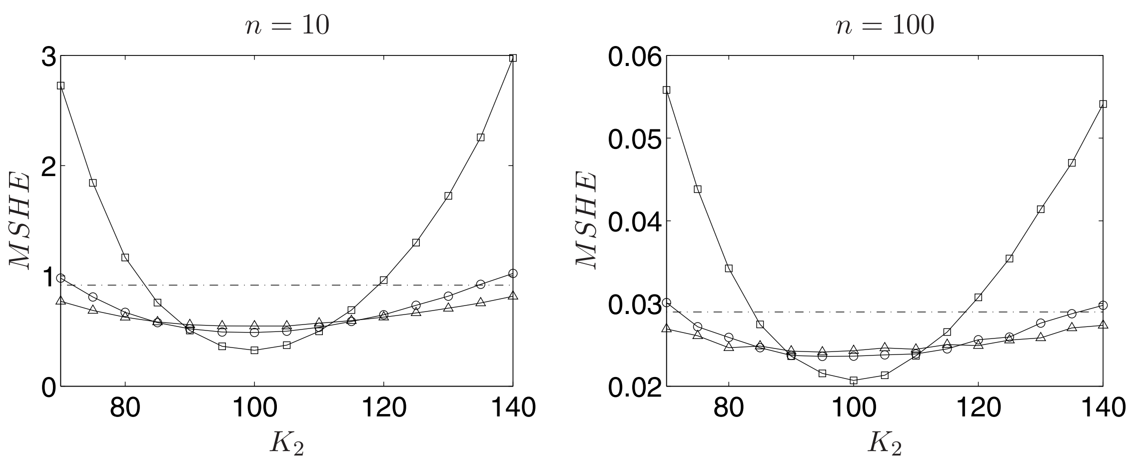


Figure 3. *MSHE as a function of K_2, where $T_2 = 0.6$ (squares), $T_2 = 0.8$ (circles), $T_2 = 1.0$ (triangles), and expression* (9) *(dash-dotted line).*

Appendix A. Proofs and technical lemmas. In Lemma A.3 the mean squared error is decomposed into seven terms A_i $(1 \leq i \leq 7)$, and in Lemma A.5 it is shown that all but term $A_1(n)$ are of higher order than $n^{-\frac{3}{2}}$. The proof of Lemma A.5 relies on the estimates from Lemmas 2.1 and 2.2, as well as on an estimate that is stated in Lemma A.2 together with a limit result. Furthermore, we need an analysis result, Lemma A.4, that provides us with the order of convergence of some integrals.

In the first part of the appendix we will prove Lemmas 2.1 and 2.2, as well as state and prove Lemma A.2. In order to prove Lemmas 2.1 and 2.2, we will make use of two results from [6] that are stated below (see (11) and (12)).

Consider a process Z with dynamics given by

$$\mathrm{d}Z_t = \mu(Z_t)\mathrm{d}t + \mathrm{d}W_t\,, \quad Z_0 = z_0. \tag{10}$$

Define the functions N and G by

$$N(t) = \int_0^t (\mu'(Z_u) + \mu^2(Z_u))\mathrm{d}u \qquad \text{and} \qquad G(y) = \int_{y_0}^y \mu(u)\mathrm{d}u,$$

where $y_0 \in \mathbb{R}$, and define the constants $\overline{m}$ and $\underline{m}$ by

$$\overline{m} = \sup(\mu'(z) + \mu^2(z)) \qquad \text{and} \qquad \underline{m} = \inf(\mu'(z) + \mu^2(z)).$$

Assuming that μ is absolutely continuous and such that (10) has unique strong solution, then, according to Proposition 1 of [6], for any $\mathcal{A} \in \mathcal{F}_t$

$$\mathbb{E}_{0,z_0}[I(A)] = \hat{\mathbb{E}}_{0,z_0}\left[e^{G(Z_t)-G(z_0)}e^{-(1/2)N(t)}I(A)\right], \tag{11}$$

where $\hat{\mathbb{E}}$ denotes expectation with respect to the law of a standard Brownian motion. Furthermore, under the same assumptions as above, according to Corollary 5 of [6],

$$e^{-t\overline{m}/2} \le \frac{p_Z(t,z,z')}{e^{G(z')-G(z)}p_W(t,z,z')} \le e^{-t\underline{m}/2}, \tag{12}$$

where p_Z is the transition density of the process Z and p_W is the transition density of a standard Brownian motion.

Another result that will be used in the calculations is the so-called put-call duality. According to Corollary 3.2 of [13], which holds under H1–H2, the call price, $F_i(t,x) = \mathbb{E}_{t,x}e^{-r(T_i-t)}[(X_{T_i} - K)^+]$, may be represented by

$$F_i(t,x) = \mathbb{E}_{t,K}[(x - \hat{X}_{T_i})^+], \tag{13}$$

where the process $\hat{X}$ is defined by the dynamics

$$\mathrm{d}\hat{X}_t = -r\hat{X}_t\mathrm{d}t + \sigma(\hat{X}_t)\hat{X}_t\mathrm{d}\hat{W}_t,$$

where $\hat{W}$ is a standard Wiener process.

In the proofs below we will let $a(x) = \sigma^2(x)x^2$, $k_i = \log(K_i)$, and $\tau_i = T_i - t$. Furthermore, we define the constants $\overline{\sigma}$ and $\underline{\sigma}$ by

$$\overline{\sigma} = \sup \sigma(z) \qquad \text{and} \qquad \underline{\sigma} = \inf \sigma(z)\,.$$

Also, throughout we will let C denote a bounded positive constant whose value may change from line to line.

Proof of Lemma 2.1.

Equation (4). Differentiating (13) once with respect to x, we get that

$$F_{i,x}(t,x) = \int_0^x p_{\hat{X}}(\tau_i, K_i, y)\mathrm{d}y = \mathbb{E}_{t,K_i}[I(\hat{X}_{T_i} < x)],$$

where $p_{\hat{X}}$ denotes the transition density of the process $\hat{X}$. It holds that

$$0 \le \mathbb{E}_{t,K_i}[I(\hat{X}_{T_i} < x)] \le 1,$$

which proves the claim.

Equation (5). Differentiating (13) twice with respect to x, we get that

$$F_{i,x}(t,xx) = p_{\hat{X}}(\tau_i, K_i, x).$$

Consider the transformation $Z_t = H(\log(\hat{X}_t))$, where $H(y) = \int_{y_0}^{y} 1/\tilde{\sigma}(u)\mathrm{d}u$ for some $y_0 \in \mathbb{R}$; then

$$\mathrm{d}Z_t = \mu(Z_t)\mathrm{d}t + \mathrm{d}W_t,$$

where

$$\mu(z) = -\frac{1}{2}\left(\tilde{\sigma}(H^{-1}(z)) + \tilde{\sigma}'(H^{-1}(z))\right).$$

Note that since $|(\partial^k/\partial y)\tilde{\sigma}(y)| < \infty$, $i \in \{0,1,2\}$ (due to H1–H2), the function μ is absolutely continuous, and thus the condition of Corollary 5 of [6] is satisfied and (12) holds.

Let $p_{\hat{X}}$ and p_Z denote the transition densities of $\hat{X}$ and Z, respectively; then the following relation between the transition densities holds:

$$p_{\hat{X}}(t,x,x') = \frac{p_Z(t, H(\log(x)), H(\log(x')))}{x'\sigma(x')}.$$

Using the inequalities (12) for the density, we get

$$0 \le F_{i,xx}(t,x) \le e^{-\frac{1}{2}\underline{m}\tau_i} e^{G(H(\log(x)))-G(H(k_i))} \frac{p_W(\tau_i, H(k_i), H(\log(x)))}{\sigma(x)x}. \tag{14}$$

Since σ is bounded, it holds that

$$-(H(\log(x)) - H(k_i))^2 \le -\frac{(\log(x)-k_i)^2}{\overline{\sigma}^2}, \tag{15}$$

which will be used to bound $p_W(\tau_i, H(k_i), H(\log(x)))$. For the function G it holds that

$$G(H(\log(x))) - G(H(k_i)) = -\int_{k_i}^{\log(x)} \frac{r}{\tilde{\sigma}^2(u)}\mathrm{d}u - \frac{1}{2}(\log(x)-k_i) - \frac{1}{2}\log\left(\frac{\sigma(x)}{\sigma(K_i)}\right).$$

Since $|\log(\sigma(x)/\sigma(K))|$ is bounded and

$$\left|\int_{k_i}^{\log(x)} \frac{r}{\tilde{\sigma}^2(u)}\mathrm{d}u\right| \le \frac{r}{\underline{\sigma}^2}|\log(x)-k_i|,$$

there is a bounded constant C such that

$$e^{G(H(\log(x)))-G(H(k_i))} \le Ce^{C|\log(x)-k_i|}. \tag{16}$$

To bound expressions of the form $x^p = \exp\{n\log(x)\}$, where p is a bounded real number, we will use

$$e^{p\log(x)} \le e^{pk_i + p|\log(x)-k_i|} \le Ce^{C|\log(x)-k_i|} \tag{17}$$

for some bounded constant C.

Using (14)–(17), we have that there is a bounded constant C such that

$$|x^p F_{i,xx}(t,x)| \leq C e^{C|\log(x)-k_i|} \frac{e^{-\frac{(\log(x)-k_i)^2}{2\overline{\sigma}^2\tau_i}}}{\sqrt{\tau_i}}.$$

The above inequality together with the inequality

$$-x^2 + cx \leq -x^2/2 + c^2/2, \tag{18}$$

which holds for all $x, c \in \mathbb{R}$, yields that there is a bounded constant C such that

$$|x^p F_{i,xx}(t,x)| \leq C \frac{e^{-\frac{(\log(x)-k_i)^2}{4\overline{\sigma}^2\tau_i}}}{\sqrt{\tau_i}},$$

which implies that there is a bounded constant C such that $|x^p F_{i,xx}(t,x)| \leq C/\sqrt{T_i - t}$, which concludes the proof. ■

Proof of Lemma 2.2. In the same way as for (14), we get

$$F_{i,xx}(t,x) \geq e^{-\frac{1}{2}\overline{m}\tau_i} e^{G(H(\log(x)))-G(H(k_i))} \frac{p_W(\tau_i, H(k_i), H(\log(x)))}{\sigma(x)x}. \tag{19}$$

Inequalities (14) and (19) yield that there is a constant C such that

$$\frac{F_{1,xx}(t,x)}{F_{2,xx}(t,x)} \leq \frac{C}{\sqrt{T_1 - t}} e^{-\frac{(H(\log(x))-H(k_1))^2}{2(T_1-t)} + \frac{(H(\log(x))-H(k_2))^2}{2(T_2-t)}}.$$

Consider the function f defined by

$$f(z) = -\frac{(z - H(k_1))^2}{2(T_1 - t)} + \frac{(z - H(k_2))^2}{2(T_2 - t)}.$$

The equation $f'(z_0) = 0$ has the solution

$$z_0 = \frac{H(k_1)\tau_2 + H(k_2)\tau_1}{T_2 - T_1},$$

and

$$f(z_0) = \frac{(H(k_1) - H(k_2))^2}{2(T_2 - T_1)},$$

which is bounded since $T_2 > T_1$ and is a maximum since $f''(z_0) = -1/\tau_1 + 1/\tau_2 < 0$. Hence, we have proved that $\exp\{f(H(\log(x)))\}$ is bounded, and thus there is a bounded constant C such that

$$\frac{F_{1,xx}(t,x)}{F_{2,xx}(t,x)} \leq \frac{C}{\sqrt{T_1 - t}},$$

whence the lemma follows. ■

Lemma A.1. *Define the function h_τ by*

$$h_\tau(z,z') = \frac{2}{\tau^{\frac{3}{2}}\sqrt{\pi}}(z-z')^2 e^{-\frac{(z-z')^2}{\tau}},$$

and let g be a continuous and integrable function; then

$$\lim_{\tau\to 0}\int_{\mathbb{R}} g(x)h_\tau(x_0,x)dx = g(x_0).$$

Proof. For some $\epsilon > 0$ decompose the integral as

$$\int_{\mathbb{R}} g(x)h_\tau(x_0,x)\mathrm{d}x = \int_{-\infty}^{x_0-\epsilon} g(x)h_\tau(x_0,x)\mathrm{d}x + \int_{x_0-\epsilon}^{x_0+\epsilon} g(x)h_\tau(x_0,x)\mathrm{d}x + \int_{x_0+\epsilon}^{\infty} g(x)h_\tau(x_0,x)\mathrm{d}x.$$

For the first term we have that

$$\left|\int_{-\infty}^{x_0-\epsilon} g(x)h_\tau(x_0,x)\mathrm{d}x\right| \leq h_\tau(x_0,\xi)\int_{-\infty}^{x_0-\epsilon} |g(x)|\mathrm{d}x,$$

where $\xi \in (-\infty, x_0-\epsilon)$. Since g is integrable, we have that $\int_{-\infty}^{x_0-\epsilon}|g(x)|\mathrm{d}x < \infty$. Furthermore, $\lim_{\tau\to 0} h_\tau(x_0,\xi) = 0$; thus

$$\lim_{\tau\to 0}\int_{-\infty}^{x_0-\epsilon} g(x)h_\tau(x_0,x)\mathrm{d}x = 0,$$

and equivalently for the last term. For the second term we have that

$$\int_{x_0-\epsilon}^{x_0+\epsilon} g(x)h_\tau(x_0,x)\mathrm{d}x = g(\xi)\int_{x_0-\epsilon}^{x_0+\epsilon} h_\tau(x_0,x)\mathrm{d}x,$$

where $\xi \in (x_0-\epsilon, x_0+\epsilon)$. Now

$$\lim_{\tau\to 0} g(\xi)\int_{x_0-\epsilon}^{x_0+\epsilon} h_\tau(x_0,x)\mathrm{d}x = g(\xi)\lim_{\tau\to 0}\int_{-\epsilon/\sqrt{\tau}}^{\epsilon/\sqrt{\tau}} \frac{2}{\sqrt{\pi}}u^2 e^{-u^2}\mathrm{d}u = g(\xi).$$

Since g is continuous and ϵ can be chosen arbitrarily small, this finishes the proof. ∎

Lemma A.2. *Assume that* H1 *and* H2 *hold; then there exists a constant C such that*

$$\mathbb{E}[e^{-2rt}F_{i,xxx}^2(t)X_t^6\sigma^6(X_t)] \leq \frac{C}{(T_i-t)^{\frac{3}{2}}}, \tag{20}$$

and

$$\lim_{t\to T_i}(T_i-t)^{\frac{3}{2}}\mathbb{E}[e^{-2rt}F_{i,xxx}^2(t)X_t^6\sigma^6(X_t)] = e^{-2rT_i}\frac{K_i^3\sigma^3(K_i)}{4\sqrt{\pi}}p_X(T_i,x_0,K_i). \tag{21}$$

Proof. Differentiating F_i twice with respect to x, we get that

$$F_{i,xx}(t,x) = \frac{\partial}{\partial x}\mathbb{E}_{t,K_i}[I(\hat{X}_{T_i} < x)]\,.$$

Using the transformation $Z_t = H(\log(\hat{X}_t))$, as in the proof of Lemma 2.2, and the change of measure from (11), we have that

$$\begin{aligned}F_{i,xx}(t,x) &= \frac{\partial}{\partial x}\hat{\mathbb{E}}_{t,H(k_i)}\Big[I(Z_{T_i} < H(\log(x)))e^{G(Z_{T_i})-G(H(k_i))}e^{-\frac{1}{2}(N(T_i)-N(t))}\Big]\\ &= \frac{\partial}{\partial x}\hat{\mathbb{E}}_{t,H(k_i)}\Big[I(Z_{T_i} < H(\log(x)))e^{G(Z_{T_i})-G(H(k_i))}\hat{\mathbb{E}}_{t,H(k_i)}\Big[e^{-\frac{1}{2}(N(T_i)-N(t))}|Z_{T_i}\Big]\Big].\end{aligned}$$

Denote

$$\phi(t,z,z') = \hat{\mathbb{E}}_{t,z}[e^{-\frac{1}{2}(N(T_i)-N(t))}|Z_{T_i} = z']\,;$$

then

$$\begin{aligned}F_{i,xx}(t,x) &= \frac{\partial}{\partial x}\mathbb{E}_{t,H(k_i)}\Big[I(Z_{T_i} < H(\log(x)))e^{G(Z_{T_i})-G(H(k_i))}\phi(t,H(k_i),Z_{T_i})\Big]\\ &= e^{G(H(\log(x)))-G(H(k_i))}\phi(t,H(k_i),H(\log(x)))\frac{p_W(\tau_i,H(k_i),H(\log(x)))}{x\sigma(x)}.\end{aligned}$$

Differentiating $F_{i,xx}(t,x)$ with respect to x yields

$$F_{i,xxx}(t,x) = f_1(t,x) + f_2(t,x) + f_3(t,x) + f_4(t,x)\,,$$

where

$$\begin{aligned}f_1(t,x) &= e^{G(H(\log(x)))-G(H(k_i))}\frac{\mu(H(\log(x)))}{\sigma(x)x}\phi(t,H(k_i),H(\log(x)))\\ &\quad\times\frac{p_W(\tau_i,H(k_i),H(\log(x)))}{\sigma(x)x}\,,\\ f_2(t,x) &= e^{G(H(\log(x)))-G(H(k_i))}\frac{\partial}{\partial z}\phi(t,H(k_i),z)|_{z=H(\log(x))}\frac{1}{x\sigma(x)}\\ &\quad\times\frac{p_W(\tau_i,H(k_i),H(\log(x)))}{x\sigma(x)}\,,\\ f_3(t,x) &= e^{G(H(\log(x)))-G(H(k_i))}\phi(t,H(k_i),H(\log(x)))\\ &\quad\times p_W(\tau,H(k_i),H(\log(x)))\frac{(-\sigma(x)-x\sigma'(x))}{x^2\sigma^2(x)}\,,\end{aligned}$$

and

$$\begin{aligned}f_4(t,x) &= e^{G(H(\log(x)))-G(H(k_i))}\phi(t,H(k_i),H(\log(x)))\\ &\quad\times\frac{\partial}{\partial z}p_W(\tau,H(k_i),z)|_{z=H(\log(x))}\frac{1}{x^2\sigma^2(x)}.\end{aligned}$$

Since the functions μ and μ' under assumptions H1 and H2 are bounded, there is a bounded constant C such that

$$\phi(t,z,z') = \hat{\mathbb{E}}_{t,z}[e^{-\frac{1}{2}\int_t^{T_i}\mu'(Z_s)+\mu^2(Z_s)ds}|Z_{T_i} = z'] \leq C. \tag{22}$$

Note that the process Z conditioned on $Z_{T_i} = z'$ is a Brownian bridge from z to z'. Let

$$h(s,z,z') = z + \frac{s-t}{T_i - t}(z'-z) + \left(W_{s-t} - \frac{s-t}{T_i-t}W_{T_i-t}\right) \quad \text{for} \quad s \in [t, T_i]\,;$$

then, under the measure $\hat{\mathbb{E}}$ and the condition $Z_{T_i} = z'$, the following equality in distribution holds: $h(s,z,z') \overset{d}{=} Z_s$. Define the function g by $g(z) = -(\mu'(z) + \mu^2(z))/2$. Differentiating ϕ we get, after applying the dominated convergence theorem twice, that

$$\begin{aligned}\frac{\partial}{\partial z}\phi(t,z,z') &= \frac{\partial}{\partial z}\mathbb{E}\left[e^{\int_t^{T_i} g(h(s,z,z'))ds}\right] \\ &= \mathbb{E}\left[e^{\int_t^{T_i} g(h(s,z,z'))ds}\int_t^{T_i} \frac{\partial}{\partial x}g(x)|_{x=h(s,z,z')}\frac{\partial}{\partial z}h(s,z,z')ds\right],\end{aligned}$$

where $\frac{\partial}{\partial z}h(s,z,z') = (T_i - s)/(T_i - t)$. Since g and g' are bounded under H1–H2, there is a bounded constant C such that

$$\left|\frac{\partial}{\partial z}\phi(t,z,z')\right| \le 2C\int_t^{T_i}\frac{T_i - s}{T_i - t}\mathrm{d}s = C(T_i - t)\,. \tag{23}$$

By (15), (16), (17), (22), (23), and the boundedness of σ, $\tilde{\sigma}'$, and μ, we have that there is a bounded constant C such that

$$|(f_1(t,x) + f_2(t,x) + f_3(t,x))x^3\sigma^3(x)| \le Ce^{C|\log(x)-k_i|}\frac{e^{-\frac{(\log(x)-k_i)^2}{2\overline{\sigma}\tau_i}}}{\sqrt{2\pi\tau_i}}\,. \tag{24}$$

Inequality (18) together with (24) yields that there is a bounded constant C such that

$$|(f_1(t,x) + f_2(t,x) + f_3(t,x))x^3\sigma^3(x)| \le C\frac{e^{-\frac{(\log(x)-k_i)^2}{4\overline{\sigma}\tau_i}}}{\sqrt{2\pi\tau_i}}\,. \tag{25}$$

The bounds (15), (16), (17), (18), and (22) yield that there is a bounded constant C such that

$$\begin{aligned}|f_4(t,x)x^3\sigma^3(x)| &\le Ce^{C|\log(x)-k_i|}\frac{|H(k_i) - H(\log(x))|}{\tau_i}\frac{e^{-\frac{(H(k_i)-H(\log(x)))^2}{2\tau_i}}}{\sqrt{2\pi\tau_i}} \\ &\le C\frac{|k_i - \log(x)|}{\underline{\sigma}\tau_i}\frac{e^{-\frac{(k_i-\log(x))^2}{4\overline{\sigma}\tau_i}}}{\sqrt{2\pi\tau_i}}\,.\end{aligned} \tag{26}$$

Using the transformation $H(\log(X_t)) = \int_{y_0}^{\log(X_t)} 1/\tilde{\sigma}(u)\mathrm{d}u$, the bound (12), and inequality (18), we get that there is a bounded constant C such that

$$\begin{aligned}p_X(t,x,x') &\le Ce^{-\frac{1}{2}(\log(x')-\log(x))}\frac{p_W(t,H(\log(x)),H(\log(x')))}{x'} \\ &\le C\frac{e^{-\frac{(\log(x')-\log(x))^2}{4\overline{\sigma}t}}}{\sqrt{2\pi t}}\,.\end{aligned} \tag{27}$$

By (25), (27), and a change of variable ($y = \log(x)$), formal integration yields that there is a bounded constant C such that

$$\mathbb{E}[e^{-2rt}(f_1(t,X_t) + f_2(t,X_t) + f_3(t,X_t))^2 X_t^6 \sigma^6(X_t)] \le \frac{C}{\sqrt{T_i - t}}\,.$$

Using (26), we get in the same way as for the bound above that there is a bounded constant C such that

$$\mathbb{E}[e^{-2rt} f_4^2(t,X_t) X_t^6 \sigma^6(X_t)] \le \frac{C}{(T_i - t)^{3/2}}\,.$$

Consequently, by the Cauchy–Bunyakovskii inequality, there is a bounded constant C such that

$$\mathbb{E}[e^{-2rt}(f_1(t,X_t) + f_2(t,X_t) + f_3(t,X_t)) f_3 X_t^6 \sigma^6(X_t)] \le \frac{C}{(T_i - t)}\,.$$

Hence, $E[f_4^2(t,X_t)X_t^6\sigma^6(X_t)]$ is the leading term, and there is a bounded constant C such that

$$\mathbb{E}[e^{-2rt} F_{i,xxx}^2(t) X_t^6 \sigma^6(X_t)] \le \frac{C}{(T_i - t)^{\frac{3}{2}}}\,,$$

which proves the first part of the lemma.

It holds that

$$\begin{aligned}
&\mathbb{E}[e^{-2rt} f_4^2(t,X_t) X_t^6 \sigma^6(X_t)] \\
&= \int_0^\infty e^{-2rt} \frac{e^{2(G(H(\log(x))) - G(H(k_i)))}\phi(t, H(k_i), H(\log(x)))}{K_i^2 \sigma^2(K_i) x^2 \sigma^2(x)} \\
&\quad \times \frac{(H(\log(x)) - H(k_i))^2}{\tau^2} \frac{e^{-\frac{(H(\log(x)) - H(k_i))^2}{\tau_i}}}{2\pi\tau_i} x^6 \sigma^6(x) p_X(t, x_0, x)\mathrm{d}x \\
&= \int_{\mathbb{R}} e^{-2rt} \frac{e^{2(G(z) - G(H(k_i)))}\phi(t, H(k_i), z)}{K_i^2 \sigma^2(K_i)} \frac{(z - H(k_i))^2}{\tau_i^2} \frac{e^{-\frac{(z - H(k_i))^2}{\tau_i}}}{2\pi\tau_i} \\
&\quad \times e^{5H^{-1}(z)} \sigma^5(e^{H^{-1}(z)}) p_X(t, x_0, e^{H^{-1}(z)})\mathrm{d}z\,.
\end{aligned}$$

Using Lemma A.1 and that $E[f_3^2(t,X_t)X_t^6\sigma^6(X_t)]$ is the leading term, we get

$$\begin{aligned}
&\lim_{\tau_i \to 0} \tau_i^{\frac{3}{2}} \mathbb{E}[e^{-2rt} F_{i,xxx}^2(X_t, t) X_t^6 \sigma^6(X_t)] \\
&= \lim_{\tau_i \to 0} \int_{\mathbb{R}} e^{-2rt} e^{2(G(z) - G(H(k_i)))} \phi(t, H(k_i), z) \frac{e^{5H^{-1}(z)} \sigma^5(e^{H^{-1}(z)})}{4\sqrt{\pi} K_i^2 \sigma^2(K_i)} \\
&\quad \times p_X(t, x_0, e^{H^{-1}(z)}) \frac{2}{\tau_i^{\frac{3}{2}} \sqrt{\pi}} (z - H(k_i))^2 e^{-\frac{(z - H(k_i))^2}{\tau_i}} \mathrm{d}y \\
&= e^{-2rT_i} \frac{K_i^3 \sigma^3(K_i)}{4\sqrt{\pi}} p_X(t, x_0, K_i)\,,
\end{aligned}$$

which finishes the proof. ■

To ease the notation in the following lemma and its proof we will occasionally suppress the second variable in the functions $F_i(t, X_t)$; i.e., $F_i(t) = F_i(t, X_t)$. We adopt the same notation for derivatives of the function F_i and write $F_{i,x}(t) = F_{i,x}(t, X_t)$ and equivalently for higher order derivatives and derivatives with respect to the variable t. Also, recall that $a(x) = x^2\sigma^2(x)$, and let the function D be defined by

$$D(t) = F_{1,x}(t) - F_{1,x}(\varphi_n(t)) - (F_{2,x}(t) - F_{2,x}(\varphi_n(t)))\frac{F_{1,xx}(\varphi_n(t))}{F_{2,xx}(\varphi_n(t))} \,.$$

Furthermore, let $D_t(t)$, $D_x(t)$, $D_{xx}(t)$, and $D_{xxx}(t)$ be defined by

$$\begin{aligned} D_t(t) &= F_{1,tx}(t) - F_{2,tx}(t)\frac{F_{1,xx}(\varphi_n(t))}{F_{2,xx}(\varphi_n(t))} \,, \\ D_x(t) &= F_{1,xx}(t) - F_{2,xx}(t)\frac{F_{1,xx}(\varphi_n(t))}{F_{2,xx}(\varphi_n(t))} \,, \\ D_{xx}(t) &= F_{1,xxx}(t) - F_{2,xxx}(t)\frac{F_{1,xx}(\varphi_n(t))}{F_{2,xx}(\varphi_n(t))} \,, \\ D_{xxx}(t) &= F_{1,xxxx}(t) - F_{2,xxxx}(t)\frac{F_{1,xx}(\varphi_n(t))}{F_{2,xx}(\varphi_n(t))} \,. \end{aligned}$$

Lemma A.3. *Let*

$$A_i(n) = \mathbb{E}\left[\int_0^{T_1}\int_{\varphi_n(t)}^t\int_{\varphi_n(s)}^s e^{-2ru} g_i(u, X_u)\,du\,ds\,dt\right]$$

for $i \in \{1, 2, 3, 4, 5\}$*, and*

$$A_i(n) = \mathbb{E}\left[\int_0^{T_1}\int_{\varphi_n(t)}^t e^{-2rs} g_i(s, X_s)\,ds\,dt\right]$$

for $i \in \{6, 7\}$*, where*

$$\begin{aligned} g_1(u, X_u) &= F_{1,xxx}^2(u)a^3(X_u)\,, \\ g_2(u, X_u) &= -2\,F_{1,xxx}(u)F_{2,xxx}(u)\frac{F_{1,xx}(\varphi_n(u))}{F_{2,xx}(\varphi_n(u))}a^3(X_u)\,, \\ g_3(u, X_u) &= F_{2,xxx}^2(u)\frac{F_{1,xx}^2(\varphi_n(u))}{F_{2,xx}^2(\varphi_n(u))}a^3(X_u)\,, \\ g_4(u, X_u) &= D_x^2(u)h_2(X_u)\,, \\ g_5(u, X_u) &= 2D_x(u)D_{xx}(u)a^2(X_u)a'(X_u)\,, \\ g_6(s, X_s) &= D^2(s)h_1(X_s)\,, \\ g_7(s, X_s) &= D(s)D_x(s)a(X_s)a'(X_s)\,, \end{aligned}$$

where

$$\begin{aligned} h_1(x) &= -2ra(x) + rxa'(x) + a(x)a''(x)/2, \\ h_2(x) &= a(x)(-4ra(x) + 2rxa'(x) + (a'(x))^2). \end{aligned}$$

Then

$$\mathbb{E}[\mathcal{R}_\Gamma^2(n)] = \sum_{i=1}^{7} A_i(n).$$

Proof. The expected squared error can be written as

$$\begin{aligned}\mathbb{E}[\mathcal{R}_\Gamma^2(n)] = \mathbb{E}\Bigg[\int_0^{T_1}\Bigg(\Bigg(F_{1,x}(t) - \frac{F_{2,xx}(t)F_{1,xx}(t)}{F_{2,xx}(t)}\Bigg) \\ - \Bigg(F_{1,x}(\varphi_n(t)) - \frac{F_{2,xx}(\varphi_n(t))F_{1,xx}(\varphi_n(t))}{F_{2,xx}(\varphi_n(t))}\Bigg) \\ + \Bigg(\frac{F_{1,xx}(t)}{F_{2,xx}(t)} - \frac{F_{1,xx}(\varphi_n(t))}{F_{2,xx}(\varphi_n(t))}\Bigg)F_{2,x}(t)\Bigg)^2 X_t^2\sigma^2(X_t)e^{-2rt}\mathrm{d}t\Bigg] \\ = \mathbb{E}\Bigg[\int_0^{T_1} D^2(t)X_t^2\sigma^2(X_t)e^{-2rt}\mathrm{d}t\Bigg].\end{aligned}$$

Let

$$f_1(t) = e^{-2rt}D^2(t)X_t^2\sigma^2(X_t) = e^{-2rt}D^2(t)a(X_t)\,,$$

and apply Itô's formula between $\varphi_n(t)$ and t. Then

$$\mathrm{d}f_1(t) = \Big\{f_{1,t}(t) + rX_tf_{1,x}(t) + \frac{1}{2}a(X_t)f_{1,xx}(t)\Big\}\mathrm{d}t + \sigma(X_t)X_tf_{1,x}(t)\mathrm{d}W_t\,,$$

where the diffusion term vanishes when taking expectations. The derivatives of f_1 are given by

$$\begin{aligned}f_{1,t}(t) &= e^{-2rt}(2D(t)D_t(t)a(X_t) - 2rD^2(t)a(X_t))\,,\\ f_{1,x}(t) &= e^{-2rt}(2D(t)D_x(t)a(X_t) + D^2(t)a'(X_t))\,,\end{aligned}$$

and

$$\begin{aligned}f_{1,xx}(t) = e^{-2rt}(&2D_x^2(t)a(X_t) + 2D(t)D_{xx}(t)a(X_t)\\ &+ 4D(t)D_x(t)a'(X_t) + D^2(t)a''(X_t))\,.\end{aligned}$$

Multiplying these terms with the corresponding coefficients and summing them together yields the drift term

$$\begin{aligned}&f_{1,t}(t) + rX_tf_{1,x}(t) + \frac{1}{2}a(X_t)f_{1,xx}(t)\\ &\quad = e^{-2rt}\Big(2D(t)D_t(t)a(X_t) - 2rD^2(t)a(X_t) + 2rX_tD(t)D_x(t)a(X_t)\\ &\qquad + rX_tD^2(t)a'(X_t) + D_x^2(t)a^2(X_t) + D(t)D_{xx}(t)a^2(X_t)\\ &\qquad + 2D(t)D_x(t)a(X_t)a'(X_t) + \frac{1}{2}D^2(t)a(X_t)a''(X_t)\Big)\,.\end{aligned}$$

Differentiating the Black and Scholes PDE

$$F_{i,t}(t,x) + rxF_{i,x}(t,x) + \frac{1}{2}a(x)F_{i,xx}(t,x) = rF_i(t,x)$$

with respect to x yields

$$F_{i,tx}(t,x) + \left(rx + \frac{1}{2}a'(x)\right) F_{i,xx}(t,x) + \frac{1}{2}a(x)F_{i,xxx}(t,x) = 0\,. \tag{28}$$

Using (28), we have that

$$D_t(t) + \left(rX_t + \frac{1}{2}a'(X_t)\right) D_x(t) + \frac{1}{2}a(X_t)D_{xx}(t) = 0\,. \tag{29}$$

This leaves us with the drift term

$$\begin{aligned} f_{1,t}(t) + rX_t f_{1,x}(t) + \frac{1}{2}a(X_t)f_{1,xx}(t) = e^{-2rt}\Big(& D_x^2(t)a^2(X_t) - 2rD^2(t)a(X_t) \\ & + rX_t D^2(t)a'(X_t) + \frac{1}{2}D^2(t)a(X_t)a''(X_t) + D(t)D_x(t)a(X_t)a'(X_t)\Big)\,. \end{aligned}$$

Thus we get the following decomposition:

$$\begin{aligned} \mathbb{E}[\mathcal{R}_\Gamma^2(n)] = {} & \mathbb{E}\left[\int_0^{T_1}\int_{\varphi_n(t)}^t e^{-2rt}D_x^2(s)a^2(X_s)\mathrm{d}s\mathrm{d}t\right] \\ & + \mathbb{E}\left[\int_0^{T_1}\int_{\varphi_n(t)}^t e^{-2rt}D^2(s)h_1(X_t)\mathrm{d}s\mathrm{d}t\right] \\ & + \mathbb{E}\left[\int_0^{T_1}\int_{\varphi_n(t)}^t e^{-2rt}D(s)D_x(s)a(X_s)a'(X_s)\mathrm{d}s\mathrm{d}t\right], \end{aligned} \tag{30}$$

where $h_1(x) = -2ra(x) + rxa'(x) + a(x)a''(x)/2$.

In the next step the first term will be decomposed once more. The first term is given by

$$\begin{aligned} & \mathbb{E}\left[\int_0^{T_1}\int_{\varphi_n(t)}^t e^{-2rs}D_x^2(s)a^2(X_s)\mathrm{d}s\mathrm{d}t\right] \\ & \quad = \mathbb{E}\left[\int_0^{T_1}\int_{\varphi_n(t)}^t e^{-2rs}\left(\frac{F_{1,xx}(s)}{F_{2,xx}(s)} - \frac{F_{1,xx}(\varphi_n(s))}{F_{2,xx}(\varphi_n(s))}\right)^2 F_{1,xx}^2(s)a^2(X_s)\,\mathrm{d}s\,\mathrm{d}t\right]. \end{aligned}$$

Let

$$f_2(t) = e^{-2rt}D_x^2(t)a^2(X_t)\,,$$

and apply Itô's lemma between $\varphi_n(s)$ and s. The derivatives of f_2 are given by

$$\begin{aligned} f_{2,t}(t) &= e^{-2rt}(2D_x(t)D_{tx}(t)a^2(X_t) - 2rD_x^2(t)a^2(X_t))\,, \\ f_{2,x}(t) &= e^{-2rt}(2D_x(t)D_{xx}(t)a^2(X_t) + 2D_x^2(t)a(X_t)a'(X_t))\,, \end{aligned}$$

and

$$f_{2,xx}(t) = e^{-2rt}\big(2D_{xx}^2(t)a^2(X_t) + 2D_x(t)D_{xxx}(t)a^2(X_t) + 2D_x^2(t)(a'(X_t))^2 + 8D_x(t)D_{xx}(t)a(X_t)a'(X_t) + 2D_x^2(t)a(X_t)a''(X_t)\big)\,.$$

Hence, the drift term is given by

$$\begin{aligned}
&f_{2,t}(t) + rX_t f_{2,x}(t) + \frac{1}{2}a(X_t)f_{2,xx}(t)\\
&\quad= e^{-2rt}\big(2D_x(t)D_{tx}(t)a(X_t) - 2rD_x^2(t)a^2(X_t)\\
&\qquad + 2rX_tD_x(t)D_{xx}(t)a^2(X_t) + 2rX_tD_x^2(t)a(X_t)a'(X_t)\\
&\qquad + D_x^2(t)a^2(X_t)a''(X_t) + 4D_x(t)D_{xx}(t)a^2(X_t)a'(X_t)\\
&\qquad + D_x(t)D_{xxx}(t)a^3(X_t) + 2D_{xx}^2(t)a^2(X_t) + 2D_x^2(t)(a'(X_t))^2\big)\,.
\end{aligned}$$

Differentiating (28) with respect to x, we have that

$$F_{i,txx}(t) + \left(r + \frac{1}{2}a''(x)\right)F_{i,xx}(t) + \big(rx + a'(x)\big)F_{i,xxx}(t) + \frac{1}{2}a(X_t)F_{i,xxxx}(t) = 0\,,$$

and, consequently,

$$D_{tx}(t) + \left(r + \frac{1}{2}a''(X_t)\right)D_x(t) + \big(rX_t + a'(X_t)\big)D_{xx}(t) + \frac{1}{2}a(X_t)D_{xxx}(t) = 0\,. \tag{31}$$

Equation (31) yields

$$\begin{aligned}
&f_{2,t}(t) + rX_tf_{2,x}(t) + \frac{1}{2}\sigma^2(X_t)X_t^2f_{2,xx}(t)\\
&\qquad= e^{-2rt}\big(D_{xx}^2(t)a^3(X_t) - 4rD_x^2(t)a^2(X_t) + 2rX_tD_x^2(t)a'(X_t)\\
&\qquad\qquad + D_x^2(t)a(X_t)a_x^2(X_t) + 2D_x(t)D_{xx}(t)a^2(X_t)a'(X_t)\big)\,,
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_\Gamma^2(n)] &= \mathbb{E}\left[\int_0^{T_1}\int_{\varphi_n(t)}^t\int_{\varphi_n(s)}^s e^{-2ru}D_{xx}^2(u)a^3(X_u)\mathrm{d}u\mathrm{d}s\mathrm{d}t\right]\\
&+ \mathbb{E}\left[\int_0^{T_1}\int_{\varphi_n(t)}^t\int_{\varphi_n(s)}^s e^{-2ru}D_x^2(u)h_2(X_u)\mathrm{d}u\mathrm{d}s\mathrm{d}t\right]\\
&+ \mathbb{E}\left[\int_0^{T_1}\int_{\varphi_n(t)}^t\int_{\varphi_n(s)}^s e^{-2ru}2D_x(u)D_{xx}(u)a^2(X_u)a'(X_u)\mathrm{d}u\mathrm{d}s\mathrm{d}t\right]\\
&+ \mathbb{E}\left[\int_0^{T_1}\int_{\varphi_n(t)}^t e^{-2rs}D^2(s)h_1(X_s)\mathrm{d}s\mathrm{d}t\right]\\
&+ \mathbb{E}\left[\int_0^{T_1}\int_{\varphi_n(t)}^t e^{-2rs}D(s)D_x(s)a(X_s)a'(X_s)\mathrm{d}s\mathrm{d}t\right],
\end{aligned}$$

where $h_2(x) = a(x)(-4ra(x) + 2rxa'(x) + (a'(x)^2))$. Furthermore, we have that

$$D^2_{xx}(t) = \left(F_{1,xxx}(t) - F_{2,xxx}(t)\frac{F_{1,xx}(\varphi_n(t))}{F_{2,xx}(\varphi_n(t))}\right)^2$$

$$= F^2_{1,xxx}(t) - 2F_{1,xxx}(t)F_{2,xxx}(t)\frac{F_{1,xx}(\varphi_n(t))}{F_{2,xx}(\varphi_n(t))} + F^2_{2,xxx}\frac{F^2_{2,xx}(\varphi_n(t))}{F^2_{1,xx}(\varphi_n(t))},$$

which completes the proof. ■

In order to show that the integrals are of the right order in the proof of Lemma A.5, we will use the following lemma, which is an extension of Lemma 4 of [10].

Lemma A.4. *Let $g : [0,T] \mapsto \mathbb{R}$ be a measurable, bounded, and continuous function, and let $a \in (1, 5/2)$. Then*

$$\int_0^T \int_{\varphi_n(u)}^u \int_{\varphi_n(t)}^t \frac{g(s)}{(T-s)^a} \mathrm{d}s\, \mathrm{d}t\, \mathrm{d}u = C_a\, g(T)\left(\frac{T}{n}\right)^{3-a} + o\left(\frac{1}{n^{3-a}}\right),$$

where

$$C_a = \sum_{j=1}^{\infty} \int_0^1 \int_0^x \int_0^w \frac{1}{(j-v)^a} \mathrm{d}v\, \mathrm{d}w\, \mathrm{d}x\,.$$

Proof. Let $\epsilon = T/n$; then we have that

$$\int_0^T \int_{\varphi_n(u)}^u \int_{\varphi_n(t)}^t \frac{g(s)}{(T-s)^a} \mathrm{d}s\, \mathrm{d}t\, \mathrm{d}u = \sum_{k=0}^{n-1} \int_{\epsilon k}^{\epsilon(k+1)} \int_{\epsilon k}^u \int_{\epsilon k}^t \frac{g(s)}{(T-s)^a} \mathrm{d}s\, \mathrm{d}t\, \mathrm{d}u$$

$$= \left| v = \frac{s-\epsilon k}{\epsilon} \right| = \sum_{k=0}^{n-1} \int_{\epsilon k}^{\epsilon(k+1)} \int_{\epsilon k}^u \int_0^{\frac{t-\epsilon k}{\epsilon}} \frac{g(\epsilon(v+k))}{(T-\epsilon(v+k))^a} \epsilon\, \mathrm{d}v\, \mathrm{d}t\, \mathrm{d}u$$

$$= \left| w = \frac{t-\epsilon k}{\epsilon} \right| = \sum_{k=0}^{n-1} \int_{\epsilon k}^{\epsilon(k+1)} \int_0^{\frac{u-\epsilon k}{\epsilon}} \int_0^w \frac{g(\epsilon(v+k))}{(T-\epsilon(x+k))^a} \epsilon^2\, \mathrm{d}v\, \mathrm{d}w\, \mathrm{d}u$$

$$= \left| x = \frac{u-\epsilon k}{\epsilon} \right| = \sum_{k=0}^{n-1} \int_0^1 \int_0^x \int_0^w \frac{g(\epsilon(v+k))}{(T-\epsilon(v+k))^a} \epsilon^3\, \mathrm{d}v\, \mathrm{d}w\, \mathrm{d}x$$

$$= \sum_{k=0}^{n-1} \int_0^1 \int_0^x \int_0^w \frac{g\left(\frac{T}{n}(v+k)\right)}{\left(\frac{T}{n}(n-(v+k))\right)^a} \left(\frac{T}{n}\right)^3 \mathrm{d}v\, \mathrm{d}w\, \mathrm{d}x$$

$$= |j = n-k| = \left(\frac{T}{n}\right)^{3-a} \sum_{j=1}^{n} \int_0^1 \int_0^x \int_0^w \frac{g\left(T - T\frac{j-v}{n}\right)}{(j-v)^a} \mathrm{d}v\, \mathrm{d}w\, \mathrm{d}x\,.$$

Since g is continuous, for every $\varepsilon > 0$ there exists a δ_0 such that $|g(T) - g(T-\delta)| < \varepsilon$ for all $\delta \leq \delta_0$. For all $n > \lceil T/\delta \rceil$ there exist $j_0(n) > 0$ such that $Tj/n \leq \delta$ for all $j \leq j_0(n)$ and

$Tj/n > \delta$ for all $j > j_0(n)$. Now we have using that g is bounded, say by C,

$$\begin{aligned}
&\sum_{j=1}^{n}\int_0^1\int_0^x\int_0^w \frac{\left|g(T)-g\left(T-T\frac{j-v}{n}\right)\right|}{(j-v)^a}\,\mathrm{d}v\,\mathrm{d}w\,\mathrm{d}x\\
&=\sum_{j=1}^{j_0(n)}\int_0^1\int_0^x\int_0^w \frac{\left|g(T)-g\left(T-T\frac{j-v}{n}\right)\right|}{(j-v)^a}\,\mathrm{d}v\,\mathrm{d}w\,\mathrm{d}x\\
&\quad+\sum_{j=j_0(n)+1}^{n}\int_0^1\int_0^x\int_0^w \frac{\left|g(T)-g\left(T-T\frac{j-v}{n}\right)\right|}{(j-v)^a}\,\mathrm{d}v\,\mathrm{d}w\,\mathrm{d}x\\
&\le \epsilon\sum_{j=1}^{j_0(n)}\int_0^1\int_0^x\int_0^w \frac{1}{(j-v)^a}\,\mathrm{d}v\,\mathrm{d}w\,\mathrm{d}x+\frac{2CT^a}{(n\delta-T)^a}n\,.
\end{aligned}$$

Since ϵ can be made arbitrarily small, we have shown that

$$\begin{aligned}
&\left(\frac{T}{n}\right)^{3-a}\sum_{j=1}^{n}\int_0^1\int_0^x\int_0^w \frac{g\left(T-\frac{j-v}{n}\right)}{(j-v)^a}\,\mathrm{d}v\,\mathrm{d}w\,\mathrm{d}x\\
&\qquad=\left(\frac{T}{n}\right)^{3-a}g(T)\sum_{j=1}^{n}\int_0^1\int_0^x\int_0^w \frac{1}{(j-v)^a}\,\mathrm{d}v\,\mathrm{d}w\,\mathrm{d}x+o\left(\frac{1}{n^{3-a}}\right).
\end{aligned}$$

Finally, we have that

$$\begin{aligned}
&\left(\frac{T}{n}\right)^{3-a}\sum_{j=1}^{n}\int_0^1\int_0^x\int_0^w \frac{g\left(T-\frac{j-v}{n}\right)}{(j-v)^a}\,\mathrm{d}v\,\mathrm{d}w\,\mathrm{d}x\\
&\qquad=\left(\frac{T}{n}\right)^{3-a}g(T)\sum_{j=1}^{\infty}\int_0^1\int_0^x\int_0^w \frac{1}{(j-v)^a}\,\mathrm{d}v\,\mathrm{d}w\,\mathrm{d}x+o\left(\frac{1}{n^{3-a}}\right),
\end{aligned}$$

which is the desired result, and thus the lemma is proved. ■

Lemma A.5. *Assume that* H1–H3 *hold; then*

$$A_i(n)\le O\left(\frac{1}{n^{\frac{13}{8}}}\right)$$

for $i\in\{2,3,4,5,6,7\}$.

Proof. Recall that C denotes a bounded constant whose value may change between the lines. Also, in the calculations below we will use that $\exp(-2rt)\le 1$, without referring to this fact.

Term A_2. Inequalities (6) and (20), the Cauchy–Bunyakovskii inequality, and Lemma A.4 with $a=5/4$ yield

$$|A_2(n)|\le C\int_0^{T_1}\int_{\varphi_n(t)}^{t}\int_{\varphi_n(s)}^{s}\frac{1}{(T_1-u)^{\frac{5}{4}}}\mathrm{d}u\,\mathrm{d}s\,\mathrm{d}t\le O\left(\frac{1}{n^{\frac{7}{4}}}\right).$$

Term A_3. Inequalities (6) and (20) together with Lemma A.4 with $a = 5/4$ yield

$$|A_3(n)| \le C \int_0^{T_1} \int_{\varphi_n(t)}^{t} \int_{\varphi_n(s)}^{s} \frac{1}{(T_1 - u)} \,\mathrm{d}u\,\mathrm{d}s\,\mathrm{d}t$$
$$\le C \int_0^{T_1} \int_{\varphi_n(t)}^{t} \int_{\varphi_n(s)}^{s} \frac{1}{(T_1 - u)^{\frac{5}{4}}} \,\mathrm{d}u\,\mathrm{d}s\,\mathrm{d}t \le O\left(\frac{1}{n^{\frac{7}{4}}}\right),$$

where C may change between the lines.

Term A_4. Taking absolute values yields

$$|A_4(n)| \le \int_0^{T_1} \int_{\varphi_n(t)}^{t} \int_{\varphi_n(s)}^{s} \mathbb{E}[D_x^2(u)h_2(X_u)]\mathrm{d}u\mathrm{d}s\mathrm{d}t\,.$$

There is a constant C such that $|h_2(x)| \le Cx^4$ (by the boundedness assumption of $\tilde{\sigma}$ and $\tilde{\sigma}'$). Using (5) and (6), we have that $X^4(t)D_x^2(t) \le C/(T_1 - t)$. Now, in the same way as for term A_3, we get that $|A_4(n)| \le O(1/n^{\frac{7}{4}})$.

Term A_5. Taking absolute values yields

$$|A_5(n)| \le \mathbb{E}\left[\int_0^{T_1} \int_{\varphi_n(t)}^{t} \int_{\varphi_n(s)}^{s} 2|D_x(u)D_{xx}(u)|a^2(X_u)a'(X_u)\,\mathrm{d}u\,\mathrm{d}s\,\mathrm{d}t\right].$$

Furthermore, we have that

$$|D_x(t)D_{xx}(t)| \le |F_{1,xxx}(t)|\left(|F_{1,xx}(t)| + \left|F_{2,xx}(t)\frac{F_{1,xx}(\varphi_n(t))}{F_{2,xx}(\varphi_n(t))}\right|\right)$$
$$+ \left|F_{2,xxx}(t)\frac{F_{1,xx}(\varphi_n(t))}{F_{2,xx}(\varphi_n(t))}\right|(|F_{1,xx}(t)| + |F_{2,xx}(t)|)\,,$$

where the first term is the leading one. Using that there is a constant C such that $|a^2(x)a'(x)| \le C|x|^5$, the Cauchy–Bunyakovskii inequality, and inequalities (6) and (20), together with Lemma A.4 with $a = 5/4$, we have that

$$|A_5(n)| \le C \int_0^{T_1} \int_{\varphi_n(t)}^{t} \int_{\varphi_n(s)}^{s} \frac{1}{(T_1 - u)^{\frac{5}{4}}} \mathrm{d}u\,\mathrm{d}s\,\mathrm{d}t \le O\left(\frac{1}{n^{\frac{7}{4}}}\right).$$

Term A_6. Recall that

$$A_6(n) = \mathbb{E}\left[\int_0^{T_1} \int_{\varphi_n(t)}^{t} D^2(s)h_1(X(s))\,\mathrm{d}s\,\mathrm{d}t\right].$$

Using Itô's formula and (29) on $D^2(s)h_1(X(s))$, we have that the leading term of A_6 is of the form $\int_0^{T_1} \int_{\varphi_n(t)}^{t} \int_{\varphi_n(s)}^{s} \mathbb{E}[D_x^2(u)X_u^4]\mathrm{d}u\,\mathrm{d}s\,\mathrm{d}t$. Thus we are in the same situation as for Term A_4. Using the same techniques as for A_4, we have that $|A_6(n)| \le O(1/n^{\frac{7}{4}})$. The assumption that

$\tilde{\sigma}^{(4)}$ is bounded (assumption H2) is used for one of the lower order terms in the expansion of A_6.

Term A_7. Consider the decomposition (30). From the calculations of the Terms A_2–A_5 and inequality (20) we have that $\mathbb{E}[D_x^2(t)X_t^4] \leq C/(T-t)^{3/2}$, and from the calculations of Term A_6 we have that $\mathbb{E}[D^2(t)X_t^2] \leq C/(T-t)^{5/4}$. The Cauchy–Bunyakovskii inequality gives us that $\mathbb{E}[D(t)D_x(t)a(X_t)a'(X_t)] \leq C/(T-t)^{11/8}$, which together with Lemma A.4 yields $|A_4(n)| \leq O(1/n^{\frac{13}{8}})$. ■

Acknowledgments. We would like to thank Roger Pettersson for many insightful comments and Emmanuel Gobet for finding and pointing out an error in a previous version of this paper. Any remaining errors are our responsibility.

REFERENCES

[1] D. Bertsimas, L. Kogan, and A. W. Lo, *When is time continuous?*, J. Financial Econ., 55 (2000), pp. 173–204.

[2] T. Björk, *Arbitrage Theory in Continuous Time*, 2nd ed., Oxford University Press, Oxford, UK, 2004.

[3] F. Black and M. Scholes, *The pricing of options and corporate liabilities*, J. Political Economy, 81 (1973), pp. 637–654.

[4] P. P. Boyle and D. Emanuel, *Discretely adjusted option hedges*, J. Financial Econ., 8 (1980), pp. 259–282.

[5] F. Delbaen and W. Schachermayer, *The Mathematics of Arbitrage*, Springer Finance, Springer-Verlag, Berlin, 2006.

[6] A. N. Downes, *Bounds for the transition density of time-homogeneous diffusion processes*, Statist. Probab. Lett., 79 (2009), pp. 835–841.

[7] A. Friedman, *Stochastic Differential Equations and Applications*, Dover Publications, Mineola, NY, 2006.

[8] S. Geiss, *Quantitative approximation of certain stochastic integrals*, Stoch. Stoch. Rep., 73 (2002), pp. 241–270.

[9] E. Gobet and A. Makhlouf, *The tracking error rate of the Delta-Gamma hedging strategy*, to appear.

[10] E. Gobet and E. Temam, *Discrete time hedging errors for options with irregular payoffs*, Finance Stoch., 5 (2001), pp. 357–367.

[11] T. Hayashi and P. A. Mykland, *Evaluating hedging errors: An asymptotic approach*, Math. Finance, 15 (2005), pp. 309–343.

[12] J. C. Hull, *Options, Futures, and Other Derivative Securities*, 7th ed., Prentice-Hall, Englewood Cliffs, NJ, 2008.

[13] B. Jourdain, *Stochastic flow approach to Dupire's formula*, Finance Stoch., 11 (2007), pp. 521–535.

[14] R. P. Robins and B. Schachter, *An analysis of the risk in discretely rebalanced option hedges and delta-based techniques*, Manage. Sci., 40 (1994), pp. 798–808.

[15] R. Zhang, *Couverture approchée des options Européennes*, Ph.D. thesis, Ecole Nationale des Ponts et Chaussées, Paris, 1999.

SIAM J. FINANCIAL MATH.
Vol. 2, pp. 79–111

Convergence of Price and Sensitivities in Carr's Randomization Approximation Globally and Near Barrier*

Sergei Levendorskiĭ†

Abstract. Barrier options under wide classes of Lévy processes with exponentially decaying jump densities, including the variance gamma model, KoBoL and CGMY models, normal inverse Gaussian processes, and β-class, are studied. The leading term of asymptotics of the option price and the leading term of asymptotics in Carr's randomization approximation to the price are calculated as the price of the underlying approaches the barrier. We prove that the order of asymptotics is the same in both cases and that the asymptotic coefficient in the asymptotic formula for Carr's randomization approximation converges to the asymptotic coefficient for the price as the number of time steps $N \to +\infty$. Also, we justify Richardson extrapolation of arbitrary order. Similar results are derived for sensitivities and the leading terms of their asymptotics in Carr's randomization approximation. The convergence of prices and sensitivities is proved in appropriate weighted Hölder spaces.

Key words. Greeks, barrier options, first-touch digitals, Lévy processes, Carr's randomization, KoBoL processes, CGMY model, normal inverse Gaussian processes, variance gamma processes, β-class, Wiener–Hopf factorization, asymptotics

AMS subject classifications. 91G20, 91G60, 65T50, 60K99

DOI. 10.1137/100788331

1. Introduction.

1.1. Carr's randomization.

In [22], Carr and Faguet applied a time discretization technique known as the "analytical method of lines" to pricing a finite-lived American option. This method reduces the pricing problem for the finite-lived option with a sequence of pricing problems for perpetual American options. The latter problems are easier to solve than the former, so that, as a result, an efficient numerical pricing procedure for finite-lived American options was developed. In [20], Carr found a conceptual interpretation of this method in the language of probability theory and christened it "Canadization" (or "maturity randomization"); we prefer the term "Carr's randomization."

Briefly speaking, Carr's idea was to replace a deterministic maturity date T of an American option on a given stock $S = \{S_t\}_{t\ge 0}$ with a sum, $\tilde{T} = \tilde{T}_1 + \cdots + \tilde{T}_N$, of exponentially distributed random maturity dates that are independent of each other and of the process S, such that each $\tilde{T}_j$ has mean T/N. Therefore, $\tilde{T}$ has mean T and variance T^2/N. As $N \to \infty$, the variance of $\tilde{T}$ approaches 0, while the mean remains fixed. Carr's insight was that the price of the option with random maturity $\tilde{T}$ should approach the price of the option with deterministic maturity T as $N \to \infty$.

*Received by the editors March 11, 2010; accepted for publication (in revised form) October 24, 2010; published electronically January 20, 2011.
http://www.siam.org/journals/sifin/2/78833.html

†Department of Mathematics, The University of Leicester, University Road, Leicester LE1 7RH, United Kingdom (sl278@le.ac.uk).

The convergence as $N \to \infty$ of the approximation scheme described above was proved fully in [6] for American options in the classical Black–Scholes model, along with results for more general classes of models under certain additional assumptions. (We recall that Carr's randomization procedure can be formulated for American options in arbitrary Lévy-driven models (see, e.g., [15]) and Markov models [6].)

1.2. Barrier options. Carr's randomization method can also be applied to barrier options; it was already used to design fast and accurate pricing procedures for single and double barrier options under various classes of Lévy processes in a number of works [31, 10, 11, 12, 8]. A formal convergence proof for these procedures was derived in [7]. However, the proof in [7] cannot justify the claim made in [31, 10, 11] that Carr's randomization method gives more accurate approximation to prices near the barrier than many other methods. The aim of this paper is to supplement [7] and prove that, for wide classes of Lévy processes, the leading term of asymptotics of the price of a barrier option in Carr's randomization approximation tends to the leading term of asymptotics of the price of the barrier option as $N \to \infty$. We prove similar results for sensitivities as well.

These results are important because qualitatively different behavior of prices of barrier options and first-touch digitals near the barrier, in different Lévy models, can be used to identify the type of underlying process. In applications to real financial markets, the prices are calculated using an approximate numerical method. Therefore, it is important to know that the method used reproduces the prices near the barrier accurately. The aim of this paper is to provide a theoretical background for Carr's randomization approximation as a reliable tool for pricing barrier options near the barrier. At the same time, there are fundamental reasons for many other numerical methods to produce a different kind of price behavior near the boundary.

The problem of pricing and hedging barrier options has attracted much attention in recent years, from both the theoretical finance side and the practitioners' side. For instance, a rather comprehensive review of the 1965–1995 literature on the pricing of barrier options given in [19] lists about 30 articles, while hundreds of new works devoted to the same topic have appeared since 1995. In the framework of the Black–Scholes market model [5], an explicit formula for the price of a barrier call option was obtained by Merton [40]. Many subsequent works on barrier and first-touch digital options also remained in the Black–Scholes framework (the interested reader may wish to consult, for example, the bibliography lists in the papers [19, 18]).

However, it is a known fact that the Black–Scholes model yields rather inaccurate prices of barrier options near expiry, especially when the spot price S is close to the barrier. The reason is that the Black–Scholes value function $V_{BS} = V_{barr,BS}(T, S)$, where T is time to maturity, for (say) a down-and-out barrier put option is continuously differentiable w.r.t. S within the closed interval $[H, +\infty)$, whereas, for many classes of Lévy processes, the delta is unbounded as the underlying tends to the barrier. The dashed line in Figure 1 is an example of such a behavior if the underlying process is Brownian motion: we can clearly see how the delta, $\partial V_{BS}/\partial S$, of the option has a finite limit as S approaches H from the right.

As a result, the relative errors of prices of barrier options computed using the Black–Scholes model versus more realistic stock pricing models can sometimes reach several dozen percent near the barrier. We refer the reader to [31], where this phenomenon was demonstrated using an example of the normal inverse Gaussian (NIG) model.

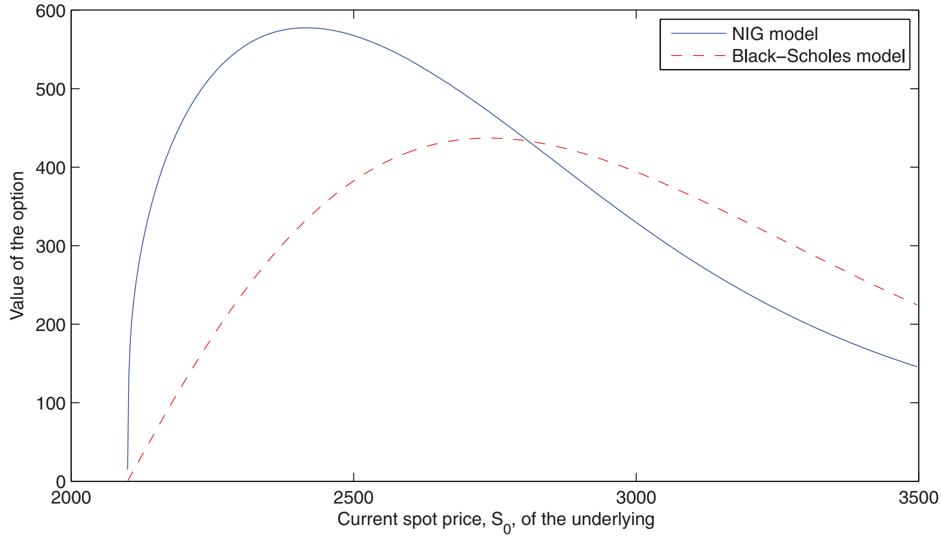


Figure 1. *The value function of a down-and-out barrier put option in the NIG and Black–Scholes models. The strike price is $K = 3500$, the barrier is $H = 2100$, the time to maturity is $T = 1$ year, the riskless rate is 3%, and the underlying stock pays no dividends. (The example is taken from [28].) Solid line: the graph of the value function calculated assuming that under a risk-neutral measure chosen by the market the log-price process $\{X_t = \ln S_t\}$ of the underlying is an NIG process with parameters $\alpha = 8.858$, $\beta = -5.808$, $\delta = 0.174$. Dashed line: same as above, except that X is assumed to be a Brownian motion with volatility $\sigma \approx 0.2136$, chosen so that the second (instantaneous) moment, σ^2, of $X = \{X_t\}$ is the same as the second (instantaneous) moment, $\delta\alpha^2(\alpha^2 - \beta^2)^{-3/2}$, of the NIG process in the first example.*

Similar problems arise if the underlying pure jump Lévy process (or a Lévy process with insignificant jump component) is approximated by a jump diffusion with a tangible diffusion component. The most convenient jump-diffusion models are the double-exponential jump-diffusion model [30, 35, 36] and its natural generalization constructed in [33] to price American options and later labelled the *hyperexponential jump-diffusion* (HEJD) *model* by Jeannin and Pistorius [28], who derived explicit formulas for the Laplace transforms w.r.t. the time variable of the value functions, deltas, gammas, and thetas of the barrier options. They further showed that, approximating other Lévy-driven models (such as variance gamma (VG) [39, 38, 37] and NIG) by suitable hyperexponential models, one can obtain accurate approximations to the prices and sensitivities of barrier and first-touch digital options in the regions not too close to the barrier. For a more recent and more detailed approach to an approximation technique of a similar nature, and for its application to the pricing of double barrier options, we refer the reader to [26, 21], respectively.

Nevertheless, hyperexponential models (with nonzero Gaussian component) also have the disadvantage that the value functions of barrier options in these models are continuously differentiable up to the barrier. In other words, qualitatively these functions exhibit behavior similar to that of the dashed line in Figure 1, whereas value functions obtained from more realistic models of stock prices, such as NIG and KoBoL, exhibit behavior similar to that of the solid line in Figure 1.

Other methods of pricing barrier options that use models with a tangible diffusion component to approximate models with zero diffusion component (such as the method of Cont and Voltchkova [25]) suffer from the same problem (see [34, 31] for an analysis of the errors that result from applying these methods). This issue is important because empirical studies of financial markets (see, e.g., [23]) show that, fairly often, the dynamics of the underlying has zero diffusion component.

1.3. Main results. In the paper, we derive the leading terms of asymptotics of the price of a barrier option and the corresponding Carr's randomization approximation, and we prove that the asymptotic coefficient in the latter converges to the asymptotic coefficient of the former as $N \to \infty$ for wide classes of options including puts, calls, and digital options, as well as wide classes of Lévy processes, which contain the processes of the extended Koponen family [13][1] (with the CGMY model [23] as a special case), NIG processes [2] and the generalization of this family (normal tempered stable processes [3]), the VG model introduced to finance by Madan and coauthors [39, 38, 37], and Kuznetsov's β-class [32]. However, in the case of VG model, we were able to calculate the leading term of asymptotics only in the case of drift pointing from the boundary. The same holds for β-class, in cases when the Lévy density has the same leading term of asymptotics at the origin as in the VG model. Hence, in these cases, we are able to prove the convergence of the asymptotic coefficient only if the drift points from the barrier.

We also prove that, in all cases, the difference between the price and Carr's randomization approximation can be represented in the form $\sum_{j=1}^{m} c_j(T,x)N^{-j} + O(N^{-m-1})$, for any positive integer m, with the constant in the O-term depending on m, T, x. Similar results can be obtained for sensitivities and their leading terms of asymptotics. This implies that Richardson extrapolation of any order is justified provided Carr's randomization approximations are calculated sufficiently accurately. We produce numerical results, which demonstrate that Richardson extrapolation of different orders agree very well, and, therefore, Richardson extrapolation of prices obtained with small numbers of time steps can be used to achieve very good accuracy.

1.4. Organization of the paper. In section 2, we present the well-known formulas that express the prices of barrier options in probabilistic terms, calculate the Laplace transforms of these expressions w.r.t. time to maturity in terms of the expected present value (EPV) operators under supremum and infimum processes,[2] and write the formulas for the prices using the inverse Laplace transform. In section 3, we derive similar general formulas for Carr's randomization approximation.

Main results for prices (formulas for the leading terms of asymptotics and convergence theorems) are formulated in section 6.1 for a wide class of regular Lévy processes of exponential type (RLPEs) of order $\nu \in (1, 2]$ and for strongly regular Lévy processes of exponential type (sRLPEs) of order $\nu \in [0+, 2]$. The definitions of RLPEs and sRLPEs are given in section 4, and necessary general formulas for the Wiener–Hopf factors are recalled in section 5. In section 6, we derive explicit formulas for the Wiener–Hopf factors and crucial estimates for the Wiener–Hopf factors, and we use these estimates to prove the convergence of Carr's

[1] These are also known as KoBoL processes [15]; we adopt this terminology.

[2] Here we use the operator form of the Wiener–Hopf factorization method developed in [14, 15, 16, 34, 17] with an important further development in [11].

randomization approximation to the price in a Hölder norm and justify Richardson extrapolation of arbitrary order. In section 7, we prove the convergence of the asymptotic coefficient of Carr's randomization approximation to the asymptotic coefficient of the price as the time step vanishes and improve the characterization of the convergence of Carr's randomization to the price. Similar statements for sensitivities are formulated and proved in section 8. Numerical examples are given in section 9; section 10 concludes. Technical proofs and necessary elements of the theory of generalized functions are relegated to the appendices.

2. General formulas for barrier options.

2.1. Lévy-driven models. For an exposition of the general theory of Lévy processes and their applications to pricing derivative securities, we refer the reader to [4, 42] and [15, 24, 43], respectively. We recall that every Lévy process $X = \{X_t\}_{t\geq 0}$ has the *characteristic exponent*, which is a continuous function $\psi : \mathbb{R} \to \mathbb{C}$ satisfying $\psi(0) = 0$ and

$$\mathbb{E}\big[e^{i\xi X_t}\big] = e^{-t\psi(\xi)} \qquad \forall\, \xi \in \mathbb{R},\, t \geq 0;$$

and, conversely, the law of a Lévy process is uniquely determined by its characteristic exponent [42, Thm. 7.10]. Some examples of Lévy processes that are commonly used in empirical studies of financial markets are listed in section 4.

We consider a model frictionless market consisting of a riskless bond and a stock, which is modelled as an exponential Lévy process $S_t = e^{X_t}$, under a chosen equivalent martingale measure (EMM) $\mathbb{Q}$. The riskless rate r is constant. We remark that, in general, $\mathbb{Q}$ is not unique. We assume that an EMM $\mathbb{Q}$ has been fixed once and for all, and all expectation operators appearing in this text will be w.r.t. this measure. The characteristic exponent ψ of X is also under this $\mathbb{Q}$.

If the stock does not pay dividends, then S_t must be a martingale under $\mathbb{Q}$. In terms of the characteristic exponent of the log-price process $\{X_t\}$, the EMM condition can be written as follows: $r + \psi(-i) = 0$, where we are implicitly assuming that $\psi(\xi)$ admits the analytic continuation into the closed strip $-1 \leq \operatorname{Im}\xi \leq 0$ (if this is not the case, then $\mathbb{E}[S_t] = \infty$ for all $t > 0$, i.e., the process $\{S_t\}$ cannot be priced; we exclude this situation from our consideration). If the stock pays dividends at constant rate δ, then the EMM condition becomes $r+\psi(-i) = \delta$.

2.2. Down-and-out options. We consider a down-and-out barrier contingent claim with barrier H, expiry date T, and terminal payoff function $G(x)$, which is a nonnegative measurable function on $\mathbb{R}$. If, at any time $t \leq T$ prior to expiry, the price, $S_t = e^{X_t}$, of the underlying reaches or falls below $H = e^h$, the claim expires worthless. Otherwise, at expiry, the claim yields payoff equal to $G(X_T)$. The payoff function is nonnegative and measurable. Contingent claims of this type provide a common generalization of down-and-out barrier call and put options. In particular, $G(x) = (e^x - K)_+$ (respectively, $G(x) = (K - e^x)_+$) for a down-and-out barrier call (respectively, put) option, and $G(x) = \mathbb{1}_{[\ln K,+\infty)}(x)$ (respectively, $G(x) = \mathbb{1}_{(-\infty,\ln K]}(x)$) for the digital call (respectively, put), with strike price $K > H$. To calculate the leading term of asymptotics, we impose an additional regularity condition, which is satisfied for down-and-out barrier (European) call and put options and digital down-and-out options. Namely, the Fourier transform of the payoff function admits the bound

$$|\hat{G}(\xi)| \leq C(1 + |\xi|)^{-1} \tag{2.1}$$

along any line $\mathrm{Im}\,\xi = \omega$, with the exception of a finite number of omegas (indeed, for puts and calls, $\hat{G}(\xi)$ has two poles at $\xi = 0, -i$, and, for digitals, it has one pole at $\xi = 0$). The results apply (and the proof simplifies) in the case of the digital no-touch option with $G = \mathbb{1}_{\mathbb{R}}$. Formally, we can satisfy (2.1) in this case as well using $G(x) = \mathbb{1}_{[\ln K, +\infty)}(x)$ with $K < H$, but this will unnecessarily complicate the proof. Finally, pricing the call option with $K < H$ reduces to pricing the digital no-touch option by the change of measure.

The time-0 value of the barrier option with payoff $G(X_T)$ is given by

$$V(T, x) = \mathbb{E}^x \left[e^{-rT} G(X_T) \mathbb{1}_{\{\tau_h > T\}} \right], \tag{2.2}$$

where x is the current log-spot price of the underlying, $h = \ln H$, and τ_h is the first entrance time of the process $\{X_t\}_{t \geq 0}$ into the interval $(-\infty, h]$.

As is well known, $\hat{V}(q, x)$, the Laplace transform of $V(T, x)$, can be calculated more explicitly. Applying Fubini's theorem, we obtain that $\hat{V}(q, x)$ is the value function of the perpetual stream $G(X_t)$, which is terminated the first moment X_t reaches the barrier or falls below it, the discounting factor being $q + r$:

$$\hat{V}(q, x) = \int_0^{+\infty} e^{-qt} \mathbb{E}^x \left[e^{-rt} G(X_t) \mathbb{1}_{\{\tau_h > t\}} \right] dt = \mathbb{E}^x \left[\int_0^{\tau_h} e^{-(q+r)t} G(X_t) dt \right].$$

Before giving a formula for the last expectation, let us introduce some notation. First, we define the *supremum process* $\overline{X}$ and the *infimum process* $\underline{X}$ of X by

$$\overline{X}_t = \sup_{0 \leq s \leq t} X_s, \qquad \underline{X}_t = \inf_{0 \leq s \leq t} X_s. \tag{2.3}$$

Given any $q > 0$, we let $T_q \sim \mathrm{Exp}\, q$ denote an exponentially distributed random variable with mean q^{-1}, and we define operators $\mathcal{E}_q^+$ and $\mathcal{E}_q^-$ acting on a nonnegative measurable (or an arbitrary bounded measurable) function f on $\mathbb{R}$ as follows:

$$(\mathcal{E}_q^+ f)(x) = \mathbb{E}^x \left[f(\overline{X}_{T_q}) \right], \qquad (\mathcal{E}_q^- f)(x) = \mathbb{E}^x \left[f(\underline{X}_{T_q}) \right], \tag{2.4}$$

where the notation $\mathbb{E}^x$ means that X, $\overline{X}$, and $\underline{X}$ start at x. Writing explicitly the right-hand sides (RHSs) in (2.4) as expectations of integrals

$$(\mathcal{E}_q^+ f)(x) = \mathbb{E}^x \left[\int_0^{+\infty} q e^{-qt} f(\overline{X}_t) dt \right], \tag{2.5}$$

$$(\mathcal{E}_q^- f)(x) = \mathbb{E}^x \left[\int_0^{+\infty} q e^{-qt} f(\underline{X}_t) dt \right], \tag{2.6}$$

we obtain the interpretation of $\mathcal{E}_q^\pm$ as the EPV operators, which calculate the EPV of the stream, under the supremum and infimum processes. Note that the EPV operator $\mathcal{E}_q f(x) = \mathbb{E}^x \left[f(X_{T_q}) \right]$ under the initial process is the normalized resolvent.

Lemma 2.1. *We have*

$$\hat{V}(q, x) = (q + r)^{-1} \left(\mathcal{E}_{q+r}^- \mathbb{1}_{(h, +\infty)} \mathcal{E}_{q+r}^+ G \right)(x). \tag{2.7}$$

Equation (2.7) was derived in [15] for RLPEs (see section 4 for the definition) and in [11] for any Lévy process.

It is evident from (2.5) and (2.6) that the EPV operators $\mathcal{E}^{\pm}_{q+r}$ admit the analytic continuation into the open half-plane $\mathrm{Re}\, q > -r$, which is uniformly bounded (as a family of operators in $L^\infty(\mathbb{R})$) in each closed half-plane $\mathrm{Re}\, q + r \geq \sigma > 0$. Hence, $\hat{V}(q,x)$ admits the bound

$$|\hat{V}(q,x)| \leq C_\sigma |q+r|^{-1}, \quad \mathrm{Re}\, q + r \geq \sigma, \tag{2.8}$$

where C_σ depends on $\sigma > 0$ but not on q. Now, one can use the inverse Laplace transform and calculate the value function of the barrier option. To simplify the result, we make the change of variables $q + r \mapsto q$:

$$V(T,x) = \frac{e^{-rT}}{2\pi i} \int_{\mathrm{Re}\, q = \sigma} e^{qT} q^{-1} \left(\mathcal{E}^-_q \mathbb{1}_{(h,+\infty)} \mathcal{E}^+_q G \right)(x) dq. \tag{2.9}$$

To prove that the RHS in (2.9) is continuous as a function of $T > 0$ and to derive the leading term of asymptotics, we integrate by parts k times using $e^{qT} dq = T^{-1} de^{qT}$:

$$V(T,x) = \frac{e^{-rT}}{2\pi i (-T)^k} \int_{\mathrm{Re}\, q = \sigma} e^{qT} \partial^k_q \left(q^{-1} \mathcal{E}^-_q \mathbb{1}_{(h,+\infty)} \mathcal{E}^+_q G \right)(x) dq. \tag{2.10}$$

In section 6, we prove that if k is sufficiently large, then the integral in (2.10) converges absolutely and uniformly w.r.t. T, and we calculate the leading term of asymptotics as $x \downarrow h$.

3. Carr's randomization for barrier options.

3.1. General formulas. We consider the simplest case of Carr's randomization, which is equivalent to the analytical method of lines. We divide the time interval $[0,T]$ into N intervals of equal length. For the calculations below, it is convenient to use the opposite direction on the time line, with T as the initial point. We set $t_s = T - sT/N$, $s = 0, 1, \dots,$ and denote by $V^s(x)$ Carr's randomization approximation to $V(t_s, x)$, the option price at time t_s. For $s = 0$, $V^0(x) = G(x)\mathbb{1}_{(h,+\infty)}(x)$ is known, and, for $s = 0, 1, \dots,$ V^{s+1} are found by induction as

$$V^{s+1}(x) = \mathbb{E}^x \left[\int_0^{\tau_h} e^{-(r+N/T)t} (N/T) V^s(X_t) dt \right], \tag{3.1}$$

where τ_h is the first entrance time by X into $(-\infty, h]$. If X satisfies the (ACP) property (absolute continuity of potential kernels (see [42]); all processes that we consider satisfy this property), then it follows from [15, Thm. 2.12] that function V^{s+1} solves the stationary generalized Black–Scholes equation

$$\frac{V^s(x) - V^{s+1}(x)}{T/N} + (L - r)V^{s+1}(x) = 0, \quad x > h, \tag{3.2}$$

which is understood in the sense of the theory of generalized functions (the analytical method of lines starts with time discretization of the nonstationary Black–Scholes equation, which also gives (3.2)). Clearly, the boundary condition

$$V^{s+1}(x) = 0, \quad x \leq h, \tag{3.3}$$

holds. Changing the measure if necessary, we can reduce to the case of a bounded G; then, for all s and x,

$$|V^s(x)| \le (1 + rT/N)^{-s}||G||_\infty. \tag{3.4}$$

For $z \in \mathbb{C}$, define

$$\hat{V}(z,x) = \sum_{s=0}^{\infty} z^s V^{s+1}(x),$$

notice that, on the strength of (3.4), $\hat{V}(z,x)$ is analytic in the open disc $|z| < e^{rT/N}$, multiply (3.2) and (3.3) by z^s, and sum up; the result is the boundary problem

$$\frac{G(x) + (z-1)\hat{V}(z,x)}{(N/T)} + (L - r)\hat{V}(z,x) = 0, \quad x > h, \tag{3.5}$$

$$\hat{V}(z,x) = 0, \quad x \le h. \tag{3.6}$$

We rewrite (3.5) as

$$(r + (N/T)(1-z) - L)\hat{V}(z,x) = (N/T)G(x), \quad x > h, \tag{3.7}$$

and set $q = r + (N/T)(1-z)$. In [14, Thm. 4.4], it was proved for RLPEs that if $q > 0$ is sufficiently large so that $q + \operatorname{Re}\psi(\xi) > 0$ for ξ in the strip $\operatorname{Im}\xi \in (\lambda_-, \lambda_+)$, for some $\lambda_- < 0 < \lambda_+$, and G is measurable and bounded, then a solution of the problem (3.7), (3.6) is unique in the class of measurable bounded functions; if the order of the process is 2 (that is, the diffusion component is nontrivial) or the process is of order $\nu < 1$ with the drift pointing to the barrier, then the uniqueness holds in the class of continuous bounded functions. Moreover, it was proved that, for sufficiently large $q > 0$,

$$\hat{V}(z,x) = (N/T)q^{-1}\mathcal{E}_q^- \mathbb{1}_{(h,+\infty)}\mathcal{E}_q^+ G(x). \tag{3.8}$$

The RHS admits the analytic continuation into an open half-plane $\operatorname{Re} q > 0$ and hence into the disc $|z| \le 1$. The left-hand side (LHS) admits the analytic continuation into this disc as well, and hence the equality (3.8) holds for z in this disc.

Since $\hat{V}(z,x)$ is analytic in the disc $|z| \le 1$, we can recover

$$V^N(x) = \frac{1}{2\pi i}\int_{|z|=1} z^{-N}\left(\frac{N}{T}\right) q^{-1}\mathcal{E}_q^- \mathbb{1}_{(h,+\infty)}\mathcal{E}_q^+ G(x)dz. \tag{3.9}$$

Change the variable $z = 1 + T(r-q)/N$, $dz = -(T/N)dq$, and denote by $\mathcal{C}_N$ the contour $\{q \mid |1 + (T/N)(r-q)| = 1\}$. Then

$$V^N(x) = -\frac{1}{2\pi i}\int_{\mathcal{C}_N}\left(1 + \frac{T(r-q)}{N}\right)^{-N} q^{-1}\mathcal{E}_q^- \mathbb{1}_{(h,+\infty)}\mathcal{E}_q^+ G(x)dq. \tag{3.10}$$

We integrate by parts k $(< N)$ times to get

$$\begin{aligned} V^N(x) = {} & -\frac{N^k(N-k-1)!}{(N-1)!}\cdot\frac{1}{2\pi i(-T)^k} \\ & \times \int_{\mathcal{C}_N}\left(1 + \frac{T(r-q)}{N}\right)^{k-N} \partial_q^k\left(q^{-1}\mathcal{E}_q^- \mathbb{1}_{(h,+\infty)}\mathcal{E}_q^+\right) G(x)dq. \end{aligned} \tag{3.11}$$

3.2. Idea of the proofs of convergence. We compare the price of the barrier option given by (2.9) and Carr's randomization approximation (3.11). Denote by $\mathcal{C}_{N,\sigma}$ the simple closed contour, which coincides with $\mathcal{C}_N$ in the open half-plane $\mathrm{Re}\, q > \sigma$ and with a segment of the line $\mathrm{Re}\, q = \sigma$ in the closed half-plane $\mathrm{Re}\, q \le \sigma$. Clearly, if σ is fixed and N is sufficiently large, we can transform $\mathcal{C}_N$ in (3.11) into $\mathcal{C}_{N,\sigma}$. Next, fix a positive integer k and $\rho \in (0, 1/2)$, and set $\mathcal{C}_{N,\sigma,\rho} = \{q \mid \mathrm{Re}\, q = \sigma,\ |\mathrm{Im}\, q| \le N^\rho\}$. It follows from estimate (6.24), which we will derive in section 6.4, that, as $N \to \infty$,

$$V^N(x) = -\frac{N^k(N-k-1)!}{(N-1)!} \cdot \frac{1}{2\pi i(-T)^k} \times \int_{\mathcal{C}_{N,\sigma,\rho}} \left(1 + \frac{T(r-q)}{N}\right)^{k-N} \partial_q^k \left(q^{-1}\mathcal{E}_q^- \mathbb{1}_{(h,+\infty)} \mathcal{E}_q^+\right) G(x) dq + o(1), \tag{3.12}$$

where the segment $\mathcal{C}_{N,\sigma,\rho}$ is passed down. Since $\rho \in (0, 1/2)$, we have for $|q| \le N^\rho$

$$(1 + T(r-q)/N)^{k-N} = \exp[(k-N)\ln(1 + T(r-q)/N)] = e^{(q-r)T} + o(1);$$

therefore, as $N \to \infty$,

$$V^N(x) = \frac{N^k(N-k-1)!}{(N-1)!} \cdot \frac{e^{-rT}}{2\pi i(-T)^k} \times \int_{\sigma - iN^\rho}^{\sigma + iN^\rho} e^{qT} \partial_q^k \left(q^{-1}\mathcal{E}_q^- \mathbb{1}_{(h,+\infty)} \mathcal{E}_q^+\right) G(x) dq + o(1). \tag{3.13}$$

Using (6.24) once again, we pass to the limit in (3.13) and obtain that

$$\lim_{N\to\infty} V^N(x) = \frac{e^{-rT}}{2\pi i(-T)^k} \int_{\mathrm{Re}\, q = \sigma} e^{qT} \partial_q^k \left(q^{-1}\mathcal{E}_q^- \mathbb{1}_{(h,+\infty)} \mathcal{E}_q^+\right) G(x) dq. \tag{3.14}$$

The absolute convergence of the integral will be proved in section 6.5.

Rigorous proofs of the asymptotic formulas above and similar formulas for sensitivities and the leading term of asymptotics require certain regularity conditions on the characteristic exponent, which are formulated in section 4. In fact, asymptotic formulas of a simple form are possible only if the characteristic exponent $\psi(\xi)$ stabilizes to a positively homogeneous function as $\xi \to \infty$ in the domain of analyticity of ψ. This is the basis for the definitions below. Luckily, the popular classes of Lévy processes satisfy this condition.

4. Regular Lévy processes of exponential type.

4.1. Definition. The main classes of processes used in the study of financial markets (see the examples in the next subsection) enjoy the following property: there exist $\lambda_- < -1 < 0 < \lambda_+$ such that the underlying Lévy process X is of exponential type (λ_-, λ_+). This means that the characteristic exponent $\psi(\xi)$ of X admits the analytic continuation into the open strip $\mathrm{Im}\, \xi \in (\lambda_-, \lambda_+)$. Moreover, $\psi(\xi)$ grows at most polynomially as $\mathrm{Re}\, \xi \to \pm\infty$ within every closed strip $\mathrm{Im}\, \xi \in [\omega_-, \omega_+] \subset (\lambda_-, \lambda_+)$. For the behavior of the value function (especially near the barrier) to be sufficiently regular, and for the leading term of the asymptotics of a simple form to exist, we need additional regularity conditions. In [15, Def. 3.2], such conditions were formulated, and the resulting class of processes was called regular Lévy processes of exponential type (RLPEs) in one dimension. Two definitions were given: in terms of the properties of the Lévy measure and in terms of the Lévy exponent. Loosely speaking, the first definition postulates that the Lévy measure behaves, asymptotically, as $c_\pm|x|^{-\nu-1}$ as $x\to\pm0$ and decays at least as fast as $e^{-\lambda^\pm x}$ as $x\to\mp\infty$. An almost equivalent definition in terms of the characteristic exponent is as follows.

Definition 4.1. *Let $\lambda_-<0<\lambda_+$, and let $\nu\in(0,2]$. We call X an RLPE (λ_-,λ_+) of order ν if its characteristic exponent admits the analytic continuation in the strip $\operatorname{Im}\xi\in(\lambda_-,\lambda_+)$, and, for some $\mu\in\mathbb{R}$, function $\psi^0(\xi):=\psi(\xi)+i\mu\xi$ is asymptotically positively homogeneous of order ν as $\operatorname{Re}\xi\to\pm\infty$ and ξ remains in any strip $\operatorname{Im}\xi\in[\omega_-,\omega_+]\subset(\lambda_-,\lambda_+)$:*

$$\psi^0(\xi)=d^0_\pm|\xi|^\nu+O(|\xi|^{\nu_1}),\tag{4.1}$$

where $\operatorname{Re}d^0_\pm>0$ and $\nu_1<\nu$. In addition, there exists $\nu_2<\nu$ such that

$$(\psi^0)'(\xi)=O(|\xi|^{\nu_2})\tag{4.2}$$

as $\xi\to\infty$ in the strip.

The following class of *strongly* regular Lévy processes of exponential type (sRLPEs), which is slightly less general than the class of RLPEs, was introduced in [9] to study the asymptotics of the prices of barrier options and first-touch digitals near the barrier.

Definition 4.2. *Let $\lambda_-<0<\lambda_+$, and let $\nu\in(0,2]$. We call X an sRLPE (λ_-,λ_+) of order ν if the following conditions hold:*

(i) *the characteristic exponent ψ admits the analytic continuation into the complex plane with two cuts $i(-\infty,\lambda_-]$ and $i[\lambda_+,+\infty)$;*
(ii) *for $z<\lambda_-$ and $z>\lambda_+$, the limits $\psi(iz\pm0)$ exist;*
(iii) *there exists $\mu\in\mathbb{R}$ such that the function $\psi^0(\xi):=\psi(\xi)+i\mu\xi$ is asymptotically positively homogeneous of order ν as $\xi\to\infty$ in the complex plane with these cuts;*
(iv) *to be more specific, there exist $\nu_1<\nu$ and $d^0_\pm$, $\operatorname{Re}d^0_\pm>0$, such that, as $\rho\to+\infty$,*

$$\psi^0(\rho e^{i\varphi})=d^0_+e^{i\varphi\nu}\rho^\nu+O(\rho^{\nu_1})\quad\forall\,\varphi\in[0,\pi/2-0];\tag{4.3}$$
$$\psi^0(\rho e^{i\varphi})=d^0_-e^{i(-\pi+\varphi)\nu}\rho^\nu+O(\rho^{\nu_1})\quad\forall\,\varphi\in[\pi/2+0,\pi];\tag{4.4}$$
$$\psi^0(\rho e^{i\varphi})=d^0_-e^{i(\pi+\varphi)\nu}\rho^\nu+O(\rho^{\nu_1})\quad\forall\,\varphi\in[-\pi,-\pi/2-0];\tag{4.5}$$
$$\psi^0(\rho e^{i\varphi})=d^0_+e^{i\varphi\nu}\rho^\nu+O(\rho^{\nu_1})\quad\forall\,\varphi\in[-\pi/2+0,0].\tag{4.6}$$

The notation $\varphi=\pi/2\pm0$ means that $\eta=\rho e^{i\varphi}$ is of the form $\eta=iz\mp0$, where $z>0$, and $\varphi=-\pi/2\pm0$ means that $\eta=\rho e^{i\varphi}$ is of the form $\eta=iz\pm0$, where $z<0$.

4.2. Model classes. It is straightforward to verify that the following popular classes of Lévy processes are sRLPEs:

(1) A Brownian motion (used in the classical Black–Scholes model [5]) is an sRLPE of order 2 and exponential type $(-\infty,\infty)$. Its characteristic exponent is given by $\frac{\sigma^2}{2}\xi^2-i\mu\xi$, where $\sigma>0$ is the volatility and $\mu\in\mathbb{R}$ is the drift of the process. We have $d^0_\pm=\sigma^2/2$.

(2) Kou's model [30], its generalization introduced in [33] and later called the hyperexponential jump-diffusion (HEJD) model in [28], and any Lévy process with the rational characteristic exponent and nontrivial Brownian motion component (in particular, Lévy processes of phase type [1]) are sRLPEs of order 2.

(3) Merton's model is an sRLPE of order 2.

(4) Lévy processes of the *extended Koponen family* (generalizing the class of processes introduced by Koponen [29]) were defined by Boyarchenko and Levendorskiĭ [13]. Later a subfamily thereof was used in [23] under the name "CGMY model," and the full family was used in [15] under the name "KoBoL processes." We use the latter term. The Lévy density of a KoBoL process has the form

$$F(dx) = c_+(-\lambda_-)x^{-\nu^- -1}e^{\lambda_- x}\mathbb{1}_{(0,+\infty)}dx + c_-\lambda_+ x^{-\nu^+ -1}e^{\lambda_+ x}\mathbb{1}_{(-\infty,0)}dx, \tag{4.7}$$

where $c_\pm \geq 0$, $\nu^\pm < 2$, and $\lambda_- < 0 < \lambda_+$; see [13, 15]. In the special case where $c_+ = c = c_-$ and $\nu^\pm = \nu \neq 0, 1$ (which corresponds to the CGMY model), the characteristic function of the process is given by

$$\psi(\xi) = -i\mu\xi + c \cdot \Gamma(-\nu) \cdot \left[(-\lambda_-)^\nu - (-\lambda_- - i\xi)^\nu + \lambda_+^\nu - (\lambda_+ + i\xi)^\nu\right]. \tag{4.8}$$

(5) Variance gamma (VG) processes were first used in empirical studies of financial markets by Madan and collaborators [39, 38, 37]. The characteristic exponent of a VG process is of the form[3]

$$\psi(\xi) = -i\mu\xi + c_+[\ln(-\lambda_- - i\xi) - \ln(-\lambda_-)] + c_-[\ln(\lambda_+ + i\xi) - \ln(\lambda_+)], \tag{4.9}$$

where $\lambda_- < 0 < \lambda_+$, $c > 0$, and $\mu \in \mathbb{R}$. A VG process with these parameters is also a Lévy process of exponential type (λ_-, λ_+); but it is not an sRLPE because the jump term in (4.9) increases at infinity as a logarithm. We will call the VG model an sRLPE of order $\nu = 0+$.

(6) Normal inverse Gaussian (NIG) processes were introduced by Barndorff-Nielsen [2]. A natural generalization of NIG was constructed in [3] and called a *normal tempered stable* (NTS) *Lévy process*. The characteristic exponent of an NTS process is of the form

$$\psi(\xi) = -i\mu\xi + \delta \cdot \left[\left(\alpha^2 - (\beta + i\xi)^2\right)^{\nu/2} - (\alpha^2 - \beta^2)^{\nu/2}\right], \tag{4.10}$$

where $\nu \in (0, 2)$, $\alpha > |\beta| > 0$, $\delta > 0$, and $\mu \in \mathbb{R}$. An NTS process with these parameters has exponential type $(\beta - \alpha, \beta + \alpha)$, order ν, and $d^0_\pm = \delta$. NIG is an NTS process of order $\nu = 1$.

(7) The β-family of Lévy processes constructed in [32] is defined by the Lévy density

$$F(dx) = c_1 \frac{e^{-\alpha_1 \beta_1 x}}{(1 - e^{-\beta_1 x})^{\gamma_1}} \mathbb{1}_{(0,+\infty)}(x) + c_2 \frac{e^{\alpha_2 \beta_2 x}}{(1 - e^{\beta_2 x})^{\gamma_2}} \mathbb{1}_{(-\infty,0)}(x), \tag{4.11}$$

where $c_j \geq 0$, $\alpha_j, \beta_j > 0$, and $\gamma_j \in (0, 3)$. For any positive integer N, $F(dx)$ can be represented as a sum of a finite number of densities of the form (4.7), with decreasing $\nu^\pm$, and

[3] What we present is not the most common way of writing the formula. Rather, we chose an expression that is equivalent to the standard one and makes the analogy with (4.8) transparent.

not necessarily positive $c_\pm$, plus a density of the form $p_N(x)dx$, where, for each $s = 0, 1, \dots,$ $p_N^{(s)}(x) = |x|^{N-s}$ as $x \to 0$ and exponentially decays as $x \to +\infty$. Therefore, the straightforward calculations (the same as those used to derive the formulas for the characteristic exponent of the processes of the extended Koponen family in [13, 15]) show that the characteristic exponent of the β model and its derivatives admit the same estimates as in the (general) KoBoL model with the steepness parameters and orders for positive and negative jumps $\lambda_+ = \alpha_2\beta_2$, $\nu^+ = 1 - \gamma_2$, $\lambda_- = -\alpha_1\beta_1$, $\nu^- = 1 - \gamma_1$.

Hyperbolic processes are sRLPEs of order 1 (see, e.g., [15]), but the verification is not so easy.

4.3. Important constants characterizing an sRLPE. For real ξ, $\overline{\psi^0(\xi)} = \psi^0(-\xi)$; hence, $\overline{d^0_-} = d^0_+$. We consider the following cases:

(i) if $\nu \in (1, 2]$ or $\nu \in (0, 1)$ and $\mu = 0$, set $d_\pm = d^0_\pm$, $d = |d_\pm|$, $\gamma_\pm = \arg d_\pm$, $\bar\nu = \nu$,

$$\nu_\pm = \nu/2 - \gamma_\pm/\pi; \tag{4.12}$$

(ii) if $\nu = 1$, set $d_\pm = \mp i\mu + d^0_\pm$, $d = |d_\pm|$, $\gamma_\pm = \arg d_\pm$, $\nu_\pm = 1/2 - \gamma_\pm/\pi$, $\bar\nu = 1$;

(iii) if $\nu \in [0+, 1)$ and $\mu > 0$, set $d_\pm = \mp i\mu$, $\nu_+ = 1$, $\nu_- = 0$, $\bar\nu = 1$;

(iv) if $\nu \in [0+, 1)$ and $\mu < 0$, set $d_\pm = \mp i\mu$, $\nu_+ = 0$, $\nu_- = 1$, $\bar\nu = 1$.

Notice that, in all cases, $\nu_+ + \nu_- = \bar\nu$, and

$$\nu_\pm = \bar\nu/2 - \gamma_\pm/\pi. \tag{4.13}$$

Furthermore, $\nu_\pm \in (0, \nu)$ if $\nu \geq 1$; if $\nu \in (0, 1)$ and $\mu = 0$, then we require that $\gamma := |\gamma_\pm|$ be in $[0, \pi\nu/2)$; then $\nu_\pm \in (0, 1)$ as well.

As we will see, constants $\nu_\pm$ characterize the rate of decay of the Wiener–Hopf factors at infinity and, as a consequence, the rate of decay of the price of the barrier option near the barrier as well.

5. Wiener–Hopf factorization.

5.1. Three forms of the Wiener–Hopf factorization. Let $q > 0$, and let T_q be an exponential random variable of mean $1/q$, independent of X. The proof of (2.7) in [11] is very close to the proof of the Wiener–Hopf factorization formula in the form used in probability (see, e.g., [41, p. 98]):

$$\mathbb{E}\left[e^{X_{T_q}}\right] = \mathbb{E}\left[e^{\overline{X}_{T_q}}\right] \cdot \mathbb{E}\left[e^{\underline{X}_{T_q}}\right]. \tag{5.1}$$

The operator form of the Wiener–Hopf factorization

$$\mathcal{E}_q = \mathcal{E}^-_q \mathcal{E}^+_q = \mathcal{E}^+_q \mathcal{E}^-_q \tag{5.2}$$

can also be proved similarly to (5.1). Finally, using the definition

$$\phi^+_q(\xi) = \mathbb{E}\left[e^{i\xi\overline{X}_{T_q}}\right], \qquad \phi^-_q(\xi) = \mathbb{E}\left[e^{i\xi\underline{X}_{T_q}}\right] \tag{5.3}$$

of the Wiener–Hopf factors and noticing that

$$\mathbb{E}\left(e^{iX_{T_q}\xi}\right) = \frac{q}{q + \psi(\xi)}, \tag{5.4}$$

we can write (5.1) in the form

$$\frac{q}{q+\psi(\xi)} = \phi_q^+(\xi)\phi_q^-(\xi). \tag{5.5}$$

Equation (5.5) is a special case of the factorization of functions on the real line into a product of two functions analytic in the upper and lower open half-planes and admitting the continuous continuation up to the real line. This is the initial factorization formula discovered by Wiener and Hopf [44] in 1931 for functions of a much more general form than in the LHS of (5.5).

5.2. Realization of the EPV operators using the Fourier transform. Decomposing a sufficiently regular function $f(x)$ as a Fourier integral and using (5.4), we obtain

$$(\mathcal{E}_q f)(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\xi x}\frac{q\widehat{f}(\xi)}{q+\psi(\xi)}\,d\xi, \tag{5.6}$$

where $\widehat{f}(\xi) = \mathcal{F}_{x\to\xi}f$ is the Fourier transform of f. Identity (5.6) can be justified under fairly weak regularity assumptions; we refer the reader to [15, sect. 2.3.3] for the details (with the notation of *op. cit.*, we have $\mathcal{E}_q = qU^q$, where U^q is referred to as the resolvent operator, or the q-potential operator, of X).

Similarly, it follows from (2.5)–(2.6) and (5.3) that

$$(\mathcal{E}_q^\pm f)(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\xi x}\phi_q^\pm(\xi)\widehat{f}(\xi)\,d\xi. \tag{5.7}$$

We conclude that $\mathcal{E}_q$ and $\mathcal{E}_q^\pm$ are pseudodifferential operators (PDOs) with the symbols $q/(q+\psi(\xi))$ and $\phi_q^\pm(\xi)$, respectively.

To realize PDOs $\mathcal{E}_q^\pm$, one needs explicit analytical expressions for the Wiener–Hopf factors $\phi^\pm(\xi)$. There exist general formulas [42, eqs. (4.5), (4.6)] for any Lévy process, but they are inefficient for both theoretical study and computational purposes.

Under a certain regularity assumption on the characteristic exponent $\psi(\xi)$ of X (see, e.g., [15, Thm. 3.2]), Boyarchenko and Levendorskiĭ obtained integral formulas for the Wiener–Hopf factors $\phi_q^\pm(\xi)$, which are efficient for computational purposes. This assumption holds in all model examples of Lévy processes of exponential type, including those listed in section 4.2, so we prefer not to state it to save space. To state these formulas, recall that (see [15, eq. (3.40)]) if the characteristic exponent is analytic in a strip (λ_-,λ_+), then, for any $q>0$, there exist $\omega_- < 0 < \omega_+$ and $\delta > 0$ such that

$$\operatorname{Re}(q+\psi(\xi)) \ge \delta, \quad \operatorname{Im}\xi \in [\omega_-,\omega_+]. \tag{5.8}$$

The formulas [15, eqs. (3.58), (3.60)] are as follows: for $\pm\operatorname{Im}\xi > \pm\omega_\mp$,

$$\phi_q^\pm(\xi) = \exp\left[\pm\frac{1}{2\pi i}\int_{\operatorname{Im}\eta=\omega_\mp}\frac{\xi\cdot\ln(q+\psi(\eta))}{\eta(\xi-\eta)}\,d\eta\right]. \tag{5.9}$$

Note that one can take any $\omega_- \in (\lambda_-,0)$ and $\omega_+ \in (0,\lambda_+)$ and, after that, choose a sufficiently large $q>0$ so that (5.8) holds. Next, to apply the inverse Laplace transform w.r.t. q, we need to consider the analytic continuation w.r.t. q as well. In addition, we need to obtain estimates for $\phi_q^\pm(\xi)$ and, whenever possible, calculate the leading term of asymptotics of $\phi_q^\pm(\xi)$ as $\xi\to\infty$. It follows from (5.8) that if $\sigma>0$ is sufficiently large, then the RHS in (5.9) defines the analytic continuation into the half-plane $\operatorname{Im} q \ge \sigma$.

6. Convergence of Carr's randomization approximation.

6.1. Main results. We consider a down-and-out barrier option with the payoff satisfying (2.1), fixed time to maturity T, and barrier at $x = 0$.

For $s \geq 0$, denote by $C^s([0+,\infty))$ the space of functions f defined on $(0,+\infty)$, which admit continuations $lf \in C^s(\mathbb{R})$, with the norm $||f||_s = \inf ||lf||_s$, and let $\nu_\pm$ be the constants defined in section 4.3.

Theorem 6.1. *Let X be either an RLPE of order $\nu \in (1,2]$ or an sRLPE of order $\nu \in (0,1]$ or an sRLPE of order $0+$ with the drift pointing from the boundary.*

Then there exists $s > 0$ such that Carr's randomization approximation $V^N(\cdot)$ converges to the option price $V(T,\cdot)$ in the topology of the Hölder space $C^{\nu_-}([0+,\infty))$.

Theorem 6.2. *For any integer $m > 0$, there exist functions $c_j(m,T,\cdot) \in C^{\nu_-}([0+,\infty))$, $j = 1,2,\dots,m$, which are independent of N, such that, as $N \to \infty$,*

$$V^N(x) = V(T,x) + \sum_{j=1}^{m} c_j(m,T,x)N^{-j} + O(N^{-m-1}), \tag{6.1}$$

where the O-term is understood in the sense of the $C^{\nu_-}([0+,\infty))$-norm.

We conclude that if, for large N, and several multiples of N, say, $N_j = n_jN$, $j = 1,2,\dots,m$, Carr's randomization approximations $V^{N_j}(x)$ are calculated, then we can use the extrapolation formula of order m in $1/N$-line to calculate $W(0) := V(T,x)$ given $W(1/N_j) := V^{N_j}(x)$, $j = 1,2,\dots,m$. In particular, the linear extrapolation gives $V(T,x) = 2V^{2N}(x) - V^N(x)$, and the quadratic extrapolation used in [20] is $V(T,x) = 0.5V^N(x) - 4V^{2N}(x) + 4.5V^{3N}(x)$. The next two versions of quadratic extrapolation $V(T,x) = (1/3)V^N(x) - 2V^{2N}(x) + (8/3)V^{4N}(x)$ and $V(T,x) = V^{2N}(x) - 3V^{3N}(x) + 3V^{6N}(x)$ are more accurate but need more CPU time. The reader can easily derive extrapolations of higher order; however, we have found that quite often even linear extrapolation with moderate $N = 20$ for $T = 1$ gives the result with the relative error less than 0.5 percent.

Theorem 6.3. *Let X be either an RLPE of order $\nu \in (1,2]$ or an sRLPE of order $\nu \in (0,1]$ or an sRLPE of order $0+$ with the drift pointing from the boundary.*

Then there exists $\sigma > 0$ such that the following hold:

(a) *function $\tilde{G}(q,0+) = (\mathcal{E}_q^+ G)(0+)$ is well defined and analytic in the half-plane $\mathrm{Re}\, q \geq \sigma$;*

(b) *for q in the half-plane $\mathrm{Re}\, q \geq \sigma$, the Wiener–Hopf factors have the following asymptotics as $\xi \to \infty$ in a strip around the real axis:*

$$\phi_q^\pm(\xi) = \phi_{q,\infty}^\pm(1 \mp i\xi)^{-\nu_\pm}(1 + O(|\xi|^{-\rho})), \tag{6.2}$$

where $\rho > 0$, and functions $q \mapsto \phi_{q,\infty}^\pm$ are analytic in the half-plane $\mathrm{Re}\, q \geq \sigma$;

(c) *there exist $C, s > 0$ and integer $k \geq 1$ such that*

$$|\partial_q^k(q^{-1}\phi_{q,\infty}^- \tilde{G}(q,0+))| \leq C|q|^{-1-s}, \quad \mathrm{Re}\, q \geq \sigma; \tag{6.3}$$

hence, the following integrals absolutely converge:

$$\kappa_k(T) = \frac{e^{-rT}}{2\pi i \Gamma(1+\nu_-)(-T)^k} \int_{\mathrm{Re}\, q = \sigma} e^{qT} \partial_q^k(q^{-1}\phi_{q,\infty}^- \tilde{G}(q,0+))dq \tag{6.4}$$

and

$$\kappa_{\mathrm{Carr};N,k}(T) = -\frac{N^k(N-k-1)!}{(N-1)!}\cdot\frac{1}{2\pi i(-T)^k} \times \int_{\mathcal{C}_N}\left(1+\frac{T(r-q)}{N}\right)^{k-N}\partial_q^k\left(q^{-1}\phi^-_{q,\infty}\tilde{G}(q,0+)\right)dq; \tag{6.5}$$

(d) *there exists* $s>0$ *such that, as* $x\downarrow 0$,

$$V(T,x) = \kappa_k(T)x^{\nu_-} + O(x^{\nu_-+s}) \tag{6.6}$$

and

$$V^N(x) = \kappa_{\mathrm{Carr};N,k}(T)x^{\nu_-} + O(x^{\nu_-+s}); \tag{6.7}$$

(e) $\kappa_{\mathrm{Carr};N,k}(T)\to\kappa_k(T)$ *as* $N\to\infty$, *and the* O*-term in* (6.7) *converges to the* O*-term in* (6.6) *in the* C^{ν_-+s}*-norm.*

Remark 6.4. (i) The proofs of Theorems 6.1 and 6.2 and part (e) of Theorem 6.3, which are the first main results of the paper for the price, are based on parts (a)–(d) of Theorem 6.3. Statements (a)–(c) and (6.6) are proved in [9] for sRLPEs of order $\nu\in[0+,2]$, with the exception of the VG model with zero drift or drift pointing to the boundary; therefore, the proof of convergence based on the crucial estimates is valid for sRLPEs (with the same reservation) as well. The proof of (6.7) is almost identical to the proof of (7.11). In the present paper, we complement the results in [9] calculating the leading term of asymptotics for all RLPEs of order $\nu>1$, in particular, for all standard classes of processes with a nontrivial Brownian motion component.

(ii) In this section, we prove that if k is sufficiently large, then the integrals in (2.10) and (3.11) converge absolutely and uniformly w.r.t. T; moreover, the latter converges to the former as $N\to\infty$ in the Hölder space of order $s<\nu_-$. We also prove Theorem 6.2. These proofs are simpler than the proof of Theorem 6.3, which we give in section 7. Theorem 6.1 follows from Theorem 6.3(e).

(iii) For processes of order $\nu>1$, we will show that it is possible to take $k=1$.

(iv) The proofs are based on several crucial estimates listed in section 6.3 and proved below for RLPEs of order $\nu>1$. Similar estimates are derived in [9] for sRLPEs of order $\nu\in(0,2]$ and for the VG model with the drift pointing from the boundary.

6.2. Formulas and estimates for the Wiener–Hopf factors. Fix $\lambda_-<\omega_-<0<\omega_+<\lambda_+$. The statements below are valid for η in the strip $\operatorname{Im}\eta\in[\omega_-,\omega_+]$, and q in the half-plane $\operatorname{Re}q\geq\sigma$, if σ is sufficiently large. ("Large" means that $\operatorname{Re}(q+\psi(\eta))>0$ for (q,η) indicated.)

Define $\nu_\pm$ as in section 4.3 for the case $\nu>1$. Then $\nu_++\nu_-=\nu$, and $\nu_\pm\in(0,\nu)$.

Lemma 6.5. *Let* X *be an RLPE of order* $\nu\in(1,2]$*. Then the following hold:*

(a) *there exist* $\rho>0$ *and* $d>0$ *such that*

$$\Psi(q,\eta) := \frac{1+\psi(\eta)/q}{(1-i\eta(d/q)^{1/\nu})^{\nu_+}(1+i\eta(d/q)^{1/\nu})^{\nu_-}} = 1+O(|\eta|^{-\rho}) \tag{6.8}$$

as $\eta\to\infty$*, with the coefficient in the* O*-term depending on* q*;*

(b) *moreover, for any* $s \in (0,\nu)$, ρ *and the coefficient in the* O*-term can be chosen the same for all* (q,η) *that satisfy* $|q| \le |\eta|^s$;

(c) *there exists* $C > 0$ *such that*

$$|\ln \Psi(q,\eta)| \le C \begin{cases} |\eta|^\nu/|q|, & |\eta|^\nu \le |q|, \\ |q|/|\eta|^\nu, & |\eta|^\nu \ge |q|; \end{cases} \tag{6.9}$$

(d) *for any positive integer* k, *there exists* $C_k > 0$ *such that*

$$|\partial_q^k(\ln \Psi(q,\eta))| \le C|q|^{-k}. \tag{6.10}$$

Proof. (a) and (b) can be easily verified (see [15, sect. 3.6.2]).[4] (c) As $c \to 0$, we have $\sup_{|\eta|^\nu \le c|q|} |\Psi(q,\eta) - 1| \to 0$, and, similarly to (6.8),

$$\sup_{|\eta|^\nu \ge C|q|} |\Psi(q,\eta) - 1| \to 0 \quad \text{as} \quad C \to +\infty.$$

Since $\ln(1+x) = O(x)$ as $x \to 0$, we conclude that (6.9) holds if $|\eta|^\nu \le c|q|$ or $|\eta|^\nu \ge C|q|$, where $c > 0$ is sufficiently small and $C > 0$ is sufficiently large. Since $\Psi(q,\eta)$ is uniformly bounded, (6.9) holds for $c \le |\eta|^\nu/|q| \le C$ as well.

(d) We take the multiplicative structure of $\Psi(q,\eta)$ into account and apply the Leibnitz rule. ■

If $q \ge \sigma$ is real, then it follows from [15, Thm. 3.3] that, for $\pm\mathrm{Im}\,\xi > \pm\omega_\mp$, the Wiener–Hopf factors (5.3) in the Wiener–Hopf factorization formula (5.1) can be represented in the form

$$\phi_q^\pm(\xi) = (1 \mp i\xi(d/q)^{1/\nu})^{-\nu_\pm} \exp[\hat{I}^\pm(q,0) - \hat{I}^\pm(q,\xi)], \tag{6.11}$$

where

$$\hat{I}^\pm(q,\xi) = \pm\frac{1}{2\pi i}\int_{\mathrm{Im}\,\eta=\omega_\mp} \frac{\ln \Psi(q,\eta)}{\eta - \xi}\,d\eta. \tag{6.12}$$

The integral in (6.12) absolutely converges for (q,ξ) under consideration due to (6.8) and defines an analytic function in the region $\mathrm{Re}\,q \ge \sigma$, $\pm\mathrm{Im}\,\xi > \pm\omega_\mp$. In the following lemma, we state important bounds for functions $\hat{I}^\pm(q,\xi)$ and the Wiener–Hopf factors. The bounds will be valid for $\pm\mathrm{Im}\,\xi \ge \pm\omega'_\mp$, where $\lambda_- < \omega_- < \omega'_- < 0 < \omega'_+ < \omega_+ < \lambda_+$.

6.3. Crucial estimates. In Appendix C, we will prove the following two lemmas.

Lemma 6.6. (a) *There exist* $C, c > 0$ *such that*

$$|\hat{I}^\pm(q,\xi)| \le C, \tag{6.13}$$

$$c(1 + |\xi|/|q|^{1/\nu})^{-\nu_\pm} \le |\phi_q^\pm(\xi)| \le C(1 + |\xi|/|q|^{1/\nu}|)^{-\nu_\pm}. \tag{6.14}$$

[4] Condition $\nu > 1$ forces $q^{1/\nu}$ to be in a sector $\Sigma_{\sigma,\theta} = \{q = \sigma + \rho e^{i\varphi} \mid \rho \ge 0,\ -\theta < \varphi < \theta\}$, where $\theta\pi\nu/2 > \pi/2$. In this paper, we will use the analytic continuation into the half-plane $\mathrm{Re}\,q \ge \sigma$ only.

(b) *For any $\epsilon > 0$ and positive integer k, there exists $C_{k,\epsilon} > 0$ such that*

$$|\partial_q^k \hat{I}^{\pm}(q,\xi)| \le C_{k,\epsilon}|q|^{\epsilon-k}, \tag{6.15}$$

$$|\partial_q^k \phi_q^{\pm}(\xi)| \le C_{k,\epsilon}|q|^{\epsilon-k}(1+|\xi|/|q|^{1/\nu})^{-\nu_\pm}. \tag{6.16}$$

Lemma 6.7. *For any $s \in (0,\nu)$, $k \in \mathbb{Z}_+$, and $\epsilon > 0$, there exists $C_{s,k,\epsilon} > 0$ such that, in the region $|q| \le |\xi|^s$,*

$$|\partial_q^k \hat{I}^{\pm}(q,\xi)| \le C_{s,k,\epsilon}|q|^{-k}(1+|\xi|)^{-1+s/\nu+\epsilon}. \tag{6.17}$$

Furthermore,

$$\phi_q^{\pm}(\xi) = \phi_{q,\infty}^{\pm}(1 \mp i\xi)^{-\nu_\pm} + R_\pm(q,\xi), \tag{6.18}$$

where

$$\phi_{q,\infty}^{\pm} = (q/d)^{\nu_\pm/\nu}\exp[\hat{I}^{\pm}(q,0)], \tag{6.19}$$

and $R_\pm(q,\xi)$ admits the following estimate:

$$|\partial_q^k R_\pm(q,\xi)| \le C_{s,k,\epsilon}|q|^{\nu_\pm/\nu-k}|\xi|^{-\nu_\pm-1+s/\nu+\epsilon}. \tag{6.20}$$

The proofs of Lemmas 6.6 and 6.7 and the form of the asymptotic coefficient $\phi_{q,\infty}^{\pm}$ rely on the assumption $\nu > 1$; however, the statements of these lemmas, with different $\nu_\pm \in [0,1]$ and expressions for $\phi_{q,\infty}^{\pm}$, are valid for sRLPEs of order $\nu \in [0+,1]$ as well, with certain modifications for the upper bounds in (6.16) for processes of order $\nu \in [0+,1)$, with nonzero drift. Therefore, the proofs of the main results in the following subsections are valid for these processes as well.

6.4. Convergence of the integrals in (2.10) and (3.11). Without loss of generality, we may assume that $h = 0$. Since $\nu \in (1,2]$, we have $\nu_+ \in (0,\nu)$, and therefore it follows from (6.16) with sign "+" and (2.1) that, for any $k \in \mathbb{Z}_+$, $\epsilon > 0$, and $a \in [0,\nu_+]$,

$$|\partial_q^k \phi_q^+(\xi)\hat{G}(\xi)| \le C_{k,\epsilon}(1+|\xi|)^{-1-a}|q|^{a/\nu-k+\epsilon}, \tag{6.21}$$

where $C_{k,\epsilon}$ is independent of $q \in \Sigma_{\sigma,\theta}$ and ξ in the half-plane $\operatorname{Im}\xi \ge \omega'_-$. Therefore, for k, a, ϵ satisfying the same conditions, function $\tilde{G}(q,\cdot) := \mathcal{E}_q^+ G = \phi_q^+(D)G$ admits the bound

$$||\partial_q^k \tilde{G}(q,\cdot)||_{1/2+a+\epsilon} = O(|q|^{a/\nu-k+\epsilon}), \tag{6.22}$$

where $||\cdot||_s$ denotes the norm in $H^s(\mathbb{R})$. Letting $a = 0$ and using the Sobolev embedding theorem, we conclude that $\partial_q^k \tilde{G}(q,\cdot)$ is continuous on $(0,+\infty)$ and has the right limit at 0, which admits the bound

$$|\partial_q^k \tilde{G}(q,0+)| \le C_{k,\epsilon}|q|^{-k+\epsilon} \tag{6.23}$$

for any $\epsilon > 0$, where $C_{k,\epsilon}$ is independent of $q \in \Sigma_{\sigma,\theta}$. Next, using the general facts collected in Appendix A, we deduce from (6.22) that, for any $s \in (-1/2,1/2)$ and $\epsilon > 0$, $\partial_q^k \mathbb{1}_{[0,+\infty)}\tilde{G}(q,\cdot) =$

$O(|q|^{\epsilon-k})$ as an element of $\mathring{H}^s(\mathbb{R}_+)$. Set $f_k(q,x) = \partial_q^k(q^{-1}\mathcal{E}_q^- \mathbb{1}_{[0,+\infty)}\tilde{G}(q,\cdot))$, and denote by $\hat{f}_k(q,\xi)$ the Fourier transform of $f_k(q,x)$ w.r.t. x. Using (6.22) and applying (6.16) with sign "$-$," we conclude that if $k \geq 1$, then, for any $\epsilon, \epsilon' > 0$, there exists $C_{\epsilon,\epsilon'}$ such that

$$|\hat{f}_k(q,\xi)| \leq C_{\epsilon,\epsilon'}(1+|\xi|)^{-1-\nu_-+\epsilon'}|q|^{-1-k+\nu_-/\nu+\epsilon}. \tag{6.24}$$

Since $0 < \nu_- < \nu$, (6.24) implies that, by the Sobolev embedding theorem, the integral (2.10) absolutely converges and defines a continuous function of T, with values in $\mathring{H}^{1/2+\nu_- -\epsilon_1}(\mathbb{R}_+) \subset C^{\nu_- -\epsilon_2}(\mathbb{R})$, for any $0 < \epsilon_2 < \epsilon_1 < \epsilon'$.

To prove that the integral in (3.11) converges absolutely and uniformly w.r.t. N and T, it suffices to note that the length of $\mathcal{C}_N$ is $O(N)$ (which is $O(|q|)$ for q on the contour of integration) and use (6.24).

6.5. Convergence of Carr's randomization. It follows from (6.24) that (3.12) holds, and (3.13) follows. Using (6.24) once again, we pass to the limit in (3.13) and obtain (3.14). We consider the case $\nu > 1$, which implies that $\nu_- > 0$. Therefore, the integral in (3.14) absolutely converges for any $k \geq 1$, and we can integrate back and obtain (3.14) with $k = 1$. In some other cases, it may be necessary to use larger k; see [9]. Finally, we note that the proof above gives the convergence of Carr's randomization in the C^s-norm for any $s < \nu_-$. To prove the convergence in C^{ν_-}, we need to calculate the leading term of asymptotics near the boundary.

6.6. Justification of Richardson extrapolation of arbitrary order. Fix a positive integer m, and choose $k > m$ and $\rho \in (0,1)$ such that $(k-\nu_-/\nu-m-1)\rho > m+1$. Then the O-term in (3.12) can be replaced with $O(N^{-m-1})$ and, in the integral in (3.12), the asymptotic expansion of the form

$$\begin{aligned}(1+T(r-q)/N)^{k-N} &= e^{T(q-r)}(1+T(r-q)/N)^k \\ &\quad \times \left(1+\sum_{j=1}^m c_j(T(r-q)/N)^j\right) + O((Tq/N)^{m+1})\end{aligned}$$

used. The O-term above can be omitted but an additional $O(N^{-m-1})$ term in (3.12) added, and then the integral over $\mathcal{C}_{N,\sigma,\rho}$ replaced with the integral over $\mathrm{Re}\, q = \sigma$, at the same cost. Since, for a fixed k, the coefficient in front of the integral in (3.12) can be represented as a convergent series in $1/N$, we arrive at (6.1), where the coefficients are independent of N and belong to the same Hölder space as indicated in the proof of convergence of Carr's randomization to the price; the asymptotic expansion can be understood in the topology of this space.

7. Asymptotics and convergence: The leading term for prices. Without loss of generality, we assume that $h = 0$, and we use the notation $f_k(q,x)$ and $\hat{f}(q,\xi)$ introduced in section 6.4.

7.1. Asymptotics of the price near the barrier. *Step* I. Take $s > 0$, and consider the set $\{(\xi,q) \mid |\xi|^s \leq |q|\}$. On this set, we have $|q|^{-k+\nu_-/\nu} \leq |\xi|^{-s(1-\nu_-/\nu)}$ since $0 < \nu_-/\nu < 1$. Now it follows from (6.24) that, for any $\epsilon, \epsilon' > 0$,

$$|\mathbb{1}_{|\xi|^s \leq |q|}\hat{f}_k(q,\xi)| \leq C_{\epsilon,\epsilon'}(1+|\xi|)^{-1-\nu_- -s(1-\nu_-/\nu)+\epsilon'}|q|^{-1-\epsilon}. \tag{7.1}$$

Hence, if we substitute the inverse Fourier transform of $\mathbb{1}_{|\xi|^s\le|q|}\hat{f}_k(q,\xi)$ for $f_k(q,x)$ in (2.10), then we obtain a continuous function of T, with values in $\mathring{H}^{1/2+\nu_-+s'}(\mathbb{R}_+)\subset C^{\nu_-+s''}(\mathbb{R})$, for any $0<s''<s'<s(1-\nu_-/\nu)$. This function decays faster than x^{ν_-} as $x\downarrow 0$. Therefore, we can continue the calculation of the leading term replacing $f_k(q,x)$ in (2.10) with $f^s_k(q,\xi)$, the inverse Fourier transform of $\mathbb{1}_{|\xi|^s\ge|q|}\hat{f}_k(q,\xi)$.

Step II. We take $s\in(0,\nu)$ and, for a multi-index $p=(p_1,p_2,p_3)$, define

$$f(p,q,x)=q^{-1-p_1}(\partial_q^{p_2}\mathcal{E}_q^-)\mathbb{1}_{(0,+\infty)}(\partial_q^{p_3}\mathcal{E}_q^+)G(x) \tag{7.2}$$

and then $f^s(p,q,\xi)$, the inverse Fourier transform of $\mathbb{1}_{|\xi|^s\ge|q|}\hat{f}(p,q,\xi)$. Applying the Leibnitz rule to $\partial_q^k\left(q^{-1}\mathcal{E}_q^-\mathbb{1}_{(0,+\infty)}\mathcal{E}_q^+G\right)(x)$ in (2.10), we obtain a linear combination of functions $f(p,q,x)$ with $|p|=k$. According to Step I, we can replace $f(p,q,x)$ with $f^s(p,q,x)$ (that is, introduce factor $\mathbb{1}_{|\xi|^s\ge|q|}$ in the dual space).

Introduce further

$$g_{\nu_-}(x)=(1+iD)^{-\nu_--1}\delta=(1/\Gamma(\nu_-+1))\mathbb{1}_{[0,+\infty)}(x)x^{\nu_-}e^{-x}, \tag{7.3}$$

$$f^1(p,q,x)=q^{-1-p_1}\partial_q^{p_2}\tilde{G}(q,0+)(\partial_q^{p_3}\mathcal{E}_q^-)(1+iD)^{-1}\delta, \tag{7.4}$$

$$f^2(p,q,\xi)=q^{-1-p_1}\partial_q^{p_2}\tilde{G}(q,0+)\partial_q^{p_3}\left((q/d)^{\nu_-/\nu}\exp[\hat{I}^-(q,0)]\right)g_{\nu_-}(x), \tag{7.5}$$

and then define $f^{j,s}(p,q,x)$ similarly to $f^s(p,q,x)$ above.

Step III. We use the equality (see (A.2))[5]

$$\mathbb{1}_{(0,+\infty)}\partial_q^k\tilde{G}=\partial_q^k\tilde{G}(q,0+)(1+iD)^{-1}\delta+(1+iD)^{-1}\mathbb{1}_{(0,+\infty)}(1+iD)\partial_q^k\tilde{G}, \tag{7.6}$$

where $\tilde{G}(q,x)=\partial_q^k\mathcal{E}_q^+G(x)$, to show that if $|p|\ge 1$, then there exist $C>0$, $s_1>1$, and $s_2>\nu_-$, independent of $q\in\partial\Sigma_{\sigma,\theta}$ and ξ in the half-plane $\operatorname{Im}\xi\le 0$, such that

$$|\widehat{f^s}(p,q,\xi)-\widehat{f^{1,s}}(p,q,\xi)|\le C|q|^{-s_1}(1+|\xi|)^{-1-s_2} \tag{7.7}$$

and conclude that, for any $\epsilon>0$,

$$\int_{\operatorname{Re}q=\sigma}e^{qT}(f^s(p,q,x)-f^{1,s}(p,q,x))dq\in\mathring{H}^{1/2+s_2-\epsilon}(\mathbb{R}_+). \tag{7.8}$$

By the Sobolev embedding theorem, the integral above defines a function of class $C^{s_2-\epsilon}$, for any $\epsilon>0$, which vanishes on $(-\infty,0]$; hence, it tends to zero as $x\to 0$ faster than x^{ν_-}. It follows that the leading term of asymptotics does not change if we replace $f^s(p,q,x)$ with $f^{1,s}(p,q,x)$.

Take $a\in(0,\min\{\nu_+,1\})$; then $-1/2<a-1/2<1/2$. Using (6.21), we conclude that, for any $\epsilon>0$,

$$|\widehat{f}(p,q,\xi)-\widehat{f^1}(p,q,\xi)|\le C_{p,\epsilon}|q|^{(a+\nu_-)/\nu-2+\epsilon}|\xi|^{-1-a-\nu_-+\epsilon}. \tag{7.9}$$

Since $\nu_\pm\in(0,\nu)$, (7.9) implies (7.7).

[5]Equation (7.6) is justified if $(1+iD)\mathcal{E}_q^+G\in H^s(\mathbb{R})$ for some $s>-1/2$; in our case, we can take $s=-1/2+\nu_+-\epsilon$ for any $\epsilon>0$ (see (6.22) and (6.21)). Since $\nu_+\in(0,1)$, we can take $s>-1/2$.

Step IV. We show that we can replace $f^{1,s}(p,q,x)$ with $f^{2,s}(p,q,x)$. As above, it suffices to prove that if s is sufficiently close to ν, then there exist $C>0$, $s_1>1$, and $s_2>\nu_-$, independent of $q\in\Sigma_{\sigma,\theta}$ and ξ in the half-plane $\operatorname{Im}\xi\le\omega'_+$, such that

$$|\widehat{f^{1,s}}(p,q,\xi)-\widehat{f^{2,s}}(p,q,\xi)|\le C|q|^{-s_1}(1+|\xi|)^{-1-s_2}. \tag{7.10}$$

But this follows from (6.19)–(6.20).

Step V. The same argument as at Step I shows that we can replace $f^{2,s}(p,q,x)$ with $f^2(p,q,x)$. After that, we use (7.3) and derive the asymptotic formula

$$V(T,x)=\kappa_k(T)x^{\nu_-}+O(x^{\nu_-+s'}),\quad x\downarrow 0, \tag{7.11}$$

for some $s'>0$, where $\kappa_k(T)$ is given by (6.4). It follows from (6.13) and (6.15) that, for $k\in\mathbb{Z}_+$ and any $\epsilon>0$,

$$|\partial_q^k\phi^\pm_{q,\infty}|\le C_{k,\epsilon}|q|^{\nu_\pm/\nu-k+\epsilon}. \tag{7.12}$$

Using (7.12) and (6.23), we conclude that, for any integer $k\ge 1$, there exist C_k and $\rho>0$ such that

$$\left|\partial_q^k\left(q^{-1}\phi^-_{q,\infty}\tilde G(q,0+)\right)\right|\le C_k|q|^{-k-\rho},\quad \operatorname{Re}q\ge\sigma. \tag{7.13}$$

Hence, we can integrate by parts back and derive (6.4) and (6.6) with $k=1$.

7.2. Asymptotics of the price in Carr's randomization approximation near the barrier. The only difference with the case of the barrier option itself is that, now, we need to estimate a series of integrals over a regular contour of length $O(|q|)$, and the functions involved admit estimates via $|q|^{-1-s}$, where $s>0$. Hence, the same argument applies, and the result is (6.7). As above, we can integrate by parts back and obtain (6.5) with $k=1$.

7.3. Convergence of the asymptotic coefficient in Carr's randomization. It follows from (7.13) that, as $N\to\infty$,

$$\begin{aligned}\kappa_{\mathrm{Carr};N,1}(T)=-\frac{N}{N-1}\cdot\frac{1}{2\pi i(-T)}\int_{\mathcal{C}_{N,\sigma,\rho}}&\left(1+\frac{T(r-q)}{N}\right)^{1-N}\\ &\times\partial_q\left(q^{-1}\phi^-_{q,\infty}\tilde G(q,0+)\right)dq+o(1),\end{aligned} \tag{7.14}$$

where $\rho\in(0,1/2)$ and $\mathcal{C}_{N,\sigma,\rho}$ are the same as in section 6.5. We have for $|q|\le N^\rho$

$$(1+T(r-q)/N)^{-N}=\exp[-N\ln(1+T(r-q)/N)]=e^{(q-r)T}+o(1);$$

therefore, as $N\to\infty$,

$$\kappa_{\mathrm{Carr};N,1}(T)=\frac{e^{-rT}}{2\pi i(-T)}\int_{\sigma-iN^\rho}^{\sigma+iN^\rho}e^{qT}\partial_q\left(q^{-1}\phi^-_{q,\infty}\tilde G(q,0+)\right)dq+o(1). \tag{7.15}$$

Finally, using (7.13) once again, we pass to the limit in (7.15) and obtain that

$$\lim_{N\to\infty}\kappa_{\mathrm{Carr};N,1}(T)=\frac{e^{-rT}}{2\pi i(-T)}\int_{\operatorname{Re}q=\sigma}e^{qT}\partial_q\left(q^{-1}\phi^-_{q,\infty}\tilde G(q,0+)\right)dq. \tag{7.16}$$

7.4. Convergence of Carr's randomization approximation to the price in the Hölder norm. A straightforward analysis of the proof of convergence above shows that the O-term in the asymptotic formula for Carr's randomization approximation converges to the O-term of the asymptotic formula for the price in the C^{ν_-+s}-norm, for some $s>0$.

8. Sensitivities. In this section, we consider four related problems for θ, Δ, and Γ of the barrier option: calculation of a sensitivity in Carr's randomization approximation; proof of convergence of Carr's sensitivity to the sensitivity of the barrier option; calculation of the leading term of the sensitivity and the leading term of its Carr's randomization approximation; convergence of the leading term. The cases of θ and Δ will be considered in more detail, and the case Γ will be only outlined, because the calculations in the last case are rather long. The detailed calculations with a systematic study of the zoology of possible shapes of sensitivity curves and surfaces will be made in a different paper. The types of processes below are the same as in Theorems 6.1 and 6.3.

8.1. Theta. By definition, $\theta=-\partial V(T,x)/\partial T$. Since we want to differentiate under the integral sign in (2.10), we need to choose k large enough so that even after we insert an additional factor q under the integral sign the integrand will remain absolutely and uniformly integrable. For such a k, which exists on the strength of estimates used to calculate the asymptotics of the price, we have

$$\begin{aligned}\theta(T,x) &= -\frac{e^{-rT}}{2\pi i(-T)^k}\\ &\quad\times\int_{\mathrm{Re}\,q=\sigma} e^{qT}\left(q-r-\frac{k}{T}\right)\partial_q^k\left(q^{-1}\mathcal{E}_q^-\mathbb{1}_{(0,+\infty)}\mathcal{E}_q^+G\right)(x)dq.\end{aligned}\tag{8.1}$$

The asymptotics of the integral in (8.1), as $x\downarrow 0$, can be calculated exactly as the asymptotics of the price was calculated, except a larger k may be necessary. Similarly, in the final formula for the asymptotic coefficient (6.4) in the formula (6.6) for the price, $k=1$ was admissible; now we need to use $k\ge 2$. The final result is the following evident analogue of (6.6) and (6.4).

Theorem 8.1. *Let k be an integer ≥ 2. There exists $s>0$ such that*

$$\theta(T,x)=\kappa^\theta(T)x^{\nu_-}+O(x^{\nu_-+s}),\quad x\downarrow 0,\tag{8.2}$$

where

$$\kappa^\theta(T)=-\frac{e^{-rT}}{2\pi i\,\Gamma(1+\nu_-)(-T)^k}\int_{\mathrm{Im}\,q=\sigma}e^{qT}\left(q-r-\frac{k}{T}\right)\partial_q^k\left(q^{-1}\phi^-_{q,\infty}\tilde{G}(q,0+)\right)dq.$$

The natural analogue of θ in Carr's randomization approximation is the finite difference

$$\theta_{\mathrm{Carr};N}(T,x)=\frac{V^{N-1}(x)-V^N(x)}{T/N},$$

where V^N is given by (3.10) and (3.11) and V^{N-1} by a similar formula:

$$V^{N-1}(x)=-\frac{1}{2\pi i}\int_{\mathcal{C}_N}\left(1+\frac{T(r-q)}{N}\right)^{1-N}q^{-1}\mathcal{E}_q^-\mathbb{1}_{(0,+\infty)}\mathcal{E}_q^+G(x)dq.\tag{8.3}$$

We integrate by parts in (8.3) k times and obtain

$$V^{N-1}(x) = -\frac{N^k(N-k-2)!}{(N-2)!}\cdot\frac{1}{2\pi i(-T)^k} \times\int_{\mathcal{C}_N}\left(1+\frac{T(r-q)}{N}\right)^{1+k-N}\partial_q^k\left(q^{-1}\mathcal{E}_q^-\mathbb{1}_{(0,+\infty)}\mathcal{E}_q^+\right)G(x)dq. \tag{8.4}$$

Then we subtract (3.11) from (8.4) and divide by T/N:

$$\begin{aligned}\theta_{\mathrm{Carr};N}(T,x) = &-\frac{N^k(N-k-1)!}{(N-1)!}\cdot\frac{1}{2\pi i(-T)^k}\\ &\times\int_{\mathcal{C}_N}\left(\frac{N-1}{N-k-1}\left(1+\frac{T(r-q)}{N}\right)-1\right)\frac{N}{T}\\ &\times\left(1+\frac{T(r-q)}{N}\right)^{k-N}\partial_q^k\left(q^{-1}\mathcal{E}_q^-\mathbb{1}_{(0,+\infty)}\mathcal{E}_q^+\right)G(x)dq.\end{aligned}$$

For a fixed k, as $N\to\infty$,

$$\left(\frac{N-1}{N-k-1}\left(1+\frac{T(r-q)}{N}\right)-1\right)\frac{N}{T}\to q-r-\frac{k}{T};$$

therefore, the proof of convergence $\theta_{\mathrm{Carr};N}(T,x)\to\theta(T,x)$ as $N\to\infty$ is essentially the same as the proof of convergence $V^N(x)\to V(T,x)$.

The asymptotics of $\theta_{\mathrm{Carr};N}(T,x)$ as $x\downarrow 0$ is calculated as the asymptotics of $\theta(T,x)$, and the proof of convergence of the asymptotic coefficient in the former to the asymptotic coefficient of the latter is proved as convergence $\theta_{\mathrm{Carr};N}(T,x)\to\theta(T,x)$. We summarize the results in the following theorem.

Theorem 8.2. (a) *Let k be an integer ≥ 2. There exists $s>0$ such that*

$$\theta_{\mathrm{Carr};N}(T,x) = \kappa^\theta_{\mathrm{Carr};N}(T)x^{\nu_-}+O(x^{\nu_-+s}),\quad x\downarrow 0, \tag{8.5}$$

where

$$\begin{aligned}\kappa^\theta_{\mathrm{Carr};N}(T) = &-\frac{N^k(N-k-1)!}{(N-1)!}\cdot\frac{1}{2\pi i(-T)^k}\\ &\times\int_{\mathcal{C}_N}\left(\frac{N-1}{N-k-1}\left(1+\frac{T(r-q)}{N}\right)-1\right)\frac{N}{T}\\ &\times\left(1+\frac{T(r-q)}{N}\right)^{k-N}\partial_q^k\left(q^{-1}\phi^-_{q,\infty}\tilde{G}(q,0+)\right)dq\end{aligned} \tag{8.6}$$

for any integer $k\geq 2$.

(b) *$\kappa^\theta_{\mathrm{Carr};N}(T)\to\kappa^\theta(T)$ as $N\to\infty$, and the O-term in (8.5) converges to the O-term in (8.2) in the C^{ν_-+s}-norm.*

8.2. Delta. Clearly, it suffices to consider the log-delta $\tilde{\Delta} = \partial V/\partial x$. It can be shown that if k is sufficiently large, then the differentiation under the integral sign in (2.10) is admissible, and the interchange of the order of differentiation $\partial_x \partial_q^k = \partial_q^k \partial_x$ is admissible as well. Thus,

$$\tilde{\Delta}(T,x) = \frac{e^{-rT}}{2\pi i(-T)^k} \int_{\mathrm{Re}\, q=\sigma} e^{qT} \partial_q^k \left(q^{-1}\partial_x \mathcal{E}_q^- \mathbb{1}_{(0,+\infty)} \mathcal{E}_q^+ G(x)\right) dq. \tag{8.7}$$

Similarly, the log-delta in Carr's randomization approximation is

$$\begin{aligned}\tilde{\Delta}_{\mathrm{Carr};N}(T,x) &= -\frac{N^k(N-k-1)!}{(N-1)!} \cdot \frac{1}{2\pi i(-T)^k} \\ &\quad \times \int_{\mathcal{C}_N} \left(1 + \frac{T(r-q)}{N}\right)^{k-N} \partial_q^k \left(q^{-1}\partial_x \mathcal{E}_q^- \mathbb{1}_{(0,+\infty)} \mathcal{E}_q^+ G(x)\right) dq.\end{aligned} \tag{8.8}$$

Theorem 8.3. (a) *Let k be a sufficiently large integer. There exists $s > 0$ such that*

$$\tilde{\Delta}(T,x) = \nu_- \kappa_k(T) x^{\nu_- - 1} + R(x), \tag{8.9}$$

where $R(x) = O(x^{\nu_- - 1 + s})$, and

$$\tilde{\Delta}_{\mathrm{Carr};N}(T,x) = \nu_- \kappa_{\mathrm{Carr};N;k}(T) x^{\nu_- - 1} + R_N(x), \tag{8.10}$$

where $R_N(x) = O(x^{\nu_- - 1 + s})$.

(b) *$xR_N(x) \to xR(x)$ in the $C^{\nu_- + s}([0+,\infty))$-norm.*

Proof. To calculate the asymptotics of (8.7) and (8.8) and study the convergence of the latter to the former, we apply (7.6) with $k = 0$:

$$\mathbb{1}_{(0,+\infty)} \tilde{G} = \tilde{G}(q, 0+)(1+iD)^{-1}\delta + (1+iD)^{-1}\mathbb{1}_{(0,+\infty)}(1+iD)\tilde{G}, \tag{8.11}$$

which leads to

$$\begin{aligned}\partial_x \mathcal{E}_q^- \mathbb{1}_{(0,+\infty)} \mathcal{E}_q^+ G(x) &= \phi_q^-(D) iD(1+iD)^{-1}[\tilde{G}(q,0+)\delta(x) + \mathbb{1}_{(0,+\infty)}(1+iD)\tilde{G}(q,x)] \\ &= F_0(q,x) + F_1(q,x) - F_2(q,x),\end{aligned}$$

where $F_0(q,x) = \tilde{G}(q,0+)\phi_q^-(D)\delta(x)$, $F_1(q,x) = \phi_q^-(D)\mathbb{1}_{(0,+\infty)}(1+iD)\tilde{G}(q,x)$, and

$$\begin{aligned}F_2(q,x) &= \phi_q^-(D)(1+iD)^{-1}[\tilde{G}(q,0+)\delta(x) + \mathbb{1}_{(0,+\infty)}(1+iD)\tilde{G}(q,x)] \\ &= \mathcal{E}_q^- \mathbb{1}_{(0,+\infty)} \mathcal{E}_q^+ G(x)\end{aligned}$$

is the expression in the formulas for the price of the option and Carr's randomization approximation V^N to the price. Denote by

$$\tilde{\Delta}(T,x) = \tilde{\Delta}_0(T,x) + \tilde{\Delta}_1(T,x) - \tilde{\Delta}_2(T,x)$$

and

$$\tilde{\Delta}_{\mathrm{Carr};N}(T,x) = \tilde{\Delta}_{\mathrm{Carr};N;0}(T,x) + \tilde{\Delta}_{\mathrm{Carr};N;1}(T,x) - \tilde{\Delta}_{\mathrm{Carr};N;2}(T,x)$$

the corresponding decompositions of the log-delta and the log-delta Carr's randomization approximation. Both $\tilde\Delta_2(T,x) = V(T,x)$ and $\tilde\Delta_{\mathrm{Carr};N;2}(T,x) = V^N(x)$ are of order $O(x^{\nu_-})$, and the latter converges to the former as stated in section 7.4.

In the case $\nu > 1$ that we consider, $\nu_+ > 0$, and $F_1(q,x) \in \mathring{H}^{\nu_-+\nu_+-1/2-\epsilon}(\mathbb{R}) = \mathring{H}^{\nu-1/2-\epsilon}(\mathbb{R})$ for any $\epsilon > 0$. (If $\nu_+ = 0$, then, taking into account the representation $\phi^+_q(D) = \phi^+_{q,\infty} + R^+(q,D)$, where $R^+(q,D)$ is a PDO of negative order, we find that $F_1(q,x) \in \mathring{H}^s(\mathbb{R})$ for some $s > -1/2$.) Moreover, for any $k \in \mathbb{Z}_+$, there exists $C_{k,\epsilon}$ such that

$$||\partial_q^k F_1(q,\cdot)||_{\nu-1/2-\epsilon} \le C_{k,\epsilon}|q|^{1-k}$$

uniformly w.r.t. q in a half-plane $\mathrm{Re}\, q \ge \sigma$. We conclude that $\tilde\Delta_1(T,x)$ and $\tilde\Delta_{\mathrm{Carr};N;1}(T,x)$ are of class $\mathring{C}^{\nu-1-\epsilon}(\mathbb{R}_+)$, for any $\epsilon > 0$, and the latter converges to the former in the $\mathring{C}^{\nu-1-\epsilon}(\mathbb{R}_+)$-norm.[6] Furthermore, $\tilde\Delta_1(T,x) = O(x^{\nu-1-\epsilon})$ as $x \downarrow 0$, which can be included in the O-term of the asymptotics in view of the study of $F_0(q,x)$ below. Integrating by parts in the Fourier inversion formula for $F_1(q,\cdot)$, as we will do in a moment for $F_0(q,\cdot)$, it is possible to obtain a better characterization of convergence: $x\tilde\Delta_1(T,x)$ and $x\tilde\Delta_{\mathrm{Carr};N;1}(T,x)$ are of class $C^{\nu-\epsilon}(\mathbb{R}_+)$, for any $\epsilon > 0$, and the latter converges to the former in the $C^{\nu-\epsilon}(\mathbb{R}_+)$-norm.[7]

For $x > 0$,

$$\phi^-_q(D)\delta(x) = \frac{1}{2\pi}\int_{\mathrm{Im}\,\xi=\omega_+} e^{ix\xi}\phi^-_q(\xi)d\xi = \frac{1}{2\pi(-ix)}\int_{\mathrm{Im}\,\xi=\omega_+} e^{ix\xi}\partial_\xi\phi^-_q(\xi)d\xi,$$

where

$$\begin{aligned}\partial_\xi\phi^-_q(\xi) &= \partial_\xi\left((1+i\xi(d/q)^{1/\nu})^{-\nu_-}\exp[\hat I^-(q,0) - \hat I^-(q,\xi)]\right)\\ &= -[i\nu_-((d/q)^{1/\nu}+i\xi)^{-1} + \partial_\xi\hat I^-(q,\xi)]\phi^-_q(\xi).\end{aligned}$$

Introduce

$$\begin{aligned}F_{00}(q,x) &= \tilde G(q,0+)\nu_-((d/q)^{1/\nu}+iD)^{-1}\phi^-_q(D)\delta(x),\\ F_{01}(q,x) &= -i\tilde G(q,0+)(\partial_\xi\hat I^-)(q,D)\phi^-_q(D)\delta(x),\end{aligned}$$

and define $\tilde\Delta_{00}(T,x)$, $\tilde\Delta_{01}(T,x)$ and their Carr's randomization analogues using $F_{00}(q,x)$ and $F_{01}(q,x)$, in the same way as $\tilde\Delta_0(T,x)$ was defined using $F_0(q,x)$. Then

$$\tilde\Delta_0(T,x) = x^{-1}(\tilde\Delta_{00}(T,x) + \tilde\Delta_{01}(T,x)).$$

In section C.3 of Appendix C, it is proved that, for any $k \in \mathbb{Z}_+$ and any $\epsilon > 0$, there exists $C_{k,\epsilon}$ such that

$$|\partial_q^k\partial_\xi\hat I^-(q,\xi)| \le C_{k,\epsilon}|q|^{-k+(k-1)\epsilon}(1+|\xi|)^{-1}. \tag{8.12}$$

Bound (8.12) means that we can repeat the argument for the price and obtain that, for any $\epsilon > 0$, $\Delta_{01}(T,x) \in C^{\nu_--\epsilon}(\mathbb{R}_+)$, and $\Delta_{\mathrm{Carr};N;01}(T,x) \to \Delta_{01}(T,x)$ in the $C^{\nu_--\epsilon}(\mathbb{R}_+)$-norm.

[6] In the case of digital puts and calls, $(1+iD)G$ is a sum of a regular function and a multiple of the delta function supported at $\ln K$. This leads to a certain drop of regularity of the theta at $\ln K$, which can be quantified; even in these cases, the theta is regular on $(\ln H, \ln K)$ as stated. For simplicity, we restrict ourselves to the case of puts and calls and the case $G(x) \equiv 1$, where this problem does not emerge.

[7] If $\nu \le 1$, then we can prove that $x\tilde\Delta_1(T,x)$ and $x\tilde\Delta_{\mathrm{Carr};N;1}(T,x)$ are of class $C^{\bar\nu-\epsilon}(\mathbb{R}_+)$, for any $\epsilon > 0$, and the latter converges to the former in the $C^{\bar\nu-\epsilon}(\mathbb{R}_+)$-norm.

The asymptotics of $\tilde{\Delta}_{00}(T,x)$ can be calculated and the convergence of $\tilde{\Delta}_{\mathrm{Carr};N;00}(T,x)$ proved exactly as for the prices, and the evident result is

$$\tilde{\Delta}_{00}(T,x) = \nu_-\kappa(T)x^{\nu_-} + O(x^{\nu_-+s})$$

for some $s>0$. Finally, summing up all the terms considered above, we derive (8.9). The remaining results are proved similarly. ■

Remark 8.4. The asymptotic formula (8.9) can be obtained by the formal differentiation of the asymptotic formula (7.11) for the price. However, if $\nu_-=0$, then $F_{00}\equiv 0$, and the result—only an upper bound for $\tilde{\Delta}$—is too weak. Studying the asymptotics of $F_{01}\equiv 0$, it is possible to prove that if $\nu_-=0$ (this is the case of processes of order $\nu<1$ with the drift pointing from the barrier), then the asymptotics of $\tilde{\Delta}$ and $\tilde{\Delta}_{\mathrm{Carr};N}(T,x)$ are of the form

$$\tilde{\Delta}(T,x) = \tilde{\kappa}(T)x^{\nu-1} + R(x), \tag{8.13}$$

$$\tilde{\Delta}_{\mathrm{Carr};N}(T,x) = \tilde{\kappa}_N(T)x^{\nu-1} + R_N(x), \tag{8.14}$$

where $R(x)=O(x^{\nu-1+s})$ and $R_N(x)=O(x^{\nu-1+s})$ for some $s>0$. Moreover, $\tilde{\kappa}_N(T)\to\tilde{\kappa}(T)$ and $xR_N(x)\to xR(x)$ in the $C^{\nu+s}$-norm.

8.3. Gamma. Since the convergence and asymptotics of the price and log-delta have already been studied, it suffices to consider the log-gamma $\tilde{\Gamma}=\partial^2V/\partial^2x$. It can be shown that if k is sufficiently large, then the differentiation under the integral sign in (2.10) is admissible, and the interchange of the order of differentiation $\partial_x^2\partial_q^k=\partial_q^k\partial_x^2$ is admissible as well. Thus,

$$\tilde{\Gamma}(T,x) = \frac{e^{-rT}}{2\pi i(-T)^k}\int_{\mathrm{Re}\,q=\sigma} e^{qT}\partial_q^k\left(q^{-1}\partial_x^2\mathcal{E}_q^-\mathbb{1}_{(0,+\infty)}\mathcal{E}_q^+G(x)\right)dq. \tag{8.15}$$

The study of $\tilde{\Gamma}$ is very similar to the study of $\tilde{\Delta}$. The main changes are as follows. We have to use an analogue of (8.11) with an additional term:

$$\begin{aligned}\mathbb{1}_{(0,+\infty)}\mathcal{E}_q^+G &= \tilde{G}(q,0+)(1+iD)^{-1}\delta\\ &\quad+(\tilde{G}(q,0+)+(\partial_x\tilde{G})(q,0+))(1+iD)^{-2}\delta\\ &\quad+(1+iD)^{-2}\mathbb{1}_{(0,+\infty)}\mathcal{E}_q^+(1+iD)^2G.\end{aligned}\tag{8.16}$$

In the case of digital puts and calls, $(1+iD)^2G$ is a sum of a regular function and a linear combination of the delta function and its derivative supported at $\ln K$; in the case of the standard calls and puts, the derivative of the delta function at $\ln K$ does not emerge. This leads to a certain drop of regularity of the gamma at $\ln K$, which can be quantified. As global statements, the statements below are valid on $(\ln H,\ln K)$, and, naturally, any loss of regularity at $\ln K>0=\ln H$ does not influence the leading term of asymptotics as $x\downarrow 0$. The number of terms, which needs to be studied separately, is larger than in the study of $\tilde{\Delta}$, and in the Fourier-inversion representation of some of those it may be necessary to integrate by parts twice. The result is the following theorem.

Theorem 8.5. (a) *Let $\nu_->0$, $\nu_-\neq 1$; then there exists $s>0$ such that, as $x\downarrow 0$,*

$$\tilde{\Gamma}(T,x) = \nu_-(\nu_--1)\kappa(T)x^{\nu_--2} + O(x^{\nu_--2+s}), \tag{8.17}$$

and $\tilde\Gamma_{\mathrm{Carr};N}(T,x)$ *converges to* $\tilde\Gamma(T,x)$ *in* $C^{\nu_-}([0+,\infty))$ *with weight* x^2*:*

$$x^2\tilde\Delta_{\mathrm{Carr};N}(T,x)\to x^2\tilde\Delta(T,x) \tag{8.18}$$

in the $C^{\nu_-}([0+,\infty))$*-norm.*

(b) *If* $\nu_-=0$*, then* $\tilde\Gamma(T,x)$ *is unbounded as* $x\downarrow 0$*, and the same is typically true if* $\nu_-=1$*.* From the list of model classes that we consider, the exceptions are Brownian motion, Kou's model, and the HEJD model, for which $\nu_-=1$ and $\Gamma(x)$ has the finite limit as $x\downarrow 0$.

9. Numerical examples. Consider the same example as in Figure 1. In Table 1, we give prices (rounded) of the down-and-out put option of maturity $T=1$, strike $K=3500$, and barrier $H=2100$ close to the barrier; the distances are shown in the $x=\ln(S/H)$ coordinate. In panel A, we show the benchmark prices calculated using Monte Carlo simulations and prices calculated using Carr's randomization approximations with moderate $N=10,20,30,40$. After that, we calculate the prices using three versions of Richardson extrapolations; we see that all three agree very well. In panel B, we show prices normalized by $x^{-\nu_-}$, that is, $V(x)x^{-\nu_-}$ instead of $V(x)$, and, in panel C, we show the relative errors with respect to the Monte Carlo prices. We observe that all three versions of Richardson extrapolation produce small errors of the same order of magnitude, which suggests that a significant part of the errors are errors of Monte Carlo simulations. The discretization and truncation errors in realizations of the EPV operators are a less likely source of errors. Indeed, the former errors must be sensitive to the choice of N and hence of q; but they are almost identical.

In Table 2, we produce similar results for the log-delta, the normalization being done with $x^{1-\nu_-}$ instead of $x^{-\nu_-}$. We observe that all three versions of Richardson extrapolation produce small errors of the same order of magnitude, which suggests that the errors are mostly the errors of Carr's randomization approximation with a too large $N=160$ used as the benchmark. This example, as do many others, indicates that better results can be obtained with a moderate number of time steps and Richardson extrapolation than with a very large number of time steps.

10. Conclusion. In the paper, we derived the leading terms of asymptotics of the prices and sensitivities of down-and-out barrier options and the corresponding Carr's randomization approximations and showed that the asymptotic coefficients in the latter converge to the asymptotic coefficients of the former as $N\to\infty$. We also proved the convergence of prices and sensitivities in appropriate Hölder norms. The result is proved for wide classes of options including puts, calls, and digital options, as well as wide classes of Lévy processes.

Finally, we justified Richardson extrapolation of arbitrary order and demonstrated that it can be more accurate than Carr's randomization with a very large number of time steps.

Appendix A. Elements of the theory of generalized functions [27, 15]. The Sobolev space $H^s(\mathbb{R})$ consists of generalized functions with the finite norm

$$||u||_s=\left(\int_{\mathbb{R}}|\hat u(\xi)|^2(1+|\xi|^2)^s d\xi\right)^{1/2},$$

where $\hat u$ is the Fourier transform of u. $\mathring{H}^s(\mathbb{R}_+)$ denotes the subspace of functions from $H^s(\mathbb{R}_+)$ supported on $[0,+\infty)$. For any $s\in(-1/2,1/2)$, the multiplication-by-$\mathbb{1}_{(0,+\infty)}$ operator extends to a bounded operator $\mathbb{1}_{(0,+\infty)}:H^s(\mathbb{R})\to\mathring{H}^s(\mathbb{R}_+)$.

Table 1

Convergence of Carr's randomization approximation in the NIG model. Price of a down-and-out put option.

A. Prices (in units of 100, rounded)

x	$N=10$	$N=20$	$N=30$	$N=40$	(a)	(b)	(c)	MC
0.0025	2.0040	1.9579	1.9428	1.9352	1.9125	1.9129	1.9127	1.9163
0.005	2.4442	2.3884	2.3701	2.3609	2.3335	2.3339	2.3338	2.3398
0.01	3.0061	2.9387	2.9165	2.9055	2.8724	2.8728	2.8728	2.8732
0.02	3.7280	3.6483	3.6220	3.6090	3.5696	3.5700	3.5699	3.5754
0.04	4.6231	4.5351	4.5058	4.4912	4.4473	4.447	4.4474	4.4570
0.06	5.1859	5.0988	5.0697	5.0551	5.0114	5.0112	5.0112	5.0182
0.08	5.5511	5.4697	5.4422	5.4284	5.3871	5.3868	5.3868	5.3877

B. Prices in units of 100, normalized by $x^{-\nu_-}$ and rounded.

x	$N=10$	$N=20$	$N=30$	$N=40$	(a)	(b)	(c)	MC
0.0025	9.6647	9.4424	9.3694	9.3329	9.2234	9.2250	9.2245	9.2415
0.005	9.8258	9.6014	9.5278	9.4912	9.3809	9.3825	9.3822	9.463
0.01	10.0736	9.8477	9.7736	9.7367	9.6257	9.6271	9.6269	9.6284
0.02	10.414	10.191	10.1179	10.0814	9.9714	9.9725	9.9724	9.988
0.04	10.7652	10.5603	10.4921	10.4581	10.3560	10.3561	10.3560	10.3785
0.06	10.8554	10.6737	10.6128	10.5823	10.4908	10.4905	10.4904	10.505
0.08	10.7751	10.6170	10.5637	10.5369	10.4568	10.4561	10.4561	10.458

C. Relative errors (rounded) w.r.t. Monte Carlo

x	$N=10$	$N=20$	$N=30$	$N=40$	(a)	(b)	(c)
0.0025	0.046	0.022	0.014	0.0010	−0.0002	−0.0002	−0.0002
0.005	0.045	0.020	0.013	0.009	−0.0027	−0.0025	−0.0026
0.01	0.046	0.023	0.015	0.011	−0.00028	−0.00015	−0.00015
0.02	0.043	0.020	0.013	0.009	−0.0016	−0.0015	−0.0015
0.04	0.037	0.018	0.011	0.008	−0.0022	−0.0022	−0.0022
0.06	0.033	0.016	0.010	0.007	−0.0014	−0.0014	−0.0014
0.08	0.030	0.015	0.010	0.008	−0.00011	−0.00018	−0.00018

NIG parameters: $\alpha = 8.858$, $\beta = -5.808$, $\delta = 0.174$, $\mu \approx 0.16074$; $\nu_- \approx 0.262593798$
Option parameters: $K = 3500$, $H = 2100$, $r = 0.03$, $T = 1$
Monte Carlo parameters: # of trajectories 1,000,000, # of time steps 20000; st. dev. varies from 0.00408 to 0.00508
Parameters of the realization of Carr's randomization in [11]: $M_2 = 2$, $M_3 = 16$, $m = 8$, $N = \#$ number of time steps; $\Delta = 0.00025$
(a) Richardson extrapolation $V(T, x) = 2V^{40}(x) - V^{20}(x)$
(b) Richardson extrapolation $V(T, x) = 0.5V^{10}(x) - 4V^{20}(x) + 4.5V^{30}(x)$
(c) Richardson extrapolation $V(T, x) = (1/3)V^{10}(x) - 2V^{20}(x) + (8/3)V^{40}(x)$

If a is a measurable function, which admits an estimate

$$|a(\xi)| \le C(1 + |\xi|^2)^{m/2}, \quad \xi \in \mathbb{R}, \tag{A.1}$$

then the PDO $a(D) : H^s(\mathbb{R}) \to H^{s-m}(\mathbb{R})$ is a bounded operator, with the norm bounded by C, and if (A.1) holds for ξ in the half-plane $\operatorname{Im} \xi \le 0$, then $a(D)$ maps $\mathring{H}^s(\mathbb{R}_+)$ to $\mathring{H}^{s-m}(\mathbb{R}_+)$.

The Sobolev embedding theorem (in one dimension) states that if $s > 1/2$, then, for any $\epsilon > 0$, $H^s(\mathbb{R})$ is continuously embedded in the Hölder space $C^{s-1/2-\epsilon}(\mathbb{R})$. Denoting by $\mathring{C}^s(\mathbb{R}_+)$

Table 2

Convergence of Carr's randomization approximation in the NIG model. Log-delta of a down-and-out put option.

A. Log-delta (in units of 10^4, rounded)

x	$N = 10$	$N = 20$	$N = 30$	$N = 40$	(a)	(b)	(c)	$N = 160$
0.0025	2.1837	2.1348	2.1189	2.1111	2.0875	2.0878	2.0880	2.0957
0.005	1.4053	1.3750	1.3651	1.3602	1.3454	1.3455	1.3456	1.3499
0.01	0.9100	0.8924	0.8865	0.8836	0.8749	0.8749	0.8749	0.8773
0.02	0.5834	0.5751	0.5723	0.5708	0.5666	0.5664	0.5665	0.5677
0.04	0.3454	0.3443	0.3438	0.3435	0.3428	0.3426	0.3426	0.3429
0.06	0.2251	0.2270	0.2276	0.2278	0.2286	0.2285	0.2285	0.2284
0.08	0.1437	0.1470	0.1480	0.1485	0.1500	0.1499	0.1499	0.1496

B. Log-delta in units of 10^2, normalized by $x^{1-\nu_-}$ and rounded.

x	$N = 10$	$N = 20$	$N = 30$	$N = 40$	(a)	(b)	(c)	$N = 160$
0.0025	2.6328	2.5738	2.5547	2.5453	2.5168	2.5172	2.5175	2.5267
0.005	2.8247	2.7638	2.7439	2.7340	2.7044	2.7046	2.7047	2.7134
0.01	3.0494	2.9905	2.9709	2.9611	2.9318	2.9318	2.9319	2.9400
0.02	3.2595	3.2132	3.1973	3.1892	3.1653	3.1647	3.1647	3.1714
0.04	3.2167	3.2066	3.2020	3.1996	3.1925	3.1911	3.1912	3.1937
0.06	2.8273	2.8516	2.8585	2.8617	2.8718	2.8703	2.8703	2.8686
0.08	2.2314	2.2820	2.2979	2.3057	2.3294	2.3284	2.3284	2.3230

C. Relative errors (rounded) w.r.t. the benchmark $N = 160$

x	$N = 10$	$N = 20$	$N = 30$	$N = 40$	(a)	(b)	(c)
0.0025	0.04	0.02	0.01	0.007	−0.0004	−0.0004	−0.0004
0.005	0.04	0.02	0.01	0.008	−0.003	−0.003	−0.003
0.01	0.04	0.02	0.01	0.007	−0.003	−0.003	−0.003
0.02	0.03	0.01	0.008	0.006	−0.0002	−0.0002	−0.0002
0.04	0.007	0.004	0.003	0.002	−0.0004	−0.0008	−0.0008
0.06	−0.014	−0.006	−0.004	−0.002	0.0011	0.0006	0.0006
0.08	−0.037	−0.018	−0.011	−0.007	0.0028	0.0023	0.0023

NIG parameters: $\alpha = 8.858$, $\beta = -5.808$, $\delta = 0.174$, $\mu \approx 0.16074$; $\nu_- \approx 0.262593798$
Option parameters: $K = 3500$, $H = 2100$, $r = 0.03$, $T = 1$
Parameters of the realization of Carr's randomization in [11]: $M_2 = 2$, $M_3 = 16$, $m = 8$, $N = \#$ number of time steps; $\Delta = 0.00025$
(a) Richardson extrapolation $V(T, x) = 2V^{40}(x) - V^{20}(x)$
(b) Richardson extrapolation $V(T, x) = 0.5V^{10}(x) - 4V^{20}(x) + 4.5V^{30}(x)$
(c) Richardson extrapolation $V(T, x) = (1/3)V^{10}(x) - 2V^{20}(x) + (8/3)V^{40}(x)$

the subspace of $C^s(\mathbb{R})$ of functions vanishing on $(-\infty, 0]$, we have $\mathring{H}^s(\mathbb{R}) \subset \mathring{C}^{s-1/2-\epsilon}(\mathbb{R}_+)$ for any $s > 1/2$.

If $u \in H^s(\mathbb{R})$, and if $s > m - 1/2$, where m is a positive integer, then

$$\begin{aligned} \mathbb{1}_{(0,+\infty)}u &= u(0+)(1+iD)^{-1}\delta + ((1+iD)u)(0+)(1+iD)^{-2}\delta \\ &\quad \cdots + (1+iD)^{-m}\mathbb{1}_{(0,+\infty)}(1+iD)^m u \\ &= u(0+)(1+iD)^{-1}\delta + (u(0+) + u'(0+))(1+iD)^{-2}\delta \\ &\quad \cdots + (1+iD)^{-m}\mathbb{1}_{(0,+\infty)}(1+iD)^m u \end{aligned} \tag{A.2}$$

(see [27, Lem. 5.5], [15, Thm. 15.15]).

Appendix B. Digital puts and calls, and the case of strike $K < H$. If strike $K < H$, then only the down-and-out call option makes sense, and we may assume that $G(x) = e^x - K$ and calculate $\tilde{G}(q,x) = \phi_q^+(-i)e^x - K$ explicitly. The rest of the calculations remain the same. Similarly, in the case of the call option with strike $K < H$, $\tilde{G}(q,x) = 1$. In the case of digitals with strike $K > H$ $(= 1)$, (2.1) fails. We take $\chi \in C^\infty(\mathbb{R})$, $\chi(x) = 1$, $x < \ln K/2$, $\chi(x) = 0$, $x > \ln K/3$, and write (A.2) in the form

$$\begin{aligned}\mathbb{1}_{(0,+\infty)}\tilde{G}(q,x) &= \tilde{G}(q,0+)(1+iD)^{-1}\delta + \cdots + ((1+iD)^m\tilde{G})(q,0+)(1+iD)^{-m}\delta \\ &\quad + (1+iD)^{-m}\mathbb{1}_{(0,+\infty)}(1+iD)^m\chi(x)\tilde{G}(q,x) + (1-\chi(x))\tilde{G}(q,x).\end{aligned} \tag{B.1}$$

The last term vanishes in a neighborhood of zero; hence, it is irrelevant for the study of the leading term of asymptotics. For the standard classes of RLPEs, the characteristic exponent admits estimates of the form

$$|\psi^{(s)}(\xi)| \le C_s(1+|\xi|)^{\nu-s}, \quad s = 0,1,\ldots,$$

and the Wiener–Hopf factors admit similar estimates. Using the estimates for $\phi_q^+(\xi)$ and integrating by parts in the oscillatory integral that defines $\chi(x)\tilde{G}(q,x)$ (composition of the Fourier transform, multiplication-by-$\phi_q^+(\xi)$, and the inverse Fourier transform applied to $G(x)$), one easily obtains that $\chi(x)\tilde{G}(q,x)$ is of class $C^\infty(\mathbb{R})$ in x, with the derivatives w.r.t. q decaying faster with each differentiation.

Hence, in (B.1), all terms but the first one do not influence the leading term of asymptotics. These terms do not influence the second term as well if the second term is of order x^{ν_-+s}, where $s \in (0,1)$. If the second term is of order x^{ν_-+1}, then the second term in (B.1) must be taken into account.

Appendix C. Technical proofs.

C.1. Proof of Lemma 6.6. In view of (6.11), bounds (6.14) and (6.16) are immediate from (6.13) and (6.15); therefore it suffices to prove the latter pair.

C.1.1. Proof of (6.13). We consider cases $|\xi| \le |q|^{1/\nu}$ and $|\xi| \ge |q|^{1/\nu}$. Both cases being similar, we consider only the case $|\xi| \le |q|^{1/\nu}$. We separate the line of integration into four parts by conditions $|\eta| \le |\xi|/2$, $|\xi|/2 \le |\eta| \le 2|\xi|$, $2|\xi| \le |\eta| \le |q|^{1/\nu}$, and $|\eta| \ge |q|^{1/\nu}$ and denote the corresponding integrals $\hat{I}_j^\pm(q,\xi)$, $j = 1,2,3,4$ (depending on (q,ξ), some of the intervals may be empty, and then the corresponding integrals equal 0). To obtain the bounds for $\hat{I}_j^\pm(q,\xi)$, $j = 1,3$ (respectively, $j = 4$), we use the first line (respectively, second line) in (6.9):

$$|\hat{I}_1^\pm(q,\xi)| \le C\int_1^{|\xi|/2} \frac{|\eta|^\nu|q|^{-1}}{1+|\xi|}d\eta \le C_1|\xi|^\nu|q|^{-1} \le C_1,$$

$$|\hat{I}_3^\pm(q,\xi)| \le C\int_{2|\xi|}^{|q|^{1/\nu}} \frac{|\eta|^\nu|q|^{-1}}{1+|\eta|}d\eta \le C_1|q|^{\nu/\nu}|q|^{-1} = C_1,$$

$$|\hat{I}_4^\pm(q,\xi)| \le C\int_{|q|^{1/\nu}}^{+\infty} |q||\eta|^{-\nu-1}d\eta \le C_1|q||q|^{-\nu/\nu} = C_1.$$

To estimate $\hat{I}_2^{\pm}$, we use (4.2) and the mean value theorem to represent $\ln\Psi(q,\eta)$ for $|\xi|/2 \le |\eta| \le 2|\xi|$ in the form

$$\ln\Psi(q,\eta) = \ln\Psi(q,\xi) + O(|\xi|^{-\rho})$$

for some $\rho > 0$ and conclude that

$$|\hat{I}_2^{\pm}(q,\xi)| \le |\ln\Psi(q,\xi)| \left|\ln\frac{2|\xi|-\xi}{|\xi|/2-\xi}\right| + O(|\xi|^{-\rho}\ln(1+|\xi|)),$$

which is bounded.

C.1.2. Proof of (6.15). We use (6.10):

$$|\partial_q^k \hat{I}^{\pm}(q,\xi)| \le C|q|^{-k+\epsilon}\int_1^{+\infty}\frac{d\eta}{|q|^{\epsilon}(1+|\eta|/|q|^{1/\nu})(1+|\xi-\eta|)};$$

notice that the last integral is uniformly bounded as a function of ξ (for a proof, divide $[1,+\infty)$ into three parts by inequalities $|\eta| \le |\xi|/2$, $|\xi|/2 \le |\eta| \le 2|\xi|$, $2|\xi| \le |\eta|$).

C.2. Proof of Lemma 6.7. Using (6.9),

$$\begin{aligned}|\hat{I}^{\pm}(q,\xi)| &\le C\left(\int_0^{|q|^{1/\nu}}\frac{|\eta|^{\nu}|q|^{-1}d\eta}{1+|\xi|} + \int_{|q|^{1/\nu}}^{|\xi|/2}\frac{|q||\eta|^{-\nu}d\eta}{1+|\xi|}\right.\\ &\qquad\left. + \int_{|\xi|/2}^{2|\xi|}\frac{|q||\xi|^{-\nu}d\eta}{1+|\xi-\eta|} + \int_{2|\xi|}^{+\infty}|q||\eta|^{-\nu-1}d\eta\right)\\ &\le C_1(|\xi|^{-1}|q|^{1/\nu} + |\xi|^{-1}|q|^{1/\nu} + |q||\xi|^{-1}\ln|\xi| + |q||\xi|^{-1}),\end{aligned}$$

which gives (6.17) with $k = 0$ because $s < \nu$ and $|q| \le |\xi|^s$. To obtain (6.17) with $k \ge 0$, we use (6.10) and argue similarly.

Representation (6.19)–(6.20) in the region $|q| \le |\xi|^s$ is immediate from (6.17).

C.3. Proof of (8.12). To simplify the proof, we assume that the characteristic exponent admits the analytic continuation not only into a strip around the real axis but into a union of two open sectors with the real half-axis as the axes of symmetry as well and obeys there the same estimates as in the strip.[8] Then there exists $\tilde\theta \in (0,\pi/2)$ such that function $\Psi(q,\eta)$ admits the analytic continuation w.r.t. η into the set bounded from below by the line $\mathrm{Im}\,\eta = \omega_-$ and from above by the contour $\mathcal{L}_{\omega_+,\tilde\theta} = \{\eta = \rho e^{i\varphi} \mid \rho \ge 0, \varphi = \pi/2 \pm \tilde\theta\}$, and into the half-plane $\mathrm{Re}\,q \ge \sigma$ w.r.t. q, and admits there estimates (6.9) and (6.10). Then

$$\hat{I}^{-}(q,\xi) = -\frac{1}{2\pi i}\int_{\mathcal{L}_{\omega_+,\tilde\theta}}\frac{\ln\Psi(q,\eta)}{\eta-\xi}d\eta,$$

and, for ξ in the half-plane $\mathrm{Im}\,\xi \le \omega_+'$ $(< \omega_+)$,

$$\partial_\xi \hat{I}^{-}(q,\xi) = \frac{1}{2\pi i}\int_{\mathcal{L}_{\omega_+,\tilde\theta}}\frac{\ln\Psi(q,\eta)}{(\eta-\xi)^2}d\eta$$

[8] All model classes of RLPEs satisfy this condition.

admits a bound

$$\begin{aligned}|\partial_\xi \hat{I}^-(q,\xi)| &\le C\int_1^\infty \frac{\min\{|q||\eta|^{-\nu}, |\eta|^\nu|q|^{-1}\}}{|\eta|^2+|\xi|^2}d\eta\\ &\le C\left[\int_1^{|q|^{1/\nu}} \frac{|q|^{-1}\eta^\nu}{|\eta|^2+|\xi|^2}d\eta + \int_{|q|^{1/\nu}}^\infty \frac{|q|\eta^{-\nu}}{|\eta|^2+|\xi|^2}d\eta\right]\end{aligned}$$

(see (6.9)). Consider two cases: $|\xi|^\nu \le |q|$ and $|\xi|^\nu \ge |q|$. In the first case,

$$\begin{aligned}|\partial_\xi \hat{I}^-(q,\xi)| &\le C\left[\int_1^{|\xi|} \frac{|q|^{-1}\eta^\nu}{|\eta|^2+|\xi|^2}d\eta + \int_{|\xi|}^{|q|^{1/\nu}} \frac{|q|^{-1}\eta^\nu}{|\eta|^2+|\xi|^2}d\eta + \int_{|q|^{1/\nu}}^\infty \frac{|q|\eta^{-\nu}}{|\eta|^2+|\xi|^2}d\eta\right]\\ &\le C_1(|q|^{-1}|\xi|^{-2+\nu+1} + |q|^{-1}|\xi|^{\nu-1} + |q|^{1-1-1/\nu})\\ &\le C_2(|q|^{-1}|\xi|^{\nu-1} + |q|^{-1/\nu})\\ &\le C_3|\xi|^{-1};\end{aligned}$$

in the second case,

$$\begin{aligned}|\partial_\xi \hat{I}^-(q,\xi)| &\le C\left[\int_1^{|q|^{1/\nu}} \frac{|q|^{-1}\eta^\nu}{|\eta|^2+|\xi|^2}d\eta + \int_{|q|^{1/\nu}}^{|\xi|} \frac{|q|\eta^{-\nu}}{|\eta|^2+|\xi|^2}d\eta + \int_{|\xi|}^\infty \frac{|q|\eta^{-\nu}}{|\eta|^2+|\xi|^2}d\eta\right]\\ &\le C_1(|q|^{-1+1+1/\nu}|\xi|^{-2} + |q||\xi|^{-2+1-\nu} + |q||\xi|^{1-\nu})\\ &\le C_2|\xi|^{-1}\end{aligned}$$

as well. This proves (8.12) with $k=0$.

To obtain the bounds for $k\ge 1$, we use (6.10). If $|\xi|^\nu \le |q|$, then

$$\begin{aligned}|\partial_q^k\partial_\xi \hat{I}^-(q,\xi)| &\le C_k\left[\int_1^{|\xi|} |q|^{-k}|\xi|^{-2}d\eta + \int_{|\xi|}^{|q|^{1/\nu}} |q|^{-k}\eta^{-2}d\eta + \int_{|q|^{1/\nu}}^\infty |q|^{-k+1/\nu}\eta^{-3}d\eta\right]\\ &\le C_{1k}(|q|^{-k}|\xi|^{-1} + |q|^{-k}|\xi|^{-1} + |q|^{-k+1/\nu-2/\nu}) \le C_{2k}|q|^{-k}|\xi|^{-1},\end{aligned}$$

and if $|\xi|^\nu \ge |q|$, then, for any $\epsilon > 0$,

$$\begin{aligned}|\partial_q^k\partial_\xi \hat{I}^-(q,\xi)| &\le C_k\left[\int_1^{|q|^{1/\nu}} |q|^{-k}|\xi|^{-2}d\eta\right.\\ &\qquad\left. + \int_{|q|^{1/\nu}}^{|\xi|} |q|^{-k+1/\nu}|\xi|^{-2}\eta^{-1}d\eta + \int_{|\xi|}^\infty |q|^{-k+1/\nu}\eta^{-3}d\eta\right]\\ &\le C_{1k}(|q|^{-k+1/\nu}|\xi|^{-2} + |q|^{-k+1/\nu}|\xi|^{-2}\ln|\xi| + |q|^{-k+1/\nu}|\xi|^{-2})\\ &\le C_{k,\epsilon}|q|^{-k+\epsilon}|\xi|^{-1}.\end{aligned}$$

Acknowledgments. The author is grateful to the participants of the Financial and Insurance Mathematics seminar at ETH Zürich (November 13, 2008) for the suggestion to calculate the asymptotics of the price of barrier options near the barrier and demonstrate that the limit of the leading term of the asymptotics of the price in Carr's randomization approximation equals the leading term of the asymptotics of the price of barrier options near the barrier, and to the participants of the 6th Congress of the Bachelier Finance Society, Toronto, June 22–27, 2010) and of Mathematical Finance seminars at the University of Edinburgh (October 4, 2009) and the University of Chicago (November 6, 2009) for useful discussions about the paper. The author is especially grateful to the two anonymous referees for valuable suggestions. All remaining errors are those of the author.

REFERENCES

[1] S. Asmussen, F. Avram, and M.R. Pistorius, *Russian and American put options under exponential phase-type Lévy models*, Stochastic Process. Appl., 109 (2004), pp. 79–111.

[2] O.E. Barndorff-Nielsen, *Processes of normal inverse Gaussian type*, Finance Stoch., 2 (1998), pp. 41–68.

[3] O.E. Barndorff-Nielsen and S. Levendorskiĭ, *Feller processes of normal inverse Gaussian type*, Quant. Finance, 1 (2001), pp. 318–331.

[4] J. Bertoin, *Lévy Processes*, Cambridge Tracts in Math. 121, Cambridge University Press, Cambridge, UK, 1996.

[5] F. Black and M. Scholes, *The pricing of options and corporate liabilities*, J. Polit. Econ., 81 (1973), pp. 637–659.

[6] B. Bouchard, N. El Karoui, and N. Touzi, *Maturity randomization for stochastic control problems*, Ann. Appl. Probab., 15 (2005), pp. 2575–2605.

[7] M. Boyarchenko, *Carr's Randomization for Finite-Lived Barrier Options: Proof of Convergence*, working paper; available online from http://ssrn.com/abstract=1275666 (2008).

[8] M. Boyarchenko and S.I. Boyarchenko, *User's Guide to Double Barrier Options. Part* I: *Kou's Model and Generalizations*, working paper; available online from http://ssrn.com/abstract=1272081 (2008).

[9] M. Boyarchenko, M. de Innocentis, and S.Z. Levendorskiĭ, *Prices of barrier and first-touch digital options in Lévy-driven models, near barrier*, Int. J. Theor. Appl. Finance, to appear (working paper version available online from http://ssrn.com/abstract=1514025).

[10] M. Boyarchenko and S.Z. Levendorskiĭ, *Refined and Enhanced Fast Fourier Transform Techniques, with an Application to the Pricing of Barrier Options*, working paper; available online from http://ssrn.com/abstract=1142833 (2008).

[11] M. Boyarchenko and S.Z. Levendorskiĭ, *Prices and sensitivities of barrier and first-touch digital options in Lévy-driven models*, Int. J. Theor. Appl. Finance, 12 (2009), pp. 1125–1170.

[12] M. Boyarchenko and S.Z. Levendorskiĭ, *Valuation of continuously monitored double barrier options and related securities*, Math. Finance, to appear (working paper version available online from http://ssrn.com/abstract=1227065).

[13] S.I. Boyarchenko and S.Z. Levendorskiĭ, *Option pricing for truncated Lévy processes*, Int. J. Theor. Appl. Finance, 3 (2000), pp. 549–552.

[14] S.I. Boyarchenko and S.Z. Levendorskiĭ, *Perpetual American options under Lévy processes*, SIAM J. Control Optim., 40 (2002), pp. 1663–1696.

[15] S.I. Boyarchenko and S.Z. Levendorksiĭ, *Non-Gaussian Merton-Black-Scholes Theory*, Adv. Ser. Stat. Sci. Appl. Probab. 9, World Scientific, River Edge, NJ, 2002.

[16] S.I. Boyarchenko and S.Z. Levendorskiĭ, *Barrier options and touch-and-out options under regular Lévy processes of exponential type*, Ann. Appl. Probab., 12 (2002), pp. 1261–1298.

[17] S.I. Boyarchenko and S.Z. Levendorksiĭ, *Irreversible Decisions under Uncertainty (Optimal Stopping Made Easy)*, Springer, Berlin, 2007.

[18] M. Broadie, P. Glasserman, and S.G. Kou, *A continuity correction for discrete barrier options*, Math. Finance, 7 (1997), pp. 325–348.

[19] P. Carr, *Two extensions to barrier option valuation*, Appl. Math. Finance, 2 (1995), pp. 173–209.

[20] P. Carr, *Randomization and the American put*, Rev. Financ. Stud., 11 (1998), pp. 597–626.

[21] P. Carr and J. Crosby, *A class of Lévy process models with almost exact calibration to both barrier and vanilla FX options*, Quant. Finance, 10 (2010), pp. 1–22.

[22] P. Carr and D. Faguet, *Valuing Finite-Lived Options as Perpetual*, working paper; available online from http://ssrn.com/abstract=706 (1996).

[23] P. Carr, H. Geman, D.B. Madan, and M. Yor, *The fine structure of asset returns: An empirical investigation*, J. Business, 75 (2002), pp. 305–332.

[24] R. Cont and P. Tankov, *Financial Modelling with Jump Processes*, Chapman & Hall/CRC Financ. Math. Ser., Chapman & Hall/CRC, Boca Raton, FL, 2004.

[25] 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.

[26] J. Crosby, N. Le Saux, and A. Mijatovič, *Approximating Lévy processes with a view to option pricing*, Int. J. Theor. Appl. Finance, 13 (2010), pp. 63–91.

[27] G.I. Eskin, *Boundary Value Problems for Elliptic Pseudodifferential Equations*, Transl. Math. Monogr. 52, American Mathematical Society, Providence, RI, 1981.

[28] M. Jeannin and M.R. Pistorius, *A transform approach to compute prices and Greeks of barrier options driven by a class of Lévy processes*, Quant. Finance, 10 (2010), pp. 629–644.

[29] I. Koponen, *Analytic approach to the problem of convergence of truncated Lévy flights towards the Gaussian stochastic process*, Phys. Rev. E (3), 52 (1995), pp. 1197–1199.

[30] S.G. Kou, *A jump-diffusion model for option pricing*, Management Sci., 48 (2002), pp. 1086–1101.

[31] O. Kudryavtsev and S.Z. Levendorskiĭ, *Fast and accurate pricing of barrier options under Lévy processes*, Finance Stoch., 13 (2009), pp. 531–562.

[32] A. Kuznetsov, *Wiener-Hopf factorization and distribution of extrema for a family of Lévy processes*, Ann. Appl. Probab., 20 (2009), pp. 1801–1830.

[33] S.Z. Levendorskiĭ, *Pricing of the American put under Lévy processes*, Int. J. Theor. Appl. Finance, 7 (2004), pp. 303–335.

[34] S.Z. Levendorskiĭ, *Early exercise boundary and option pricing in Lévy driven models*, Quant. Finance, 4 (2004), pp. 525–547.

[35] A. Lipton, *Path-Dependent Options on Assets with Jumps*, talk at Columbia-Jaffe Conference, New York, 2002; available online from http://www.math.columbia.edu/~lrb/columbia2002.pdf.

[36] A. Lipton, *Assets with jumps*, Risk, September (2002), pp. 149–153.

[37] D.B. Madan, P. Carr, and E.C. Chang, *The variance gamma process and option pricing*, Eur. Finance Rev., 2 (1998), pp. 79–105.

[38] D.B. Madan and F. Milne, *Option pricing with v.g. martingale components*, Math. Finance, 1 (1991), pp. 39–55.

[39] D.B. Madan and E. Seneta, *The variance gamma (v.g.) model for share market returns*, J. Business, 63 (1990), pp. 511–524.

[40] R.C. Merton, *Theory of rational option pricing*, Bell J. Econ. Management Sci., 4 (1973), pp. 141–183.

[41] L.C.G. Rogers and D. Williams, *Diffusions, Markov Processes, and Martingales. Volume* 1. *Foundations*, 2nd ed., John Wiley & Sons, Chichester, UK, 1994.

[42] K. Sato, *Lévy Processes and Infinitely Divisible Distributions*, Cambridge Stud. Adv. Math. 68, Cambridge University Press, Cambridge, UK, 1999.

[43] P. Tankov, *Lévy processes in finance and risk management*, Wilmott Magazine, September-October (2007).

[44] N. Wiener and E. Hopf, *Über eine Klasse singulärer Integralgleichungen*, S.-B. Preuss. Akad. Wiss.-Berlin, Phys.-Math. Kl., 30 (1931), pp. 696–706.

SIAM J. FINANCIAL MATH.
Vol. 2, pp. 112–140

Dynamic Hedging of Portfolio Credit Derivatives*

Rama Cont†‡ and Yu Hang Kan‡

Abstract. We compare the performance of various hedging strategies for index collateralized debt obligation (CDO) tranches across a variety of models and hedging methods during the recent credit crisis. Our empirical analysis shows evidence for market incompleteness: a large proportion of risk in the CDO tranches appears to be unhedgeable. We also show that, unlike what is commonly assumed, dynamic models do not necessarily perform better than static models, nor do high-dimensional bottom-up models perform better than simpler top-down models. When it comes to hedging, top-down and regression-based hedging with the index provide significantly better results during the credit crisis than bottom-up hedging with single-name credit default swap (CDS) contracts. Our empirical study also reveals that while significantly large moves—"jumps"—do occur in CDS, index, and tranche spreads, these jumps do not necessarily occur on the default dates of index constituents, an observation which shows the insufficiency of some recently proposed portfolio credit risk models.

Key words. hedging, credit default swaps, portfolio credit derivatives, index default swaps, collateralized debt obligations, portfolio credit risk models, default contagion, spread risk, sensitivity-based hedging, variance minimization

AMS subject classifications. 91G20, 91G70, 60G55

DOI. 10.1137/090750937

1. Introduction.

Static factor models, in particular the Gaussian copula model [18], have been widely used for hedging portfolio credit derivatives such as collateralized debt obligations (CDOs). In such models, the risk of a CDO tranche is characterized in terms of sensitivities to shifts in risk factors [10, 24]. Accordingly, hedging practices have typically been based on such measures of sensitivity. The most common hedging approach has been to "delta hedge" spread fluctuations using credit default swaps (CDSs).

However, the recent turmoil in credit derivatives markets shows that these commonly used hedging approaches are inefficient. One of the main criticisms has been the lack of well-defined dynamics for the risk factors in such static models, which prevents any model-based assessment of hedging strategies. In particular, delta hedging of spread risk is loosely justified using a Black–Scholes analogy which does not necessarily hold, and the corresponding hedge ratios, the spread-deltas, are in fact computed from a static model without spread risk. Indeed, delta hedging of spread risk is not deduced from any theory of derivative replication. Furthermore, delta hedging of spread risk ignores default risk and jumps in the spreads, which appeared to be critical for risk management during the difficult market environment in 2008. Although gamma hedging can improve performance slightly, it is not sufficient to solve these issues.

*Received by the editors February 27, 2009; accepted for publication (in revised form) November 4, 2010; published electronically February 1, 2011.
http://www.siam.org/journals/sifin/2/75093.html
†Laboratoire de Probabilités et Modèles Aléatoires, CNRS-UPMC, 4 Place Jussieu, Paris 75252, France (Rama.Cont@upmc.fr).
‡IEOR Department, Columbia University, New York, NY (yk2246@columbia.edu).

Finally, the common approach to price portfolio credit derivatives using copula-based models does not guarantee the absence of arbitrage. Cont, Deguest, and Kan [6] show that pricing CDO tranches based on linear interpolation of the base correlations in a one-factor Gaussian copula model can lead to static arbitrage.

Given the deficiencies of copula-based hedging methods, alternatives have been proposed to tackle the problem of hedging portfolio credit derivatives. Durand and Jouanin [10] describe common hedging practices for credit derivatives and correctly point out the inconsistency between most of the pricing models, where the sole risk is in the occurrence of defaults, and delta hedging strategies, where the trader seeks to protect his/her portfolio against small movements in CDS spreads. Bielecki, Jeanblanc, and Rutkowski [3] show that, in a bottom-up hazard process framework driven by a Brownian motion, perfect replication is possible by continuously trading a sufficient number of liquid CDS contracts. Bielecki, Crépey, and Jeanblanc [2] discuss hedging performance in bottom-up and top-down models using simulation but do not comment on the performance of such strategies in a real market setting.

Laurent, Cousin, and Fermanian [17] study hedging of synthetic CDO tranches in a local intensity framework without spread risk, and show that CDO tranches can then be replicated by a self-financing portfolio consisting of the index default swap and a risk-free bond. However, as we will show in section 3, spread fluctuation is a major source of risk even in the absence of defaults, so failure to incorporate spread risk can lead to unrealistic conclusions.

Using a more realistic approach which acknowledges market incompleteness and incorporates both spread risk and default contagion, Frey and Backhaus [14] observe significant differences between the sensitivity-based hedging strategies computed in the Gaussian copula framework and the dynamic hedging strategies derived in their setup. They also show that variance-minimization hedging provides a model-based endogenous interpolation between the hedging against spread risk and default risk.

Giesecke, Goldberg, and Ding [15] discuss an alternative hedging approach based on a self-exciting process for portfolio defaults and compare the hedging performance for equity CDO tranches in September 2008 with a Gaussian copula model.

The hedging methods in these studies approach the problem from different, often incompatible, standpoints, and a systematic comparison of the resulting hedge ratios and the subsequent hedging performance has not been done in a realistic setting with market data. Needless to say, in order for such a comparison to be meaningful, the models need to be calibrated to the same data set. The very feasibility of this calibration is a serious (computational) constraint which excludes many models discussed in the literature, leading us to focus on the class of tractable models.

Motivated by previous studies indicating the impact of model uncertainty on the pricing and hedging derivative instruments [5], our objective is to assess the performance of hedging strategies of index CDO tranches derived under various model assumptions. We compare the performance of different dynamic hedging strategies across a range of models including the Gaussian copula model, a multiname reduced-form model introduced by Duffie and Gârleanu [9], a Markovian portfolio default model [16], and a two-factor model with spread and default risk [1]. Strategies considered include delta hedging of spread risk, hedging of default risk, variance minimization (quadratic hedging), and regression-based hedging.

In particular we shall attempt to address some important questions which have been left unanswered by previous studies:

- How did various hedging strategies for CDOs perform during major credit events in 2008?
- Do complete market models provide the right insight for hedging credit derivatives?
- How good are delta hedging strategies for CDO tranches?
- Does gamma hedging improve hedging performance?
- Do hedge ratios based on jump-to-default fare better than sensitivity-based hedge ratios?
- Are dynamic models better for hedging than static models?
- Do hedging strategies using single-name CDSs perform better than hedges using the index?
- Are bottom-up models more suitable for hedging than top-down models?

This article is structured as follows. Section 2 describes the cash flow structure of credit default swaps, index default swaps, and index CDO tranches. Section 3 presents the dataset used for the empirical analysis and describes some important statistical features of the CDO and CDS markets. Section 4 introduces the models under consideration and discusses procedures used for parameter calibration. Section 5 discusses the hedging strategies under consideration. Section 6 compares the performance of different strategies for the hedging of index tranches in 2008. Section 7 summarizes our main findings and discusses some implications.

2. Credit derivatives. A credit derivative is a financial instrument whose payoff depends on the losses due to defaults of the reference obligors (debt instruments). A portfolio credit derivative is a credit derivative whose payoffs depend on default losses in a reference portfolio of obligors. We will consider here index credit derivatives, for which the underlying portfolio is an equally weighted portfolio, such as the CDX or iTraxx indices. Typically, the payoffs depend only on the aggregate loss due to defaults in the index, not on the identity of the defaulting firm.

Consider an equally weighted portfolio consisting of n obligors, and assume for simplicity a constant recovery rate R (typically assumed to be 40%) and deterministic interest rates. Let τ_i be the default time of obligor i. The portfolio loss (in fraction of total notional value) at time t is equal to

$$L_t = \frac{1-R}{n}\sum_{i=1}^{n} 1_{\tau_i \le t} = \frac{1-R}{n} N_t,$$

where N_t is the number of defaults by time t. The portfolio loss (L_t) is modeled as a stochastic process on a (filtered) probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{Q})$, where Ω is the set of market scenarios, $(\mathcal{F}_t)$ represents the flow of information, and $\mathbb{Q}$ is a risk-neutral probability measure representing the market pricing rule.

We will consider the three most commonly traded credit derivatives: credit default swaps (CDSs), index default swaps (index), and collateralized debt obligations (CDOs). All three derivatives are swap contracts between two parties, a protection buyer and a protection seller, whereby the protection buyer is compensated for the loss generated by the default of a reference obligor (CDS) or defaults from a pool of obligors (index and CDO). In return, the protection buyer pays a premium to the protection seller. A more detailed description of these products can be found in [7, 14, 20].

2.1. Credit default swaps. Consider a reference obligor i and its corresponding CDS contract initiated at time 0 with unit notional and payment dates $T_1 < T_2 < \cdots < T_M$, where $T = T_M$ is the maturity date. Assume that the default payments are made on the next payment date;[1] then if obligor i defaults between time T_{m-1} and T_m, the default payment at time T_m is equal to $1 - R$. On the other hand, if obligor i has not defaulted yet at time T_m, the protection seller will receive a premium payment $s_0^i(T_m - T_{m-1})$, where s_0^i is the CDS spread that has been determined at the inception.

The par CDS spread s_t^i quoted in the market at date t is defined as the value of the spread which sets the present values of the default leg and the premium leg equal. The mark-to-market value of a protection seller's position at time t is equal to the difference between the net present values of the two legs:

$$V_t^i = (s_0^i - s_t^i) \sum_{T_m > t} B(t, T_m)(T_m - T_{m-1})\mathbb{Q}(\tau_i > T_m | \mathcal{F}_t), \tag{1}$$

where $B(t, T_m)$ is the discount factor from time t to T_m. In what follows we will refer to the value of the protection seller's position as the mark-to-market value, and we will use $\mathbf{s}_t^{cds} = (s_t^1, \ldots, s_t^n)$ to denote the vector of constituent CDS spreads and $\mathbf{D}_t = (1_{\tau_1 \le t}, \ldots, 1_{\tau_n \le t})$ to denote the vector of default indicators at time t.

2.2. Index default swap. Index default swaps are now commonly traded on various credit indices such as iTraxx and CDX series which are equally weighted indices of CDSs. In an index default swap transaction initiated at time 0, a protection seller agrees to pay all default losses in the index in return for a fixed periodic spread s_0^{idx} paid on the total notional of obligors remaining in the index.

The index default swap par spread s_t^{idx} quoted in the market at time t is defined as the value of the spread which balances the present values of the default leg and the premium leg. The mark-to-market value of a protection seller's position at time t is equal to the difference between the two legs, which can be expressed as

$$V_t^{idx} = \left(s_0^{idx} - s_t^{idx}\right) \sum_{T_m > t} B(t, T_m)(T_m - T_{m-1}) E^{\mathbb{Q}}\left[1 - \frac{N_{T_m}}{n} \middle| \mathcal{F}_t\right]. \tag{2}$$

Here, we assume that the outstanding notional value is calculated at payment dates.[2]

2.3. Collateralized debt obligations. Consider a tranche defined by an interval $[a, b]$, $0 \le a < b \le 1$, for the portfolio loss normalized by the total notional value of the underlying portfolio. We call a (resp., b) the attachment (resp., detachment) point of the tranche. 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 return for a periodic spread $s_0^{[a,b]}$, which is determined at inception $t = 0$, on the remaining tranche notional value.

[1] The default payment is sometimes assumed to be made immediately after the default. Nevertheless, the choice of payment schedule has negligible effects on our analysis in this paper.

[2] A more precise valuation would consider the average outstanding notional value over the time period between the payment dates [4], but the approximation above has negligible effects on our analysis.

The par tranche spread $s_t^{[a,b]}$ quoted in the market at time t is defined as the spread which sets the present values of the default leg and the premium leg to be equal. The mark-to-market value of a protection seller's position (normalized by the total tranche notional value) at time t is equal to the difference between the two legs, which can be expressed as

$$V_t^{[a,b]} = \left(s_0^{[a,b]} - s_t^{[a,b]}\right) \sum_{T_m > t} \frac{B(t,T_m)}{b-a}(T_m - T_{m-1}) E^{\mathbb{Q}}\left[(b - L_{T_m})^+ - (a - L_{T_m})^+ \middle| \mathcal{F}_t\right]. \tag{3}$$

3. Data analysis. Our dataset contains the 5-year CDX North America Investment Grade Series 10 (CDX) index spreads; the standard tranche spreads with attachment/detachment points $0\%, 3\%, 7\%, 10\%, 15\%, 30\%, 100\%$; and the constituent 5-year CDS spreads, all obtained from Bloomberg. The time series runs from 25 March, 2008 until 25 September, 2008. Figure 1 illustrates the time series of the index, tranche $[10\%, 15\%]$, CDS of IBM and Disney Corp.

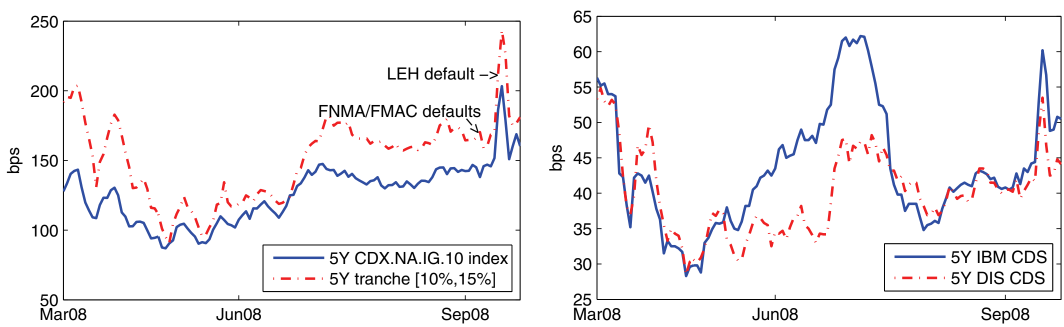


Figure 1. *Left: 5-year CDX.NA.IG.*10 *index and tranche* $[10\%, 15\%]$ *spreads. Right: 5-year CDS spreads of IBM and Disney Corp. (DIS).*

3.1. Comovements in CDSs and CDO tranches. A CDO hedging strategy should be based on a good understanding of the relation between the profit and loss (P&L) of the hedging instruments, namely the CDSs and the index, and that of the target instruments, the CDO tranches. Given that the P&L is driven mainly by the changes in spreads, this requires a correct representation of comovements in the credit spreads. Figure 2 shows the tranche $[10\%, 15\%]$ daily spread returns against the index and IBM CDS daily spread returns. The crosses and circles represent the data points where spread returns of the two credit derivatives move in the same and opposite directions, respectively. Here we use IBM CDS and the $[10\%, 15\%]$ tranche data for illustration, but similar results are obtained by looking at other constituent CDSs and tranches. From the figure, we can immediately observe two important properties:

1. CDS/index spreads tend to move *together* with the tranche spreads when the movements are *large*.
2. In many cases, CDS/index spreads and tranche spreads move in *opposite* directions, especially when the movements are *small*.

Large comovements of the spreads, or common jumps, can be explained by the exposures of the credit derivatives to common risk factors which undergo large movements. This phe-

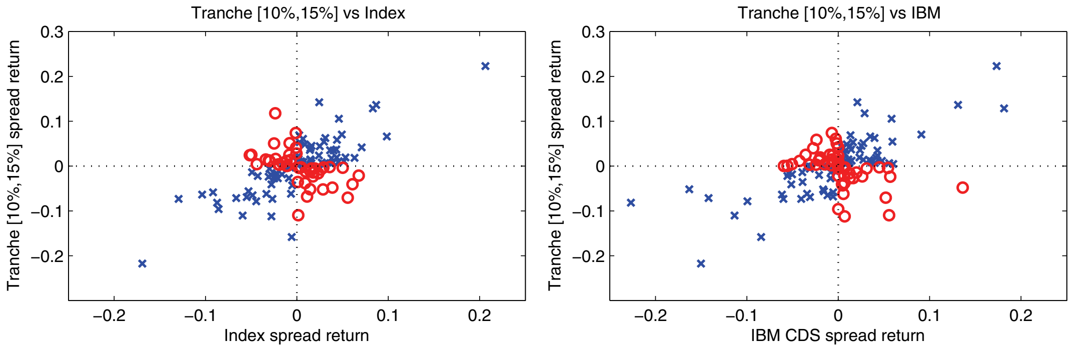


Figure 2. *Tranche* [10%, 15%] *daily spread returns versus index daily spread returns and IBM CDS daily spread returns. Crosses represent data points where spread returns have the same signs (movements in the same direction), and circles represent data points where spread returns have opposite signs (movements in opposite directions).*

Table 1

Conditional correlations between daily spread returns of the index and tranche [10%, 15%] *show evidence of a common heavy-tailed factor.*

Index spread return	Unconditional	$> 8\%$	$> 5\%$	$> 1\%$	$< -1\%$	$< -5\%$	$< -8\%$
Correlation	0.63	0.82	0.82	0.60	0.68	0.71	0.81
Observations	126	4	12	52	43	11	6

nomenon can be seen more clearly in Table 1, which shows that the correlation between the index and tranche spread returns increases substantially if we condition on larger observations.

From a hedging perspective, frequent opposite movements between the CDS/index and the tranche spreads can lead to serious problems, because most hedging strategies imply positive hedge ratios with respect to the CDSs or the index. When the values of the hedging instruments and the tranche move in opposite directions, those strategies may fail to reduce the exposure of the tranche positions or, more seriously, can substantially amplify the overall exposure. As we will see in our empirical study in section 6, this problem frequently arises in common hedging strategies.

3.2. Impact of defaults on credit spreads. Our sample period covers several important credit events: the takeover of Fannie Mae and Freddie Mac, which led to losses in the CDX, and the bankruptcy of Lehman Brothers, which led to a significant shock to the market.

During the sample period, Fannie Mae and Freddie Mac were taken over by the U.S. government on 7 September, 2008, which generated a credit event in the CDX reference portfolio. According to Bloomberg, the recovery rates of the 5-year senior CDS contracts of Fannie Mae and Freddie Mac were 92% and 94%, respectively, which will be used to determine the losses in our empirical study. On the other hand, although Lehman Brothers is *not* a reference obligor in the CDX, Figure 1 shows that there is considerable upward movement of the spreads on the next business day after Lehman Brothers announced bankruptcy.

Table 2 shows the daily spread returns on the next business day after Fannie Mae/Freddie Mac and Lehman Brothers credit events in units of sample standard deviation. Interestingly, we observe that the IBM, the index, and the super senior tranche [30%, 100%] spreads *decrease*

Table 2

Daily spread returns on the next business day after Fannie Mae/Freddie Mac (8 September, 2008) and Lehman Brothers (16 September, 2008) credit events, normalized by unconditional sample standard deviations.

	IBM	Index	0%–3%	3%–7%	7%–10%	10%–15%	15%–30%	30%–100%
8-Sep-08	−0.66	−1.10	0.16	0.05	0.04	0.44	0.02	−0.06
16-Sep-08	3.23	4.55	3.51	4.92	4.45	3.91	4.14	4.05

on the next business day after Fannie Mae/Freddie Mac were taken over. Although the spreads of other tranches do increase, the magnitudes of these changes are rather small, less than 0.5 standard deviations.

On the other hand, we *do* observe jumps in CDX spreads, but not necessarily on the dates corresponding to constituent defaults. The typical example is on 16 September, 2008 when Lehman Brothers filed for bankruptcy. Although Lehman Brothers is not a constituent of the CDX, the IBM, index, and tranche spreads increase by as much as 4.9 standard deviations, which are substantial upward moves and can be attributed to jumps.

These observations have two important implications. First, they show that jumps in the spreads are not necessarily tied to defaults in the underlying portfolio, as is the case in Markovian contagion models [1, 16] and self-exciting models [12, 15], where jumps occur only on portfolio default dates. Jumps may be caused by information external to the portfolio, such as macroeconomic events, of which the Lehman credit event is an example. Second, jump sizes at default dates appear to depend on the severity of the events, with lower recovery rate implying fewer or no upward jumps in the spreads. This suggests that models with constant jumps in the default intensity at each default are insufficient for capturing the impact of defaults on the spread movements: this impact should depend on the severity of loss in the given default, as suggested in [15]. We note that this may be difficult in practice, since recovery rates are usually not observable immediately after default and can be determined only after liquidation.

4. Models for portfolio credit derivatives. We will consider four different modeling approaches in our analysis: the one-factor Gaussian copula model [18], a bottom-up affine jump-diffusion model [9], a local intensity model [6, 7, 16, 21, 25], and a top-down bivariate spread-loss model [1].

4.1. Gaussian copula model. The one-factor Gaussian copula model [18] is a standard market reference for pricing CDO tranches, in which the default times are constructed as

$$\tau_i = F_i^{-1}(\Phi(\rho M_0 + \sqrt{1-\rho^2} M_i)),$$

where M_0, M_i are independent standard normal random variables, $\Phi(.)$ is the standard normal distribution function, F_i is the marginal distribution of τ_i, and ρ is a correlation parameter. The distribution function $F_i(.)$ is calibrated to the single-name CDS spreads by assuming a constant hazard rate.[3] Then, we fit one correlation to each tranche, which is a situation known as *compound correlations*. If multiple correlations give the same tranche spread, we will choose the smallest one.

[3] The hazard rate term structure is usually assumed to be piecewise constant. Since we consider only CDS with one maturity for each obligor, it reduces to a constant hazard rate.

There are two reasons why we consider compound correlations instead of the base correlations [19] for calibration. As noted by Morgan and Mortensen [22], we found that computing the spread-deltas while keeping base correlations fixed can lead to a negative sensitivity of a tranche with respect to a change in the CDS spreads. Therefore, even if the CDS and the tranche spreads move in the same direction, especially when the movement is large, the negative spread-deltas will have the wrong sign and give poor hedging results. Second, unlike the Black–Scholes implied volatility, which is in a one-to-one correspondence to the vanilla options prices, base correlations are not guaranteed to exist. For instance, we were not able to calibrate the base correlation for 15% strike on many of the dates in our sample.

In order to express the hedging positions in later sections, it is convenient to write the mark-to-market values of the credit derivatives as functions of the modeling variables. Given the CDS spreads $\mathbf{s}_t^{cds} = \mathbf{s}^{cds} = (s^1, \ldots, s^n)$, the default indicators $\mathbf{D}_t = \mathbf{D}$, and the set of compound correlations $\boldsymbol{\rho}_t = \boldsymbol{\rho}$, we write the mark-to-market values of CDS i, the index, and a tranche $[a, b]$ at time t computed under the Gaussian copula model as $V_{gc}^i(t, s^i)$, $V_{gc}^{idx}(t, \mathbf{s}^{cds}, \boldsymbol{\rho}, \mathbf{D})$, and $V_{gc}^{[a,b]}(t, \mathbf{s}^{cds}, \boldsymbol{\rho}, \mathbf{D})$, respectively.

4.2. Affine jump-diffusion model. Various dynamic reduced-form models have been proposed to overcome some of the shortcomings of static copula-based models. An example of such a model used in industry is the affine jump-diffusion model introduced by Duffie and Gârleanu [9]. In this model the default time τ_i of an obligor i is modeled as a random time with a stochastic intensity (λ_t^i) given by

$$\lambda_t^i = X_t^i + a^i X_t^0, \tag{4}$$

where the idiosyncratic risk factors $(X_t^1, \ldots, X_t^n)$ and the common (macro) risk factor (X_t^0) are independent affine jump-diffusion processes

$$dX_t^i = \kappa(\theta - X_t^i)dt + \sigma\sqrt{X_t^i}dW_t^i + dJ_t^i,$$

where (W_t^i) are standard Brownian motions and (J_t^i) are compound Poisson process with exponentially distributed jump sizes. The conditional survival probability is then given by

$$\mathbb{Q}(\tau_i > T|\mathcal{F}_t) = E^{\mathbb{Q}}\left[\exp\left(-\int_t^T \lambda_u^i du\right)\Big|\mathcal{F}_t\right].$$

We will denote by $\mathbf{X}_t = (X_t^0, X_t^1, \ldots, X_t^n)$ the risk factor values at time t. In order to calibrate the model, we follow the algorithm proposed by Eckner [11]. The tractability of this model relies on the conditional independence assumption of the default processes and also an analytical formula for the characteristic function of the affine jump-diffusion process. We refer readers to [9, 11] for the details of the calibration procedure.

Since $(\mathbf{X}_t, \mathbf{D}_t)$ is a Markov process, given the values $\mathbf{X}_t = \mathbf{X} = (X^0, X^1, \ldots, X^n)$ and the default indicators $\mathbf{D}_t = \mathbf{D}$, we can write the mark-to-market values of CDS i, the index, and a tranche $[a, b]$ at time t computed under the affine jump-diffusion model as $V_{af}^i(t, X^0, X^i)$, $V_{af}^{idx}(t, \mathbf{X}, \mathbf{D})$, and $V_{af}^{[a,b]}(t, \mathbf{X}, \mathbf{D})$, respectively. The mark-to-market values are computed as given in (1), (2), and (3).

4.3. Local intensity models. Local intensity models [6, 7, 16, 21, 25] are top-down models in which the number of defaults N_t in a reference portfolio is modeled as a Markov point process with an intensity $\lambda_t = f(t, N_{t-})$: the *portfolio default intensity* is a (positive) function of time and the number of defaulted obligors.

We consider the following parametrization of the local intensity function, introduced by Herbertsson [16]:

$$\lambda_t = (n - N_{t-}) \sum_{k=0}^{N_{t-}} b_k, \tag{5}$$

where $\{b_k\}$ are the parameters. The interpretation of (5) is that the portfolio default intensity jumps by an amount b_k when the kth default happens. There is no sign restriction on $\{b_k\}$ as long as the portfolio default intensity remains positive. As in [16], we parameterize $\{b_k\}$ as

$$b_k = \begin{cases} b^{(1)}, & 1 \le k < \mu_1, \\ b^{(2)}, & \mu_1 \le k < \mu_2, \\ \vdots & \\ b^{(I)}, & \mu_{I-1} \le k < \mu_I = n, \end{cases} \tag{6}$$

where $1, \mu_1, \ldots, \mu_I$ is a partition of $\{1, \ldots, n\}$ which includes the attachment points of the tranches. The local intensity model is a Markovian top-down model in which the only risk factor is the loss process. Therefore we can express the mark-to-market values of the index and a tranche $[a, b]$ at date t computed as functions $V_{lo}^{idx}(t, N)$ and $V_{lo}^{[a,b]}(t, N)$ of the number of defaults and time.

4.4. Bivariate spread-loss model. One major shortcoming of the local intensity model is that spreads have piecewise-deterministic dynamics—i.e., no “volatility”—between defaults. As we have seen in Figure 1, credit derivative positions fluctuate substantially in value even in the absence of defaults in the underlying credit portfolio, so a hedging strategy based on the jump-to-default ratio may lead to poor performance. A more realistic picture is given by a two-factor top-down model [1, 21] which accounts for both default risk and spread volatility by allowing the portfolio default intensity to depend on the number of defaults and a factor driving spread volatility:

$$\lambda_t = F(t, N_{t-}, Y_t).$$

Arnsdorff and Halperin [1] model the number of defaults N_t in a reference portfolio as a point process which has a (portfolio default) intensity (λ_t) that follows

$$\lambda_t = e^{Y_t}(n - N_{t-}) \sum_{k=0}^{N_{t-}} b_k, \tag{7}$$

where $\{b_k\}$ are parameters and the second factor (Y_t) generates spread volatility between default dates which follows an Ornstein–Uhlenbeck process,

$$dY_t = -\kappa Y_t dt + \sigma dW_t,$$

Table 3

Relative calibration error (RMSE), as a percentage of market spreads.

Model	CDS	Index	0%–3%	3%–7%	7%–10%	10%–15%	15%–30%
Gaussian copula	6.04	0.00	0.00	0.18	0.05	0.00	0.00
Affine jump-diffusion	14.53	21.67	14.56	5.05	10.40	13.91	6.17
Local intensity	-	6.29	1.60	0.95	0.46	0.34	1.71
Bivariate spread-loss	-	6.67	1.67	1.02	0.42	0.39	1.79

where (W_t) is a standard Brownian motion. Notice that the parameters $\{b_k\}$ will provide enough degrees of freedom to fit the CDO tranche spreads on a given date, so the remaining parameters (κ, σ) are estimated from time series of tranche spreads as follows:

- On the first sample day: Set $Y_0 = 0$, $\kappa = 0.3$, and $\sigma = 0.7$ and calibrate $\{b_k\}$ to index and tranche spreads on day 1.
- On the jth sample day:
  1. Fix $\{b_k\}$ as those calibrated on day $j-1$.
  2. Calibrate Y_0, κ, and σ by minimizing the mean square pricing error of day $j-10$, $j-9, \ldots, j$.
  3. Calibrate $\{b_k\}$ to the index and the tranche spreads on day j.

Since (N_t, Y_t) is a Markov process, given the values $N_t = N$ and $Y_t = Y$, we can write the mark-to-market values of the index and a tranche $[a, b]$ at time t computed under the bivariate spread-loss model as functions $V_{bi}^{idx}(t, Y, N)$ and $V_{bi}^{[a,b]}(t, Y, N)$ of the state variables.

4.5. Calibration results. All models are calibrated to the same market data using a 40% recovery rate. Table 3 shows the root mean square calibration error (RMSE). Since we calibrate the Gaussian copula model to each tranche with a different correlation (compound correlations), it by design gives good fits to the tranche spreads. The discrepancy of CDS spreads is due to the adjustment to match the index spreads. Top-down models are amendable to calibration to the market data as well. The RMSE are well within 2% for all tranches and around 5% for the CDS and the index. On the other hand, the Duffie–Gârleanu affine jump-diffusion does not calibrate market data as well as the top-down models. The CDS and the tranche spreads have RMSE at about 10%, which is still reasonable, but the fit to the index spread has RMSE larger than 20%, which is a poor fit. This is due to the fact that its calibration involves a high dimensional nonlinear optimization problem which is not guaranteed to converge. Therefore, we will consider only a hedging strategy using the single-name CDS as the hedging instruments in the affine jump-diffusion framework, so that the poor calibration to the index will not affect our analysis significantly. Note that, for all models under consideration, we have experienced a poor fit to the super senior tranche $[30\%, 100\%]$, even with the Gaussian copula model. Since poor calibration leads to inaccurate computation of the mark-to-market values and the hedge ratios, we will omit the $[30\%, 100\%]$ tranche in what follows.

5. Hedging strategies. Our objective is to hedge a position in a tranche $[a, b]$ using the constituent CDSs and the index, and sometimes with an additional tranche $[l, u]$. We will now introduce different dynamic hedging strategies that aim to achieve this task.

We assume a continuously rebalancing framework and let (ϕ_t^i), (ϕ_t^{idx}), and $(\phi_t^{[l,u]})$ be predictable processes which denote the hedging positions in CDS i, the index, and a tranche

$[l, u]$, respectively. In addition, we will use the same notation as in section 4 to represent the mark-to-market values of the credit derivatives computed under different models. All hedging strategies are implemented using daily rebalancing.

5.1. Delta hedging of single-name spread movements. The most common approach for hedging CDO tranches is to hedge against small changes in the single-name CDS spreads [10, 24]. In practice, traders usually consider delta hedging under the Gaussian copula model where the corresponding hedging position in CDS i is known as the *spread-delta*:

$$\phi_t^i = \frac{\delta_{s^i} V_{gc}^{[a,b]}(t, \mathbf{s}_{t-}^{cds}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-})}{\delta_{s^i} V_{gc}^i(t, \mathbf{s}_{t-}^i)}, \tag{8}$$

where

$$\delta_{s^i} V_{gc}^{[a,b]}(t, \mathbf{s}^{cds}, \boldsymbol{\rho}, \mathbf{D}) = V_{gc}^{[a,b]}(t, \mathbf{s}^{cds} + \mathbf{e}_i, \boldsymbol{\rho}, \mathbf{D}) - V_{gc}^{[a,b]}(t, \mathbf{s}^{cds}, \boldsymbol{\rho}, \mathbf{D}),$$
$$\delta_{s^i} V_{gc}^i(t, s^i) = V_{gc}^i(t, s^i + 1bp) - V_{gc}^i(t, s^i)$$

are the changes in value of a tranche $[a, b]$ and CDS i with respect to an increase in the CDS spread of obligor i by 1 basis point, while the correlations and other CDS spreads remain unchanged. $\mathbf{e}_i \in \mathbb{R}^n$ is a vector with all entries equal to 0 except for the ith entry equal to 1 basis point.

In order to compute the spread-deltas, we first calibrate the one-factor Gaussian copula model [18] to the market CDS and the tranche spreads as described in section 4.1. After that, we perturb the CDS spread of, say, obligor i by 1 basis point while keeping all other CDS spreads and the correlations unchanged. Then, we *recalibrate* the hazard rate function of obligor i and compute the new values for CDS i and the tranche. The spread-delta defined by (8) is the ratio of the change in the tranche value to the change in the CDS value.

The main drawback of implementing the spread-deltas (8) is the absence of well-defined dynamics for the single-name CDS spreads in the Gaussian copula framework. On the other hand, we can consider delta hedging under the dynamic affine jump-diffusion model [9]. Hedging moves in the single-name CDS spreads is then equivalent to hedging against changes in the idiosyncratic risk factor. The corresponding position in CDS i is equal to

$$\phi_t^i = \frac{\partial_{X^i} V_{af}^{[a,b]}(t, \mathbf{X}_{t-}, \mathbf{D}_{t-})}{\partial_{X^i} V_{af}^i(t, X_{t-}^0, X_{t-}^i)}, \tag{9}$$

where $\partial_{X^i} V_{af}^{[a,b]}(t, \mathbf{X}, \mathbf{D})$ and $\partial_{X^i} V_{af}^i(t, X^0, X^i)$ are the partial derivatives with respect to X^i which can be approximated by finite differences. Note that one of the main differences between the hedge ratio (9) and the spread-delta (8) is that there is no recalibration involved when computing (9) under the affine jump-diffusion model.

5.2. Delta hedging of index spread movements. In section 3.1, we observe that the CDS and the tranche spreads appear to be driven by some common risk factors. Therefore, we may argue that it is also important to hedge against *global* movements in the CDS spreads. We use the Gaussian copula model and enter positions in the index to neutralize the *index spread-delta*:

$$\phi_t^{idx} = \frac{\Delta_{gc}^{[a,b]}(t)}{\Delta_{gc}^{idx}(t)}, \tag{10}$$

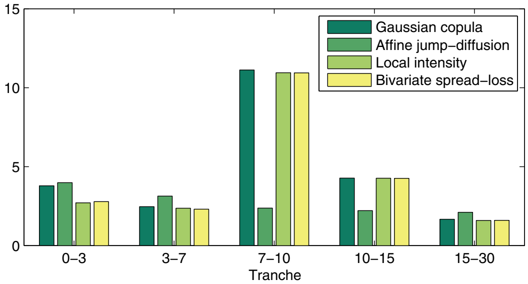


Figure 3. *Index spread-deltas. Data: CDX.NA.IG.S*10 *on* 25 *March,* 2008.

where

$$\Delta_{gc}^{[a,b]}(t) = V_{gc}^{[a,b]}(t, \mathbf{s}_{t-}^{cds} + \mathbf{e}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-}) - V_{gc}^{[a,b]}(t, \mathbf{s}_{t-}^{cds}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-}),$$
$$\Delta_{gc}^{idx}(t) = V_{gc}^{idx}(t, \mathbf{s}_{t-}^{cds} + \mathbf{e}, \mathbf{D}_{t-}) - V_{gc}^{idx}(t, \mathbf{s}_{t-}^{cds}, \mathbf{D}_{t-}),$$

and $\mathbf{e} \in \mathbb{R}^n$ is a vector with all entries equal to 1 basis point. Notice that $\Delta_{gc}^{[a,b]}(t)$ and $\Delta_{gc}^{idx}(t)$ are the changes of a tranche $[a, b]$ and the index values with respect to a parallel shift in all CDS spreads by 1 basis point while keeping the correlations unchanged. Computation of the index spread-delta is the same as for the spread-deltas, except that we need to shift all CDS spreads by 1 basis point.

This strategy also has the advantage of being cost-effective. Unlike delta hedging individual CDS fluctuations, which requires rebalancing multiple hedging positions, this strategy only requires adjusting the position in the index.

If we consider the CDS/index spreads and the correlations as the market inputs, which is analogous to the stock price and implied volatility moves in equity derivatives markets, the index spread-delta (10) can also be computed by models other than the Gaussian copula model. The procedure is similar to the case described above in which we first calibrate the models to the CDS/index spreads and the correlations. Then, we recalibrate the models to the perturbed CDS/index spreads while keeping the correlations unchanged. The index spread-delta is the ratio of the change in tranche value over the change in the index value.

Figure 3 shows the index spread-deltas (10) computed under different models. Interestingly, we observe that the index spread-deltas are very similar across the models, except those for tranches [7%, 10%] and [10%, 15%] computed from the affine jump-diffusion model. In fact, this discrepancy is due only to the fact that the affine jump-diffusion model does not calibrate well to the market data.

The similarity of the index spread-deltas across the models implies that there is no point in using a more sophisticated model if its only use is to delta hedge spread risk. The standard one-factor Gaussian copula model would be sufficient to carry out this strategy. Indeed, a more meaningful hedging strategy for the dynamic models is to hedge against the underlying risk factors specified in the modeling framework, taking into account the dynamics of these factors and their correlations.

5.3. Delta and gamma hedging of index spread movements. By analogy with gamma hedging of equity derivatives, one may consider hedging the second-order changes in the tranche values due to fluctuations in the CDS spreads. We consider positions in the index and a tranche $[l,u]$ such that

$$\Delta_{gc}^{[a,b]}(t) = \phi_t^{[l.u]}\,\Delta_{gc}^{[l,u]}(t) + \phi_t^{idx}\,\Delta_{gc}^{idx}(t),$$
$$\Gamma_{gc}^{[a,b]}(t) = \phi_t^{[l.u]}\,\Gamma_{gc}^{[l,u]}(t) + \phi_t^{idx}\,\Gamma_{gc}^{idx}(t),$$

where

(11)
$$\Gamma_{gc}^{[a,b]}(t) = V_{gc}^{[a,b]}(t, \mathbf{s}_{t-}^{cds} + \mathbf{e}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-}) - 2V_{gc}^{[a,b]}(t, \mathbf{s}_{t-}^{cds}, \boldsymbol{\rho}_t, \mathbf{D}_{t-}) + V_{gc}^{[a,b]}(t, \mathbf{s}_{t-}^{cds} - \mathbf{e}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-}),$$
(12)
$$\Gamma_{gc}^{idx}(t) = V_{gc}^{idx}(t, \mathbf{s}_{t-}^{cds} + \mathbf{e}, \mathbf{D}_{t-}) - 2V_{gc}^{idx}(t, \mathbf{s}_{t-}^{cds}, \mathbf{D}_{t-}) + V_{gc}^{idx}(t, \mathbf{s}_{t-}^{cds} - \mathbf{e}, \mathbf{D}_{t-})$$

are the *gammas* of a tranche $[a,b]$ and the index, or equivalently the second-order finite differences in the values with respect to 1 basis point shifting of all CDS spreads. Solving for the hedge ratios, we have

$$\phi_t^{idx} = \frac{\Delta_{gc}^{[a,b]}(t)\Gamma_{gc}^{[l,u]}(t) - \Delta_{gc}^{[l,u]}(t)\Gamma_{gc}^{[a,b]}(t)}{\Delta_{gc}^{idx}(t)\Gamma_{gc}^{[l,u]}(t) - \Delta_{gc}^{[l,u]}(t)\Gamma_{gc}^{idx}(t)}, \tag{13}$$

$$\phi_t^{[l,u]} = \frac{\Delta_{gc}^{[a,b]}(t)\Gamma_{gc}^{idx}(t) - \Delta_{gc}^{idx}(t)\Gamma_{gc}^{[a,b]}(t)}{\Delta_{gc}^{[l,u]}(t)\Gamma_{gc}^{idx}(t) - \Delta_{gc}^{idx}(t)\Gamma_{gc}^{[l,u]}(t)}. \tag{14}$$

Note that we can also delta hedge against movements in the single-name CDS spreads in the case of gamma hedging. However, we need to solve an ill-conditioned linear system which may lead to unstable hedge ratios. Moreover, as we will see in section 6, the main component of changes in the single-name CDS spreads is a parallel move which is already reflected in the index spread. Thus, the inclusion of single-name CDS corresponds to hedging higher-order principal components, which have a smaller impact on the variance of the portfolio. Therefore, we do not include single-name CDS hedges in our gamma hedging analysis.

Unlike the situation in a Black–Scholes model, where the gamma of a call or put option is always positive and the gamma of a long position in a call or put option can be neutralized by shorting another call or put option, such simple relations fail to hold in the Gaussian copula framework for CDO tranches. Figure 4 shows the gammas of tranches $[0\%, 3\%]$ and $[3\%, 7\%]$ computed for various days in the sample. Observe that the gammas can be positive or negative, even for the equity tranche. Moreover, the gammas of the two tranches do not have any clear relationship, in the sense that they do not always have the same sign. These results also suggest that the empirical performance of gamma hedging may depend on the choice of the hedging tranches, so we will consider below different choices of tranches as hedging instruments.

5.4. Hedging parallel shifts in correlations. In addition to hedging against changes in spreads, by analogy with "vega" hedging in the Black–Scholes model, one can argue that it is also important to manage the risk of the fluctuation in another parameter of the Gaussian copula model, the implied correlation. We will consider scenarios where all (compound) correlations shift by the same magnitude.

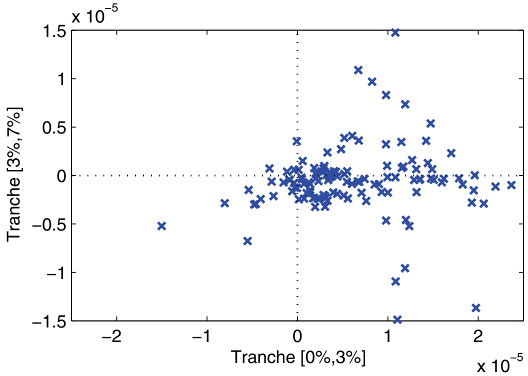


Figure 4. *Values of gammas* $\Gamma_{gc}^{[a,b]}(t)$ *for* $[0\%, 3\%]$ *and* $[3\%, 7\%]$ *CDX tranches. Each point represents one day in the sample.*

A joint hedge with respect to small changes in spreads and correlation then requires us to enter positions in the CDSs and a tranche $[l, u]$ such that

$$\delta_{s^i} V_{gc}^{[a,b]}(t, \mathbf{s}_{t-}^{cds}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-}) = \phi_t^{[l,u]}\, \delta_{s^i} V_{gc}^{[l,u]}(t, \mathbf{s}_{t-}^{cds}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-}) + \phi_t^i\, \delta_{s^i} V_{gc}^i(t, s_{t-}^i), \quad i = 1, \ldots, n,$$
$$\delta_\rho V_{gc}^{[a,b]}(t, \mathbf{s}_{t-}^{cds}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-}) = \phi_t^{[l,u]}\, \delta_\rho V_{gc}^{[l,u]}(t, \mathbf{s}_{t-}^{cds}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-}),$$

where

$$\delta_\rho V_{gc}^{[a,b]}(t, \mathbf{s}^{cds}, \boldsymbol{\rho}, \mathbf{D}) = V_{gc}^{[a,b]}(t, \mathbf{s}^{cds}, \boldsymbol{\rho} + 0.1\%, \mathbf{D}) - V_{gc}^{[a,b]}(t, \mathbf{s}^{cds}, \boldsymbol{\rho}, \mathbf{D})$$

is the change in the tranche value with respect to an increase in all compound correlations by 0.1%. Since the index is insensitive to the correlations, we must consider another tranche as a hedging instrument. Solving for the hedge ratios, we have

$$\phi_t^{[l,u]} = \frac{\delta_\rho V_{gc}^{[a,b]}(t, \mathbf{s}_{t-}^{cds}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-})}{\delta_\rho V_{gc}^{[l,u]}(t, \mathbf{s}_{t-}^{cds}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-})},$$
$$\phi_t^i = \frac{\delta_{s^i} V_{gc}^{[a,b]}(t, \mathbf{s}_{t-}^{cds}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-}) \delta_\rho V_{gc}^{[l,u]}(t, \mathbf{s}_{t-}^{cds}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-}) - \delta_\rho V_{gc}^{[a,b]}(t, \mathbf{s}_{t-}^{cds}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-}) \delta_{s^i} V_{gc}^{[l,u]}(t, \mathbf{s}_{t-}^{cds}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-})}{\delta_{s^i} V_{gc}^i(t, s_{t-}^i)\, \delta_\rho V_{gc}^{[l,u]}(t, \mathbf{s}_{t-}^{cds}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-})}$$

Note that this strategy does not take into account the *comovements* in the index and in correlations. This is simply a first-order sensitivity-based hedge against moves in the CDS spreads and the correlation movements.

5.5. Hedging default risk. Although the common hedging practice is to protect against CDS spread fluctuations, the occurrence of defaults is also a major source of risk of a CDO tranche. The natural strategy to hedge against constituent defaults is to enter positions in the CDS or the index according to the *jump-to-default ratio*, which is defined as the ratio of the change in the tranche value over the change in the index (or CDS) value with respect to one additional default.

Since the top-down models, such as the local intensity model, focus on the next-to-default rather than the default of a specific obligor, there is only one jump-to-default ratio to consider,

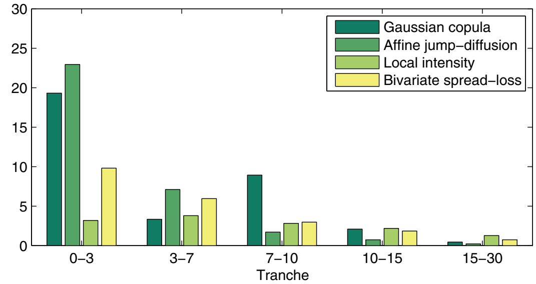


Figure 5. *Jump-to-default ratios. Data: CDX.NA.IG.S*10 *on* 25 *March,* 2008*. Ratios of Gaussian copula model and affine jump-diffusion model are computed by assuming that IBM defaults.*

which is equal to

$$\phi_t^{idx} = \frac{V_{lo}^{[a,b]}(t, N_{t-} + 1) - V_{lo}^{[a,b]}(t, N_{t-})}{V_{lo}^{idx}(t, N_{t-} + 1) - V_{lo}^{idx}(t, N_{t-})}. \tag{15}$$

The jump-to-default ratio under the bivariate spread-loss model is defined in the same manner.

On the other hand, since the bottom-up models specify the default probabilities of each obligor, there are n possible jump-to-default ratios to be considered. For the Gaussian copula model, the jump-to-default ratio corresponding to obligor i, using the index as the hedging instrument, is equal to

$$\phi_t^{idx} = \frac{V_{gc}^{[a,b]}(t, \mathbf{s}_{t-}^{cds}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-} + \mathbf{u}_i) - V_{gc}^{[a,b]}(t, \mathbf{s}_{t-}^{cds}, \boldsymbol{\rho}_{t-}, \mathbf{D}_{t-})}{V_{gc}^{idx}(t, \mathbf{s}_{t-}^{cds}, \mathbf{D}_{t-} + \mathbf{u}_i) - V_{gc}^{idx}(t, \mathbf{s}_{t-}^{cds}, \mathbf{D}_{t-})}, \tag{16}$$

where $\mathbf{u}_i \in \mathbb{R}^n$ is a vector with all entries equal to 0 except for the ith entry equal to 1. We can define the jump-to-default ratio for the affine jump-diffusion model in the same manner.

Figure 5 shows the jump-to-default ratios computed from different models using the index as the hedging instrument as in (15) and (16). For the bottom-up models, the ratios are computed in the scenario where IBM defaults. Unlike the index spread-deltas, the jump-to-default ratios are substantially different across models. This implies that hedging against occurrence of default is also exposed to substantial model risk, because the jump-to-default ratio, which is computed by assuming the occurrence of one additional default, depends on the credit portfolio loss dynamic in the modeling framework. Although each model is calibrated to the same CDS, index, and CDO market data, those credit derivatives provide information only on the *marginal distribution* of the loss process at some fixed times. Therefore, the dynamic of the loss process cannot be uniquely determined by the market data, and thus the jump-to-default ratios are substantially different across models.

5.6. Variance minimization. When spread risk and default risk are considered simultaneously, we are in an incomplete market setting, and hedging strategies in this setting need to be determined by an optimality criterion which takes both spread risk and default risk into account. A well-known approach to hedging in incomplete markets is the *variance-minimizing* *strategy*, introduced by Föllmer and Sondermann [13]. Unlike many other approaches to hedging in incomplete markets, it has been shown that this approach actually leads to analytically tractable hedging strategies [8, 14].

Definition 5.1. *Let H be a square-integrable contingent claim at maturity T, and (X_t) be the discounted price process of the hedging instrument which is a square-integrable martingale under $\mathbb{Q}$. Let $\mathcal{S}$ be the set of admissible self-financing strategies with $E^{\mathbb{Q}}[(\int_0^T \phi_t dX_t)^2] < \infty$. A variance-minimizing strategy is a choice of initial capital c and a self-financing trading strategy $(\phi_t) \in \mathcal{S}$ which minimizes the quadratic hedging error:*

$$\inf_{c\in\mathbb{R},(\phi_t)\in\mathcal{S}} E^{\mathbb{Q}}\left[\left(c + \int_0^T \phi_t dX_t - H\right)^2\right]. \tag{17}$$

The variance (17) is computed under the risk-neutral probability measure $\mathbb{Q}$, because the models are calibrated to the observed credit spreads which only provide information on the risk factor dynamics under the risk-neutral measure. If we want to minimize the hedging error under the real-world measure, we will need a statistical model. One example is the regression-based hedging strategy in section 5.7.

Föllmer and Sondermann [13] characterize the variance-minimizing strategy in terms of a Galtchouk–Kunita–Watanabe projection of the claim on the set of replicable payoffs (see Appendix A). One nice property of the variance-minimizing strategy is that it coincides with the self-financing hedging strategy, which replicates the contingent claim in a complete market. Furthermore, in many Markovian models with jumps, the variance-minimizing hedge ratios can be explicitly computed [8, 14, 23]. The justification for this approach, which is not specific to credit derivatives, is discussed in [8, 14] from a methodological standpoint.

We will show that these variance-minimizing hedge ratios can be explicitly computed for the local intensity model and the bivariate spread-loss model. Our analysis will omit the Gaussian copula model and the affine jump-diffusion model. The reason is that since the Gaussian copula model defines the marginal distribution of the portfolio losses only at fixed times, there is no intrinsic dynamic for the loss process. For the affine jump-diffusion model, the computation requires inverting a high dimensional matrix (125×125) which is numerically unstable.

5.6.1. Local intensity model. Laurent, Cousin, and Fermanian [17] show that the local intensity framework generates a complete market in the sense that we can replicate the payoff of a tranche $[a, b]$ by means of a self-financing portfolio with positions in the index default swap and a numeraire. The hedge ratio in the replication strategy then coincides with the variance-minimizing hedge ratio, which is equal to the jump-to-default ratio.

Proposition 5.2. *Consider a local intensity model in which the portfolio default intensity (λ_t) under the risk-neutral measure $\mathbb{Q}$ has the form $\lambda_t = f(t, N_{t-})$ for some positive function $f(.,.) > 0$. Let $V_{lo}^{idx}(t, N)$ and $V_{lo}^{[a,b]}(t, N)$ be the mark-to-market values of the index and a tranche $[a, b]$ at time t given N number of defaults which satisfy the following:*

- *For $N = 0, \dots, n$, the functions $t \mapsto V_{lo}^{idx}(t, N)$ and $t \mapsto V_{lo}^{[a,b]}(t, N)$ belong to $C^1([0, T))$.*
- *$|V_{lo}^{idx}(t, N+1) - V_{lo}^{idx}(t, N)| > 0$ for all $t \in [0, T]$, $N = 0, \dots, n-1$.*

Then, the variance-minimizing hedge defined in Definition 5.1 for a tranche $[a, b]$ using the index as the only hedging instrument is given by

$$\phi_t = \frac{V_{lo}^{[a,b]}(t, N_{t-}+1) - V_{lo}^{[a,b]}(t, N_{t-})}{V_{lo}^{idx}(t, N_{t-}+1) - V_{lo}^{idx}(t, N_{t-})}. \tag{18}$$

The second condition in Proposition 5.2 implies that the index default swap value is always sensitive to defaults in the underlying portfolio. The proof of this proposition is shown in Appendix B.1.

5.6.2. Bivariate spread-loss model. In the bivariate spread-loss model, the portfolio default intensity is driven by the loss process (L_t) and a diffusion process (Y_t). Since we consider the index as the only hedging instrument, the market is incomplete in this two-factor framework. We use variance minimization to compute a trade-off between default risk and spread risk, as follows.

Proposition 5.3. *Consider the bivariate spread-loss model [1] in which the portfolio default intensity (λ_t) under the risk-neutral measure $\mathbb{Q}$ follows (7). Let $V_{bi}^{idx}(t, Y, N)$ and $V_{bi}^{[a,b]}(t, Y, N)$ be the values of the index and a tranche $[a, b]$ at time t given N number of defaults and risk factor value Y which satisfy the following:*

- *For $N = 0, \ldots, n$, the functions $(t, Y) \mapsto V_{bi}^{idx}(t, Y, N)$ and $(t, Y) \mapsto V_{bi}^{[a,b]}(t, Y, N)$ belong to $C^{1,2}([0, T] \times \mathbb{R})$;*
- $[\partial_Y V_{bi}^{idx}(t, Y, N)]^2 + [\delta_N V_{bi}^{idx}(t, Y, N)]^2 \lambda > 0$ *for all $t \in [0, T]$, $Y \in \mathbb{R}$, $N = 0, \ldots, n-1$;*
- $E^{\mathbb{Q}}\left[\int_0^T \left(\partial_Y V_{bi}^{[a,b]}(t, Y_{t-}, N_{t-})\right)^2 dt\right] < \infty$,

where $\lambda = (n - N)\sum_{k=0}^{N} b_k$, $\partial_Y V_{bi}^{\cdot}(t, Y, N)$ is the partial derivative with respect to Y, and $\delta_N V_{bi}^{\cdot}(t, Y, N) = V_{bi}^{\cdot}(t, Y, N+1) - V_{bi}^{\cdot}(t, Y, N)$ is the change of value with respect to one additional default. Then, the variance-minimizing hedge defined in Definition 5.1 for a tranche $[a, b]$ using the index as the only hedging instrument is given by

$$\phi_t = \frac{\partial_Y V_{bi}^{[a,b]}(t, Y_{t-}, N_{t-})\, \partial_Y V_{bi}^{idx}(t, Y_{t-}, N_{t-})\, \sigma^2 + \delta_N V_{bi}^{[a,b]}(t, Y_{t-}, N_{t-})\, \delta_N V_{bi}^{idx}(t, Y_{t-}, N_{t-})\, \lambda_t}{\left[\partial_Y V_{bi}^{idx}(t, Y_{t-}, N_{t-})\, \sigma\right]^2 + \left[\delta_N V_{bi}^{idx}(t, Y_{t-}, N_{t-})\right]^2 \lambda_t}. \tag{19}$$

The second condition in Proposition 5.3 implies that the index default swap is always sensitive to either the change of the risk factor (Y_t) or defaults in the underlying portfolio. The third condition is an integrability condition to ensure that the optimal hedging portfolio has finite variance. The proof of this proposition is shown in Appendix B.2.

Unlike the case of the local intensity model, the variance-minimizing hedge ratio for the bivariate spread-loss model involves not only the jump-to-default values but also the partial derivatives of the values with respect to the additional risk factor (Y_t). This reflects the fact that variance-minimization hedging is a strategy that takes both default risk and spread risk into account.

Figure 6 shows the variance-minimizing hedge ratios of the local intensity model and the bivariate spread-loss model. Similar to the case of comparing the jump-to-default ratios, the variance-minimizing hedge ratios are substantially different across the models, especially for the junior tranches. The reason for the differences is the same as the case for the jump-to-

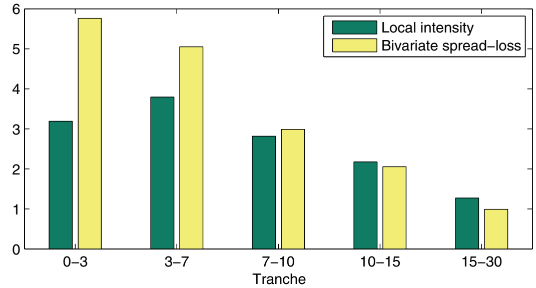


Figure 6. *Variance-minimization hedge ratios. Data: CDX.NA.IG.S10 on 25 March, 2008.*

default ratios: from (18) and (19), we can see that the variance-minimizing hedge ratio is a model-dependent quantity.

5.7. Regression-based hedging. One drawback of hedging strategies based on pricing models is that it is not clear how well the models can capture the dynamics of the credit spreads under the real-world measure, which is an important issue for hedging in practice. We now discuss a model-free, regression-based hedging strategy based on the observed dynamics of the credit spreads.

Consider a simple regression model relating the daily changes in the index and tranche values:

$$\delta V_{t_j}^{[a,b]} = \alpha^{[a,b]} + \beta^{[a,b]}\, \delta V_{t_j}^{idx} + \epsilon_j,$$

where $\delta V_{t_j}^{\cdot} = V_{t_j}^{\cdot} - V_{t_{j-1}}^{\cdot}$ is the daily change of value from time t_{j-1} to t_j. $\alpha^{[a,b]}$ and $\beta^{[a,b]}$ can be estimated by the ordinary least squares regression over a rolling window. Choosing the hedging position in the index default swap as

$$\hat{\beta}_t^{[a,b]} := \frac{\sum_{t_j \le t}(\delta V_{t_j}^{idx} - \overline{\delta V_t^{idx}})(\delta V_{t_j}^{[a,b]} - \overline{\delta V_t^{[a,b]}})}{\sum_{t_j \le t}(\delta V_{t_j}^{idx} - \overline{\delta V_t^{idx}})^2} \tag{20}$$

yields a model-free hedging strategy which we call the *regression-based hedge*. Here $\overline{\delta V_t^{\cdot}}$ is the average of daily P&L on $[0, t]$, and $\hat{\beta}_t^{[a,b]}$ is the estimate of $\beta^{[a,b]}$ using observations over the period $[0, t]$.

The main advantages of this strategy are its ability to directly capture the actual dynamics of comovements in credit spreads and its simplicity.

6. Empirical performance of hedging strategies. We now present an empirical assessment of the performance of the eight hedging strategies described in section 5 using the dataset in section 3. Table 4 summarizes all the strategies that will be considered. Note that we will consider two choices of hedging tranches for gamma hedging where the details will be presented in the following subsections.

We consider the hedging of the protection seller's position on a CDX tranche, initiated on the first day of the sample period. On each day, we calibrate the models to the market data, as stated in section 5, and compute the corresponding hedging positions. A successful

Table 4
Overview of hedging strategies.

Strategy	Underlying risk	Model	Type	Nature
1	Single-name CDS fluctuation	Gaussian copula	Bottom-up	Static
2	Global CDS/index fluctuation	Gaussian copula	Bottom-up	Static
3	1st- and 2nd-order global CDS fluctuation	Gaussian copula	Bottom-up	Static
4	Single-name CDS fluctuation + correlations shifts	Gaussian copula	Bottom-up	Static
5	Single-name CDS fluctuation	Affine jump-diffusion	Bottom-up	Dynamic
6	Default risk	Local intensity	Top-down	Dynamic
7	Variance minimization (risk-neutral measure)	Bivariate spread-loss	Top-down	Dynamic
8	Variance minimization (statistical measure)	Ordinary least squares regression	Statistical	-

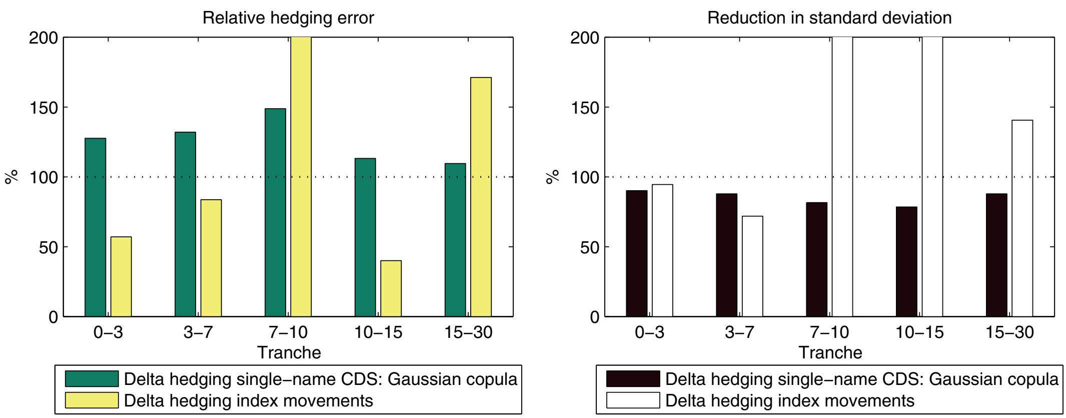


Figure 7. *Comparison among delta hedging strategies based on the Gaussian copula model.*

strategy should (substantially) reduce dispersion of the P&L distribution with respect to an unhedged position. Here, we use the following metrics to assess the reduction in magnitude and volatility of the daily P&L:

$$\text{Relative hedging error} = \left|\frac{\text{Average daily P\&L of hedged position}}{\text{Average daily P\&L of unhedged tranche position}}\right|,$$
$$\text{Residual volatility} = \frac{\text{Daily P\&L volatility of hedged position}}{\text{Daily P\&L volatility of unhedged tranche position}}.$$

Note that the two ratios should be close to 0 for a good hedging strategy.

6.1. Does delta hedging work? Our first analysis is to check whether the commonly criticized delta hedging strategies under the Gaussian copula model work. In Figure 7, we see that delta hedging does not work well. Indeed, the only effective strategy is a delta hedge against index movements, which reduces the absolute exposures of tranches $[0\%, 3\%]$, $[3\%, 7\%]$, and $[10\%, 15\%]$. On the other hand, delta hedging against single-name CDS movements fails to reduce absolute exposures. In terms of reduction in P&L volatility, delta hedging of single-

Table 5
Hedging tranches for gamma hedging strategy.

Tranche being hedged	0%–3%	3%–7%	7%–10%	10%–15%	15%–30%
Hedging tranche: Case 1	7%–10%	7%–10%	10%–15%	15%–30%	10%–15%
Hedging tranche: Case 2	15%–30%	15%–30%	15%–30%	7%–10%	7%–10%

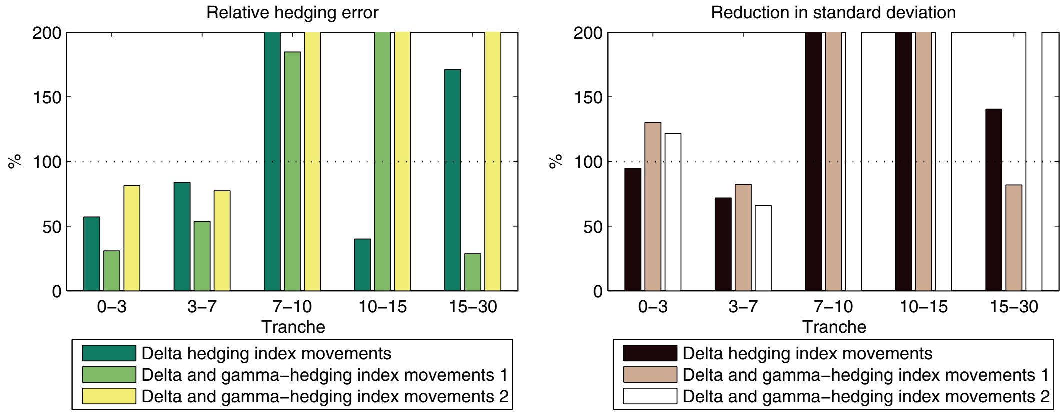


Figure 8. *Comparison among delta and gamma hedging strategies.*

name CDS movements results in a consistent reduction of volatility across all tranches, while delta hedging index movements fails to do so for the three most senior tranches.

6.2. Does gamma hedging improve hedging performance? In order to examine whether gamma hedging can improve performance, we consider two different choices of tranches as the hedging instruments, as illustrated in Table 5.

Hedging tranches in Case 1 are chosen such that they give the best performance in our sample in terms of the relative hedging errors, and those in Case 2 are chosen as a comparison with Case 1. Figure 8 shows that gamma hedging can help reduce the hedging error for the $[0\%, 3\%]$, $[3\%, 7\%]$, and $[15\%, 30\%]$ tranches and reduce the P&L volatility for the tranches $[3\%, 7\%]$ and $[15\%, 30\%]$. However, gamma hedging worsens the hedging performance in all other cases. Moreover, its performance is very sensitive to the choice of hedging tranches.

In summary, we conclude that gamma hedging, while very sensitive to the choice of hedging instruments, does not necessarily perform well. Indeed, observations in section 3.1 suggest that the appropriate correction to delta hedging is not the second-order sensitivity with respect to the CDS spread movements but a correction taking into account jumps in spreads. This situation also arises in other contexts when jump risk is present [8].

6.3. Can hedging parallel shifts in correlations improve performance? As shown in Figure 9, immunizing the portfolio against parallel shifts in the (implied) correlation does not improve performance: for almost all tranches neither the hedging error nor the P&L volatility is reduced. This suggests that the (observed) changes in the compound correlations are typically not parallel. Hedging performance may be improved if we consider other scenarios for changes in the correlations.

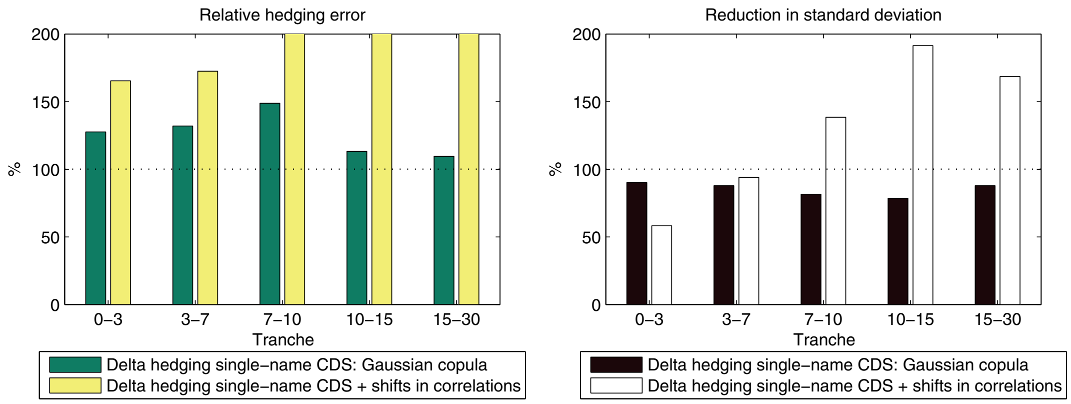


Figure 9. *Comparison between delta hedging and the addition of hedging parallel shifts in compound correlations.*

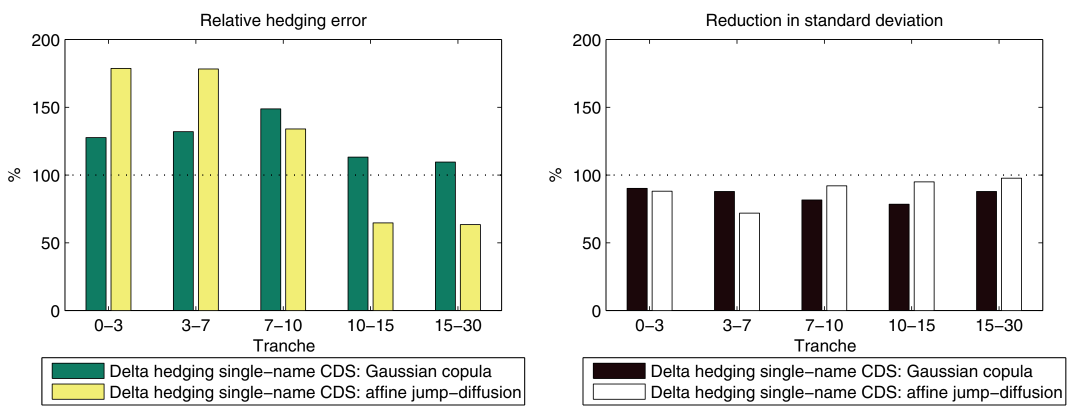


Figure 10. *Comparison between delta hedging under the static Gaussian copula and the dynamic affine jump-diffusion models.*

6.4. Do dynamic models have better hedging performance than static models?. Copula-based factor models have faced a lot of criticism for their insufficiency for hedging, and one popular explanation is that this is due to their static nature. In order to verify this claim, we compare the hedging performance of delta hedging under the static Gaussian copula model and under the dynamic affine jump-diffusion model [9].

Interestingly, the hedging error and the reduction in volatility ratios in Figure 10 do not provide any evidence that dynamic models perform better than this simple static model. Although the dynamic model successfully reduces the hedging error for tranches $[10\%, 15\%]$ and $[15\%, 30\%]$, it amplifies the hedging error significantly for the two most junior tranches. Moreover, the residual volatilities from the dynamic model are larger than those from the static model for tranches $[7\%, 10\%]$, $[10\%, 15\%]$, and $[15\%, 30\%]$.

6.5. Do bottom-up models perform better than top-down models? Although top-down models are more flexible for calibration, it has been suggested that they may be inadequate for hedging [2]. However, this claim has not been backed by any empirical evidence: we

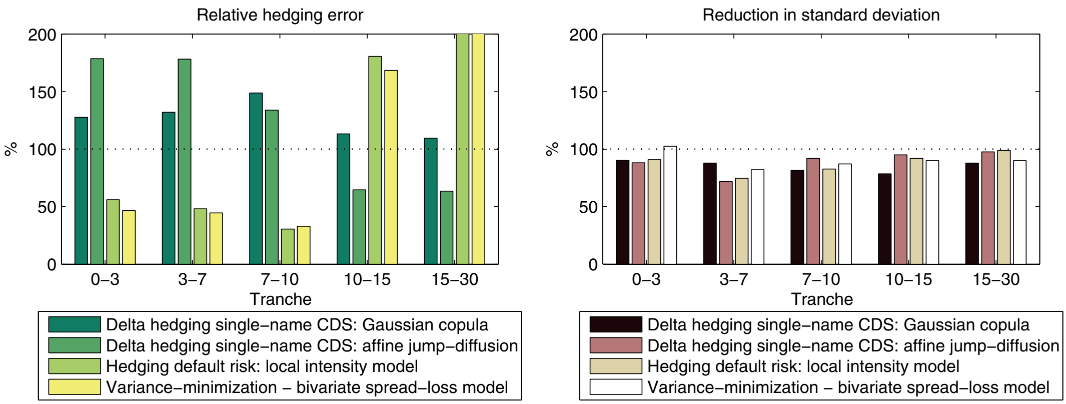


Figure 11. *Comparison of top-down and bottom-up hedging.*

will attempt to verify whether there is indeed an advantage in using bottom-up models for hedging.

Figure 11 compares the performance of various strategies based on bottom-up models, delta hedging based on the Gaussian copula model and the affine jump-diffusion model, versus the top-down strategies, hedging default risk based on the local intensity model and variance minimization based on the bivariate spread-loss model. First, we observe that the reduction of volatility is similar across the strategies, which does not provide much information to distinguish them. By comparing the hedging error, we find that the top-down models perform better than the bottom-up models for the three junior tranches, $[0\%, 3\%]$, $[3\%, 7\%]$, and $[7\%, 10\%]$, while bottom-up models fare better for the other two senior tranches.

Overall, there is no strong evidence that hedging based on the bottom-up models must perform better than that based on the top-down models. This observation contradicts the statements of Bielecki, Crépey, and Jeanblanc [2], who compare bottom-up and top-down hedging based on simulation. Although bottom-up models provide additional degrees of freedom, the effectiveness of a hedging strategy is not about goodness of fit but depends on how well the model can predict short-term comovements of the target instrument and the hedging instruments. From our results it appears that existing bottom-up models do a poor job at predicting such short-term comovements.

6.6. How does model-based hedging compare to regression-based hedging? Given the simplicity and intuitiveness of regression-based hedging, it is interesting to examine how well it performs relative to the model-based strategies. In Figure 12, we observe that regression-based hedging performs well across all tranches, consistently reducing both the hedging error and the daily P&L volatility. In particular, it reduces the volatilities for all tranches more than do the model-based strategies which are theoretically "optimal" in the respective models. This suggests that model misspecification is nonnegligible in all the models considered above.

6.7. Performance on credit event dates. Of particular interest is the performance of the hedging strategies on the next business day after the Lehman Brothers and Fannie Mae/ Freddie Mac credit events. Figure 13 shows the hedging error on the next business days after the credit events under various hedging strategies. During the Lehman Brothers event, we

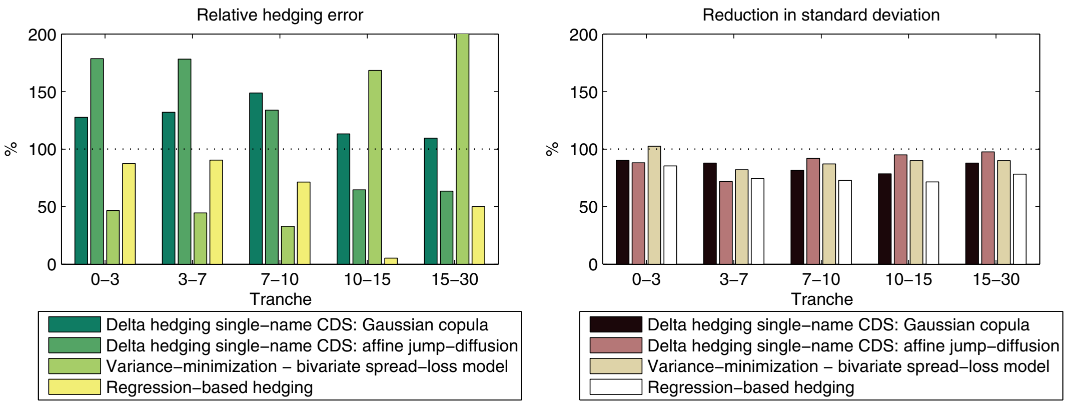


Figure 12. *Comparison among top-down, bottom-up, and regression-based hedging.*

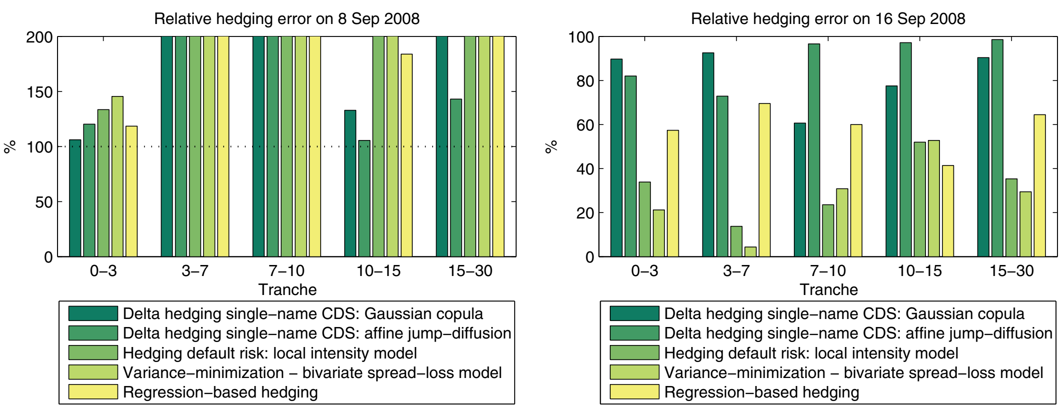


Figure 13. *Comparison of strategies on the next business day after Fannie Mae/Freddie Mac (8 September, 2008) and Lehman Brothers (16 September, 2008) credit events.*

observe that the top-down and regression-based hedging outperform the bottom-up hedging for all tranches. In particular, variance minimization based on the two-factor top-down model provides the best hedge for most tranches. This suggests that a macro event is better captured by top-down hedging.

On the other hand, all strategies failed to reduce the P&L during the Fannie Mae and Freddie Mac credit event, because, as we saw in section 3, the market happened to have anticipated the event, and there are no significant movements in the spreads. In particular, we observe a negative change in the CDS and index spreads, which leads to an increase in the overall exposure when we try to hedge the tranche positions with positive hedge ratios.

7. Conclusion. We have presented theoretical and empirical comparison of a wide range of hedging strategies for portfolio credit derivatives, with a detailed analysis of the hedging of index CDO tranches. By comparing the performance of these strategies in 2008, our analysis reveals several interesting features:

- Our analysis reveals a large proportion of unhedgeable risk in CDO tranches. This suggests that market completeness is by no means an acceptable approximation, and toy models which assume a complete market may fail to provide useful insight for issues related to hedging of CDO tranches.

- Delta hedging of CDO tranche positions based on the Gaussian copula model is not effective.
- Although gamma hedging can improve the performance for certain tranches, its effectiveness is very sensitive to the choice of hedging instruments and is inconsistent across tranches.
- We do not find strong evidence that the Duffie and Gârleanu [9] bottom-up dynamic model performs better than the static Gaussian copula model when it comes to delta hedging with credit default swaps.
- Moreover, bottom-up models ([9] and [18]) do not appear to perform consistently better than top-down models ([1] and [16]), in contrast to what has been asserted (without justification) in the literature [2]. In fact, during the period around the Lehman Brothers default, hedging strategies based on top-down models performed substantially better than those based on bottom-up models. This leads us to question the need for computationally costly dynamic bottom-up models instead of the top-down models for hedging portfolio credit derivatives.
- Model-free regression-based hedging appears to be surprisingly effective when compared to other hedging strategies. This suggests—not surprisingly—that incorporating the statistical behavior of credit spreads is an important criterion for a successful hedging strategy.
- We find evidence for common jumps, or large comovements, in spreads. However, index and tranche spreads did not appear to have upward jumps at the default dates of index constituents. This observation goes against models, such as Markovian contagion models [1, 16] and self-exciting models [15], in which jumps in spreads occur only at constituent default dates. Jumps in spreads may also arise from unexpected events not necessarily related to defaults inside the portfolio.

We have left out many practically important considerations, such as liquidity, transaction costs, and computational issues, when assessing hedging performance. The *(il)liquidity* of CDS contracts leads to questions of feasibility of hedging strategies which require frequent rebalancing of positions in single-name CDS. *Transaction costs*—as reflected, for instance, in bid-ask spreads—are known to be higher for single-name CDS contracts than for the index, and taking them into account would favor top-down/index hedging strategies as opposed to hedging with single-name CDS, which requires rebalancing more than a hundred single-name CDS positions. Finally, computational costs are much lower for the top-down models, especially when it comes to calibration: various fast calibration methods have been proposed for top-down models [1, 6, 7, 21], whereas parameter calibration, especially if it needs to be done on a periodic basis, remains nontrivial for bottom-up models [11]. Therefore it should be clear that taking these aspects into account would tilt the comparison even more in favor of top-down/index hedging as opposed to hedging with single-name CDS.

Appendix A. Variance-minimizing hedge and Galtchouk–Kunita–Watanabe decomposition. Our work in this section is similar to the earlier work by Frey and Backhaus [14]. We first define the *gain process* for the index and tranches. Let $P^{idx}(T_m)$ be the net payment received from an index default swap contract at time T_m which is bounded by definition. The mark-to-market value of an index default swap at time t is equal to

$$V_t^{idx} = \sum_{T_m > t} B(t, T_m) E^{\mathbb{Q}}[P^{idx}(T_m)|\mathcal{F}_t].$$

We define the discounted value process as

$$\tilde{V}_t^{idx} := B(0,t)V^{idx}(t).$$

Then, the gain process is defined as the present value of all cash-flows:

$$G_t^{idx} = \sum_{T_m \le t} B(0,T_m)P^{idx}(T_m) + \tilde{V}_t^{idx} = E^{\mathbb{Q}}\left[\sum_{T_m > 0} B(0,T_m)P^{idx}(T_m)\middle|\mathcal{F}_t\right], \tag{21}$$

where (G_t^{idx}) is a square-integrable (bounded) martingale under $\mathbb{Q}$. Similarly, we can derive the expression for the gain process of an $[a,b]$ tranche $(G_t^{[a,b]})$ which is also a square-integrable (bounded) martingale under $\mathbb{Q}$.

Consider the variance minimization setting in Definition 5.1. Our goal is to hedge a tranche $[a,b]$ using the index default swap. Although we consider a terminal payoff H at maturity in Definition 5.1, the results in [23] allow us to replace the conditional expected value of the terminal payoff $E^{\mathbb{Q}}[H|\mathcal{F}_t]$ by the gain process value $G_t^{[a,b]}$. Let (ϕ_t) be the variance-minimizing hedging strategy which represents positions in the index default swap. Then, it can be shown that (see [13]) (ϕ_t) satisfies the Galtchouk–Kunita–Watanabe decomposition

$$G_t^{[a,b]} = G_0^{[a,b]} + \int_0^t \phi_s dG_s^{idx} + Z_t, \tag{22}$$

where the process $(Z_t G_t^{idx})$ is a martingale under $\mathbb{Q}$. Therefore, the variance-minimizing strategy satisfies

$$d\langle G^{[a,b]}, G^{idx}\rangle_t = \phi_t d\langle G^{idx}\rangle_t, \quad 0 \le t \le T, \tag{23}$$

where $(\langle G\rangle_t)$ denotes the unique predictable process with $\langle G\rangle_0 = 0$ and right-continuous increasing paths such that $(G_t^2 - \langle G\rangle_t)$ is a martingale under $\mathbb{Q}$.

Remark 1. Instead of variance minimization in Definition 5.1, Föllmer and Sondermann [13] introduce a stronger optimality condition which is known as *risk-minimization*. However, since the gain processes of the index and CDO tranches are bounded martingales under $\mathbb{Q}$, the two optimality criteria lead to the same optimal hedging strategies.

Appendix B. Derivation of the variance-minimizing hedge ratio. The key to computing the variance-minimizing hedge under a particular model is to express (G_t^{idx}) and $(G_t^{[a,b]})$ as stochastic integrals, use the Ito isometry formula for these stochastic integrals, then solve (23) (see, e.g., [8]). In particular, Frey and Backhaus [14] show the derivation for the *convex counterparty risk model*, and our following computations are similar to those in [14].

B.1. Proof of Proposition 5.2. Let (U_t) be a deterministic counting process which jumps by 1 at payment dates, i.e., $U_{T_m} - U_{T_{m-1}} = 1$ for all $m = 1, \ldots, M$. Using Ito's lemma, the dynamics of the discounted value process under $\mathbb{Q}$ for the index follows

$$d\tilde{V}_{lo}^{idx}(t, N_{t-}) = \frac{\partial}{\partial t}\tilde{V}_{lo}^{idx}(t, N_{t-})dt + \delta_N \tilde{V}_{lo}^{idx}(t, N_{t-})dN_t - B(0,t)P^{idx}(t)dU_t,$$

where $P^{idx}(t)$ is the net payment received at time t. Using Ito's lemma again and from (21), the dynamics of the gain process under $\mathbb{Q}$ for the index follows

$$\begin{aligned} dG_{lo}^{idx}(t) &= B(0,t)P^{idx}(t)dU_t + d\tilde{V}_{lo}^{idx}(t, N_{t-}) \\ &= \frac{\partial}{\partial t}\tilde{V}_{lo}^{idx}(t, N_{t-})dt + \delta_N \tilde{V}_{lo}^{idx}(t, N_{t-})dN_t \\ &= \left(\frac{\partial}{\partial t}\tilde{V}_{lo}^{idx}(t, N_{t-}) + \delta_N \tilde{V}_{lo}^{idx}(t, N_{t-})\lambda_t\right) dt + \delta_N \tilde{V}_{lo}^{idx}(t, N_{t-})dN_t^c \\ &= \delta_N \tilde{V}_{lo}^{idx}(t, N_{t-})dN_t^c, \end{aligned}$$

where $(N_t^c) = (N_t - \int_0^t \lambda_s ds)$ is the compensated version of (N_t). The last equality comes from the fact that $(G_{lo}^{idx}(t))$ is a martingale under $\mathbb{Q}$. Similarly, the gain process of a tranche $[a,b]$ under $\mathbb{Q}$ follows

$$dG_{lo}^{[a,b]}(t) = \delta_N \tilde{V}_{lo}^{[a,b]}(t, N_{t-})dN_t^c.$$

We now compute the compensators involved in (23) and obtain

$$\begin{aligned} \phi_t &= \frac{\delta_N \tilde{V}_{lo}^{[a,b]}(t, N_{t-})\delta_N \tilde{V}_{lo}^{idx}(t, N_{t-})\lambda_t}{(\delta_N \tilde{V}_{lo}^{idx}(t, N_{t-}))^2\lambda_t} \\ &= \frac{B(0,t)\delta_N V_{lo}^{[a,b]}(t, N_{t-})}{B(0,t)\delta_N V_{lo}^{idx}(t, N_{t-})} \\ &= \frac{V_{lo}^{[a,b]}(t, N_{t-}+1) - V_{lo}^{[a,b]}(t, N_{t-})}{V_{lo}^{idx}(t, N_{t-}+1) - V_{lo}^{idx}(t, N_{t-})}. \end{aligned}$$

Then, we want to show that the hedging portfolio obtained by implementing the above strategy has finite variance, i.e., $E^{\mathbb{Q}}\left[\int_0^T \phi_t dG_{lo}^{idx}(t)\right]^2 < \infty$, and it is sufficient to show that $E^{\mathbb{Q}}\left[\int_0^T \phi_t^2 d\langle G_{lo}^{idx}\rangle_t\right] < \infty$. Since all the cash flows of tranche $[a,b]$ are bounded, there exists a $K > 0$ such that $|V_{lo}^{[a,b]}(t,N)| \le K/2$ for all $t \in [0,T]$, $N = 0, \ldots, n$, which implies that $|\delta_N V_{lo}^{[a,b]}(t,N)| \le K$ for all $t \in [0,T]$, $N = 0, \ldots, n$. Consider

$$\begin{aligned} E^{\mathbb{Q}}\left[\int_0^T \phi_t^2 d\langle G_{lo}^{idx}\rangle_t\right] &= E^{\mathbb{Q}}\left[\int_0^T \left(\frac{\delta_N V_{lo}^{[a,b]}(t, N_{t-})}{\delta_N V_{lo}^{idx}(t, N_{t-})}\right)^2 \left(\delta_N \tilde{V}_{lo}^{idx}(t, N_{t-})\right)^2 \lambda_t dt\right] \\ &\le E^{\mathbb{Q}}\left[\int_0^T \left(\delta_N V_{lo}^{[a,b]}(t, N_{t-})\right)^2 \lambda_t dt\right] \\ &\le K^2 E^{\mathbb{Q}}\left[\int_0^T \lambda_t dt\right] \\ &\le K^2 n < \infty, \end{aligned}$$

which gives the desired result.

Note that if we implement this hedging strategy, we have

$$dG_{lo}^{[a,b]}(t) = \phi_t dG_{lo}^{idx}(t).$$

The tranche is perfectly hedged in this case, which is consistent with the results of [17].

B.2. Proof of Proposition 5.3. Defining the deterministic counting process (U_t) as in Appendix B.1 and applying Ito's lemma, the dynamic of the discounted value process for the index becomes

$$\begin{aligned} d\tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-}) &= \frac{\partial}{\partial t}\tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-})dt + \frac{\partial}{\partial Y}\tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-})dY_t \\ &\quad + \frac{1}{2}\frac{\partial^2}{\partial Y^2}\tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-})\sigma^2 dt + \delta_N \tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-})dN_t - B(0,t)P^{idx}(t)dU_t, \end{aligned}$$

where $P^{idx}(t)$ is the net payment received at time t. Then, using Ito's lemma again and from (21), the dynamics of the gain process for the index follows

$$\begin{aligned} dG_{bi}^{idx}(t) &= B(0,t)P^{idx}(t)dU_t + d\tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-}) \\ &= \frac{\partial}{\partial t}\tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-})dt + \frac{\partial}{\partial Y}\tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-})dY_t \\ &\quad + \frac{1}{2}\frac{\partial^2}{\partial Y^2}\tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-})\sigma^2 dt + \delta_N \tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-})dN_t \\ &= \Bigg(\frac{\partial}{\partial t}\tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-}) - \frac{\partial}{\partial Y}\tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-})\,\kappa Y_t \\ &\qquad + \frac{1}{2}\frac{\partial^2}{\partial Y^2}\tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-})\,\sigma^2 + \delta_N \tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-})\,\lambda_t \Bigg) dt \\ &\quad + \frac{\partial}{\partial Y}\tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-})\,\sigma dW_t + \delta_N \tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-})dN_t^c \\ &= \frac{\partial}{\partial Y}\tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-})\,\sigma dW_t + \delta_N \tilde{V}_{bi}^{idx}(t, Y_{t-}, N_{t-})dN_t^c. \end{aligned}$$

Similarly, the gain process of a tranche $[a,b]$ follows

$$dG_{bi}^{[a,b]}(t) = \frac{\partial}{\partial Y}\tilde{V}_{bi}^{[a,b]}(t, Y_{t-}, N_{t-})\,\sigma dW_t + \delta_N \tilde{V}_{bi}^{[a,b]}(t, Y_{t-}, N_{t-})dN_t^c.$$

We can now compute the compensators of $(G_{bi}^{[a,b]}(t))$ and $(G_{bi}^{idx}(t))$ and use (23) to compute (ϕ_t) which gives (19).

Let us now show that the hedging portfolio based on (19) has finite variance. Consider the following:

$$
\begin{aligned}
&E^{\mathbb{Q}}\left[\int_0^T \phi_t d\langle G_{bi}^{idx}\rangle_t\right]\\
&= E^{\mathbb{Q}}\left[\int_0^T \frac{\left[\partial_Y \tilde{V}_{bi}^{[a,b]}(t,Y_{t-},N_{t-})\,\partial_Y \tilde{V}_{bi}^{idx}(t,Y_{t-},N_{t-})\,\sigma^2 + \delta_N \tilde{V}_{bi}^{[a,b]}(t,Y_{t-},N_{t-})\delta \tilde{V}_{bi}^{idx}(t,Y_{t-},N_{t-})\lambda_t\right]^2}{(\partial_Y \tilde{V}_{bi}^{idx}(t,Y_{t-},N_{t-})\,\sigma)^2 + (\delta_N \tilde{V}_{bi}^{idx}(t,Y_{t-},N_{t-}))^2\,\lambda_t} dt\right]\\
&\le E^{\mathbb{Q}}\left[\int_0^T \frac{\left[\partial_Y V_{bi}^{[a,b]}(t,Y_{t-},N_{t-})\,\partial_Y V_{bi}^{idx}(t,Y_{t-},N_{t-})\,\sigma^2 + \delta_N V_{bi}^{[a,b]}(t,Y_{t-},N_{t-})\delta_N V_{bi}^{idx}(t,Y_{t-},N_{t-})\lambda_t\right]^2}{(\partial_Y V_{bi}^{idx}(t,Y_{t-},N_{t-})\,\sigma)^2 + (\delta_N V_{bi}^{idx}(t,Y_{t-},N_{t-}))^2\,\lambda_t} dt\right]\\
&\le 2E^{\mathbb{Q}}\left[\int_0^T \frac{\left[\partial_Y V_{bi}^{[a,b]}(t,Y_{t-},N_{t-})\,\partial_Y V_{bi}^{idx}(t,Y_{t-},N_{t-})\,\sigma^2\right]^2}{(\partial_Y V_{bi}^{idx}(t,Y_{t-},N_{t-})\,\sigma)^2 + (\delta_N V_{bi}^{idx}(t,Y_{t-},N_{t-}))^2\,\lambda_t} dt\right]\\
&\quad + 2E^{\mathbb{Q}}\left[\int_0^T \frac{\left[\delta_N V_{bi}^{[a,b]}(t,Y_{t-},N_{t-})\delta_N V_{bi}^{idx}(t,Y_{t-},N_{t-})\lambda_t\right]^2}{(\partial_Y V_{bi}^{idx}(t,Y_{t-},N_{t-})\,\sigma)^2 + (\delta_N V_{bi}^{idx}(t,Y_{t-},N_{t-}))^2\,\lambda_t} dt\right].
\end{aligned}
$$

Consider the first term. For a fixed time t, the integrand is equal to zero if $\partial_Y V_{bi}^{idx}(t,Y_{t-},N_{t-}) = 0$. If $\partial_Y V_{bi}^{idx}(t,Y_{t-},N_{t-}) \neq 0$, we have

$$
\frac{\left[\partial_Y V_{bi}^{[a,b]}(t,Y_{t-},N_{t-})\,\partial_Y V_{bi}^{idx}(t,Y_{t-},N_{t-})\,\sigma^2\right]^2}{(\partial_Y V_{bi}^{idx}(t,Y_{t-},N_{t-})\,\sigma)^2 + (\delta_N V_{bi}^{idx}(t,Y_{t-},N_{t-}))^2\,\lambda_t} \le \left[\partial_Y V_{bi}^{[a,b]}(t,Y_{t-},N_{t-})\,\sigma\right]^2.
$$

For the second expectation, for a fixed time t, the integrand is equal to zero if $\delta_N V_{bi}^{idx}(t,Y_{t-},N_{t-})\lambda_t = 0$. Otherwise, we have

$$
\frac{\left[\delta_N V_{bi}^{[a,b]}(t,Y_{t-},N_{t-})\delta_N V_{bi}^{idx}(t,Y_{t-},N_{t-})\lambda_t\right]^2}{(\partial_Y V_{bi}^{idx}(t,Y_{t-},N_{t-})\,\sigma)^2 + (\delta_N V_{bi}^{idx}(t,Y_{t-},N_{t-}))^2\,\lambda_t} \le \left(\delta_N V_{bi}^{[a,b]}(t,Y_{t-},N_{t-})\right)^2 \lambda_t.
$$

Using the fact that $\delta_N V_{bi}^{[a,b]}(t,Y,N)$ is bounded, we obtain

$$
\begin{aligned}
E^{\mathbb{Q}}\left[\int_0^T \phi_t d\langle G_{bi}^{idx}\rangle_t\right] &\le 2E^{\mathbb{Q}}\left[\int_0^T \left[\partial_Y V_{bi}^{[a,b]}(t,Y_{t-},N_{t-})\,\sigma\right]^2 + \left(\delta_N V_{bi}^{[a,b]}(t,Y_{t-},N_{t-})\right)^2 \lambda_t dt\right]\\
&\le 2\sigma^2 E^{\mathbb{Q}}\left[\int_0^T \left(\partial_Y V_{bi}^{[a,b]}(t,Y_{t-},N_{t-})\right)^2 dt\right] + 2K^2 n < \infty.
\end{aligned}
$$

Therefore, $E^{\mathbb{Q}}\left[\int_0^T \phi_t dG_{bi}^{idx}(t)\right]^2 < \infty$.

Acknowledgments. We thank Pierre Collin-Dufresne, Rüdiger Frey, Alexander Herbertsson, Jean-Paul Laurent, and the other seminar participants at University of Evry Credit Risk Conference 2008, the Bachelier Congress 2008, INFORMS 2008 Financial Services Section, SIAM Conference on Financial Mathematics and Engineering 2008, the Courant Institute Mathematical Finance Seminar 2009, the Conference on Theory and Practice of Credit Derivatives (Nice 2009), the Bachelier Seminar (Paris), and the S&P Credit Risk Summit (2009) for helpful remarks.

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] T. Bielecki, S. Crépey, and M. Jeanblanc, *Up and down credit risk*, Quant. Finance, 10 (2010), pp. 1137–1151.

[3] T. Bielecki, M. Jeanblanc, and M. Rutkowski, *Pricing and trading credit default swaps in a hazard process model*, Ann. Appl. Probab., 18 (2008), pp. 2495–2529.

[4] D. Brigo and M. Morini, *Arbitrage-free Pricing of Credit Index Options. The No-Armageddon-Pricing Measure and the Role of Correlation after the Subprime Crisis*, SSRN working paper, 2008.

[5] R. Cont, *Model uncertainty and its impact on the pricing of derivative instruments*, Math. Finance, 16 (2006), pp. 519–547.

[6] R. Cont, R. Deguest, and Y. H. Kan, *Default intensities implied by CDO spreads: Inversion formula and model calibration*, SIAM J. Financial Math., 1 (2010), pp. 555–585.

[7] R. Cont and A. Minca, *Recovering Portfolio Default Intensities Implied by CDO Tranches*, Financial Engineering Report 2008-01, Columbia University, New York, 2008.

[8] R. Cont, P. Tankov, and E. Voltchkova, *Hedging with options in presence of jumps*, in Stochastic Analysis and Applications: The Abel Symposium 2005 in Honor of Kiyosi Ito, F. Benth, G. Di Nunno, T. Lindstrom, B. Oksendal, and T. Zhang, eds., Springer, New York, 2007, pp. 197–218.

[9] D. Duffie and N. Gârleanu, *Risk and valuation of collateralized debt obligations*, Financial Anal. J., 57 (2001), pp. 41–59.

[10] P. Durand and J.-F. Jouanin, *Some short elements on hedging credit derivatives*, ESAIM Probab. Statist., 11 (2007), pp. 23–34.

[11] A. Eckner, *Computational techniques for basic affine models of portfolio credit risk*, J. Comput. Finance, 13 (2009), pp. 63–97.

[12] E. Errais, K. Giesecke, and L. R. Goldberg, *Affine point processes and portfolio credit risk*, SIAM J. Finan. Math., 1 (2010), pp. 642–665.

[13] H. Föllmer and D. Sondermann, *Hedging of nonredundant contingent claims*, in Contributions to Mathematical Economics, North–Holland, Amsterdam, 1986, pp. 205–223.

[14] R. Frey and J. Backhaus, *Dynamic hedging of synthetic CDO tranches with spread and contagion risk*, J. Econom. Dynam. Control, 34 (2010), pp. 710–724.

[15] K. Giesecke, L. Goldberg, and X. Ding, *A top-down approach to multi-name credit*, Oper. Res., 59 (2011), to appear.

[16] A. Herbertsson, *Pricing synthetic CDO tranches in a model with default contagion using the matrix-analytic approach*, J. Credit Risk, 4 (2008), pp. 3–35.

[17] J.-P. Laurent, A. Cousin, and J.-D. Fermanian, *Hedging Default Risks of CDOs in Markovian Contagion Models*, working paper, 2008.

[18] D. Li, *On default correlation: A copula function approach*, J. Fixed Income, 9 (2000), pp. 43–54.

[19] D. Li, *Base correlation*, in Encyclopedia of Quantitative Finance, R. Cont, ed., Wiley, New York, 2010.

[20] A. Lipton and A. Rennie, eds., *The Oxford Handbook of Credit Derivatives*, Oxford University Press, London, 2011.

[21] 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, Amsterdam, 2008, pp. 243–274.

[22] S. Morgan and A. Mortensen, *CDO Hedging Anomalies in the Base Correlation Approach*, Technical report, 2007.

[23] T. Møller, *Risk-minimizing hedging strategies for insurance payment process*, Finance Stoch., 5 (2001), pp. 419–446.

[24] M. Neugebauer, J. Carter, R. Hrvatin, T. Cunningham, and R. Hardee, *Understanding and Hedging Risks in Synthetic CDO Tranches*, Technical report, Fitch Ratings, New York, 2006.

[25] 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.

SIAM J. FINANCIAL MATH.
Vol. 2, pp. 141–182

Robust Hedging of Double Touch Barrier Options*

A. M. G. Cox† and Jan Obłój‡

Abstract. We consider robust pricing of digital options, which pay out if the underlying asset has crossed both upper and lower barriers. We make only weak assumptions about the underlying process (typically continuity), but assume that the initial prices of call options with the same maturity and all strikes are known. In such circumstances, we are able to give upper and lower bounds on the arbitrage-free prices of the relevant options and show that these bounds are tight. Moreover, pathwise inequalities are derived, which provide the trading strategies with which we are able to realize any potential arbitrages. These super- and subhedging strategies have a simple quasi-static structure, their associated hedging error is bounded below, and in practice they carry low transaction costs. We show that, depending on the risk aversion of the investor, they can outperform significantly the standard delta/vega-hedging in presence of market frictions and/or model misspecification. We make use of embeddings techniques; in particular, we develop two new solutions to the (optimal) Skorokhod embedding problem.

Key words. double barrier option, robust hedging, no-arbitrage pricing, Skorokhod embedding, risk neutral distribution, superhedging, subhedging

AMS subject classifications. Primary, 91B28, 91B70; Secondary, 60G40, 60G44

DOI. 10.1137/090777487

1. Introduction.

In the standard approach to pricing and hedging, one postulates a model for the underlying asset, calibrates it to the market prices of liquidly traded vanilla options, and then uses the model to derive prices and associated hedges for exotic over-the-counter products. Prices and hedges will be correct only if the model describes the real world perfectly, which is unlikely. The robust (model-independent) approach uses market data to deduce bounds on the prices consistent with *no-arbitrage* and the associated super- and subreplicating strategies, which are robust to model misspecification. More precisely, we start with quoted prices of some liquid options and assume that this market input is consistent with *no-arbitrage*. Then we want to answer two questions. First, for a given exotic option, what is the range of prices that we can charge for it without introducing a model-independent arbitrage? Second, if we see a price outside this range, how do we exploit it to make a riskless profit? In this paper we adopt such an approach to pricing and hedging for digital double barrier options.

The general methodology, which we now outline, is based on solving the *Skorokhod embedding problem* (SEP). We assume *no-arbitrage* and suppose that we know the market prices of calls and puts for all strikes at one maturity T. We are interested in pricing an exotic option

*Received by the editors November 17, 2009; accepted for publication (in revised form) September 27, 2010; published electronically February 3, 2011. This research was partially supported by a Marie Curie Intra-European Fellowship at Imperial College London within the 6th European Community Framework Programme.
http://www.siam.org/journals/sifin/2/77748.html

†Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK (A.M.G.Cox@bath.ac.uk).
‡Mathematical Institute and Oxford-Man Institute of Quantitative Finance, University of Oxford, Oxford OX1 3LB, UK (obloj@maths.ox.ac.uk).

with payoff given by a path-dependent functional $O(S)_T$. The example we consider here is a *digital double touch* barrier option struck at $(\underline{b}, \overline{b})$ which pays 1 if the stock price reaches both $\underline{b}$ and $\overline{b}$ before maturity T. Our aim is to construct a robust superreplicating strategy of the form

$$O(S)_T \leq F(S_T) + N_T, \tag{1.1}$$

where $F(S_T)$ is the payoff of a finite portfolio of European puts and calls quoted at time zero and N_T are gains from a self-financing trading strategy (typically forward transactions). Furthermore, we want (1.1) to be tight in the sense that we can construct a market model which matches the market prices of calls and puts and in which we have equality in (1.1). The initial price of the portfolio $F(S_T)$ is then the least upper bound on the price of the exotic $O(S)_T$, and the right-hand side (RHS) of (1.1) gives a simple superreplicating strategy at that cost. There is an analogous argument for the lower bound and an analogous subreplicating strategy. We stress that the RHS in (1.1) makes sense as a model-independent superhedge. It requires an initial capital, the price of $F(S_T)$, which is uniquely specified by the prices quoted in the market, and the rest is carried out in a self-financing way. Typically, for any specific payoff $O(S)_T$, one will be able to come up with a variety of random variables X which satisfy $O(S)_T \leq X \leq F(S_T) + N_T$, and hence, in some market models, X may be cheaper than $F(S_T)$. However, such X has no interpretation as a model-independent superreplicating strategy. Indeed, if X is a valid model-independent superreplicating strategy, it has a uniquely specified price at time $t = 0$ from the market quoted prices. This price is independent of the market model and hence, since we required (1.1) to be tight, is equal to the price of $F(S_T)$, as in the extreme model both are equal to the price of $O(S)_T$.

In fact, in order to construct (1.1), we first construct the market model which induces the upper bound on the price of $O(S)_T$ and hence will attain equality in (1.1). For this construction we rely on the theory of Skorokhod embeddings (cf. Obłój [34]). We assume *no-arbitrage* and consider a market model in the risk-neutral measure so that the forward price process $(S_t : t \leq T)$ is a martingale.[1] It follows from the work of Monroe [32] that $S_t = B_{\rho_t}$, for a Brownian motion (B_t) with $B_0 = S_0$ and some increasing sequence of stopping times $\{\rho_t : t \leq T\}$ (possibly relative to an enlarged filtration). Let us further assume that the payoff of the exotic option is invariant under the time change: $O(S)_T = O(B)_{\rho_T}$ a.s. Knowing the market prices of calls and puts for all strikes at maturity T is equivalent to knowing the distribution μ of S_T (cf. Breeden and Litzenberger [6]). Thus, we can see the stopping time $\rho = \rho_T$ as a solution to the SEP for μ. Conversely, let τ be a solution to the SEP for μ; i.e., $B_\tau \sim \mu$ and $(B_{t\wedge\tau} : t \geq 0)$ is a uniformly integrable martingale. Then the process $\tilde{S}_t := B_{\tau \wedge \frac{t}{T-t}}$ is a model for the stock-price process consistent with the observed prices of calls and puts at maturity T. In this way, we obtain a correspondence which allows us to identify market models with solutions to the SEP and vice versa. In consequence, to estimate the fair price of the exotic option $\mathbb{E}O(S)_T$, it suffices to bound $\mathbb{E}O(B)_\tau$ among all solutions τ to the

[1] Equivalently, under a simplifying assumption of zero interest rates, S_t is simply the stock price process. See section 3.2 for further discussion.

SEP. More precisely, we have

$$\inf_{\tau: B_\tau \sim \mu} \mathbb{E}O(B)_\tau \leq \mathbb{E}O(S)_T \leq \sup_{\tau: B_\tau \sim \mu} \mathbb{E}O(B)_\tau, \tag{1.2}$$

where all stopping times τ are such that $(B_{t\wedge\tau})_{t\geq 0}$ is uniformly integrable. Once we compute the above bounds and the stopping times which achieve them, we usually have a good intuition into how to construct the super- (and sub-)replicating strategies (1.1).

A more detailed description of the SEP-driven methodology outlined above can be found in Hobson [27] or in Obłój [35]. The idea of *no-arbitrage* bounds on prices goes back to Merton [31], and a recent survey of the literature can be found in Cox [12]. The methods for robust pricing and hedging of options sketched above go back to the works of Hobson [28] (lookback option) and Brown, Hobson, and Rogers [8] (single barrier options). More recently, Dupire [20] investigated volatility derivatives using the SEP, and Cox, Hobson, and Obłój [13] designed pathwise inequalities to derive price range and robust superreplicating strategies for derivatives paying a convex function of the local time.

Unlike in previous works, e.g., [8], we don't find a unique inequality (1.1) for a given barrier option. Instead we find that, depending on the market input (i.e., prices of calls and puts) and the pair of barriers, different strategies may be optimal. We characterize all of them and give precise conditions for deciding which one should be used. This new difficulty is coming from the dependence of the payoff on both the running maximum and minimum of the process. Solutions to the SEP which maximize or minimize $\mathbb{P}(\sup_{u\leq\tau} B_u \geq \overline{b}, \inf_{u\leq\tau} B_u \leq \underline{b})$ have not been developed previously and are introduced in this paper. These are new probabilistic results of independent interest which we derive here as tools to study our financial problem. As one might suspect, our new solutions are considerably more involved than those by Perkins [36] or Azéma and Yor [2] exploited by Brown, Hobson, and Rogers [8].

From a practical point of view, the no-arbitrage price bounds which we obtain are too wide to be used for pricing. However, our super- or subhedging strategies can still be used. Specifically, suppose an agent sells a double touch barrier option $O(S)_T$ for a premium p. She can then set up our superhedge (1.1) for an initial premium $\overline{p} > p$. At maturity T she holds $H = -O(S)_T + F(S)_T + N_T + p - \overline{p}$, which on average is worth zero, $\mathbb{E}H = 0$, but is also bounded below: $H \geq p - \overline{p}$. In reality, in the presence of model uncertainty and market frictions, this can be an appealing alternative to the standard delta/vega-hedging. Indeed, our numerical simulations in section 3.3 suggest that in the presence of transaction costs a risk averse agent will generally prefer the hedging strategy we construct to a (monitored daily) delta/vega-hedge.

The paper is structured as follows. First we present the setup: our assumptions and terminology and the types of double barriers considered in this paper. Then in section 2 we consider digital double touch barrier options introduced above. We first present super- and subreplicating strategies and then prove in section 2.3 that they induce tight robust bounds on the admissible prices of the double touch options. In section 3 we reconsider our assumptions and investigate some applications. Specifically, in section 3.1 we consider the case when calls and puts with only a finite number of strikes are observed, and in section 3.2 we discuss discontinuities in the price process (S_t), nonzero interest rates, and further additions to the set of available market quoted prices. In section 3.3 we present a numerical investigation of the performance of our super- and subhedging strategies. Section 4 contains the proofs of main theorems. In particular, it contains new solutions to the SEP which are necessary to prove results in section 2.3.

1.1. Setup. In what follows, $(S_t)_{t\geq 0}$ is the forward price process. Equivalently, we can think of the underlying asset with zero interest rates, or an asset with zero cost of carry. In particular, our results can be directly applied in Foreign Exchange markets for currency pairs from economies with equal interest rates. Moving to the spot market with nonzero interest rates is not immediate as our barriers become time-dependent; see section 3.2.

We assume that $(S_t)_{t\geq 0}$ has continuous paths. We comment in section 3.2 that this assumption can be removed or weakened to a requirement that given barriers are crossed continuously. We fix a maturity $T > 0$ and assume that we observe the initial spot price S_0 and the market prices of European calls for all strikes $K > 0$ and maturity T,

$$\Big(C(K) : K \geq 0\Big), \tag{1.3}$$

which we call the *market input*. For simplicity we assume that $C(K)$ is twice differentiable and strictly convex on $(0, \infty)$. Further, we assume that we can enter a forward transaction at no cost. More precisely, let ρ be a stopping time relative to the natural filtration of $(S_t)_{t\leq T}$ such that $S_\rho = \overline{b}$. Then the portfolio corresponding to selling a forward at time ρ has final payoff $(\overline{b} - S_T)\mathbf{1}_{\rho\leq T}$, and we assume its initial price is zero. The initial price of a portfolio with a constant payoff K is K. We denote by $\mathcal{X}$ the set of all calls, forward transactions, and constants, and $\mathrm{Lin}(\mathcal{X})$ is the space of their finite linear combinations, which is precisely the set of portfolios with given initial market prices. For convenience we introduce a pricing operator $\mathcal{P}$ which, to a portfolio with payoff X at maturity T, associates its initial (time zero) price, e.g., $\mathcal{P}K = K$, $\mathcal{P}(S_T - K)^+ = C(K)$, and $\mathcal{P}(\overline{b} - S_T)\mathbf{1}_{\rho\leq T} = 0$. We also assume that $\mathcal{P}$ is linear, whenever defined. Initially, $\mathcal{P}$ is given only on $\mathrm{Lin}(\mathcal{X})$. One of the aims of the paper is to understand extensions of $\mathcal{P}$ which do not introduce arbitrage to $\mathrm{Lin}(\mathcal{X} \cup \{Y\})$, for double touch barrier derivatives Y. Note that linearity of $\mathcal{P}$ on $\mathrm{Lin}(\mathcal{X})$ implies that call-put parity holds, and in consequence we also know the market prices of all European put options with maturity T:

$$P(K) := \mathcal{P}(K - S_T)^+ = K - S_0 + C(K).$$

Finally, we assume that the market admits no *model-independent arbitrage* in the sense that any portfolio of initially traded assets with a nonnegative payoff has a nonnegative price:

$$\forall X \in \mathrm{Lin}(\mathcal{X}) : X \geq 0 \Longrightarrow \mathcal{P}X \geq 0. \tag{1.4}$$

As we do not yet have any probability measure, by $X \geq 0$ we mean that the payoff is nonnegative for any continuous nonnegative stock price path $(S_t)_{t\leq T}$.

By a *market model* we mean a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ with a continuous $\mathbb{P}$-martingale (S_t) which matches the market input (1.3). Note that we consider the model under the risk-neutral measure, and the pricing operator is then just the expectation $\mathcal{P} = \mathbb{E}$. Saying that (S_t) matches the market input is equivalent to saying that it starts in the initial spot S_0 a.s. and that $\mathbb{E}(S_T - K)^+ = C(K)$, $K > 0$. This in turn is equivalent to knowing the distribution of S_T (cf. [6, 8]). We denote this distribution by μ and often refer to it as the *law of S_T implied by the call prices*. Our regularity assumptions on $C(K)$ imply that

$$\mu(\mathrm{d}K) = C''(K)\mathrm{d}K, \quad K > 0, \tag{1.5}$$

so that μ has a positive density on $(0,\infty)$. We could relax this assumption and take the support of μ to be any interval $[a, b]$. Introducing atoms would complicate our formulae (essentially without introducing new difficulties).

The running maximum and minimum of the price process are denoted respectively by $\overline{S}_t = \sup_{u\le t} S_u$ and $\underline{S}_t = \inf_{u\le t} S_u$. We are interested in this paper in derivatives whose payoff depends on both $\overline{S}_T$ and $\underline{S}_T$. It is often convenient to express events involving the running maximum and minimum in terms of the first hitting times $H_x = \inf\{t : S_t = x\}$, $x \ge 0$. As an example, note that $\mathbf{1}_{\overline{S}_T\ge\overline{b},\, \underline{S}_T\le\underline{b}} = \mathbf{1}_{H_{\overline{b}}\vee H_{\underline{b}}\le T}$.

We use the notation $a \ll b$ to indicate that a is *much smaller* than b. This is only used to give intuition and is not formal. The minimum and maximum of two numbers are denoted $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$, respectively, and the positive part is denoted by $a^+ = a \vee 0$.

1.2. Connections to other barrier options. The barrier options considered in this paper are fairly specific: we are interested in a double touch option which pays out $\mathbf{1}_{\overline{S}_T\ge\overline{b},\, \underline{S}_T\le\underline{b}}$ at maturity T. It is natural to ask how the problem we consider is connected to similar problems for related barrier options, and also whether the results can be generalized to a wider class of options. One question is: can our results on double barrier options be expressed in terms of the results for single barrier options due to Brown, Hobson, and Rogers [8]? The answer is negative: we are inspired by their paper and we use similar methodology but to solve a different problem, and we cannot apply their results in our setting. More specifically, in [8], the authors develop arbitrage-free bounds on the price of a one-touch digital option (that is, an option which pays out 1 if a given level is crossed before maturity). At first sight one might want to price our double touch option as a sort of compound option which, upon hitting the first barrier, pays out a one-touch option struck at the second barrier. This intuition would work in a model-specific framework, but it breaks down entirely in the model-independent framework that we consider. Specifically, the bounds given in [8] depend on knowing the call prices at the time the option is issued. In our setting, however, we know the call prices initially but make no assumption about how they behave (or even if they are quoted) at intermediate times. In particular we have no a priori information about future call prices at the time the first barrier is hit and so cannot use the bounds derived in [8] for the options we study. On the other hand, we will recover results from [8] as limiting cases of double touch options when one of the barriers degenerates to the spot S_0.

An alternative question that may be asked is: can one use our results to say something about different types of digital barrier options? In this work, we are interested in the option with payoff $\mathbf{1}_{\overline{S}_T\ge\overline{b},\, \underline{S}_T\le\underline{b}}$, but the identity $\mathbf{1}_{\overline{S}_T\ge\overline{b},\, \underline{S}_T\le\underline{b}} = 1 - \mathbf{1}_{\overline{S}_T<\overline{b} \text{ or } \underline{S}_T>\underline{b}}$ immediately allows us to convert a superhedge of $\mathbf{1}_{\overline{S}_T\ge\overline{b},\, \underline{S}_T\le\underline{b}}$ into a subhedge of $\mathbf{1}_{\overline{S}_T<\overline{b} \text{ or } \underline{S}_T>\underline{b}}$, and consequently we can convert an upper bound on the price of $\mathbf{1}_{\overline{S}_T\ge\overline{b},\, \underline{S}_T\le\underline{b}}$ to a lower bound on the price of $\mathbf{1}_{\overline{S}_T<\overline{b} \text{ or } \underline{S}_T>\underline{b}}$. There are also identities which connect the double touch to other double barrier options, for example, $\mathbf{1}_{\overline{S}_T\ge\overline{b},\, \underline{S}_T\le\underline{b}} = \mathbf{1}_{\overline{S}_T\ge\overline{b}} - \mathbf{1}_{\overline{S}_T\ge\overline{b},\, \underline{S}_T>\underline{b}}$. A natural conjecture would then be that an upper bound on the price of $\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}}$ might translate into a lower bound on the price of $\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T > \underline{b}}$. However, in the setup we consider this is not the case since the price of the one-touch option, $\mathbf{1}_{\overline{S}_T \geq \overline{b}}$, is not specified under our assumptions, and the lowest possible price of $\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T > \underline{b}}$ will typically not occur at the same time as the price of the one-touch option is maximized. This situation would alter if we assumed that the one-touch option was traded at a given initial price, in which case the lower bound on the price of $\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}}$ would correspond to an upper bound on $\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T > \underline{b}}$. However, then the additional information given in the price of the one-touch option changes the setup of the initial problem and would, in all likelihood, change the bounds we derive in this paper. See section 3.2 for a further discussion of this point. As a consequence, the results in this paper will not extend to other double barrier options beyond the bounds given by the identity $\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}} = 1 - \mathbf{1}_{\overline{S}_T < \overline{b} \text{ or } \underline{S}_T > \underline{b}}$.

The question of bounds for the double no-touch option $\mathbf{1}_{\overline{S}_T \leq \overline{b},\, \underline{S}_T \geq \underline{b}}$ is considered in Cox and Obłój [14]. In this case, the analysis of the hedges and bounds is relatively straightforward, but the paper focuses much more on subtleties concerning different classes of arbitrage with which we do not concern ourselves in this paper.

1.3. Probabilistic interpretation. The bounds on prices of double touch options developed in Theorems 2.2 and 2.4 correspond, in probabilistic terms, to computing

$$\sup_M \mathbb{P}\left(\sup_t M_t \geq \overline{b},\ \inf_t M_t \leq \underline{b}\right) \quad \text{and} \quad \inf_M \mathbb{P}\left(\sup_t M_t \geq \overline{b},\ \inf_t M_t \leq \underline{b}\right) \tag{1.6}$$

over all uniformly integrable continuous martingales $M = (M_t : 0 \leq t \leq \infty)$ with $M_0 = S_0$ and M_∞ distributed according to μ. To the best of our knowledge, such bounds have not been studied before and hence are of independent interest. As mentioned above, in order to compute these we develop new solutions to the Skorokhod embedding problem. The bounds that we obtain depend in a complex way on μ and $(\underline{b}, \overline{b})$ and are considerably more involved than in the single-sided case $\underline{b} = S_0$, which goes back to Blackwell and Dubins [4] and which was exploited in [8]. More precisely, $\sup_M \mathbb{P}(\sup_t M_t \geq \overline{b})$ is attained by the Azéma–Yor martingale (see Azéma and Yor [3]) simultaneously for all $\overline{b} \geq S_0$. In contrast, the supremum in (1.6) is attained by a different martingale for each pair $(\overline{b}, \underline{b})$. The bounds in (1.6) can be seen as a first step toward studying admissible laws for the triplet $(M_\infty, \sup_t M_t, \inf_t M_t)$, in a similar way as the single-sided case led to studies of admissible laws for $(M_\infty, \sup_t M_t)$ in Rogers [37] and Vallois [40].

2. Robust pricing and hedging. We now investigate robust pricing and hedging of a double touch option which pays 1 if and only if the stock price goes above $\overline{b}$ and below $\underline{b}$ before maturity: $\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}}$. We present simple quasi-static super- and subreplicating strategies which prove to be optimal (i.e., replicating) in some market model.[2] Sometimes, by a slight abuse, we refer to these robust strategies as *model-independent*. This emphasises that they work universally under our setup outlined above and do not depend on specific modeling assumptions.

[2] At first how the strategies below were derived may appear rather mysterious. In fact, as explained in the introduction, we first identified these extreme models and analyzed hedging strategies within these models. This way the super- and subreplicating strategies arise quite naturally.

It seems to us that the hedges are most easily expressed by considering four special cases. Each case will provide a super- or subhedge. We will see, however, that, depending on the values of $\underline{b}, \overline{b}$ relative to S_0, a different one will be the smallest superhedge/largest subhedge. In sections 2.1 and 2.2 we will outline the super- and subhedges, and in section 2.3 we will give criteria that allow us to determine exactly into which case a given set of parameter values falls.

The fact that we have four different hedges is rather intuitive. Imagine a trader who has a long position in a digital double touch barrier option and needs to hedge it in a robust way.[3] Then he is likely to think differently about the option depending on where the barriers are relative to the spot. If one of the barriers is very close to the spot, then he can effectively approximate the double touch with a simple one-touch struck at the other barrier. We will see that for some parameter values this is indeed the case, and the double touch has robust prices and superhedges that are identical to the one-touch. These cases are given below as $\overline{H}^I$ and $\overline{H}^{II}$. When barriers are approximately symmetric around the spot, our rough estimation above becomes too costly, and the trader hedges a genuine double touch option. When the barriers are close to the spot—relative to trader's belief about the volatility of the market (which is here inferred from the quoted call prices)—then it is reasonable to build the hedging strategy around the assumption that at least one of them will be hit. The optimal strategy then is described in our hedge $\overline{H}^{IV}$. On the other hand, if the barriers are far away, there will also be situations when neither barrier is struck, and the strategy has to account for that. This is done in $\overline{H}^{III}$. Finally, an analogous story holds for subhedging strategies.

We note also that there are strong similarities between some of the cases that we separate, to the extent that it is natural to ask, for example: can we express $\overline{H}^{IV}$ as a special case of $\overline{H}^{III}$? To some degree, the answer to this is that we can, with a suitable interpretation of some of the parameter values. However, it does not appear to us that making such a change would simplify the analysis in any way, since the special cases would need to be treated separately in any subsequent analysis anyway; rather, we have chosen to express the different super- and subhedges in the manner that appears to convey the most intuitive picture of the differing possible behavior.

2.1. Superhedging. We present here four superreplicating strategies. All our strategies have the same simple structure: we buy an initial portfolio of calls and puts, and when the stock price reaches $\underline{b}$ or $\overline{b}$ we buy or sell forward contracts. Naturally our goal is not only to write a superreplicating strategy but to write *the smallest superreplicating strategy*, and to do so we have to choose the parameters judiciously. As we will see in section 2.3, for a given pair of barriers $\underline{b}, \overline{b}$ exactly one of the superreplicating strategies will induce a tight bound on the derivative's price. We will provide an explicit criterion determining which strategy to use.

$\overline{H}^I$: *superhedge for* $\underline{b} \ll S_0 < \overline{b}$. We buy α puts with strike $K \in (\underline{b}, \infty)$, and when the stock price reaches $\underline{b}$ we buy β forward contracts; see Figure 1. The values of α, β are chosen so that the final payoff on $(0, K)$, provided that the stock price has reached $\underline{b}$, is constant and equal to 1. One easily computes that $\alpha = \beta = (K - \underline{b})^{-1}$. Formally, the superreplication

[3] This can be due to, e.g., high uncertainty about a market model, an illiquid market, or high transaction costs; see section 3.3 for a detailed discussion.

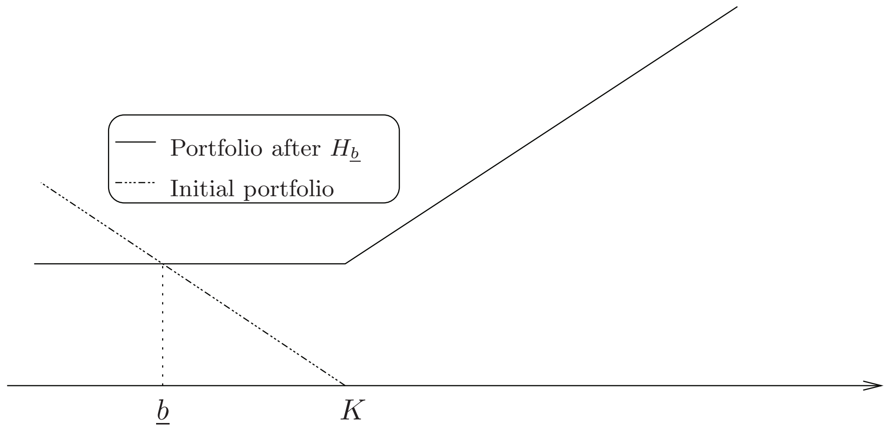


Figure 1. *Superhedge* $\overline{H}^I$.

follows from the following inequality:

$$\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}} \leq \mathbf{1}_{\underline{S}_T \leq \underline{b}} \leq \frac{(K-S_T)^+}{K-\underline{b}} + \frac{S_T-\underline{b}}{K-\underline{b}} \mathbf{1}_{\underline{S}_T \leq \underline{b}} =: \overline{H}^I(K), \tag{2.1}$$

where the last term corresponds to a forward contract entered into, at no cost, when $S_t = \underline{b}$. Note that $\mathbf{1}_{\underline{S}_T \leq \underline{b}} = \mathbf{1}_{H_{\underline{b}} \leq T}$.

$\overline{H}^{II}$*: superhedge for* $\underline{b} < S_0 \ll \overline{b}$. This is a mirror image of $\overline{H}^I$: we buy α calls with strike $K \in (0, \overline{b})$, and when the stock price reaches $\overline{b}$ we sell β forward contracts. The values of α, β are chosen so that the final payoff on (K, ∞), provided that the stock price reaches $\overline{b}$, is constant and equal to 1. One easily computes that $\alpha = \beta = (\overline{b} - K)^{-1}$. Formally, the superreplication follows from the following inequality:

$$\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}} \leq \mathbf{1}_{\overline{S}_T \geq \overline{b}} \leq \frac{(S_T-K)^+}{\overline{b}-K} + \frac{\overline{b}-S_T}{\overline{b}-K} \mathbf{1}_{\overline{S}_T \geq \overline{b}} =: \overline{H}^{II}(K). \tag{2.2}$$

$\overline{H}^{III}$*: superhedge for* $\underline{b} \ll S_0 \ll \overline{b}$. This superhedge involves a static portfolio of four calls and puts and at most four dynamic trades. The choice of parameters is judicious, which makes the strategy the most complex to describe. Choose

$$0 < K_4 < \underline{b} < K_3 < K_2 < \overline{b} < K_1 \tag{2.3}$$

and buy α_i calls with strike K_i, $i = 1, 2$, and α_j puts with strike K_j, $j = 3, 4$. If the stock price reaches $\overline{b}$ without having hit $\underline{b}$ before, that is when $H_{\overline{b}} < H_{\underline{b}} \wedge T$, sell β_1 forward. If $H_{\underline{b}} < H_{\overline{b}} \wedge T$, at $H_{\underline{b}}$ buy β_2 forwards. When the stock price, having hit $\overline{b}$, first reaches $\underline{b}$, that is, at $H_{\underline{b}} \in (H_{\overline{b}}, T]$, buy $\beta_3 = \alpha_3 + \beta_1$ forwards. Finally, at $H_{\overline{b}} \in (H_{\underline{b}}, T]$ sell $\beta_4 = \alpha_2 + \beta_2$ forwards. The choice of β_3 and β_4 is such that the final payoff after hitting $\overline{b}$ and then $\underline{b}$ (resp., $\underline{b}$ and then $\overline{b}$) is constant and equal to 1 on $[K_4, K_3]$ (resp., $[K_2, K_1]$). We now proceed to impose conditions which determine other parameters. A pictorial representation of the superhedge is given in Figure 2.

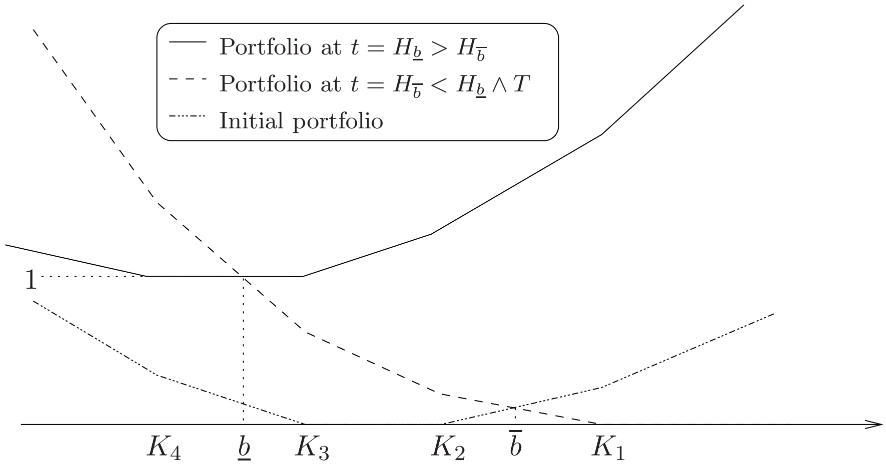


Figure 2. *Superhedge* $\overline{H}^{III}$.

Note that the initial payoff on $[K_3, K_2]$ is zero. After hitting $\overline{b}$ and before hitting $\underline{b}$, the payoff should be zero on $[K_1, \infty)$ and equal to 1 at $\underline{b}$. Likewise, after hitting $\underline{b}$ and before hitting $\overline{b}$, the payoff should be zero on $[0, K_4]$ and equal to 1 at $\overline{b}$. This yields six equations:

$$\begin{cases} \alpha_1 + \alpha_2 - \beta_1 = 0, \\ \alpha_2(K_1 - K_2) - \beta_1(K_1 - \overline{b}) = 0, \\ \alpha_3(K_3 - \underline{b}) - \beta_1(\underline{b} - \overline{b}) = 1, \end{cases} \qquad \begin{cases} \alpha_3 + \alpha_4 - \beta_2 = 0, \\ \alpha_3(K_3 - K_4) + \beta_2(K_4 - \underline{b}) = 0, \\ \alpha_2(\overline{b} - K_2) + \beta_2(\overline{b} - \underline{b}) = 1. \end{cases} \tag{2.4}$$

The superhedging strategy corresponds to an almost-sure inequality

$$\begin{aligned} \mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}} \leq{} & \alpha_1(S_T - K_1)^+ + \alpha_2(S_T - K_2)^+ + \alpha_3(K_3 - S_T)^+ + \alpha_4(K_4 - S_T)^+ \\ & - \beta_1(S_T - \overline{b})\mathbf{1}_{H_{\overline{b}} < H_{\underline{b}} \wedge T} + \beta_2(S_T - \underline{b})\mathbf{1}_{H_{\underline{b}} < H_{\overline{b}} \wedge T} \\ & + \beta_3(S_T - \underline{b})\mathbf{1}_{H_{\overline{b}} < H_{\underline{b}} \leq T} - \beta_4(S_T - \overline{b})\mathbf{1}_{H_{\underline{b}} < H_{\overline{b}} \leq T} \\ & =: \overline{H}^{III}(K_1, K_2, K_3, K_4), \end{aligned} \tag{2.5}$$

where the parameters, after solving (2.4), are given by

$$\begin{cases} \alpha_3 = \frac{(K_1-K_2)(\underline{b}-K_4)(\overline{b}-\underline{b}) - (K_1-\overline{b})(\overline{b}-K_2)(\underline{b}-K_4)}{(K_1-K_2)(K_3-K_4)(\overline{b}-\underline{b})^2 - (K_3-\underline{b})(K_1-\overline{b})(\overline{b}-K_2)(\underline{b}-K_4)}, \\ \alpha_1 = \left(1 - \alpha_3 \frac{K_3-K_4}{\underline{b}-K_4}(\overline{b} - \underline{b})\right)(K_1 - \overline{b})^{-1}, \\ \alpha_2 = \left(1 - \alpha_3 \frac{K_3-K_4}{\underline{b}-K_4}(\overline{b} - \underline{b})\right)(\overline{b} - K_2)^{-1}, \\ \alpha_4 = \frac{K_3-\underline{b}}{\underline{b}-K_4}\alpha_3, \end{cases} \qquad \begin{cases} \beta_1 = \alpha_1 + \alpha_2, \\ \beta_2 = \alpha_3 + \alpha_4. \end{cases} \tag{2.6}$$

Using (2.3), one can verify that α_3 and α_1 are nonnegative and thus also α_2 and α_4 and all $\beta_1, \dots, \beta_4$.

$\overline{H}^{IV}$*: superhedge for* $\underline{b} < S_0 < \overline{b}$. Choose $0 < K_2 < \underline{b} < S_0 < \overline{b} < K_1$. The initial portfolio is composed of α_1 calls with strike K_1, α_2 puts with strike K_2, α_3 forward contracts, and α_4 in cash. If we hit $\overline{b}$ before hitting $\underline{b}$, we sell β_1 forwards, and if we hit $\underline{b}$ before hitting $\overline{b}$, we

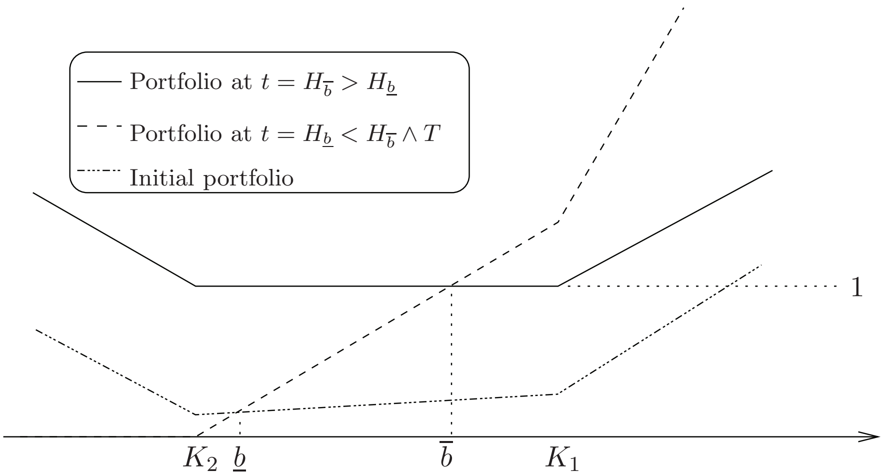


Figure 3. *Superhedge* $\overline{H}^{IV}$.

buy β_2 forwards. The payoff of the portfolio should be zero on $[K_1, \infty)$ (resp., $[0, K_2]$) and equal to 1 at $\underline{b}$ (resp., $\overline{b}$) in the first (resp., second) case. Finally, when we first hit $\underline{b}$ after having hit $\overline{b}$, we buy β_3 forwards, and when we first hit $\overline{b}$ having previously hit $\underline{b}$, we sell β_4 forwards. In both cases the final payoff should then be equal to 1 on $[K_2, K_1]$; see Figure 3. The superhedging strategy corresponds to the following almost-sure inequality:

$$\begin{aligned} \mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}} \leq\ & \alpha_1 (S_T - K_1)^+ + \alpha_2 (K_2 - S_T)^+ + \alpha_3 (S_T - S_0) + \alpha_4 \\ & - \beta_1 (S_T - \overline{b}) \mathbf{1}_{H_{\overline{b}} < H_{\underline{b}} \wedge T} + \beta_2 (S_T - \underline{b}) \mathbf{1}_{H_{\underline{b}} < H_{\overline{b}} \wedge T} \\ & + \beta_3 (S_T - \underline{b}) \mathbf{1}_{H_{\overline{b}} < H_{\underline{b}} \leq T} - \beta_4 (S_T - \overline{b}) \mathbf{1}_{H_{\underline{b}} < H_{\overline{b}} \leq T} \\ & =: \overline{H}^{IV}(K_1, K_2), \end{aligned} \tag{2.7}$$

where, working out the conditions on α_i, β_i, the parameters are

$$\begin{cases} \alpha_1 = 1/(K_1 - \underline{b}), \\ \alpha_2 = 1/(\overline{b} - K_2), \\ \alpha_3 = \frac{(K_1 - \underline{b}) - (\overline{b} - K_2)}{(K_1 - \underline{b})(\overline{b} - K_2)}, \\ \alpha_4 = \frac{\underline{b}\overline{b} - K_1 K_2}{(K_1 - \underline{b})(\overline{b} - K_2)} + \alpha_3 S_0, \end{cases} \qquad \begin{cases} \beta_1 = \alpha_1 + \alpha_3 = 1/(\overline{b} - K_2), \\ \beta_2 = \alpha_2 - \alpha_3 = 1/(K_1 - \underline{b}), \\ \beta_3 = \alpha_1 = 1/(K_1 - \underline{b}), \\ \beta_4 = \alpha_2 = 1/(\overline{b} - K_2). \end{cases} \tag{2.8}$$

As we highlighted at the beginning of the section, all our superhedges have the same general structure: they consist of an initial portfolio of cash, puts, and calls and then involve some forward transactions. We presented above four distinct strategies of this type, and one could ask: it is possible to unify them into one general parametric strategy? It is not too difficult to see that the inequalities (2.1), (2.2), and (2.7) can, with the correct modifications of the parameters, be rewritten in the form (2.5); however, in general the relationships in (2.4) and (2.3) will not hold (for (2.7), one needs to take $K_3 = K_1 > \overline{b}$ and $K_2 = K_4 < \underline{b}$).

However, if we suppose simply that

$$K_4 < \underline{b}, \quad K_3 > \underline{b}, \quad K_2 < \overline{b}, \quad K_1 > \overline{b},$$

one can derive the following conditions on the parameters in (2.5), which preserve the inequality: we require $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2 \geq 0$ and

$$\begin{cases} \alpha_1 + \alpha_2 - \beta_1 \geq 0, \\ \alpha_2(K_1 - K_2) - \beta_1(K_1 - \overline{b}) + \alpha_3(K_3 - K_1)^+ \geq 0, \\ \alpha_3(K_3 - \underline{b}) - \beta_1(\underline{b} - \overline{b}) + +\alpha_2(\underline{b} - K_2)^+ \geq 1, \\ \alpha_3 + \alpha_4 - \beta_2 \geq 0, \\ \alpha_3(K_3 - K_4) + \beta_2(K_4 - \underline{b}) + \alpha_2(K_4 - K_2)^+ \geq 0, \\ \alpha_2(\overline{b} - K_2) + \beta_2(\overline{b} - \underline{b}) + \alpha_3(K_3 - \overline{b})^+ \geq 1. \end{cases} \tag{2.9}$$

It can then be verified that the four different cases are each specific solutions of this system where the inequalities are tight in differing manners. Of course, checking that these are the only interesting such solutions is nontrivial and will be the content of Theorem 2.2. We also believe that, while perhaps this is presentationally more concise, the unified presentation hides the true nature of the superhedging strategies.

2.2. Subhedging. We present now three constructions of robust subhedges. Depending on the relative distance of barriers from the spot, one of them will turn out to be the most expensive (model-independent) subhedge. We note, however, that there is also a fourth (trivial) subhedge, which has payoff zero and corresponds to an empty portfolio. In fact this will be the most expensive subhedge when $\underline{b} \ll S_0 \ll \overline{b}$, and we can construct a market model in which both barriers are never hit. Details will be given in Theorem 2.4.

$\underline{H}_I$*: subhedge for* $\underline{b} < S_0 < \overline{b}$. Choose $0 < K_2 < \underline{b} < S_0 < \overline{b} < K_1$. The initial portfolio will contain a cash amount, a forward, and calls with five different strikes and will also include two digital options, which pay 1 provided that S_T is above a specified level. Figure 4 demonstrates graphically the hedging strategy, and we note that the effect of the digital options is to provide a jump in the payoff at the points $\underline{b}$, $\overline{b}$.

As in the previous cases, the optimality of the construction will follow from an almost-sure inequality. The relevant inequality is now

$$\begin{aligned} \mathbf{1}_{\overline{S}_T \geq \overline{b}, \underline{S}_T \leq \underline{b}} \geq\ & \alpha_0 + \alpha_1(S_T - S_0) - \alpha_2(S_T - K_2)^+ + \alpha_3(S_T - \underline{b})^+ - \alpha_3(S_T - K_3)^+ \\ & + \alpha_3(S_T - \overline{b})^+ - (\alpha_3 - \alpha_2)(S_T - K_1)^+ - \gamma_1 \mathbf{1}_{\{S_T > \underline{b}\}} + \gamma_2 \mathbf{1}_{\{S_T \geq \overline{b}\}} \\ & + (\alpha_2 - \alpha_1)(S_T - \underline{b}) \mathbf{1}_{\{H_{\underline{b}} < H_{\overline{b}} \wedge T\}} - \alpha_2(S_T - \overline{b}) \mathbf{1}_{\{H_{\underline{b}} < H_{\overline{b}} < T\}} \\ & - (\alpha_3 - \alpha_2 + \alpha_1)(S_T - \overline{b}) \mathbf{1}_{\{H_{\overline{b}} < H_{\underline{b}} \wedge T\}} + (\alpha_3 - \alpha_2)(S_T - \underline{b}) \mathbf{1}_{\{H_{\overline{b}} < H_{\underline{b}} < T\}}. \end{aligned} \tag{2.10}$$

Specifically, we can see that the hedging strategy consists of a portfolio which contains cash α_0, has α_1 forwards, is short α_2 calls at strike K_2, etc. The novel terms here are the digital options; we note further that the digital options can be considered also as the limit of portfolios of calls (see, for example, Bowie and Carr [5]). In our context, we can use their limiting argument to deduce $\mathcal{P}\mathbf{1}_{\{S_T \geq \overline{b}\}} = -C'(K)$.

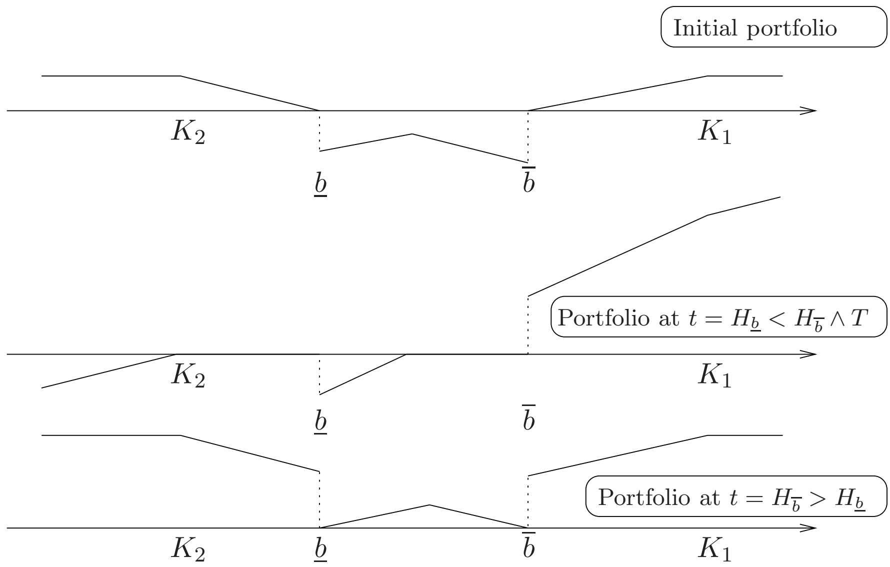


Figure 4. *Subhedge* $\underline{H}_I$.

The strategy to be employed is then as follows: initially, run to either $\underline{b}$ or $\overline{b}$; supposing that $\underline{b}$ is hit first, we buy $(\alpha_2 - \alpha_1)$ forwards, then if we later hit $\overline{b}$, we sell α_2 forwards. A similar strategy is followed if $\overline{b}$ is hit first. As previously, the structure imposes some constraints on the parameters. The relevant constraints are

$$0 = \alpha_0 + \alpha_1(\underline{b} - S_0) - \alpha_2(\underline{b} - K_2), \tag{2.11}$$

$$0 = \alpha_0 + \alpha_1(\overline{b} - S_0) - \alpha_2(\overline{b} - K_2) + \alpha_3(K_3 - \underline{b}) - \gamma_1 + \gamma_2, \tag{2.12}$$

$$1 = \alpha_0 + \alpha_1(K_2 - S_0) + (\alpha_2 - \alpha_1)(K_2 - \underline{b}) - \alpha_2(K_2 - \overline{b}), \tag{2.13}$$

$$1 = \alpha_0 + \alpha_1(K_2 - S_0) - (\alpha_3 - \alpha_2 + \alpha_1)(K_2 - \overline{b}) + (\alpha_3 - \alpha_2)(K_2 - \underline{b}), \tag{2.14}$$

$$\gamma_1 = (K_3 - \underline{b})\alpha_3, \tag{2.15}$$

$$\gamma_2 = (\overline{b} - K_3)\alpha_3, \tag{2.16}$$

$$\frac{K_3 - \underline{b}}{K_1 - \underline{b}} = \frac{\overline{b} - K_3}{\overline{b} - K_2}. \tag{2.17}$$

Equations (2.11) and (2.12) arise from the constraint that initially the payoff is zero at $\underline{b}$, $\overline{b}$; constraints (2.13) and (2.14) come from the constraint that the final payoff is 1 at K_2 when both barriers are hit (in either order); (2.15) and (2.16) represent the fact that, in the intermediate step, at K_3 the gap at $\underline{b}$ (resp., $\overline{b}$) is the size of the respective digital option. The final constraint, (2.17), follows from noting that K_3 is the intersection point of the lines from $(\underline{b}, 0)$ to $(K_1, 1)$ and from $(K_2, 1)$ to $(\overline{b}, 0)$. Note that it follows that the initial payoffs on $(0, K_1)$ and (K_2, ∞) are colinear and that the final payoff in K_1 is 1 when both barriers are hit.

The given equations can be solved to deduce

$$
\left\{\begin{array}{l}
\alpha_0 = \frac{S_0(K_1+K_2-\overline{b}-\underline{b})+\overline{b}\underline{b}-K_1K_2}{(\overline{b}-K_2)(K_1-\underline{b})},\\
\alpha_1 = \frac{K_1+K_2-\overline{b}-\underline{b}}{(\overline{b}-K_2)(K_1-\underline{b})},\\
\alpha_2 = \frac{1}{\overline{b}-K_2},\\
\alpha_3 = \frac{\overline{b}-K_2+K_1-\underline{b}}{(\overline{b}-K_2)(K_1-\underline{b})},
\end{array}\right.
\qquad
\left\{\begin{array}{l}
K_3 = \frac{\overline{b}K_1-\underline{b}K_2}{\overline{b}-K_2-\underline{b}+K_1},\\
\gamma_1 = \frac{\overline{b}-\underline{b}}{\overline{b}-K_2},\\
\gamma_2 = \frac{\overline{b}-\underline{b}}{K_1-\underline{b}}.
\end{array}\right. \tag{2.18}
$$

We note from the above that $\alpha_2, \alpha_3, \gamma_1$, and γ_2 are all (strictly) positive; further, it can be checked that the quantities $(\alpha_3 - \alpha_2), (\alpha_2 - \alpha_1), (\alpha_3 - \alpha_2 + \alpha_1)$ are all positive. It follows that the construction holds for all choices of K_1, K_2 with $K_2 < \underline{b}$ and $K_1 > \overline{b}$.

For future reference, we define $\underline{H}_I(K_1, K_2)$ to be the random variable given by the RHS of (2.10), where the coefficients are given by the solutions of (2.11)–(2.17).

$\underline{H}_{II}$*: subhedge for* $\underline{b} < S_0 \ll \overline{b}$. While the above hedge can be considered to be the "typical" subhedge for the option, there are two further cases that need to be considered when the initial stock price, S_0, is much closer to one of the barriers than the other. The resulting subhedge will share many of the features of the previous construction; however, the main difference concerns the behavior in the tails: we now have the hedge taking the value 1 in the tails under only one of the possible ways of knocking in (specifically, in the case where $\underline{b} < S_0 \ll \overline{b}$, we get equality in the tails only when $\underline{b}$ is hit first).

A graphical representation of the construction is given in Figure 5. In this case, rather than specifying only K_1 and K_2, we also need to specify $K_3 \in (\underline{b}, \overline{b})$ satisfying

$$\frac{\overline{b} - K_3}{\overline{b} - K_2} \geq \frac{K_3 - \underline{b}}{K_1 - \underline{b}}.$$

This identity implies that, for the initial portfolio, the value of the function just below $\overline{b}$ is strictly smaller than the value of the function just above $\underline{b}$. This can be rearranged to get

$$K_3 \leq \overline{b}\frac{K_1 - \underline{b}}{(K_1 - \underline{b}) + (\overline{b} - K_2)} + \underline{b}\frac{\overline{b} - K_2}{(K_1 - \underline{b}) + (\overline{b} - K_2)}.$$

The actual inequality we use remains the same as in the previous case (2.10), as do some of the constraints:

$$0 = \alpha_0 + \alpha_1(\underline{b} - S_0) - \alpha_2(\underline{b} - K_2), \tag{2.19}$$

$$0 = \alpha_0 + \alpha_1(\overline{b} - S_0) - \alpha_2(\overline{b} - K_2) + \alpha_3(K_3 - \underline{b}) - \gamma_1 + \gamma_2, \tag{2.20}$$

$$1 = \alpha_0 + \alpha_1(K_2 - S_0) + (\alpha_2 - \alpha_1)(K_2 - \underline{b}) - \alpha_2(K_2 - \overline{b}), \tag{2.21}$$

$$1 = \alpha_0 + \alpha_1(\underline{b} - S_0) + \alpha_2(K_2 - K_1 + \overline{b} - \underline{b}) + \alpha_3(K_3 + K_1 - \underline{b} - \overline{b}) - \gamma_1 + \gamma_2, \tag{2.22}$$

$$\gamma_1 = (K_3 - \underline{b})\alpha_3, \tag{2.23}$$

$$\gamma_2 = (\overline{b} - K_3)\alpha_3; \tag{2.24}$$

(2.19) and (2.20) refer still to having an initial payoff of 0 at $\overline{b}$ and $\underline{b}$, and (2.23) and (2.24) also still relate the size of the digital options to the slopes. The change is in the constraints

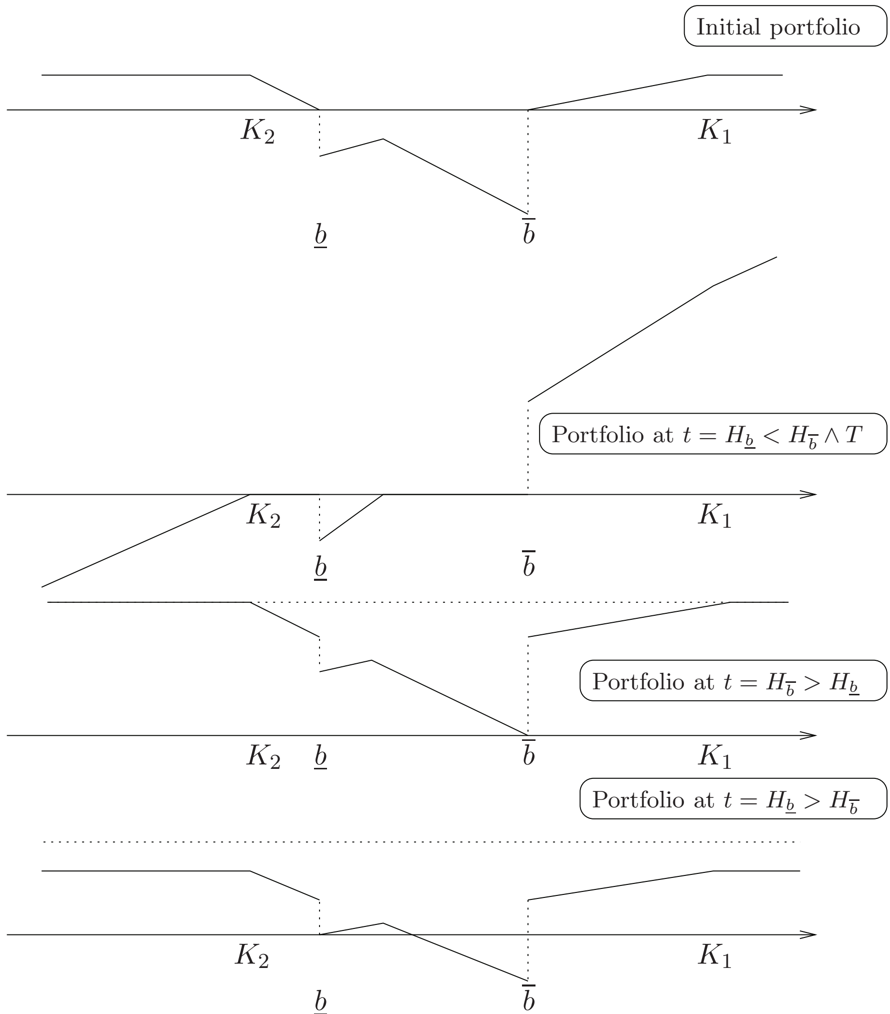


Figure 5. *Subhedge* $\underline{H}_{II}$.

(2.21) and (2.22), which now ensure that the function at K_1 and K_2, after hitting first $\underline{b}$ and then $\overline{b}$, takes the value 1. We note that, in the previous example, where (2.17) held, these are in fact equivalent to (2.13) and (2.14); the fact that (2.17) no longer holds means that we need to be more specific about the constraints.

The solutions to the above are now

(2.25)
$$\begin{cases} \alpha_0 = \frac{(\overline{b}\underline{b}+S_0K_3)(K_1-K_2)-(\underline{b}K_1+S_0\overline{b})(K_3-K_2)-(\overline{b}K_2+S_0\underline{b})(K_1-K_3)}{(\overline{b}-\underline{b})(K_1-K_3)(\overline{b}-K_2)}, \\ \alpha_1 = \frac{K_3(K_1-K_2)-\underline{b}(K_1-K_3)-\overline{b}(K_3-K_2)}{(\overline{b}-\underline{b})(K_1-K_3)(\overline{b}-K_2)}, \\ \alpha_2 = \frac{1}{\overline{b}-K_2}, \\ \alpha_3 = \frac{K_1-K_2}{(K_1-K_3)(\overline{b}-K_2)}, \end{cases} \qquad \begin{cases} \gamma_1 = \frac{(K_3-\underline{b})(K_1-K_2)}{(\overline{b}-K_2)(K_1-K_3)}, \\ \gamma_2 = \frac{(\overline{b}-K_3)(K_1-K_2)}{(\overline{b}-K_2)(K_1-K_3)}. \end{cases}$$

As before, we write $\underline{H}_{II}(K_1, K_2, K_3)$ for the random variable on the RHS of (2.10), where the constants are now specified by (2.25).

$\underline{H}_{III}$*: subhedge for* $\underline{b} \ll S_0 < \overline{b}$. The third case here is the corresponding version of the above where we have a large value of K_3, specifically,

$$K_3 \geq \overline{b}\frac{K_1 - \underline{b}}{(K_1 - \underline{b}) + (\overline{b} - K_2)} + \underline{b}\frac{\overline{b} - K_2}{(K_1 - \underline{b}) + (\overline{b} - K_2)},$$

and we need to modify (2.21) and (2.22) appropriately:

$$\begin{aligned} 1 &= \alpha_0 + \alpha_1(\overline{b} - S_0) + (\alpha_3 - \alpha_2)(\overline{b} - \underline{b}), \\ 1 &= \alpha_0 + \alpha_1(\overline{b} - S_0) + \alpha_2(K_2 - K_1 + \underline{b} - \overline{b}) + \alpha_3(K_3 + K_1 - 2\underline{b}) - \gamma_1 + \gamma_2. \end{aligned}$$

The solutions are now

(2.26)

$$\begin{cases} \alpha_0 = \frac{(\overline{b}\underline{b}+S_0K_3)(K_1-K_2)-(\underline{b}K_1+S_0\overline{b})(K_3-K_2)-(\overline{b}K_2+S_0\underline{b})(K_1-K_3)}{(\overline{b}-\underline{b})(K_3-K_2)(K_1-\underline{b})}, \\ \alpha_1 = \frac{K_3(K_1-K_2)-\underline{b}(K_1-K_3)-\overline{b}(K_3-K_2)}{(\overline{b}-\underline{b})(K_3-K_2)(K_1-\underline{b})}, \\ \alpha_2 = \frac{K_1-K_3}{(K_3-K_2)(K_1-\underline{b})}, \\ \alpha_3 = \frac{K_1-K_2}{(K_3-K_2)(K_1-\underline{b})}, \end{cases} \qquad \begin{cases} \gamma_1 = \frac{(K_3-\underline{b})(K_1-K_2)}{(K_3-K_2)(K_1-\underline{b})}, \\ \gamma_2 = \frac{(\overline{b}-K_3)(K_1-K_2)}{(K_3-K_2)(K_1-\underline{b})}. \end{cases}$$

As before, we write $\underline{H}_{III}(K_1, K_2, K_3)$ for the random variable on the RHS of (2.10), where the constants are now specified by (2.26).

We note that all three subhedges have very similar structure, and it was convenient to represent them using a common inequality (2.10). We could further combine them into a "general" lower bound consisting of (2.10) together with a set of inequalities constraining the parameter choice, out of which one finds three differing possible extremal sets of inequalities. However, similarly to the superhedge, we do not think this would offer any new insights or simplify the presentation.

2.3. Pricing. Consider the double touch digital barrier option with the payoff $\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}}$. As an immediate consequence of the superhedging strategies described in section 2.1 we get an upper bound on the price of this derivative, as follows.

Proposition 2.1. *Given the market input* (1.3)*, no-arbitrage* (1.4) *in the class of portfolios* $\mathrm{Lin}(\mathcal{X} \cup \{\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}}\})$ *implies the following inequality between the prices:*

$$\mathcal{P}\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}} \leq \inf\left\{\mathcal{P}\overline{H}^{I}(K), \mathcal{P}\overline{H}^{II}(K'), \mathcal{P}\overline{H}^{III}(K_1, K_2, K_3, K_4), \mathcal{P}\overline{H}^{IV}(K_1, K_4)\right\}, \tag{2.27}$$

where the infimum is taken over $K > \underline{b}$*,* $K' < \overline{b}$*, and* $0 < K_4 < \underline{b} < K_3 < K_2 < \overline{b} < K_1$ *and where* $\overline{H}^{I}$*,* $\overline{H}^{II}$*,* $\overline{H}^{III}$*, and* $\overline{H}^{IV}$ *are given by* (2.1)*,* (2.2)*,* (2.5)*, and* (2.7)*, respectively.*

The purpose of this section is to show that, given the law of S_T and the pair of barriers $\underline{b}, \overline{b}$, we can determine explicitly which superhedges and with what strikes the infimum on the RHS of (2.27) is achieved. We present formal criteria, but we also use labels, e.g., $\underline{b} \ll S_0 < \overline{b}$, which provide an intuitive classification. Furthermore, we will show that we can always construct a market model in which the infimum in (2.27) is the actual price of the double barrier option and therefore exhibit the model-independent least upper bound for the price of the derivative. Subsequently, an analogous reasoning for subhedging and the lower bound is presented.

Let μ be the market implied law of S_T given by (1.5). The barycenter of μ associates with a nonempty Borel set $\Gamma \subset \mathbb{R}$ the mean of μ over Γ via

$$\mu_B(\Gamma) = \frac{\int_\Gamma u\mu(\mathrm{d}u)}{\int_\Gamma \mu(\mathrm{d}u)}. \tag{2.28}$$

For $w < \underline{b}$ and $z > \overline{b}$ let $\rho_-(w) > \underline{b}$ and $\rho_+(z) < \overline{b}$ be the unique points such that the intervals $[w, \rho_-(w)]$ and $[\rho_+(z), z]$ are centered respectively around $\underline{b}$ and $\overline{b}$, that is,

$$\begin{cases} \rho_- : [0, \underline{b}] \to [\underline{b}, \infty) & \text{defined via } \mu_B([w, \rho_-(w)]) = \underline{b}, \\ \rho_+ : [\overline{b}, \infty) \to [0, \overline{b}] & \text{defined via } \mu_B([\rho_+(z), z]) = \overline{b}. \end{cases} \tag{2.29}$$

Note that $\rho_\pm$ are decreasing and well defined as $\mu_B([0,\infty)) = S_0 \in (\underline{b}, \overline{b})$. We need to define two more functions,

$$\begin{cases} \gamma_+(w) \geq \overline{b} & \text{defined via } \mu_B\big([0, w] \cup [\rho_+(\gamma_+(w)), \gamma_+(w)]\big) = \underline{b}, \ w \leq \underline{b}, \\ \gamma_-(z) \leq \underline{b} & \text{defined via } \mu_B\big([\gamma_-(z), \rho_-(\gamma_-(z))] \cup [z, \infty)\big) = \overline{b}, \ z \geq \overline{b}, \end{cases} \tag{2.30}$$

so that $\gamma_+(\cdot)$ is increasing, $\gamma_-(\cdot)$ is decreasing, and

$$\gamma_+(w) \downarrow \overline{b} \quad \text{as } w \downarrow 0, \qquad \gamma_-(z) \uparrow \underline{b} \quad \text{as } z \uparrow \infty.$$

Note that γ_+ is defined on $[0, w_0]$, where $w_0 = \underline{b} \wedge \sup\{w < \underline{b} : \gamma_+(w) < \infty\}$, and similarly γ_- is defined on $[z_0, \infty]$. We are now ready to state our main theorem.

Theorem 2.2. *Let μ be the law of S_T inferred from the prices of vanillas via (1.5), and consider the double barrier derivative paying $\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}}$ for a fixed pair of barriers $\underline{b} < S_0 < \overline{b}$. Then exactly one of the following is true:*

$\boxed{\text{I}}$ *"$\underline{b} \ll S_0 < \overline{b}$": There exists $z_0 > \overline{b}$ such that*[4]

$$\gamma_-(z) \downarrow 0 \textit{ as } z \downarrow z_0 \quad \textit{and} \quad \rho_-(0) \leq \overline{b}. \tag{2.31}$$

Then there is a market model in which $\mathbb{E}\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}} = \mathbb{E}\overline{H}^{I}(\rho_-(0)) = \frac{P(\rho_-(0))}{\rho_-(0) - \underline{b}}$.

$\boxed{\text{II}}$ *"$\underline{b} < S_0 \ll \overline{b}$": There exists $w_0 < \underline{b}$ such that*

$$\gamma_+(w) \uparrow \infty \textit{ as } w \uparrow w_0 \quad \textit{and} \quad \rho_+(\infty) \geq \underline{b}.$$

Then there is a market model in which $\mathbb{E}\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}} = \mathbb{E}\overline{H}^{II}(\rho_+(\infty)) = \frac{C(\rho_+(\infty))}{\overline{b} - \rho_+(\infty)}$.

[4]Note that here and subsequently, we use $\uparrow \infty$ and $\downarrow 0$ as meaning only the case where the increasing/decreasing sequence is itself finite/strictly positive, so that in (2.31) we strictly mean $\gamma_-(z) \to 0$ as $z \downarrow z_0$ and $\gamma_-(z) > 0$ for $z > z_0$.

$\boxed{\text{III}}$ *"$\underline{b} \ll S_0 \ll \overline{b}$": There exists $0 \le w_0 \le \underline{b}$ such that $\gamma_-(\gamma_+(w_0)) = w_0$ and $\rho_-(w_0) \le \rho_+(\gamma_+(w_0))$. Then there is a market model in which*

$$\begin{aligned} \mathbb{E}\mathbf{1}_{\overline{S}_T \ge \overline{b},\, \underline{S}_T \le \underline{b}} &= \mathbb{E}\overline{H}^{III}\big(\gamma_+(w_0), \rho_+(\gamma_+(w_0)), \rho_-(w_0), w_0\big) \\ &= \alpha_1 C\big(\gamma_+(w_0)\big) + \alpha_2 C\big(\rho_+(\gamma_+(w_0))\big) + \alpha_3 P\big(\rho_-(w_0)\big) + \alpha_4 P\big(w_0\big), \end{aligned} \tag{2.32}$$

where α_i are given in (2.6).

$\boxed{\text{IV}}$ *"$\underline{b} < S_0 < \overline{b}$": We have $\overline{b} < \rho_-(0)$, $\underline{b} > \rho_+(\infty)$, and $\rho_+(\rho_-(0)) < \rho_-(\rho_+(\infty))$. Then there is a market model in which*

$$\begin{aligned} \mathbb{E}\mathbf{1}_{\overline{S}_T \ge \overline{b},\, \underline{S}_T \le \underline{b}} &= \mathbb{E}\overline{H}^{IV}\big(\rho_-(0), \rho_+(\infty)\big) \\ &= \alpha_1 C\big(\rho_-(0)\big) + \alpha_2 P\big(\rho_+(\infty)\big) + \alpha_4, \end{aligned} \tag{2.33}$$

where α_i are given in (2.8).

We present now the analogues of Proposition 2.1 and Theorem 2.2 for the subhedging case. Whereas above we find an upper bound on the price of the derivative, in this case we will construct a lower bound.

Proposition 2.3. *Given the market input (1.3), no-arbitrage (1.4) in the class of portfolios* $\mathrm{Lin}(\mathcal{X} \cup \{\mathbf{1}_{\overline{S}_T \ge \overline{b},\, \underline{S}_T \le \underline{b}}, \mathbf{1}_{\{S_T > \underline{b}\}}, \mathbf{1}_{\{S_T \ge \overline{b}\}}\})$ *implies the following inequality between the prices:*

$$\mathcal{P}\mathbf{1}_{\overline{S}_T \ge \overline{b},\, \underline{S}_T \le \underline{b}} \ge \sup\{\mathcal{P}\underline{H}_I(K_1, K_2), \mathcal{P}\underline{H}_{II}(K_1, K_2, K_3), \mathcal{P}\underline{H}_{III}(K_1, K_2, K_3), 0\}, \tag{2.34}$$

where the supremum is taken over $0 < K_2 < \underline{b} < K_3 < \overline{b} < K_1$ and $\underline{H}_I, \underline{H}_{II}, \underline{H}_{III}$ are given by (2.10) and the solutions to the relevant set of equations: (2.18), (2.25), and (2.26).

Again, an important aspect of (2.34) is that we can in fact show that the bound is tight—that is, given a set of call prices, there exists a market model under which equality is attained. Recall that under no-arbitrage the prices of digital calls are essentially specified by our market input via $\mathcal{P}\mathbf{1}_{\{S_T \ge \overline{b}\}} = -C'(K)$.

In order to classify the different states, we make the following definitions. Let μ be the law of S_T implied by the call prices. Fix $\underline{b} < S_0 < \overline{b}$, and, given $v \in [\underline{b}, \overline{b}]$, define

$$\begin{aligned} \psi(v) = \inf\Bigg\{ z \in [0, \underline{b}] : \int_{(z,\underline{b}) \cup (v,\overline{b})} u\,\mu(\mathrm{d}u) + \overline{b}\left(\frac{\overline{b} - S_0}{\overline{b} - \underline{b}} - \mu((z,\underline{b}) \cup (v,\overline{b}))\right) \\ = \underline{b}\frac{\overline{b} - S_0}{\overline{b} - \underline{b}} \text{ and } \mu((z,\underline{b}) \cup (v,\overline{b})) \le \frac{\overline{b} - S_0}{\overline{b} - \underline{b}} \Bigg\}, \end{aligned} \tag{2.35}$$

$$\begin{aligned} \theta(v) = \sup\Bigg\{ z \ge \overline{b} : \int_{(\underline{b},v) \cup (\overline{b},z)} u\,\mu(\mathrm{d}u) + \underline{b}\left(\frac{S_0 - \underline{b}}{\overline{b} - \underline{b}} - \mu((\underline{b},v) \cup (\overline{b},z))\right) \\ = \overline{b}\frac{S_0 - \underline{b}}{\overline{b} - \underline{b}} \text{ and } \mu((\underline{b},v) \cup (\overline{b},z)) \le \frac{S_0 - \underline{b}}{\overline{b} - \underline{b}} \Bigg\}, \end{aligned} \tag{2.36}$$

where we use the convention $\sup\{\emptyset\} = -\infty$, $\inf\{\emptyset\} = \infty$. The functions ψ and θ have a natural interpretation in terms of embedding properties used in the proofs in section 4. For example, the definition of ψ ensures that, on the set where $\psi(v) \ne \infty$, we can diffuse all the mass initially from S_0 to $\{\underline{b}, \overline{b}\}$ and then embed from $\underline{b}$ to $(\psi(v), \underline{b}) \cup (v, \overline{b})$ and a compensating atom at $\overline{b}$ with the remaining mass.

The functions ψ and θ are both decreasing on the sets $\{v \in [\underline{b}, \overline{b}] : \psi(v) < \infty\}$ and $\{v \in [\underline{b}, \overline{b}] : \theta(v) > -\infty\}$, which are both closed intervals. Specifically, we will be interested in the region where both the functions allow for a suitable embedding; define

$$\begin{aligned} \overline{v} &= \min\left\{\sup\{v \in [\underline{b}, \overline{b}] : \psi(v) < \infty\}, \sup\{v \in [\underline{b}, \overline{b}] : \theta(v) > -\infty\}\right\}, \\ \underline{v} &= \max\left\{\inf\{v \in [\underline{b}, \overline{b}] : \psi(v) < \infty\}, \inf\{v \in [\underline{b}, \overline{b}] : \theta(v) > -\infty\}\right\}, \\ \kappa(v) &= \overline{b}\frac{\theta(v) - \underline{b}}{\theta(v) - \underline{b} + \overline{b} - \psi(v)} + \underline{b}\frac{\overline{b} - \psi(v)}{\theta(v) - \underline{b} + \overline{b} - \psi(v)}, \end{aligned} \tag{2.37}$$

where $\sup\{\emptyset\} = -\infty$ and $\inf\{\emptyset\} = \infty$.

Theorem 2.4. *Let μ be the law of S_T inferred from the prices of vanillas via (1.5), consider the double barrier derivative paying $\mathbf{1}_{\overline{S}_T \geq \overline{b}, \underline{S}_T \leq \underline{b}}$ for a fixed pair of barriers $\underline{b} < S_0 < \overline{b}$, and recall (2.35)–(2.37). Then exactly one of the following is true:*

$\boxed{\text{I}}$ *"$\underline{b} < S_0 < \overline{b}$": We have $\overline{v} \geq \underline{v}$, and there exists $v_0 \in [\underline{v}, \overline{v}]$ such that $\kappa(v_0) = v_0$. Then there exists a market model in which*

$$\begin{aligned} \mathbb{E}\mathbf{1}_{\overline{S}_T \geq \overline{b}, \underline{S}_T \leq \underline{b}} &= \mathbb{E}\underline{H}_I(\theta(v_0), \psi(v_0)) \\ &= \alpha_0 + \alpha_2(C(\theta(v_0)) - C(\psi(v_0))) + \gamma_2 D(\overline{b}) - \gamma_1 D(\underline{b}) \\ &\quad + \alpha_3\left[C(\underline{b}) + C(\overline{b}) - C(v_0) - C(\theta(v_0))\right], \end{aligned} \tag{2.38}$$

where $D(x)$ is the price of a digital option with payoff $\mathbf{1}_{\{S_T \geq x\}}$ and the values of $\alpha_0, \alpha_2, \alpha_3, \gamma_1, \gamma_2$ are given by (2.18).

$\boxed{\text{II}}$ *"$\underline{b} < S_0 \ll \overline{b}$": We have $\overline{v} \geq \underline{v}$ and $\overline{v} < \kappa(\overline{v})$. Then there exists a market model in which*

$$\begin{aligned} \mathbb{E}\mathbf{1}_{\overline{S}_T \geq \overline{b}, \underline{S}_T \leq \underline{b}} &= \mathbb{E}\underline{H}_{II}(\theta(\overline{v}), \psi(\overline{v}), \overline{v}) \\ &= \alpha_0 + \alpha_2(C(\theta(\overline{v})) - C(\psi(\overline{v}))) + \gamma_2 D(\overline{b}) - \gamma_1 D(\underline{b}) \\ &\quad + \alpha_3\left[C(\underline{b}) + C(\overline{b}) - C(\overline{v}) - C(\theta(\overline{v}))\right], \end{aligned} \tag{2.39}$$

where $D(x)$ is the price of a digital option with payoff $\mathbf{1}_{\{S_T \geq x\}}$ and the values of $\alpha_0, \alpha_2, \alpha_3, \gamma_1, \gamma_2$ are given by (2.25).

$\boxed{\text{III}}$ *"$\underline{b} \ll S_0 < \overline{b}$": We have $\overline{v} \geq \underline{v}$ and $\underline{v} > \kappa(\underline{v})$. Then there exists a market model in which*

$$\begin{aligned} \mathbb{E}\mathbf{1}_{\overline{S}_T \geq \overline{b}, \underline{S}_T \leq \underline{b}} &= \mathbb{E}\underline{H}_{III}(\theta(\underline{v}), \psi(\underline{v}), \underline{v}) \\ &= \alpha_0 + \alpha_2(C(\theta(\underline{v})) - C(\psi(\underline{v}))) + \gamma_2 D(\overline{b}) - \gamma_1 D(\underline{b}) \\ &\quad + \alpha_3\left[C(\underline{b}) + C(\overline{b}) - C(\underline{v}) - C(\theta(\underline{v}))\right], \end{aligned}$$

where $D(x)$ is the price of a digital option with payoff $\mathbf{1}_{\{S_T \geq x\}}$ and the values of $\alpha_0, \alpha_2, \alpha_3, \gamma_1, \gamma_2$ are given by (2.26).

$\boxed{\text{IV}}$ *"$\underline{b} \ll S_0 \ll \overline{b}$": We have $\overline{v} < \underline{v}$. Then there exists a market model in which $\mathbb{E}\mathbf{1}_{\overline{S}_T \geq \bar{b},\, \underline{S}_T \leq \underline{b}} = 0$.*

Furthermore, in cases $\boxed{\text{I}}$–$\boxed{\text{III}}$ we have $\underline{v} = \inf\{v \in [\underline{b}, \overline{b}] : \psi(v) < \infty\} \leq \sup\{v \in [\underline{b}, \overline{b}] : \theta(v) > -\infty\} = \overline{v}$.

3. Applications and practical considerations.

3.1. Finitely many strikes. One important practical aspect where reality differs from the theoretical situation described above concerns the availability of calls with arbitrary strikes. Generally, calls will only trade at a finite set of strikes, $0 = x_0 \leq x_1 \leq \cdots \leq x_N$ (with $x_0 = 0$ corresponding to the asset itself). It is then natural to ask: how does this affect the hedging strategies introduced above? In full generality, this question results in a rather large number of "special" cases that need to be considered separately (for example, the case where no strikes are traded above $\overline{b}$, or the case where there are no strikes traded with $\underline{b} < K < \overline{b}$). In addition, there are differing cases, dependent on whether the digital options at $\underline{b}$ and $\overline{b}$ are traded. Consequently, we will not attempt to give a complete answer to this question, but will consider only the cases where there are "comparatively many" traded strikes and will assume that digital calls are not available to trade. Furthermore, we will apply the theorems of section 2 to measures with atoms. It should be clear how to do this, but a formal treatment would be rather lengthy and tedious, with some extra care needed when the atoms are at the barriers. In addition, as noted in Cox and Obłój [14], there are some rather technical issues relating to forms of arbitrage that need to be carefully considered for some boundary cases. For that reason we state the results of this section only informally.

Mathematically, the presence of atoms in the measure μ means that the call prices are no longer twice differentiable. The function is still convex, but we now have possibly differing left and right derivatives for the function. The implication for the call prices is the following:

$$\mu([x, \infty)) = -C'_-(K) \quad \text{and} \quad \mu((x, \infty)) = -C'_+(K).$$

In particular, atoms of μ will correspond to "kinks" in the call prices.

The first remark to make in the finite-strike case is that, if we replace the supremum/infimum over strikes that appear in expressions such as (2.27) and (2.34) by the maximum/minimum over traded strikes, then the arguments that conclude that these are lower/upper bounds on the price are still valid. The argument breaks down only when we wish to show that these are the best possible bounds. To try to replace the latter, we now need to consider which models might be possible under the given call prices. Our approach will be based on the following type of argument:

(i) suppose that, using only calls and puts with traded strikes, we may construct $\tilde{H}^i$ for $i \in \{I, \ldots, IV\}$ such that $\tilde{H}^i \geq \overline{H}^i$ as a function of S_T;

(ii) suppose further that we can find an admissible call price function $C(K)$, $K > 0$, which agrees with the traded prices and such that in the market model $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P}^*)$ associated by Theorem 2.2 with the upper bound (2.27) we have $\tilde{H}^i(S_T) = \overline{H}^i(S_T)$, $\mathbb{P}^*$-a.s.;

then the smallest upper bound on the price of a digital double touch barrier option is the cost of the cheapest portfolio $\tilde{H}^i$. This is fairly easy to see: clearly the price is an upper

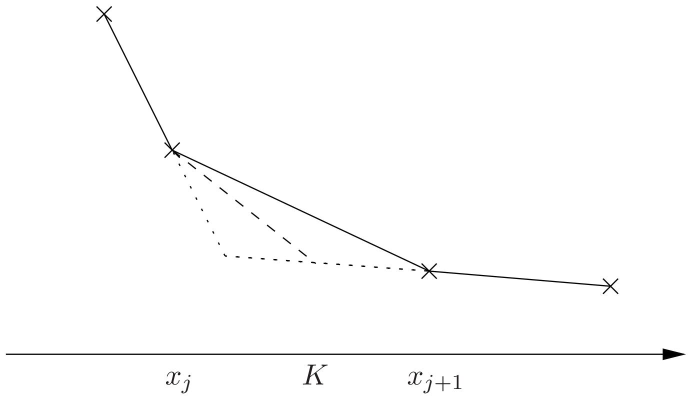


Figure 6. *Possible call price surfaces as a function of the strike. The crosses indicate the prices of traded calls, and the solid line corresponds to the upper bound, which corresponds to placing all possible mass at traded strikes. The lower bound on the interval (x_j, x_{j+1}) is indicated by the dotted line, and the dashed line indicates the surface we will choose when we wish to minimize the call price at K. In this case, we note that there will be mass at K and x_j but not at x_{j+1}. There is a second case where K is below the kink in the dotted line, when the resulting surface would place mass at K and x_{j+1} but not at x_j .*

bound on the price of the option, since $\tilde{H}^i$ superhedges, and under $\mathbb{P}^*$ this upper bound is attained. Indeed, by assumption on $\tilde{H}^i$, in the market model associated with $\mathbb{P}^*$ we have $\mathcal{P}\tilde{H}^i = \mathbb{E}^*\tilde{H}^i = \mathbb{E}^*\overline{H}^i = \mathcal{P}\overline{H}^i$. Consequently, by Theorem 2.2, the price of the traded portfolio $\tilde{H}^i$ and the price of the digital double touch barrier option are equal under the market model $\mathbb{P}^*$. Note that in (ii) above it is in fact enough to have $\tilde{H}^i(S_T) = \overline{H}^i(S_T)$ just for the $\overline{H}^i$ which attains equality in (2.27).

We now wish to understand the possible models that might correspond to a given set of call prices $\{C(x_i); 0 \leq i \leq N\}$. Simple arbitrage constraints (see, e.g., Carr and Madan [11] or Theorem 3.1 in Davis and Hobson [16][5]) require that the call prices at other strikes (if traded) are such that the function $C(K)$ is convex and decreasing. This allows us to deduce that, for K such that $x_j < K < x_{j+1}$ for some j, we must have

$$C(K) \leq C(x_j)\frac{x_{j+1} - K}{x_{j+1} - x_j} + C(x_{j+1})\frac{K - x_j}{x_{j+1} - x_j}, \tag{3.1}$$

$$C(K) \geq C(x_j) + \frac{C(x_j) - C(x_{j-1})}{x_j - x_{j-1}}(K - x_j), \tag{3.2}$$

$$C(K) \geq C(x_{j+1}) - \frac{C(x_{j+2}) - C(x_{j+1})}{x_{j+2} - x_{j+1}}(x_{j+1} - K); \tag{3.3}$$

see Figure 6 for a graphical representation. These inequalities therefore provide upper and lower bounds on the call price at strike K, and it can be seen that the upper bound and lower bound are tight by choosing suitable models: in the upper bound, the corresponding

[5]We suppose also that our call prices do not exhibit what is termed here as weak arbitrage.

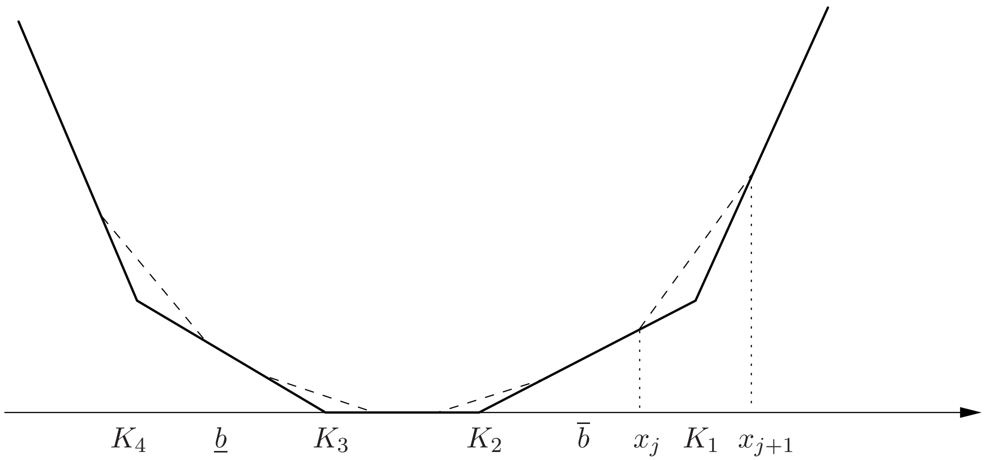


Figure 7. *An optimal superhedge $\overline{H}^{III}$ in the case where only finitely many strikes are traded. The lower (solid) payoff denotes the optimal construction under the chosen extension of call prices to all strikes, and the upper (dashed) payoff denotes the payoff actually constructed. Note that, for example, x_j is the largest traded strike below K_1, and x_{j+1} is the smallest traded strike greater than K_1.*

model places all mass of the law of S_T at the strikes x_i; in the lower bounds, the larger of the two possible terms can be attained with a law that places mass at K, and at other x_i's except, in (3.2), at x_j, and, in (3.3), at x_{j+1}. Moreover, provided that there is at least one traded call between two strikes K, K', we can (for example) choose a law that attains the maximum possible call price at K and the minimum possible price at K', while we need at least two traded strikes between K and K' should we wish these both to be able to attain their minimum possible price. In general, two intermediate strikes will be sufficient for all constructions, and we will assume that this property holds for all "relevant" points in what follows.

Consider first the case where we wish to superhedge the double touch option. We will consider the model $\mathbb{P}^*$ which corresponds (through Theorem 2.2) to the call prices obtained by linearly interpolating the prices at $x_0, \ldots, x_n$—in particular, $C(K)$ is the maximum possible value call prices may take under the assumption of no arbitrage—and assume we are in case $\boxed{\text{III}}$ of Theorem 2.2 so that (2.32) holds. In particular, $\overline{H}^{III}$ would be a perfect hedge if all call options were traded. The key idea is to consider the portfolio depicted in Figure 7, where the smaller payoff is the static part of $\overline{H}^{III}$ and the upper payoff is one that may be constructed using only strikes that are actually traded. Then, although the upper payoff is strictly larger between x_j and x_{j+1}, say, where $x_j < K_1 < x_{j+1}$, the points at which this occurs are not attained by S_T since, under $\mathbb{P}^*$, the law of S_T is supported only by the points x_i. Consequently, both the upper and lower payoff have a.s. the same payoff under $\mathbb{P}^*$ and therefore the same expectation and price. Define $\tilde{H}^{III}$ to be $\overline{H}^{III}$ with its static part replaced with the upper payoff in Figure 7. Then $\tilde{H}^{III}$ is a superhedge and a perfect hedge under $\mathbb{P}^*$ and hence is the least expensive superhedge. Analogous arguments (with the same choice of $C(K)$) hold for the other hedges $\overline{H}^{I}$, $\overline{H}^{II}$, and $\overline{H}^{IV}$.

Consider now the lower bound. To keep things simple we begin by altering slightly the problem: rather than the payoff $\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}}$, we consider a subhedge of the option with payoff $\mathbf{1}_{\overline{S}_T > \overline{b},\, \underline{S}_T < \underline{b}}$.[6] Then (2.10) still holds[7] with the new term $\mathbf{1}_{\overline{S}_T > \overline{b},\, \underline{S}_T < \underline{b}}$ on the left-hand side (LHS), provided that we modify the digital options on the RHS to $\mathbf{1}_{\{S_T \geq \underline{b}\}}$ and $\mathbf{1}_{\{S_T > \overline{b}\}}$. So we may still consider the optimal portfolio (in all three cases) as being short a collection of calls at strikes K_1, K_2, and K_3, long calls at $\underline{b}$ and $\overline{b}$, and holding digital options at each of these points. Intuitively, we should look for a model which will maximize the cost of the calls at K_1, K_2, and K_3 and minimize the cost at $\overline{b}$ and $\underline{b}$, as well as maximizing the cost of a digital call at $\underline{b}$ and minimizing the cost of the digital call at $\overline{b}$. The former conditions correspond to choosing the call prices which give the upper bound (3.1), so we choose the call price which linearly interpolates $C(x_i)$ except when $x_i \leq \underline{b} \leq x_{i+1}$ and $x_i \leq \overline{b} \leq x_{i+1}$. In the latter cases, we wish to minimize the call price, so we choose the prices corresponding to the appropriate lower bound (3.2) or (3.3), which have a kink at $\underline{b}$ and at (exactly) one of its two adjacent traded strikes, and likewise for $\overline{b}$. We note that the prices of the digital calls (which are either minus the left gradient or minus the right gradient of the call prices at the barrier) will also now be optimized when they trade in exactly the forms specified above (that is, the digital call at $\underline{b}$ pays out only if the asset is greater than or equal to $\underline{b}$, while the digital call at $\overline{b}$ will pay out if the asset is strictly larger than $\overline{b}$ at maturity).

The above procedure specifies uniquely a complete set of call prices $C(K)$, which match the market input and which are our candidate for the largest lower bound among the possible models. Then we note that a construction similar to that given in Figure 8 will work—the main difference from the superhedge case is that, at the discontinuity, there are two possible cases that need to be considered, and the optimal subhedge will depend on behavior of $C(K)$ at the strikes adjacent to $\underline{b}$ and $\overline{b}$. More precisely, the portfolio given by the dotted line in Figure 8 corresponds to the case when $C(K)$ has no kinks at the traded strikes to the immediate right of both barriers (i.e., S_T has no atoms at these strikes). The other three possibilities are straightforward modifications. The argument then proceeds as above: in the model given by Theorem 2.4, subhedge $\underline{H}_i$ achieves equality in (2.34) (modified to account for the changes from inequalities to strict inequalities and vice versa), and the portfolio constructed in Figure 8 is a.s. equal[8] to $\underline{H}_i$, so that they must have the same price. The resulting subhedge, constructed using only calls and puts with traded strikes, is therefore a hedge under the chosen model and therefore the optimal lower bound.

---

[6]Under the optimal model (see the proofs in section 4), the distribution of S_T has atoms at the barriers which are embedded by mass which has already hit the other barrier. With the modified payoff, $\mathbf{1}_{\overline{S}_T > \overline{b},\, \underline{S}_T < \underline{b}}$, this will still give equality in the subhedge, but this would not have been the case with the original payoff. One could now consider the option $\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}}$ by approximating it with a sequence of payoffs of the form $\mathbf{1}_{\overline{S}_T > \overline{b}-\varepsilon,\, \underline{S}_T < \underline{b}+\varepsilon}$, which would (in the optimal embedding) place atoms "just inside the barriers"—it is this behavior of the "optimal" construction that required us to consider the modified payoff in this case. A more careful analysis of the behavior in the limit is nontrivial, but the consequences for a similar example can be seen in Cox and Obłój [14].

[7]Technically, for (2.10) to still hold, we actually need to modify the hitting times so that we consider the entrance times of, e.g., $(0, \underline{b})$ rather than the hitting time of $\underline{b}$, and be able to trade forward with strike $\underline{b}$ at this stopping time. In practice, this will not be crucial, since, for example, continuity of the asset price means that this entrance time may be suitably approximated, and it will commonly be the case that the two stopping times are in fact equal.

[8]Assuming that the optimal K_3 chosen is separated from $\underline{b}$ and $\overline{b}$ by a traded call price.

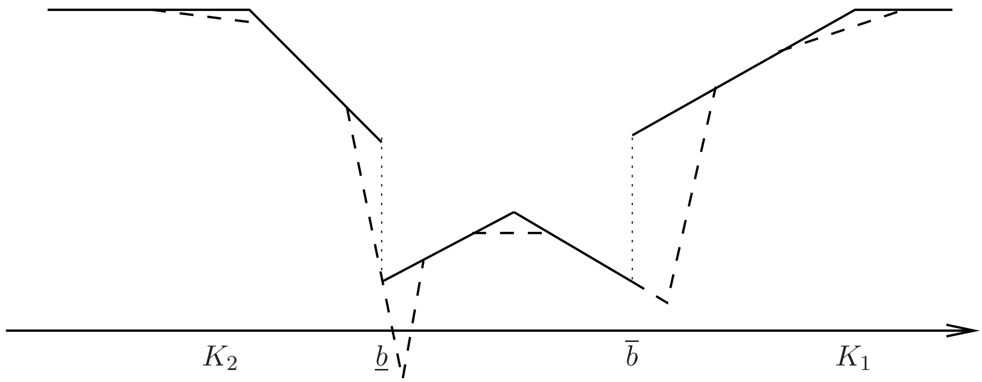


Figure 8. *An optimal subhedge in the case where only finitely many strikes are traded. The upper (solid) payoff denotes the optimal construction under the chosen extension of call prices to all strikes, and the lower (dashed) payoff denotes the payoff actually constructed. Observe that there are two possible constructions at the discontinuity—which type of construction is optimal will depend on whether the optimal lower bound on the call price at $\underline{b}$ or $\overline{b}$ corresponds to the bound* (3.2) *or* (3.3).

3.2. Toward relaxation of our market assumptions. So far, we have made a number of idealized assumptions about the setup that will not necessarily be true in an application. In the previous section, we described how having only finitely many traded strikes may affect our results; in this section we briefly describe how several other aspects that are not captured immediately by our assumptions might alter the results presented.

Continuity of paths. Throughout the paper, we have assumed that $(S_t)_{t\geq 0}$ has continuous paths. In fact we can relax this assumption considerably. First, it is relatively simple to see that if we assume only that barriers $\underline{b}, \overline{b}$ are crossed in a continuous manner, then all of our results remain true. Second, if we make no continuity assumptions,[9] then all our superhedges still work—jumping over the barrier only makes the appropriate forward transaction more profitable. In contrast, our subhedges do not work, and lower bounds on prices are trivially zero and are attained in the model where $S_t = S_0$ for all $t < T$, and then it jumps to the final position S_T.

Zero interest rates. The assumption that we are working with a forward price (or, more specifically, that interest rates are zero) is an important assumption for the methods we have used. Without assuming that the cost-of-carry is zero, the position of the barriers, after discounting, will alter over the lifetime of the option, and our methods are not easily adapted to this setting. There are perhaps two possible resolutions: the first is to look at the “worst-case” barrier values and compute the upper and lower bounds using these values, resulting in upper and lower estimates on the value of the option. Note, however, that there may not be models that correspond to these worst-case scenarios. Second, a more practical response may be to carry out a model-specific adjustment to the model-independent bounds, which may involve something along the following lines: choose an “average” interest adjusted value for each of the barriers and solve the robust problem for these barriers (and discounted stock price process).

[9]We always suppose that the processes have right-continuous paths with left limits.

The idea is to use this as a basic hedge, and, in general, one should expect that the performance of the hedge may be fairly close to a super- or subhedge. Now choose your favorite model, look at the paths where the super- or subhedge fails, and add in a model-specific adjustment. If the choice of model which one makes is moderately good, one would expect an adjustment that is of relatively small transactional size compared to the model-independent component, but which performed relatively well across most models, thus maintaining the relatively robust nature of the initial hedge.

Additional traded options. In practice, in most markets where digital double barrier options are traded, there will be other options traded—for example, one-touch barrier options or call options with earlier maturities. In both of these cases, one would expect these options to include further information about the price of the double-barrier option which, one may hope, would impact the resulting price bounds and hedges. In theory, we believe this to be the case; however, it would seem that a complete analysis along the lines carried out in this work that also included these extra options would be much more involved—this additional complexity can be seen, for example, in the work of Brown, Hobson, and Rogers [7], where including calls traded at a single intermediate time considerably complicates the picture established in Hobson [28]. In this case, the most natural question to ask would appear to be: how might the bounds and hedges alter if, in addition to the call prices with the same maturity, one could also trade in one-touch barrier options with the same maturity? From a theoretical point of view, experience with Skorokhod embeddings suggests this may be a very hard problem, but it is an interesting direction for further research.

3.3. Hedging comparisons. In practice, one would expect that the prices we derive as upper and lower bounds using the techniques of this paper will be rather far apart and well outside typical bid/ask spreads. Consequently, the use of these techniques as a method for pricing is unlikely to be successful, although it will give a good indication of the size of the model-risk associated with a given model-based price. However, our techniques also provide superhedging and subhedging strategies that may be helpful. Consider a trader who has sold a double barrier option (at a price determined by some model perhaps) and who wishes to hedge the resulting risk. In a Black–Scholes world, the trader could remove the risk from his position by delta-hedging the short position. However, there are a number of practical considerations that would interfere with such an approach:

Discrete hedging. A notable source of errors in the hedge will be the fact that the hedging portfolio cannot be continuously adjusted; rather the delta of the position might be adjusted on a periodic basis, resulting in an inexact hedge of the position. While including a gamma hedge could improve this, the hedge will never be perfect. In addition, there is an organizational cost (and risk) to setting up such a hedging operation that might be important.

Transaction costs. A second consideration is that each trade will incur a certain level of transaction costs. These might be tiny for delta-hedging, but this is no longer the case for vega-hedging. To minimize the total transaction costs, the trader would like to be able to trade as infrequently as possible. Of course, this means that there will be a necessary trade-off with the discretization errors incurred above.

Model risk. The final concern for the trader would be: am I hedging with the correct model? Using an incorrect model will of course result in systematic hedging errors due to, e.g., incorrectly estimated volatility, but could also lead to large losses should the model fail to incorporate structural effects such as jumps. A delta-hedge would typically be improved with a vega-hedge, which in turn raises the issue of transaction costs as mentioned above.

We claim that the constructions developed in section 2 address some of these issues: there is no need for regular recalculation of the Greeks of the position, although the breaching of the barrier still requires monitoring; since there are only a small number of transactions, it seems likely that the transaction costs may be reduced; our hedge has been derived using robust techniques, so that we will still be hedged even if the market does not behave according to our initial model, and behavior such as jumps (at least for the upper bound of the double touch) will not affect this.

A further consideration that will be of importance to a hedger is the distribution of the returns, even before the transaction costs are deducted. Under the hedging strategy suggested above, if the trader has sold the option using the "correct" price (plus a small profit) and set up the robust superhedging strategy suggested at a higher price, on average the trader will come out even, as he will if he delta hedges. His comparison between the approaches would then come down to the respective risk involved in the different hedges. For a delta/vega-hedge this is typically symmetric about zero; however, the robust hedge will be very asymmetric since it is bounded below. If the trader is particularly worried about the possible tail of his trading losses as a measure of risk, this strict cut-off could be very advantageous. The delta/vega-hedge, on the other hand, has the appeal of having a lower variance of hedging errors.

Of course a variety of such strategies (typically known as static, semistatic, or robust) have been suggested in the literature, under a variety of more or less restrictive assumptions on the price process, and mostly for single barrier options and variants such as knock-out calls. We have already mentioned the paper by Brown, Hobson, and Rogers [8], which makes very limited restrictions on the underlying price process. More restrictive is the work of Bowie and Carr [5], and subsequent papers of Carr and Chou [9] and Carr, Ellis, and Gupta [10]. Here the authors assume that the volatility satisfies a symmetry assumption, and as a consequence one can, for example, hedge a knock-out call with the barrier above the strike by holding the vanilla call and being short a call at a certain strike above the barrier. By the assumption on the volatility, whenever the underlying asset hits the barrier, both calls have the same value, and the position may be closed out for zero value. A related technique is due to Derman, Ergener, and Kani [18], followed up by Andersen, Andreasen, and Eliezer [1] and Fink [23]. The idea here is to use other traded options to make the value of the hedging portfolio equal to zero along the barrier when liquidated. In the simplest form, a portfolio of calls above the barrier at different strikes and/or maturities is purchased so that the portfolio value at selected times before maturity is zero. Extensions allow this idea to be used for stochastic volatility, and even to cover jumps, at the expense of needing possibly a very large portfolio of options. More recently, work of Nalholm and Poulsen [33] unifies both these approaches and allows a fairly general set of asset dynamics, as does work by Giese and Maruhn [25], where the authors find an optimal portfolio by setting up an optimization problem. Note, however, that all these strategies assume a known model for the underlying asset and also that the hedging assets will be liquid enough for the portfolio to be liquidated at the price specified under the model. In addition, since some of the hedging portfolios can involve a large number of options, it is not clear that the static hedges here will be efficient at resolving the issues of transaction costs (and operational simplicity) or model risk. A numerical investigation of the performance of a range of these hedges in practice has been conducted in Engelmann et al. [21], and similar investigations for a different class of static hedges appear in Davis, Schachermayer, and Tompkins [17] and Tompkins [38].

There are also more classical, theoretical approaches to the problem of hedging where the problem is considered in an incomplete market (without which, of course, a perfect hedge would be possible). In this situation (see, for example, Föllmer and Schweizer [24]), one wants to solve an optimal control problem where the aim is to minimize the "risk" of the hedging error, where risk is interpreted suitably (perhaps with regards to a utility function or a risk measure). More recently, in a combination of the static and dynamic approaches, Ilhan, Jonsson, and Sircar [30] have considered the problem of risk minimization over an initial static portfolio and a dynamic trading strategy in the underlying asset.

The most notable difference between the studies described above and the ideas of the previous sections is that we make very few modeling assumptions on the underlying asset, whereas most of the approaches listed require a single model to be specified, with respect to which the results will then be optimal. In particular, these techniques are unable to say anything about hedging losses should the assumed model actually be incorrect.

Of course, the criteria under which we have constructed our hedges—that they are the smallest robust superhedging strategy, or the greatest robust subhedging strategy—do not necessarily mean that the behavior of the hedges in "normal" circumstances will be particularly suitable: we would expect that the hedge would perform best in extreme market conditions; however, in order for it to be suitable as a hedge against model risk, one would also want the performance of the hedge to generally be reasonable. To see how this strategy compares, we will now consider some Monte Carlo–based comparisons with the standard delta/vega-hedging techniques. The comparisons will take the following form:

(i) We choose the Heston model for the "true" underlying asset and compute the time-0 call prices under this model at a range of strike prices, and the time-0 price of a double barrier option.

(ii) We compute the optimal super- and subhedges for the digital double touch barrier option based on the observed call prices, and suppose that the hedger purchases these portfolios using the cash received from the buyer (and borrowing/investing the difference between the portfolios).

(iii) For comparison purposes, we also hedge the option using a suitable delta/vega-hedge with daily updating. For the robust hedges, we will consider both the case where hitting of the barrier is monitored daily and the case where it is monitored exactly.

In the numerical examples, we assume that the underlying process is the Heston stochastic volatility model (Heston [26])

$$\begin{cases} dS_t = \sqrt{v_t} S_t dW^1_t, & S_0 = S_0,\ v_0 = \sigma_0 \\ dv_t = \kappa(\theta - v_t)dt + \xi\sqrt{v_t} dW^2_t, & d\langle W^1, W^2\rangle_t = \rho dt, \end{cases} \tag{3.4}$$

with parameters

$$S_0 = 1.4990,\ \sigma_0^2 = 0.0110,\ \kappa = 3.8626,\ \theta = 0.0169,\ \xi = 0.5004,\ \text{and } \rho = -0.1850. \tag{3.5}$$

These parameters, taken from Ulmer [39], resulted from the calibration of a Heston model to market prices of European options with different strikes and maturities (a total of 176 options) on the EUR/USD foreign exchange spot rate on January 14, 2010. This gives us realistic dynamics for our hypothetical forward market, which in fact is not far from the real spot market, given that 1-year interest rates were then similar for EUR and USD (deposit rates reported by Bloomberg were respectively 1.1% and 0.905%).

Transactions in S_t carry a 0.15% transaction cost, and buying or selling call/put options carries a 1% transaction cost.[10] The delta/vega-hedge is constructed using the Black–Scholes delta of the option, but using the at-the-money implied volatility assuming that the call prices are correct (i.e., they follow the Heston model). While not perfect as a hedge, empirical evidence as in Dumas, Fleming, and Whaley [19] or Engelmann, Fengler, and Schwendner [22] suggests that the hedge is reasonable even without the vega component, although it is also the case that more sophisticated methods should result in an improvement of this benchmark.

We consider a short and a long position in a digital double touch barrier option with payoff $\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}}$ for each combination of upper and lower barriers $\overline{b} = 1.47, 1.52, 1.57$ and $\underline{b} = 1.35, 1.39, 1.43$. We then compare hedging performance of our robust super- and subhedges and the standard delta/vega-hedges by running 20,000 Monte Carlo simulations for each of the nine combinations of upper and lower barrier, and analyzing the resulting loss/gains from trading.[11] We note that whether the barrier is monitored exactly or daily (so that the trades in the robust hedges may not occur exactly at the barrier) will have a noticeable difference on the cumulative distribution function. To highlight this difference, the daily hedge will be type (A) and the exact hedge type (B).

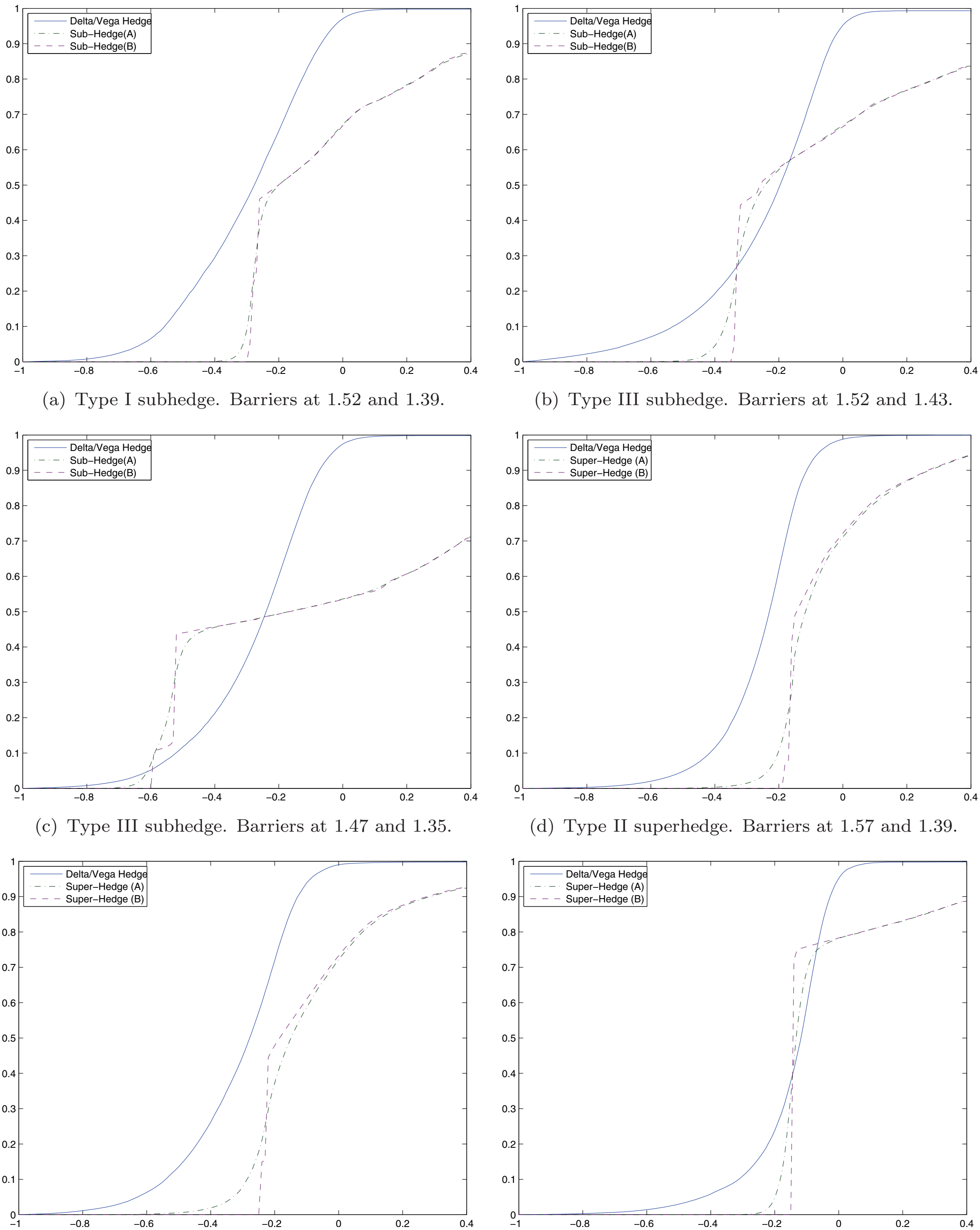
The cumulative distributions of hedging errors for a selection of barrier pairs are given in the graphs in Figure 9. Figure 9 shows that the robust super- and subhedges introduced in this paper can incur losses in a manner similar to the delta/vega strategies. However, the comparatively large losses are less frequent for the robust strategies in comparison to the delta/vega-hedging strategy. Indeed, in Table 1 we show that an agent with an exponential utility[12] $U(x) = 1 - \exp(-x)$ would systematically, strongly prefer the error distribution of our hedges to that of the delta/vega-hedge. Note also that if we allow our hedges to monitor the barrier crossings exactly, the corresponding cumulative distributions of hedging errors have losses which are bounded below. It is interesting, however, to note that in terms of utility of hedging errors (cf. Table 1) this doesn't really change the performance of our hedges.

Finally, we consider how the performance of our hedges compares if we vary the barrier levels. The hedge for the short position fairly consistently outperforms delta/vega-hedging. In the long position, on the other hand, the hedge appears to perform worse as the barriers

[10] The latter has a predominant effect in our simulations, and the results below remain essentially the same even when transactions in S_t carry no transaction costs.

[11] The resulting hedging errors were mean-adjusted as in Tompkins [38] for consistency. The adjustments are of order 0.01 and have no qualitative influence on our results.

[12] We take the parameter α in the utility function $U(x) = 1 - \exp(-\alpha x)$ to be 1, but a sensible parametrization may be to take $\alpha = 10^{-6}$ and to consider a contract with notional value £10^6.



(a) Type I subhedge. Barriers at 1.52 and 1.39.

(b) Type III subhedge. Barriers at 1.52 and 1.43.

(c) Type III subhedge. Barriers at 1.47 and 1.35.

(d) Type II superhedge. Barriers at 1.57 and 1.39.

(e) Type III superhedge. Barriers at 1.52 and 1.39.

(f) Type IV superhedge. Barriers at 1.47 and 1.43.

Figure 9. *Cumulative distributions of hedging errors under different scenarios of a long position* (a)–(c) *and a short position* (d)–(f) *in a double touch option with barriers at different levels under the Heston model* (3.4)–(3.5). *The robust hedge* (A) *monitors barrier crossing daily, while the robust hedge* (B) *is allowed to monitor exactly the moments of barrier crossings.*

Table 1

Comparison of exponential utilities of hedging errors of positions (Pos) in double touch options under Heston model (3.4)–(3.5) resulting from delta/vega ($\Delta/\mathcal{V}$) hedging and our robust (Rob) super- or subhedging strategies. The robust hedge (A) *monitors barrier crossing daily, while the robust hedge* (B) *is allowed to monitor exactly the moments of barrier crossings. Type of the robust hedge and the strikes used in its construction are reported.*

Barriers	Pos	$\Delta/\mathcal{V}$	Rob (A)	Rob (B)	Type	K_1	K_2	K_3	K_4
1.47–1.35	Short	−0.3258	−0.0690	−0.0674	IV	1.5017	1.1611		
1.47–1.35	Long	−0.3278	−0.1774	−0.1767	III	1.7421	1.2546	1.416	
1.47–1.39	Short	−0.3183	−0.0605	−0.0589	IV	1.5818	1.1611		
1.47–1.39	Long	−0.3180	−0.1139	−0.1117	III	1.6753	1.2947	1.4416	
1.47–1.43	Short	−0.1666	−0.0414	−0.0406	IV	1.7487	1.1611		
1.47–1.43	Long	−0.1698	−0.1550	−0.1495	I	1.5751	1.3214	1.4549	
1.52–1.35	Short	−0.3272	−0.0501	−0.0483	III	1.5551	1.4883	1.4015	1.2880
1.52–1.35	Long	−0.3263	−0.0609	−0.0623	III	1.9558	1.0275	1.4549	
1.52–1.39	Short	−0.3750	−0.0824	−0.0786	III	1.5885	1.4616	1.4416	1.3214
1.52–1.39	Long	−0.3799	−0.0779	−0.0795	I	1.7287	1.1477	1.4549	
1.52–1.43	Short	−0.3121	−0.0668	−0.0654	IV	1.7487	1.3414		
1.52–1.43	Long	−0.3169	−0.1107	−0.1082	II	1.6018	1.2078	1.4549	
1.57–1.35	Short	−0.2363	−0.0313	−0.0303	III	1.6152	1.5351	1.3748	1.3214
1.57–1.35	Long	−0.2348	−0.0421	−0.0445	IV				
1.57–1.39	Short	−0.2850	−0.0441	−0.0423	III	1.6486	1.5150	1.4149	1.3614
1.57–1.39	Long	−0.2875	−0.0603	−0.0617	II	1.9758	0.9341	1.4349	
1.57–1.43	Short	−0.2702	−0.0660	−0.0636	III	1.7755	1.4683	1.4416	1.4149
1.57–1.43	Long	−0.2795	−0.0841	−0.0838	II	1.6619	1.1277	1.4349	

get closer together. However, for the parameters given, the robust strategy appears to consistently outperform the delta/vega strategy. Naturally when the barriers vary, the types of super-
and subhedges which we use change. Table 1 also shows which types are optimal depending on the values of the barriers, and indicates the values of the corresponding strikes. This provides an illustration of the intuitive labels we gave to each case in section 2.

The results presented in this section clearly show that the hedges we advocate are in many circumstances an improvement on the classical hedges. We stress that this was meant to be an indicative enquiry and not a comprehensive numerical study of performance of our robust hedging methods. Naturally, there is a large literature (e.g., Hodges and Neuberger [29], Cvitanić and Karatzas [15], Whalley and Wilmott [41]) on more sophisticated techniques that might offer a considerable improvement over the classical hedge we have implemented. In our case the delta/vega-hedge was severely impacted by high transaction costs associated with trading options, and it is possible that such improvements would reverse the relative performance of the hedges. Further, it would be interesting to compare the performance of our hedges against various static and semistatic hedges cited above. Such study could feature different market scenarios and levels of model misspecification as well as more involved performance measures than just expected utility for one value of risk aversion. We believe this is an interesting avenue for further research.

However, we hope that the numerical evidence presented above does at least convince the reader that the robust hedges are competitive with dynamic hedging, and that their differing nature means that they might well prove a more suitable approach in situations where market conditions are dramatically different from the idealized Black–Scholes world—e.g., large transaction costs, illiquidity, or parameter uncertainty. There also seems scope for a more sophisticated approach based on the robust hedges but allowing for some model-based trading: for example, a hybrid of a robust portfolio and some dynamic trading could be used to reduce some of the overhedge in the simple robust hedge. Finally, our simulations include only a very simple form of model misspecification. We would expect that if a trader believes in a model which is significantly different from the real world model, in particular if real world dynamics change dramatically within the time horizon $[0, T]$, then our robust hedging strategies would outperform hedging "using the wrong model."

4. Proofs. Let $(B_t)_{t\geq 0}$ be a standard real valued Brownian motion starting from B_0, defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ satisfying the usual hypothesis. We recall that for any probability measure ν on $\mathbb{R}_+$ with $\nu_B(\mathbb{R}_+) = B_0$ we can find a stopping time τ such that $B_\tau \sim \nu$ and $(B_{t\wedge\tau})$ is a uniformly integrable martingale. Such stopping is simply a solution to the Skorokhod embedding problem (SEP), and a number of different explicit solutions are known; see Obłój [34] for an overview of the domain. Note also that when $\nu([a, b]) = 1$ then $(B_{t\wedge\tau})$ is a uniformly integrable martingale if and only if $B_t \in [a, b]$, $t \leq \tau$, a.s. In the remainder, when speaking about embedding a measure we implicitly mean embedding it in a uniformly integrable (UI) manner in (B_t).

Recall that if $B_0 = S_0$ and τ is an embedding of μ, then $S_t := B_{\tau\wedge\frac{t}{T-t}}$, $t \leq T$, is a market model which matches the market input (1.3). In what follows we will be constructing embeddings τ of μ such that the associated market model attains equality in our super- or subhedging inequalities. Stopping times τ will often be compositions of other stopping times embedding (rescaled) restrictions of μ or some other intermediary measures. Unless specified otherwise, the choice of particular intermediary stopping times has no importance, and we do not specify it—one's favorite solution to the SEP can be used.

Proof of Theorem 2.2. We start with some preliminary lemmas and then prove Theorem 2.2. In the body of the proof, cases $\boxed{\text{I}}$–$\boxed{\text{IV}}$ refer to the cases stated in Theorem 2.2. We note that, by considering $z_0 = \gamma_+(w_0)$, $\boxed{\text{III}}$ is equivalent to

$$\text{There exists } z_0 \geq \overline{b} \text{ such that } \gamma_+(\gamma_-(z_0)) = z_0 \text{ and } \rho_+(z_0) \geq \rho_-(\gamma_-(z_0)).$$

We recall, without proof, straightforward properties of the barycenter function (2.28) which will be useful in what follows.

Lemma 4.1. *The barycenter function defined in* (2.28) *satisfies*

(i) $\mu_B(\Gamma) \geq a$, $\Gamma_1 \subset (0, a) \Longrightarrow \mu_B(\Gamma \setminus \Gamma_1) \geq a$,

(ii) $a \leq \mu_B(\Gamma) \leq \mu_B(\Gamma_1) \leq b$ *and* $\mu(\Gamma \cap \Gamma_1) = 0 \Longrightarrow a \leq \mu_B(\Gamma \cup \Gamma_1) \leq b$.

We separate the proof into two steps. In the first step we prove that exactly one of $\boxed{\text{I}}$–$\boxed{\text{IV}}$ holds. In the second step we construct the appropriate embeddings and market models which achieve the upper bounds on the prices.

Step 1. This step is divided into four possible cases, (A)–(D), corresponding to four regions in which $\underline{b}, \overline{b}$ may lie. For each region, we shall show that exactly one of the instances $\boxed{\text{I}}$–$\boxed{\text{IV}}$ will hold, and therefore in general exactly one them can hold. We start with technical lemmas which are proved after the cases are considered.

Lemma 4.2. *If $\rho_+(\gamma_+(w_0)) < \rho_-(w_0)$ for some w_0, then $\rho_+(\gamma_+(w)) < \rho_-(w)$ for all $w \leq w_0$. Similarly, if $\rho_-(\gamma_-(z_0)) > \rho_+(z_0)$ for some z_0, then $\rho_-(\gamma_-(z)) > \rho_+(z)$ for all $z \geq z_0$.*

Lemma 4.3. *If $\overline{b} \geq \rho_-(0)$ and $\underline{b} \leq \rho_+(\infty)$, then at least of one of the functions $\gamma_\pm$ is bounded on its domain. In particular at most one of $\boxed{\text{I}}$ and $\boxed{\text{II}}$ may be true.*

Lemma 4.4. *$\boxed{\text{IV}}$ implies not $\boxed{\text{III}}$. $\boxed{\text{III}}$ implies not ($\boxed{\text{I}}$ or $\boxed{\text{II}}$).*

Case (a): $\overline{b} \geq \rho_-(0)$, $\underline{b} \leq \rho_+(\infty)$. We note first that the case $\boxed{\text{IV}}$ is not possible, and that the second half of the conditions for $\boxed{\text{I}}$ and $\boxed{\text{II}}$ are trivially true. Suppose that neither $\boxed{\text{I}}$ nor $\boxed{\text{II}}$ holds. If we have $\gamma_-(\overline{b}) = 0$, then $\boxed{\text{III}}$ holds with $w_0 = 0$, and if $\gamma_+(\underline{b}) = \infty$, then $\boxed{\text{III}}$ holds with $w_0 = \underline{b}$. We may thus assume that $\gamma_+(\cdot)$ is bounded above and $\gamma_-(\cdot)$ is bounded away from zero, which in turn implies

$$\gamma_-(\gamma_+(\underline{b})) < \underline{b} \quad \text{and} \quad \gamma_-(\gamma_+(0)) > 0.$$

The function $\gamma_-(\gamma_+(\cdot))$ is continuous and increasing on $(0, \underline{b}]$, and thus we must have w_0 such that $\gamma_-(\gamma_+(w_0)) = w_0$. Finally, suppose that for such a w_0 we in fact have $\rho_-(w_0) > \rho_+(\gamma_+(w_0))$; then Lemma 4.2 implies $\rho_-(0) > \rho_+(\gamma_+(0)) = \overline{b}$, contradicting our assumptions. That only one of the cases $\boxed{\text{I}}$–$\boxed{\text{III}}$ holds now follows from Lemmas 4.3 and 4.4.

Case (b): $\overline{b} \geq \rho_-(0)$, $\underline{b} > \rho_+(\infty)$. It follows that neither $\boxed{\text{II}}$ nor $\boxed{\text{IV}}$ is possible. Observe that Lemma 4.2 implies that $\rho_-(w) \leq \rho_+(\gamma_+(w))$ for all $w \leq \underline{b}$—if this were not true, then $\overline{b} = \rho_+(\gamma_+(0)) < \rho_-(0)$. Suppose further that $\boxed{\text{I}}$ does not hold. If $\gamma_-(\overline{b}) = 0$, then $\boxed{\text{III}}$ holds with $w_0 = 0$.

So assume instead that $\gamma_-(\cdot)$ is bounded away from zero, and therefore that $w < \gamma_-(\gamma_+(w))$ for w close to zero. If we show also that $\gamma_-(\gamma_+(\rho_+(\infty))) \leq \rho_+(\infty)$, then by continuity of $\gamma_-(\gamma_+(\cdot))$ there exists a suitable w_0 for which $\boxed{\text{III}}$ holds. Let $\Gamma_1 = (\rho_+(\infty), \infty)$ and $\Gamma_2 = (\rho_+(\gamma_+(\rho_+(\infty))), \gamma_+(\rho_+(\infty)))$. We have by definition $\mu_B(\Gamma_1) = \overline{b} = \mu_B(\Gamma_2)$ so that $\mu_B(\Gamma_1 \setminus \Gamma_2) = \overline{b}$, since $\Gamma_2 \subset \Gamma_1$. Let $\Gamma = (\rho_+(\infty), \rho_-(\rho_+(\infty))) \cup (\gamma_+(\rho_+(\infty)), \infty)$ and note that $\mu_B(\Gamma) = \overline{b}$ is equivalent to $\gamma_-(\gamma_+(\rho_+(\infty))) = \rho_+(\infty)$. Noting that $\rho_-(w) \leq \rho_+(\gamma_+(w))$ for all $w \leq \underline{b}$ implies $\rho_-(\rho_+(\infty)) \leq \rho_+(\gamma_+(\rho_+(\infty)))$, and using Lemma 4.1, we have

$$\mu_B(\Gamma) = \mu_B\Big((\Gamma_1 \setminus \Gamma_2) \setminus \big(\rho_-(\rho_+(\infty)), \rho_+(\gamma_+(\rho_+(\infty)))\big)\Big) \geq \overline{b},$$

which implies that $\gamma_-(\gamma_+(\rho_+(\infty))) \leq \rho_+(\infty)$. As previously, it remains to note that Lemma 4.4 implies exclusivity of $\boxed{\text{III}}$ and $\boxed{\text{I}}$.

Case (c): $\overline{b} < \rho_-(0)$, $\underline{b} \leq \rho_+(\infty)$. This case is essentially identical to Case (b) above.

Case (d): $\overline{b} < \rho_-(0)$, $\underline{b} > \rho_+(\infty)$. Note that we now cannot have either of $\boxed{\text{I}}$ or $\boxed{\text{II}}$. Suppose further that $\boxed{\text{IV}}$ does not hold—or rather, the weaker:

$$\rho_+(\rho_-(0)) > \underline{b},$$
$$\rho_-(\rho_+(\infty)) < \overline{b}.$$

Let $\Gamma_w = (w, \rho_-(w)) \cup (\overline{b}, \infty)$, and observe that $\mu_B(\Gamma_w)$ decreases as w decreases, provided that $\rho_-(w) < \overline{b}$. We have

$$\mu_B(\Gamma_{\rho_+(\infty)}) \geq \mu_B((\rho_+(\infty), \infty)) = \overline{b}.$$

Our assumption $\rho_-(\rho_+(\infty)) < \overline{b} < \rho_-(0)$ implies that $\rho_-^{-1}(\overline{b}) \in (0, \rho_+(\infty))$ so that $\Gamma_{\rho_-^{-1}(\overline{b})} \supsetneq (\rho_+(\infty), \infty)$ and in consequence $\mu_B(\Gamma_{\rho_-^{-1}(\overline{b})}) < \overline{b}$. Using continuity of $w \to \mu_B(\Gamma_w)$, we conclude that there exists a $w_1 \in (\rho_-^{-1}(\overline{b}), \rho_+(\infty)]$ with $\mu_B(\Gamma_{w_1}) = \overline{b}$, or equivalently $\gamma_-(\overline{b}) = w_1$. A symmetric argument implies that $\gamma_+(\underline{b}) > 0$. We conclude, as in Case (A), that there exists a w_0 such that $\gamma_-(\gamma_+(w_0)) = w_0$.

It remains to show that if $\boxed{\text{IV}}$ does not hold, and the first half of the condition for $\boxed{\text{III}}$ holds, then so too does the second condition. Suppose that w_0 is a point satisfying $\gamma_-(\gamma_+(w_0)) = w_0$, and suppose for a contradiction that $\rho_-(w_0) > \rho_+(\gamma_+(w_0))$. Since the sets $(\rho_+(\gamma_+(w_0)), \gamma_+(w_0))$ and $(w_0, \rho_-(w_0)) \cup (\gamma_+(w_0), \infty)$ are both centered at $\overline{b}$ and overlap, it follows that $\mu_B([w_0, \infty)) > \overline{b}$ and in consequence $\rho_+(\infty) < w_0$. Symmetric arguments imply that $\rho_-(0) > \gamma_+(w_0)$. Applying $\rho_-(\cdot), \rho_+(\cdot)$ to these inequalities, we further deduce that

$$\rho_-(w_0) < \rho_-(\rho_+(\infty)),$$
$$\rho_+(\rho_-(0)) < \rho_+(\gamma_+(w_0)),$$

which, together with the assumption that $\rho_-(w_0) > \rho_+(\gamma_+(w_0))$, implies

$$\rho_+(\rho_-(0)) < \rho_+(\gamma_+(w_0)) < \rho_-(w_0) < \rho_-(\rho_+(\infty)),$$

contradicting $\boxed{\text{IV}}$ not holding. Hence Step 1 is complete once we prove the lemmas stated at the start, as follows.

Proof of Lemma 4.2. Consider $w < w_0$ with $\rho_+(\gamma_+(w_0)) < \rho_-(w_0)$. The latter implies $\mu_B((w_0, \rho_+(\gamma_+(w_0)))) < \underline{b}$, so that $\mu_B((0, \gamma_+(w_0))) < \underline{b}$. Suppose now that $\rho_+(\gamma_+(w)) \geq \rho_-(w)$. As ρ_- is decreasing and γ_+ is increasing, we have $\rho_-(w_0) < \rho_-(w) \leq \rho_+(\gamma_+(w))$ and $\overline{b} < \gamma_+(w) < \gamma_+(w_0)$. We then have

$$\begin{aligned}\underline{b} &= \mu_B\big((0, \rho_-(w)) \cup (\rho_+(\gamma_+(w)), \gamma_+(w))\big) \\ &= \mu_B\Big((0, \gamma_+(w_0)) \setminus \big[\big(\rho_-(w_0), \rho_+(\gamma_+(\omega))\big) \cup \big((\gamma_+(w), \gamma_+(w_0))\big)\big]\Big) \\ &\leq \mu_B\big((0, \gamma_+(w_0))\big) < \underline{b},\end{aligned}$$

which gives the desired contradiction. ■

Proof of Lemma 4.3. Define $\Gamma = (0, \underline{b}) \cup (\overline{b}, \infty)$ and consider $\mu_B(\Gamma)$. If both $\boxed{\text{I}}$ and $\boxed{\text{II}}$ hold, or more generally if $\gamma_-(z_0) = 0$ and $\gamma_+(w_0) = \infty$ for some $z_0 \geq \overline{b}$, $w_0 \leq \underline{b}$, we have

$$\mu_B((0, \rho_-(0)) \cup (\overline{b}, \infty)) \geq \mu_B((0, \rho_-(0)) \cup (z_0, \infty)) = \overline{b} \tag{4.1}$$

and

$$\mu_B((0, \underline{b}) \cup (\rho_+(\infty), \infty)) \leq \mu_B((0, w_0) \cup (\rho_+(\infty), \infty)) = \underline{b}. \tag{4.2}$$

Now suppose $\mu_B(\Gamma) < \overline{b}$. Then

$$\mu_B(\Gamma \cup (\underline{b}, \rho_-(0))) < \overline{b},$$

contradicting (4.1), and similarly if $\mu_B(\Gamma) > \underline{b}$,

$$\mu_B(\Gamma \cup (\rho_+(\infty), \overline{b})) > \underline{b}$$

contradicts (4.2). ■

Proof of Lemma 4.4. $\boxed{\text{IV}} \implies$ not $\boxed{\text{III}}$: Assume that both $\boxed{\text{IV}}$ and $\boxed{\text{III}}$ hold. From the definition of $\gamma_+(w_0)$ and $\rho_-(w_0)$ we have that $\mu_B(\Gamma_+) = \underline{b}$ for $\Gamma_+ = (0, \rho_-(w_0)) \cup (\rho_+(\gamma_+(w_0)), \gamma_+(w_0))$. This implies that $\gamma_+(w_0) \geq \rho_-(0)$ since otherwise

$$\mu_B(\Gamma_+) = \mu_B\Big((0, \rho_-(0)) \setminus \big\{ \big(\rho_-(w_0), \rho_+(\gamma_+(w_0))\big) \cup \big(\gamma_+(w_0), \rho_-(0)\big) \big\} \Big) > \underline{b},$$

where we also used the assumption $\rho_-(w_0) \leq \rho_+(\gamma_+(w_0))$. Likewise, using $\gamma_-(\gamma_+(w_0)) = w_0$, we see that $w_0 \leq \rho_+(\infty)$. Applying ρ_- to the last inequality and using our assumptions, we obtain

$$\rho_+(\rho_-(0)) < \rho_-(\rho_+(\infty)) \leq \rho_-(w_0) \leq \rho_+(\gamma_+(w_0)).$$

In consequence, $\gamma_+(w_0) < \rho_-(0)$, which gives the desired contradiction.

$\boxed{\text{III}} \implies$ not ($\boxed{\text{I}}$ or $\boxed{\text{II}}$): Suppose that $\boxed{\text{III}}$ and $\boxed{\text{II}}$ hold together. Let $w_1 < \underline{b}$ be the point given by $\boxed{\text{II}}$ such that $\gamma_+(w_1) = \infty$, and w_0 the point in $\boxed{\text{III}}$ such that $\gamma_-(\gamma_+(w_0)) = w_0$. Naturally, as $\gamma_-(\gamma_+(w_1)) = \gamma_-(\infty) = \underline{b} > w_1$ we have that $w_0 < w_1$. Observe also that $\mu_B((0, \rho_-(w_1)) \cup (\rho_+(\infty), \infty)) = \underline{b}$ and $\mu_B(\mathbb{R}) = S_0 \in (\underline{b}, \overline{b})$, which readily imply $\underline{b} < \rho_-(w_1) < \rho_+(\infty) < \overline{b}$. Let us further denote $r = \min\{\rho_-(w_0), \rho_+(\infty)\}$ and $R = \max\{\rho_-(w_0), \rho_+(\infty)\}$ so that finally, using our assumptions,

$$w_0 < w_1 < \underline{b} < \rho_-(w_1) < r \leq R \leq \rho_+(\gamma_+(w_0)) \leq \overline{b} \leq \gamma_+(w_0). \tag{4.3}$$

By definition we have

$$\int_{\rho_+(\infty)}^{\infty} (\overline{b} - u)\mu(\mathrm{d}u) = 0 = \int_{w_0}^{\rho_-(w_0)} (\overline{b} - u)\mu(\mathrm{d}u) + \int_{\rho_+(\gamma_+(w_0))}^{\infty} (\overline{b} - u)\mu(\mathrm{d}u).$$

Subtracting these two quantities, we arrive at

$$\int_{R}^{\rho_+(\gamma_+(w_0))} (\overline{b} - u)\mu(\mathrm{d}u) = \int_{w_0}^{r} (\overline{b} - u)\mu(\mathrm{d}u), \quad \text{and using (4.3), we deduce}$$

$$\mu\Big(\big(R, \rho_+(\gamma_+(w_0))\big) \Big) > \mu\Big((w_0, r) \Big). \tag{4.4}$$

Using the properties of our functions again, we have

$$\int_{-\infty}^{w_1} (u - \underline{b})\mu(\mathrm{d}u) + \int_{\rho_+(\infty)}^{\infty} (u - \underline{b})\mu(\mathrm{d}u) = 0 = \int_{-\infty}^{w_0} (u - \underline{b})\mu(\mathrm{d}u) + \int_{\rho_+(\gamma_+(w_0))}^{\gamma_+(w_0)} (u - \underline{b})\mu(\mathrm{d}u),$$

which after subtracting, using $\int_{w_0}^{w_1} (u - \underline{b})\mu(\mathrm{d}u) = -\int_{\rho_-(w_1)}^{\rho_-(w_0)} (u - \underline{b})\mu(\mathrm{d}u)$, yields

$$\int_{\rho_+(\infty)}^{\rho_+(\gamma_+(w_0))} (u - \underline{b})\mu(\mathrm{d}u) - \int_{\rho_-(w_1)}^{\rho_-(w_0)} (u - \underline{b})\mu(\mathrm{d}u) + \int_{\gamma_+(w_0)}^{\infty} (u - \underline{b})\mu(\mathrm{d}u) = 0. \tag{4.5}$$

The last term in (4.5) is positive, and for the first two terms, using (4.4), we have

$$\begin{aligned} \int_{R}^{\rho_+(\gamma_+(w_0))} (u - \underline{b})\mu(\mathrm{d}u) &\geq (R - \underline{b})\mu\Big(\big(R, \rho_+(\gamma_+(w_0))\big) \Big) \\ &> (R - \underline{b})\mu\Big((\rho_-(w_1), r) \Big) \geq \int_{\rho_-(w_1)}^{r} (u - \underline{b})\mu(\mathrm{d}u). \end{aligned} \tag{4.6}$$

This readily implies that the LHS of (4.5) is strictly positive, leading to the desired contradiction.

The case when $\boxed{\text{III}}$ and $\boxed{\text{I}}$ hold together is similar. ■

Step 2. Construction of relevant embeddings. Our strategy is now as follows. For each of the four exclusive cases $\boxed{\text{I}}$–$\boxed{\text{IV}}$ we construct a stopping time τ which solves the SEP for μ and such that for the price process $S_t := B_{\frac{t}{T-t}\wedge\tau}$ the appropriate superhedge $\overline{H}^I - \overline{H}^{IV}$ is in fact a perfect hedge. The stopping time τ will be a composition of stopping times, each of which is a solution to an embedding problem for a (rescaled) restriction of μ to appropriate intervals.

Suppose that $\boxed{\text{II}}$ holds. This embedding is closely related to the classical embedding of Azéma and Yor [2] used in the work of Brown, Hobson, and Rogers [8] on one-sided barrier options. There, it is used to achieve equality in the second inequality in (2.2). However, in order to also achieve equality in the first inequality we need to modify the embedding slightly. Let τ_1 be a UI embedding, in $(B_t)_{t\geq 0}$ with $B_0 = S_0$, of

$$\nu^1 = \mu|_{(w_0,\rho_+(\infty))} + p\delta_{\underline{b}}, \quad \text{where } p = 1 - \mu((w_0,\rho_+(\infty))),$$

which is centered in S_0. Let $\nu^2 = \frac{1}{p}\mu|_{\mathbb{R}_+\setminus(w_0,\rho_+(\infty))}$, which is a probability measure with $\nu^2_B(\mathbb{R}_+) = \underline{b}$, and let τ_2 be the Azéma–Yor embedding (cf. section 5 in [34]) of ν^2, i.e.,

$$\tau_2 = \inf\left\{t > 0 : \overline{B}_t \geq \nu^2_B([B_t,\infty))\right\},$$

which is a UI embedding of ν^2 when $B_0 = \underline{b}$. Note that $\nu^2_B([x,\infty)) = \mu_B([x,\infty))$ for $x \geq \rho_+(\infty)$ and that

$$\{\overline{B}_{\tau_2} \geq \overline{b}\} = \{B_{\tau_2} \geq \rho_+(\infty)\}, \quad \text{since } \nu^2_B((\rho_+(\infty),\infty)) = \overline{b}.$$

We define our final embedding as follows: we first embed ν^1, and then the atom in $\underline{b}$ is diffused into ν^2 using the Azéma–Yor procedure, i.e.,

$$\tau := \tau_1\mathbf{1}_{B_{\tau_1}\neq\underline{b}} + \tau_2\circ\tau_1\mathbf{1}_{B_{\tau_1}=\underline{b}}, \tag{4.7}$$

where $B_0 = S_0$. Clearly, τ is a UI embedding of μ, and $S_t := B_{\frac{t}{T-t}\wedge\tau}$ defines a model for the stock price which matches the given prices of calls and puts; i.e., $S_T \sim \mu$. Furthermore, $\{\overline{S}_T \geq \overline{b}\} = \{\overline{S}_T \geq \overline{b}, \underline{S}_T \leq \underline{b}\} = \{S_T \geq \rho_+(\infty)\}$, since $\rho_+(\infty) < \overline{b}$ implies $\{\tau = \tau_1\} \subseteq \{\overline{S}_T < \overline{b}\}$ and $\{\tau \neq \tau_1\} \subseteq \{\underline{S}_T \leq \underline{b}\}$. It follows that

$$\mathbf{1}_{\overline{S}_T\geq\overline{b},\,\underline{S}_T\leq\underline{b}} = \overline{H}^{II}(\rho_+(\infty)).$$

Suppose that $\boxed{\text{I}}$ holds. This is a mirror image of $\boxed{\text{II}}$. We first embed $\nu^1 = \mu|_{(\rho_-(0),z_0)} + p\delta_{\overline{b}}$, with $p = 1-\mu((\rho_-(0),z_0))$. Then the atom in $\overline{b}$ is diffused into $\mu|_{\mathbb{R}_+\setminus(\rho_-(0),z_0)}$ using the reversed Azéma–Yor stopping time (cf. section 5.3 in [34]). The resulting stopping time τ and the stock price model $S_t := B_{\frac{t}{T-t}\wedge\tau}$ satisfy $\mathbf{1}_{\overline{S}_T\geq\overline{b},\,\underline{S}_T\leq\underline{b}} = \overline{H}^I(\rho_-(0))$.

Suppose that $\boxed{\text{III}}$ holds. We describe the embedding in words before writing it formally. We first embed μ on $(\rho_-(w_0),\rho_+(\gamma_+(w_0)))$, or we stop when we hit $\overline{b}$ or $\underline{b}$. If we hit $\overline{b}$, then we embed μ on $(\gamma_+(w_0), \infty)$ or we run until we hit $\underline{b}$. Likewise, if we first hit $\underline{b}$, then we embed μ on $(0, w_0)$ or we run till we hit $\overline{b}$. Finally, from $\underline{b}$ and $\overline{b}$ we embed the remaining bits of μ.

We now formalize these ideas. Let

$$\nu^1 = p\delta_{\underline{b}} + \mu|_{(\rho_-(w_0),\rho_+(\gamma_+(w_0)))} + \big(1 - p - \mu(\rho_-(w_0), \rho_+(\gamma_+(w_0)))\big)\delta_{\overline{b}}, \tag{4.8}$$

where p is chosen so that $\nu^1_B(\mathbb{R}_+) = S_0$. Define two more measures,

$$\begin{aligned}\nu^2 &= \mu\big([w_0, \rho_-(w_0)]\big)\delta_{\underline{b}} + \mu|_{(\gamma_+(w_0),\infty)},\\ \nu^3 &= \mu|_{(0,w_0)} + \mu\big([\rho_+(\gamma_+(w_0)), \gamma_+(w_0)]\big)\delta_{\overline{b}},\end{aligned} \tag{4.9}$$

and note that by definition $\nu^2_B(\mathbb{R}_+) = \overline{b}$ and $\nu^3_B(\mathbb{R}_+) = \underline{b}$. Furthermore, as the barycenter of

$$\nu^3 + \mu|_{((\rho_-(w_0),\rho_+(\gamma_+(w_0)))} + \nu^2$$

is equal to the barycenter of μ, and from the uniqueness of p in (4.8), we deduce that

$$\nu^3(\mathbb{R}_+) = p \quad \text{and} \quad \nu^2(\mathbb{R}_+) = q = \big(1 - p - \mu(\rho_-(w_0), \rho_+(\gamma_+(w_0)))\big).$$

Let τ_1 be a UI embedding of ν^1 (for $B_0 = S_0$), τ_2 be a UI embedding of $\frac{1}{q}\nu^2$ (for $B_0 = \overline{b}$), and τ_3 be a UI embedding of $\frac{1}{p}\nu^3$ (for $B_0 = \underline{b}$). Further, let τ_4 and τ_5 be UI embeddings respectively of

$$\frac{1}{\mu\big((w_0, \rho_-(w_0))\big)}\mu|_{(w_0,\rho_-(w_0))} \quad \text{and} \quad \frac{1}{\mu\big((\rho_+(\gamma_+(w_0)), \gamma_+(w_0))\big)}\mu|_{(\rho_+(\gamma_+(w_0)),\gamma_+(w_0))},$$

where the starting points are respectively $B_0 = \underline{b}$ and $B_0 = \overline{b}$. We are ready to define our stopping time. Let $B_0 = S_0$, and write $H_z = \inf\{t : B_t = z\}$. We put

$$\begin{aligned}\tau := {} & \tau_1 \mathbf{1}_{\tau_1 < H_{\underline{b}} \wedge H_{\overline{b}}}\\ & + \tau_2 \circ \tau_1 \mathbf{1}_{H_{\overline{b}} = \tau_1} \mathbf{1}_{\tau_2 \circ \tau_1 < H_{\underline{b}}}\\ & + \tau_4 \circ \tau_2 \circ \tau_1 \mathbf{1}_{H_{\overline{b}} = \tau_1} \mathbf{1}_{H_{\underline{b}} = \tau_2 \circ \tau_1}\\ & + \tau_3 \circ \tau_1 \mathbf{1}_{H_{\underline{b}} = \tau_1} \mathbf{1}_{\tau_3 \circ \tau_1 < H_{\overline{b}}}\\ & + \tau_5 \circ \tau_3 \circ \tau_1 \mathbf{1}_{H_{\underline{b}} = \tau_1} \mathbf{1}_{H_{\overline{b}} = \tau_3 \circ \tau_1},\end{aligned} \tag{4.10}$$

and it is immediate from the properties of our measures that $B_\tau \sim \mu$ and $(B_{t\wedge\tau})$ is a UI martingale. Furthermore, with $S_t := B_{\frac{t}{T-t}\wedge\tau}$, we see that

$$\mathbf{1}_{\overline{S}_T \geq \overline{b},\, \underline{S}_T \leq \underline{b}} = \overline{H}^{III}\big(\gamma_+(w_0), \rho_+(\gamma_+(w_0)), \rho_-(w_0), w_0\big), \quad \text{a.s.}$$

Finally, suppose that $\boxed{\text{IV}}$ holds. In this case, we initially run to $\{\underline{b}, \overline{b}\}$ without stopping any mass. Then, from $\overline{b}$, we either run to $\underline{b}$ or embed μ on $(\rho_-(0), \infty)$. The mass which is at $\underline{b}$ after the first step is either run to $\overline{b}$ or used to embed μ on $(0, \rho_+(\infty))$. The mass which remains at $\underline{b}$ and $\overline{b}$ is then used to embed the remaining part of μ on $(\rho_+(\infty), \rho_-(0))$.

To begin with, we define the measures

$$\nu^1 = \left[\frac{S_0 - \underline{b}}{\overline{b} - \underline{b}} - \mu((\rho_-(0), \infty))\right] \delta_{\underline{b}} + \mu|_{(\rho_-(0),\infty)},$$
$$\nu^2 = \left[\frac{\overline{b} - S_0}{\overline{b} - \underline{b}} - \mu((0, \rho_+(\infty))\right] \delta_{\overline{b}} + \mu|_{(0,\rho_+(\infty))}.$$

Then ν^1 is a measure, since $\mu((\rho_-(0), \infty)) < \frac{S_0 - \underline{b}}{\overline{b} - \underline{b}}$: noting that $\overline{b} < \rho_-(0)$, we get

$$\begin{aligned} 0 &= \int_0^{\rho_-(0)} (u - S_0)\mu(\mathrm{d}u) + \int_{\rho_-(0)}^{\infty} (u - S_0)\mu(\mathrm{d}u) \\ &\geq (\underline{b} - S_0)\mu((0, \rho_-(0))) + (\overline{b} - S_0)\mu((\rho_-(0), \infty)), \end{aligned}$$

and the statement follows. Moreover, we can see that $\nu^1_B(\mathbb{R}_+) = \overline{b}$:

$$\begin{aligned} &\int_{\rho_-(0)}^{\infty} (u - \overline{b})\mu(\mathrm{d}u) + \frac{S_0 - \underline{b}}{\overline{b} - \underline{b}}(\underline{b} - \overline{b}) - \mu((\rho_-(0), \infty))(\underline{b} - \overline{b}) \\ &\quad = \int_{\rho_-(0)}^{\infty} (u - \underline{b})\mu(\mathrm{d}u) + (\underline{b} - S_0) \\ &\quad = (S_0 - \underline{b}) - \int_0^{\rho_-(0)} (u - \underline{b})\mu(\mathrm{d}u) + (\underline{b} - S_0) = 0. \end{aligned}$$

Similar results hold for ν^2.

Consequently, we can construct the first stages of the embedding. The final stage is to run from $\underline{b}$ and $\overline{b}$ to embed the remaining mass. Of course, it does not matter exactly how we do this from the optimality point of view, since these paths have already struck both barriers, but we do need to check that the embedding is possible. It is clear that the means and probabilities match, but unlike the previously considered cases, we now have initial mass in two places, and the existence of a suitable embedding is not trivial. To resolve this, we note the following: suppose we can find a point $z^* \in (\rho_+(\rho_-(0)), \rho_-(\rho_+(\infty)))$ such that $\nu^1(\{\underline{b}\}) = \mu((\rho_+(\infty), z^*))$. Then because $\mu_B((\rho_+(\infty), \rho_-(\rho_+(\infty)))) = \underline{b}$, we can find $z_1 \in (\rho_+(\infty), z^*)$ such that the measure

$$\nu^3 = \mu|_{(\rho_+(\infty),z_1)} + (\nu^1(\{\underline{b}\}) - \mu((\rho_+(\infty), z_1)))\delta_{z^*}$$

has barycenter $\underline{b}$. There is a similar construction for ν^4, a point z_2 which will embed mass from ν^2 at $\overline{b}$ to μ on $(z_2, \rho_-(0))$, and an atom at z^*. In the final stage, we can then embed the mass from z^* to (z_1, z_2).

It remains to show that we can find such a point z^*. To do this, we check that there is sufficient mass being stopped at $\underline{b}$ at the end of the second step (i.e., paths which have first hit $\overline{b}$ and then $\underline{b}$). Specifically, we need to show that

$$\frac{S_0 - \underline{b}}{\overline{b} - \underline{b}} - \mu((\rho_-(0), \infty)) \geq \mu((\rho_+(\infty), \rho_+(\rho_-(0)))).$$

Rearranging and using the definitions of the functions ρ_+ and ρ_-, this is equivalent to

$$\begin{aligned}(S_0 - \underline{b}) &\geq \int_{\rho_+(\infty)}^{\infty} (u - \underline{b})\, \mu(\mathrm{d}u) - \int_{\rho_+(\rho_-(0))}^{\rho_-(0)} (u - \underline{b})\mu(\mathrm{d}u) \\ &\geq \int_{(\rho_+(\infty),\rho_+(\rho_-(0)))\cup(\rho_-(0),\infty)} (u - \underline{b})\, \mu(\mathrm{d}u) \\ &\geq (S_0 - \underline{b}) - \int_{(0,\rho_+(\infty))\cup(\rho_+(\rho_-(0)),\rho_-(0))} (u - \underline{b})\, \mu(\mathrm{d}u).\end{aligned}$$

Using the definitions of the appropriate functions, this can be seen to be equivalent to

$$0 \leq \int_{\rho_+(\rho_-(0))}^{\rho_-(\rho_+(\infty))} (u - \underline{b})\, \mu(dy),$$

which follows since $\rho_+(\rho_-(0)) > \underline{b}$. The construction of the appropriate stopping time, and its optimality, follow as previously. This ends the proof of Theorem 2.2. ■

In order to prove Theorem 2.4 we start with an auxiliary lemma.

Lemma 4.5. *Either we may construct an embedding of μ under which the process never hits both $\overline{b}$ and $\underline{b}$, or*

$$\inf\{v \in [\underline{b}, \overline{b}] : \psi(v) < \infty\} \geq \inf\{v \in [\underline{b}, \overline{b}] : \theta(v) > -\infty\}, \tag{4.11}$$

$$\sup\{v \in [\underline{b}, \overline{b}] : \psi(v) < \infty\} \geq \sup\{v \in [\underline{b}, \overline{b}] : \theta(v) > -\infty\}, \tag{4.12}$$

and we may then write

$$\underline{v} = \inf\{v \in [\underline{b}, \overline{b}] : \psi(v) < \infty\} \leq \sup\{v \in [\underline{b}, \overline{b}] : \theta(v) > -\infty\} = \overline{v}, \tag{4.13}$$

where $\underline{v}, \overline{v}$ are given in (2.37).

Proof. We begin by showing that if $\theta(v) = -\infty$ for all $v \in [\underline{b}, \overline{b}]$, then there exists an embedding of μ which does not hit both $\underline{b}$ and $\overline{b}$.

For $w \geq \overline{b}$, define $\alpha_*(w)$ to be the mass that must be placed at $\underline{b}$ in order for the barycenter of this mass plus μ on $(\overline{b}, w)$ to be $\overline{b}$, so that $\alpha_*(w)$ satisfies

$$\alpha_*(w)\underline{b} + \int_{\overline{b}}^{w} u\, \mu(\mathrm{d}u) = \overline{b}\left(\alpha_*(w) + \mu((\overline{b}, w))\right).$$

It follows that $\alpha_*(w)$ exists, although there is no guarantee that it is less than $1 - \mu((\overline{b}, w))$. In addition, define $\beta^*(w)$ to be

$$\beta^*(w) = \inf\left\{\beta \in [\underline{b}, \overline{b}] : \int_{(\underline{b},\beta)\cup(\overline{b},w)} u\, \mu(\mathrm{d}u) = \overline{b} \int_{(\underline{b},\beta)\cup(\overline{b},w)} \mu(\mathrm{d}u)\right\}.$$

If this is finite, then it is the point at which $\mu_B((\underline{b}, \beta^*(w)) \cup (\overline{b}, w)) = \overline{b}$. Note also that $\beta^*(w)$ is increasing as a function of w and is continuous when $\beta^*(w) < \infty$. Define

$$\begin{aligned}p_*(w) &= \mu((\overline{b}, w)) + \alpha_*(w), \\ p^*(w) &= \mu((\underline{b}, \beta^*(w)) \cup (\overline{b}, w)).\end{aligned}$$

Suppose initially that $\beta^*(w) \leq \overline{b}$ for all $w \geq \overline{b}$. Then we may assign the following interpretations to these quantities: $p_*(w)$ is the smallest amount of mass that we can start at $\overline{b}$ and run to embed μ on $(\overline{b}, w)$, $(\underline{b}, \cdot)$ and an atom at $\underline{b}$, and $p^*(w)$ is the largest amount of mass that we may do this with; the smallest amount is attained by running all the mass below $\overline{b}$ to $\underline{b}$, while the largest probability is attained by running all this mass to $(\underline{b}, \beta^*(w))$. The assumption that $\beta^*(w) \leq \overline{b}$ implies that this upper bound does not run out of mass to embed. Moreover, by adjusting the size of the atom at $\underline{b}$, we can embed an atom of any size between $p_*(w)$ and $p^*(w)$ from $\overline{b}$ in this way. Recalling the definition of $\theta(v)$, we conclude that there exists v such that $\theta(v) = w$ if and only if $p_*(w) \leq \frac{S_0 - \underline{b}}{\overline{b} - \underline{b}} \leq p^*(w)$. Finally, note that the functions $p_*(w)$ and $p^*(w)$ are both increasing in w, and further that $p_*(\overline{b}) = 0$. Consequently, if there is no v such that $\theta(v) > -\infty$, and if $\beta^*(w) \leq \overline{b}$ for all $w \geq \overline{b}$, we must have $p^*(\infty) := \lim_{w \to \infty} p^*(w) < \frac{S_0 - \underline{b}}{\overline{b} - \underline{b}}$.

So suppose $p^*(\infty) < \frac{S_0 - \underline{b}}{\overline{b} - \underline{b}}$. We now construct an embedding as follows: from S_0, we initially run to either $\overline{b}$ or

$$b_* = \frac{S_0 - p^*(\infty)\overline{b}}{1 - p^*(\infty)}.$$

Since $p^*(\infty) < \frac{S_0 - \underline{b}}{\overline{b} - \underline{b}}$, then $b_* \in (\underline{b}, S_0)$, and the probability that we hit $\overline{b}$ before b_* is $p^*(\infty)$. In addition, by the definition of $\beta^*(w)$, we deduce that the set $(\underline{b}, \beta^*(\infty)] \cup [\overline{b}, \infty)$ is given mass $p^*(\infty)$ by μ, and that the barycenter of μ on this set is $\overline{b}$. We may therefore embed the paths from $\overline{b}$ to this set, and the paths from b_* to the remaining intervals, $[0, \underline{b}] \cup (\beta^*(\infty), \overline{b})$, and we note that no paths will hit both $\overline{b}$ and $\underline{b}$.

Suppose instead that $\beta^*(w_0) = \overline{b}$ for some w_0, with $p^*(w_0) \leq \frac{S_0 - \underline{b}}{\overline{b} - \underline{b}}$. (If the latter condition does not hold, then using the fact that $\beta^*(w)$ is left-continuous and increasing, we can find a w such that $\beta^*(w) < \overline{b}$ and $p^*(w) = \frac{S_0 - \underline{b}}{\overline{b} - \underline{b}}$, and therefore, by the arguments above, there exists v with $\theta(v) > -\infty$.) We may then continue to construct measures with barycenter $\overline{b}$, which are equal to μ on $(\underline{b}, w)$ for $w > w_0$ and have a compensating atom at $\underline{b}$. As we increase w, eventually either w reaches ∞ or the mass of the measure reaches $\frac{S_0 - \underline{b}}{\overline{b} - \underline{b}}$. In the latter case, we know $\theta(\overline{b}) = w$, contradicting $\theta(v) = -\infty$ for all $v \in [\underline{b}, \overline{b}]$. So consider the former case: we obtained that the measure which is μ on $(\underline{b}, \infty)$ with a further atom at $\underline{b}$ to give barycenter $\overline{b}$ has total mass (p, say) less than $\frac{S_0 - \underline{b}}{\overline{b} - \underline{b}}$. We show that this is impossible: divide μ into its restriction to $(0, z)$ and $[z, \infty)$, where z is chosen so that $\mu([z, \infty)) = p$. Then $z < \underline{b}$, and the barycenter of the restriction to $[z, \infty)$ is strictly smaller than the barycenter of the measure with the mass on $[z, \underline{b})$ placed at $\underline{b}$, which is the measure described above and which has barycenter $\overline{b}$. Additionally, the barycenter of the lower restriction of μ must be strictly smaller than $\underline{b}$. Moreover, we may calculate the barycenter of μ by considering the barycenter of the two restrictions; since μ has mean (and therefore barycenter) S_0, we must have

$$S_0 = (1-p)\mu_B((0,z)) + p\mu_B([z,\infty)) < (1-p)\underline{b} + p\overline{b} < \underline{b}\frac{\overline{b} - S_0}{\overline{b} - \underline{b}} + \overline{b}\frac{S_0 - \underline{b}}{\overline{b} - \underline{b}} = S_0,$$

which is a contradiction.

We conclude that, if $\{v \in [\underline{b}, \overline{b}] : \theta(v) > -\infty\}$ is empty, there is an embedding of μ which does not hit both $\overline{b}$ and $\underline{b}$. A similar result follows for $\psi(v)$. In particular, if we assume that there is no such embedding, then there exists v such that $\psi(v) < \infty$, and (not necessarily the same) v such that $\theta(v) > -\infty$. We now wish to show that (4.11) holds. Suppose not. Then

$$v_* := \inf\{v \in [\underline{b}, \overline{b}] : \psi(v) < \infty\} < \inf\{v \in [\underline{b}, \overline{b}] : \theta(v) > -\infty\} =: v^*.$$

Moreover, we can deduce from the definition of $\theta(v)$ that since $v^* > \underline{b}$, we must have $\theta(v^*) = \infty$. Now consider the barycenter of the measure which is taken by running from S_0 to $\underline{b}$ and $\overline{b}$, and then from $\underline{b}$ to $(\psi(v_*), \underline{b}) \cup (v_*, \overline{b})$, with a compensating mass at $\overline{b}$, so that the measure has barycenter $\underline{b}$, and from $\overline{b}$ to $(\underline{b}, v^*) \cup (\overline{b}, \theta(v^*)) = (\underline{b}, v^*) \cup (\overline{b}, \infty)$, with a compensating mass at $\underline{b}$, so that the measure has barycenter $\overline{b}$. Then the whole law of the resulting process must have mean S_0, since this can be done in a uniformly integrable way, but the resulting distribution is at least μ on $(\psi(v_*), \infty)$ (it is twice μ on (v_*, v^*), has atoms at $\underline{b}$ and $\overline{b}$, and is μ elsewhere), and zero on $(0, \psi(v_*))$, so must have mean greater than S_0, which is a contradiction. A similar argument shows (4.12). Hence, we may conclude (still under the assumption that there is no embedding which never hits both $\underline{b}$ and $\overline{b}$) that the equalities in (4.13) hold. It remains to show the inequality when $\underline{v}, \overline{v} \in (\underline{b}, \overline{b})$. However, this is now almost immediate: the forms of $\overline{v}, \underline{v}$ imply that μ gives mass $\frac{\overline{b}-S_0}{\overline{b}-\underline{b}}$ to the set $(\psi(\underline{v}), \underline{b}) \cup (\underline{v}, \overline{b})$ and gives mass $\frac{S_0-\underline{b}}{\overline{b}-\underline{b}}$ to the set $(\underline{b}, \overline{v}) \cup (\overline{b}, \theta(\overline{v}))$. If $\overline{v} < \underline{v}$, this implies that μ gives mass 1 to the set $(\psi(\underline{v}), \overline{v}) \cup (\underline{v}, \theta(\overline{v})) \subsetneq [0, \infty)$, contradicting the positivity of μ. ■

Proof of Theorem 2.4. From Lemma 4.5 case $\boxed{\text{IV}}$ and the last statement of the theorem follow. Assume from now on that $\underline{v} \le \overline{v}$. We note first that $\psi(v)$ and $\theta(v)$ are both continuous and decreasing on $[\underline{v}, \overline{v}]$, and consequently $\kappa(v)$ is also continuous and decreasing as a function of v on $[\underline{v}, \overline{v}]$. It follows that the three cases $\kappa(\underline{v}) < \underline{v}$, $\kappa(\overline{v}) > \overline{v}$, and the existence of $v_0 \in [\underline{v}, \overline{v}]$ such that $\kappa(v_0) = v_0$ are exclusive and exhaustive. We consider each case separately:

$\boxed{\text{I}}$ Suppose that there exists $v_0 \in [\underline{v}, \overline{v}]$ such that $\kappa(v_0) = v_0$. By the definition of $\psi(v)$, we can run all the mass initially from S_0 to $\{\underline{b}, \overline{b}\}$ and then embed (in a UI way) from $\underline{b}$ to $(\psi(v_0), \underline{b}) \cup (v_0, \overline{b})$ and a compensating atom at $\overline{b}$ with the remaining mass, and similarly from $\overline{b}$ to $(\underline{b}, v_0) \cup (\overline{b}, \theta(v_0))$ with an atom at $\underline{b}$. The mass now at $\underline{b}$ and $\overline{b}$ can then be embedded in the remaining tails in a suitable way—the means and masses must agree, since the initial stages were embedded in a UI manner, and the remaining mass all lies outside $[\underline{b}, \overline{b}]$. We denote by τ the stopping time which achieves the embedding.

Now we compare both sides of the inequality in (2.10), where we choose $K_1 = \theta(v_0), K_2 = \psi(v_0)$, and therefore, as a consequence of the definition of $\kappa(v)$, we also have $K_3 = v_0$. The key observation is now that the mass is stopped only at points where the inequality is an equality: mass which hits $\overline{b}$ initially either stops in the interval $(\underline{b}, K_3) \cup (\overline{b}, K_1)$, when there is equality in (2.10), or it goes on to hit $\overline{b}$, and from this point also continues to the tails $(0, K_2) \cup (K_1, \infty)$, where there is again equality between both sides of (2.10). Taking expectations on the RHS of (2.10), we get the terms on the RHS of (2.38), and we conclude that (2.38) holds in the market model $S_t := B_{\tau \wedge \frac{t}{T-t}}$.

$\boxed{\text{II}}$ Suppose now that $\kappa(\overline{v}) > \overline{v}$. Then we must have $\overline{v} = \sup\{v \in [\underline{b}, \overline{b}] : \theta(v) > -\infty\}$ by Lemma 4.5, and then $\overline{v} < \kappa(\overline{v}) \le \overline{b}$. So, by the definition of $\theta(v)$, since $\overline{v}$ exists and is less than $\overline{b}$, we must have

$$\int_{(\underline{b},\overline{v})\cup(\overline{b},\theta(\overline{v}))} u\,\mu(\mathrm{d}u) = \overline{b}\frac{S_0-\underline{b}}{\overline{b}-\underline{b}} \quad \text{and} \quad \mu((\underline{b},\overline{v})\cup(\overline{b},\theta(\overline{v}))) = \frac{S_0-\underline{b}}{\overline{b}-\underline{b}},$$

and we can embed from $\overline{b}$ (having initially run to $\{\underline{b},\overline{b}\}$) to $(\underline{b},\overline{v})\cup(\overline{b},\theta(\overline{v}))$ *without* leaving an atom at $\underline{b}$. Similarly, we can also run from $\underline{b}$ to $(\psi(\overline{v}),\underline{b})\cup(\overline{v},\overline{b})$ with an atom at $\overline{b}$. The atom can then be embedded in the tails $(0,\psi(\overline{v}))\cup(\theta(\overline{v}),\infty)$ in a UI manner. We now need to show that when we take $K_3=\overline{v}, K_1=\theta(\overline{v})$, and $K_2=\psi(\overline{v})$ we get the required equality in (2.39). The main difference from the above case occurs in the case where we hit $\overline{b}$ initially and then hit $\underline{b}$: we no longer need equality in (2.10), since this no longer occurs in our optimal construction; however, what remains to be checked is that the inequality does hold on this set. Specifically, we need to show that

$$1 \geq \alpha_0+\alpha_1(K_2-S_0)-(\alpha_3-\alpha_3+\alpha_1)(K_2-\overline{b})+(\alpha_3-\alpha_2)(K_2-\underline{b}).$$

Using (2.20) and (2.25), we see that this occurs when

$$\frac{(K_3-K_2)(K_1-\underline{b})}{(\overline{b}-K_2)(K_1-K_3)} \leq 1,$$

which rearranges to give

$$K_3 \leq \overline{b}\frac{K_1-\underline{b}}{(K_1-\underline{b})+(\overline{b}-K_2)}+\underline{b}\frac{\overline{b}-K_2}{(K_1-\underline{b})+(\overline{b}-K_2)}.$$

This is satisfied by our choice of $\overline{v}$ as K_3, and $K_1=\theta(\overline{v}), K_2=\psi(\overline{v})$.

$\boxed{\text{III}}$ This is symmetric to case $\boxed{\text{II}}$. ■

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. Azéma and M. Yor, *Une solution simple au problème de Skorokhod*, in Séminaire de Probabilités, XIII, Lecture Notes in Math. 721, Springer, Berlin, 1979, pp. 90–115.

[3] J. Azéma and M. Yor, *Le problème de Skorokhod: Compléments à "Une solution simple au problème de Skorokhod"*, in Séminaire de Probabilités, XIII, Lecture Notes in Math. 721, Springer, Berlin, 1979, pp. 625–633.

[4] D. Blackwell and L. E. Dubins, *A converse to the dominated convergence theorem*, Illinois J. Math., 7 (1963), pp. 508–514.

[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 option prices*, J. Business, 51 (1978), pp. 621–651.

[7] H. Brown, D. Hobson, and L. C. G. Rogers, *The maximum maximum of a martingale constrained by an intermediate law*, Probab. Theory Related Fields, 119 (2001), pp. 558–578.

[8] H. Brown, D. Hobson, and L. C. G. Rogers, *Robust hedging of barrier options*, Math. Finance, 11 (2001), pp. 285–314.

[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 D. B. Madan, *A note on sufficient conditions for no arbitrage*, Finance Res. Lett., 2 (2005), pp. 125–130.

[12] A. M. G. Cox, *Arbitrage bounds*, in Encyclopedia of Quantitative Finance, R. Cont, ed., Wiley, New York, 2010, pp. 53–61.

[13] A. M. G. Cox, D. G. Hobson, and J. Obłój, *Pathwise inequalities for local time: Applications to Skorokhod embeddings and optimal stopping*, Ann. Appl. Probab., 18 (2008), pp. 1870–1896.

[14] A. M. G. Cox and J. Obłój, *Robust pricing and hedging of double no-touch options*, Finance Stoch., (2010), to appear.

[15] J. Cvitanić and I. Karatzas, *Hedging and portfolio optimization under transaction costs: A martingale approach*, Math. Finance, 6 (1996), pp. 133–165.

[16] M. H. A. Davis and D. G. Hobson, *The range of traded option prices*, Math. Finance, 17 (2007), pp. 1–14.

[17] M. H. A. Davis, W. Schachermayer, and R. G. Tompkins, *Pricing, no-arbitrage bounds and robust hedging of installment options*, Quant. Finance, 1 (2001), pp. 597–610.

[18] E. Derman, D. Ergener, and I. Kani, *Static options replication*, J. Derivatives, 2 (1995), pp. 78–95.

[19] B. Dumas, J. Fleming, and R. E. Whaley, *Implied volatility functions: Empirical tests*, J. Finance, 53 (1998), pp. 2059–2106.

[20] B. Dupire, *Arbitrage bounds for volatility derivatives as a free boundary problem*, lecture, 2005; available online at http://www.math.kth.se/pde_finance/presentations/Bruno.pdf.

[21] B. Engelmann, M. R. Fengler, M. Nalholm, and P. Schwendner, *Static versus dynamic hedges: An empirical comparison for barrier options*, Rev. Deriv. Res., 9 (2006), pp. 239–264.

[22] B. Engelmann, M. R. Fengler, and P. Schwendner, *Better Than Its Reputation: An Empirical Hedging Analysis of the Local Volatility Model for Barrier Options*, manuscript, 2006; available online at http://ssrn.com/abstract=935951.

[23] J. Fink, *An examination of the effectiveness of static hedging in the presence of stochastic volatility*, J. Futures Markets, 23 (2003), pp. 859–890.

[24] H. Föllmer and M. Schweizer, *Hedging of contingent claims under incomplete information*, in Applied Stochastic Analysis, M. H. A. Davis and R. J. Elliott, eds., Stochastics Monogr. 5, Gordon and Breach, London, New York, 1991, pp. 389–414.

[25] A. Giese and J. Maruhn, *Cost-optimal static super-replication of barrier options: An optimisation approach*, J. Comput. Finance, 10 (2007), pp. 71–97.

[26] 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.

[27] D. Hobson, *The Skorokhod embedding problem and model-independent bounds for option prices*, in Paris-Princeton Lectures on Mathematical Finance 2010, R. Carmona, I. Çinlar, E. Ekeland, E. Jouini, J. Scheinkman, and N. Touzi, eds., Lecture Notes in Math. 2003, Springer, New York, 2010, pp. 267–318.

[28] D. G. Hobson, *Robust hedging of the lookback option*, Finance Stoch., 2 (1998), pp. 329–347.

[29] S. D. Hodges and A. Neuberger, *Optimal replication of contingent claims under transaction costs*, Rev. Futures Markets, 8 (1989), pp. 222–239.

[30] A. Ilhan, M. Jonsson, and R. Sircar, *Optimal static-dynamic hedges for exotic options under convex risk measures*, Stochastic Process. Appl., 119 (2009), pp. 3608–3632.

[31] R. C. Merton, *Theory of rational option pricing*, Bell J. Econom. Management Sci., 4 (1973), pp. 141–183.

[32] I. Monroe, *Processes that can be embedded in Brownian motion*, Ann. Probab., 6 (1978), pp. 42–56.

[33] M. Nalholm and R. Poulsen, *Static hedging of barrier options under general asset dynamics: Unification and application*, J. Derivatives, 13 (2006), pp. 46–60.

[34] J. Obłój, *The Skorokhod embedding problem and its offspring*, Probab. Surv., 1 (2004), pp. 321–390.

[35] J. Obłój, *The Skorokhod embedding problem*, in Encyclopedia of Quantitative Finance, R. Cont, ed., Wiley, New York, 2010, pp. 1653–1657.

[36] E. Perkins, *The Cereteli-Davis solution to the H^1-embedding problem and an optimal embedding in Brownian motion*, in Proceedings of the Seminar on Stochastic Processes (Gainesville, FL, 1985), Birkhäuser Boston, Cambridge, MA, 1986, pp. 172–223.

[37] L. C. G. Rogers, *The joint law of the maximum and terminal value of a martingale*, Probab. Theory Related Fields, 95 (1993), pp. 451–466.

[38] R. Tompkins, *Static versus dynamic hedging of exotic options: An evaluation of hedge performance via simulation*, Netexposure, 1 (1997), pp. 1–28.

[39] F. Ulmer, *Performance of Robust Model-free Hedging via Skorokhod Embeddings of Digital Double Barrier Options*, Master's thesis, Mathematical Institute, University of Oxford, Oxford, UK, 2010.

[40] P. Vallois, *On the joint distribution of the supremum and terminal value of a uniformly integrable martingale*, in Stochastic Processes and Optimal Control (Friedrichroda, 1992), Stochastics Monogr. 7, Gordon and Breach, Montreux, France, 1993, pp. 183–199.

[41] 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.

SIAM J. Financial Math.
Vol. 2, pp. 183–212

Optimal Execution in a General One-Sided Limit-Order Book*

Silviu Predoiu†, Gennady Shaikhet‡, and Steven Shreve†

Abstract. We construct an optimal execution strategy for the purchase of a large number of shares of a financial asset over a fixed interval of time. Purchases of the asset have a nonlinear impact on price, and this is moderated over time by resilience in the limit-order book that determines the price. The limit-order book is permitted to have arbitrary shape. The form of the optimal execution strategy is to make an initial lump purchase and then purchase continuously for some period of time during which the rate of purchase is set to match the order book resiliency. At the end of this period, another lump purchase is made, and following that there is again a period of purchasing continuously at a rate set to match the order book resiliency. At the end of this second period, there is a final lump purchase. Any of the lump purchases could be of size zero. A simple condition is provided that guarantees that the intermediate lump purchase is of size zero.

Key words. optimal execution, limit-order book, price impact

AMS subject classifications. 91B26, 91G80, 49K45, 90C25

DOI. 10.1137/10078534X

1. Introduction.

We consider optimal execution over a fixed time interval of a large asset purchase in the face of a one-sided limit-order book. We assume that the *ask price* (sometimes called the *best ask price*) for the underlying asset is a continuous martingale that undergoes two adjustments during the period of purchase. The first adjustment is that orders consume a part of the limit-order book, and this increases the ask price for subsequent orders. The second adjustment is that resilience in the limit-order book causes the effect of these prior orders to decay over time. In this paper, there is no permanent effect from the purchase we model. However, the temporary effect requires infinite time to disappear completely.

We assume that there is a fixed shadow limit-order book shape toward which resilience returns the limit-order book. At any time, the actual limit-order book relative to the martingale component of the ask price has this shape but with some left-hand part missing due to prior purchases. An investor is given a period of time and a target amount of asset to be purchased within that period. His goal is to distribute his purchasing over the period in order to minimize the expected cost of purchasing the target. We permit purchases to occur in lumps or to be spread continuously over time. We show that the optimal execution strategy consists of three lump purchases, one or more of which may be of size zero, i.e., does not occur. One of these lump purchases is made at the initial time, one at an intermediate time,

*Received by the editors February 8, 2010; accepted for publication (in revised form) December 14, 2010; published electronically March 9, 2011.

http://www.siam.org/journals/sifin/2/78534.html

†Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (spredoiu@andrew.cmu.edu, shreve@andrew.cmu.edu). The third author's research was partially supported by the National Science Foundation under grant DMS-0903475.

‡Department of Mathematics and Statistics, Carleton University, Ottawa, ON, K1S 5B6, Canada (gennady@math.carleton.ca).

and one at the final time. Between these lump purchases, the optimal strategy purchases at a constant rate matched to the limit-order book recovery rate so that the ask price minus its martingale component remains constant. We provide a simple condition under which the intermediate lump purchase is of size zero (see Theorem 4.2 and Remark 4.4).

Bouchaud, Farmer, and Lillo [9] provide a survey of the empirical behavior of limit-order books. Dynamic models for optimal execution designed to capture some of this behavior have been developed by several authors, including Bertsimas and Lo [8], Almgren and Chriss [6, 7], Grinold and Kahn [15] (Chapter 16), Almgren [5], Obizhaeva and Wang [10], and Alfonsi, Fruth, and Schied [1, 4]. Trading in [8] is on a discrete-time grid, and the price impact of a trade is linear in the size of the trade and is permanent. In [8], the expected-cost-minimizing liquidation strategy for an order is to divide the order into equal pieces, one for each trading date. Trading in [6, 7] is also on a discrete-time grid, and there are linear permanent and temporary price impacts. In [6, 7], the variance of the cost of execution is taken into account. This leads to the construction of an efficient frontier of trading strategies. In [15] and [5], trading takes place continuously, and finding the optimal trading strategy reduces to a problem in the calculus of variations.

Other authors focus on the possibility of price manipulation, an idea that traces back to Huberman and Stanzl [16]. Price manipulation is a way of starting with zero shares and using a strategy of buying and selling so as to end with zero shares while generating income. Gatheral, Schied, and Slynko [13] permit continuous trading and use an integral of a kernel with respect to the trading strategy to capture the resilience of the book. In such a model, Gatheral [12] shows that exponential decay of market impact and absence of price manipulation opportunities are compatible only with linear market impact. In [14], this result is reconciled with the nonlinear market impact in models such as [2, 3, 4] and this paper. Alfonsi, Schied, and Slynko [3] discover in a discrete-time version of the model of [13] that, even under conditions that prevent price manipulation, it may still be optimal to execute intermediate sells while trying to execute an overall buy order, and they provide conditions to rule out this phenomenon.

For the type of model we consider in this paper, based on a shadow limit-order book, Alfonsi and Schied [2] show that price manipulation is not possible under very general conditions. Furthermore, it is never advantageous to execute intermediate sells while trying to execute an overall buy order. In [2], trading takes place at finitely many stopping times, and execution is optimized over these stopping times. In the present paper, where trading is continuous, we do not permit intermediate sells. This simplification of the model is justified by Remark 3.1, which argues that intermediate sells cannot reduce the total cost.

The present paper is inspired by Obizhaeva and Wang [10], who explicitly model the one-sided limit-order book as a means of capturing the price impact of order execution. Empirical evidence for the model of [10] and its generalizations by Alfonsi, Fruth, and Schied [1, 4] and Alfonsi and Schied [2] are reported in [1, 2, 4, 10]. In [10] and [1], the limit-order book has a block shape, and in this case the price impact of a purchase is linear, the same as in [8, 7]. However, the change of mindset is important because it focuses attention on the shape of the limit-order book as the determinant of price impact, rather than making assumptions about the price impact directly. This change of mindset was exploited in [2, 4], where more general limit-order book shapes are permitted, subject to the condition discussed in Remark 4.4.

In [2, 4], trading is on a discrete-time grid, and it is shown that for an optimal purchasing strategy all purchases except the first and last ones are of the same size. Furthermore, the size of the intermediate purchases is chosen so that the price impact of each purchase is exactly offset by the order book resiliency before the next purchase. Similar results are obtained in [2], although here trades are executed at stopping times.

In contrast to [2, 4, 10], we permit the order book shape to be completely general. However, in our model all price impact is transient; [4, 10] also include the possibility of a permanent linear price impact. In contrast to [2, 4], we do not assume that the limit-order book has a positive density. It can be discrete or continuous and can have gaps. In contrast to [2, 4, 10], we permit the resilience in the order book to be a function of the adjustments to the martingale component of the ask price. Weiss [18] argues in a discrete-time model that this conforms better to empirical observations.

Finally, we set up our model so as to allow for both discrete-time and continuous-time trading, whereas [4, 10] begin with discrete-time trading and then study the limit of their optimal strategies as trading frequency approaches infinity. The simplicity afforded by a fully continuous model is evident in the analysis below. In particular, we provide constructive proofs of Theorems 4.2 and 4.5 that describe the form of the optimal purchasing strategies.

Section 2 of this paper presents our model. It contains the definition of the cost of purchasing in our more general framework, and that is preceded by a justification of the definition. Section 3 shows that randomness can be removed from the optimal purchasing problem and reformulates the cost function into a convenient form. In section 4, we solve the problem, first in the case that is analogous to the one solved by [4] and then in full generality. Sections 4.1 and 4.3 contain examples.

2. The model. Let T be a positive constant. We assume that the ask price of some asset, in the absence of the large investor modeled by this paper, is a continuous nonnegative martingale A_t, $0 \le t \le T$, relative to some filtration $\{\mathcal{F}_t\}_{0 \le t \le T}$ satisfying the usual conditions. We assume that

$$\mathbb{E}\left[\max_{0 \le t \le T} A_t\right] < \infty. \tag{2.1}$$

We show below that for the optimal execution problem of this paper one can assume without loss of generality that this martingale is identically zero. We make this assumption beginning in section 3 in order to simplify the presentation.

For some extended positive real number M, let μ be an infinite measure on $[0, M)$ that is finite on each compact subset of $[0, M)$. Denote the associated left-continuous cumulative distribution function by

$$F(x) \triangleq \mu\big([0, x)\big), \quad x \ge 0.$$

This is the *shadow limit-order book*, in the sense described below. We assume $F(x) > 0$ for every $x > 0$. If B is a measurable subset of $[0, M)$, then, in the absence of the large investor modeled in this paper, at time $t \ge 0$ the number of limit orders with prices in $B + A_t \triangleq \{b + A_t; b \in B\}$ is $\mu(B)$.

There is a strictly positive constant $\overline{X}$ such that our large investor must purchase $\overline{X}$ shares over the time interval $[0, T]$. His *purchasing strategy* is a nondecreasing right-continuous adapted process X with $X_T = \overline{X}$. We interpret X_t to be the cumulative amount of purchasing done by time t. We adopt the convention $X_{0-} = 0$, so that $X_0 = \Delta X_0$ is the number of shares purchased at time zero. Here and elsewhere, we use the notation ΔX_t to denote the jump $X_t - X_{t-}$ in X at time t.

The effect of the purchasing strategy X on the limit-order book is determined by a *resilience function* h, a strictly increasing, locally Lipschitz function defined on $[0, \infty)$ and satisfying

$$h(0) = 0, \quad h(\infty) \triangleq \lim_{x\to\infty} h(x) > \frac{\overline{X}}{T}. \tag{2.2}$$

The function h together with X determines the *volume effect process*[1] E satisfying

$$E_t = X_t - \int_0^t h(E_s)\,ds, \quad 0 \le t \le T. \tag{2.3}$$

It is shown in Appendix A that there is a unique nonnegative right-continuous finite-variation adapted process E satisfying (2.3). As with X, we adopt the convention $E_{0-} = 0$. We note that $\Delta X_t = \Delta E_t$ for $0 \le t \le T$.

Let B be a measurable subset of $[0, M)$. The interpretation of E is that, in the presence of the large investor using strategy X, at time $t \ge 0$ the number of limit orders with prices in $B + A(t)$ is $\mu_t(B)$, where μ_t is the σ-finite infinite measure on $[0, M)$ with left-continuous cumulative distribution function $(F(x) - E_t)^+$, $x \ge 0$. In other words, E_t units of mass have been removed from the shadow limit-order book μ. In any interval in which no purchases are made, (2.3) implies $\frac{d}{dt}E_t = -h(E_t)$. Hence, in the absence of purchases, the volume effect process decays toward zero and the limit-order book tends toward the shadow limit-order book μ, displaced by the ask price A.

To calculate the cost to the investor of using the strategy X, we introduce the following notation. We first define the left-continuous inverse of F,

$$\psi(y) \triangleq \sup\{x \ge 0 | F(x) < y\}, \quad y > 0.$$

We set $\psi(0) \triangleq \psi(0+) = 0$, where the second equality follows from the assumption that $F(x) > 0$ for every $x > 0$. The *ask price* in the presence of the large investor is defined to be $A_t + D_t$, where

$$D_t \triangleq \psi(E_t), \quad 0 \le t \le T. \tag{2.4}$$

This is the price after any lump purchases by the investor at time t (see Figure 1). We give some justification for calling $A_t + D_t$ the ask price after the following three examples.

Example 2.1 (block order book). Let q be a fixed positive number. If q is the density of shares available at each price, then for each $x \ge 0$ the quantity available at prices in $[0, x]$ is $F(x) = qx$. This is the *block order book* considered by [10]. In this case, $\psi(y) = y/q$ and $F(\psi(y)) = y$ for all $y \ge 0$.

---

[1] The case that resilience is based on price rather than volume is also considered in [2, 4].

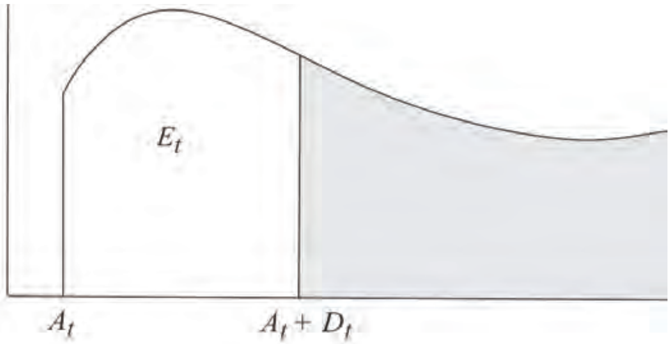


Figure 1. *Limit-order book at time t. The shaded region corresponds to the remaining shares. The white area E_t corresponds to the amount of shares missing from the order book at time t. The current ask price is $A_t + D_t$.*

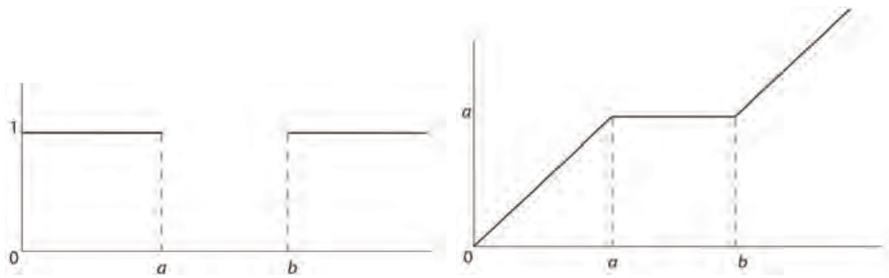


Figure 2. *Density and cumulative distribution of the modified block order book.*

Example 2.2 (modified block order book). Let $0 < a < b < \infty$ be given, and suppose

$$F(x) = \begin{cases} x, & 0 \le x \le a, \\ a, & a \le x \le b, \\ x - (b - a), & b \le x < \infty. \end{cases} \tag{2.5}$$

This is a block order book, except that the orders with prices between a and b are not present (see Figure 2). In this case,

$$\psi(y) = \begin{cases} y, & 0 \le y \le a, \\ y + b - a, & a < y < \infty. \end{cases} \tag{2.6}$$

We have $F(\psi(y)) = y$ for all $y \ge 0$.

Example 2.3 (discrete order book). Suppose that

$$F(x) = \sum_{i=0}^{\infty} \mathbb{I}_{(i,\infty)}(x), \quad x \ge 0, \tag{2.7}$$

which corresponds to an order of size 1 at each of the nonnegative integers (see Figure 3). Then

$$\psi(y) = \sum_{i=1}^{\infty} \mathbb{I}_{(i,\infty)}(y), \quad y \ge 0. \tag{2.8}$$

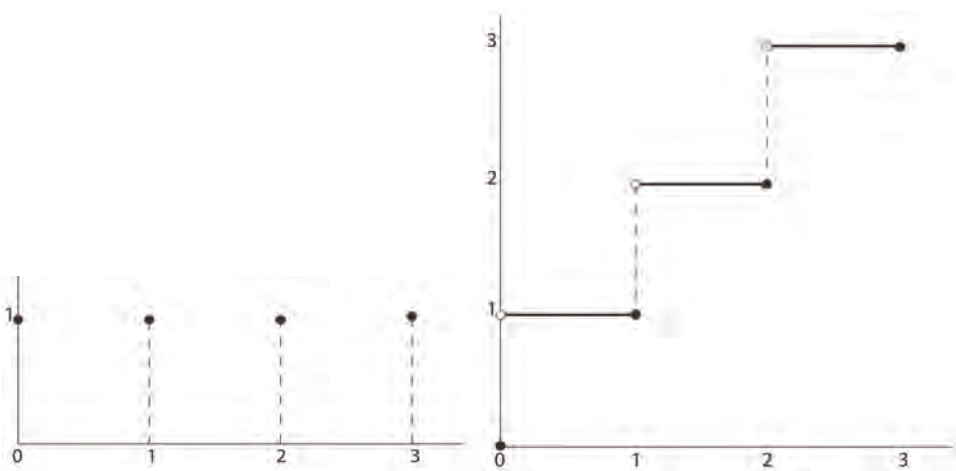


Figure 3. *Measure and cumulative distribution function of the discrete order book.*

For every nonnegative integer j, we have $F(j) = j$, $F(j+) = j+1$, $\psi(j+1) = j$, $\psi(j+) = j$, $F(\psi(j)+) = j$, and $\psi(F(j)+) = j$.

We return to the definition of the ask price as $A_t + D_t$ to provide some justification, leading up to Definition 2.4, for the total cost of a purchasing strategy. Suppose, as in Example 2.2, F is constant on an interval $[a, b]$ but strictly increasing to the left of a and to the right of b. Let $y = F(x)$ for $a \le x \le b$. Then $\psi(y) = a$ and $\psi(y+) = b$. Suppose, at time t, we have $E_t = y$. Then $D_t = a$, but the measure μ_t assigns mass zero to $[a, b)$. The ask price is $A_t + D_t$, but there are no shares for sale at this price, nor in an interval to the right of this price. Nonetheless, it is reasonable to call $A_t + D_t$ the ask price for an infinitesimal purchase because if the agent will wait an infinitesimal amount of time before making this purchase, shares will appear at the price $A_t + D_t$ due to resilience. We make this argument more precise.

Suppose the agent wishes to purchase a small number $\varepsilon > 0$ shares at time t at the ask price $A_t + D_t$. This purchase can be approximated by first purchasing zero shares in the time interval $[t, t+\delta]$, where δ is chosen so that $\int_t^{t+\delta} h(E_s)\,ds = \varepsilon$ and

$$E_s = X_t - \int_0^s h(E_u)\,du, \quad t \le s < t+\delta.$$

In other words, E_s for $t \le s < t+\delta$ is given by (2.3) with X held constant (no purchases) over this interval. With δ chosen this way, $E_{(t+\delta)-} = E_t - \varepsilon$. Resilience in the order book has created ε shares. Suppose the investor purchases these shares at time $t+\delta$, which means that $\Delta X_{t+\delta} = \Delta E_{t+\delta} = \varepsilon$ and $E_{t+\delta} = E_t$. Immediately before the purchase, the ask price is $A_{t+\delta} + \psi(E_t - \varepsilon)$; immediately after the purchase, the ask price is $A_{t+\delta} + \psi(E_t) = A_{t+\delta} + a$. The cost of purchasing these shares is

$$\varepsilon A_{t+\delta} + \int_{[\psi(E_t-\varepsilon),a]} \xi\, d\big(F(\xi) - E_t + \varepsilon\big)^+. \tag{2.9}$$

Because $\int_{[\psi(E_t-\varepsilon),a]} d\big(F(\xi) - E_t + \varepsilon\big)^+ = \varepsilon$, the integral in (2.9) is bounded below by $\varepsilon\psi(E_t - \varepsilon)$ and bounded above by εa. But $a = \psi(E_t) = D_t$ and ψ is left continuous, so the cost per share obtained by dividing (2.9) by ε converges to $A_t + a = A_t + D_t$ as ε (and hence δ) converges down to zero.

On the other hand, an impatient agent who does not wait before purchasing shares could choose a different method of approximating an infinitesimal purchase at time t that leads to a limiting cost per share $A_t + b$. In particular, it is not the case that our definition of ask price is consistent with all limits of discrete-time purchasing strategies. Our definition is designed to capture the limit of discrete-time purchasing strategies that seek to minimize cost.

To simplify calculations of the type just presented, we define the functions

$$\varphi(x) = \int_{[0,x)} \xi \, dF(\xi), \quad x \geq 0, \tag{2.10}$$

$$\Phi(y) = \varphi\big(\psi(y)\big) + \big[y - F\big(\psi(y)\big)\big]\psi(y), \quad y \geq 0. \tag{2.11}$$

We note that $\Phi(0) = 0$, and we extend Φ to be zero on the negative half-line. In the absence of the large investor, the cost one would pay to purchase all the shares available at prices in the interval $[A(t), A(t) + x)$ at time t would be $A(t) + \varphi(x)$. The function $\Phi(y)$ captures the cost, in excess of A_t, of purchasing y shares in the absence of the large investor. The first term on the right-hand side of (2.11) is the cost less A_t of purchasing all the shares with prices in the interval $[A_t, A_t + \psi(y))$. If F has a jump at $\psi(y)$, this might be fewer than y shares. The difference, $y - F(\psi(y))$ shares, can be purchased at price $A_t + \psi(y)$, and this explains the second term on the right-hand side of (2.11). We present these functions in the three examples considered earlier.

Example 2.1 (block order book, continued). We have simply $\varphi(x) = q\int_0^x \xi \, d\xi = \frac{q}{2}x^2$ for all $x \geq 0$, and $\Phi(y) = \frac{q}{2}\psi^2(y) = \frac{1}{2q}y^2$ for all $y \geq 0$. Note that Φ is convex and $\Phi'(y) = \psi(y)$ for all $y \geq 0$, including at $y = 0$ because we define Φ to be identically zero on the negative half-line.

Example 2.2 (modified block order book, continued). With F and ψ given by (2.5) and (2.6), we have

$$\varphi(x) = \begin{cases} \dfrac{1}{2}x^2, & 0 \leq x \leq a, \\[2ex] \dfrac{1}{2}a^2, & a \leq x \leq b, \\[2ex] \dfrac{1}{2}(x^2 + a^2 - b^2), & b \leq x < \infty, \end{cases}$$

and

$$\Phi(y) = \begin{cases} \dfrac{1}{2}y^2, & 0 \leq y \leq a, \\[2ex] \dfrac{1}{2}\big((y + b - a)^2 + a^2 - b^2\big), & a \leq y < \infty. \end{cases}$$

Note that Φ is convex with subdifferential

$$\partial\Phi(y) = \begin{cases} \{y\}, & 0 \leq y < a, \\ [a, b], & y = a, \\ \{y + b - a\}, & a < y < \infty. \end{cases} \tag{2.12}$$

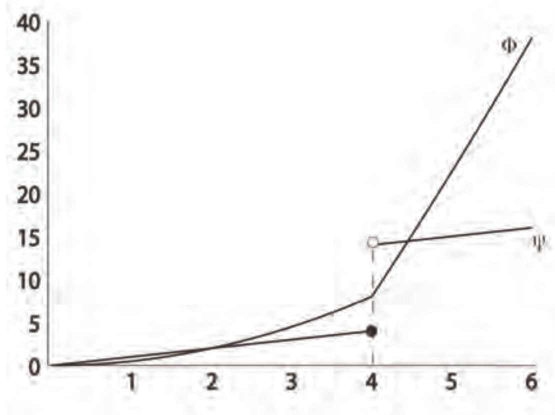


Figure 4. *Functions Φ and ψ for the modified block order book with parameters $a = 4$ and $b = 14$.*

In particular, $\partial\Phi(y) = [\psi(y), \psi(y+)]$ for all $y \geq 0$ (see Figure 4).

Example 2.3 (discrete order book, continued). With F given by (2.7), we have $\varphi(x) = \sum_{i=0}^{\infty} i\mathbb{I}_{(i,\infty)}(x)$. In particular, $\varphi(0) = 0$, and for integers $k \geq 1$ and $k - 1 < x \leq k$,

$$\varphi(x) = \sum_{i=0}^{k-1} i = \frac{k(k-1)}{2}.$$

For $0 \leq y \leq 1$, $\psi(y) = 0$ and hence $\varphi(\psi(y)) = 0$, $[y - F(\psi(y))]\psi(y) = 0$, and $\Phi(y) = 0$. For integers $k \geq 1$ and $k < y \leq k+1$, (2.8) gives $\psi(y) = k$, and hence $\varphi(\psi(y)) = \frac{k(k-1)}{2}$. Finally, for y in this range, $[y - F(\psi(y))]\psi(y) = k(y - k)$. We conclude that

$$\Phi(y) = \sum_{k=1}^{\infty} k\left(y - \frac{1}{2}k - \frac{1}{2}\right)\mathbb{I}_{(k,k+1]}(y). \tag{2.13}$$

For each positive integer k, $\Phi(k-) = \Phi(k+) = \frac{1}{2}k(k-1)$, so Φ is continuous. Furthermore, $\partial\Phi(k) = [k-1, k] = [\psi(k), \psi(k+)]$. For nonnegative integers k and $k < y < k+1$, $\Phi'(y)$ is defined and is equal to $\psi(y) = k$. Furthermore, $\Phi'(0) = \psi(0) = 0$. Once again we have $\partial\Phi(y) = [\psi(y), \psi(y+)]$ for all $y \geq 0$, and because ψ is nondecreasing, Φ is convex (see Figure 5).

We decompose the purchasing strategy X into its continuous and pure jump parts $X_t = X_t^c + \sum_{0 \leq s \leq t} \Delta X_s$. The investor pays price $A_t + D_t$ for infinitesimal purchases at time t, and hence the total cost of these purchases is $\int_0^T (A_t + D_t)\, dX_t^c$. On the other hand, if $\Delta X_t > 0$, the investor makes a lump purchase of size $\Delta X_t = \Delta E_t$ at time t. Because mass E_{t-} is missing in the shadow order book immediately prior to time t, the cost of this purchase is the difference between purchasing E_t and purchasing E_{t-} from the shadow order book, i.e., the difference in what the costs of these purchases would be in the absence of the large investor. Therefore, the cost of the purchase ΔX_t at time t is $A_t\Delta X_t + \Phi(E_t) - \Phi(E_{t-})$. These considerations lead to the following definition.

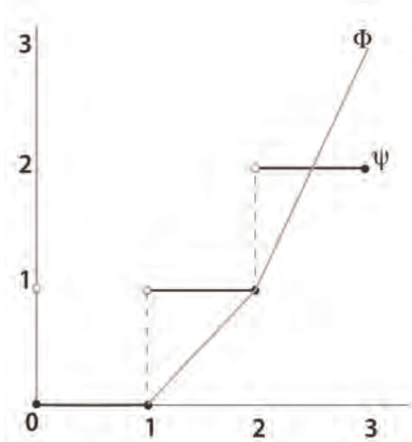


Figure 5. *Functions Φ and ψ for the discrete order book.*

Definition 2.4. *The* total cost *incurred by the investor using purchasing strategy X over the interval $[0,T]$ is*

$$C(X) \triangleq \int_0^T (A_t + D_t)\, dX_t^c + \sum_{0\le t\le T} \big[A_t \Delta X_t + \Phi(E_t) - \Phi(E_{t\_})\big]$$

$$(2.14) \qquad = \int_0^T D_t\, dX_t^c + \sum_{0\le t\le T} \big[\Phi(E_t) - \Phi(E_{t-})\big] + \int_{[0,T]} A_t\, dX_t.$$

Our goal is to determine the purchasing strategy X that minimizes $\mathbb{E}C(X)$.

3. Problem simplifications.

To compute the expectation of $C(X)$ defined by (2.14), we invoke the integration by parts formula

$$\int_{[0,T]} A_t\, dX_t = A_T X_T - A_0 X_{0-} - \int_0^T X_t\, dA_t$$

for the bounded variation process X and the continuous martingale A. Our investor's strategies must satisfy $0 = X_{0-} \le X_t \le X_T = \overline{X}$, $0 \le t \le T$, and hence $\mathbb{E}\int_0^T X_t\, dA_t = 0$ (see Appendix B) and $\mathbb{E}\int_0^T A_t\, dX_t = \overline{X}\mathbb{E}A_T = \overline{X}A_0$. It follows that

$$(3.1) \qquad \mathbb{E}C(X) = \mathbb{E}\int_0^T D_t\, dX_t^c + \mathbb{E}\sum_{0\le t\le T} \big[\Phi(E_t) - \Phi(E_{t-})\big] + \overline{X}A_0.$$

Since the third term on the right-hand side of (3.1) does not depend on X, minimization of $\mathbb{E}C(X)$ is equivalent to minimization of the first two terms. But the first two terms do not depend on A, and hence we may assume without loss of generality that A is identically zero. Under this assumption, the cost of using strategy X is

$$(3.2) \qquad C(X) = \int_0^T D_t\, dX_t^c + \sum_{0\le t\le T} \big[\Phi(E_t) - \Phi(E_{t-})\big].$$

But, with $A \equiv 0$, there is no longer a source of randomness in the problem. Consequently, without loss of generality we may restrict the search for an optimal strategy to nonrandom functions of time. Once we find a nonrandom purchasing strategy minimizing (3.2), then even if A is a continuous nonzero nonnegative martingale, we have found a purchasing strategy that minimizes the expected value of (2.14) over all (possibly random) purchasing strategies.

Remark 3.1. We do not allow our agent to make intermediate sells in order to achieve the ultimate goal of purchasing $\overline{X}$ shares because doing so would not decrease the cost, at least when the total amount of buying and selling is bounded. Indeed, in addition to the purchasing strategy X, suppose the agent has a selling strategy Y, which we take to be a nondecreasing right-continuous adapted process with $Y_{0-} = 0$. We assume that both X and Y are bounded. For each t, X_t represents the number of shares bought by time t and Y_t is the number of shares sold. These processes must be chosen so that $X_T - Y_T = \overline{X}$. We have not modeled the limit-buy-order book, but if we did so in a way analogous to the model of the limit-sell-order book, then the bid price at each time t would be less than or equal to A_t. Therefore, the net cost of executing the strategy (X, Y) would satisfy

$$C(X,Y) \geq \int_0^T D_t\, dX_t^c + \sum_{0 \leq t \leq T} \big[\Phi(E_t) - \Phi(E_{t-})\big] + \int_{[0,T]} A_t\, dX_t - \int_{[0,T]} A_t\, dY_t.$$

The integration by parts formula implies

$$\begin{aligned}\int_{[0,T]} A_t\, dX_t - \int_{[0,T]} A_t\, dY_t &= A_T(X_T - Y_T) - A_0(X_{0-} - Y_{0-}) - \int_0^T (X_t - Y_t)\, dA_t \\ &= A_T\overline{X} - \int_0^T (X_t - Y_t)\, dA_t.\end{aligned}$$

Because we can apply Lemma B.1 to both X and Y, the expectation of $\int_0^T (X_t - Y_t)\, dA_t$ is zero and

$$\mathbb{E}C(X,Y) \geq \mathbb{E}\int_0^T D_t\, dX_t^c + \mathbb{E}\sum_{0 \leq t \leq T} \big[\Phi(E_t) - \Phi(E_{t-})\big] + \overline{X}A_0. \tag{3.3}$$

The right-hand side of (3.3) is the formula (3.1) obtained for the cost of using the purchasing strategy X alone, but the X in inequality (3.3) makes a total purchase of $X_T = \overline{X} + Y_T \geq \overline{X}$. If we replace X by $\min\{X, \overline{X}\}$, we obtain a feasible purchasing strategy whose total cost is less than or equal to the right-hand side of (3.3).

Theorem 3.2. *Under the assumption (made without loss of generality) that A is identically zero, the cost* (3.2) *associated with a nonrandom nondecreasing right-continuous function X_t, $0 \leq t \leq T$, satisfying $X_{0-} = 0$ and $X_T = \overline{X}$ is equal to*

$$C(X) = \Phi(E_T) + \int_0^T D_t h(E_t)\, dt. \tag{3.4}$$

Proof. The proof proceeds in two steps. In Step 1 we show that, as we have seen in the examples, Φ is a convex function with subdifferential

$$\partial\Phi(y) = [\psi(y), \psi(y+)], \quad y \geq 0. \tag{3.5}$$

In Step 2 we justify the integration formula

$$\Phi(E_T) = \int_0^T D^-\Phi(E_t)\,dE_t^c + \sum_{0\le t\le T}\big[\Phi(E_t) - \Phi(E_{t-})\big], \tag{3.6}$$

where $D^-\Phi(E_t)$ denotes the left-hand derivative $\psi(E_t) = D_t$ of Φ at E_t, and E^c is the continuous part of E: $E_t^c = E_t - \sum_{0\le s\le t}\Delta E_s$. From (2.3) and (3.6) we have immediately that

$$\Phi(E_T) = \int_{[0,T]} D_t\,dX_t^c - \int_0^T D_t h(E_t)\,dt + \sum_{0\le t\le T}\big[\Phi(E_t) - \Phi(E_{t-})\big],$$

and (3.4) follows from (3.2).

Step 1. Using the integration by parts formula $xF(x) = \int_{[0,x)}\xi\,dF(\xi) + \int_0^x F(\xi)\,d\xi$, we write

$$\begin{aligned}
\Phi(y) &= \int_{[0,\psi(y))}\xi\,dF(\xi) + [y - F(\psi(y))]\psi(y)\\
&= \int_0^{\psi(y)}\big(y - F(\xi)\big)\,d\xi\\
&= \int_0^{\psi(y)}\int_{F(\xi)}^{y} d\eta\,d\xi\\
&= \int_0^{y}\int_0^{\psi(\eta)} d\xi d\eta,
\end{aligned}$$

where the last step follows from the fact that the symmetric difference of the sets $\{(\eta,\xi)|\xi \in [0,\psi(y)], \eta\in[F(\xi),y]\}$ and $\{(\eta,\xi)|\eta\in[0,y],\xi\in[0,\psi(\eta)]\}$ is an at most countable union of line segments and thus has two-dimensional Lebesgue measure 0. Therefore,

$$\Phi(y) = \int_0^y \psi(\eta)\,d\eta, \tag{3.7}$$

and by Problem 3.6.20 on p. 213 of [17], with ψ and Φ extended to be 0 for the negative reals, we conclude that Φ is convex and that $\partial\Phi(y) = [\psi(y),\psi(y+)]$, as desired.

Step 2. We mollify ψ, taking ρ to be a nonnegative C^∞ function with support on $[-1,0]$ and integral 1, defining $\rho_n(\eta) = n\rho(n\eta)$, and defining

$$\psi_n(y) = \int_{\mathbb{R}}\psi(y+\eta)\rho_n(\eta)\,d\eta = \int_{\mathbb{R}}\psi(\zeta)\,\rho_n(\zeta - y)\,d\zeta.$$

Then each ψ_n is a C^∞ function satisfying $0\le\psi_n(y)\le\psi(y)$ for all $y\ge 0$. Furthermore, $\psi(y) = \lim_{n\to\infty}\psi_n(y)$ for every $y\in\mathbb{R}$. We set $\Phi_n(y) = \int_0^y\psi_n(\eta)\,d\eta$, so that each Φ_n is also a C^∞ function and $\lim_{n\to\infty}\Phi_n'(y) = D^-\Phi(y)$.

Because $\Phi_n(E_{0-}) = \Phi(0) = 0$, we have

$$\Phi_n(E_T) = \int_0^T \Phi_n'(E_t)\,dE_t^c + \sum_{0\le t\le T}[\Phi_n(E_t) - \Phi_n(E_{t-})]; \tag{3.8}$$

see, e.g., [11, p. 78]. The function E_t, $0 \le t \le T$, is bounded. Letting $n \to \infty$ in (3.8) and using the bounded convergence theorem, we obtain

$$\Phi(E_T) = \int_0^T D^-\Phi(E_t)\, dE_t^c + \lim_{n\to\infty} \sum_{0\le t\le T} \big[\Phi_n(E_t) - \Phi_n(E_{t-})\big]. \tag{3.9}$$

To conclude the proof of (3.6), we divide the sum in (3.9) into two parts. Given $\delta > 0$, we define $S_\delta = \{t \in [0,T] : 0 < \Delta E_t \le \delta\}$ and $S'_\delta = \{t \in [0,T] : \Delta E_t > \delta\}$. The sum in (3.9) is over $t \in S_\delta \cup S'_\delta$, and because E has finite variation, $\sum_{t\in S_\delta\cup S'_\delta} \Delta E_t < \infty$. Let $\varepsilon > 0$ be given. We choose $\delta > 0$ so small that $\sum_{t\in S_\delta} \Delta E_t \le \varepsilon$. Because ψ (and hence each ψ_n) is bounded on $[0, E_T]$, the function Φ and each Φ_n is Lipschitz continuous on $[0, E_T]$ with the same Lipschitz constant $L = \psi(E_T)$. It follows that

$$\sum_{t\in S_\delta} \big[\Phi(E_t) - \Phi(E_{t-})\big] \le L \sum_{t\in S_\delta} \Delta E_t \;\le\; L\varepsilon,$$
$$\sum_{t\in S_\delta} \big[\Phi_n(E_t) - \Phi_n(E_{t-})\big] \le L \sum_{t\in S_\delta} \Delta E_t \;\le\; L\varepsilon, \quad n = 1, 2, \ldots.$$

Hence the difference between $\sum_{t\in S_\delta}\big[\Phi(E_t) - \Phi(E_{t-})\big]$ and any limit point as $n \to \infty$ of $\sum_{t\in S_\delta}\big[\Phi_n(E_t) - \Phi_n(E_{t-})\big]$ is at most $2L\varepsilon$. On the other hand, the set S'_δ contains only finitely many elements, and thus

$$\lim_{n\to\infty} \sum_{t\in S'_\delta} \big[\Phi_n(E_t) - \Phi_n(E_{t-})\big] = \sum_{t\in S'_\delta} \big[\Phi(E_t) - \Phi(E_{t-})\big].$$

Since $\varepsilon > 0$ is arbitrary, (3.9) reduces to (3.6). ■

4. Solution of the optimization problem. In view of Theorem 3.2, we want to minimize $\Phi(E_T) + \int_0^T D_t h(E_t)\, dt$ over the set of deterministic purchasing strategies. The main result of this paper is that there exists an optimal strategy X under which the trader buys a lump quantity $X_0 = E_0$ of shares at time 0, then buys at a constant rate $dX_t = h(E_0)\, dt$ up to time t_0 (so as to keep $E_t = E_0$ for $t \in [0, t_0)$), then buys another lump quantity of shares at time t_0, subsequently trades again at a constant rate $dX_t = h(E_{t_0})\, dt$ until time T (so as to keep $E_t = E_{t_0}$ for $t \in [t_0, T)$), and finally buys the remaining shares at time T. We shall call this strategy a *Type B strategy*. We further show that if the nonnegative function

$$g(y) \triangleq y\psi\big(h^{-1}(y)\big) \tag{4.1}$$

is convex, then the purchase at time t_0 consists of 0 shares (so X has jumps only at times 0 and T). We call such a strategy a *Type A strategy*. Clearly the latter is a special case of the former.

Although g is naturally defined on $[0, h(\infty))$ by (4.1), we will want it to be defined on a compact set. Therefore we set

$$\overline{Y} = \max\left\{h(\overline{X}), \frac{\overline{X}}{T}\right\} \tag{4.2}$$

and note that, because of assumption (2.2), h^{-1} is defined on $[0,\overline{Y}]$. We specify the domain of the function g to be $[0,\overline{Y}]$. For future reference, we make three observations about the function g. First,

$$\lim_{y\downarrow 0} g(y) = g(0) = 0. \tag{4.3}$$

Second, using the definition (2.4) of D_t, we can rewrite the cost function formula (3.4) as

$$C(X) = \Phi(E_T) + \int_0^T g\big(h(E_t)\big)\,dt. \tag{4.4}$$

Lemma A.1(iv) in Appendix A shows that $0 \le E_t \le \overline{X}$, so the domain $[0,\overline{Y}]$ of g is large enough in order for (4.4) to make sense. Because h^{-1} is strictly increasing and continuous and ψ is nondecreasing and left continuous, the function g is nondecreasing and left continuous and hence lower semicontinuous. In particular,

$$g\big(\overline{Y}\big) = \lim_{y\uparrow\overline{Y}} g(y). \tag{4.5}$$

4.1. Convexity and Type A strategies.

Remark 4.1. A Type A strategy X^A can be characterized in terms of the terminal value E_T^A of the process E^A related to X^A by (2.3), and the cost of using a Type A strategy can be written as a function of E_T^A. It is this function of E_T^A we will minimize. To see that this is possible, let X^A be a Type A strategy and let E^A be related to X^A via (2.3), so that $E_t^A = X_0^A$ for $0 \le t < T$. Then

$$X_{T-}^A = E_{T-}^A + \int_0^T h(E_t^A)\,dt \;=\; X_0^A + h(X_0^A)T, \tag{4.6}$$

$$\Delta X_T^A = \overline{X} - X_{T-}^A \;=\; \overline{X} - X_0^A - h(X_0^A)T, \tag{4.7}$$

$$E_T^A = E_{T-}^A + \Delta X_T^A \;=\; \overline{X} - h(X_0^A)T. \tag{4.8}$$

A Type A strategy is fully determined by its initial condition X_0^A, and from (4.8) we now see that choosing X_0^A is equivalent to choosing E_T^A. According to (4.4) and (4.8), the cost of this strategy

$$C(X^A) = \Phi(E_T^A) + Tg\big(h(X_0^A)\big) = \Phi(E_T^A) + Tg\left(\frac{\overline{X} - E_T^A}{T}\right) \tag{4.9}$$

can be written as a function of E_T^A.

We conclude this remark by determining the range of values that E_T^A can take for a Type A strategy. We must choose X_0^A so that $X_0^A \ge 0$ and X_{T-}^A given by (4.6) does not exceed $\overline{X}$. The function $k(x) \triangleq x + h(x)T$ is strictly increasing and continuous on $[0,\infty)$, and $k(\overline{X}) > \overline{X}$. Therefore, there exists a unique $\overline{e} \in (0,\overline{X})$ such that $k(\overline{e}) = \overline{X}$, i.e.,

$$\overline{e} + h(\overline{e})T = \overline{X}. \tag{4.10}$$

The constraint on the initial condition of Type A strategies that guarantees that the strategy is feasible is $0 \le X_0^A \le \overline{e}$. From (4.8) and (4.10) we see that the corresponding feasibility condition on E_T^A for Type A strategies is

$$\overline{e} \le E_T^A \le \overline{X}. \tag{4.11}$$

Theorem 4.2. *If g given by (4.1) is convex on $[0,\overline{Y}]$, then there exists a Type A purchasing strategy that minimizes $C(X)$ over all purchasing strategies X. If g is strictly convex, this is the unique optimal strategy.*

Proof. Assume that g is convex, and let X be a purchasing strategy. Jensen's inequality applied to (4.4) yields the lower bound

$$C(X) = \Phi(E_T) + T\int_0^T g\big(h(E_t)\big)\frac{dt}{T} \ge \Phi(E_T) + Tg\left(\int_0^T h(E_t)\frac{dt}{T}\right).$$

From (2.3) we further have $\int_0^T h(E_t)\,dt = \overline{X} - E_T$, and thus the lower bound can be rewritten as

$$C(X) \ge \Phi(E_T) + Tg\left(\frac{\overline{X} - E_T}{T}\right). \tag{4.12}$$

Recall that $0 \le E_T \le \overline{X}$, so the argument of g in (4.12) is in $[0,\overline{Y}]$.

This leads us to consider minimization of the function

$$G(e) \triangleq \Phi(e) + Tg\left(\frac{\overline{X} - e}{T}\right)$$

over $e \in [0,\overline{X}]$. By assumption, the function g is convex on $[0,\overline{Y}]$ and hence continuous on $(0,\overline{Y})$. Equations (4.3) and (4.5) show that g is also continuous at the endpoints of its domain. Because Φ has the integral representation (3.7), it also is convex and continuous on $[0,\overline{X}]$. Therefore, G is a convex continuous function on $[0,\overline{X}]$, and hence the minimum is attained.

We show next that the minimum of G over $[0,\overline{X}]$ is attained in $[\overline{e},\overline{X}]$. For this, we first observe that, because g is convex,

$$D^+g(y) \ge \frac{g(y) - g(0)}{y} = \psi\big(h^{-1}(y)\big), \quad 0 < y \le \overline{Y}.$$

This inequality together with (3.5) and (4.10) implies that

$$D^-G(\overline{e}) = \psi(\overline{e}) - D^+g(y)\Big|_{y=\frac{\overline{X}-\overline{e}}{T}} \le \psi(\overline{e}) - \psi\left(h^{-1}\left(\frac{\overline{X}-\overline{e}}{T}\right)\right) = 0. \tag{4.13}$$

Therefore, the minimum of the convex function G over $[0,\overline{X}]$ is obtained in $[\overline{e},\overline{X}]$.

Let $e^* \in [\overline{e},\overline{X}]$ attain the minimum of G over $[0,\overline{X}]$. The Type A strategy X^A with initial condition $X_0^A = h^{-1}(\frac{\overline{X}-e^*}{T})$ satisfies $E_T^A = e^*$ (see (4.8)), and hence the strategy is feasible (see (4.11)). The cost associated with this strategy is less than or equal to the right-hand side of (4.12) (see (4.9)). This strategy is therefore optimal.

If g is strictly convex at the point $\frac{\overline{X}-e^*}{T}$, where e^* minimizes G, then G is strictly convex at e^*, and this point is thus the unique minimizer of G. Therefore, every optimal strategy strategy must satisfy $E_T = e^*$. By strict convexity of g, a strategy that does not keep $h(E)$ equal to $\frac{\overline{X}-e^*}{T}$ almost everywhere in $(0, T)$ would result in strict inequality in (4.12). Since h is strictly increasing, we conclude that the only optimal strategy is the Type A strategy constructed above. ■

If g is not strictly convex at the point $\frac{\overline{X}-e^*}{T}$ found in the proof of Theorem 4.2, then G might still be strictly convex at e^*, in which case there would be only one optimal strategy of Type A, but there could be optimal strategies that are not of Type A. We demonstrate this phenomenon with an example.

Example 4.3 (nonuniqueness of optimal purchasing strategy). Suppose

$$F(x) = \begin{cases} x, & 0 \le x \le 2, \\ \frac{4}{4-x}, & 2 \le x \le 3, \\ 4 + \frac{1}{8}(x-3), & x \ge 3. \end{cases}$$

This function is continuous and strictly increasing, and hence

$$\psi(y) = \begin{cases} y, & 0 \le y \le 2, \\ 4 - \frac{4}{y}, & 2 \le y \le 4, \\ 8y - 29, & y \ge 4, \end{cases}$$

is also continuous and strictly increasing. This implies that

$$\Phi(y) = \int_0^y \psi(\eta) d\eta = \begin{cases} \frac{1}{2}y^2, & 0 \le y \le 2, \\ 4y - 6 - 4\log\frac{y}{2}, & 2 \le y \le 4, \\ 4y^2 - 29y + 62 - 4\log 2, & y \ge 4. \end{cases}$$

We take $h(x) = x$, so that

$$g(y) = y\psi(y) = \begin{cases} y^2, & 0 \le y \le 2, \\ 4y - 4, & 2 \le y \le 4, \\ 8y^2 - 29y, & y \ge 4, \end{cases}$$

and

$$g'(y) = \begin{cases} 2y, & 0 \le y \le 2, \\ 4, & 2 \le y < 4, \\ 16y - 29, & y > 4. \end{cases}$$

Note that g' is nondecreasing, so g is convex, but g is not strictly convex on the interval $[2, 4]$. Finally, we take $\overline{X} = 10\frac{1}{8}$ and $T = 2$.

In the notation of the proof of Theorem 4.2, we have $e^* = 4\frac{1}{8}$ and hence $\frac{\overline{X}-e^*}{T} = 3$. Indeed, $G'\left(4\frac{1}{8}\right) = \psi\left(4\frac{1}{8}\right) - g'(3) = 0$, and because ψ is strictly increasing, G is strictly convex, and hence $4\frac{1}{8}$ is the unique minimizer of G.

The Type A strategy with $E_T^A = 4\frac{1}{8}$ begins with an initial purchase of $X_0^A = 3$ and then consumes at rate 3 over the interval $[0, 2]$, so that $E_t^A = 3$ for $0 \le t < T$. At the final time

$T = 2$, there is an additional lump purchase of $1\frac{1}{8}$, so that $E_T^A = 4\frac{1}{8}$. The total cost of this strategy is

$$\Phi(E_T^A) + \int_0^T g(E_t^A)\,dt = \Phi\left(4\frac{1}{8}\right) + \int_0^2 \left(4E_t^A - 4\right)dt = \Phi\left(4\frac{1}{8}\right) + 16.$$

In particular, $\int_0^2 E_t^A\,dt = 6$.

In fact, any policy that satisfies $2 \le E_t \le 4$, $0 \le t < 2$, and $\int_0^2 E_t\,dt = 6$ will result in the same cost. Indeed, for such a policy, we will have

$$E_T = X_T - \int_0^T E_t\,dt = 10\frac{1}{8} - 6 = 4\frac{1}{8} = E_T^A$$

and

$$\int_0^T g(E_t)\,dt = \int_0^T (4E_t - 4)\,dt = 16 = \int_0^T g(E_t^A)\,dt,$$

so $\Phi(E_T) + \int_0^T g(E_t)\,dt = \Phi(E_T^A) + \int_0^T g(E_t^A)\,dt$. There are infinitely many policies like this. One such policy is to make an initial lump purchase of size 2 and then purchase at rate 2 up to time $\frac{1}{2}$ so that $E_t = 2$, $0 \le t < \frac{1}{2}$, make a lump purchase of size 1 at time $\frac{1}{2}$ and then purchase at rate 3 up to time $\frac{3}{2}$ so that $E_t = 3$, $\frac{1}{2} \le t < \frac{3}{2}$, make a lump purchase of size 1 at time $\frac{3}{2}$ and then purchase at rate 4 up to time 2 so that $E_t = 4$, $\frac{3}{2} \le t < 2$, and conclude with a lump purchase of size $\frac{1}{8}$ at time 2 so that $E_2 = 4\frac{1}{8}$.

Remark 4.4. Alfonsi, Fruth, and Schied [4] consider the case that the measure μ has a strictly positive density f. In this case, the function $F(x) = \int_0^x f(\xi)\,d\xi$ is strictly increasing and continuous with derivative $F'(x) = f(x)$, and its inverse ψ is likewise strictly increasing and continuous with derivative $\psi'(y) = 1/f(\psi(y))$. Furthermore, in [4], the resilience function is $h(x) = \rho x$, where ρ is a positive constant. In this case,

$$g'(y) = \psi(y/\rho) + \frac{y/\rho}{f\big(\psi(y/\rho)\big)},$$

and Theorem 4.2 guarantees the existence of a Type A strategy under the assumption that g' is nondecreasing. This is equivalent to the condition that

$$\psi(y) + \frac{y}{f\big(\psi(y)\big)}$$

is nondecreasing.

Alfonsi, Fruth, and Schied [4] obtain a discrete-time version of a Type A strategy under the assumption that

$$h_1(y) \triangleq \psi(y) - e^{-\rho\tau}\psi(e^{-\rho\tau}y)$$

is strictly increasing, where τ is the time between trading dates. In order to study the limit of their model as $\tau \downarrow 0$, they observe that

$$\lim_{\tau\downarrow 0} \frac{h_1(y)}{1 - e^{-\rho\tau}} = \psi(y) + \frac{y}{f\big(\psi(y)\big)},$$

which is thus nondecreasing. Thus g given by (4.1) is convex in their model.

To find a simpler formulation of the hypothesis of Theorem 4.2 under the assumption that μ has a strictly positive density f and $h(x) = \rho x$ for a positive constant ρ, we compute

$$\frac{d}{dy}\left(\psi(y) + \frac{y}{f\big(\psi(y)\big)}\right) = \frac{2}{f\big(\psi(y)\big)} - \frac{yf'\big(\psi(y)\big)}{f^3\big(\psi(y)\big)}.$$

This is nonnegative if and only if $2f^2\big(\psi(y)\big) \geq yf'\big(\psi(y)\big)$. Replacing y by $F(x)$, we obtain the condition

$$2f^2(x) \geq F(x)f'(x), \quad x \geq 0.$$

This is clearly satisfied under the assumption of [10] that f is a positive constant.

Example 2.1 (block order book, continued). In the case of the block order book with $h(x) = \rho x$, where ρ is a strictly positive constant,

$$g(y) = \frac{yh^{-1}(y)}{q} = \frac{y^2}{\rho q},$$

which is strictly convex. Theorem 4.2 implies that there is an optimal strategy of Type A, and this is the unique optimal strategy. From the formula $\Phi(e) = \frac{1}{2q}e^2$, we have

$$G(e) = \frac{e^2}{2q} + \frac{(\overline{X} - e)^2}{\rho q T}.$$

The minimizer is $e^* = \frac{2\overline{X}}{2+\rho T}$, which lies between $\overline{e} = \frac{\overline{X}}{1+\rho T}$ and $\overline{X}$, as expected. According to Remark 4.1, the optimal strategy of Type A is to make an initial purchase of size

$$X_0^A = h^{-1}\left(\frac{\overline{X} - e^*}{T}\right) = \frac{\overline{X}}{2 + \rho T},$$

then purchase continuously at rate $dX_t^A = h(X_0^A)\,dt = \frac{\rho \overline{X}}{2+\rho T}\,dt$ over the time interval $[0,T]$, and conclude with a lump purchase

$$e^* - X_0^A = \frac{\overline{X}}{2 + \rho T}$$

at the final time T. In particular, the initial and final lump purchases are of the same size, and there is no intermediate lump purchase.

4.2. Type B strategies.

Theorem 4.5. *In the absence of the assumption that g given by* (4.1) *is convex, there exists a Type B purchasing strategy that minimizes $C(X)$ over all purchasing strategies X.*

The proof of Theorem 4.5 depends on the following lemma, whose proof is given in Appendix C.

Lemma 4.6. *The convex hull of g, defined by*

$$\widehat{g}(y) \triangleq \sup\big\{\ell(y) : \ell \text{ is an affine function and } \ell(\eta) \leq g(\eta)\ \forall \eta \in [0,\overline{Y}]\big\}, \tag{4.14}$$

is the largest convex function defined on $[0,\overline{Y}]$ that is dominated by g there. It is continuous and nondecreasing on $[0,\overline{Y}]$, $\widehat{g}(0)=g(0)=0$, and $\widehat{g}(\overline{Y})=g(\overline{Y})$. If $y^*\in(0,\overline{Y})$ satisfies $\widehat{g}(y^*)<g(y^*)$, then there exists a unique affine function ℓ lying below g on $[0,\overline{Y}]$ and agreeing with $\widehat{g}$ at y^*. In addition, there exist numbers α and β satisfying

$$0\le\alpha<y^*<\beta\le\overline{Y}, \tag{4.15}$$

$$\ell(\alpha)=\widehat{g}(\alpha)=g(\alpha),\quad \ell(\beta)=\widehat{g}(\beta)=g(\beta), \tag{4.16}$$

$$\ell(y)=\widehat{g}(y)<g(y),\quad \alpha<y<\beta. \tag{4.17}$$

Proof of Theorem 4.5. Using $\widehat{g}$ in place of g in (4.4), we define the *modified cost function*

$$\widehat{C}(X)\triangleq\Phi(E_T)+\int_0^T\widehat{g}\big(h(E_t)\big)\,dt.$$

For any purchasing strategy X, we obviously have $\widehat{C}(X)\le C(X)$. Analogously to (4.12), for any purchasing strategy X, the lower bound

$$\widehat{C}(X)\ge\Phi(E_T)+T\widehat{g}\left(\frac{\overline{X}-E_T}{T}\right)$$

holds. This leads us to consider minimization of the function

$$\widehat{G}(e)\triangleq\Phi(e)+T\widehat{g}\left(\frac{\overline{X}-e}{T}\right) \tag{4.18}$$

over $e\in[0,\overline{X}]$. As in the proof of Theorem 4.2, this function attains its minimum at some $e^*\in[0,\overline{X}]$.

For the remainder of the proof, we use the notation

$$y^*=\frac{\overline{X}-e^*}{T},\quad x^*=h^{-1}(y^*), \tag{4.19}$$

where it is assumed without loss of generality that e^* is the largest minimizer of $\widehat{G}$ in $[0,\overline{X}]$. There are two cases. In both cases, we construct a strategy that satisfies $E_T^B=e^*$ and

$$C(X^B)=\widehat{G}(e^*). \tag{4.20}$$

In the first case, the strategy is a Type A strategy, and it is Type B in the second case. In both cases, we exhibit the strategy explicitly.

Case I. $\widehat{g}(y^*)=g(y^*)$. It is tempting to claim that we are now in the situation of Theorem 4.2 with the convex function $\widehat{g}$ replacing g. However, the proof needed here that $e^*\ge\overline{e}$, where $\overline{e}$ is determined by (4.10), cannot follow the proof of Theorem 4.2. In the proof of Theorem 4.2, this inequality was a consequence of (4.13), which ultimately depended on the definition (4.1) of $g(\overline{e})$. But we have only $\widehat{g}(\overline{e})\le\overline{e}\psi(h^{-1}(\overline{e}))$; we do not have an equation analogous to (4.1) for $\widehat{g}$. We thus provide a different proof, which depends on e^* being the largest minimizer of $\widehat{G}$ in $[0,\overline{X}]$.

If $x^* = 0$, then $y^* = 0$, $e^* = \overline{X}$, and $\widehat{G}(e^*) = G(e^*)$. The Type A strategy that waits until the final time T and then purchases $\overline{X}$ is optimal. In particular, this strategy satisfies the initial condition $X_0^A = x^*$.

If $x^* > 0$, we must consider two subcases. It could be that $0 < x^* \leq F(0+)$. In this subcase, $\widehat{g}(y^*) = g(y^*) = y^*\psi(x^*) = 0$ because $\psi \equiv 0$ on $[0, F(0+)]$. But $\widehat{g}(0) = 0$ and $\widehat{g}$ is nondecreasing, so $\widehat{g} \equiv 0$ on $[0, y^*]$. Furthermore, x^* is positive, so $e^* < \overline{X}$. For $e \in (e^*, \overline{X})$, the number $\frac{\overline{X}-e}{T}$ is in $(0, y^*)$, and by (3.5), $D^+\widehat{G}(e) = D^+\Phi(e) = \psi(e+)$. On the other hand, e^* is the largest minimizer of $\widehat{G}$ in $[0, \overline{X}]$, which implies that $D^+\widehat{G}(e) > 0$. This shows that $\psi(e+) > 0$ for every $e \in (e^*, \overline{X})$, which implies that $\psi(e) > 0$ for every $e \in (e^*, \overline{X})$ and further implies that $e > F(0+)$ for every $e \in (e^*, \overline{X})$. We conclude that $e^* \geq F(0+)$. Applying h to this inequality and using the subcase assumption $x^* \leq F(0+)$, we obtain

$$h(e^*) \geq h\big(F(0+)\big) \geq h(x^*) = \frac{\overline{X} - e^*}{T}. \tag{4.21}$$

In other words, $e^* + h(e^*)T \geq \overline{X}$, and by the defining equation (4.10) of $\overline{e}$ we conclude that $e^* \geq \overline{e}$. The corresponding optimal strategy, which is Type A, satisfies $X_0^A = x^*$ and $E_T^A = e^*$. The proof of optimality of this strategy follows the proof of Theorem 4.2 with $\widehat{g}$ replacing g.

Finally, we consider the subcase $x^* > F(0+)$. Because $y^* = h(x^*)$ is positive, the left-hand derivative of $\widehat{g}$ at y^* is defined, and it satisfies

$$D^-\widehat{g}(y^*) \geq \frac{\widehat{g}(y^*) - \widehat{g}(0)}{y^*} = \frac{g(y^*)}{y^*} = \psi(x^*). \tag{4.22}$$

In fact, the inequality in (4.22) is strict. It it were not, the affine function

$$\ell(y) = \psi(x^*)\big(y - y^*\big) + \widehat{g}(y^*) = y\psi(x^*)$$

would describe a tangent line to the graph of $\widehat{g}$ at $(y^*, \widehat{g}(y^*))$ lying below $\widehat{g}(y)$, and hence below $g(y)$, for all $y \in [0, \overline{Y}]$. But the resulting inequality $y\psi(x^*) \leq g(y) = y\psi(h^{-1}(y))$ yields $\psi(x^*) \leq \psi(h^{-1}(y))$ for all $y \in (0, \overline{Y}]$, and letting $y \downarrow 0$ we would conclude that $\psi(x^*) = 0$. This violates the subcase assumption $x^* > F(0+)$. We conclude that $D^-\widehat{g}(y^*) > \psi(x^*)$. The strict inequality, the fact that e^* minimizes $\widehat{G}$, and (3.5) further imply that

$$0 \leq D^+\widehat{G}(e^*) = D^+\Phi(e^*) - D^-\widehat{g}(y^*) < \psi(e^*+) - \psi(x^*).$$

But $\psi(x^*) < \psi(e^*+)$ implies that $x^* \leq e^*$. Consequently, $h(e^*) \geq h(x^*) = \frac{\overline{X}-e^*}{T}$. This is the essential part of inequality (4.21), and we conclude as above, constructing an optimal Type A strategy with $X_0^A = x^*$ and $E_T^A = e^*$.

Case II. $\widehat{g}(y^*) < g(y^*)$. Recall from Lemma 4.6 that this case can occur only if $0 < y^* < \overline{Y}$. In particular, $x^* > 0$. We let ℓ be the affine function and α and β be numbers as described in Lemma 4.6, and we construct a Type B strategy. To do this, we define $t_0 \in (0, T)$ by

$$t_0 = \frac{(\beta - y^*)T}{\beta - \alpha}, \tag{4.23}$$

so that $\alpha t_0 + \beta(T - t_0) = y^* T$. Consider the Type B strategy that makes an initial purchase $X_0^B = h^{-1}(\alpha)$, then purchases at rate $dX_t^B = \alpha\, dt$ for $0 \le t < t_0$ (so $E_t^B = h^{-1}(\alpha)$ for $0 \le t < t_0$), then follows this with a purchase $\Delta X_{t_0}^B = h^{-1}(\beta) - h^{-1}(\alpha)$ at time t_0, thereafter purchases at rate $dX_t^B = \beta\, dt$ for $t_0 \le t < T$ (so $E_t^B = h^{-1}(\beta)$ for $t_0 \le t < T$), and makes a final purchase $\overline{X} - X_{T-}^B$ at time T. According to (2.3),

$$X_t^B = \begin{cases} h^{-1}(\alpha) + \alpha t, & 0 \le t < t_0, \\ h^{-1}(\beta) + \alpha t_0 + \beta(t - t_0), & t_0 \le t < T, \\ \overline{X}, & t = T. \end{cases}$$

In particular,

$$\Delta X_T^B = \overline{X} - h^{-1}(\beta) - \alpha t_0 - \beta(T - t_0) = \overline{X} - h^{-1}(\beta) - y^* T = e^* - h^{-1}(\beta). \tag{4.24}$$

We show at the end of this proof that

$$h^{-1}(\beta) \le e^*. \tag{4.25}$$

This will ensure that ΔX_T^B is nonnegative, and since X^B is obviously nondecreasing on $[0, T)$, this will establish that X^B is a feasible purchasing strategy.

Accepting (4.25) for the moment, we note that (4.24) implies that

$$E_T^B = E_{T-}^B + \Delta E_T^B = h^{-1}(\beta) + \Delta X_T^B = e^*. \tag{4.26}$$

Using (4.4), (4.26), (4.16), the affine property of ℓ, and (4.17) in that order, we compute

$$\begin{aligned} C(X^B) &= \Phi(E_T^B) + \int_0^T g\big(h(E_t^B)\big)\, dt \\ &= \Phi(e^*) + g(\alpha) t_0 + g(\beta)(T - t_0) \\ &= \Phi(e^*) + \ell(\alpha) t_0 + \ell(\beta)(T - t_0) \\ &= \Phi(e^*) + T\ell\left(\frac{\alpha t_0 + \beta(T - t_0)}{T}\right) \\ &= \Phi(e^*) + T\ell(y^*) \\ &= \Phi(e^*) + T\widehat{g}(y^*) \\ &= \widehat{G}(e^*). \end{aligned}$$

This is (4.20).

Finally, we turn to the proof of (4.25). Because e^* is the largest minimizer of the convex function $\widehat{G}$ in $[0, \overline{X}]$ and $e^* < \overline{X}$ (because $x^* > 0$), the right-hand derivative of $\widehat{G}$ at e^* must be nonnegative. Indeed, for all $e \in (e^*, \overline{X})$, this right-hand derivative must in fact be strictly positive. For e greater than but sufficiently close to e^*, $\frac{\overline{X} - e}{T}$ is in (α, y^*), where $\widehat{g}$ is linear with slope $\frac{g(\beta)-g(\alpha)}{\beta-\alpha}$. For such e,

$$\begin{aligned}
0 &< D^+\widehat{G}(e) \\
&= D^+\Phi(e) - D^-\widehat{g}(y)\Big|_{y=\frac{\overline{X}-e}{T}} \\
&= \psi(e+) - \frac{g(\beta)-g(\alpha)}{\beta-\alpha} \\
&= \psi(e+) - \frac{\beta\psi\big(h^{-1}(\beta)\big) - \alpha\psi\big(h^{-1}(\alpha)\big)}{\beta-\alpha} \\
&\leq \psi(e+) - \frac{\beta\psi\big(h^{-1}(\beta)\big) - \alpha\psi\big(h^{-1}(\beta)\big)}{\beta-\alpha} \\
&= \psi(e+) - \psi\big(h^{-1}(\beta)\big).
\end{aligned}$$

This inequality $\psi\big(h^{-1}(\beta)\big) < \psi(e+)$ for all e greater than but sufficiently close to e^* implies (4.25). ■

Remark 4.7 (uniqueness). In Case I of the proof of Theorem 4.5, when $\widehat{g}(y^*) = g(y^*)$, strict convexity of $\widehat{g}$ at y^* implies uniqueness of the optimal purchasing strategy. The proof is similar to the uniqueness proof in Theorem 4.2.

However, in Case II, $\widehat{g}$ is not strictly convex at y^*. In this case, if ψ is strictly increasing at e^* and if the affine function ℓ of Lemma 4.6 agrees with g only at α and β, then the optimal purchasing strategy is unique. Indeed, if ψ is strictly increasing at e^*, then Φ (and hence $\widehat{G}$) is strictly convex at e^*, which implies that e^* is the unique minimizer of $\widehat{G}$. In order to be optimal, a purchasing strategy must satisfy the two inequalities

$$\int_0^T g\big(h(E_t)\big)\,dt \geq \int_0^T \widehat{g}\big(h(E_t)\big)\,dt \geq T\widehat{g}\left(\int_0^T h(E_t)\frac{dt}{T}\right) \tag{4.27}$$

with equality, as we explain below, and must also satisfy $E_T = e^*$. When the inequalities (4.27) hold, we can use (2.3) to obtain a lower bound on the cost of an arbitrary purchasing strategy X by the relations

$$\begin{aligned}
C(X) &= \Phi(E_T) + \int_0^T g(h(E_t))dt \\
&\geq \Phi(E_T) + T\widehat{g}\left(\int_0^T h(E_t)\frac{dt}{T}\right) \\
&= \Phi(E_T) + T\widehat{g}\left(\frac{\overline{X}-E_T}{T}\right) \\
&= \widehat{G}(E_T).
\end{aligned}$$

The minimal cost is $\widehat{G}(e^*) = \Phi(e^*) + T\widehat{g}(\frac{\overline{X}-e^*}{T}) = \Phi(e^*) + T\widehat{g}(y^*)$, and hence optimality of a strategy requires that equality hold in both parts of (4.27). The second inequality in (4.27) is Jensen's inequality, and equality holds if and only if $h(E_t)$, $0 \leq t < T$, stays in the region in which $\widehat{g}$ is affine. But the average value of $h(E_t)$, $\frac{1}{T}\int_0^T h(E_t)dt$, is equal to y^*, and hence

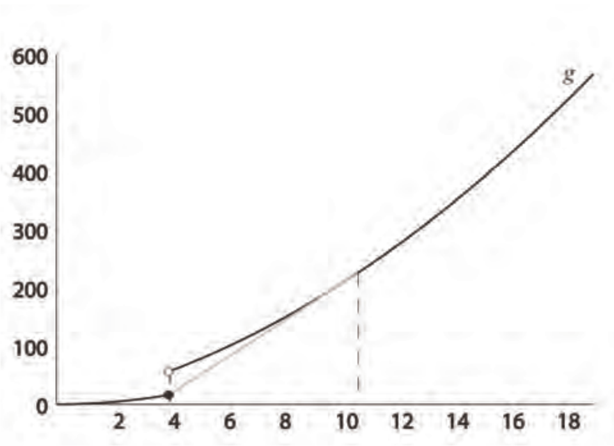


Figure 6. *Function g for the modified block order book with parameters $a = 4$ and $b = 14$. The convex hull $\widehat{g}$ is constructed by replacing a part $\{g(y)\ , y \in (a, \beta)\}$ by a straight line connecting $g(a)$ and $g(\beta)$. Here $\beta = 10.3246$.*

we cannot have $h(E_t) < y^*$ for all $t \in [0,T)$, nor can we have $h(E_t) > y^*$ for all $t \in [0,T)$. Hence the region in which $h(E_t)$ stays must be the region in which $\widehat{g}$ agrees with ℓ. To get an equality in the first inequality in (4.27), $h(E_t)$, $0 \le t < T$, must stay in the region where $\widehat{g}$ agrees with g. If ℓ agrees with g only at the two points α and β, then $h(E_t)$, $0 \le t < T$, must stay in the two-point set $\{\alpha, \beta\}$. Because $\Delta E_t = \Delta X_t \ge 0$ for all t, there must be some initial time interval $[0, t_0)$ on which $h(E_t) = \alpha$ and there must be some final time interval $[t_0, T)$ on which $h(E_t) = \beta$. In order to achieve this and to also have $\frac{1}{T}\int_0^T h(E_t) = y^*$, t_0 must be given by (4.23).

4.3. Examples of Type B optimal strategies.

Example 2.2 (modified block order book, continued). We continue Example 2.2 under the simplifying assumptions $T = 1$ and $h(x) = x$ for all $x \ge 0$, so $h^{-1}(y) = y$ for all $y \ge 0$ and $\overline{Y} = \overline{X}$. Recalling (2.6) and (4.1), we see that

$$g(y) = \begin{cases} y^2, & 0 \le y \le a, \\ y^2 + (b-a)y, & a < y < \infty. \end{cases}$$

The convex hull of g over $[0,\infty)$, given by (4.14), is

$$\widehat{g}(y) = \begin{cases} y^2, & 0 \le y \le a, \\ (2\beta + b - a)(y-a) + a^2, & a \le y \le \beta, \\ y^2 + (b-a)y, & \beta \le y < \infty, \end{cases}$$

where

$$\beta = a + \sqrt{a(b-a)} \tag{4.28}$$

(see Figure 6). We take $\overline{X} = \overline{Y} > \beta$, so that this is also the convex hull of g over $[0, \overline{Y}]$.

For $a < y^* < \beta$, we have $\widehat{g}(y^*) < g(y)$. For constants α and β from the statement of Lemma 4.6 (see (C.1)–(C.2) in Appendix C), we have that α of (C.1) is a, and β of (C.2) is given by (4.28). In order to illustrate a case in which a Type B purchasing strategy is optimal, we assume

$$a + 2\beta < \overline{X} < 3\beta. \tag{4.29}$$

The function $\widehat{G}$ of (4.18) is minimized over $[0, \overline{X}]$ at e^* if and only if

$$0 \in \partial\widehat{G}(e^*) = \partial\Phi(e^*) - \partial\widehat{g}(\overline{X} - e^*),$$

which is equivalent to $\partial\Phi(e^*) \cap \partial\widehat{g}(\overline{X} - e^*) \neq \emptyset$. We show below that the largest value of e^* satisfying this condition is $e^* = 2\beta$. According to (4.29), $e^* = 2\beta$ is in $(\overline{X} - \beta, \overline{X} - a)$. Because $\beta > a$, e^* is also in (a, ∞). We compute (recall (2.12))

$$\partial\Phi(e) = \begin{cases} \{e\}, & 0 \leq e < a, \\ [a, b], & e = a, \\ \{e + b - a\}, & a < e < \infty, \end{cases}$$

$$\partial\widehat{g}(\overline{X} - e) = \begin{cases} \big\{2(\overline{X} - e) + b - a\big\}, & 0 \leq e \leq \overline{X} - \beta, \\ \big\{2\beta + b - a\big\}, & \overline{X} - \beta \leq e < \overline{X} - a, \\ [2a, 2\beta + b - a], & e = \overline{X} - a, \\ \big\{2(\overline{X} - e)\big\}, & \overline{X} - a < e \leq \overline{X}, \end{cases}$$

and then evaluate

$$\partial\Phi(e^*) = \{e^* + b - a\} = \{2\beta + b - a\} = \partial\widehat{g}(\overline{X} - e^*).$$

Therefore, $\widehat{G}$ attains its minimum at e^*.

To see that there is no $e \in (2\beta, \overline{X}]$ where $\widehat{G}$ attains its minimum, we observe that for $e \in (2\beta, \overline{X} - a)$, $\partial\Phi(e) \cap \partial\widehat{g}(\overline{X} - e) = \{e + b - a\} \cap \{2\beta + b - a\} = \emptyset$. For $e \in [\overline{X} - a, \overline{X}]$, all points in $\partial\widehat{g}(\overline{X} - e)$ lie in the interval $[0, 2a]$, whereas the only point in $\partial\Phi(e)$, which is $e + b - a$, lies in the interval $[\overline{X} + b - 2a, \overline{X} + b - a]$. Because of (4.29), we have $2a < \overline{X} + b - 2a$, and hence $\partial\Phi(e) \cap \partial\widehat{g}(\overline{X} - e) = \emptyset$ for $e \in [\overline{X} - a, \overline{X}]$.

As in the proof of Theorem 4.5, we set $y^* = \overline{X} - e^* = \overline{X} - 2\beta$, $x^* = h^{-1}(y^*) = \overline{X} - 2\beta$. Condition (4.29) is equivalent to $a < y^* < \beta$, which in turn is equivalent to $\widehat{g}(y^*) < g(y^*)$. The first inequality in (4.29) shows that $x^* > 0$, and we are thus in Case II of the proof of Theorem 4.5. In this case, we define

$$t_0 = \frac{\beta - y^*}{\beta - a} = \frac{3\beta - \overline{X}}{\beta - a}.$$

The optimal purchasing strategy is

$$X_t^B = \begin{cases} a(t + 1), & 0 \leq t < t_0, \\ at_0 + \beta(t + 1 - t_0), & t_0 \leq t < 1, \\ \overline{X}, & t = 1. \end{cases}$$

In particular, $\Delta X_0 = a$, $\Delta X_{t_0} = \beta - a$, $\Delta X_1 = \beta$ (see (4.24) for the last equality). The corresponding E^B process is

$$E_t^B = \begin{cases} a, & 0 \le t < t_0, \\ \beta, & t_0 \le t < 1, \\ 2\beta, & t = 1. \end{cases}$$

The initial lump purchase moves the ask price to the left endpoint a of the gap in the order book. Purchasing is done to keep the ask price at a until time t_0, when another lump purchase moves the ask price to β beyond the right endpoint b of the gap in the order book. Purchasing is done to keep the ask price at β until the final time, when another lump purchase is executed.

Example 2.3 (discrete order book, continued). We continue Example 2.3 under the simplifying assumptions that $T = 1$ and $h(x) = x$ for all $x \ge 0$, so that $h^{-1}(y) = y$ for all $y \ge 0$ and $\overline{Y} = \overline{X}$. From (2.8) and (4.1) we see that $g(0) = 0$, and $g(y) = ky$ for integers $k \ge 0$ and $k < y \le k+1$. In particular, $g(k) = (k-1)k$ for nonnegative integers k. The convex hull of g interpolates linearly between the points $(k, (k-1)k)$ and $(k+1, k(k+1))$, i.e., $\widehat{g}(y) = k(2y - (k+1))$ for $k \le y \le k+1$, where k ranges over the nonnegative integers (see Figure 7).

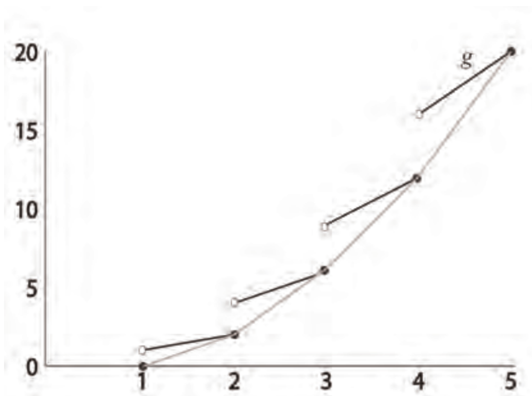


Figure 7. *Function g for the discrete order book. The convex hull $\widehat{g}$ interpolates linearly between the points $(k, (k-1)k)$ and $(k+1, k(k+1))$.*

Therefore,

$$\partial \widehat{g}(y) = \begin{cases} \{0\}, & y = 0, \\ [2(k-1), 2k], & y = k \text{ and } k \text{ is a positive integer}, \\ \{2k\}, & k < y < k+1 \text{ and } k \text{ is a nonnegative integer}. \end{cases}$$

Recall from the discussion following (2.13) that

$$\partial \Phi(y) = \begin{cases} \{0\}, & y = 0, \\ [k-1, k], & y = k \text{ and } k \text{ is a positive integer}, \\ \{k\}, & k < y < k+1 \text{ and } k \text{ is a nonnegative integer}. \end{cases}$$

We seek the largest number $e^* \in [0, \overline{X}]$ for which $\partial\Phi(e^*) \cap \partial\widehat{g}(\overline{X} - e^*) \neq \emptyset$. This is the largest minimizer of $\widehat{G}(e) = \Phi(e) + \widehat{g}(\overline{X} - e)$ in $[0, \overline{X}]$. We define k^* to be the largest integer less than or equal to $\frac{\overline{X}}{3}$, so that

$$3k^* \leq \overline{X} < 3k^* + 3.$$

We divide the analysis into three cases:

Case A. $3k^* \leq \overline{X} \leq 3k^* + 1$.

Case B. $3k^* + 1 < \overline{X} < 3k^* + 2$.

Case C. $3k^* + 2 \leq \overline{X} < 3k^* + 3$.

We show below that in Cases A and B the optimal strategy makes an initial lump purchase of size k^*, which executes the orders at prices $0, 1, \ldots, k^* - 1$. In Case A, the optimal strategy then purchases at rate k^* over the interval $(0, 1)$ and at time 1 makes a final lump purchase of size $\overline{X} - 2k^*$, which is in the interval $[k^*, k^* + 1]$. This is a Type A strategy. In Case B, there is an intermediate lump purchase of size 1 at time $3k^* + 2 - \overline{X}$. Before this intermediate purchase, the rate of purchase is k^*, and after this purchase the rate of purchase is $k^* + 1$. In Case B, at time 1, there is a final lump purchase of size k^*. In Case B, we have a Type B strategy. In Case C, the optimal strategy makes a lump purchase of size $k^* + 1$ at time 0, which executes the orders at prices $0, 1, \ldots, k^* - 1, k^*$. The optimal strategy then purchases continuously at rate $k^* + 1$ over the interval $(0, 1)$ and at time 1 makes a final lump purchase of size $\overline{X} - 2k^* - 2$, which is in the interval $[k^*, k^* + 1)$. This is a Type A strategy.

Case A. $3k^* \leq \overline{X} \leq 3k^* + 1$. We define $e^* = \overline{X} - k^*$, so that $2k^* \leq e^* \leq 2k^* + 1$ and $k^* = \overline{X} - e^*$. Then $2k^* \in \partial\Phi(e^*)$ and $\partial\widehat{g}(\overline{X} - e^*) = [2(k^* - 1), 2k^*]$, so the intersection of $\partial\Phi(e^*)$ and $\partial\widehat{g}(\overline{X} - e^*)$ is nonempty, as desired. On the other hand, if $e > e^*$, then $\partial\Phi(e) \subset [2k^*, \overline{X}]$ and $\partial\widehat{g}(\overline{X} - e) \subset [0, 2(k^* - 1)]$, so the intersection of these two sets is empty.

In this case, y^* and x^* defined by (4.19) are both equal to k^*, and hence $\widehat{g}(y^*) = g(y^*)$. If $k^* = 0$, we are in the first subcase of Case I of the proof of Theorem 4.5. The optimal purchasing strategy is to do nothing until time 1 and then make a lump purchase of size $\overline{X}$. If $k^* = 1$, which is equal to $F(0+)$, we are in the second subcase of Case I of the proof of Theorem 4.5. We should make an initial purchase of size $x^* = 1$, purchase continuously over the time interval $(0, 1)$ at rate 1 so that $E_t \equiv 1$ and $D_t \equiv 0$, and make a final purchase of size $\overline{X} - 2$. If $k^* \geq 2$, we are in the third subcase of Case I of the proof of Theorem 4.5. We should make an initial purchase of size k^*, purchase continuously over the time interval $(0, 1)$ at rate k^* so that $E_t \equiv k^*$ and $D_t \equiv k^* - 1$, and make a final purchase of size $\overline{X} - 2k^*$.

Case B. $3k^* + 1 < \overline{X} < 3k^* + 2$. We define $e^* = 2k^* + 1$, so that $k^* < \overline{X} - e^* < k^* + 1$. Then $\partial\Phi(e^*) = [2k^*, 2k^* + 1]$ and $\partial\widehat{g}(\overline{X} - e^*) = \{2k^*\}$, so the intersection of $\partial\Phi(e^*)$ and $\partial\widehat{g}(\overline{X} - e^*)$ is nonempty, as desired. On the other hand, if $e > e^*$, then $\partial\Phi(e) \subset [2k^* + 1, \overline{X}]$ and $\partial\widehat{g}(\overline{X} - e) \subset [0, 2k^*]$, so the intersection of these two sets is empty.

In this case, y^* and x^* defined by (4.19) are both equal to $\overline{X} - e^*$. Hence $k^* < y^* < k^* + 1$, $\widehat{g}(y^*) < g(y^*)$, and we are in Case II of the proof of Theorem 4.5 with $\alpha = k^*$ and $\beta = k^* + 1$ (see (4.14)–(4.17) and (C.1)–(C.2)). The optimal purchasing strategy is Type B. In particular, with $t_0 = \beta - y^* = k^* + 1 - x^* = 3k^* + 2 - \overline{X}$, the optimal purchasing strategy makes an initial lump purchase $\alpha = k^*$, which executes the orders at prices $0,\ 1, \ldots, k^* - 1$, then purchases continuously over the interval $(0, t_0)$ at rate k^* so that $E_t \equiv k^*$ and $D_t \equiv k^* - 1$, at time t_0 makes a lump purchase of size $\beta - \alpha = 1$, which consumes the order at price k^*, then purchases continuously over the interval $(t_0, 1)$ at rate k^*+1 so that $E_t \equiv k^*+1$ and $D_t \equiv k^*$, and finally executes a lump purchase of size $e^* - \beta = k^*$ (see (4.24)) at time 1. The total quantity purchased is

$$k^* + k^* t_0 + 1 + (k^*+1)(1-t_0) + k^* = \overline{X},$$

as required.

Case C. $3k^*+2 \le \overline{X} < 3k^*+3$. We define $e^* = \overline{X} - k^* - 1$, so that $2k^*+1 \le e^* < 2k^*+2$ and $\overline{X} - e^* = k^*+1$. Then $2k^*+1 \in \partial\Phi(e^*)$ and $\widehat{g}(\overline{X} - e^*) = [2k^*, 2k^*+2]$, and the intersection of $\partial\Phi(e^*)$ and $\partial\widehat{g}(\overline{X} - e^*)$ is nonempty, as desired. On the other hand, if $e > e^*$, then $\partial\Phi(e) \subset [2k^*+1, \overline{X}]$ and $\partial\widehat{g}(\overline{X} - e) \subset [0, 2k^*]$, so the intersection of these two sets is empty. In this case, y^* and x^* are both equal to k^*+1. The optimal purchasing strategy falls into either the second (if $k^*=0$) or third (if $k^* \ge 1$) subcase of Case I of the proof of Theorem 4.5.

Appendix A. The process E. In this appendix we prove that there exists a unique adapted process E satisfying (2.3) pathwise, and we provide a list of its properties.

Lemma A.1. *Let h be a nondecreasing, real-valued, locally Lipschitz function defined on $[0,\infty)$ such that $h(0)=0$. Let X be a purchasing strategy. Then there exists a unique bounded adapted process E depending pathwise on X such that (2.3) is satisfied. Furthermore, the following hold:*

(i) *E is right continuous with left limits;*
(ii) *$\Delta E_t = \Delta X_t$ for all t;*
(iii) *E has finite variation on $[0,T]$;*
(iv) *E takes values in $[0, \overline{X}]$.*

Proof. Because we do not know a priori that E is nonnegative, we extend h to all of $\mathbb{R}$ by defining $h(x) = 0$ for $x < 0$. This extended h is nondecreasing and locally Lipschitz.

In section 2 we introduced the filtration $\{\mathcal{F}_t\}_{0 \le t \le T}$. The purchasing strategy X is right continuous and adapted to this filtration and hence is an optional process; i.e., $(t,\omega) \mapsto X_t(\omega)$ is measurable with respect to the optional σ-algebra, the σ-algebra generated by the right-continuous adapted processes. For any bounded optional process Y, $h(Y)$ and $\int_0^{\cdot} h(Y_s)\,ds$ are also bounded optional processes. Optional processes are adapted, and hence $\int_0^t h(Y_s)\,ds$ is $\mathcal{F}_t$-measurable for each $t \in [0,T]$.

We first prove uniqueness. If E and $\widehat{E}$ are bounded processes satisfying (2.3), then there is a local Lipschitz constant K, chosen taking the bounds on E and $\widehat{E}$ into account, such that

$$|E_t - \widehat{E}_t| = \left| \int_0^t \big(h(E_s) - h(\widehat{E}_s)\big)\,ds \right| \le K \int_0^t |E_s - \widehat{E}_s|\,ds.$$

Gronwall's inequality implies $E = \widehat{E}$.

For the existence part of the proof, we assume for the moment that h is globally Lipschitz with Lipschitz constant K, and we construct E as a limit of a recursion. Let $E_t^0 \equiv X_0$. For $n = 1, 2, \ldots$, define recursively

$$E_t^n = X_t - \int_0^t h(E_s^{n-1})\,ds, \quad 0 \le t \le T.$$

Since X is bounded and optional, each E^n is bounded and optional. For $n = 1, 2, \dots,$ let $Z_t^n = \sup_{0\le s\le t} |E_s^n - E_s^{n-1}|$. A proof by induction shows that

$$Z_t^n \le \frac{K^{n-1}t^{n-1}}{(n-1)!} \max\{\overline{X}, Th(X_0) + X_0\}.$$

Because this sequence of nonrandom bounds is summable, E^n converges uniformly in $t \in [0,T]$ and ω to a bounded optional process E that satisfies (2.3). In particular, E_t is $\mathcal{F}_t$-measurable for each t, and since X is nondecreasing and right continuous with left limits and the integral in (2.3) is continuous, (i), (ii), and (iii) hold.

It remains to prove (iv). For $\varepsilon > 0$, let $X_t^\varepsilon = X_t + \varepsilon t$ and define $t_0^\varepsilon = \inf\{t \in [0,T] : E_t^\varepsilon < 0\}$. Assume this set is not empty. Then the right continuity of E^ε combined with the fact that E^ε has no negative jumps implies that $E_{t_0^\varepsilon}^\varepsilon = 0$. Let $t_n^\varepsilon \downarrow t_0^\varepsilon$ be such that $E_{t_n^\varepsilon}^\varepsilon < 0$ for all n. Then

$$\int_{t_0^\varepsilon}^{t_n^\varepsilon} h(E_s^\varepsilon)\, ds = X_{t_n^\varepsilon}^\varepsilon - X_{t_0^\varepsilon}^\varepsilon - (E_{t_n^\varepsilon}^\varepsilon - E_{t_0^\varepsilon}^\varepsilon) > X_{t_n^\varepsilon}^\varepsilon - X_{t_0^\varepsilon}^\varepsilon \ge \varepsilon(t_n^\varepsilon - t_0^\varepsilon).$$

But, since

$$\int_{t_0^\varepsilon}^{t_n^\varepsilon} h(E_s^\varepsilon)\, ds \le K \left(\max_{t_0^\varepsilon \le s \le t_n^\varepsilon} E_s^\varepsilon \right) (t_n^\varepsilon - t_0^\varepsilon),$$

there must exist $s_n^\varepsilon \in (t_0^\varepsilon, t_n^\varepsilon)$ such that $E_{s_n^\varepsilon}^\varepsilon \ge \frac{\varepsilon}{K}$. This contradicts the right continuity of E^ε at t_0^ε. Consequently, the set $\{t \in [0,T] : E_t^\varepsilon < 0\}$ must be empty. We conclude that $E_t^\varepsilon \ge 0$ for all $t \in [0,T]$.

Now notice that for $0 \le t \le T$,

$$E_t^\varepsilon - E_t = \varepsilon t - \int_0^t (h(E_s^\varepsilon) - h(E_s))\, ds,$$

and hence

$$|E_t^\varepsilon - E_t| \le \varepsilon t + K \int_0^t |E_s^\varepsilon - E_s|\, ds.$$

Gronwall's inequality implies that $E^\varepsilon \to E$ as $\varepsilon \downarrow 0$. Since $E_t^\varepsilon \ge 0$, we must have $E_t \ge 0$ for all t. Equation (2.3) now implies that $E_t \le X_t$, and therefore $E_t \le \overline{X}$. The proof of (iv) is complete.

When h is locally but not globally Lipschitz, we let $\tilde{h}$ be equal to h on $[0, \overline{X}]$, $\tilde{h}(x) = 0$ for $x < 0$, and $\tilde{h}(x) = h(\overline{X})$ for $x > \overline{X}$. We apply the previous arguments to $\tilde{h}$, and we observe that the resulting $\tilde{E}$ satisfies the equation corresponding to h. ■

Remark A.2. The pathwise construction of E in the proof of Lemma A.1 shows that if X is deterministic, then so is E.

Appendix B. $\mathbb{E} \int_0^T X_t\, dA_t = 0$.

Lemma B.1. *Under the assumptions that* $0 \le X_t \le \overline{X}$, $0 \le t \le T$, *and that the continuous nonnegative martingale* A *satisfies* (2.1), *we have* $\mathbb{E} \int_0^T X_t\, dA_t = 0$.

Proof. The Burkholder–Davis–Gundy inequality implies that the continuous local martingale $M_t = \int_0^t X_s\,dA_s$ satisfies

$$\begin{aligned}\mathbb{E}\left[\max_{0\le t\le T}|M_t|\right] &\le C\mathbb{E}\big[\langle M\rangle_T^{1/2}\big]\\ &= C\mathbb{E}\left[\left(\int_0^T X_t^2\,d\langle A\rangle_t\right)^{1/2}\right]\\ &\le C\overline{X}\mathbb{E}\big[\langle A\rangle_T^{1/2}\big]\\ &\le C'\overline{X}\mathbb{E}\left[\max_{0\le t\le T}A_t\right],\end{aligned}$$

where C and C' are positive constants. By virtue of being a local martingale, M has the property that $\mathbb{E}M_{\tau_n}=0$ for a sequence of stopping times $\tau_n\uparrow T$. The dominated convergence theorem implies that $\mathbb{E}M_T=0$. ■

Appendix C. Convex hull of g.

Proof of Lemma 4.6. Recall the definition

$$\widehat{g}(y) \triangleq \sup\left\{\ell(y) : \ell \text{ is an affine function and } \ell(\eta)\le g(\eta)\ \forall \eta\in[0,\overline{Y}]\right\} \tag{4.14}$$

of the convex hull of g, defined for $y\in[0,\overline{Y}]$. The function $\widehat{g}$ is the largest convex function defined on $[0,\overline{Y}]$ that is dominated by g there.

For each $0\le y<\overline{Y}$, the supremum in (4.14) is obtained by the support line of $\widehat{g}$ at y. For $y=0$, the zero function is such a support line, and hence $0\le\widehat{g}(0)\le g(0)=0$ (recall (4.3)). At $y=\overline{Y}$ the only support line might be vertical, in which case the supremum in (4.14) is not attained. Because $\widehat{g}(0)=0$, $\widehat{g}$ is nonnegative, and $\widehat{g}$ is convex, we know that $\widehat{g}$ is also nondecreasing. Being convex, $\widehat{g}$ is continuous on $(0,\overline{Y})$ and upper semicontinuous on $[0,\overline{Y}]$, and we have continuity at 0 because of (4.3). We also have continuity of $\widehat{g}$ at $\overline{Y}$, as we now show. Given $\varepsilon>0$, the definition of $\widehat{g}$ implies that there exists an affine function $\ell\le g$ such that $\ell(\overline{Y})\ge\widehat{g}(\overline{Y})-\varepsilon$. But $\widehat{g}\ge\ell$, and thus $\liminf_{y\uparrow\overline{Y}}\widehat{g}(y)\ge\lim_{y\uparrow\overline{Y}}\ell(y)=\ell(\overline{Y})\ge\widehat{g}(\overline{Y})-\varepsilon$. Since $\varepsilon>0$ is arbitrary, we must in fact have $\liminf_{y\uparrow\overline{Y}}\widehat{g}(y)\ge\widehat{g}(\overline{Y})$. Coupled with the upper semicontinuity of $\widehat{g}$ at $\overline{Y}$, this gives us continuity.

We next argue that $\widehat{g}(\overline{Y})=g(\overline{Y})$. Suppose, on the contrary, we had $\widehat{g}(\overline{Y})<g(\overline{Y})$. The function g is continuous at $\overline{Y}$ (see (4.5)), and $\widehat{g}$ is upper semicontinuous. Therefore, there is a one-sided neighborhood $[\gamma,\overline{Y}]$ of $\overline{Y}$ (with $\gamma<\overline{Y}$) on which $g-\widehat{g}$ is bounded away from zero by a positive number ε. The function

$$\widehat{g}(y)+\frac{\varepsilon(y-\gamma)}{\overline{Y}-\gamma},\quad 0\le y\le\overline{Y},$$

is convex, lies strictly above $\widehat{g}$ at $\overline{Y}$, and lies below g everywhere. This contradicts the fact that $\widehat{g}$ is the largest convex function dominated by g. We must therefore have $\widehat{g}(\overline{Y})=g(\overline{Y})$.

Finally, we describe the situation when for some $y^*\in[0,\overline{Y}]$ we have $\widehat{g}(y^*)<g(y^*)$. We have shown that this can happen only if $0<y^*<\overline{Y}$. Let ℓ be a support line of $\widehat{g}$ at y^*, which is an affine function that attains the maximum in (4.14) at the point y^*. In particular, $\ell \le \widehat{g} \le g$ and $\ell(y^*) = \widehat{g}(y^*)$. Define

$$\alpha = \sup\{\eta \in [0, y^*] : g(\eta) - \ell(\eta) = 0\}, \tag{C.1}$$

$$\beta = \inf\{\eta \in [y^*, \overline{Y}] : g(\eta) - \ell(\eta) = 0\}. \tag{C.2}$$

Because g is lower semicontinuous, the minimum of $g - \ell$ over $[0, \overline{Y}]$ is attained. This minimum cannot be a positive number ε, for then $\ell + \varepsilon$ would be an affine function lying below g. Therefore, either the supremum in (C.1) or the infimum in (C.2) is taken over a nonempty set. In the former case, we must have $g(\alpha) = \ell(\alpha)$, whereas in the latter case $g(\beta) = \ell(\beta)$.

Let us consider first the case that $g(\alpha) = \ell(\alpha)$. Define $\gamma = \frac{1}{2}(\alpha + y^*)$. Like α, γ is strictly less than y^*. The function $g - \ell$ attains its minimum over $[\gamma, \overline{Y}]$. If this minimum were a positive number ε, then the affine function

$$\ell(y) + \frac{\varepsilon(y - \gamma)}{\overline{Y} - \gamma}, \quad 0 \le y \le \overline{Y},$$

would lie below g but have a larger value at y^* than ℓ, violating the choice of ℓ. It follows that $g - \ell$ attains the minimum value zero on $[\gamma, \overline{Y}]$, and since this function is strictly positive on $[\gamma, y^*]$, the minimum is attained to the right of y^*. This implies that $g(\beta) = \ell(\beta)$. Similarly, if we begin with the assumption that $g(\beta) = \ell(\beta)$, we can argue that $g(\alpha) = \ell(\alpha)$.

In conclusion, α and β defined by (C.1) and (C.2) satisfy (4.15) and (4.16). Finally, (4.16) shows that ℓ restricted to $[\alpha, \beta]$ is the largest affine function lying below g on this interval, and hence (4.17) holds.

Because of (4.16), every affine function lying below g on $[0, \overline{Y}]$ must lie below ℓ on $[\alpha, \beta]$. If such an affine function agrees with $\widehat{g}$ and hence with ℓ at y^*, it must in fact agree with ℓ everywhere. Hence, ℓ is the only function lying below g on $[0, \overline{Y}]$ and agreeing with $\widehat{g}$ at y^*. ∎

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 and A. Schied, *Optimal trade execution and absence of price manipulations in limit order book models,* SIAM J. Financial Math., 1 (2010), pp. 490–522.

[3] A. Alfonsi, A. Schied, and A. Slynko, *Order Book Resilience, Price Manipulation, and the Positive Portfolio Problem*, http://ssrn.com/abstract=1498514 (2011).

[4] 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.

[5] R. Almgren, *Optimal execution with nonlinear impact functions and trading-enhanced risk*, Appl. Math. Finance, 10 (2003), pp. 1–18.

[6] R. Almgren and N. Chriss, *Value under liquidation*, Risk, 12 (1999), pp. 61–63.

[7] R. Almgren and N. Chriss, *Optimal execution of portfolio transactions*, J. Risk, 3 (2001), pp. 5–39.

[8] D. Bertsimas and A. W. Lo, *Optimal control of liquidation costs*, J. Financial Markets, 1 (1998), pp. 1–50.

[9] J.-P. Bouchaud, J. D. Farmer, and F. Lillo, *How markets slowly digest changes in supply and demand*, in Handbook of Financial Markets: Dynamics and Evolution, North–Holland, Amsterdam, 2009.

[10] A. Obizhaeva and J. Wang, *Optimal Trading Strategy and Supply/Demand Dynamics*, EFA 2005 Moscow Meetings Paper; available online from http://ssrn.com/abstract=666541 (2005).
[11] P. Protter, *Stochastic Integration and Differential Equations*, 2nd ed., Springer, Berlin, 2005.
[12] J. Gatheral, *No-dynamic-arbitrage and market impact*, Quant. Finance, 10 (2010), pp. 749–759.
[13] J. Gatheral, A. Schied, and A. Slynko, *Transient linear price impact and Fredholm integral equations*, Math. Finance, to appear; available online from http://ssrn.com/abstract=1531466 (2010).
[14] J. Gatheral, A. Schied, and A. Slynko, *Exponential resilience and decay of market impact*, in Econophysics of Order-Driven Markets: Proc. Econophys-Kolkata V Conference; available online from http://ssrn.com/abstract=1650937 (2010).
[15] R. C. Grinold and R. N. Kahn, *Active Portfolio Management*, 2nd ed., McGraw–Hill, New York, 1999.
[16] G. Huberman and W. Stanzl, *Price manipulation and quasi-arbitrage*, Econometrica, 72 (2004), pp. 1247–1275.
[17] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*, Springer, New York, 1991.
[18] A. Weiss, *Executing Large Orders in a Microscopic Market Model*, http://arxiv.org/abs/0904.4131v1 (2009).

SIAM J. FINANCIAL MATH.
Vol. 2, pp. 213–220

Is the Minimum Value of an Option on Variance Generated by Local Volatility?*

Mathias Beiglböck†, Peter Friz‡, and Stephan Sturm§

Abstract. We discuss the possibility of obtaining model-free bounds on volatility derivatives, given present market data in the form of a calibrated local volatility model. A counterexample to a widespread conjecture is given.

Key words. local volatility, Dupire's formula

AMS subject classification. 91G99

JEL subject classification. G10

DOI. 10.1137/100800166

1. Introduction.

"... *it has been conjectured that the minimum possible value of an option on variance is the one generated from a local volatility model fitted to the volatility surface.*" (Gatheral [Gat06, page 155]).

Leaving precise definitions to below, let us clarify that an *option on variance* refers to a derivative whose payoff is a convex function f of *total realized variance.* Turning from convex to concave, this conjecture, if true, would also imply that the maximum possible value of a *volatility swap* ($f(x) = x^{1/2}$) is the one generated from a local volatility model fitted to the volatility surface. Given the well-documented model risk in pricing volatility swaps, such bounds are of immediate practical interest.

The mathematics of local volatility theory (à la Dupire, Derman, Kani, ...) is intimately related to the following.

Theorem 1 (see [Gyö86]). *Assume $dY_t = \mu(t,\omega)\,dt + \sigma(t,\omega)\,dB_t$ is a multidimensional Itô process, where B is a multidimensional Brownian motion, μ, σ are progressively measurable and bounded, and $\sigma\sigma^T \geq \varepsilon^2 I$ for some $\varepsilon > 0$ (σ^T denotes the transpose of σ). Then*

$$d\tilde{Y}_t = \mu_{loc}\left(t, \tilde{Y}_t\right) dt + \sigma_{loc}\left(t, \tilde{Y}_t\right) d\tilde{B}_t, \qquad \tilde{Y}_0 = Y_0, \tag{1.1}$$

*Received by the editors June 25, 2010; accepted for publication (in revised form) December 28, 2010; published electronically March 9, 2011.
http://www.siam.org/journals/sifin/2/80016.html

†Fakultät für Mathematik, Universität Wien, Nordbergstraße 15, 1090 Wien, Austria (mathias.beiglboeck@univie.ac.at). The research of this author was supported by the Austrian Science Fund (FWF) under grant P21209.

‡Institut für Mathematik, TU Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany, and Weierstraß–Institut für Angewandte Analysis und Stochastik, Mohrenstraße 39, 10117 Berlin, Germany (friz@math.tu-berlin.de, friz@wias-berlin.de). The research of this author was supported by MATHEON.

§Department of Operations Research and Financial Engineering, 116 Sherrerd Hall, Princeton University, Princeton, NJ 08544 (ssturm@princeton.edu). The research of this author (affiliated with TU Berlin while this work was started) was supported by MATHEON.

$$\mu_{loc}(t,y) = E\left[\mu(t,\omega) \mid Y_t = y\right],$$
$$\sigma_{loc}(t,y) = E\left[\sigma(t,\omega)\,\sigma^T(t,\omega) \mid Y_t = y\right]^{\frac{1}{2}}$$

(where the power $\frac{1}{2}$ denotes the positive square root of a positive definite matrix) has a weak solution $\tilde{Y}_t$ such that $\tilde{Y}_t \overset{law}{=} Y_t$ for all fixed t.

We will apply Theorem 1 only in the simple one-dimensional (resp., two-dimensional in section 4) setting, where it is well known that the solution to (1.1) is unique (cf. [Kry67] or [SV06, Chapter 7]).

A generic stochastic volatility model (already written under the appropriate equivalent martingale measure and with a suitable choice of numéraire) is of the form $dS = S\sigma dB$, where $\sigma = \sigma(t,\omega)$ is the (progressively measurable) instantaneous volatility process. (It will suffice for our application to assume σ to be bounded from above and below by positive constants.) Arguing on log-price $X = \log S$ rather than S,

$$dX_t = \sigma(t,\omega)\,dB_t - \left(\sigma^2(t,\omega)/2\right)dt, \tag{1.2}$$

a classical application of Theorem 1 yields the following Markovian projection result:[1] the (weak) solution to

$$d\tilde{X}_t = \sigma_{loc}\left(t,\tilde{X}\right)d\tilde{B}_t - \left(\sigma^2_{loc}\left(t,\tilde{X}_t\right)/2\right)dt, \qquad \tilde{X}_0 = X_0, \tag{1.3}$$
$$\left(\sigma_{loc}(t,x) = E\left[\sigma^2(t,\omega) \mid X_t = x\right]^{\frac{1}{2}}\right)$$

has the one-dimensional marginals of the original process X_t. Equivalently,[2] the process $\tilde{S} = \exp\tilde{X}$,

$$d\tilde{S}_t = \sigma_{loc}(t,\tilde{S}_t)\tilde{S}_t\,dB_t,$$

known as (Dupire's) *local volatility model*, gives rise to identical prices of all European call options $C(T,K)$.[3] It easily follows that $\sigma^2_{loc}(t,\tilde{S})$ is given by *Dupire's formula*:

$$\sigma^2_{loc}(T,\tilde{S})|_{\tilde{S}=K} = 2\frac{\partial_T C}{K^2\partial_{KK}C}. \tag{1.4}$$

Volatility derivatives are options on realized variance; that is, the payoff is given by some function f of realized variance. The latter is given by

$$V_T := \langle \log S\rangle_T = \langle X\rangle_T = \int_0^T \sigma^2(t,\omega)\,dt$$

in the model $dS = \sigma(t,\omega)\,S\,dB$ and by

$$\tilde{V}_T := \langle \log \tilde{S}\rangle_T = \langle \tilde{X}\rangle_T = \int_0^T \sigma^2_{loc}(t,\tilde{X}_t)\,dt$$

in the corresponding local volatility model.

[1] Let us quickly remark that Markovian projection techniques have recently led to a number of new applications (see [Pit06], for instance).

[2] The abuse of notation, by writing both $\sigma_{loc}(t,\tilde{X})$ and $\sigma_{loc}(t,\tilde{S})$, will not cause confusion.

[3] We emphasize that $C(T,K)$ denotes the price at time $t = 0$ of a European call with maturity T and strike K.

Common choices of f are $f(x) = x$, the *variance swap*, $f(x) = x^{1/2}$, the volatility swap, and simply $f(x) = (x - K)^+$, a *call option on realized variance*. See [FG05], for instance. As is well known (see, e.g., [Gat06]), the pricing of a variance swap, assuming continuous dynamics of S such as those specified above, is model free in the sense that it can be priced in terms of a log-contract, that is, a European option with payoff $\log S_T$. In particular, it follows that

$$E\left[\tilde{V}_T\right] = E\left[V_T\right].$$

Of course this can also be seen from (1.3), after exchanging E and integration over $[0, T]$. Passing from V_T to $f(V_T)$ for general f, this is not true, and the resulting differences are known in the industry as *convexity adjustment*. We can now formalize the conjecture given in the first lines of the introduction.[4]

Conjecture 1. *For any convex f, one has $E\big[f\big(\tilde{V}_T\big)\big] \le E\big[f\big(V_T\big)\big]$.*

Our contribution is twofold: first, we discuss a simple (toy) example which provides a counterexample to the above conjecture; second, we refine our example using a two-dimensional Markovian projection (which may be interesting in its own right) and thus construct a perfectly sensible Markovian stochastic volatility model in which the conjectured result fails. All this narrows the class of possible dynamics for S for which the conjecture can hold true and thus should be a useful step towards positive answers.

2. Idea and numerical evidence.

Example 2. Consider a Black–Scholes “mixing” model $dS = S\sigma dB$, $S_0 = 1$ with time horizon $T = 3$ in which $\sigma^2(t, \omega)$ is given by $\sigma_+^2(t)$ or $\sigma_-^2(t)$,

$$\sigma_+^2(t) := \begin{cases} 2 & \text{if } t \in [0,1], \\ 3 & \text{if } t \in]1,2], \\ 1 & \text{if } t \in]2,3], \end{cases} \qquad \sigma_-^2(t) := \begin{cases} 2 & \text{if } t \in [0,1], \\ 1 & \text{if } t \in]1,2], \\ 3 & \text{if } t \in]2,3], \end{cases}$$

depending on a fair coin flip $\epsilon = \pm 1$ (independent of B). Obviously $V = V_3 = \int_0^3 \sigma^2 \, dt \equiv 6$ in this example; hence $E\big[(V-6)^+\big] = (V-6)^+ = 0$. On the other hand, the local volatility is explicitly computable (cf. the following section), and one can see from simple Monte Carlo simulations that for $\tilde{V} = \tilde{V}_3$

$$E\left[(\tilde{V} - 6)^+\right] \approx 0.026 > 0,$$

thereby (numerically) contradicting Conjecture 1, with $f(x) = (x-6)^+$.

Our analysis of this toy model is simple enough: in section 3 we prove that $P[\tilde{V} = 6] \neq 1$. Since $E[\tilde{V}] = E[V] = 6$ and $(x-6)^+$ is strictly convex at $x = 6$, Jensen's inequality then tells us that $E[(\tilde{V} - 6)^+] > 0 = E[(V-6)^+]$.

The reader may note that an even simpler construction would be possible; i.e., one could simply leave out the interval $[0,1]$ where σ_+^2 and σ_-^2 coincide. We decided not to do so for two reasons. First, insisting on $\sigma_+^2(t) = \sigma_-^2(t)$ for $t \in [0,1]$ leads to well-behaved coefficients of the SDE describing the local volatility model. Second, we will use the present setup to obtain a complete model contradicting Conjecture 1 at the end of the next section.

---

[4] It is tacitly assumed that $f(V_T), f(\tilde{V}_T)$ are integrable.

3. Analysis of the toy example. We recall that it suffices to show that $\tilde{V} = \int_0^3 \sigma_{loc}^2(t, \tilde{X}_t)\, dt$ is not a.s. equal to $V \equiv 6$. The distribution of X_t is simply the mixture of two normal distributions. More explicitly, $X_t = I_{\{\epsilon=+1\}} X_{t,+} + I_{\{\epsilon=-1\}} X_{t,-}$,

$$X_{t,\pm} = \int_0^t \sigma_\pm(s)\, dB_s - \tfrac{1}{2}\int_0^t \sigma_\pm^2(s)\, ds \quad \sim \quad N\Big(\tfrac{1}{2}\Sigma_\pm(t), \Sigma_\pm(t)\Big),$$

where $\Sigma_\pm(t) := \int_0^t \sigma_\pm^2(s)\, ds$. Thus $\sigma_{loc}^2(t,x) = E[\sigma^2(t,\omega) \mid X_t = x]$ is given by[5]

$$\sigma_{loc}^2(t,x) = \frac{\frac{\sigma_+^2(t)}{\sqrt{\Sigma_+(t)}} \exp\left[-\frac{(x+\Sigma_+(t)/2)^2}{2\Sigma_+(t)}\right] + \frac{\sigma_-^2(t)}{\sqrt{\Sigma_-(t)}} \exp\left[-\frac{(x+\Sigma_-(t)/2)^2}{2\Sigma_-(t)}\right]}{\frac{1}{\sqrt{\Sigma_+(t)}} \exp\left[-\frac{(x+\Sigma_+(t)/2)^2}{2\Sigma_+(t)}\right] + \frac{1}{\sqrt{\Sigma_-(t)}} \exp\left[-\frac{(x+\Sigma_-(t)/2)^2}{2\Sigma_-(t)}\right]}. \tag{3.1}$$

Since $\sigma_{loc} = \sigma_{loc}(s,x)$ is bounded and measurable in t and Lipschitz in x (uniformly w.r.t. t) and bounded away from zero, it follows from [SV06, Theorem 5.1.1] that the SDE

$$d\tilde{X}_t = \sigma_{loc}\left(t, \tilde{X}\right) dB_t - \tfrac{1}{2}\sigma_{loc}^2\left(t, \tilde{X}_t\right) dt$$

has a unique strong solution (started from $\tilde{X}_0 = 0$, say). Since σ_{loc} is uniformly bounded away from 0, it follows that the process $(\tilde{X}_t)$ has *full support*; i.e., for every continuous $\varphi : [0,3] \to \mathbb{R}$, $\varphi(0) = 0$ and every $\varepsilon > 0$,

$$P[\|\tilde{X}_t - \varphi(t)\|_{\infty;[0,T]} \leq \varepsilon] > 0.$$

Indeed, there are various ways to see this: one can apply the Stroock–Varadhan support theorem in the form of [Pin95, Theorem 6.3] (several simplifications arise in the proof thanks to the one-dimensionality of the present problem); alternatively, one can employ localized lower heat kernel bounds (à la Fabes and Stroock [FS86]) or exploit that the Itô map is continuous here (thanks to Doss and Sussman; see, for instance, [RW00, page 180]) and deduce the support statement from the full support of B.

Figure 1 illustrates the dependence of $\sigma_{loc}^2(t,x)$ on time t and log-moneyness x. To gain our end of proving that $\tilde{V}(\omega) = \int_0^3 \sigma_{loc}^2(t, \tilde{X}_t)\, dt$ is not constantly equal to 6, we can determine a set of paths $(\tilde{X}_t(\omega))$ for which $\tilde{V}$ is strictly larger than 6. In view of Figure 1 it is natural to consider paths which are large, i.e., $\tilde{X}_t(\omega) \in [8, 10]$, for $t \in]1, 2 - \frac{1}{10}]$ and small, i.e., $|\tilde{X}_t(\omega)| \leq 1$, on the interval $]2,3]$. A short *Mathematica* calculation reveals that $\tilde{V}(\omega) \gtrsim 6.65 > 6$ for each such path, and according to the full-support statement the set of all such paths has positive probability; hence $\tilde{V}$ is indeed not deterministic.

Using elementary analysis it is not difficult to turn numerical evidence into rigorous mathematics. Making (3.1) explicit yields that $\sigma_{loc}^2(t,x) \equiv 2$ for $t \in [0,1]$ and that

$$\sigma_{loc}^2(t+1,x) = \frac{\frac{3}{\sqrt{2+3t}} e^{-\frac{(2x+2+3t)^2}{8(2+3t)}} + \frac{1}{\sqrt{2+t}} e^{-\frac{(2x+2+t)^2}{8(2+t)}}}{\frac{1}{\sqrt{2+3t}} e^{-\frac{(2x+2+3t)^2}{8(2+3t)}} + \frac{1}{\sqrt{2+t}} e^{-\frac{(2x+2+t)^2}{8(2+t)}}}, \tag{3.2}$$

[5] More general expressions for local volatility are found in [BM06, Chapter 4] and [Lee01, HL09]. Note the necessity to keep $\sigma^2(.,\omega)$ constant on some interval $[0,\varepsilon]$; for otherwise the local volatility surface is not Lipschitz in x uniformly as $t \to 0$.

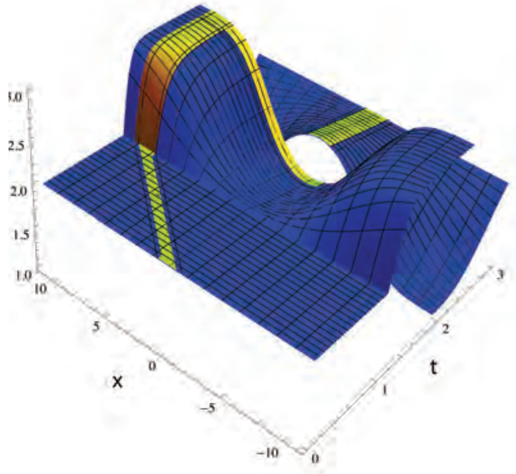


Figure 1. *Time evolution of local variance $\sigma^2_{loc}(t,x)$ in dependence of log-moneyness. The bright strip indicates a set of paths with realized variance strictly larger than* 6.

$$\sigma^2_{loc}(t+2,x) = \frac{\frac{1}{\sqrt{5+t}} e^{-\frac{(2x+5+t)^2}{8(5+t)}} + \frac{3}{\sqrt{3+3t}} e^{-\frac{(2x+3+3t)^2}{8(3+3t)}}}{\frac{1}{\sqrt{5+t}} e^{-\frac{(2x+5+t)^2}{8(5+t)}} + \frac{1}{\sqrt{3+3t}} e^{-\frac{(2x+3+3t)^2}{8(3+3t)}}} \tag{3.3}$$

for $t \in]0,1]$. We fix $\varepsilon \in]0,1]$ and observe that it is simple to see that $\lim_{x\to\infty} \sigma^2_{loc}(t+1,x) = 3$ uniformly w.r.t. $t \in [\varepsilon, 1]$ and that $\lim_{x\to 0} \sigma^2_{loc}(t+1,x) \geq 2$ uniformly w.r.t. $t \in]0,1]$. It follows that there exists some $\delta > 0$ such that

$$\begin{aligned} &\sigma^2_{loc}(t+1,x) \geq 3-\varepsilon \text{ for } x > \tfrac{1}{\delta}, \quad t \in [\varepsilon,1], \\ &\sigma^2_{loc}(t+1,x) \geq 2-\varepsilon \text{ for } |x| < \delta, \quad t \in]0,1]. \end{aligned}$$

Thus we obtain

$$\tilde{V}(\omega) = \int_0^3 \sigma^2_{loc}(t, \tilde{X}_t(\omega))\, dt \geq 1\cdot 2 + (1-2\varepsilon)\cdot(3-\varepsilon) + 1\cdot(2-\varepsilon) \tag{3.4}$$

for every path $\tilde{X}(\omega)$ satisfying $\tilde{X}_t(\omega) > \frac{1}{\delta}$ for $t \in [1+\varepsilon, 2-\varepsilon]$ and $|\tilde{X}_t(\omega)| < \delta$ for $t \in [2,3]$. This set of paths $\tilde{X}(\omega)$ has positive probability, and the quantity on the right-hand side of (3.4) is strictly larger than 6, provided that ε was chosen sufficiently small. Hence we find that $\tilde{V}$ is not constantly equal to 6 as required.

For what it's worth, the example can be modified such that volatility is adapted to the filtration of the driving Brownian motion.

The trick is to choose a random sign $\hat{\epsilon}$, $P(\hat{\epsilon} = +1) = P(\hat{\epsilon} = -1) = \frac{1}{2}$ depending solely on the behavior of $(B_t)_{0\leq t\leq 1}$ and in such a way that S_1 is *independent* of $\hat{\epsilon}$. For instance, if we let $m(s)$ be the unique number satisfying $P(S_{1/2} > m(s) \mid S_1 = s) = P(S_{1/2} \leq m(s) \mid S_1 = s) = \frac{1}{2}$, it is sensible to define $\hat{\epsilon} := +1$ if $S_{1/2} > m(S_1)$ and $\hat{\epsilon} := -1$ otherwise.

We then leave the stock price process unchanged on $[0,1]$; i.e., we define $\hat{\sigma}^2(t) = \sigma^2(t) = 2$ and $\hat{S}_t = S_t$ for $t \in [0,1]$. On $]1,2]$ (resp., $]2,3]$) we set $\hat{\sigma}^2(t) := 2 + \hat{\epsilon}$ (resp., $\hat{\sigma}^2(t) := 2 - \hat{\epsilon}$) and define $\hat{S}_t$, $t \in]1,3]$ as the solution of the SDE

$$d\hat{S}_t = \hat{\sigma}(t)\hat{S}_t\, dB_t, \quad \hat{S}_1 = S_1. \tag{3.5}$$

Here (3.5) depends only on S_1 and the process $(B_t - B_1)_{1\le t\le 3}$; since both are independent of $\hat{\epsilon}$, we obtain that $(\hat{S}_t)_{1\le t\le 3}$ and $(S_t)_{1\le t\le 3}$ are equivalent in law. It follows that $\hat{V} = \int_0^3 \hat{\sigma}^2(t,\omega)\, dt \equiv 6$, and since $\hat{S}_t$ and S_t have the same law for each $t \in [0,3]$, they induce the same local volatility model and in particular the same (nondeterministic) $\tilde{V}$.

4. Counterexample for a Markovian stochastic volatility model.

Recall that X denotes the log-price process of a general stochastic volatility model,

$$dX_t = \sigma(t,\omega)\, dB_t - \left(\sigma^2(t,\omega)/2\right) dt,$$

where $\sigma = \sigma(t,\omega)$ is the (progressively measurable) instantaneous volatility process.[6] Recall also our standing assumption that σ is bounded from above and below by positive constants. We would like to apply Theorem 1 to the two-dimensional diffusion (X,V), where $dV = \sigma^2 dt$ keeps track of the running realized variance.[7] We can do so only after elliptic regularization. That is, we consider

$$\begin{aligned} dX_t &= \sigma(t,\omega)\, dB_t - \left(\sigma^2(t,\omega)/2\right) dt,\\ da_t^\varepsilon &= \sigma^2(t,\omega)\, dt + \varepsilon^{1/2} dZ_t, \end{aligned}$$

where Z is a Brownian motion, independent of the filtration generated by B and σ. It follows that the "double-local" volatility model

$$\begin{aligned} d\tilde{X}_t^\varepsilon &= \sigma_{dloc}\left(t, \tilde{X}_t^\varepsilon, \tilde{a}_t^\varepsilon\right) dB_t - \left(\sigma_{dloc}^2\left(t, \tilde{X}_t^\varepsilon, \tilde{a}_t^\varepsilon\right)/2\right) dt,\\ d\tilde{a}_t^\varepsilon &= \sigma_{dloc}^2\left(t, \tilde{X}_t^\varepsilon, \tilde{a}_t^\varepsilon\right) dt + \varepsilon^{1/2} dZ_t \end{aligned}$$

(with $\sigma_{dloc}^2(t,x,a) = E[\sigma^2(t,\omega) \mid X_t = x,\ a_t^\varepsilon = a]$) has the one-dimensional marginals of the original process (X_t, a_t^ε). That is, for all fixed t and ε,

$$X_t \overset{law}{=} \tilde{X}_t^\varepsilon \quad \text{and} \quad \tilde{a}_t^\varepsilon \overset{law}{=} a_t^\varepsilon.$$

Let us also note that the law of a_t^ε is the law of $V_t = a_t^0$ convolved with a standard Gaussian of mean 0 and variance ε. The log-price processes X and $\tilde{X}^\varepsilon$ induce the same local volatility surface. To this end, just observe that $X_t \overset{law}{=} \tilde{X}_t^\varepsilon$ implies identical call option prices for all strikes and maturities and hence (by Dupire's formula) the same local volatility:

$$\sigma_{loc}^2(t,x) = E\left[\sigma^2(t,\omega) \mid X_t = x\right] = E\left[\sigma_{dloc}^2(t, \tilde{X}_t^\varepsilon, \tilde{a}_t^\varepsilon) \mid \tilde{X}_t^\varepsilon = x\right].$$

---

[6] We note that we intend to insert the volatility of Example 2 later, but so far our considerations hold in general.

[7] In other words, $V_T = \int_0^T \sigma^2(t,\omega)\, dt$ if $V_0 = 0$, which we shall assume from here on.

Since the law of a time inhomogeneous Markov process is fully specified by its generator, it follows that the law of the local volatility process associated with (X) has the same law as the local volatility process associated with $(\tilde{X}^{\varepsilon})$.

We apply this to the toy model discussed earlier. Recall that in this example, with $T = 3$,

$$V_T = \int_0^T \sigma^2(t,\omega)\, dt = 6,$$

whereas realized variance under the corresponding local volatility model,

$$\tilde{V}_T = \int_0^T \sigma^2_{loc}\left(t, \tilde{X}_t\right) dt,$$

was seen to be random (but with mean V_T, thanks to the matching variance swap prices). As a particular consequence, using Jensen,

$$\begin{aligned} E\left(\int_0^T \sigma^2_{loc}(t,\tilde{X}_t)\, dt - 6\right)^+ &> \left(E\int_0^T \sigma^2_{loc}(t,\tilde{X}_t)\, dt - 6\right)^+ \\ &= \left(E\int_0^T \sigma^2(t,\omega)\, dt - 6\right)^+ = (V_T - 6)^+ = 0. \end{aligned}$$

We claim that this persists when replacing the abstract stochastic volatility model (X) by $(\tilde{X}^{\varepsilon})$, the first component of a two-dimensional Markov diffusion, for any $\varepsilon > 0$. Indeed, thanks to the identical laws of the respective local volatility processes, the left-hand side above does not change when replacing $(\tilde{X})$ by the local volatility process associated with $(\tilde{X}^{\varepsilon})$. On the other hand,

$$\begin{aligned} E\int_0^T \sigma^2_{dloc}(t,\tilde{X}^{\varepsilon}_t,\tilde{a}^{\varepsilon}_t)\, dt &= E(\tilde{a}^{\varepsilon}_T - \varepsilon^{1/2} Z_T) \\ &= E\left(\tilde{a}^{\varepsilon}_T\right) = E\left(a^{\varepsilon}_T\right) \\ &= E\big(V_T + \varepsilon^{1/2} Z_T\big) = V_T. \end{aligned}$$

Thus, insisting again that the process $\tilde{X}$ be (in law) the local volatility model associated with the double-local volatility model $(\tilde{X}^{\varepsilon}, \tilde{a}^{\varepsilon})$, we see that

$$c = E\left(\int_0^T \sigma^2_{loc}(t,\tilde{X}_t)\, dt - 6\right)^+ > \left(E\int_0^T \sigma^2_{dloc}(t,\tilde{X}^{\varepsilon}_t,\tilde{a}^{\varepsilon}_t)\, dt - 6\right)^+ = 0.$$

(Observe that $c > 0$ is independent of ε.) Using the Lipschitz property of the hockeystick function, again Gyöngy, and the fact that a^{ε}_T is normally distributed with mean V_T and variance εT, we can conclude that

$$\begin{aligned} E\left(\int_0^T \sigma^2_{dloc}(t,\tilde{X}^{\varepsilon}_t,\tilde{a}^{\varepsilon}_t)\, dt - 6\right)^+ &= E\left(\int_0^T \sigma^2_{dloc}(t,\tilde{X}^{\varepsilon}_t,\tilde{a}^{\varepsilon}_t)\, dt + \varepsilon^{\frac{1}{2}} Z_T - 6 - \varepsilon^{\frac{1}{2}} Z_T\right)^+ \\ &\le E\big((\tilde{a}^{\varepsilon}_T - 6)^+ + |\varepsilon^{\frac{1}{2}} Z_T|\big) \\ &= E(a^{\varepsilon}_T - 6)^+ + E|\varepsilon^{\frac{1}{2}} Z_T| = 3\sqrt{\varepsilon T/2\pi}. \end{aligned}$$

Now we choose ε small enough such that $3\sqrt{\varepsilon T/2\pi} < c$, whence the conjecture fails to hold true in the double-local volatility model for $\varepsilon > 0$ small enough.

5. Conclusion. Summing up, the double-local volatility model constitutes an example of a continuous two-dimensional Markovian stochastic volatility model, where stochastic volatility is a function of both state variables, in which Conjecture 1 fails, i.e., in which the minimal possible value of a call option is not generated by a local volatility model.

Acknowledgment. All authors thank Gerard Brunick, Johannes Muhle-Karbe, and Walter Schachermayer for useful comments.

REFERENCES

[BM06] D. Brigo and F. Mercurio, *Interest Rate Models—Theory and Practice: With Smile, Inflation and Credit*, 2nd ed., Springer Finance, Springer-Verlag, Berlin, 2006.

[FS86] E. B. Fabes and D. W. Stroock, *A new proof of Moser's parabolic Harnack inequality using the old ideas of Nash*, Arch. Rational Mech. Anal., 96 (1986), pp. 327–338.

[FG05] P. Friz and J. Gatheral, *Valuation of volatility derivatives as an inverse problem*, Quant. Finance, 5 (2005), pp. 531–542.

[Gat06] J. Gatheral, *The Volatility Surface: A Practitioner's Guide*, Wiley, New York, 2006.

[Gyö86] I. Gyöngy, *Mimicking the one-dimensional marginal distributions of processes having an Itô differential*, Probab. Theory Related Fields, 71 (1986), pp. 501–516.

[HL09] P. Henry-Labordère, *Calibration of local stochastic volatility models to market smiles*, Risk Magazine, September (2009).

[Kry67] N. V. Krylov, *The first boundary value problem for elliptic equations of second order*, Differencial'nye Uravnenija, 3 (1967), pp. 315–326.

[Lee01] R. Lee, *Implied and local volatilities under stochastic volatility*, Int. J. Theor. Appl. Finance, 4 (2001), pp. 45–89.

[Pin95] R. G. Pinsky, *Positive Harmonic Functions and Diffusion*, Cambridge Stud. Adv. Math. 45, Cambridge University Press, Cambridge, UK, 1995.

[Pit06] V. Piterbarg, *Markovian Projection Method for Volatility Calibration*, http://ssrn.com/abstract=906473 (2006).

[RW00] L. C. G. Rogers and D. Williams, *Diffusions, Markov Processes, and Martingales. Vol.* 1. *Foundations*, Cambridge Math. Lib., Cambridge University Press, Cambridge, UK, 2000.

[SV06] D. W. Stroock and S. R. Srinivasa Varadhan, *Multidimensional Diffusion Processes*, Classics Math., Springer-Verlag, Berlin, 2006.

SIAM J. FINANCIAL MATH.
Vol. 2, pp. 221–254

A Fast Mean-Reverting Correction to Heston's Stochastic Volatility Model*

Jean-Pierre Fouque† and Matthew J. Lorig†

Abstract. We propose a multiscale stochastic volatility model in which a fast mean-reverting factor of volatility is built on top of the Heston stochastic volatility model. A singular perturbative expansion is then used to obtain an approximation for European option prices. The resulting pricing formulas are semianalytic, in the sense that they can be expressed as integrals. Difficulties associated with the numerical evaluation of these integrals are discussed, and techniques for avoiding these difficulties are provided. Overall, it is shown that computational complexity for our model is comparable to the case of a pure Heston model, but our correction brings significant flexibility in terms of fitting to the implied volatility surface. This is illustrated numerically and with option data.

Key words. stochastic volatility, Heston model, fast mean reversion, asymptotics, implied volatility smile/skew

AMS subject classifications. 60F99, 91B70

DOI. 10.1137/090761458

1. Introduction.

Since its publication in 1993, the Heston model [12] has received considerable attention from academics and practitioners alike. The Heston model belongs to a class of models known as stochastic volatility models. Such models relax the assumption of constant volatility in the stock price process and, instead, allow volatility to evolve stochastically through time. As a result, stochastic volatility models are able to capture some of the well-known features of the implied volatility surface, such as the volatility smile and skew (slope at the money). Among stochastic volatility models, the Heston model enjoys wide popularity because it provides an explicit, easy-to-compute, integral formula for calculating European option prices. In terms of the computational resources needed to calibrate a model to market data, the existence of such a formula makes the Heston model extremely efficient compared to models that rely on Monte Carlo techniques for computation and calibration.

Yet, despite its success, the Heston model has a number of documented shortcomings. For example, it has been statistically verified that the model misprices far in-the-money and out-of-the-money European options [6], [21]. In addition, the model is unable to simultaneously fit implied volatility levels across the full spectrum of option expirations available on the market [10]. In particular, the Heston model has difficulty fitting implied volatility levels for options with short expirations [11]. In fact, such problems are not limited to the Heston model. Any stochastic volatility model in which the volatility is modeled as a one-factor diffusion (as is the case in the Heston model) has trouble fitting implied volatility levels across all strikes and maturities [11].

*Received by the editors June 8, 2009; accepted for publication (in revised form) January 4, 2011; published electronically March 9, 2011.
http://www.siam.org/journals/sifin/2/76145.html

†Department of Statistics & Applied Probability, University of California, Santa Barbara, CA 93106-3110 (fouque@pstat.ucsb.edu, lorig@pstat.ucsb.edu). The work of the first author was partially supported by NSF grant DMS-0806461.

One possible explanation for why such models are unable to fit the implied volatility surface is that a single factor of volatility, running on a single time scale, is simply not sufficient for describing the dynamics of the volatility process. Indeed, the existence of several stochastic volatility factors running on different time scales has been well documented in literature that uses empirical return data [1], [2], [3], [5], [8], [13], [16], [18], [19]. Such evidence has led to the development of multiscale stochastic volatility models, in which instantaneous volatility levels are controlled by multiple diffusions running on different time scales (see, for example, [7]). We see value in this line of reasoning and thus develop our model accordingly.

Multiscale stochastic volatility models represent a struggle between two opposing forces. On one hand, adding a second factor of volatility can greatly improve a model's fit to the implied volatility surface of the market. On the other hand, adding a second factor of volatility often results in the loss of some, if not all, analytic tractability. Thus, in developing a multiscale stochastic volatility model, one seeks to model market dynamics as accurately as possible, while at the same time retaining a certain level of analyticity. Because the Heston model provides explicit integral formulas for calculating European option prices, it is an ideal template on which to build a multiscale model and accomplish this delicate balancing act.

In this paper, we show one way to bring the Heston model into the realm of multiscale stochastic volatility models without sacrificing analytic tractability. Specifically, we add a fast mean-reverting component of volatility on top of the Cox–Ingersoll–Ross (CIR) process that drives the volatility in the Heston model. Using the multiscale model, we perform a singular perturbation expansion, as outlined in [7], in order to obtain a correction to the Heston price of a European option. This correction is easy to implement, as it has an integral representation that is quite similar to that of the European option pricing formula produced by the Heston model.

This paper is organized as follows. In section 2 we introduce the multiscale stochastic volatility model, and we derive the resulting pricing partial differential equation (PDE) and boundary condition for the European option pricing problem. In section 3 we use a singular perturbative expansion to derive a PDE for a correction to the Heston price of a European option, and in section 4 we obtain a solution for this PDE. A proof of the accuracy of the pricing approximation is provided in section 5. In section 6 we examine how the implied volatility surface, as obtained from the multiscale model, compares with that of the Heston model, and in section 7 we present an example of calibration to market data. In Appendix A we review the dynamics of the Heston stochastic volatility model under the risk-neutral measure and present the pricing formula for European options. An explicit formula for the correction is given in Appendix B, and the issues associated with numerically evaluating the integral representations of option prices obtained from the multiscale model are explored in Appendix D.

2. Multiscale model and pricing PDE.

Consider the price X_t of an asset (stock, index, etc.) whose dynamics under the pricing risk-neutral measure are described by the following system of stochastic differential equations:

$$dX_t = rX_t dt + \Sigma_t X_t dW_t^x, \tag{2.1}$$

$$\Sigma_t = \sqrt{Z_t}\, f(Y_t), \tag{2.2}$$

$$dY_t = \frac{Z_t}{\epsilon}(m - Y_t)dt + \nu\sqrt{2}\sqrt{\frac{Z_t}{\epsilon}}\, dW_t^y, \tag{2.3}$$

$$dZ_t = \kappa(\theta - Z_t)dt + \sigma\sqrt{Z_t}\, dW_t^z. \tag{2.4}$$

Here, W_t^x, W_t^y, and W_t^z are one-dimensional Brownian motions with the correlation structure

$$d\left\langle W^x, W^y \right\rangle_t = \rho_{xy} dt, \tag{2.5}$$

$$d\left\langle W^x, W^z \right\rangle_t = \rho_{xz} dt, \tag{2.6}$$

$$d\left\langle W^y, W^z \right\rangle_t = \rho_{yz} dt, \tag{2.7}$$

where the correlation coefficients ρ_{xy}, ρ_{xz}, and ρ_{yz} are constants satisfying $\rho_{xy}^2 < 1, \rho_{xz}^2 < 1$, $\rho_{yz}^2 < 1$, and $\rho_{xy}^2 + \rho_{xz}^2 + \rho_{yz}^2 - 2\rho_{xy}\rho_{xz}\rho_{yz} < 1$ in order to ensure positive definiteness of the covariance matrix of the three Brownian motions.

As it should be, in (2.1), the stock price discounted by the risk-free rate r is a martingale under the pricing risk-neutral measure. The volatility Σ_t is driven by two processes Y_t and Z_t, through the product $\sqrt{Z_t}\, f(Y_t)$. The process Z_t is a CIR process with long-run mean θ, rate of mean reversion κ, and "CIR volatility" σ. We assume that κ, θ, and σ are positive, and that $2\kappa\theta \geq \sigma^2$, which ensures that $Z_t > 0$ at all times, under the condition $Z_0 > 0$.

Note that, given Z_t, the process Y_t in (2.3) appears as an Ornstein–Uhlenbeck (OU) process evolving on the time scale ϵ/Z_t, and with the invariant (or long-run) distribution $\mathcal{N}(m, \nu^2)$. This way of "modulating" the rate of mean reversion of the process Y_t by Z_t has also been used in [4] in the context of interest rate modeling.

Multiple time scales are incorporated in this model through the parameter $\epsilon > 0$, which is intended to be small, so that Y_t is fast reverting.

We do not specify the precise form of $f(y)$, which will not play an essential role in the asymptotic results derived in this paper. However, in order to ensure Σ_t has the same behavior at zero and infinity as in the case of a pure Heston model, we assume there exist constants c_1 and c_2 such that $0 < c_1 \leq f(y) \leq c_2 < \infty$ for all $y \in \mathbb{R}$. Likewise, the particular choice of an OU-like process for Y_t is not crucial in the analysis. The mean-reversion aspect (or ergodicity) is the important property. In fact, we could have chosen Y_t to be a CIR-like process instead of an OU-like process without changing the nature of the correction to the Heston model presented in this paper.

Here, we consider the unique strong solution to (2.1)–(2.4) for a fixed parameter $\epsilon > 0$. Existence and uniqueness are easily obtained by (i) using the classical existence and uniqueness result for the CIR process Z_t defined by (2.4), (ii) using the representation (5.18) of the process Y_t to derive moments for a fixed $\epsilon > 0$, and (iii) using the exponential formula for X_t:

$$X_t = x \exp\left(\int_0^t \left(r - \frac{1}{2}\Sigma_s^2 \right) ds + \int_0^t \Sigma_s dW_s^x \right).$$

We note that if one chooses $f(y) = 1$, the multiscale model becomes ϵ-independent and reduces to the pure Heston model expressed under the risk-neutral measure with stock price

X_t and stochastic variance Z_t:

$$\begin{aligned} dX_t &= rX_t dt + \sqrt{Z_t} X_t dW_t^x, \\ dZ_t &= \kappa(\theta - Z_t)dt + \sigma\sqrt{Z_t}\, dW_t^z, \\ d\left\langle W^x, W^z \right\rangle_t &= \rho_{xz} dt. \end{aligned}$$

Thus, the multiscale model can be thought of as a Heston-like model with a fast-varying factor of volatility, $f(Y_t)$, built on top of the CIR process Z_t, which drives the volatility in the Heston model.

We consider a European option expiring at time $T > t$ with payoff $h(X_T)$. As the dynamics of the stock in the multiscale model are specified under the risk-neutral measure, the price of the option, denoted by P_t, can be expressed as an expectation of the option payoff, discounted at the risk-free rate:

$$P_t = \mathbb{E}\left[e^{-r(T-t)} h(X_T) \middle| X_t, Y_t, Z_t \right] =: P^\epsilon(t, X_t, Y_t, Z_t),$$

where we have used the Markov property of (X_t, Y_t, Z_t) and defined the pricing function $P^\epsilon(t,x,y,z)$, the superscript ϵ denoting the dependence on the small parameter ϵ. Using the Feynman–Kac formula, $P^\epsilon(t,x,y,z)$ satisfies the following PDE and boundary condition:

$$\mathcal{L}^\epsilon P^\epsilon(t,x,y,z) = 0, \tag{2.8}$$

$$\mathcal{L}^\epsilon = \frac{\partial}{\partial t} + \mathcal{L}_{(X,Y,Z)} - r\,, \tag{2.9}$$

$$P^\epsilon(T,x,y,z) = h(x), \tag{2.10}$$

where the operator $\mathcal{L}_{(X,Y,Z)}$ is the infinitesimal generator of the process (X_t, Y_t, Z_t):

$$\begin{aligned} \mathcal{L}_{(X,Y,Z)} = rx\frac{\partial}{\partial x} &+ \frac{1}{2} f^2(y) z x^2 \frac{\partial^2}{\partial x^2} + \rho_{xz}\sigma f(y) z x \frac{\partial^2}{\partial x \partial z} \\ &+ \kappa(\theta - z)\frac{\partial}{\partial z} + \frac{1}{2}\sigma^2 z \frac{\partial^2}{\partial z^2} \\ &+ \frac{z}{\epsilon}\left((m-y)\frac{\partial}{\partial y} + \nu^2 \frac{\partial^2}{\partial y^2} \right) \\ &+ \frac{z}{\sqrt{\epsilon}}\left(\rho_{yz}\sigma\nu\sqrt{2}\frac{\partial^2}{\partial y \partial z} + \rho_{xy}\nu\sqrt{2} f(y) x \frac{\partial^2}{\partial x \partial y} \right). \end{aligned}$$

It will be convenient to separate $\mathcal{L}^\epsilon$ into groups of like powers of $1/\sqrt{\epsilon}$. To this end, we define the operators $\mathcal{L}_0$, $\mathcal{L}_1$, and $\mathcal{L}_2$ as follows:

$$\mathcal{L}_0 := \nu^2 \frac{\partial^2}{\partial y^2} + (m-y)\frac{\partial}{\partial y}, \tag{2.11}$$

$$\mathcal{L}_1 := \rho_{yz}\sigma\nu\sqrt{2}\frac{\partial^2}{\partial y \partial z} + \rho_{xy}\nu\sqrt{2}\, f(y) x \frac{\partial^2}{\partial x \partial y}, \tag{2.12}$$

$$\begin{aligned} \mathcal{L}_2 := \frac{\partial}{\partial t} &+ \frac{1}{2} f^2(y) z x^2 \frac{\partial^2}{\partial x^2} + r\left(x\frac{\partial}{\partial x} - \cdot \right) \\ &+ \frac{1}{2}\sigma^2 z \frac{\partial^2}{\partial z^2} + \kappa(\theta - z)\frac{\partial}{\partial z} + \rho_{xz}\sigma f(y) z x \frac{\partial^2}{\partial x \partial z}. \end{aligned} \tag{2.13}$$

With these definitions, $\mathcal{L}^\epsilon$ is expressed as

$$\mathcal{L}^\epsilon = \frac{z}{\epsilon}\mathcal{L}_0 + \frac{z}{\sqrt{\epsilon}}\mathcal{L}_1 + \mathcal{L}_2. \tag{2.14}$$

Note that $\mathcal{L}_0$ is the infinitesimal generator of an OU process with unit rate of mean reversion, and $\mathcal{L}_2$ is the pricing operator of the Heston model with volatility and correlation modulated by $f(y)$.

3. Asymptotic analysis.

For a general function f, there is no analytic solution to the Cauchy problem (2.8)–(2.10). Thus, we proceed with an asymptotic analysis as developed in [7]. Specifically, we perform a singular perturbation with respect to the small parameter ϵ, expanding our solution in powers of $\sqrt{\epsilon}$:

$$P^\epsilon = P_0 + \sqrt{\epsilon}P_1 + \epsilon P_2 + \cdots . \tag{3.1}$$

We now plug (3.1) and (2.14) into (2.8) and (2.10) and collect terms of equal powers of $\sqrt{\epsilon}$.

The order $1/\epsilon$ terms. Collecting terms of order $1/\epsilon$ we have the following PDE:

$$0 = z\mathcal{L}_0 P_0. \tag{3.2}$$

We see from (2.11) that both terms in $\mathcal{L}_0$ take derivatives with respect to y. In fact, $\mathcal{L}_0$ is an infinitesimal generator, and consequently zero is an eigenvalue with constant eigenfunctions. Thus, we seek P_0 of the form

$$P_0 = P_0(t, x, z)$$

so that (3.2) is satisfied.

The order $1/\sqrt{\epsilon}$ terms. Collecting terms of order $1/\sqrt{\epsilon}$ leads to the following PDE:

$$\begin{aligned} 0 &= z\mathcal{L}_0 P_1 + z\mathcal{L}_1 P_0 \\ &= z\mathcal{L}_0 P_1. \end{aligned} \tag{3.3}$$

Note that we have used that $\mathcal{L}_1 P_0 = 0$, since both terms in $\mathcal{L}_1$ take derivatives with respect to y and P_0 is independent of y. As above, we seek P_1 of the form

$$P_1 = P_1(t, x, z)$$

so that (3.3) is satisfied.

The order 1 terms. Matching terms of order 1 leads to the following PDE and boundary condition:

$$\begin{aligned} 0 &= z\mathcal{L}_0 P_2 + z\mathcal{L}_1 P_1 + \mathcal{L}_2 P_0 \\ &= z\mathcal{L}_0 P_2 + \mathcal{L}_2 P_0, \end{aligned} \tag{3.4}$$

$$h(x) = P_0(T, x, z). \tag{3.5}$$

In deriving (3.4) we have used that $\mathcal{L}_1 P_1 = 0$, since $\mathcal{L}_1$ takes derivative with respect to y and P_1 is independent of y.

Note that (3.4) is a Poisson equation in y with respect to the infinitesimal generator $\mathcal{L}_0$ and with source term $\mathcal{L}_2 P_0$; in solving this equation, (t, x, z) are fixed parameters. In order for this equation to admit solutions with reasonable growth at infinity (polynomial growth), we impose that the source term satisfy the following *centering condition*:

$$0 = \langle \mathcal{L}_2 P_0 \rangle = \langle \mathcal{L}_2 \rangle P_0, \tag{3.6}$$

where we have used the notation

$$\langle g \rangle := \int g(y) \Phi(y) dy; \tag{3.7}$$

here Φ denotes the density of the invariant distribution of the process Y_t, which we remind the reader is $\mathcal{N}(m, \nu^2)$. Note that in (3.6) we have pulled $P_0(t, x, z)$ out of the linear $\langle \cdot \rangle$ operator, since it does not depend on y.

Note that the PDE (3.6) and the boundary condition (3.5) jointly define a Cauchy problem that $P_0(t, x, z)$ must satisfy.

Using (3.4) and the centering condition (3.6), we deduce

$$P_2 = -\frac{1}{z} \mathcal{L}_0^{-1} \left(\mathcal{L}_2 - \langle \mathcal{L}_2 \rangle \right) P_0, \tag{3.8}$$

where $\mathcal{L}_0^{-1}$ is the inverse operator of $\mathcal{L}_0$ acting on the centered functions.

The order $\sqrt{\epsilon}$ terms. Collecting terms of order $\sqrt{\epsilon}$, we obtain the following PDE and boundary condition:

$$0 = z \mathcal{L}_0 P_3 + z \mathcal{L}_1 P_2 + \mathcal{L}_2 P_1, \tag{3.9}$$

$$0 = P_1(T, x, z). \tag{3.10}$$

We note that $P_3(t, x, y, z)$ solves the Poisson equation (3.9) in y with respect to $\mathcal{L}_0$. Thus, we impose the corresponding centering condition on the source $z \mathcal{L}_1 P_2 + \mathcal{L}_2 P_1$, leading to

$$\langle \mathcal{L}_2 \rangle P_1 = - \langle z \mathcal{L}_1 P_2 \rangle . \tag{3.11}$$

Plugging P_2, given by (3.8), into (3.11) gives

$$\langle \mathcal{L}_2 \rangle P_1 = \mathcal{A} P_0, \tag{3.12}$$

$$\mathcal{A} := \left\langle z \mathcal{L}_1 \frac{1}{z} \mathcal{L}_0^{-1} \left(\mathcal{L}_2 - \langle \mathcal{L}_2 \rangle \right) \right\rangle . \tag{3.13}$$

Note that the PDE (3.12) and the zero boundary condition (3.10) define a Cauchy problem that $P_1(t, x, z)$ must satisfy.

Summary of the key results. We summarize the key results of our asymptotic analysis. We have written the expansion (3.1) for the solution of the PDE problem (2.8)–(2.10). Along the way, he have chosen solutions for P_0 and P_1 which are of the forms $P_0 = P_0(t,x,z)$ and $P_1 = P_1(t,x,z)$. These choices lead us to conclude that $P_0(t,x,z)$ and $P_1(t,x,z)$ must satisfy the following Cauchy problems:

$$\langle \mathcal{L}_2 \rangle P_0 = 0, \tag{3.14}$$

$$P_0(T,x,z) = h(x), \tag{3.15}$$

and

$$\langle \mathcal{L}_2 \rangle P_1(t,x,z) = \mathcal{A} P_0(t,x,z), \tag{3.16}$$

$$P_1(T,x,z) = 0, \tag{3.17}$$

where

$$\begin{aligned} \langle \mathcal{L}_2 \rangle = {} & \frac{\partial}{\partial t} + \frac{1}{2}\left\langle f^2 \right\rangle z x^2 \frac{\partial^2}{\partial x^2} + r\left(x\frac{\partial}{\partial x} - \cdot \right) \\ & + \frac{1}{2}\sigma^2 z \frac{\partial^2}{\partial z^2} + \kappa(\theta - z)\frac{\partial}{\partial z} + \rho_{xz}\sigma \langle f \rangle zx \frac{\partial^2}{\partial x \partial z}, \end{aligned} \tag{3.18}$$

and $\mathcal{A}$ is given by (3.13). Recall that the bracket notation is defined in (3.7).

4. Formulas for $P_0(t,x,z)$ and $P_1(t,x,z)$.

In this section we use the results of our asymptotic calculations to find explicit solutions for $P_0(t,x,z)$ and $P_1(t,x,z)$.

4.1. Formula for $P_0(t,x,z)$.

Recall that $P_0(t,x,z)$ satisfies a Cauchy problem defined by (3.14) and (3.15).

Without loss of generality, we normalize f so that $\left\langle f^2 \right\rangle = 1$. Thus, we rewrite $\langle \mathcal{L}_2 \rangle$ given by (3.18) as follows:

$$\begin{aligned} \langle \mathcal{L}_2 \rangle = {} & \frac{\partial}{\partial t} + \frac{1}{2} z x^2 \frac{\partial^2}{\partial x^2} + r\left(x\frac{\partial}{\partial x} - \cdot \right) \\ & + \frac{1}{2}\sigma^2 z \frac{\partial^2}{\partial z^2} + \kappa(\theta - z)\frac{\partial}{\partial z} + \rho\sigma z x \frac{\partial^2}{\partial x \partial z} \\ := {} & \mathcal{L}_H, \end{aligned} \tag{4.1}$$

$$\rho := \rho_{xz} \langle f \rangle . \tag{4.2}$$

We note that $\rho^2 \leq 1$, since $\langle f \rangle^2 \leq \left\langle f^2 \right\rangle = 1$. So, ρ can be thought of as an *effective correlation* between the Brownian motions in the Heston model obtained in the limit $\epsilon \to 0$, where $\langle \mathcal{L}_2 \rangle = \mathcal{L}_H$, the pricing operator for European options as calculated in the Heston model. Thus, we see that $P_0(t,x,z) =: P_H(t,x,z)$ is the classical solution for the price of a European option as calculated in the Heston model with effective correlation $\rho = \rho_{xz} \langle f \rangle$.

The derivation of pricing formulas for the Heston model is given in Appendix A. Here, we simply state the main result:

$$P_H(t,x,z) = e^{-r\tau}\frac{1}{2\pi}\int e^{-ikq}\widehat{G}(\tau,k,z)\widehat{h}(k)dk, \tag{4.3}$$

$$\tau(t) = T-t, \tag{4.4}$$

$$q(t,x) = r(T-t)+\log x, \tag{4.5}$$

$$\widehat{h}(k) = \int e^{ikq}h(e^q)dq, \tag{4.6}$$

$$\widehat{G}(\tau,k,z) = e^{C(\tau,k)+zD(\tau,k)}, \tag{4.7}$$

$$C(\tau,k) = \frac{\kappa\theta}{\sigma^2}\left(\left(\kappa+\rho ik\sigma+d(k)\right)\tau - 2\log\left(\frac{1-g(k)e^{\tau d(k)}}{1-g(k)}\right)\right), \tag{4.8}$$

$$D(\tau,k) = \frac{\kappa+\rho ik\sigma+d(k)}{\sigma^2}\left(\frac{1-e^{\tau d(k)}}{1-g(k)e^{\tau d(k)}}\right), \tag{4.9}$$

$$d(k) = \sqrt{\sigma^2(k^2-ik)+(\kappa+\rho ik\sigma)^2}, \tag{4.10}$$

$$g(k) = \frac{\kappa+\rho ik\sigma+d(k)}{\kappa+\rho ik\sigma-d(k)}. \tag{4.11}$$

We note that, for certain choices of h, the integral in (4.6) may not converge. For example, a European call with strike K has $h(e^q) = (e^q-K)^+$. In this case, the integral in (4.6) converges only if we set $k = k_r + ik_i$, where $k_i > 1$. Hence, when evaluating (4.3), (4.6), one must impose $k = k_r + ik_i$, $k_r > 1$, and $dk = dk_r$.

4.2. Formula for $P_1(t,x,z)$. Recall that $P_1(t,x,z)$ satisfies a Cauchy problem defined by (3.16) and (3.17). In order to find a solution for $P_1(t,x,z)$ we must first identify the operator $\mathcal{A}$. To this end, we introduce two functions, $\phi(y)$ and $\psi(y)$, which solve the following Poisson equations in y with respect to the operator $\mathcal{L}_0$:

$$\mathcal{L}_0\phi = \frac{1}{2}\left(f^2-\langle f^2\rangle\right), \tag{4.12}$$

$$\mathcal{L}_0\psi = f-\langle f\rangle. \tag{4.13}$$

From (3.13) we have

$$\begin{aligned}
\mathcal{A} &= \left\langle z\mathcal{L}_1\frac{1}{z}\mathcal{L}_0^{-1}\left(\mathcal{L}_2-\langle\mathcal{L}_2\rangle\right)\right\rangle \\
&= \left\langle z\mathcal{L}_1\frac{1}{z}\mathcal{L}_0^{-1}\frac{z}{2}\left(f^2-\langle f^2\rangle\right)x^2\frac{\partial^2}{\partial x^2}\right\rangle \\
&\quad + \left\langle z\mathcal{L}_1\frac{1}{z}\mathcal{L}_0^{-1}\rho_{xz}\sigma z\left(f-\langle f\rangle\right)x\frac{\partial^2}{\partial x\partial z}\right\rangle \\
&= z\left\langle \mathcal{L}_1\phi(y)x^2\frac{\partial^2}{\partial x^2}\right\rangle + \rho_{xz}\sigma z\left\langle\mathcal{L}_1\psi(y)x\frac{\partial^2}{\partial x\partial z}\right\rangle.
\end{aligned}$$

Using the definition (2.12) of $\mathcal{L}_1$, one deduces the following expression for $\mathcal{A}$:

$$\mathcal{A} = V_1 z x^2 \frac{\partial^3}{\partial z \partial x^2} + V_2 z x \frac{\partial^3}{\partial z^2 \partial x} + V_3 z x \frac{\partial}{\partial x}\left(x^2 \frac{\partial^2}{\partial x^2}\right) + V_4 z \frac{\partial}{\partial z}\left(x \frac{\partial}{\partial x}\right)^2, \tag{4.14}$$

$$V_1 = \rho_{yz}\sigma\nu\sqrt{2}\left\langle \phi' \right\rangle, \tag{4.15}$$

$$V_2 = \rho_{xz}\rho_{yz}\sigma^2\nu\sqrt{2}\left\langle \psi' \right\rangle, \tag{4.16}$$

$$V_3 = \rho_{xy}\nu\sqrt{2}\left\langle f\phi' \right\rangle, \tag{4.17}$$

$$V_4 = \rho_{xy}\rho_{xz}\sigma\nu\sqrt{2}\left\langle f\psi' \right\rangle. \tag{4.18}$$

Note that we have introduced four group parameters, V_i, $i = 1, \dots, 4$, which are constants that can be obtained by calibrating our model to the market, as will be done in section 7.

Now that we have expressions for $\mathcal{A}$, P_H, and $\mathcal{L}_H$, we are in a position to solve for $P_1(t, x, z)$, which is the solution to the Cauchy problem defined by (3.16) and (3.17). We leave the details of the calculation to Appendix B. Here, we simply present the main result:

$$P_1(t,x,z) = \frac{e^{-r\tau}}{2\pi}\int_{\mathbb{R}} e^{-ikq}\left(\kappa\theta\widehat{f}_0(\tau,k) + z\widehat{f}_1(\tau,k)\right) \times \widehat{G}(\tau,k,z)\widehat{h}(k)dk, \tag{4.19}$$

$$\tau(t) = T - t,$$

$$q(t,x) = r(T-t) + \log x,$$

$$\widehat{h}(k) = \int e^{ikq}h(e^q)dq,$$

$$\widehat{G}(\tau,k,z) = e^{C(\tau,k)+zD(\tau,k)},$$

$$\widehat{f}_0(\tau,k) = \int_0^\tau \widehat{f}_1(s,k)ds, \tag{4.20}$$

$$\widehat{f}_1(\tau,k) = \int_0^\tau b(s,k)e^{A(\tau,k,s)}ds, \tag{4.21}$$

$$C(\tau,k) = \frac{\kappa\theta}{\sigma^2}\left(\left(\kappa + \rho ik\sigma + d(k)\right)\tau - 2\log\left(\frac{1 - g(k)e^{\tau d(k)}}{1 - g(k)}\right)\right),$$

$$D(\tau,k) = \frac{\kappa + \rho ik\sigma + d(k)}{\sigma^2}\left(\frac{1 - e^{\tau d(k)}}{1 - g(k)e^{\tau d(k)}}\right),$$

$$A(\tau,k,s) = \left(\kappa + \rho\sigma ik + d(k)\right)\frac{1 - g(k)}{d(k)g(k)}\log\left(\frac{g(k)e^{\tau d(k)} - 1}{g(k)e^{sd(k)} - 1}\right)$$

$$
\begin{aligned}
&\qquad + d(k)\,(\tau - s)\,, && (4.22)\\
d(k) &= \sqrt{\sigma^2(k^2 - ik) + (\kappa + \rho i k \sigma)^2}\,,\\
g(k) &= \frac{\kappa + \rho i k \sigma + d(k)}{\kappa + \rho i k \sigma - d(k)},\\
b(\tau,k) &= -\left(V_1 D(\tau,k)\left(-k^2 + ik\right) + V_2 D^2(\tau,k)\,(-ik)\right.\\
&\qquad \left. + V_3\left(ik^3 + k^2\right) + V_4 D(\tau,k)\left(-k^2\right)\right). && (4.23)
\end{aligned}
$$

Once again, we note that, depending on the option payoff, evaluating (4.19) may require setting $k = k_r + ik_i$ and $dk = dk_r$, as described at the end of subsection 4.1.

5. Accuracy of the approximation. In this section, we prove that the approximation $P^\epsilon \sim P_0 + \sqrt{\epsilon}P_1$, where P_0 and P_1 are defined in the previous sections, is accurate to order ϵ^α for any given $\alpha \in (1/2, 1)$. Specifically, for a European option with a payoff h such that $h(e^\xi)$ belongs to the Schwartz class of rapidly decaying functions with respect to the log-price variable $\xi = \log x$, we will show

$$
|P^\epsilon(t,x,y,z) - \left(P_0(t,x,z) + \sqrt{\epsilon}P_1(t,x,z)\right)| \le C\,\epsilon^\alpha, \tag{5.1}
$$

where C is a constant which depends on (y,z) but is independent of ϵ.

We start by defining the remainder term $R^\epsilon(t,x,y,z)$:

$$
R^\epsilon = \left(P_0 + \sqrt{\epsilon}P_1 + \epsilon P_2 + \epsilon\sqrt{\epsilon}P_3\right) - P^\epsilon. \tag{5.2}
$$

Recalling that

$$
\begin{aligned}
0 &= \mathcal{L}^\epsilon P^\epsilon,\\
0 &= z\mathcal{L}_0 P_0,\\
0 &= z\mathcal{L}_0 P_1 + z\mathcal{L}_1 P_0,\\
0 &= z\mathcal{L}_0 P_2 + z\mathcal{L}_1 P_1 + \mathcal{L}_2 P_0,\\
0 &= z\mathcal{L}_0 P_3 + z\mathcal{L}_1 P_2 + \mathcal{L}_2 P_1
\end{aligned}
$$

and applying $\mathcal{L}^\epsilon$ to R^ϵ, we obtain that R^ϵ must satisfy the following PDE:

$$
\begin{aligned}
\mathcal{L}^\epsilon R^\epsilon &= \mathcal{L}^\epsilon\left(P_0 + \sqrt{\epsilon}P_1 + \epsilon P_2 + \epsilon\sqrt{\epsilon}P_3\right) - \mathcal{L}^\epsilon P^\epsilon\\
&= \left(\frac{z}{\epsilon}\mathcal{L}_0 + \frac{z}{\sqrt{\epsilon}}\mathcal{L}_1 + \mathcal{L}_2\right)\left(P_0 + \sqrt{\epsilon}P_1 + \epsilon P_2 + \epsilon\sqrt{\epsilon}P_3\right)\\
&= \epsilon\left(z\mathcal{L}_1 P_3 + \mathcal{L}_2 P_2 + \sqrt{\epsilon}\mathcal{L}_2 P_3\right)\\
&= \epsilon\, F^\epsilon, && (5.3)\\
F^\epsilon &:= z\mathcal{L}_1 P_3 + \mathcal{L}_2 P_2 + \sqrt{\epsilon}\mathcal{L}_2 P_3, && (5.4)
\end{aligned}
$$

where we have defined the ϵ-dependent source term $F^\epsilon(t,x,y,z)$. Recalling that

$$
\begin{aligned}
P^\epsilon(T,x,y,z) &= h(x),\\
P_0(T,x,z) &= h(x),\\
P_1(T,x,z) &= 0,
\end{aligned}
$$

we deduce from (5.2) that

$$
\begin{aligned}
R^\epsilon(T,x,y,z) &= \epsilon P_2(T,x,y,z) + \epsilon\sqrt{\epsilon}P_3(T,x,y,z) \\
&= \epsilon\, G^\epsilon(x,y,z), \qquad (5.5)\\
G^\epsilon(x,y,z) &:= P_2(T,x,y,z) + \sqrt{\epsilon}P_3(T,x,y,z), \qquad (5.6)
\end{aligned}
$$

where we have defined the ϵ-dependent boundary term $G^\epsilon(x,y,z)$.

Using the expression (2.9) for $\mathcal{L}^\epsilon$, we find that $R^\epsilon(t,x,y,z)$ satisfies the following Cauchy problem with source:

$$
\left(\frac{\partial}{\partial t} + \mathcal{L}_{X,Y,Z} - r\right) R^\epsilon = \epsilon\, F^\epsilon, \qquad (5.7)
$$

$$
R^\epsilon(T,x,y,z) = \epsilon\, G^\epsilon(x,y,z). \qquad (5.8)
$$

Therefore R^ϵ admits the following probabilistic representation:

$$
\begin{aligned}
R^\epsilon(t,x,y,z) = \epsilon\,\mathbb{E}\Bigg[&e^{-r(T-t)}G^\epsilon(X_T,Y_T,Z_T) \\
&- \int_t^T e^{-r(s-t)}F^\epsilon(s,X_s,Y_s,Z_s)ds \mid X_t = x, Y_t = y, Z_t = z\Bigg]. \qquad (5.9)
\end{aligned}
$$

In order to bound $R^\epsilon(T,x,y,z)$, we need bounds on the growth of $F^\epsilon(t,x,y,z)$ and $G^\epsilon(x,y,z)$. From (5.6) we see that $G^\epsilon(x,y,z)$ contains the functions $P_2(t,x,y,z)$ and $P_3(t,x,y,z)$. And from (5.4) we see that $F^\epsilon(t,x,y,z)$ contains terms with the linear operators, $\mathcal{L}_1$ and $\mathcal{L}_2$, acting on $P_2(t,x,y,z)$ and $P_3(t,x,y,z)$. Thus, to bound $F^\epsilon(t,x,y,z)$ and $G^\epsilon(x,y,z)$, we need to obtain growth estimates for $P_2(t,x,y,z)$ and $P_3(t,x,y,z)$ and growth estimates for $P_2(t,x,y,z)$ and $P_3(t,x,y,z)$ when linear operators act upon them. To do this, we use the following classical result, which can be found in Chapter 5 of [7].

Lemma 5.1. *Suppose $\mathcal{L}_0\chi = g$, $\langle g\rangle = 0$, and $|g(y)| < C_1(1+|y|^n)$; then $|\chi(y)| < C_2(1+|y|^n)$ for some C_2. When $n = 0$ we have $|\chi(y)| < C_2(1+\log(1+|y|))$.*

Now, by continuing the asymptotic analysis of section 3, we find that $P_2(t,x,y,z)$ and $P_3(t,x,y,z)$ satisfy Poisson equations in y with respect to the operator, $\mathcal{L}_0$. We have

$$
\begin{aligned}
\mathcal{L}_0 P_2(t,x,y,z) &= \frac{1}{z}\left(-\mathcal{L}_2 + \langle\mathcal{L}_2\rangle\right) P_0(t,x,z), \\
\mathcal{L}_0 P_3(t,x,y,z) &= \frac{1}{z}\left(-\mathcal{L}_2 + \langle\mathcal{L}_2\rangle\right) P_1(t,x,z) + \left(-\mathcal{L}_1 P_2(t,x,y,z) + \langle\mathcal{L}_1 P_2(t,x,y,z)\rangle\right).
\end{aligned}
$$

Also note that, for any operator, $\mathcal{M}$, of the form

$$
\mathcal{M} = \frac{\partial^m}{\partial z^m}\prod_{j=1}^N x^{n(j)}\frac{\partial^{n(j)}}{\partial x^{n(j)}}, \qquad (5.10)
$$

we have $\mathcal{M}\mathcal{L}_0 = \mathcal{L}_0\mathcal{M}$ because $\mathcal{L}_0$ does not contain x or z. Hence, $\mathcal{M}P_2(t,x,y,z)$ and $\mathcal{M}P_3(t,x,y,z)$ satisfy the following Poisson equations in y with respect to the operator, $\mathcal{L}_0$:

$$\begin{aligned} \mathcal{L}_0\left(\mathcal{M}P_2(t,x,y,z)\right) &= \mathcal{M}\frac{1}{z}\left(-\mathcal{L}_2+\langle\mathcal{L}_2\rangle\right)P_0(t,x,z), \\ \mathcal{L}_0\left(\mathcal{M}P_3(t,x,y,z)\right) &= \mathcal{M}\frac{1}{z}\left(-\mathcal{L}_2+\langle\mathcal{L}_2\rangle\right)P_1(t,x,z) \\ &\quad + \mathcal{M}\left(-\mathcal{L}_1P_2(t,x,y,z)+\langle\mathcal{L}_1P_2(t,x,y,z)\rangle\right). \end{aligned} \tag{5.11}$$

Let us bound functions of the form $\mathcal{M}P_0(t,x,z)$. Using (4.3) and (5.10) and recalling that $q = r\tau + \log x$ and $\widehat{G} = e^{C+zD}$, we have

$$\begin{aligned} \mathcal{M}P_0 &= \frac{e^{-r\tau}}{2\pi}\int\left(\prod_{j=1}^{N}x^{n(j)}\frac{\partial^{n(j)}}{\partial x^{n(j)}}e^{-ikq}\right)\left(\frac{\partial^m}{\partial z^m}e^{C(\tau,k,z)+zD(\tau,k,z)}\right)\widehat{h}(k)dk \\ &= \frac{e^{-r\tau}}{2\pi}\int e^{-ikq}\left(\prod_{j=1}^{N}\prod_{l=1}^{n(j)}(-ik-l+1)\right)\left((D(\tau,k,z))^m e^{C(\tau,k,z)+zD(\tau,k,z)}\right)\widehat{h}(k)dk \\ &= \frac{e^{-r\tau}}{2\pi}\int\left(\prod_{j=1}^{N}\prod_{l=1}^{n(j)}(-ik-l+1)\right)(D(\tau,k,z))^m e^{-ikq}\widehat{G}(\tau,k,z)\widehat{h}(k)dk. \end{aligned}$$

We note the following:

- By assumption, the option payoff $h(e^q) \in \mathcal{S}$, the Schwartz class of rapidly decreasing functions, so that the Fourier transform $\widehat{h}(k) \in \mathcal{S}$, and therefore $\|k^m\widehat{h}(k)\|_\infty < \infty$ for all integers, m.
- $|\widehat{G}(\tau,k,z)| \le 1$ for all $\tau \in [0,T]$, $k \in \mathbb{R}$, $z \in \mathbb{R}^+$. This follows from the fact that $\widehat{G}(\tau,k,z)$ is the characteristic function, $\mathbb{E}[\exp(ikQ_T)|X_t = x, Z_t = z]$.
- There exists a constant, C, such that $|D(\tau,k)| \le C(1+|k|)$ for all $\tau \in [0,T]$.

It follows that for any $\mathcal{M}$ of the form (5.10) we have the following bound on $\mathcal{M}P_0(t,x,z)$:

$$\begin{aligned} |\mathcal{M}P_0(t,x,z)| &\le \frac{e^{-r\tau}}{2\pi}\int\left|\prod_{j=1}^{N}\prod_{l=1}^{n(j)}(-ik-l+1)\right||D(\tau,k)|^m\left|e^{-ikq}\right|\left|\widehat{G}(\tau,k,z)\right|\left|\widehat{h}(k)\right|dk \\ &\le \int\left|\prod_{j=1}^{N}\prod_{l=1}^{n(j)}(-ik-l+1)\right||D(\tau,k)|^m\left|\widehat{h}(k)\right|dk := C < \infty. \end{aligned} \tag{5.12}$$

The constant C depends on $\mathcal{M}$ but is independent of (t,x,z). Using similar techniques, a series of tedious but straightforward calculations leads to the following bounds:

$$\begin{aligned} |\mathcal{M}P_1(t,x,z)| &\le C(1+z), \\ \left|\frac{\partial}{\partial t}\mathcal{M}P_0(t,x,z)\right| &\le C(1+z), \\ \left|\frac{\partial}{\partial t}\mathcal{M}P_1(t,x,z)\right| &\le C(1+z^2), \end{aligned}$$

where, in each case, C is some finite constant which depends on $\mathcal{M}$ but is independent of (t,x,z). We are now in a position to bound functions of the form $\mathcal{M}P_2(t,x,y,z)$ and $\mathcal{M}P_3(t,x,y,z)$. From (5.11) we have

$$\begin{aligned}\mathcal{L}_0\left(\mathcal{M}P_2(t,x,y,z)\right) &= \mathcal{M}\frac{1}{z}\left(-\mathcal{L}_2+\langle\mathcal{L}_2\rangle\right)P_0(t,x,z)\\ &= \frac{1}{2}\left(-f^2(y)+\left\langle f^2\right\rangle\right)\mathcal{M}_1P_0(t,x,z)\\ &\quad+\rho_{xz}\sigma\left(-f(y)+\langle f\rangle\right)\mathcal{M}_2P_0(t,x,z)\\ &=: g(t,x,y,z),\end{aligned}$$

where $\mathcal{M}_i$ are of the form (5.10). Now, using the fact that $f(y)$ is bounded and using (5.12), we have

$$|g(t,x,y,z)| \le C,$$

where C is a constant which is independent of (t,x,y,z). Hence, using Lemma 5.1, there exists a constant, C, such that

$$|\mathcal{M}P_2(t,x,y,z)| \le C(1+\log(1+|y|)).$$

Similar, but more involved calculations lead to the following bounds:

$$\begin{aligned}|\mathcal{M}P_3(t,x,y,z)|,\left|\frac{\partial}{\partial t}\mathcal{M}P_2(t,x,y,z)\right| &\le C(1+\log(1+|y|))(1+z), \quad (5.13)\\ \left|\frac{\partial}{\partial y}\mathcal{M}P_2(t,x,y,z)\right| &\le C,\\ \left|\frac{\partial}{\partial y}\frac{\partial}{\partial t}\mathcal{M}P_2(t,x,y,z)\right|,\left|\frac{\partial}{\partial y}\mathcal{M}P_3(t,x,y,z)\right| &\le C(1+z),\\ \left|\frac{\partial}{\partial t}\mathcal{M}P_3(t,x,y,z)\right| &\le C(1+\log(1+|y|))(1+z^2),\\ \left|\frac{\partial}{\partial y}\frac{\partial}{\partial t}\mathcal{M}P_3(t,x,y,z)\right| &\le C(1+z^2). \quad (5.14)\end{aligned}$$

We can now bound $G^\epsilon(x,y,z)$. Using (5.6), we have

$$\begin{aligned}|G^\epsilon(x,y,z)| &\le |P_2(T,x,y,z)|+\sqrt{\epsilon}|P_3(T,x,y,z)|\\ &\le C_1(1+\log(1+|y|))+\sqrt{\epsilon}C_2(1+\log(1+|y|))(1+z)\\ &\le C(1+\log(1+|y|))(1+z). \quad (5.15)\end{aligned}$$

Likewise, using (5.4), we have

$$|F^\epsilon(t,x,y,z)| \le z\left|\mathcal{L}_1P_3(t,x,y,z)\right|+\left|\mathcal{L}_2P_2(t,x,y,z)\right|+\sqrt{\epsilon}\left|\mathcal{L}_2P_3(t,x,y,z)\right|.$$

Each of the above terms can be bounded using (5.13)–(5.14). In particular we find that there exists a constant, C, such that

$$|F^\epsilon(t,x,y,z)| \le C(1+\log(1+|y|))(1+z^2). \quad (5.16)$$

Using (5.9), bounds (5.15) and (5.16), the Cauchy–Schwarz inequality, and moments of the ϵ-independent CIR process Z_t (see, for instance, [15]), one obtains

$$|R^\epsilon(t,x,y,z)| \leq \epsilon\, C(z) \left(1 + \mathbb{E}_{t,y,z}|Y_T| + \int_t^T \mathbb{E}_{t,y,z}|Y_s| ds\right), \tag{5.17}$$

where $\mathbb{E}_{t,y,z}$ denotes the expectation starting at time t from $Y_t = y$ and $Z_t = z$ under the dynamics (2.3)–(2.4). Under these dynamics, starting at time zero from y, we have

$$Y_t = m + (y-m)e^{-\frac{1}{\epsilon}\int_0^t Z_s ds} + \frac{\nu\sqrt{2}}{\sqrt{\epsilon}} e^{-\frac{1}{\epsilon}\int_0^t Z_u du} \int_0^t e^{\frac{1}{\epsilon}\int_0^s Z_u du} \nu \sqrt{Z_s}\, dW_s^y. \tag{5.18}$$

Using the bound established in Appendix C, we have that for any given $\alpha \in (1/2, 1)$ there is a constant C such that

$$\mathbb{E}|Y_t| \leq C\, \epsilon^{\alpha - 1}, \tag{5.19}$$

and the error estimate (5.1) follows.

Numerical illustration for call options. The result of accuracy above is established for smooth and bounded payoffs. The case of call options, important for implied volatilities and calibration described in the following sections, would require regularizing the payoff as was done in [9] in the Black–Scholes case with fast mean-reverting stochastic volatility. Here, in the case of the multiscale Heston model, we simply provide a numerical illustration of the accuracy of approximation. The full model price is computed by Monte Carlo simulation, and the approximated price is given by the formula for the Heston price P_0 given in section 4.1 and our formulas for the correction $\sqrt{\epsilon}\, P_1$ given in section 4.2. Note that the group parameters V_i needed to compute the correction are calculated from the parameters of the full model.

In Table 1, we summarize the results of a Monte Carlo simulation for a European call option. We use a standard Euler scheme, with a time step of 10^{-5} years, which is short enough to ensure that Z_t never becomes negative. We run 10^6 sample paths with $\epsilon = 10^{-3}$ so that $\sqrt{\epsilon} V_3 = 0.0303$ is of the same order as V_3^ϵ, the largest of the V_i^ϵ's obtained in the calibration example presented in section 7. The parameters used in the simulation are

$$\begin{gathered} x = 100, \quad z = 0.24, \quad r = 0.05, \quad \kappa = 1, \quad \theta = 1, \quad \sigma = 0.39, \quad \rho_{xz} = -0.35, \\ y = 0.06, \quad m = 0.06, \quad \nu = 1, \quad \rho_{xy} = -0.35, \quad \rho_{yz} = 0.35, \\ \tau = 1, \quad K = 100, \end{gathered}$$

and $f(y) = e^{y-m-\nu^2}$ so that $\langle f^2 \rangle = 1$. Note that, although f is not bounded, it is a convenient choice because it allows for analytic calculation of the four group parameters V_i given by (4.15)–(4.18).

Note that the error $|P_0 + \sqrt{\epsilon}\, P_1 - \widehat{P}_{MC}|$ is smaller than $\widehat{\sigma}_{MC}$ while the correction $\sqrt{\epsilon}\, P_1$ is statistically significant. This illustrates the accuracy of our approximation for call options in a realistic parameter regime.

Table 1
Results of a Monte Carlo simulation for a European call option.

ϵ	$\sqrt{\epsilon}\,P_1$	$P_0+\sqrt{\epsilon}\,P_1$	$\widehat{P}_{MC}$	$\widehat{\sigma}_{MC}$	$\lvert P_0+\sqrt{\epsilon}\,P_1-\widehat{P}_{MC}\rvert$
0	0	21.0831	-	-	-
10^{-3}	-0.2229	20.8602	20.8595	0.0364	0.0007

6. The multiscale implied volatility surface. In this section, we explore how the implied volatility surface produced by our multiscale model compares to that produced by the Heston model. To begin, we remind the reader that an approximation to the price of a European option in the multiscale model can be obtained through the formula

$$\begin{aligned} P^\epsilon &\sim P_0+\sqrt{\epsilon}P_1 \\ &= P_H+P_1^\epsilon, \\ P_1^\epsilon &:= \sqrt{\epsilon}\,P_1, \end{aligned}$$

where we have absorbed the $\sqrt{\epsilon}$ into the definition of P_1^ϵ and used $P_0=P_H$, the Heston price. Form the formulas for the correction P_1, given in section 4.2, it can be seen that P_1 is linear in V_i, $i=1,\ldots,4$. Therefore, by setting

$$V_i^\epsilon=\sqrt{\epsilon}\,V_i, \qquad i=1,\ldots,4\,,$$

the small correction P_1^ϵ is given by the same formulas as P_1 with the V_i replaced by the V_i^ϵ.

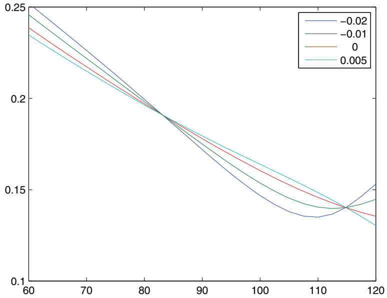
It is important to note that, although adding a fast mean-reverting factor of volatility on top of the Heston model introduces five new parameters (ν, m, ϵ, ρ_{xy}, ρ_{yz}) plus an unknown function f to the dynamics of the stock (see (2.2) and (2.3)), neither knowledge of the values of these five parameters nor the specific form of the function f is required to price options using our approximation. The effect of adding a fast mean-reverting factor of volatility on top of the Heston model is *entirely* captured by the four group parameters V_i^ϵ, which are constants that can be obtained by calibrating the multiscale model to option prices on the market.

By setting $V_i^\epsilon=0$ for $i=1,\ldots,4$, we see that $P_1^\epsilon=0$, $P^\epsilon=P_H$, and the resulting implied volatility surface, obtained by inverting the Black–Scholes formula, corresponds to the implied volatility surface produced by the Heston model. If we then vary a single V_i^ϵ while holding $V_j^\epsilon=0$ for $j\neq i$, we can see exactly how the multiscale implied volatility surface changes as a function of each of the V_i^ϵ. The results of this procedure are plotted in Figure 1.

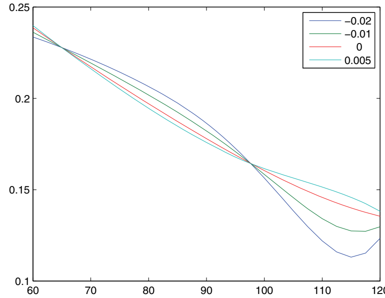
Because they are on the order of $\sqrt{\epsilon}$, typical values of the V_i^ϵ are quite small. However, in order to highlight their effect on the implied volatility surface, the range of values plotted for the V_i^ϵ in Figure 1 was intentionally chosen to be large. It is clear from Figure 1 and from (4.23) that each V_i^ϵ has a distinct effect on the implied volatility surface. Thus, the multiscale model provides considerable flexibility when it comes to calibrating the model to the implied volatility surface produced by options on the market.

7. Calibration. Denote by Θ and Φ the vectors of unobservable parameters in the Heston and multiscale approximation models, respectively:

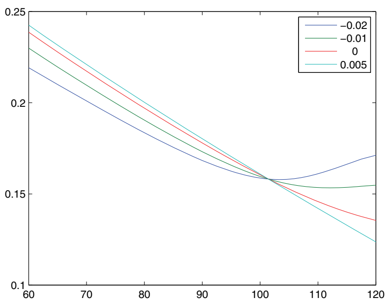
$$\begin{aligned} \Theta &= (\kappa,\rho,\sigma,\theta,z), \\ \Phi &= (\kappa,\rho,\sigma,\theta,z,V_1^\epsilon,V_2^\epsilon,V_3^\epsilon,V_4^\epsilon). \end{aligned}$$



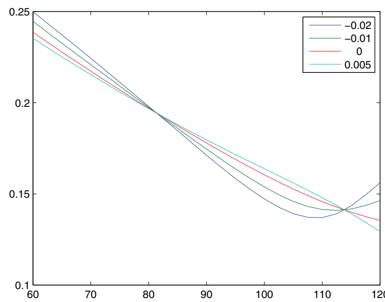
(a)



(b)



(c)



(d)

Figure 1. *Implied volatility curves are plotted as a function of the strike price for European calls in the multiscale model. In this example the initial stock price is $x = 100$. The Heston parameters are set to $z = 0.04$, $\theta = 0.024$, $\kappa = 3.4$, $\sigma = 0.39$, $\rho_{xz} = -0.64$, and $r = 0.0$. In subfigure (a) we vary only V_1^ϵ, fixing $V_i^\epsilon = 0$ for $i \neq 1$. Likewise, in subfigures (b), (c), and (d), we vary only V_2^ϵ, V_3^ϵ, and V_4^ϵ, respectively, fixing all other $V_i^\epsilon = 0$. We remind the reader that, in all four plots, $V_i^\epsilon = 0$ corresponds to the implied volatility curve of the Heston model.*

Let $\sigma(T_i, K_{j(i)})$ be the implied volatility of a call option on the market with maturity date T_i and strike price $K_{j(i)}$. Note that, for each maturity date, T_i, the set of available strikes, $\{K_{j(i)}\}$, varies. Let $\sigma_H(T_i, K_{j(i)}, \Theta)$ be the implied volatility of a call option with maturity date T_i and strike price $K_{j(i)}$ as calculated in the Heston model using parameters Θ. And let $\sigma_M(T_i, K_{j(i)}, \Phi)$ be the implied volatility of the call option with maturity date T_i and strike price $K_{j(i)}$ as calculated in the multiscale approximation using parameters Φ.

We formulate the calibration problem as a constrained, nonlinear, least squares optimization. Define the objective functions as

$$\Delta_H^2(\Theta) = \sum_i \sum_{j(i)} \left(\sigma(T_i, K_{j(i)}) - \sigma_H(T_i, K_{j(i)}, \Theta)\right)^2,$$

$$\Delta_M^2(\Phi) = \sum_i \sum_{j(i)} \left(\sigma(T_i, K_{j(i)}) - \sigma_M(T_i, K_{j(i)}, \Phi)\right)^2.$$

We consider Θ^* and Φ^* to be optimal if they satisfy

$$\Delta_H^2(\Theta^*) = \min_{\Theta} \Delta_H^2(\Theta),$$
$$\Delta_M^2(\Phi^*) = \min_{\Phi} \Delta_M^2(\Phi).$$

It is well known that the objective functions, Δ_H^2 and Δ_M^2, may exhibit a number of local minima. Therefore, if one uses a local gradient method to find Θ^* and Φ^* (as we do in this paper), there is a danger of ending up in a local minima rather than the global minimum. Therefore, it becomes important to make a good initial guess for Θ and Φ, which can be done by visually tuning the Heston parameters to match the implied volatility surface and setting each of the $V_i^\epsilon = 0$. In this paper, we calibrate the Heston model first to find Θ^*. Then, for the multiscale model, we make an initial guess $\Phi = (\Theta^*, 0, 0, 0, 0)$ (i.e., we set the $V_i^\epsilon = 0$ and use Θ^* for the rest of the parameters of Φ). This is a logical calibration procedure because the V_i^ϵ, being of order $\sqrt{\epsilon}$, are intended to be small parameters.

The data we consider consists of call options on the S&P 500 index (SPX) taken from May 17, 2006. We limit our data set to options with maturities greater than 45 days and with open interest greater than 100. We use the yield on the nominal 3-month, constant maturity, U.S. Government treasury bill as the risk-free interest rate. And we use a dividend yield on the SPX taken directly from the Standard & Poor's website (http://www.standardandpoors.com). In Figures 2 through 8, we plot the implied volatilities of call options on the market as well as the calibrated implied volatility curves for the Heston and multiscale models. We would like to emphasize that, although the plots are presented maturity by maturity, they are the result of a single calibration procedure that uses the entire data set.

From Figures 2 through 8, it is apparent to the naked eye that the multiscale model represents a vast improvement over the Heston model, especially for call options with the shortest maturities. In order to quantify this result, we define marginal residual sum of squares

$$\bar{\Delta}_H^2(T_i) = \frac{1}{N(T_i)} \sum_{j(i)} \left(\sigma(T_i, K_{j(i)}) - \sigma_H(T_i, K_{j(i)}, \Theta^*)\right)^2,$$
$$\bar{\Delta}_M^2(T_i) = \frac{1}{N(T_i)} \sum_{j(i)} \left(\sigma(T_i, K_{j(i)}) - \sigma_M(T_i, K_{j(i)}, \Phi^*)\right)^2,$$

where $N(T_i)$ is the number of different calls in the data set that expire at time T_i (i.e., $N(T_i) = \#\{K_{j(i)}\}$). A comparison of $\bar{\Delta}_H^2(T_i)$ and $\bar{\Delta}_M^2(T_i)$ is given in Table 2. The table confirms what is apparent to the naked eye, namely, that the multiscale model fits the market data significantly better than the Heston model for the two shortest maturities as well as the longest maturity.

In general, as explained in [7], the calibrated parameters are sufficient to compute approximated prices of exotic options. The leading order price P_0 is obtained by solving (eventually numerically) the homogenized PDE appropriate for a given exotic option (for instance with an additional boundary condition in the case of a barrier option). The correction P_1^ϵ is obtained as the solution of the PDE with source where the source can be computed with P_0 and the V_i's calibrated on European options.

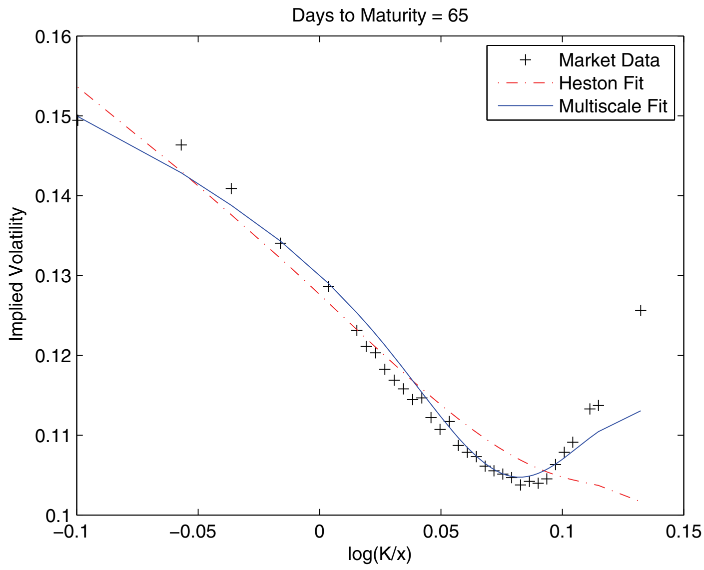


Figure 2. *SPX implied volatilities from May* 17, 2006.

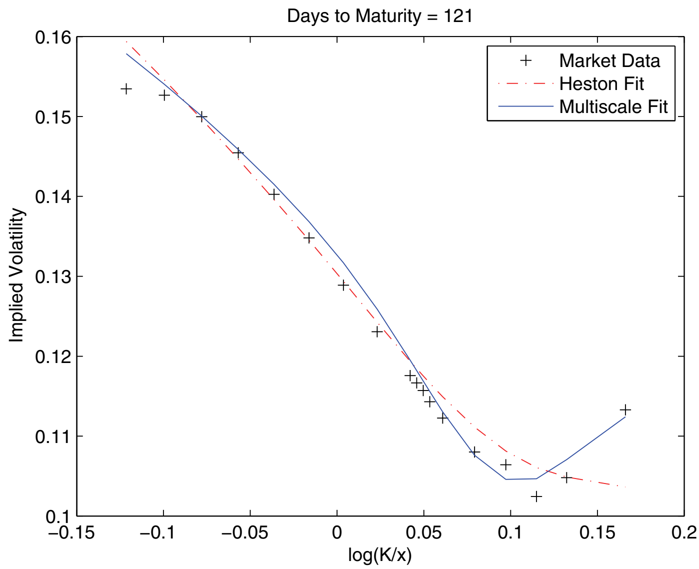


Figure 3. *SPX implied volatilities from May* 17, 2006.

Finally, we remark that V_3^ϵ, the largest calibrated V_i^ϵ, was found to be 0.025. In the Monte Carlo simulation presented at the end of section 5, we chose ϵ so that the value of $\sqrt{\epsilon}V_3$ was of the same order of magnitude as the calibrated V_3^ϵ.

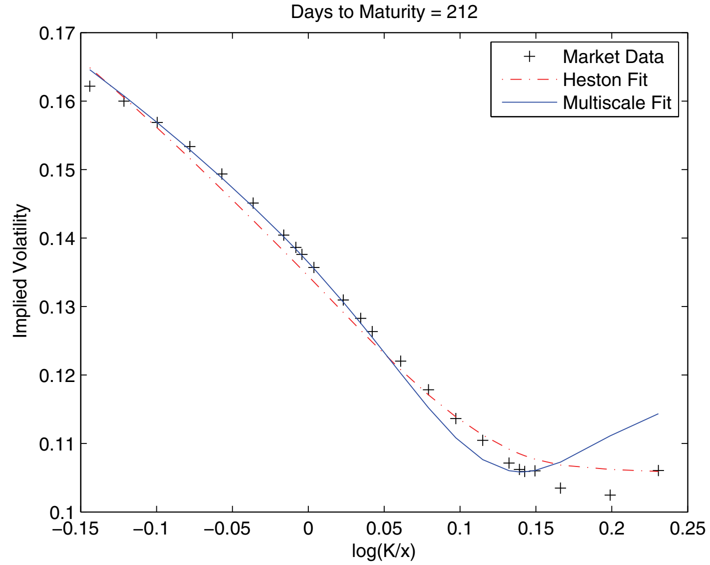


Figure 4. *SPX implied volatilities from May* 17, 2006*.*

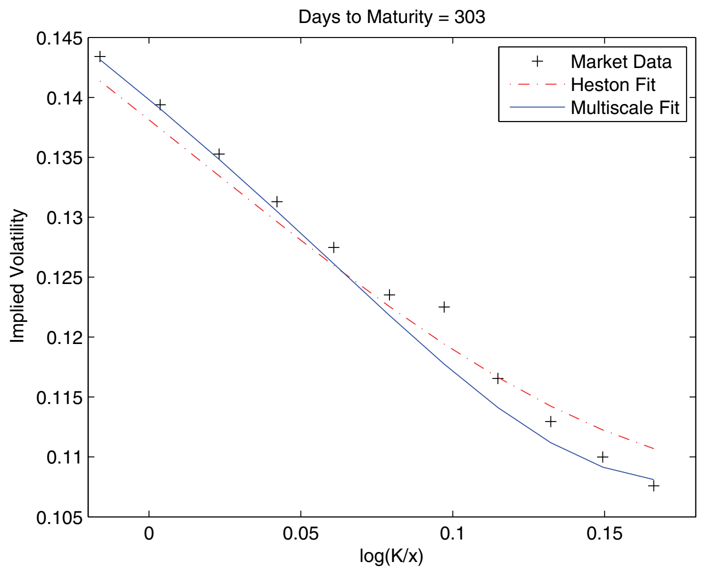


Figure 5. *SPX implied volatilities from May* 17, 2006*.*

Appendix A. Heston stochastic volatility model. There are a number of excellent resources where one can read about the Heston stochastic volatility model—so many, in fact, that a detailed review of the model would seem superfluous. However, in order to establish some notation, we will briefly review the dynamics of the Heston model here as well as show our preferred method for solving the corresponding European option pricing problem. The

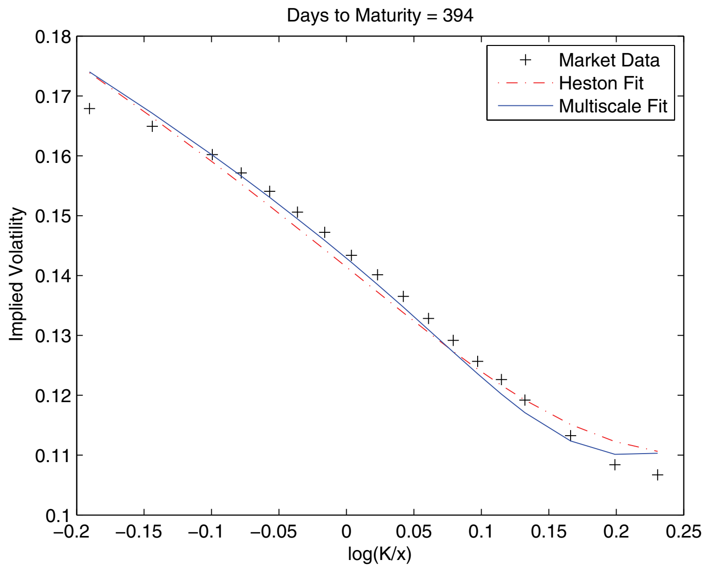


Figure 6. *SPX implied volatilities from May* 17*,* 2006*.*

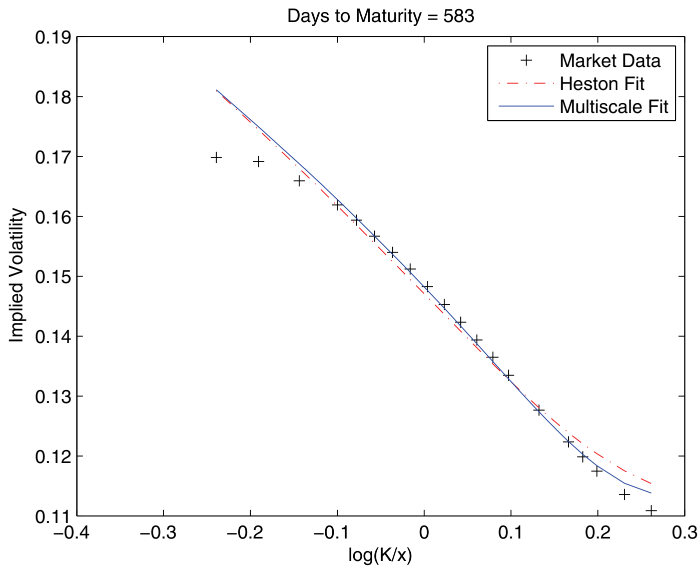


Figure 7. *SPX implied volatilities from May* 17*,* 2006*.*

notes from this section closely follow [20]. The reader should be aware that a number of the equations developed in this section are referred to throughout the main text of this paper.

Let X_t be the price of a stock. And denote by r the risk-free rate of interest. Then, under

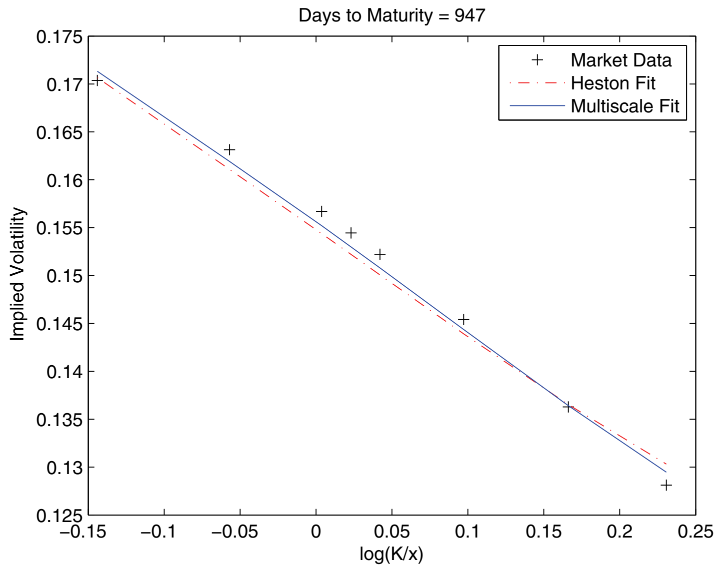


Figure 8. *SPX implied volatilities from May 17, 2006.*

Table 2
Residual sum of squares for the Heston and the multiscale models at several maturities.

$T_i - t$ (days)	$\bar{\Delta}_H^2(T_i)$	$\bar{\Delta}_M^2(T_i)$	$\bar{\Delta}_H^2(T_i)/\bar{\Delta}_M^2(T_i)$
65	29.3×10^{-6}	7.91×10^{-6}	3.71
121	10.2×10^{-6}	3.72×10^{-6}	2.73
212	4.06×10^{-6}	8.11×10^{-6}	0.51
303	3.93×10^{-6}	3.51×10^{-6}	1.12
394	7.34×10^{-6}	5.17×10^{-6}	1.42
583	11.3×10^{-6}	9.28×10^{-6}	1.22
947	3.31×10^{-6}	1.47×10^{-6}	2.25

the risk-neutral probability measure, $\mathbb{P}$, the Heston model takes the following form:

$$\begin{aligned} dX_t &= rX_t dt + \sqrt{Z_t}\, X_t dW_t^x\,, \\ dZ_t &= \kappa\,(\theta - Z_t)\,dt + \sigma\sqrt{Z_t}\,dW_t^z\,, \\ d\,\langle W^x, W^z\rangle_t &= \rho dt. \end{aligned}$$

Here, W_t^x and W_t^z are one-dimensional Brownian motions with correlation ρ such that $|\rho| \le 1$. The process, Z_t, is the stochastic variance of the stock. And κ, θ, and σ are positive constants satisfying $2\kappa\theta \ge \sigma^2$; assuming $Z_0 > 0$, this ensures that Z_t remains positive for all t.

We denote by P_H the price of a European option, as calculated under the Heston framework. As we are already under the risk-neutral measure, we can express P_H as an expectation of the option payoff, $h(X_T)$, discounted at the risk-free rate:

$$P_H(t,x,z) = \mathbb{E}\left[e^{-r(T-t)} h(X_T) \middle| X_t = x, Z_t = z \right].$$

Using the Feynman–Kac formula, we find that $P_H(t,x,z)$ must satisfy the following PDE and boundary condition:

$$\mathcal{L}_H P_H(t,x,z) = 0, \tag{A.1}$$

$$P_H(T,x,z) = h(x), \tag{A.2}$$

$$\begin{aligned}\mathcal{L}_H = {} & \frac{\partial}{\partial t} - r + rx\frac{\partial}{\partial x} + \frac{1}{2}zx^2\frac{\partial^2}{\partial x^2} \\ & + \kappa(\theta - z)\frac{\partial}{\partial z} + \frac{1}{2}\sigma^2 z\frac{\partial^2}{\partial z^2} \\ & + \rho\sigma z x\frac{\partial^2}{\partial x \partial z}.\end{aligned} \tag{A.3}$$

In order to find a solution for $P_H(t,x,z)$, it will be convenient to transform variables as follows:

$$\begin{aligned}\tau(t) &= T - t, \\ q(t,x) &= r(T-t) + \log x, \\ P_H(t,x,z) &= P_H'(\tau(t), q(t,x), z)e^{-r\tau(t)}.\end{aligned}$$

This transformation leads us to the following PDE and boundary condition for $P_H'(\tau,q,z)$:

$$\begin{aligned}\mathcal{L}_H' P_H'(\tau,q,z) &= 0, \\ \mathcal{L}_H' &= -\frac{\partial}{\partial \tau} + \frac{1}{2}z\left(\frac{\partial^2}{\partial q^2} - \frac{\partial}{\partial q}\right) + \rho\sigma z\frac{\partial^2}{\partial q\partial z} \\ & \quad + \frac{1}{2}\sigma^2 z\frac{\partial^2}{\partial z^2} + \kappa(\theta - z)\frac{\partial}{\partial z}, \\ P_H'(0,q,z) &= h(e^q).\end{aligned} \tag{A.4}$$

We will find a solution for P_H' through the method of Green's functions. Denote by $\delta(q)$ the Dirac delta function, and let $G(\tau,q,z)$, the Green's function, be the solution to the following Cauchy problem:

$$\mathcal{L}_H' G(\tau,q,z) = 0, \tag{A.5}$$

$$G(0,q,z) = \delta(q). \tag{A.6}$$

Then,

$$P_H'(\tau,q,z) = \int_{\mathbb{R}} G(\tau, q-p, z)h(e^p)dp.$$

Now, let $\widehat{P}_H(\tau,k,z)$, $\widehat{G}(\tau,k,z)$, and $\widehat{h}(k)$ be the Fourier transforms of $P_H'(\tau,q,z)$, $G(\tau,q,z)$, and $h(e^q)$, respectively:

$$\begin{aligned}\widehat{P}_H(\tau,k,z) &= \int_{\mathbb{R}} e^{ikq}P_H'(\tau,q,z)dq, \\ \widehat{G}(\tau,k,z) &= \int_{\mathbb{R}} e^{ikq}G(\tau,q,z)dq, \\ \widehat{h}(k) &= \int_{\mathbb{R}} e^{ikq}h(e^q)dq.\end{aligned}$$

Then, using the convolution property of Fourier transforms, we have

$$
\begin{aligned}
P'_H(\tau, q, z) &= \frac{1}{2\pi}\int_{\mathbb{R}} e^{-ikq}\widehat{P}_H(\tau, k, z)dk \\
&= \frac{1}{2\pi}\int_{\mathbb{R}} e^{-ikq}\widehat{G}(\tau, k, z)\widehat{h}(k)dk.
\end{aligned}
$$

Multiplying (A.5) and (A.6) by $e^{ikq'}$ and integrating over $\mathbb{R}$ in q', we find that $\widehat{G}(\tau, k, z)$ satisfies the following Cauchy problem:

$$
\widehat{\mathcal{L}}_H\widehat{G}(\tau, k, z) = 0, \tag{A.7}
$$

$$
\begin{aligned}
\widehat{\mathcal{L}}_H = &-\frac{\partial}{\partial\tau} + \frac{1}{2}z\left(-k^2 + ik\right) + \frac{1}{2}\sigma^2 z\frac{\partial^2}{\partial z^2} \\
&+ \left(\kappa\theta - (\kappa + \rho\sigma ik)\, z\right)\frac{\partial}{\partial z},
\end{aligned}
$$

$$
\widehat{G}(0, k, z) = 1. \tag{A.8}
$$

Now, we give an ansatz: suppose $\widehat{G}(\tau, k, z)$ can be written as follows:

$$
\widehat{G}(\tau, k, z) = e^{C(\tau,k)+zD(\tau,k)}. \tag{A.9}
$$

Substituting (A.9) into (A.7) and (A.8) and collecting terms of like powers of z, we find that $C(\tau, k)$ and $D(\tau, k)$ must satisfy the following ordinary differential equations (ODEs):

$$
\frac{dC}{d\tau}(\tau, k) = \kappa\theta D(\tau, k), \tag{A.10}
$$

$$
C(0, k) = 0, \tag{A.11}
$$

$$
\frac{dD}{d\tau}(\tau, k) = \frac{1}{2}\sigma^2 D^2(\tau, k) - (\kappa + \rho\sigma ik)\, D(\tau, k) + \frac{1}{2}\left(-k^2 + ik\right), \tag{A.12}
$$

$$
D(0, k) = 0. \tag{A.13}
$$

Equations (A.10), (A.11), (A.12), and (A.13) can be solved analytically. Their solutions, as well as the final solution to the European option pricing problem in the Heston framework, are given in (4.3)–(4.11).

Appendix B. Detailed solution for $P_1(t, x, z)$. In this section, we show how to solve for $P_1(t, x, z)$, which is the solution to the Cauchy problem defined by (3.16) and (3.17). For convenience, we repeat these equations here with the notation $\mathcal{L}_H = \langle\mathcal{L}_2\rangle$ and $P_H = P_0$:

$$
\mathcal{L}_H P_1(t, x, z) = \mathcal{A}P_H(t, x, z), \tag{B.1}
$$

$$
P_1(T, x, z) = 0. \tag{B.2}
$$

We remind the reader that $\mathcal{A}$ is given by (4.14), $\mathcal{L}_H$ is given by (4.1), and $P_H(t, x, z)$ is given by (4.3). It will be convenient in our analysis to make the following variable transformation:

$$
\begin{aligned}
P_1(t, x, z) &= P'_1(\tau(t), q(t, x), z)e^{-r\tau}, \\
\tau(t) &= T - t, \\
q(t, x) &= r(T - t) + \log x.
\end{aligned} \tag{B.3}
$$

We now substitute (4.3), (4.14), and (B.3) into (B.1) and (B.2), which leads us to the following PDE and boundary condition for $P_1'(\tau, q, z)$:

$$\mathcal{L}_H' P_1'(\tau, q, z) = \mathcal{A}' \frac{1}{2\pi} \int e^{-ikq} \widehat{G}(\tau, k, z) \widehat{h}(k) dk, \tag{B.4}$$

$$\mathcal{L}_H' = -\frac{\partial}{\partial \tau} + \frac{1}{2} z \left(\frac{\partial^2}{\partial q^2} - \frac{\partial}{\partial q} \right) + \rho \sigma z \frac{\partial^2}{\partial q \partial z} + \frac{1}{2} \sigma^2 z \frac{\partial^2}{\partial z^2} + \kappa (\theta - z),$$

$$\mathcal{A}' = V_1 z \frac{\partial}{\partial z} \left(\frac{\partial^2}{\partial q^2} - \frac{\partial}{\partial q} \right) + V_2 z \frac{\partial^3}{\partial z^2 \partial q} + V_3 z \left(\frac{\partial^3}{\partial q^3} - \frac{\partial^2}{\partial q^2} \right) + V_4 z \frac{\partial^3}{\partial z \partial q^2},$$

$$P_1'(0, q, z) = 0. \tag{B.5}$$

Now, let $\widehat{P}_1(\tau, k, z)$ be the Fourier transform of $P_1'(\tau, q, z)$:

$$\widehat{P}_1(\tau, k, z) = \int_{\mathbb{R}} e^{ikq} P_1'(\tau, q, z) dq.$$

Then,

$$P_1'(\tau, q, z) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-ikq} \widehat{P}_1(\tau, k, z) dk. \tag{B.6}$$

Multiplying (B.4) and (B.5) by $e^{ikq'}$ and integrating in q' over $\mathbb{R}$, we find that $\widehat{P}_1(\tau, k, z)$ satisfies the following Cauchy problem:

$$\widehat{\mathcal{L}}_H \widehat{P}_1(\tau, k, z) = \widehat{\mathcal{A}} \widehat{G}(\tau, k, z) \widehat{h}(k), \tag{B.7}$$

$$\widehat{\mathcal{L}}_H = -\frac{\partial}{\partial \tau} + \frac{1}{2} z \left(-k^2 + ik \right) + \frac{1}{2} \sigma^2 z \frac{\partial^2}{\partial z^2} + \left(\kappa \theta - \left(\kappa + \rho \sigma i k \right) z \right) \frac{\partial}{\partial z},$$

$$\widehat{\mathcal{A}} = z \left(V_1 \frac{\partial}{\partial z} \left(-k^2 + ik \right) + V_2 \frac{\partial^2}{\partial z^2} \left(-ik \right) + V_3 \left(ik^3 + k^2 \right) + V_4 \frac{\partial}{\partial z} \left(-k^2 \right) \right),$$

$$\widehat{P}_1(0, k, z) = 0. \tag{B.8}$$

Now, we give an ansatz: we suppose that $\widehat{P}_1(\tau, k, z)$ can be written as

$$\widehat{P}_1(\tau, k, z) = \left(\kappa \theta \widehat{f}_0(\tau, k) + z \widehat{f}_1(\tau, k) \right) \widehat{G}(\tau, k, z) \widehat{h}(k). \tag{B.9}$$

We substitute (B.9) into (B.7) and (B.8). After a good deal of algebra (and, in particular, making use of (A.10) and (A.12)), we find that $\widehat{f}_0(\tau, k)$ and $\widehat{f}_1(\tau, k)$ satisfy the following system of ODEs:

$$\frac{d\widehat{f}_1}{d\tau}(\tau,k) = a(\tau,k)\widehat{f}_1(\tau,k) + b(\tau,k), \tag{B.10}$$

$$\widehat{f}_1(0,k) = 0, \tag{B.11}$$

$$\frac{d\widehat{f}_0}{d\tau}(\tau,k) = \widehat{f}_1(\tau,k), \tag{B.12}$$

$$\widehat{f}_0(0,k) = 0, \tag{B.13}$$

$$\begin{aligned} a(\tau,k) &= \sigma^2 D(\tau,k) - (\kappa + \rho\sigma i k)\,, \\ b(\tau,k) &= -\left(V_1 D(\tau,k)\left(-k^2 + ik\right) + V_2 D^2(\tau,k)\left(-ik\right)\right. \\ &\qquad \left. + V_3\left(ik^3 + k^2\right) + V_4 D(\tau,k)\left(-k^2\right)\right), \end{aligned}$$

where $D(\tau,k)$ is given by (4.9).

Equations (B.10)–(B.13) can be solved analytically (to the extent that their solutions can be written down in integral form). The solutions for $\widehat{f}_0(\tau,k)$ and $\widehat{f}_1(\tau,k)$, along with the final solution for $P_1(t,x,z)$, are given by (4.19)–(4.23).

Appendix C. Moment estimate for Y_t. In this section we will derive a moment estimate for Y_t, whose dynamics under the pricing measure are given by (2.3), (2.4), (2.7). Specifically, we will show that for all $\alpha \in (1/2, 1)$ there exists a constant, C (which depends on α but is independent of ϵ), such that $\mathbb{E}|Y_t| \le C\,\epsilon^{\alpha-1}$.

We will begin by establishing some notation. First we define a continuous, strictly increasing, nonnegative process, β_t, as

$$\beta_t := \int_0^t Z_s ds.$$

Next, we note that W_t^y may be decomposed as

$$W_t^y = \rho_{yz} W_t^z + \sqrt{1-\rho_{yz}^2}\, W_t^{\perp}, \tag{C.1}$$

where $W_t^{\perp}$ is a Brownian motion which is independent of W_t^z. Using (5.18) and (C.1), we derive

$$|Y_t| \le C_1 + \frac{C_2}{\sqrt{\epsilon}}\left[e^{\frac{-1}{\epsilon}\beta_t}\left|\int_0^t e^{\frac{1}{\epsilon}\beta_s}\sqrt{Z_s}dW_s^z\right| + e^{\frac{-1}{\epsilon}\beta_t}\left|\int_0^t e^{\frac{1}{\epsilon}\beta_s}\sqrt{Z_s}dW_s^{\perp}\right|\right],$$

where C_1 and C_2 are constants. We will focus on bounding the first moment of the second stochastic integral. We have

$$\begin{aligned}
\frac{1}{\epsilon}\mathbb{E}\left[\left(e^{\frac{-1}{\epsilon}\beta_t}\int_0^t e^{\frac{1}{\epsilon}\beta_s}\sqrt{Z_s}dW_s^{\perp}\right)^2\right] &= \frac{1}{\epsilon}\mathbb{E}\left[e^{-2\beta_t/\epsilon}\mathbb{E}\left[\left(\int_0^t e^{\beta_s/\epsilon}\sqrt{Z_s}dW_s^{\perp}\right)^2 \mid \beta_t\right]\right] \\
&= \frac{1}{\epsilon}\mathbb{E}\left[e^{-2\beta_t/\epsilon}\mathbb{E}\left[\int_0^t e^{2\beta_s/\epsilon}Z_s ds \mid \beta_t\right]\right] \\
&= \frac{1}{\epsilon}\mathbb{E}\left[e^{-2\beta_t/\epsilon}\mathbb{E}\left[\int_0^t e^{2\beta_s/\epsilon}d\beta_s \mid \beta_t\right]\right] \\
&= \frac{1}{\epsilon}\mathbb{E}\left[e^{-2\beta_t/\epsilon}\frac{\epsilon}{2}\left(e^{2\beta_t/\epsilon}-1\right)\right] \\
&= \frac{1}{2}\mathbb{E}\left[1-e^{-\frac{2}{\epsilon}\beta_t}\right] \le \frac{1}{2}.
\end{aligned}$$

Then, by the Cauchy–Schwarz inequality, we see that

$$\frac{1}{\sqrt{\epsilon}}\mathbb{E}\left[e^{\frac{-1}{\epsilon}\beta_t}\left|\int_0^t e^{\frac{1}{\epsilon}\beta_s}\sqrt{Z_s}dW_s^{\perp}\right|\right] \le \frac{1}{\sqrt{2}}.$$

What remains is to bound the first moment of the other stochastic integral,

$$A := \frac{1}{\sqrt{\epsilon}}\mathbb{E}\left[e^{\frac{-1}{\epsilon}\beta_t}\left|\int_0^t e^{\frac{1}{\epsilon}\beta_s}\sqrt{Z_s}dW_s^{z}\right|\right].$$

Naively, one might try to use the Cauchy–Schwarz inequality in the following manner:

$$\begin{aligned}
A &\le \frac{1}{\sqrt{\epsilon}}\sqrt{\mathbb{E}\left[e^{-2\beta_t/\epsilon}\right]}\sqrt{\mathbb{E}\left[\int_0^t e^{2\beta_s/\epsilon}Z_s ds\right]} \\
&= \frac{1}{\sqrt{\epsilon}}\sqrt{\mathbb{E}\left[e^{-2\beta_t/\epsilon}\right]}\sqrt{\mathbb{E}\left[\frac{\epsilon}{2}\left(e^{2\beta_t/\epsilon}\right)\right]}.
\end{aligned}$$

However, this approach does not work, since $\mathbb{E}\left[e^{2\beta_t/\epsilon}\right] \to \infty$ as $\epsilon \to 0$. Seeking a more refined approach of bounding A, we note that

$$\begin{aligned}
\frac{1}{\sqrt{\epsilon}}e^{\frac{-1}{\epsilon}\beta_t}\int_0^t e^{\frac{1}{\epsilon}\beta_s}\sqrt{Z_s}dW_s^{z} &= \frac{1}{\sigma\sqrt{\epsilon}}e^{-\frac{1}{\epsilon}\beta_t}(Z_t - z) - \frac{\kappa}{\sigma\sqrt{\epsilon}}e^{-\frac{1}{\epsilon}\beta_t}\int_0^t e^{\frac{1}{\epsilon}\beta_s}(\theta - Z_s)ds \\
&\quad + \frac{1}{\sigma\epsilon^{3/2}}e^{-\frac{1}{\epsilon}\beta_t}\int_0^t e^{\frac{1}{\epsilon}\beta_s}Z_s(Z_t - Z_s)ds,
\end{aligned}$$

which can be derived by replacing t by s in (2.4), multiplying by $e^{\beta_s/\epsilon}$, integrating the result from 0 to t, and using $Z_s^2 = Z_tZ_s - Z_s(Z_t - Z_s)$ and $\int_0^t e^{-(\beta_t-\beta_s)/\epsilon}Z_s ds = \epsilon(1-e^{-\beta_t/\epsilon})$. From the equation above, we see that

$$\begin{aligned}
A &\le \frac{1}{\sigma\sqrt{\epsilon}}\mathbb{E}\left[e^{-\frac{1}{\epsilon}\beta_t}|Z_t - z|\right] + \frac{\kappa}{\sigma\sqrt{\epsilon}}\mathbb{E}\left[e^{-\frac{1}{\epsilon}\beta_t}\left|\int_0^t e^{\frac{1}{\epsilon}\beta_s}(\theta - Z_s)ds\right|\right] \\
&\quad + \frac{1}{\sigma\epsilon^{3/2}}\mathbb{E}\left[e^{-\frac{1}{\epsilon}\beta_t}\left|\int_0^t e^{\frac{1}{\epsilon}\beta_s}Z_s(Z_t - Z_s)ds\right|\right]. \qquad \text{(C.2)}
\end{aligned}$$

At this point, we need the moment generating function of (Z_t, β_t). From [15], we have

$$\text{(C.3)} \qquad \begin{aligned} \mathbb{E}\left[e^{-\lambda Z_t - \mu\beta_t}\right] &= e^{-\kappa\theta\phi_{\lambda,\mu}(t) - z\psi_{\lambda,\mu}(t)}, \\ \phi_{\lambda,\mu}(t) &= \frac{-2}{\sigma^2}\log\left[\frac{2\gamma e^{(\gamma+\kappa)t/2}}{\lambda\sigma^2\left(e^{\gamma t}-1\right)+\gamma-\kappa+e^{\gamma t}(\gamma+\kappa)}\right], \\ \psi_{\lambda,\mu}(t) &= \frac{\lambda\left(\gamma+\kappa+e^{\gamma t}(\gamma-\kappa)\right)+2\mu\left(e^{\gamma t}-1\right)}{\lambda\sigma^2\left(e^{\gamma t}-1\right)+\gamma-\kappa+e^{\gamma t}(\gamma+\kappa)}, \\ \gamma &= \sqrt{\kappa^2+2\sigma^2\mu}. \end{aligned}$$

Now, let us focus on the first term in (C.2). Using the Cauchy–Schwarz inequality, we have

$$\frac{1}{\sigma\sqrt{\epsilon}}\mathbb{E}\left[e^{-\beta_t/\epsilon}|Z_t - z|\right] \le \frac{1}{\sigma\sqrt{\epsilon}}\sqrt{\mathbb{E}\left[e^{-2\beta_t/\epsilon}\right]}\sqrt{\mathbb{E}\left[|Z_t-z|^2\right]}.$$

From (C.3) one can verify

$$\begin{aligned} \mathbb{E}\left[|Z_t - z|^2\right] &\le C_3, \\ \mathbb{E}\left[e^{-2\beta_t/\epsilon}\right] &= e^{-\kappa\theta\phi_{0,2/\epsilon}(t) - z\phi_{0,2/\epsilon}(t)} \sim e^{C_4/\sqrt{\epsilon}}, \end{aligned}$$

where C_3 and $C_4 < 0$ are constants. Since $\frac{1}{\sqrt{\epsilon}}e^{C_4/\sqrt{\epsilon}} \to 0$ as $\epsilon \to 0$, we see that

$$\frac{1}{\sigma\sqrt{\epsilon}}\mathbb{E}\left[e^{-\beta_t/\epsilon}|Z_t - z|\right] \le C_5$$

for some constant C_5.

We now turn our attention to the second term in (C.2). We have

$$\begin{aligned} &\frac{\kappa}{\sigma\sqrt{\epsilon}}\mathbb{E}\left[\left|\int_0^t e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}(\theta - Z_s)ds\right|\right] \\ &\le \frac{\kappa}{\sigma\sqrt{\epsilon}}\mathbb{E}\left[\int_0^t e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}Z_s ds\right] + \frac{\kappa\theta}{\sigma\sqrt{\epsilon}}\mathbb{E}\left[\int_0^t e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}ds\right] \\ &\le \frac{\kappa}{\sigma\sqrt{\epsilon}}\mathbb{E}\left[\int_0^t e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}d\beta_s\right] + \frac{\kappa\theta}{\sigma\sqrt{\epsilon}}\mathbb{E}\left[\int_0^t e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}ds\right] \\ &\le \frac{\kappa}{\sigma\sqrt{\epsilon}}\mathbb{E}\left[\epsilon\left(1-e^{-\beta_t/\epsilon}\right)\right] + \frac{\kappa\theta}{\sigma\sqrt{\epsilon}}\mathbb{E}\left[\int_0^t e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}ds\right] \\ &\le C_6 + \frac{\kappa\theta}{\sigma\sqrt{\epsilon}}\int_0^t \mathbb{E}\left[e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}\right]ds \end{aligned}$$

for some constant C_6. To bound the remaining integral we calculate

$$\text{(C.4)} \qquad \begin{aligned} \mathbb{E}\left[e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}\middle| Z_s\right]\right] \\ &= \mathbb{E}\left[e^{-\kappa\theta\phi_{0,1/\epsilon}(t-s) - Z_s\psi_{0,1/\epsilon}(t-s)}\right] \\ &= \exp\Big(-\kappa\theta\phi_{0,1/\epsilon}(t-s) - \kappa\theta\phi_{\bar{\psi},0}(s) - z\psi_{\bar{\psi},0}(s)\Big), \\ \bar{\psi}(s) &:= \psi_{0,1/\epsilon}(t-s). \end{aligned}$$

Using the fact that $\phi_{\lambda,\mu}(t), \psi_{\lambda,\mu}(t) > 0$ for any $\lambda, \mu, t > 0$, we see that

$$\mathbb{E}\left[e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}\right] \leq \exp\Big(- z\psi_{\bar{\psi},0}(s)\Big).$$

Hence

$$\begin{aligned}\int_0^t \mathbb{E}\left[e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}\right] ds &\leq \int_0^{t-\epsilon^\alpha} \exp\Big(- z\psi_{\bar{\psi},0}(s)\Big) ds + \int_{t-\epsilon^\alpha}^t \exp\Big(- z\psi_{\bar{\psi},0}(s)\Big) ds \\ &=: I_1 + I_2,\end{aligned} \tag{C.5}$$

where $\alpha \in (1/2, 1)$. Using $\psi_{\lambda,\mu}(t) > 0$ again, we deduce $\psi_{\bar{\psi},0}(s) > 0$ and therefore

$$I_2 \leq \epsilon^\alpha. \tag{C.6}$$

As for I_1, we claim

$$I_1 \leq C_7 \exp\left(-C_8 \epsilon^{\alpha-1}\right), \tag{C.7}$$

which is equivalent to showing there exists a constant C such that

$$\psi_{\bar{\psi},0}(s) \geq C\epsilon^{\alpha-1} \tag{C.8}$$

for all $s \in [0, t-\epsilon^\alpha]$. To prove this claim, we note that for small ϵ

$$\bar{\psi}(s) = \psi_{0,1/\epsilon}(t-s) \sim \frac{\sigma\sqrt{2}}{\sqrt{\epsilon}} \left(\frac{\exp\left[\frac{\sigma\sqrt{2}}{\sqrt{\epsilon}}(t-s)\right] - 1}{\exp\left[\frac{\sigma\sqrt{2}}{\sqrt{\epsilon}}(t-s)\right] + 1} \right),$$

where we have used $\gamma = \sqrt{\kappa^2 + 2\sigma^2/\epsilon} \sim \sigma\sqrt{2}/\sqrt{\epsilon}$. A direct computation shows that $\bar{\psi}(s)$ is strictly decreasing in s with

$$\bar{\psi}(t-\epsilon^\alpha) = \psi_{0,1/\epsilon}(\epsilon^\alpha) \sim \sigma^2\epsilon^{\alpha-1}.$$

Now, we note that $\psi_{\bar{\psi},0}(s)$ is given by

$$\psi_{\bar{\psi},0}(s) = \frac{2\kappa\bar{\psi}(s)}{\sigma^2\left(e^{\kappa s}-1\right)\bar{\psi}(s) + 2\kappa e^{\kappa s}} = \frac{2\kappa}{\sigma^2\left(e^{\kappa s}-1\right) + 2\kappa e^{\kappa s}/\bar{\psi}(s)}.$$

Since $e^{\kappa s} < e^{\kappa t}$, and since, at worst, $\bar{\psi}(s) \sim \sigma^2\epsilon^{\alpha-1}$, we conclude that there exists a constant C such that (C.8), and therefore (C.7), holds. Hence, using (C.5)–(C.7), we have

$$\begin{aligned}\int_0^t \mathbb{E}\left[e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}\right] ds &= \int_0^{t-\epsilon^\alpha} \mathbb{E}\left[e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}\right] ds + \int_{t-\epsilon^\alpha}^t \mathbb{E}\left[e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}\right] ds \\ &\leq C_7 e^{-C_8\epsilon^{\alpha-1}} + \epsilon^\alpha.\end{aligned}$$

This implies that there exists a constant C_9 such that for any $\alpha \in (1/2, 1)$

$$\frac{\kappa\theta}{\sigma\sqrt{\epsilon}} \int_0^t \mathbb{E}\left[e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}\right] ds \leq C_9.$$

Having established a uniform bound on the first two terms in (C.2), we turn our attention toward the third and final term. For $\alpha \in (1/2, 1)$ we have

$$\begin{aligned}\frac{1}{\sigma\epsilon^{3/2}}\mathbb{E}\left[\left|\int_0^t e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}Z_s(Z_t-Z_s)ds\right|\right] &\leq \frac{1}{\sigma\epsilon^{3/2}}\mathbb{E}\left[\int_0^t e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}Z_s|Z_t-Z_s|ds\right]\\ &= \frac{1}{\sigma\epsilon^{3/2}}\mathbb{E}\left[\int_0^{t-\epsilon^\alpha} e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}Z_s|Z_t-Z_s|ds\right]\\ &\quad+\frac{1}{\sigma\epsilon^{3/2}}\mathbb{E}\left[\int_{t-\epsilon^\alpha}^t e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}Z_s|Z_t-Z_s|ds\right].\end{aligned}$$

For the integral from 0 to $(t-\epsilon^\alpha)$ we compute

$$\begin{aligned}&\frac{1}{\sigma\epsilon^{3/2}}\mathbb{E}\left[\int_0^{t-\epsilon^\alpha} e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}Z_s|Z_t-Z_s|ds\right]\\ &\leq \frac{1}{\sigma\epsilon^{3/2}}\sqrt{\mathbb{E}\left[\left(\sup_{0\leq s\leq t} Z_s|Z_t-Z_s|\right)^2\right]}\sqrt{\mathbb{E}\left[\int_0^{t-\epsilon^\alpha} e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}ds\right]}\\ &\leq \frac{1}{\epsilon^{3/2}}C_{10}e^{-C_{11}\epsilon^{\alpha-1}} \leq C_{12}\end{aligned}$$

for some constants C_{10}, C_{11}, and C_{12}. For the integral from $(t-\epsilon^\alpha)$ to t we have

$$\begin{aligned}&\frac{1}{\sigma\epsilon^{3/2}}\mathbb{E}\left[\int_{t-\epsilon^\alpha}^t e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}Z_s|Z_t-Z_s|ds\right]\\ &\leq \frac{1}{\sigma\epsilon^{3/2}}\mathbb{E}\left[\sup_{t-\epsilon^\alpha\leq s\leq t}|Z_t-Z_s|\int_{t-\epsilon^\alpha}^t e^{-\frac{1}{\epsilon}(\beta_t-\beta_s)}Z_s ds\right]\\ &= \frac{1}{\sigma\epsilon^{3/2}}\mathbb{E}\left[\sup_{t-\epsilon^\alpha\leq s\leq t}|Z_t-Z_s|\,\epsilon\left(1-e^{-\beta_{(t-\epsilon^\alpha)/\epsilon}}\right)\right]\\ &\leq \frac{1}{\sigma\epsilon^{1/2}}\mathbb{E}\left[\sup_{t-\epsilon^\alpha\leq s\leq t}|Z_t-Z_s|\right] \leq C_{13}\epsilon^{\alpha-1}\end{aligned}$$

for some constant C_{13}. With this result, we have established that for all $\alpha \in (1/2, 1)$ there exists a constant, C, such that $\mathbb{E}|Y_t| \leq C\,\epsilon^{\alpha-1}$.

Appendix D. Numerical computation of option prices. The formulas (4.3) and (4.19) for $P_H(t,x,z)$ and $P_1(t,x,z)$ cannot be evaluated analytically. Therefore, in order for these formulas to be useful, an efficient and reliable numerical integration scheme is needed. Unfortunately, numerical evaluation of the integral in (4.3) is notoriously difficult. And the double and triple integrals that appear in (4.19) are no easier to compute. In this section, we point out some of the difficulties associated with evaluating these expressions numerically and show

how these difficulties can be addressed. We begin by establishing some notation:

$$\begin{aligned}
P^\epsilon(t,x,z) &\sim P_H(t,x,z) + \sqrt{\epsilon} P_1(t,x,z) \\
&= \frac{e^{-r\tau}}{2\pi} \int_{\mathbb{R}} e^{-ikq} \left(1 + \sqrt{\epsilon}\left(\kappa\theta \widehat{f}_0(\tau,k) + z\widehat{f}_1(\tau,k)\right)\right) \\
&\quad \times \widehat{G}(\tau,k,z)\widehat{h}(k)dk \\
&= \frac{e^{-r\tau}}{2\pi}\left(P_{0,0}(t,x,z) + \kappa\theta\sqrt{\epsilon}P_{1,0}(t,x,z) + z\sqrt{\epsilon}P_{1,1}(t,x,z)\right),
\end{aligned}$$

where we have defined

$$P_{0,0}(t,x,z) := \int_{\mathbb{R}} e^{-ikq}\widehat{G}(\tau,k,z)\widehat{h}(k)dk, \tag{D.1}$$

$$P_{1,0}(t,x,z) := \int_{\mathbb{R}} e^{-ikq}\widehat{f}_0(\tau,k)\widehat{G}(\tau,k,z)\widehat{h}(k)dk, \tag{D.2}$$

$$P_{1,1}(t,x,z) := \int_{\mathbb{R}} e^{-ikq}\widehat{f}_1(\tau,k)\widehat{G}(\tau,k,z)\widehat{h}(k)dk. \tag{D.3}$$

As they are written, (D.1), (D.2), and (D.3) are general enough to accommodate any European option. However, in order to make progress, we now specify an option payoff. We will limit ourselves to the case of a European call, which has payoff $h(x) = (x-K)^+$. Extension to other European options is straightforward.

We remind the reader that $\widehat{h}(k)$ is the Fourier transform of the option payoff, expressed as a function of $q = r(T-t) + \log(x)$. For the case of the European call, we have

$$\widehat{h}(k) = \int_{\mathbb{R}} e^{ikq}(e^q - K)^+ dq = \frac{K^{1+ik}}{ik - k^2}. \tag{D.4}$$

We note that (D.4) will not converge unless the imaginary part of k is greater than 1. Thus, we decompose k into its real and imaginary parts and impose the following condition on the imaginary part of k:

$$\begin{aligned} k &= k_r + ik_i, \\ k_i &> 1. \end{aligned} \tag{D.5}$$

When we integrate over k in (D.1), (D.2), and (D.3), we hold $k_i > 1$ fixed and integrate k_r over $\mathbb{R}$.

Numerical evaluation of $P_{0,0}(t,x,z)$. We rewrite (D.1) here, explicitly using expressions (4.7) and (D.4) for $\widehat{G}(\tau,k,z)$ and $\widehat{h}(k)$, respectively:

$$P_{0,0}(t,x,z) = \int_{\mathbb{R}} e^{-ikq} e^{C(\tau,k)+zD(\tau,k)} \frac{K^{1+ik}}{ik-k^2} dk_r. \tag{D.6}$$

In order for any numerical integration scheme to work, we must verify the continuity of the integrand in (D.6). First, by (D.5), the poles at $k = 0$ and $k = i$ are avoided. The only other worrisome term in the integrand of (D.6) is $e^{C(\tau,k)}$, which may be discontinuous due to the presence of the log in $C(\tau,k)$.

We recall that any $\zeta \in \mathbb{C}$ can be represented in polar notation as $\zeta = r\exp(i\theta)$, where $\theta \in [-\pi,\pi)$. In this notation, $\log\zeta = \log r + i\theta$. Now, suppose we have a map $\zeta(k_r) : \mathbb{R} \to \mathbb{C}$. We see that whenever $\zeta(k_r)$ crosses the negative real axis, $\log\zeta(k_r)$ will be discontinuous (due to θ jumping from $-\pi$ to π or from π to $-\pi$). Thus, in order for $\log\zeta(k_r)$ to be continuous, we must ensure that $\zeta(k_r)$ does not cross the negative real axis.

We now return our attention to $C(\tau,k)$. We note that $C(\tau,k)$ has two algebraically equivalent representations, (4.8) and the following representation:

$$C(\tau,k) = \frac{\kappa\theta}{\sigma^2}\left((\kappa + \rho i k\sigma - d(k))\,\tau - 2\log\zeta(\tau,k)\right), \tag{D.7}$$

$$\zeta(\tau,k) := \frac{e^{-\tau d(k)}/g(k) - 1}{1/g(k) - 1}. \tag{D.8}$$

It turns out that, under most reasonable conditions, $\zeta(\tau,k)$ does not cross the negative real axis [17]. As such, as one integrates over k_r, no discontinuities will arise from the $\log\zeta(\tau,k)$ which appears in (D.7). Therefore, if we use expression (D.7) when evaluating (D.6), the integrand will be continuous.

Numerical evaluation of $P_{1,1}(t,x,z)$ and $P_{1,0}(t,x,z)$. The integrands in (D.3) and (D.2) are identical to that of (D.1), except for the additional factor of $\widehat{f_1}(\tau,k)$. Using (4.21) for $\widehat{f_1}(\tau,k)$, we have the following expression for $P_{1,1}(t,x,z)$:

$$\begin{aligned} &P_{1,1}(t,x,z) \\ &= \int_{\mathbb{R}} e^{-ikq}\left(\int_0^\tau b(s,k)e^{A(\tau,k,s)}ds\right)e^{C(\tau,k)+zD(\tau,k)}\frac{K^{1+ik}}{ik-k^2}dk_r \\ &= \int_0^\tau \int_{\mathbb{R}} e^{-ikq} b(s,k) e^{A(\tau,k,s)+C(\tau,k)+zD(\tau,k)}\frac{K^{1+ik}}{ik-k^2}dk_r ds. \end{aligned} \tag{D.9}$$

Similarly,

$$\begin{aligned} &P_{1,0}(t,x,z) \\ &= \int_0^\tau \int_0^t \int_{\mathbb{R}} e^{-ikq} b(s,k) e^{A(t,k,s)+C(\tau,k)+zD(\tau,k)}\frac{K^{1+ik}}{ik-k^2}dk_r ds dt. \end{aligned} \tag{D.10}$$

We already know, from our analysis of $P_{0,0}(t,x,z)$, how to deal with the log in $C(\tau,k)$. It turns out that the log in $A(\tau,k,s)$ can be dealt with in a similar manner. Consider the following representation for $A(\tau,k,s)$, which is algebraically equivalent to expression (4.22):

$$\begin{aligned} A(\tau,k,s) = (\kappa + \rho i k\sigma + d(k))&\left(\frac{1-g(k)}{d(k)g(k)}\right) \\ &\times (d(k)(\tau - s) + \log\zeta(\tau,k) - \log\zeta(s,k)) \\ &+ d(k)(\tau - s), \end{aligned} \tag{D.11}$$

where $\zeta(\tau,k)$ is defined in (D.8). As expressed in (D.11), $A(\tau,k,s)$ is, under most reasonable conditions, a continuous function of k_r. Thus, if we use (D.11) when numerically evaluating (D.9) and (D.10), their integrands will be continuous.

Transforming the domain of integration. Aside from using (D.7) and (D.11) for $C(\tau,k)$ and $A(\tau,k,s)$, there are a few other tricks we can use to facilitate the numerical evaluation of (D.6), (D.10), and (D.9). Denote by $I_0(k)$ and $I_1(k,s)$ the integrands appearing in (D.6), (D.9), and (D.10):

$$\begin{aligned} P_{0,0} &= \int_{\mathbb{R}} I_0(k) dk_r, \\ P_{1,1} &= \int_0^\tau \int_{\mathbb{R}} I_1(k,s) dk_r ds, \\ P_{1,0} &= \int_0^\tau \int_0^t \int_{\mathbb{R}} I_1(k,s) dk_r ds dt. \end{aligned}$$

First, we note that the real and imaginary parts of $I_0(k)$ and $I_1(k,s)$ are even and odd functions of k_r, respectively. As such, instead of integrating in k_r over $\mathbb{R}$, we can integrate in k_r over $\mathbb{R}_+$, drop the imaginary part, and multiply the result by 2.

Second, numerically integrating in k_r over $\mathbb{R}_+$ requires that one arbitrarily truncate the integral at some k_{cutoff}. Rather than doing this, we can make the following variable transformation, suggested by [14]:

$$\begin{aligned} k_r &= \frac{-\log u}{C_\infty}, \\ C_\infty &:= \frac{\sqrt{1-\rho^2}}{\sigma}(z+\kappa\theta\tau). \end{aligned} \tag{D.12}$$

Then, for some arbitrary $I(k)$, we have

$$\int_0^\infty I(k) dk_r = \int_0^1 I\left(\frac{-\log u}{C_\infty} + ik_i\right) \frac{1}{uC_\infty} du.$$

Thus, we avoid having to establish a cutoff value, k_{cutoff} (and avoid the error that comes along with doing so).

Finally, evaluating (D.10) requires that one integrate over the triangular region parameterized by $0 \le s \le t \le \tau$. Unfortunately, most numerical integration packages facilitate integration only over a rectangular region. We can overcome this difficulty by performing the following transformation of variables:

$$\begin{aligned} s &= tv, \\ ds &= tdv. \end{aligned}$$

Then, for some arbitrary $I(s)$, we have

$$\int_0^\tau \int_0^t I(s) ds dt = \int_0^\tau \int_0^1 I(tv) t dv dt. \tag{D.13}$$

Pulling everything together we obtain

$$P_{0,0} = 2\mathrm{Re} \int_0^1 I_0 \left(\frac{-\log u}{C_\infty} + ik_i \right) \frac{1}{uC_\infty} du,$$

$$P_{1,1} = 2\mathrm{Re} \int_0^\tau \int_0^1 I_1 \left(\frac{-\log u}{C_\infty} + ik_i, s \right) \frac{1}{uC_\infty} duds,$$

$$P_{1,0} = 2\mathrm{Re} \int_0^\tau \int_0^1 \int_0^1 I_1 \left(\frac{-\log u}{C_\infty} + ik_i, tv \right) \frac{t}{uC_\infty} dudvdt,$$

where C_∞ is given by (D.12). These three changes allow one to efficiently and accurately numerically evaluate (D.6), (D.9), and (D.10).

Numerical tests show that for strikes ranging from 0.5 to 1.5 times the spot price, and for expirations ranging from 3 months to 3 years, it takes roughly 100 times longer to calculate a volatility surface using the multiscale model than it does to calculate the same surface using the Heston model.

Acknowledgments. The authors would like to thank Ronnie Sircar and Knut Sølna for earlier discussions on the model studied in this paper. They also thank the two anonymous referees for their suggestions that greatly helped improve the paper.

REFERENCES

[1] S. Alizadeh, M. W. Brandt, and F. X. Diebold, *Range-Based Estimation of Stochastic Volatility Models*, SSRN eLibrary, 2001.

[2] T. G. Andersen and T. Bollerslev, *Intraday periodicity and volatility persistence in financial markets*, J. Empir. Finance, 4 (1997), pp. 115–158.

[3] M. Chernov, A. R. Gallant, E. Ghysels, and G. Tauchen, *Alternative models for stock price dynamics*, J. Econometrics, 116 (2003), pp. 225–257.

[4] P. Cotton, J.-P. Fouque, G. Papanicolaou, and R. Sircar, *Stochastic volatility corrections for interest rate derivatives*, Math. Finance, 14 (2004), pp. 173–200.

[5] R. F. Engle and A. J. Patton, *What good is a volatility model?*, Quant. Finance, 1 (2001), pp. 237–245.

[6] G. Fiorentini, A. Leon, and G. Rubio, *Estimation and empirical performance of Heston's stochastic volatility model: The case of a thinly traded market*, J. Empir. Finance, 9 (2002), pp. 225–255.

[7] J.-P. Fouque, G. Papanicolaou, and R. Sircar, *Derivatives in Financial Markets with Stochastic Volatility*, Cambridge University Press, Cambridge, UK, 2000.

[8] J.-P. Fouque, G. Papanicolaou, R. Sircar, and K. Solna, *Short time-scale in S&P* 500 *volatility*, J. Comput. Finance, 6 (2003), pp. 1–23.

[9] J.-P. Fouque, G. Papanicolaou, R. Sircar, and K. Solna, *Singular perturbations in option pricing*, SIAM J. Appl. Math., 63 (2003), pp. 1648–1665.

[10] J. Gatheral, *Modeling the implied volatility surface*, in Global Derivatives and Risk Management, Barcelona, 2003.

[11] J. Gatheral, *The Volatility Surface: A Practitioner's Guide*, John Wiley and Sons, New York, 2006.

[12] S. Heston, *A closed-form solution for options with stochastic volatility with applications to bond and currency options*, Rev. Financ. Stud., 6 (1993), pp. 327–343.

[13] E. Hillebrand, *Overlaying time scales in financial volatility data*, in Econometric Analysis of Financial and Economic Time Series/Part B, T. B. Fomby and D. Terrell, eds., Adv. Econom. 20, Elsevier, Amsterdam, 2006, pp. 153–178.

[14] P. Jackel and C. Kahl, *Not-so-complex logarithms in the Heston model*, Wilmott Magazine, September (2005), pp. 94–103.

[15] D. Lamberton and B. Lapeyre, *Introduction to Stochastic Calculus Applied to Finance*, Chapman & Hall, London, 1996.

[16] B. D. Lebaron, *Stochastic Volatility as a Simple Generator of Financial Power-Laws and Long Memory*, SSRN eLibrary, 2001.

[17] R. Lord and C. Kahl, *Why the Rotation Count Algorithm Works*, SSRN eLibrary, 2006.

[18] A. Melino and S. M. Turnbull, *Pricing foreign currency options with stochastic volatility*, J. Econometrics, 45 (1990), pp. 239–265.

[19] U. A. Muller, M. M. Dacorogna, R. D. Dave, R. B. Olsen, O. V. Pictet, and J. E. von Weizsacker, *Volatilities of different time resolutions—analyzing the dynamics of market components*, J. Empir. Finance, 4 (1997), pp. 213–239.

[20] W. Shaw, *Stochastic Volatility, Models of Heston Type*, http://www.mth.kcl.ac.uk/~shaww/web_page/papers/StoVolLecture.pdf.

[21] J. Zhang and J. Shu, *Pricing Standard & Poor's* 500 *index options with Heston's model*, in Proceedings of the IEEE Conference on Computational Intelligence for Financial Engineering, 2003, pp. 85–92.

SIAM J. FINANCIAL MATH.
Vol. 2, pp. 255–286

On the Heston Model with Stochastic Interest Rates*

Lech A. Grzelak† and Cornelis W. Oosterlee‡

Abstract. We discuss the Heston model [*Rev. Financ. Stud.*, 6 (1993), pp. 327–343] with stochastic interest rates driven by Hull–White (HW) [*J. Derivatives*, 4 (1996), pp. 26–36] or Cox–Ingersoll–Ross (CIR) [*Econometrica*, 53 (1985), pp. 385–407] processes. Two projection techniques to derive affine approximations of the original hybrid models are presented. In these approximations we can prescribe a nonzero correlation structure between all underlying processes. The affine approximate models admit pricing basic derivative products by Fourier techniques [P. P. Carr and D. B. Madan, *J. Comput. Finance*, 2 (1999), pp. 61–73, F. Fang and C. W. Oosterlee, *SIAM J. Sci. Comput.*, 31 (2008), pp. 826–848] and can therefore be used for fast calibration of the hybrid model.

Key words. equity-interest rate hybrid models, stochastic volatility, Heston–Hull–White and Heston–Cox–Ingersoll–Ross processes, approximation by affine diffusion process

AMS subject classifications. 91G60, 91G30, 91G20

DOI. 10.1137/090756119

1. Introduction.

Modelling derivative products in finance usually starts with the specification of a system of stochastic differential equations (SDEs) that correspond to state variables like stock, interest rate, and volatility. By correlating the SDEs from the different asset classes, one can define so-called hybrid models and use them for pricing multiasset derivatives. Even if each of these SDEs yields a closed-form solution, a nonzero correlation structure between the processes may cause difficulties for modeling and product pricing. Typically, a closed-form solution of the hybrid models is not known, and numerical approximation by means of Monte Carlo (MC) simulation or discretization of the corresponding partial differential equations (PDEs) has to be employed for model evaluation and derivative pricing. The speed of pricing European products is, however, crucial, especially for the calibration. Several theoretically attractive SDE models that cannot fulfill the speed requirements are not used in practice.

The aim of this paper is to define hybrid SDE models that fit in the class of affine diffusion processes (AD), as in Duffie, Pan, and Singleton [13]. For processes within this class a closed-form solution of the characteristic function (ChF) exists. Suppose we have given a system of SDEs, i.e.,

$$\mathrm{d}\mathbf{X}(t) = \mu(\mathbf{X}(t))\mathrm{d}t + \sigma(\mathbf{X}(t))\mathrm{d}\mathbf{W}(t). \tag{1.1}$$

*Received by the editors April 15, 2009; accepted for publication (in revised form) January 12, 2011; published electronically March 15, 2011.
http://www.siam.org/journals/sifin/2/75611.html

†Delft University of Technology, Delft Institute of Applied Mathematics, Mekelweg 4, 2628 CD, Delft, The Netherlands, and Derivatives Research and Validation Group, Rabobank, Jaarbeursplein 22, 3521 AP, Utrecht, The Netherlands (L.A.Grzelak@tudelft.nl).

‡CWI – Centrum Wiskunde & Informatica, Amsterdam, The Netherlands, and Delft University of Technology, Delft Institute of Applied Mathematics Mekelweg 4, 2628 CD, Delft, The Netherlands (C.W.Oosterlee@cwi.nl).

This system (1.1) is said to be of affine form if

$$\mu(\mathbf{X}(t)) = a_0 + a_1\mathbf{X}(t) \quad \text{for any } (a_0, a_1) \in \mathbb{R}^n \times \mathbb{R}^{n\times n}, \tag{1.2}$$

$$(\sigma(\mathbf{X}(t))\sigma(\mathbf{X}(t))^{\mathrm{T}})_{i,j} = (c_0)_{ij} + (c_1)_{ij}^{\mathrm{T}}\mathbf{X}(t), \quad \text{with } (c_0, c_1) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n\times n}, \tag{1.3}$$

$$r(\mathbf{X}(t)) = r_0 + r_1^{\mathrm{T}}\mathbf{X}(t) \quad \text{for } (r_0, r_1) \in \mathbb{R} \times \mathbb{R}^n, \tag{1.4}$$

for $i, j = 1, \ldots, n$, with $r(\mathbf{X}(t))$ being an interest rate component. Then, the discounted ChF is of the following form [13]:

$$\phi(\mathbf{u}, \mathbf{X}(t), t, T) = \mathbb{E}^{\mathbb{Q}}\left(\exp\left(-\int_t^T r(s)\mathrm{d}s + i\mathbf{u}^{\mathrm{T}}\mathbf{X}(T)\right)\Big|\mathcal{F}(t)\right) = \mathrm{e}^{A(\mathbf{u},\tau)+\mathbf{B}^{\mathrm{T}}(\mathbf{u},\tau)\mathbf{X}(t)},$$

where the expectation is taken under the risk-neutral measure, $\mathbb{Q}$. For a time lag, $\tau := T - t$, the coefficients $A(\mathbf{u}, \tau)$ and $\mathbf{B}^{\mathrm{T}}(\mathbf{u}, \tau)$ have to satisfy the following complex-valued ordinary differential equations (ODEs):

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}\tau}\mathbf{B}(\mathbf{u},\tau) = & -r_1 + a_1^{\mathrm{T}}\mathbf{B}(\mathbf{u},\tau) + & \dfrac{1}{2}\mathbf{B}^{\mathrm{T}}(\mathbf{u},\tau)c_1\mathbf{B}(\mathbf{u},\tau), \\ \dfrac{\mathrm{d}}{\mathrm{d}\tau}A(\mathbf{u},\tau) = & -r_0 + \mathbf{B}^{\mathrm{T}}(\mathbf{u},\tau)a_0 + & \dfrac{1}{2}\mathbf{B}^{\mathrm{T}}(\mathbf{u},\tau)c_0\mathbf{B}(\mathbf{u},\tau), \end{cases} \tag{1.5}$$

with $a_i, c_i, r_i,\ i = 0, 1$, as in (1.2), (1.3), and (1.4).

In this article we focus our attention specifically on a hybrid model which combines the equity and interest rate asset classes. Brigo and Mercurio [8] have shown that the assumption of constant interest rates in the classical Black–Scholes model [7] can be generalized, and by including the stochastic interest rate process of Hull and White [23], one is still able to obtain a closed-form solution for European-style option prices. Originally, the Black–Scholes–Hull–White model in [8] was not dedicated to pricing hybrid products but to increasing the accuracy for long-maturity options. The model is not, however, able to describe any smile and skew shapes present in the equity markets.

In [37] a hybrid model was presented which could provide the skew pattern for the equity and included a stochastic (but uncorrelated) interest rate process. Generalizations were presented in [19, 3], where the Heston stochastic volatility model [22] was used. Some form of correlation was indirectly modeled by including additional terms in the SDEs (this approach is discussed in some detail in section 3.1.1).

In [20, 21] the Heston stochastic volatility model was replaced by the Schöbel–Zhu (SZ) model [35], while the interest rate was still driven by a Hull–White (HW) process (SZHW model). In this model a full matrix of correlations can be directly imposed on the driving Brownian motions. The model is well defined under the class of AD processes, but since the SZHW model is based on a Vašiček-type process [36] for the stochastic volatility, the volatilities can become negative.

A different approach to modelling equity-interest rate hybrids was presented by Benhamou, Rivoira, and Gruz [6], extending the local volatility framework of Dupire [15] and Derman and Kani [12] and incorporating stochastic interest rates.

Here, we investigate the Heston–Hull–White (HHW) and the Heston–Cox–Ingersoll–Ross (HCIR) hybrid models and propose approximations so that we can obtain their ChFs. The framework presented is relatively easy to understand and implement. It is inspired by the techniques in [19, 2].

Our approximations do not require several preliminary calculations of expectations like in the case of Markovian projection methods [4, 5]. The resulting option pricing method benefits greatly from the speed of ChF evaluations.

The interest rate models studied here cannot generate *interest rate* implied volatility smiles or skews. They can therefore mainly be used for long-term equity options and for "not too complicated" equity-interest rate hybrid products. As described in [24], for accurate modeling of hybrid derivatives, it is necessary to be able to describe a nonzero correlation between equity and interest rate. This is possible in the approximations presented here.

This paper is organized as follows. In section 2 we discuss the full-scale Heston hybrid models with stochastic interest rate processes. Section 3 presents a deterministic approximation of the HHW hybrid model, together with the corresponding ChF, and section 4 gives the ChF based on another stochastic approximation of that hybrid model. In section 5 we deal with the HCIR model. In section 6 the calibration based on the approximations of the full-scale hybrid models is applied. Section 7 offers concluding remarks. Details of proofs and tests are in the appendices, where, in particular, in Appendix C we compare the performance of the approximations developed with the Markovian projection method studied in [4, 5].

2. Heston hybrid models with stochastic interest rate. With state vector $\mathbf{X}(t) = [S(t), v(t)]^{\mathrm{T}}$, under the risk-neutral pricing measure, the Heston stochastic volatility model [22], which is our point of departure, is specified by the following system of SDEs:

$$\begin{cases} \mathrm{d}S(t)/S(t) = & r\mathrm{d}t + \sqrt{v(t)}\mathrm{d}W_x(t), \quad S(0) > 0, \\ \mathrm{d}v(t) = & \kappa(\bar{v} - v(t))\mathrm{d}t + \gamma\sqrt{v(t)}\mathrm{d}W_v(t), \quad v(0) > 0, \end{cases} \tag{2.1}$$

with $r > 0$ a constant interest rate, correlation $\mathrm{d}W_x(t)\mathrm{d}W_v(t) = \rho_{x,v}\mathrm{d}t$, and $|\rho_{x,v}| < 1$. The variance process, $v(t)$, of the stock, $S(t)$, is a mean reverting square-root process, in which $\kappa > 0$ determines the speed of adjustment of the volatility towards its theoretical mean $\bar{v} > 0$, and $\gamma > 0$ is the second-order volatility, i.e., the volatility of the volatility.

As already indicated in [22], the model given in (2.1) is not in the class of affine processes, whereas under the log-transform for the stock, $x(t) = \log S(t)$, it is. Then, the discounted ChF is given by

$$\phi_{\mathrm{H}}(u, \mathbf{X}(t), \tau) = \exp\left(A(u,\tau) + B(u,\tau)x(t) + C(u,\tau)v(t)\right), \tag{2.2}$$

where the functions $A(u,\tau)$, $B(u,\tau)$, and $C(u,\tau)$ are known in closed form (see [22]).

The ChF is explicit, but its inverse also has to be found for pricing purposes. Because of the form of the ChF, we cannot get its inverse analytically, and a numerical method for integration has to be used; see, for example, [10, 16, 29, 30] for Fourier methods.

2.1. Full-scale hybrid models. A constant interest rate, r, may be insufficient for pricing interest rate sensitive products. Therefore, we extend our state vector with an additional stochastic quantity, i.e., $\mathbf{X}(t) = [S(t), v(t), r(t)]^{\mathrm{T}}$. This model corresponds to a *hybrid stochastic volatility equity model with a stochastic interest rate* process, $r(t)$. In particular, we add to the Heston model the HW interest rate [23] or the square-root CIR process [11]. The extended model can be presented in the following way:

$$\begin{cases} \mathrm{d}S(t)/S(t) = & r(t)\mathrm{d}t + \sqrt{v(t)}\mathrm{d}W_x(t), \quad S(0) > 0, \\ \mathrm{d}v(t) = & \kappa(\bar{v} - v(t))\mathrm{d}t + \gamma\sqrt{v(t)}\mathrm{d}W_v(t), \quad v(0) > 0, \\ \mathrm{d}r(t) = & \lambda(\theta(t) - r(t))\mathrm{d}t + \eta r^p(t)\mathrm{d}W_r(t), \quad r(0) > 0, \end{cases} \tag{2.3}$$

where exponent $p = 0$ in (2.3) represents the HHW model and for $p = \frac{1}{2}$ it becomes the HCIR model. For both models the correlations are given by $\mathrm{d}W_x(t)\mathrm{d}W_v(t) = \rho_{x,v}\mathrm{d}t$, $\mathrm{d}W_x(t)\mathrm{d}W_r(t) = \rho_{x,r}\mathrm{d}t$, and $\mathrm{d}W_v(t)\mathrm{d}W_r(t) = \rho_{v,r}\mathrm{d}t$, and κ, γ, and $\bar{v}$ are as in (2.1); $\lambda > 0$ determines the speed of mean reversion for the interest rate process; $\theta(t)$ is the interest rate term structure; and η controls the volatility of the interest rate. We note that the interest rate process in (2.3) for $p = \frac{1}{2}$ is of the same form as the variance process $v(t)$.

System (2.3) is not in the affine form, not even with $x(t) = \log S(t)$. In particular, the symmetric instantaneous covariance matrix is given by

$$\sigma(\mathbf{X}(t))\sigma(\mathbf{X}(t))^{\mathrm{T}} = \begin{bmatrix} v(t) & \rho_{x,v}\gamma v(t) & \rho_{x,r}\eta r^p(t)\sqrt{v(t)} \\ * & \gamma^2 v(t) & \rho_{r,v}\gamma\eta r^p(t)\sqrt{v(t)} \\ * & * & \eta^2 r^{2p}(t) \end{bmatrix}_{(3\times 3)}. \tag{2.4}$$

Setting the correlation $\rho_{r,v}$ to zero would still not make the system affine. Matrix (2.4) is of the linear form with respect to state vector $[x(t) = \log S(t), v(t), r(t)]^{\mathrm{T}}$ if two correlations, $\rho_{r,v}$ and $\rho_{x,r}$, are set to zero.[1] Models with two correlations equal to zero are covered in [31].

Since for pricing equity-interest rate products a nonzero correlation between stock and interest rate is crucial (see, for example, [24]), approximations to the Heston hybrid models need to be formulated, so that correlations can be imposed. Variants are discussed in the sections to follow. These approximate models are evaluated with the help of the Cholesky decomposition of a correlation matrix.

We can decompose a given general symmetric correlation matrix, $\mathbf{C}$, denoted by

$$\mathbf{C} = \begin{bmatrix} 1 & \rho_1 & \rho_2 \\ * & 1 & \rho_3 \\ * & * & 1 \end{bmatrix}, \tag{2.5}$$

as $\mathbf{C} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$, where $\mathbf{L}$ is a lower triangular matrix, with

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ \rho_1 & \sqrt{1-\rho_1^2} & 0 \\ \rho_2 & \frac{\rho_3 - \rho_2\rho_1}{\sqrt{1-\rho_1^2}} & \sqrt{1 - \rho_2^2 - \left(\frac{\rho_3 - \rho_2\rho_1}{\sqrt{1-\rho_1^2}}\right)^2} \end{bmatrix}. \tag{2.6}$$

We then rewrite the system of SDEs in terms of the independent Brownian motions, $\mathrm{d}\widetilde{\mathbf{W}}(t)$, with the help of the lower triangular matrix $\mathbf{L}$.

---

[1] Here we assume positive parameters.

Since our main objective is to derive a closed-form ChF while assuming a nonzero correlation between the equity process, $S(t)$, and the interest rate, $r(t)$, we first assume that the Brownian motions for the interest rate $r(t)$ and the variance $v(t)$ are not correlated (the case of a full correlation structure is discussed in detail in Appendix B).

By exchanging the order of the state variables $\mathbf{X}(t) = [S(t), v(t), r(t)]^{\mathrm{T}}$ to $\mathbf{X}^*(t) = [r(t), v(t), S(t)]^{\mathrm{T}}$, the HHW and HCIR models in (2.3) then have $\rho_1 \equiv \rho_{r,v} = 0$, $\rho_2 \equiv \rho_{x,r} \neq 0$, and $\rho_3 \equiv \rho_{x,v} \neq 0$ in (2.5) and read as

$$\begin{bmatrix} \mathrm{d}r(t) \\ \mathrm{d}v(t) \\ \dfrac{\mathrm{d}S(t)}{S(t)} \end{bmatrix} = \begin{bmatrix} \lambda(\theta(t) - r(t)) \\ \kappa(\bar{v} - v(t)) \\ r(t) \end{bmatrix} \mathrm{d}t + \sigma(\mathbf{X}^*(t)) \begin{bmatrix} \mathrm{d}\widetilde{W}_r(t) \\ \mathrm{d}\widetilde{W}_v(t) \\ \mathrm{d}\widetilde{W}_x(t) \end{bmatrix}, \tag{2.7}$$

with

$$\sigma(\mathbf{X}^*(t)) = \begin{bmatrix} \eta r^p(t) & 0 & 0 \\ 0 & \gamma\sqrt{v(t)} & 0 \\ \rho_{x,r}\sqrt{v(t)} & \rho_{x,v}\sqrt{v(t)} & \sqrt{1-\rho_{x,v}^2-\rho_{x,r}^2}\sqrt{v(t)} \end{bmatrix}. \tag{2.8}$$

2.2. Reformulated Heston hybrid models. In the previous section we have seen that for the HHW and HCIR models with a full matrix of correlations given in (2.3), the affinity relations [13] are not satisfied, so that the ChF cannot be obtained by standard techniques.

In order to obtain a well-defined Heston hybrid model with an *indirectly imposed correlation*, $\rho_{x,r}$, we propose the following system of SDEs:

$$\mathrm{d}S(t)/S(t) = r(t)\mathrm{d}t + \sqrt{v(t)}\mathrm{d}W_x(t) + \Omega(t)r^p(t)\mathrm{d}W_r(t) + \Delta\sqrt{v(t)}\mathrm{d}W_v(t),\ S(0) > 0, \tag{2.9}$$

with

$$\begin{aligned} \mathrm{d}v(t) &= \kappa(\bar{v} - v(t))\mathrm{d}t + \gamma\sqrt{v(t)}\mathrm{d}W_v(t), \quad v(0) > 0, \\ \mathrm{d}r(t) &= \lambda(\theta(t) - r(t))\mathrm{d}t + \eta r^p(t)\mathrm{d}W_r(t), \quad r(0) > 0, \end{aligned} \tag{2.10}$$

where

$$\mathrm{d}W_x(t)\mathrm{d}W_v(t) = \hat{\rho}_{x,v}\mathrm{d}t, \quad \mathrm{d}W_x(t)\mathrm{d}W_r(t) = 0, \quad \mathrm{d}W_v(t)\mathrm{d}W_r(t) = 0, \tag{2.11}$$

where $p = 0$ for HHW and $p = \frac{1}{2}$ for HCIR. We have included a function,[2] $\Omega(t)$, and a constant parameter, Δ. Note that we still assume independence between the instantaneous short rate, $r(t)$, and the variance process, $v(t)$, i.e., $\hat{\rho}_{r,v} = 0$.

By exchanging the order of the state variables to $\mathbf{X}^*(t) = [r(t), v(t), S(t)]^{\mathrm{T}}$, system (2.9) is given, in terms of the independent Brownian motions, by

$$\begin{bmatrix} \mathrm{d}r(t) \\ \mathrm{d}v(t) \\ \dfrac{\mathrm{d}S(t)}{S(t)} \end{bmatrix} = \begin{bmatrix} \lambda(\theta(t) - r(t)) \\ \kappa(\bar{v} - v(t)) \\ r(t) \end{bmatrix} \mathrm{d}t + \hat{\sigma}(\mathbf{X}^*(t)) \begin{bmatrix} \mathrm{d}\widetilde{W}_r(t) \\ \mathrm{d}\widetilde{W}_v(t) \\ \mathrm{d}\widetilde{W}_x(t) \end{bmatrix}, \tag{2.12}$$

[2] This under certain conditions can also be stochastic.

where

$$\hat{\sigma}(\mathbf{X}^*(t)) = \begin{bmatrix} \eta r^p(t) & 0 & 0 \\ 0 & \gamma\sqrt{v(t)} & 0 \\ \Omega(t) r^p(t) & \sqrt{v(t)}\,(\hat{\rho}_{x,v}+\Delta) & \sqrt{v(t)}\sqrt{1-\hat{\rho}^2_{x,v}} \end{bmatrix}. \tag{2.13}$$

In the following lemma we show that the model (2.9) is equivalent to the full-scale HHW model in (2.3), with nonzero correlation $\rho_{x,r}$.

Lemma 2.1. *Model (2.9) satisfies the system in (2.3) with nonzero correlation $\rho_{x,r}$ for*

$$\Omega(t) = \rho_{x,r}\frac{\sqrt{v(t)}}{r^p(t)}, \quad \hat{\rho}^2_{x,v} = \rho^2_{x,v} + \rho^2_{x,r}, \quad \Delta = \rho_{x,v} - \hat{\rho}_{x,v}, \tag{2.14}$$

where correlation $\hat{\rho}_{x,v}$ is as in model (2.9) and $\rho_{x,v}$ as in model (2.3).

Proof. We presented the two models (2.3) and (2.9) in terms of the independent Brownian motions, (2.7) and (2.12), respectively. By matching the appropriate coefficients in (2.7) and (2.12), we find that the following relations should hold:

$$\begin{cases} \Omega(t) r^p(t) S(t) = & \rho_{x,r}\sqrt{v(t)}S(t), \\ \sqrt{1-\hat{\rho}^2_{x,v}}\sqrt{v(t)}S(t) = & \sqrt{1-\rho^2_{x,v}-\rho^2_{x,r}}\sqrt{v(t)}S(t), \\ (\hat{\rho}_{x,v}+\Delta)\sqrt{v(t)}S(t) = & \rho_{x,v}\sqrt{v(t)}S(t). \end{cases} \tag{2.15}$$

By simplifying (2.15) the proof is finished. ∎

If results (2.14) were directly included in the main system (2.9), the affinity property of the system would be lost. So, in order to satisfy the affinity constraints, appropriate *approximations* need to be introduced.

2.3. Log-transform.

Before going into the details of the approximations of the HHW and HCIR models, let us first find the dynamics for the log-transform of the reformulated Heston hybrid models. By applying Itô's lemma, model (2.9) in log-equity space, $x(t) = \log S(t)$, with a constant parameter, Δ, and a function, $\Omega(t)$, is given by

$$\begin{aligned} \mathrm{d}x(t) &= \left[r(t) - \frac{1}{2}\left(\Omega^2(t)r^{2p}(t) + v(t)\left(1+\Delta^2+2\hat{\rho}_{x,v}\Delta\right)\right)\right]\mathrm{d}t + \sqrt{v(t)}\mathrm{d}W_x(t) \\ &\quad + \Omega(t)r^p(t)\mathrm{d}W_r(t) + \Delta\sqrt{v(t)}\mathrm{d}W_v(t) \\ &= \left(r(t) - \frac{1}{2}v(t)\right)\mathrm{d}t + \sqrt{v(t)}\mathrm{d}W_x(t) + \Omega(t)r^p(t)\mathrm{d}W_r(t) + \Delta\sqrt{v(t)}\mathrm{d}W_v(t), \end{aligned}$$

because of (2.14).

For a given state vector $\mathbf{X}^*(t) = [r(t), v(t), x(t)]^{\mathrm{T}}$, the symmetric instantaneous covariance matrix (1.3) is given by

$$\mathbf{\Sigma} := \begin{bmatrix} \eta^2 r^{2p}(t) & 0 & \eta\Omega(t)r^{2p}(t) \\ * & \gamma^2 v(t) & \gamma v(t)\left(\hat{\rho}_{x,v}+\Delta\right) \\ * & * & \Omega^2(t)r^{2p}(t) + v(t)\left(1+\Delta^2+2\hat{\rho}_{x,v}\Delta\right) \end{bmatrix}. \tag{2.16}$$

As we consider two cases for parameter $p = \{0, 1/2\}$, the affinity issue appears in only one term of matrix (2.16), namely, in element $(1, 3)$:

$$\boldsymbol{\Sigma}_{(1,3)} = \eta\Omega(t)r^{2p}(t) = \eta\rho_{x,r}\sqrt{v(t)}r^p(t) = \begin{cases} \eta\rho_{x,r}\sqrt{v(t)} & \text{for} \quad \text{HHW}, \\ \eta\rho_{x,r}\sqrt{v(t)}\sqrt{r(t)} & \text{for} \quad \text{HCIR}. \end{cases} \tag{2.17}$$

Although term $\boldsymbol{\Sigma}_{(3,3)}$ does not seems to be of the affine form, by (2.14), it equals $\boldsymbol{\Sigma}_{(3,3)} = v(t)$, and therefore it is linear in the state variables.

Remark 1. We see that, in order to make either the HHW or the HCIR model affine, one does not necessarily need to approximate function $\Omega(t)$; only the nonaffine terms in the corresponding instantaneous covariance matrix[3] need be approximated. By approximation of the nonaffine covariance term, $\boldsymbol{\Sigma}_{(1,3)}$, the corresponding pricing PDE also changes. The Kolmogorov backward equation for the log-stock price (see, for example, [33]) is now given by

$$\begin{aligned} 0 = {} & \frac{\partial\phi}{\partial t} + \left(r - \frac{1}{2}v\right)\frac{\partial\phi}{\partial x} + \kappa(\bar{v} - v)\frac{\partial\phi}{\partial v} + \lambda(\theta(t) - r)\frac{\partial\phi}{\partial r} + \frac{1}{2}v\frac{\partial^2\phi}{\partial x^2} \\ & + \frac{1}{2}\gamma^2 v\frac{\partial^2\phi}{\partial v^2} + \frac{1}{2}\eta^2 r^{2p}\frac{\partial^2\phi}{\partial r^2} + \rho_{x,v}\gamma v\frac{\partial^2\phi}{\partial x\partial v} + \boldsymbol{\Sigma}_{(1,3)}\frac{\partial^2\phi}{\partial x\partial r} - r\phi, \end{aligned} \tag{2.18}$$

subject to terminal condition $\phi(u, \mathbf{X}(T), T, T) = \exp(iux(T))$.

The derivations in section 2.3 show that system (2.9) is nothing but a reformulation of the original HHW system under the conditions in (2.14). It is therefore sufficient to linearize the nonaffine terms in the covariance matrix to determine an affine approximation of the full-scale model. In the sections to follow we discuss two possible approximations for $\boldsymbol{\Sigma}_{(1,3)}$.

3. Deterministic approximation for hybrid models.

In order to make the Heston hybrid model affine, we provide a first approximation for the expressions in (2.17) in section 3.1. The corresponding ChF is derived in section 3.2.

3.1. Deterministic approach: The H1-HW model.

The first approach to finding an approximation for the term $\boldsymbol{\Sigma}_{(1,3)} = \eta\rho_{x,r}\sqrt{v(t)}r^p(t)$ in matrix (2.16) is to replace it by its expectation; i.e.,

$$\boldsymbol{\Sigma}_{(1,3)} \approx \eta\rho_{x,r}\mathbb{E}\left(r^p(t)\sqrt{v(t)}\right) \stackrel{\perp\!\!\!\perp}{=} \eta\rho_{x,r}\mathbb{E}(r^p(t))\mathbb{E}(\sqrt{v(t)}), \tag{3.1}$$

assuming independence between $r(t)$ and $v(t)$.

The approximation for $\boldsymbol{\Sigma}_{(1,3)}$ in (3.1) consists of two expectations: one with respect to $\sqrt{v(t)}$ and another with respect to $r^p(t)$. $\mathbb{E}(r^p(t)) = 1$ for $p = 0$, and it is $\mathbb{E}(\sqrt{r(t)})$ for $p = 1/2$. Since the processes for $v(t)$ and $r(t)$ are then of the same type, the approximations are analogous. By taking the expectations of the stochastic variables, the model becomes of affine form, so that we can obtain the corresponding ChF.

In Lemma 3.1 the closed-form expressions for the expectation and variance of $\sqrt{v(t)}$ (a CIR-type process) are presented.

---

[3] The drifts and the interest rate are already in the affine form, presented in (1.2) and (1.4).

Lemma 3.1 (expectation and variance for CIR-type process). *For a given time $t > 0$ the expectation and variance of $\sqrt{v(t)}$, where $v(t)$ is a CIR-type process (2.1), are given by*

$$\mathbb{E}(\sqrt{v(t)}) = \sqrt{2c(t)}\mathrm{e}^{-\lambda(t)/2}\sum_{k=0}^{\infty}\frac{1}{k!}\left(\frac{\lambda(t)}{2}\right)^k\frac{\Gamma\left(\frac{1+d}{2}+k\right)}{\Gamma(\frac{d}{2}+k)} \tag{3.2}$$

and

$$\mathbb{V}ar\left(\sqrt{v(t)}\right) = c(t)(d+\lambda(t)) - 2c(t)\mathrm{e}^{-\lambda(t)}\left(\sum_{k=0}^{\infty}\frac{1}{k!}\left(\frac{\lambda(t)}{2}\right)^k\frac{\Gamma\left(\frac{1+d}{2}+k\right)}{\Gamma\left(\frac{d}{2}+k\right)}\right)^2, \tag{3.3}$$

where

$$c(t) = \frac{1}{4\kappa}\gamma^2(1-\mathrm{e}^{-\kappa t}), \quad d = \frac{4\kappa\bar{v}}{\gamma^2}, \quad \lambda(t) = \frac{4\kappa v(0)\mathrm{e}^{-\kappa t}}{\gamma^2(1-\mathrm{e}^{-\kappa t})}, \tag{3.4}$$

with $\Gamma(k)$ being the gamma function defined by

$$\Gamma(k) = \int_0^\infty t^{k-1}\mathrm{e}^{-t}\mathrm{d}t.$$

Proof. By [14] one can find the closed-form expression for the expectation $\mathbb{E}(\sqrt{v(t)})$, which by the principle of Kummer [28] can be simplified. ∎

The analytic expression for the expectation, either of $\sqrt{v(t)}$ or $\sqrt{r(t)}$ in (3.1), is involved and requires rather expensive numerical operations.

In order to find a first-order approximation, we can apply the so-called *delta method* (see, for example, [1, 32]), which states that a function $\varphi(X)$ can be approximated by a first-order Taylor expansion at $\mathbb{E}(X)$, for a given random variable, X, with expectation, $\mathbb{E}(X)$, and variance, $\mathbb{V}\mathrm{ar}(X)$, assuming that for $\varphi(X)$ its first derivative with respect to X exists and is sufficiently smooth.

The result below provides details of the approximation.

Result 3.2. *The expectation, $\mathbb{E}(\sqrt{v(t)})$, with stochastic process $v(t)$ given by (2.3), can be approximated by*

$$\mathbb{E}(\sqrt{v(t)}) \approx \sqrt{c(t)(\lambda(t)-1) + c(t)d + \frac{c(t)d}{2(d+\lambda(t))}} =: \Lambda(t), \tag{3.5}$$

with $c(t)$, d, and $\lambda(t)$ given in Lemma 3.1, and where κ, $\bar{v}$, γ, and $v(0)$ are the parameters given in (2.3).

In order to find the approximation in Result 3.2, we can use the delta method as follows. Assuming the function φ to be sufficiently smooth and the first two moments of X to exist, we obtain by first-order Taylor expansion

$$\varphi(X) \approx \varphi(\mathbb{E}X) + (X - \mathbb{E}X)\frac{\partial\varphi}{\partial X}(\mathbb{E}X). \tag{3.6}$$

Since the variance of $\varphi(X)$ can be approximated by the variance of the right-hand side of (3.6), we have

$$\mathbb{V}\mathrm{ar}(\varphi(X)) \approx \mathbb{V}\mathrm{ar}\left(\varphi(\mathbb{E}X) + (X - \mathbb{E}X)\frac{\partial\varphi}{\partial X}(\mathbb{E}X)\right)$$

$$= \left(\frac{\partial\varphi}{\partial X}(\mathbb{E}X)\right)^2 \mathbb{V}\mathrm{ar}X. \tag{3.7}$$

Now, by using this result for function $\varphi(v(t)) = \sqrt{v(t)}$, we find

$$\mathbb{V}\mathrm{ar}(\sqrt{v(t)}) \approx \left(\frac{1}{2}\frac{1}{\sqrt{\mathbb{E}(v(t))}}\right)^2 \mathbb{V}\mathrm{ar}(v(t)) = \frac{1}{4}\frac{\mathbb{V}\mathrm{ar}(v(t))}{\mathbb{E}(v(t))}. \tag{3.8}$$

However, from the definition of the variance, we also have

$$\mathbb{V}\mathrm{ar}(\sqrt{v(t)}) = \mathbb{E}(v(t)) - \left(\mathbb{E}(\sqrt{v(t)})\right)^2, \tag{3.9}$$

and by combining (3.8) and (3.9) we obtain the following approximation:

$$\mathbb{E}(\sqrt{v(t)}) \approx \sqrt{\mathbb{E}(v(t)) - \frac{1}{4}\frac{\mathbb{V}\mathrm{ar}(v(t))}{\mathbb{E}(v(t))}}. \tag{3.10}$$

Since $v(t)$ is a square-root process, as in (2.9), we have

$$v(t) = v(0)\mathrm{e}^{-\kappa t} + \bar{v}(1 - \mathrm{e}^{-\kappa t}) + \gamma\int_0^t \mathrm{e}^{\kappa(s-t)}\sqrt{v(s)}\mathrm{d}W_v(s). \tag{3.11}$$

The expectation reads $\mathbb{E}(v(t)) = c(t)(d + \lambda(t))$, and for the variance we get $\mathbb{V}\mathrm{ar}(v(t)) = c^2(t)(2d + 4\lambda(t))$, with $c(t)$, d, and $\lambda(t)$ as given in (3.4).

Now, by substituting these expressions in (3.10), the result is proved. Since Result 3.2 provides an explicit approximation for $\mathbf{\Sigma}_{(1,3)}$ in (3.1) in terms of a deterministic function for $\mathbb{E}(\sqrt{v(t)})$, we are, in principle, able to derive the corresponding ChF.

We show here for which parameters the expression under the square root in approximation (3.5) is nonnegative.

Consider the following inequality:

$$c(t)(\lambda(t) - 1) + c(t)d + \frac{c(t)d}{2(d + \lambda(t))} \geq 0. \tag{3.12}$$

Division by $c(t) > 0$ gives

$$\frac{2\,(\lambda(t) + d)\,(d + \lambda(t)) + d}{2(d + \lambda(t))} \geq 1. \tag{3.13}$$

So,

$$2\,(\lambda(t) + d)^2 - 2(\lambda(t) + d) + d \geq 0. \tag{3.14}$$

By setting $y = \lambda(t) + d$ we find $2y^2 - 2y + d \geq 0$. The parabola is nonnegative for the discriminant $4 - 4 \cdot 2 \cdot d \leq 0$, so that the expression in (3.5) is nonnegative for $d \geq \frac{1}{2}$ (i.e., $2d \geq 1$). With $d = 4\kappa\bar{v}/\gamma^2$ we can compare the inequality obtained to the Feller condition. If the Feller condition is satisfied, the expression under the square root is certainly well defined. If $8\kappa\bar{v}/\gamma^2 \geq 1$ but the Feller condition is not satisfied, the approximation is also valid. If the expression under the square root in (3.5) becomes negative, we suggest using the expressions in Lemma 3.1 instead.

Remark 2. We assume that the first-order linear terms around the parameter values in the Taylor expansion give an accurate representation. However, this may not work satisfactorily for "flat" density functions, like those from a uniform distribution. In order to increase the accuracy, higher-order terms can be included in the expansion [1]. More discussion on the conditions for the delta method to perform well can be found in [32].

The approximation for $\mathbb{E}(\sqrt{v(t)})$ in (3.5) is still nontrivial and may cause difficulties when deriving the corresponding ChFs. In order to find the coefficients of the ChF, a routine for numerically solving the corresponding ODEs has to be incorporated. Numerical integration, however, slows down the option pricing engine and would make the SDE model less attractive. As we aim to find a closed-form expression for the ChF, we further simplify $\Lambda(t)$ in (3.5). Expectation $\mathbb{E}(\sqrt{v(t)})$ can be further approximated by a function of the following form:

$$\mathbb{E}(\sqrt{v(t)}) \approx a + b\mathrm{e}^{-ct} =: \widetilde{\Lambda}(t), \tag{3.15}$$

with a, b, and c constant. Appropriate values for a, b, and c in (3.15) can be obtained via an optimization problem of the form $\min_{a,b,c} ||\Lambda(t) - \widetilde{\Lambda}(t)||_n$, where $||\cdot||_n$ is any nth norm.

We propose here, instead of a numerical approximation for these coefficients, a simple analytic expression in Result 3.3.

Result 3.3. *By matching functions $\Lambda(t)$ and $\widetilde{\Lambda}(t)$ for $t \to +\infty$, $t \to 0$, and $t = 1$, we find*

$$\begin{aligned} \lim_{t\to+\infty} \Lambda(t) = \quad \sqrt{\bar{v} - \frac{\gamma^2}{8\kappa}} &= \quad a = \lim_{t\to+\infty} \widetilde{\Lambda}(t), \\ \lim_{t\to 0} \Lambda(t) = \quad \sqrt{v(0)} &= \quad a + b = \lim_{t\to 0} \widetilde{\Lambda}(t), \\ \lim_{t\to 1} \Lambda(t) = \quad \Lambda(1) &= \quad a + b\mathrm{e}^{-c} = \lim_{t\to 1} \widetilde{\Lambda}(t). \end{aligned} \tag{3.16}$$

The values a, b, and c can now be estimated by

$$a = \sqrt{\bar{v} - \frac{\gamma^2}{8\kappa}}, \quad b = \sqrt{v(0)} - a, \quad c = -\log\left(b^{-1}(\Lambda(1) - a)\right), \tag{3.17}$$

where $\Lambda(t)$ is given by (3.5).

The approximation given in Result 3.3 may give difficulties for $\bar{v} < \gamma^2/8\kappa$ in (3.17) (the expression under the square root then becomes negative). We recognize that this expression is well defined, as the expression under the square root in the function $\Lambda(t)$ in Result 3.2 is positive.

In order to measure the quality of approximation (3.17) to $\mathbb{E}(\sqrt{v(t)})$ in (3.2), we perform a numerical experiment (see the results in Figure 1). For randomly chosen sets of parameters the approximation (3.17) resembles $\mathbb{E}(\sqrt{v(t)})$ in (3.2) very closely.

We call the resulting model the H1-HW model (HHW model-1).

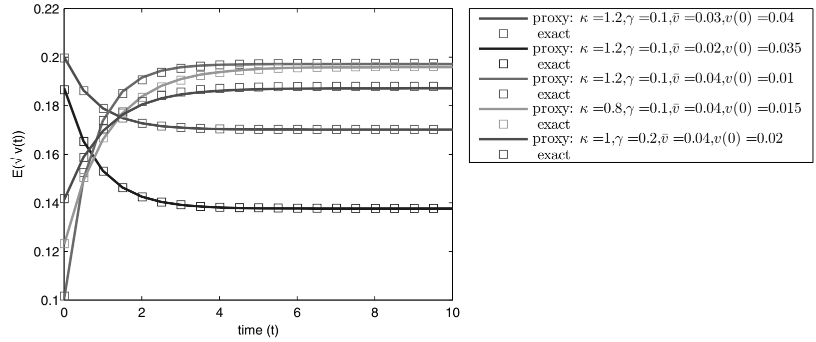


Figure 1. *The quality of the approximation* $\mathbb{E}(\sqrt{v(t)}) \approx a + b\mathrm{e}^{-ct}$ *(continuous line) versus the exact solution given in* (3.2) *(squares) for* 5 *random* κ, γ, $\bar{v}$, *and* $v(0)$.

3.1.1. The case $\Delta = 0$ and $\Omega(t) \equiv$ const. With $\Delta = 0$ in systems (2.9) and (2.12), the model resembles the one in [19, 3]. There, a constant parameter $\bar{\Omega} = \Omega(t)$ was prescribed, and an instantaneous correlation was imposed *indirectly*.

The following lemma, however, shows that this model with $\Delta = 0$ resembles the full-scale HHW and HCIR models only for correlation $\rho_{x,r} = 0$.

Lemma 3.4. *The hybrid models* (2.9) *with* $\Delta = 0$ *are full-scale HHW and HCIR models, in the sense of system* (2.3), *only if the instantaneous correlation between the stock and the interest rate processes in system* (2.3) *equals zero, i.e.,* $\rho_{x,r} = 0$.

Proof. The proof is analogous to the proof of Lemma 2.1. We see from the equalities in (2.14) that system (2.7) resembles system (2.12) with $\Delta = 0$ only if

$$\bar{\Omega} = \rho_{x,r}\frac{\sqrt{v(t)}}{r^p(t)}, \quad \hat{\rho}_{x,v} = \rho_{x,v}, \quad \hat{\rho}^2_{x,v} = \rho^2_{x,v} + \rho^2_{x,r}. \tag{3.18}$$

The equations (3.18) hold only for $\rho_{x,r} = 0$. So, the models with $\Delta = 0$ are not full-scale HHW and HCIR models with nonzero correlation $\rho_{x,r}$. ■

Although the model with $\Delta = 0$ is not a properly defined Heston hybrid model, one can still proceed with the analysis. Parameter $\bar{\Omega}$ was derived based on the following equality (see [19]), using the definition of the instantaneous correlation:

$$\hat{\rho}_{x,r} = \frac{\mathbb{E}\left(\mathrm{d}S(t)\mathrm{d}r(t)\right) - \mathbb{E}(\mathrm{d}S(t))\mathbb{E}(\mathrm{d}r(t))}{\sqrt{v(t)S^2(t)\,\mathrm{d}t + \bar{\Omega}^2 r^{2p}(t)S^2(t)\,\mathrm{d}t}\sqrt{\eta^2 r^{2p}(t)\,\mathrm{d}t}} = \frac{\bar{\Omega}r^p(t)}{\sqrt{v(t) + \bar{\Omega}^2 r^{2p}(t)}}. \tag{3.19}$$

To deal with the affinity issue a constant approximation for $\bar{\Omega}$ was proposed, given by

$$\bar{\Omega} \approx \frac{\hat{\rho}_{x,r}}{\sqrt{1-\hat{\rho}^2_{x,r}}}\mathbb{E}\left(\frac{1}{T}\int_0^T v(t)\mathrm{d}t\right)^{\frac{1}{2}} \Big/ \mathbb{E}\left(\frac{1}{T}\int_0^T r(t)\mathrm{d}t\right)^p. \tag{3.20}$$

By choosing $\bar{\Omega} = 0$ the model collapses to the well-known HHW model ($p = 0$) or the HCIR model ($p = \frac{1}{2}$) with zero correlation $\rho_{x,r}$.

In Figure 2 we present the behavior of the instantaneous correlation between the equity and the interest rates. We see that for time-dependent $\Omega(t)$, as defined in Lemma 2.1, the instantaneous correlations are stable and oscillate around the exact value, chosen to be $\rho_{x,r} = 60\%$, whereas for the model with $\Omega(t) = \bar{\Omega}$ a different correlation pattern is observed. For the latter model, initially the correlation is significantly higher than 60%, and it decreases in time. These results show that a constant $\bar{\Omega}$ in the model with $\Delta = 0$ may give an average correlation close to the exact value, although the instantaneous correlation is not stable in time.

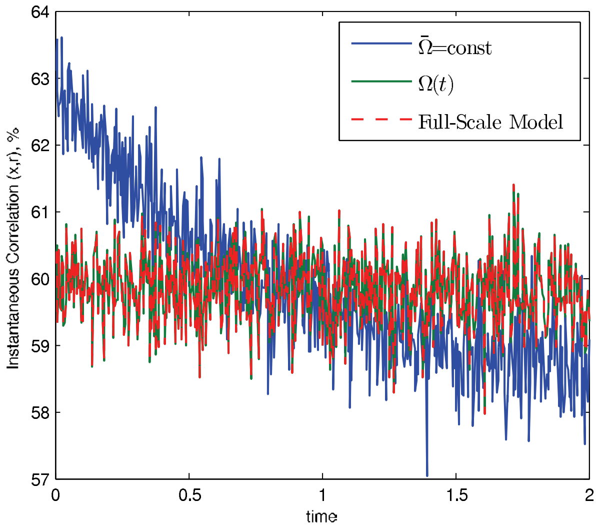


Figure 2. *The instantaneous correlations for different models. The blue line represents the model with $\Delta = 0$ with constant $\bar{\Omega}$, the dotted-red line corresponds to the full-scale HHW model, and the green line corresponds to the model with time-dependent $\Omega(t)$. Maturity is chosen to $\tau = 2$ years.*

The assumptions of constant $\bar{\Omega}$ and $\Delta = 0$ also have an impact on the corresponding pricing PDE. With the Feynman–Kac theorem the corresponding PDE is given by

$$0 = \frac{\partial \phi}{\partial t} + \left[r - \frac{1}{2}\left(v + r^{2p}\bar{\Omega}^2\right)\right] \frac{\partial \phi}{\partial x} + \kappa(\bar{v} - v)\frac{\partial \phi}{\partial v} + \lambda(\theta(t) - r)\frac{\partial \phi}{\partial r} + \frac{1}{2}\left(v + r^{2p}\bar{\Omega}^2\right)\frac{\partial^2 \phi}{\partial x^2}$$

$$+\frac{1}{2}\gamma^2 v \frac{\partial^2 \phi}{\partial v^2} + \frac{1}{2}\eta^2 r^{2p}\frac{\partial^2 \phi}{\partial r^2} + \hat{\rho}_{x,v}\gamma v \frac{\partial^2 \phi}{\partial x \partial v} + \eta \bar{\Omega} r^{2p} \frac{\partial^2 \phi}{\partial x \partial r} - r\phi, \tag{3.21}$$

with the same terminal condition as for (2.18). The assumption of constant $\bar{\Omega}$ and $\Delta = 0$ gives rise to additional terms in the convection and diffusion parts of PDE (3.21).

By means of a numerical experiment, we check the accuracy of the model with $\Delta = 0$ and determine whether the model approximates the full-scale HHW hybrid model sufficiently well. We consider here the following set of parameters: $S(0) = 1$, $\kappa = 2$, $v(0) = \bar{v} = 0.05$, $\gamma = 0.1$, $\lambda = 1.2$, $r(0) = \theta = 0.05$, $\eta = 0.01$, and correlation $\rho_{x,v} = -40\%$. In the simulation we choose two different values for correlation $\rho_{x,r} = \{30\%, 50\%\}$.

We compare the following three models: the full-scale HHW model (with MC simulation), the model with $\Delta = 0$, and our approximation for $\mathbf{\Sigma}_{(1,3)}$ in PDE (2.18) with the projection according to (3.1).

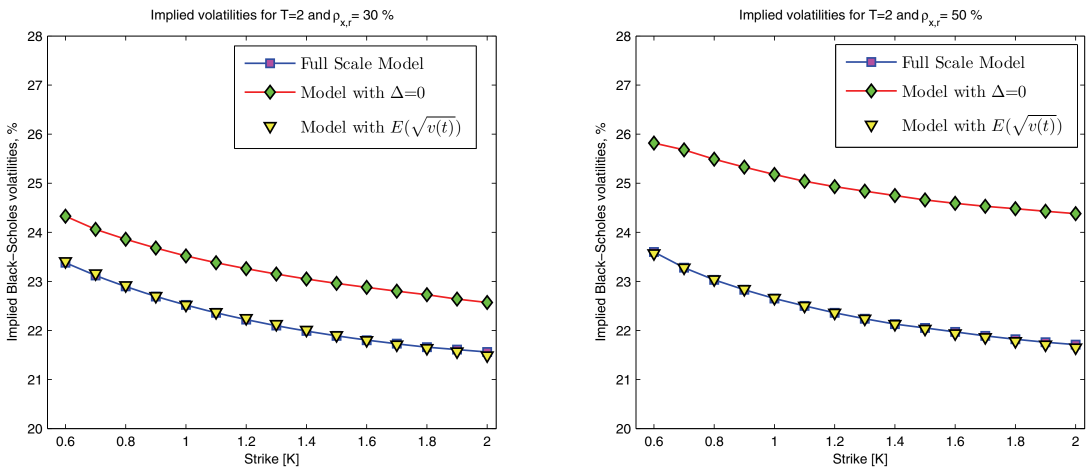


Figure 3. *The implied Black–Scholes volatilities for the full-scale Heston model and two approximations: deterministic approach (model* (2.18) *with* (3.1)*), and model with* $\Delta = 0$ *(model* (3.21)*).*

In Figure 3 the implied volatilities obtained are compared. The model with $\Delta = 0$ in (3.21) does not provide a satisfactory fit to the full-scale HHW model, whereas the implied volatilities obtained with the deterministic hybrid approximation compare very well (they essentially overlap) with the full-scale reference results; see Figure 3. The *volatility compensator* Δ, as defined in Lemma 2.1, cannot be neglected when approximating the full-scale HHW model, as was stated in Lemma 3.4.

3.2. ChF for the H1-HW model. We derive a ChF for the HHW hybrid model given in (2.18). For $p = 0$, the nonaffine term, $\mathbf{\Sigma}_{(1,3)}$, in matrix (2.18) equals $\mathbf{\Sigma}_{(1,3)} = \eta\rho_{x,r}\sqrt{v(t)}$ and will be approximated by $\mathbf{\Sigma}_{(1,3)} \approx \eta\rho_{x,r}\mathbb{E}(\sqrt{v(t)})$.

We assume here that the term structure for the interest rate $\theta(t)$ is constant: $\theta(t) = \theta$. A generalization can be found in [8].

According to [13], the discounted ChF for the H1-HW model is of the following form:

$$\phi_{\text{H1-HW}}(u, \mathbf{X}(t), \tau) = \exp\left(A(u,\tau) + B(u,\tau)x(t) + C(u,\tau)r(t) + D(u,\tau)v(t)\right), \tag{3.22}$$

with final conditions $A(u,0) = 0$, $B(u,0) = iu$, $C(u,0) = 0$, $D(u,0) = 0$, and $\tau := T - t$.

The ChF for the H1-HW model can be derived in closed form, with the help of the following lemmas.

Lemma 3.5 (ODEs related to the H1-HW model). *The functions* $B(u,\tau) =: B(\tau)$, $C(u,\tau) =: C(\tau)$, $D(u,\tau) =: D(\tau)$, *and* $A(u,\tau) =: A(\tau)$ *for* $u \in \mathbb{R}$ *and* $\tau \geq 0$ *in* (3.22) *for the H1-HW model satisfy the following system of ODEs:*

$$\begin{aligned}
&B'(\tau) = 0, \quad B(u,0) = iu,\\
&C'(\tau) = -1 - \lambda C(\tau) + B(\tau), \quad C(u,0) = 0,\\
&D'(\tau) = B(\tau)(B(\tau)-1)/2 + \left(\gamma\rho_{x,v}B(\tau) - \kappa\right)D(\tau) + \gamma^2 D^2(\tau)/2, \quad D(u,0) = 0,\\
&A'(\tau) = \lambda\theta C(\tau) + \kappa\bar{v}D(\tau) + \eta^2 C^2(\tau)/2 + \eta\rho_{x,r}\mathbb{E}(\sqrt{v(t)})B(\tau)C(\tau), \quad A(u,0) = 0,
\end{aligned}$$

with $\tau = T - t$, and where κ, λ, and θ and η, $\rho_{x,r}$, and $\rho_{x,v}$ correspond to the parameters in the HHW model (2.3).

Proof. The proof can be found in section A.1. ∎

The following lemma gives the closed-form solution for the functions $B(u,\tau)$, $C(u,\tau)$, $D(u,\tau)$, and $A(u,\tau)$ in (3.22).

Lemma 3.6 (ChF for the H1-HW model). *The solution of the ODE system in Lemma 3.5 is given by*

$$B(u,\tau) = iu, \tag{3.23}$$
$$C(u,\tau) = (iu-1)\lambda^{-1}(1-\mathrm{e}^{-\lambda\tau}), \tag{3.24}$$
$$D(u,\tau) = \frac{1-\mathrm{e}^{-D_1\tau}}{\gamma^2\left(1-g\mathrm{e}^{-D_1\tau}\right)}\left(\kappa - \gamma\rho_{x,v}iu - D_1\right), \tag{3.25}$$
$$A(u,\tau) = \lambda\theta I_1(\tau) + \kappa\bar{v}I_2(\tau) + \frac{1}{2}\eta^2 I_3(\tau) + \eta\rho_{x,r}I_4(\tau), \tag{3.26}$$

with $D_1 = \sqrt{(\gamma\rho_{x,v}iu-\kappa)^2 - \gamma^2 iu(iu-1)}$, and where $g = \frac{\kappa-\gamma\rho_{x,v}iu-D_1}{\kappa-\gamma\rho_{x,v}iu+D_1}$, κ, θ, λ, and γ are as in (2.10).

The integrals $I_1(\tau)$, $I_2(\tau)$, and $I_3(\tau)$ admit an analytic solution, and $I_4(\tau)$ admits a semianalytic solution:

$$\begin{aligned}
I_1(\tau) &= \frac{1}{\lambda}(iu-1)\left(\tau + \frac{1}{\lambda}(\mathrm{e}^{-\lambda\tau}-1)\right),\\
I_2(\tau) &= \frac{\tau}{\gamma^2}\left(\kappa-\gamma\rho_{x,v}iu-D_1\right) - \frac{2}{\gamma^2}\log\left(\frac{1-g\mathrm{e}^{-D_1\tau}}{1-g}\right),\\
I_3(\tau) &= \frac{1}{2\lambda^3}(i+u)^2\left(3+\mathrm{e}^{-2\lambda\tau}-4\mathrm{e}^{-\lambda\tau}-2\lambda\tau\right),\\
I_4(\tau) &= iu\int_0^\tau \mathbb{E}(\sqrt{v(T-s)})C(u,s)\mathrm{d}s\\
&= -\frac{1}{\lambda}(iu+u^2)\int_0^\tau \mathbb{E}(\sqrt{v(T-s)})\left(1-\mathrm{e}^{-\lambda s}\right)\mathrm{d}s.
\end{aligned}$$

Proof. The proof can be found in section A.2. ∎

Note that by taking $\mathbb{E}(\sqrt{v(T-s)}) \approx a + b\mathrm{e}^{-c(T-s)}$, with a, b, and c as given in (3.15), we obtain a closed-form expression:

$$I_4(\tau) = -\frac{1}{\lambda}(iu+u^2)\left[\frac{b}{c}\left(\mathrm{e}^{-ct}-\mathrm{e}^{-cT}\right) + a\tau + \frac{a}{\lambda}\left(\mathrm{e}^{-\lambda\tau}-1\right) + \frac{b}{c-\lambda}\mathrm{e}^{-cT}\left(1-e^{-\tau(\lambda-c)}\right)\right].$$

In Appendix B we present the generalization to a *full matrix of nonzero correlations* between the processes.

4. Stochastic approximation for hybrid models. In the previous section a rather straightforward way to approximate the nonaffine elements in the instantaneous covariance matrix was presented. Here, we model those elements alternatively by stochastic processes and call the resulting approximate model H2-HW (HHW model-2).

4.1. Stochastic approach: The H2-HW model. In the result below an approximation for finite time t and a nonzero centrality parameter is presented.

Result 4.1 (normal approximation for $\sqrt{v(t)}$, for $0 < t < \infty$). *For any time $t < \infty$, the square root of $v(t)$ in (2.9) can be approximated by*

$$\sqrt{v(t)} \approx \mathcal{N}\left(\sqrt{c(t)(\lambda(t)-1)+c(t)d+\frac{c(t)d}{2(d+\lambda(t))}}, c(t)-\frac{c(t)d}{2(d+\lambda(t))}\right), \tag{4.1}$$

with $c(t)$, d, and $\lambda(t)$ from (3.4). Moreover, for a fixed value of x in the cumulative distribution function $F_{\sqrt{v(t)}}(x)$ and a fixed value for parameter d, the error is of order $\mathcal{O}(\lambda^2(t))$ for $\lambda(t) \to 0$ and $\mathcal{O}(\lambda(t)^{-\frac{1}{2}})$ for $\lambda(t) \to \infty$.

To show the validity of the approximations presented above, we follow Patnaik in [34], who found that an accurate approximation for the noncentral chi-square distribution, $\chi_d^2(\lambda(t))$, can be obtained by an approximation with a centralized chi-square distribution, i.e.,

$$\chi^2(d, \lambda(t)) \approx a(t)\chi^2(f(t)), \tag{4.2}$$

with $a(t)$ and $f(t)$ in (4.2) chosen so that the first two moments match, i.e.,

$$a(t) = \frac{d+2\lambda(t)}{d+\lambda(t)}, \quad f(t) = d + \frac{\lambda^2(t)}{d+2\lambda(t)}. \tag{4.3}$$

It was shown in [11, 9] that, for a given time $t > 0$, $v(t)$ is distributed as $c(t)$ times a noncentral chi-squared random variable, $\chi^2(d, \lambda(t))$, with degrees of freedom parameter d and noncentrality parameter $\lambda(t)$, i.e., $v(t) = c(t)\chi^2\left(d, \lambda(t)\right),\ t > 0$. By combining this with (4.2) we have

$$\sqrt{v(t)} \approx \sqrt{c(t)}\sqrt{a(t)\chi^2(f(t))}. \tag{4.4}$$

Now, we use a result by Fisher [17] that for a given central chi-square random variable, $\chi^2(d)$, the expression $\sqrt{2\chi^2(d)}$ is approximately normally distributed with mean $\sqrt{2d-1}$ and unit variance, i.e.,

$$F_{\chi^2(d)}(x) \approx \Phi\left(\sqrt{2x} - \sqrt{2d-1}\right), \tag{4.5}$$

which implies

$$\sqrt{v(t)} \approx \mathcal{N}\left(\sqrt{\left(f(t)-\frac{1}{2}\right)c(t)a(t)}, \frac{1}{2}c(t)a(t)\right). \tag{4.6}$$

The order of this approximation can be found in [26].

Remark 3. Also in [34] it was indicated that the normal approximation resembles the noncentral chi-square distribution very well for either a large number of degrees of freedom, d, or a large noncentrality, $\lambda(t)$. For $t \to 0$, the noncentrality parameter, $\lambda(t)$, tends to infinity. Therefore, accurate approximations are expected.

In the case of long maturities, the noncentrality parameter converges to 0, which may give an inaccurate approximation. In this case, satisfactory results depend on the size of the degrees of freedom parameter d. It is clear that d in (3.4) is directly related to the Feller condition. In practical applications, however, $2\kappa\bar{v}$ is often smaller than γ^2. In the numerical experiments to follow, we will study the impact of *not* satisfying the Feller condition.

In Result 4.1 we have shown that $\sqrt{v(t)}$ can be well approximated by a normally distributed random variable. As the application of Itô's lemma to find the dynamics for $\sqrt{v(t)}$ is not allowed (the square-root process is not twice differentiable at the origin [25]), we construct here a stochastic process, $\xi(t)$, so that *equality in distribution* holds, i.e., $\xi(t) \overset{\mathrm{d}}{\approx} \sqrt{v(t)}$. Since a normal random variable is completely described by the first two moments, we need to ensure that $\mathbb{E}(\xi(t)) = \mathbb{E}(\sqrt{v(t)})$ and $\mathbb{V}\mathrm{ar}(\xi(t)) = \mathbb{V}\mathrm{ar}(\sqrt{v(t)})$. For this purpose we propose the following dynamics:

$$\mathrm{d}\xi(t) = \mu^{\xi}(t)\mathrm{d}t + \psi^{\xi}(t)\mathrm{d}W_v(t), \quad \xi(0) = \sqrt{v(0)}, \tag{4.7}$$

with some deterministic, time-dependent functions $\mu^{\xi}(t)$ and $\psi^{\xi}(t)$, determined so that the first two moments match. By moment matching, the unknown functions $\mu^{\xi}(t)$ and $\psi^{\xi}(t)$ in (4.7) read as

$$\mu^{\xi}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}(\sqrt{v(t)}), \quad \psi^{\xi}(t) = \sqrt{\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{V}\mathrm{ar}(\sqrt{v(t)})}. \tag{4.8}$$

Using the results from Lemma 3.1, the expectation, $\mathbb{E}(\sqrt{v(t)})$, and the variance, $\mathbb{V}\mathrm{ar}(\sqrt{v(t)})$, can be derived:

$$\begin{aligned}
\mu^{\xi}(t) &= \frac{1}{2\sqrt{2}}\frac{\Gamma\left(\frac{1+d}{2}\right)}{\sqrt{c(t)}}\left[{}_1\widetilde{F}_1\left(-\frac{1}{2},\frac{d}{2},-\frac{\lambda(t)}{2}\right)\frac{1}{2}\gamma^2\mathrm{e}^{-\kappa t}\right.\\
&\quad \left. + {}_1\widetilde{F}_1\left(\frac{1}{2},\frac{2+d}{2},-\frac{\lambda(t)}{2}\right)\frac{v(0)\kappa}{1-\mathrm{e}^{\kappa t}}\right],\\
\psi^{\xi}(t) &= \left(\kappa(\bar{v}-v(0))\mathrm{e}^{-\kappa t} - 2\mathbb{E}(\sqrt{v(t)})\,\mu^{\xi}(t)\right)^{\frac{1}{2}}.
\end{aligned} \tag{4.9}$$

Here, $\mathbb{E}(\sqrt{v(t)})$ and d, $c(t)$, and $\lambda(t)$ are as in (3.2), and the regularized hypergeometric function ${}_1\widetilde{F}_1(a;b;z) =: {}_1F_1(a;b;z)/\Gamma(b)$.

The expressions for $\mu^{\xi}(t)$ and $\psi^{\xi}(t)$ in (4.9) are exact. However, since those expressions are not cheap to compute, one can find suitable approximations based on the results in Result 3.2, which are, however, not guaranteed to be well defined for all sets of parameters.

Since the approximate hybrid models are to be used for the calibration to European-style options (with one terminal payment), we do not need pathwise equality between processes $\xi(t)$ and $\sqrt{v(t)}$; only equality in terminal distribution is needed.

Remark 4. In section 4.1 we projected $\sqrt{v(t)}$ onto a normal process, $\xi(t)$. As is common with approximations by normal processes (a nonnegative random variable is projected onto another variable $\in \mathbb{R}$), this approximation comes with an error (as we indicated in Result 4.1). During stress testing, examples of which are presented in section 4.3 and in Appendix C, we did not encounter any problems with this approximation. Typically, the stochastic approximation is somewhat more accurate than the deterministic approach (which is not based on a normal approximation).[4]

4.2. ChF for the H2-HW model. We now use the (stochastic) approximation for the term $\mathbf{\Sigma}_{(1,3)}$, with the process $\mathrm{d}\xi(t)$ given by (4.7), and the time-dependent functions $\mu^{\xi}(t)$ and $\psi^{\xi}(t)$ as in (4.9).

This approximation gives rise to an extension of the three-dimensional space variable $\mathbf{X}(t) = [S(t), v(t), r(t)]^{\mathrm{T}}$ to a four-dimensional space $\widetilde{\mathbf{X}}(t) = [S(t), v(t), r(t), \xi(t)]^{\mathrm{T}}$, with the following system of SDEs:

$$
\left\{
\begin{array}{rrrl}
\mathrm{d}S(t)/S(t) = & r(t)\mathrm{d}t + & \sqrt{v(t)}\mathrm{d}W_x(t), & S(0) > 0, \\
\mathrm{d}v(t) = & \kappa(\bar{v} - v(t))\mathrm{d}t + & \gamma\sqrt{v(t)}\mathrm{d}W_v(t), & v(0) > 0, \\
\mathrm{d}r(t) = & \lambda(\theta(t) - r(t))\mathrm{d}t + & \eta\mathrm{d}W_r(t), & r(0) > 0, \\
\mathrm{d}\xi(t) = & \mu^{\xi}(t)\mathrm{d}t + & \psi^{\xi}(t)\mathrm{d}W_v(t), & \xi(0) = \sqrt{v(0)},
\end{array}
\right. \tag{4.10}
$$

where

$$
\left\{
\begin{array}{rcl}
\mathrm{d}W_x(t)\mathrm{d}W_v(t) & = & \rho_{x,v}\mathrm{d}t, \\
\mathrm{d}W_x(t)\mathrm{d}W_r(t) & = & \rho_{x,r}\mathrm{d}t, \\
\mathrm{d}W_v(t)\mathrm{d}W_r(t) & = & 0,
\end{array}
\right. \tag{4.11}
$$

with $\sqrt{v(t)} \approx \xi(t)$ and $\mu^{\xi}(t)$,$\psi^{\xi}(t)$ as defined in (4.9).

By taking the log-transform, $x(t) = \log S(t)$, in the model above all the drift terms are linear, and the symmetric instantaneous covariance matrix, with $\xi(t) \approx \sqrt{v(t)}$, is given by

$$
\widetilde{\mathbf{\Sigma}} = \left[
\begin{array}{cccc}
v(t) & \gamma\rho_{x,v}v(t) & \rho_{x,r}\eta\xi(t) & \rho_{x,v}\psi^{\xi}(t)\xi(t) \\
* & \gamma^2 v(t) & 0 & \gamma\psi^{\xi}(t)\xi(t) \\
* & * & \eta^2 & 0 \\
* & * & * & \left(\psi^{\xi}(t)\right)^2
\end{array}
\right], \tag{4.12}
$$

which, since $\psi^{\xi}(t)$ is a deterministic time-dependent function, is now affine.

Since the system of SDEs (4.10) is affine, we derive the corresponding ChF:

$$
\phi_{\text{H2-HW}}(u, \mathbf{X}(t), \tau) = \exp\left(A(u,\tau) + B(u,\tau)x(t) + C(u,\tau)r(t) + D(u,\tau)v(t) + E(u,\tau)\xi(t)\right), \tag{4.13}
$$

with final conditions $\phi_{\text{H2-HW}}(u, \mathbf{X}(T), 0) = \exp(iux(T))$ and $\xi(t) \approx \sqrt{v(t)}$.

The functions $A(u,\tau)$, $B(u,\tau)$, $C(u,\tau)$, $D(u,\tau)$, and $E(u,\tau)$ satisfy the complex-valued ODEs given by the following lemma.

Lemma 4.2 (ODEs related to the H2-HW model). *The functions $B(u,\tau) =: B(\tau)$, $C(u,\tau) := C(\tau)$, $D(u,\tau) =: D(\tau)$, $E(u,\tau) =: E(\tau)$, and $A(u,\tau) =: A(\tau)$ for $u \in \mathbb{R}$*

[4] The method by Antonov from [4, 5] is also not based on normal approximations.

and $\tau = T - t > 0$ *in* (4.13) *satisfy*

$$
\begin{aligned}
&B'(\tau) = 0, \quad B(u,0) = iu,\\
&C'(\tau) = -1 + B(\tau) - \lambda C(\tau), \quad C(u,0) = 0,\\
&D'(\tau) = (B(\tau) - 1)\,B(\tau)/2 + (\gamma\rho_{x,v}B(\tau) - \kappa)\,D(\tau) + \gamma^2 D^2(\tau)/2, \quad D(u,0) = 0,\\
&E'(\tau) = \rho_{x,r}\eta B(\tau)C(\tau) + \psi^{\xi}(t)\rho_{x,v}B(\tau)E(\tau) + \gamma\psi^{\xi}(t)D(\tau)E(\tau), \quad E(u,0) = 0,\\
&A'(\tau) = \kappa\bar{v}D(\tau) + \lambda\theta C(\tau) + \mu^{\xi}(t)E(\tau) + \eta^2 C^2(\tau)/2 + \left(\psi^{\xi}(t)\right)^2 E^2(\tau)/2, \quad A(u,0) = 0,
\end{aligned}
$$

with $\mu^{\xi}(t)$, $\psi^{\xi}(t)$ *as given in* (4.9).

Proof. The proof is very similar to the proof of Lemma 3.5. ■

Solutions to the ODEs for $B(u,\tau)$, $C(u,\tau)$, and $D(u,\tau)$ can be found in Lemma 3.6, where the deterministic linearization was applied.

Note that the remaining two functions, $E(u,\tau)$ and $A(u,\tau)$, contain the rather complicated functions $\mu^{\xi}(t)$ and $\psi^{\xi}(t)$. We leave these equations to be solved numerically by a basic ODE routine.

A detailed analysis of the properties of the ChF will be studied in a followup article with theoretical research.

4.3. Numerical experiment. Here we check the performance of the deterministic (section 3.2) and the stochastic (section 4.2) approximations to the full-scale HHW model in terms of differences in implied volatilities. The HHW benchmark prices were obtained by MC simulation, as in [2].

In Table 1 we present the errors for Black–Scholes implied volatilities, $\epsilon(\rho_{x,r})$, for different correlations between the stock, $S(t)$, and the short rate, $r(t)$, and different strikes. We show results for a maturity of 10 years, $\tau = 10$, and for parameters that do not satisfy the Feller condition.[5]

Both approximations give very similar, highly accurate results for low correlations, $\rho_{x,r}$. This is different for high values of $\rho_{x,r}$. The deterministic approach generates somewhat more bias for high strikes, whereas the stochastic approach is essentially bias-free. The errors presented in Table 1 depend on the size of the volatility parameter of the interest rate process, η. For very low volatility, the two approximations provide a similar level of accuracy. As the volatility of the short rate process increases, a higher accuracy is expected for the stochastic approximation.

Calibration results will be presented in section 6. The performance of the methods developed is also presented in Appendix C, where our schemes are compared to the Markovian projection method [5].

5. HCIR hybrid model. We also present the ChF for an HCIR hybrid model, $p = 1/2$ in (2.3), which is more involved than the HW-based hybrid models. In the HCIR model the nonaffine term is given in (2.17). Again we use two approximations to obtain the ChF. In the first model, H1-CIR, we use the deterministic setup, and for the second model, H2-CIR, we determine the stochastic approximation.

[5] For short maturities, $\tau < 10$, and for model parameters for which the Feller condition is satisfied, we did not find any significant differences between the two approximations and the full-scale model.

Table 1

The implied volatilities and errors for the deterministic approximation (Approx 1) from (2.18) with approximation (3.2) and the stochastic approximation (Approx 2) from section 4.1 of the HHW model compared to the MC simulation performed with $20T$ steps and 100.000 paths. The error is defined as a difference between reference implied volatilities and the approximations. The parameters were chosen as $\kappa = 0.3$, $\gamma = 0.6$, $v(0) = \bar{v} = 0.05$, $\lambda = 0.01$, $r(0) = \theta = 0.02$, $\eta = 0.01$, and $S(0) = 100$ and the correlations as $\rho_{x,v} = -30\%$ and $\rho_{x,r} \in \{20\%, 60\%\}$. Numbers in brackets indicate standard deviations.

$\rho_{x,r}$	Strike	MC Imp. vol. [%]	Approx 1	Approx 2	Err. 1	Err. 2
	40	26.26 (0.22)	25.87	25.99	0.39	0.27
	80	20.07 (0.22)	20.03	20.02	0.04	0.05
20%	100	18.43 (0.24)	18.55	18.36	−0.12	0.07
	120	17.51 (0.20)	17.74	17.42	−0.23	0.09
	180	17.40 (0.22)	17.55	17.36	−0.15	0.04
	40	26.27 (0.14)	26.21	26.61	0.06	−0.34
	80	20.59 (0.11)	21.00	20.91	−0.41	−0.32
60%	100	19.11 (0.10)	19.84	19.22	−0.72	−0.10
	120	18.31 (0.10)	19.21	18.18	−0.90	0.13
	180	18.25 (0.11)	18.92	18.34	−0.67	−0.09

5.1. ChF for the H1-CIR model. The dynamics for the stock, $S(t)$, in the HCIR model read as

$$(5.1)\qquad \begin{cases} \mathrm{d}S(t)/S(t) = r(t)\mathrm{d}t + \sqrt{v(t)}\mathrm{d}W_x(t), & S(0) > 0, \\ \mathrm{d}v(t) = \kappa(\bar{v} - v(t))\mathrm{d}t + \gamma\sqrt{v(t)}\mathrm{d}W_v(t), & v(0) > 0, \\ \mathrm{d}r(t) = \lambda(\theta(t) - r(t))\mathrm{d}t + \eta\sqrt{r(t)}\mathrm{d}W_r(t), & r(0) > 0, \end{cases}$$

with $\mathrm{d}W_x(t)\mathrm{d}W_v(t) = \rho_{x,v}\mathrm{d}t$, $\mathrm{d}W_x(t)\mathrm{d}W_r(t) = \rho_{x,r}\mathrm{d}t$, and $\mathrm{d}W_v(t)\mathrm{d}W_r(t) = 0$.

Here, we assume that the nonaffine term in the pricing PDE (2.18), $\mathbf{\Sigma}_{(1,3)}$, in (2.17) can be approximated as

$$(5.2)\qquad \mathbf{\Sigma}_{(1,3)} \approx \eta\rho_{x,r}\mathbb{E}\left(\sqrt{r(t)}\sqrt{v(t)}\right) \stackrel{\perp\!\!\!\perp}{=} \eta\rho_{x,r}\mathbb{E}(\sqrt{r(t)})\mathbb{E}(\sqrt{v(t)}).$$

Since the processes involved are of the same type, the expectations in (5.2) can be determined as presented in section 3.1. For the log-stock, $x(t) = \log S(t)$, the ChF and the corresponding Riccati ODEs are defined as below:

$$(5.3)\qquad \phi_{\text{H1-CIR}}(u, \mathbf{X}(t), \tau) = \exp\left(A(u,\tau) + B(u,\tau)x(t) + C(u,\tau)r(t) + D(u,\tau)v(t)\right).$$

Lemma 5.1 (ODEs related to the H1-CIR model). *The functions $B(u,\tau) =: B(\tau)$, $C(u,\tau) =: C(\tau)$, $D(u,\tau) =: D(\tau)$, and $A(u,\tau) =: A(\tau)$ for $u \in \mathbb{R}$ and $\tau > 0$ in (5.3) satisfy*

$$B'(\tau) = 0, \quad B(u,0) = iu,$$
$$C'(\tau) = -1 + B(\tau) - \lambda C(\tau) + \eta^2 C^2(\tau)/2, \quad C(u,0) = 0,$$
$$D'(\tau) = (B(\tau) - 1)B(\tau)/2 + (\gamma\rho_{x,v}B(\tau) - \kappa)D(\tau) + \gamma^2 D^2(\tau)/2, \quad D(u,0) = 0,$$
$$(5.4)\qquad A'(\tau) = \kappa\bar{v}D(\tau) + \lambda\theta C(\tau) + \eta\rho_{x,r}\mathbb{E}(\sqrt{v(t)})\mathbb{E}(\sqrt{r(t)})B(\tau)C(\tau), \quad A(u,0) = 0,$$

with $\tau = T - t$, $\mathbb{E}(\sqrt{v(t)})$, *and* $\mathbb{E}(\sqrt{r(t)})$ *from Lemma* 3.1.

Proof. The proof is very similar to the proof of Lemma 3.5 in section A.1. ∎

Lemma 5.2 (solutions for the ChF coefficients of the H1-CIR model). *The solutions for the ODEs for* $B(u,\tau)$, $C(u,\tau)$, $D(u,\tau)$, *and* $A(u,\tau)$, *defined in Lemma* 5.1, *are given by*

$$B(u,\tau) = iu, \tag{5.5}$$

$$C(u,\tau) = \frac{1-\mathrm{e}^{-D_1\tau}}{\eta^2\left(1-G_1\mathrm{e}^{-D_1\tau}\right)}\left(\lambda - D_1\right), \tag{5.6}$$

$$D(u,\tau) = \frac{1-\mathrm{e}^{-D_2\tau}}{\gamma^2\left(1-G_2\mathrm{e}^{-D_2\tau}\right)}\left(\kappa - \gamma\rho_{x,v}iu - D_2\right), \tag{5.7}$$

and

$$A(u,\tau) = \int_0^\tau \Big(\kappa\bar{v}D(u,s) + \lambda\theta C(u,s) + \rho_{x,r}\eta iu\mathbb{E}(\sqrt{v(T-s)})\mathbb{E}(\sqrt{r(T-s)})C(u,s)\Big)\,\mathrm{d}s,$$

with $D_1 = \sqrt{\lambda^2 + 2\eta^2(1-iu)}$, $D_2 = \sqrt{\left(\gamma\rho_{x,v}iu - \kappa\right)^2 - (iu-1)iu\gamma^2}$, $G_1 = \frac{\lambda - D_1}{\lambda + D_1}$, *and* $G_2 = \frac{\kappa - \gamma\rho_{x,v}iu - D_2}{\kappa - \gamma\rho_{x,v}iu + D_2}$.

Proof. The proof is very similar to the proof of Lemma 3.6 in section A.2. ∎

The integral for $A(u,\tau)$ in Lemma 5.2 can be determined analytically only for constant approximations of the two expectations involved.

5.2. ChF for the H2-CIR model. As before, we aim to find an approximation of the instantaneous covariance matrix for which the affinity of the approximation model is obtained, but now with the stochastic approximation.

$\mathbf{\Sigma}_{(1,3)}$ now consists of two stochastic components, $\sqrt{v(t)}$ and $\sqrt{r(t)}$. We approximate both and obtain

$$\mathbf{\Sigma}_{(1,3)} \approx \widetilde{\mathbf{\Sigma}}_{(1,3)} = \rho_{x,r}\eta\xi(t)R(t), \quad R(t) \approx \sqrt{r(t)}, \quad \xi(t) \approx \sqrt{v(t)}. \tag{5.8}$$

This form, based on the product of two random variables, is not affine. To linearize (5.8) we need to specify the joint dynamics, $\mathrm{d}(\sqrt{v(t)}\sqrt{r(t)})$. If we assume that the dynamics for $\mathrm{d}(\sqrt{v(t)})$ and $\mathrm{d}(\sqrt{r(t)})$ can be approximated by normally distributed processes, we find, by Itô's lemma, that the dynamics of $z(t) = \xi(t)R(t)$ are given by

$$\mathrm{d}z(t) = \Big(\mu^R(t)\xi(t) + \mu^\xi(t)R(t)\Big)\,\mathrm{d}t + \psi^\xi(t)R(t)\mathrm{d}W_v(t) + \psi^R(t)\xi(t)\mathrm{d}W_r(t). \tag{5.9}$$

With three additional variables, $\xi(t)$, $R(t)$, and $z(t)$, the state vector $\mathbf{X}(t)$, with log-stock process $x(t) = \log S(t)$, is expanded to $\mathbf{X}(t) = [x(t), v(t), r(t), \xi(t), R(t), z(t)]^{\mathrm{T}}$, with the following corresponding system of SDEs:

$$\begin{cases} \mathrm{d}x(t) = & \left(r(t) - \frac{1}{2}v(t)\right)\mathrm{d}t + & \sqrt{v(t)}\mathrm{d}W_x(t), & x(0) = \log(S(0)), \\ \mathrm{d}v(t) = & \kappa(\bar{v} - v(t))\mathrm{d}t + & \gamma\sqrt{v(t)}\mathrm{d}W_v(t), & v(0) > 0, \\ \mathrm{d}r(t) = & \lambda(\theta(t) - r(t))\mathrm{d}t + & \eta\sqrt{r(t)}\mathrm{d}W_r(t), & r(0) > 0, \end{cases} \tag{5.10}$$

with the linearizing variables $\xi(t)$, $R(t)$, and $z(t)$ given by

(5.11)
$$\begin{aligned}
\mathrm{d}\xi(t) &= & \mu^{\xi}(t)\mathrm{d}t + && \psi^{\xi}(t)\mathrm{d}W_v(t), && \xi(0) = \sqrt{v(0)},\\
\mathrm{d}R(t) &= & \mu^{R}(t)\mathrm{d}t + && \psi^{R}(t)\mathrm{d}W_r(t), && R(0) = \sqrt{r(0)},\\
\mathrm{d}z(t) &= & \left(\mu^{R}(t)\xi(t) + \mu^{\xi}(t)R(t)\right)\mathrm{d}t + && \psi^{\xi}(t)\sqrt{r(t)}\mathrm{d}W_v(t) + && \psi^{R}(t)\sqrt{v(t)}\mathrm{d}W_r(t),
\end{aligned}$$

with $z(0) = \sqrt{r(0)}\sqrt{v(0)}$, $\xi(t) \approx \sqrt{v(t)}$, $R(t) \approx \sqrt{r(t)}$, $z(t) \approx \sqrt{v(t)}\sqrt{r(t)}$, and the other parameters as in (2.3). The symmetric instantaneous covariance matrix reads as

$$(5.12)\qquad \widetilde{\mathbf{\Sigma}}^* = \begin{bmatrix}
v(t) & \rho_{x,v}\gamma v(t) & \rho_{x,r}\eta z(t) & \rho_{x,v}\psi^{\xi}(t)\xi(t) & \rho_{x,r}\psi^{R}(t)\xi(t) & s_1(t)\\
* & \gamma^2 v(t) & 0 & \psi^{\xi}(t)\gamma\xi(t) & 0 & \gamma\psi^{\xi}(t)z(t)\\
* & * & \eta^2 r(t) & 0 & \psi^{R}(t)\eta R(t) & \eta\psi^{R}(t)z(t)\\
* & * & * & (\psi^{\xi}(t))^2 & 0 & (\psi^{\xi}(t))^2 R(t)\\
* & * & * & * & (\psi^{R}(t))^2 & (\psi^{R}(t))^2\xi(t)\\
* & * & * & * & * & s_2(t)
\end{bmatrix},$$

with $s_1(t) = \rho_{x,v}\psi^{\xi}(t)z(t) + \rho_{x,r}\psi^{R}(t)v(t)$ and $s_2(t) = (\psi^{\xi}(t))^2 r(t) + (\psi^{R}(t))^2 v(t)$.

Since $\psi^{\xi}(t)$ and $\psi^{R}(t)$ are deterministic time-dependent functions, the approximate H2-CIR model is now affine, and we can derive the corresponding ChF:

$$(5.13)\qquad \begin{aligned}
\phi_{\text{H2-CIR}}(u, \mathbf{X}(t), \tau) = \exp\big(&A(u,\tau) + B(u,\tau)x(t) + C(u,\tau)r(t) + D(u,\tau)v(t)\\
&+ E(u,\tau)\xi(t) + F(u,\tau)R(t) + G(u,\tau)z(t)\big),
\end{aligned}$$

with $\xi(t) = \sqrt{v(t)}$, $R(t) = \sqrt{r(t)}$, and $z(t) = \sqrt{v(t)}\sqrt{r(t)}$, and where the functions $A(u,\tau)$, $B(u,\tau)$, $C(u,\tau)$, $D(u,\tau)$, $E(u,\tau)$, $F(u,\tau)$, and $G(u,\tau)$ satisfy the ODEs given by the following lemma.

Lemma 5.3 (ODEs related to the H2-CIR model). *The functions $B(u,\tau) =: B(\tau)$, $C(u,\tau) =: C(\tau)$, $D(u,\tau) =: D(\tau)$, $E(u,\tau) =: E(\tau)$, $F(u,\tau) =: F(\tau)$, $G(u,\tau) =: G(\tau)$, and $A(u,\tau) =: A(\tau)$ for $u \in \mathbb{R}$ and $\tau > 0$ in (5.13) satisfy*

$$\begin{aligned}
B'(\tau) &= 0,\\
C'(\tau) &= -1 + B(\tau) - \lambda C(\tau) + \eta^2 C^2(\tau)/2 + (\psi^{\xi}(t))^2 G^2(\tau)/2,\\
F'(\tau) &= \mu^{\xi}(t)G(\tau) + \psi^{R}(t)\eta C(\tau)G(\tau) + (\psi^{\xi}(t))^2 E(\tau)G(\tau),\\
G'(\tau) &= \eta\rho_{x,r}B(\tau)C(\tau) + \rho_{x,v}\psi^{\xi}(t)B(\tau)G(\tau) + \gamma\psi^{\xi}(t)D(\tau)G(\tau) + \eta\psi^{R}(t)C(\tau)G(\tau)
\end{aligned}$$

and

$$\begin{aligned}
D'(\tau) &= B(\tau)\left(B(\tau)-1\right)/2 - \kappa D(\tau) + \gamma\rho_{x,v}B(\tau)D(\tau) + \gamma^2 D^2(\tau)/2\\
&\quad + \rho_{x,r}\psi^{R}(t)B(\tau)G(\tau) + (\psi^{R}(t))^2 G^2(t)/2,\\
E'(\tau) &= \mu^{R}(t)G(\tau) + \psi^{\xi}(t)\rho_{x,v}B(\tau)E(\tau) + \gamma\psi^{\xi}(t)D(\tau)E(\tau)\\
&\quad + \rho_{x,r}\psi^{R}(t)B(\tau)F(\tau) + (\psi^{R}(t))^2 F(\tau)G(\tau),\\
A'(\tau) &= \kappa\bar{v}D(\tau) + \lambda\theta C(\tau) + \mu^{\xi}(t)E(\tau) + \mu^{R}(t)F(\tau)\\
&\quad + (\psi^{\xi}(t))^2 E^2(\tau)/2 + (\psi^{R}(\tau))^2 F^2(\tau)/2,
\end{aligned}$$

with the final conditions $B(u,0) = iu$, $C(u,0) = 0$, $D(u,0) = 0$, $E(u,0) = 0$, $F(u,0) = 0$, $G(u,0) = 0$, *and* $A(u,0) = 0$. *Parameters* $\mu^{\xi}(t)$, $\mu^{R}(t)$, $\psi^{\xi}(t)$, $\psi^{R}(t)$ *are specified in* (4.9), *and the remaining parameters are in* (5.10).

Proof. The proof is very similar to the proof of Lemma 3.5 in section A.1. ■

The system of the ODEs given in Lemma 5.3 is difficult to solve analytically. To find the solution we have used an explicit Runge–Kutta method [18, 27], ode45 from the MATLAB package. Numerical results are presented in the next subsection.

The extension of the H2-CIR model to the case of a full matrix of correlations is a trivial exercise.

5.3. Numerical experiment. We compare the performance of the approximations H1-CIR and H2-CIR with the full-scale HCIR model. As in the case of the HHW models, we have chosen here $T = 10$, and the model parameters are chosen so that the Feller condition does not hold. The results, presented in Table 2, are very satisfactory. Both approximation models, H1-CIR and H2-CIR, provide an error, $\epsilon(\rho_{x,r})$, for a call option within the confidence bounds. For higher correlation $\rho_{x,r}$, the error grows, but it is still small.

Table 2

The implied volatilities and errors for the deterministic approximation (Approx 1) from (2.18) *with approximation* (3.2) *and the stochastic approximation (Approx 2) from section* 4.1 *of the HCIR model compared to the MC simulation performed with* $20T$ *steps and* 100.000 *paths. The error is defined as a difference between the reference implied volatilities and the approximation. The parameters were chosen as* $\kappa = 0.3$, $\gamma = 0.6$, $v(0) = \bar{v} = 0.05$, $\lambda = 0.01$, $r(0) = \theta = 0.02$, $\eta = 0.01$, *and* $S(0) = 100$ *and the correlations as* $\rho_{x,v} = -30\%$ *and* $\rho_{x,r} \in \{20\%, 60\%\}$. *Numbers in brackets indicate standard deviations.*

$\rho_{x,r}$	Strike	MC Imp. vol. [%]	Approx 1	Approx 2	Err. 1	Err. 2
20%	40	25.66 (0.17)	25.68	25.74	−0.02	−0.08
	80	19.17 (0.15)	19.21	19.25	−0.04	−0.08
	100	17.10 (0.18)	17.19	17.09	−0.09	−0.01
	120	15.77 (0.17)	15.90	15.85	−0.14	−0.08
	180	15.84 (0.18)	15.90	15.86	−0.06	−0.02
60%	40	24.95 (0.14)	25.72	25.79	−0.77	−0.84
	80	18.93 (0.12)	19.32	19.32	−0.39	−0.39
	100	16.92 (0.13)	17.37	17.08	−0.44	−0.15
	120	15.60 (0.13)	16.17	15.93	−0.57	−0.32
	180	15.57 (0.14)	16.10	15.98	−0.53	−0.41

We also present the time needed for obtaining the plain vanilla option prices, with the ChFs H2-HW (section 4.2) and H2-CIR (section 5.2) based on the numerical solution for the system of Riccati ODEs. Table 3 shows that, although the ODEs in Lemma 5.3 need to be solved numerically, the time for obtaining European option prices, by the COS pricing method [16], is often less than 0.1 seconds. The pricing of the options by means of the COS method, a method based on Fourier cosine series expansions, was performed with a fixed number of 250 terms, which guaranteed highly accurate option prices (up to machine precision).

The tolerance for the ODE solves, by ode45 from MATLAB, is varied in the experiments shown in the table.

Table 3

Time in seconds for pricing a call option based on an explicit Runge–Kutta method combined with the COS method [16].

Model	Accuracy	Maturity				
		$\tau = 0.5$	$\tau = 1$	$\tau = 2$	$\tau = 5$	$\tau = 10$
H2-HW	10^{-2}	4.37e-2	4.80e-2	6.41e-2	7.49e-2	8.10e-2
	10^{-5}	5.32e-2	5.82e-2	8.05e-2	9.74e-2	1.21e-1
H2-CIR	10^{-2}	7.78e-2	7.80e-2	8.38e-2	8.48e-2	8.90e-2
	10^{-5}	8.33e-2	8.97e-2	1.05e-1	1.34e-1	1.62e-1

6. Calibration of the Heston hybrid models. Here, we evaluate the performance of the approximations H1-HW and H2-HW for the HHW hybrid model in a calibration setting.

We reference call option prices, based on synthetic data that is representative for the skew and smile patterns observed in real-life applications. For all models the simulation was performed with an a priori defined speed of mean reversion for the variance process, $\kappa = 0.3$ (which is set small on purpose). The calibration is performed here with constant correlation $\rho_{x,r} = 20\%$. In practice, this correlation can be obtained from historical data.

The calibration procedure is performed in two stages. First, the parameters for the short rate process are determined (independent of the equity part). In the second stage, the calibrated $r(t)$ is included in the Heston model, and the remaining parameters are determined. The parameters for the interest rate part are found to be $\lambda_{\mathrm{HW}} = 0.501$, $\eta_{\mathrm{HW}} = 0.005$, and $r(0) = 0.04$.

First, we also perform, as a benchmark, the calibration of the pure Heston model with constant interest rate; see Table 4. SSE stands stands for the "sum-squared error." We calibrate the models for different maturities, τ.

Table 4

Calibration results for the Heston stochastic volatility model with deterministic interest rate. The mean reversion parameter is $\kappa = 0.3$.

Model	γ	$\bar{v}$	$\rho_{x,v}$	$v(0)$	r	SSE
Heston ($\tau = 0.5$)	0.5992	0.0823	-58.32%	0.0407	0.04	4.9063E-4
Heston ($\tau = 10$)	0.6019	0.0828	-48.49%	0.0411	0.04	1.2182E-4

In Table 5 the calibration results for the HHW approximations, H1-HW and H2-HW, are presented. For both models a highly satisfactory fit is obtained, with a slightly better performance of the stochastic approximation H2-HW. For $\rho_{x,r} = 20\%$ the calibration procedure gives roughly the same sets of parameters for both models. When comparing the calibration results for HHW with those for the pure Heston model, we see that the inclusion of stochastic interest rates in the model results in a lower vol-vol parameter, γ, and a more negative correlation, $\rho_{x,v}$. The lower value of parameter γ can be explained by the additional volatility which comes from the interest rate process.

In Figure 4 the corresponding implied volatilities, for the full-scale model, for a short and a long maturity time ($\tau = 0.5y$ and $\tau = 10y$) are presented. The left-hand sides of the figure

Table 5

Calibration results for the H1-HW model from section 3.2, and the H2-HW model from section 4.2, with $\kappa = 0.3$ and correlation $\rho_{x,r} = 20\%$.

Model	τ	γ	$\bar{v}$	$\rho_{x,v}$	$v(0)$	SSE
H1-HW	$\tau = 0.5$	0.5840	0.0822	-60.06%	0.0407	4.4581E-4
	$\tau = 10$	0.4921	0.0826	-61.50%	0.0418	3.2912E-4
H2-HW	$\tau = 0.5$	0.5879	0.0930	-60.10%	0.0398	4.9677E-4
	$\tau = 10$	0.4884	0.0820	-60.72%	0.0421	8.5934E-5

present the implied volatilities and their errors for H1-HW and H2-HW. The related implied volatilities of the full-scale HHW model, with the parameters from H1-HW and H2-HW, are shown in the right-hand sides of the figure.

Both hybrid models perform very well. For long maturities a higher accuracy for the hybrid models compared to the plain Heston model can be observed.

7. Concluding remarks. In this article we have presented the extension of the Heston stochastic volatility equity model by stochastic interest rates. We have focused our attention on two hybrid models, the HHW and the HCIR models.

By approximations of the nonaffine terms in the corresponding instantaneous covariance matrix, we placed the approximation hybrid models in the framework of AD processes. The approximations in the models have been validated by comparing the implied volatilities to the full-scale hybrid models.

The approximations in the HHW and the HCIR models lead to highly efficient determination of the corresponding ChFs. The more sophisticated approximation is based on a transformation of the three-dimensional HCIR model to a six-dimensional representation.

The deterministic and the stochastic approaches for approximating the instantaneous covariance matrix of the hybrid model provide highly satisfactory approximations for prices for European options.

Appendix A. Proofs and details.

A.1. Proof of Lemma 3.5.

Proof. For a given state vector $X(t) = [x(t), r(t), v(t)]^{\mathrm{T}}$ and $\phi := \phi(u, X(t), t, T)$, we find the system of the ODEs satisfying the following pricing PDE:

$$
\begin{aligned}
0 = {} & \frac{\partial \phi}{\partial t} + \left(r - \frac{1}{2}v\right)\frac{\partial \phi}{\partial x} + \kappa(\bar{v} - v)\frac{\partial \phi}{\partial v} + \lambda(\theta(t) - r)\frac{\partial \phi}{\partial r} + \frac{1}{2}v\frac{\partial^2 \phi}{\partial x^2} \\
& + \frac{1}{2}\gamma^2 v\frac{\partial^2 \phi}{\partial v^2} + \frac{1}{2}\eta^2\frac{\partial^2 \phi}{\partial r^2} + \rho_{x,v}\gamma v\frac{\partial^2 \phi}{\partial x \partial v} + \eta\rho_{x,r}\mathbb{E}(\sqrt{v(t)})\frac{\partial^2 \phi}{\partial x \partial r} - r\phi,
\end{aligned} \tag{A.1}
$$

subject to terminal condition $\phi(u, \mathbf{X}(T), T, T) = \exp\left(iux(T)\right)$.

Since the PDE in (A.1) is affine, its solution is of the following form:

$$
\phi := \phi(u, X(t), t, T) = \exp\left(A(u,t,T) + B(u,t,T)x(t) + C(u,t,T)r(t) + D(u,t,T)v(t)\right).
$$

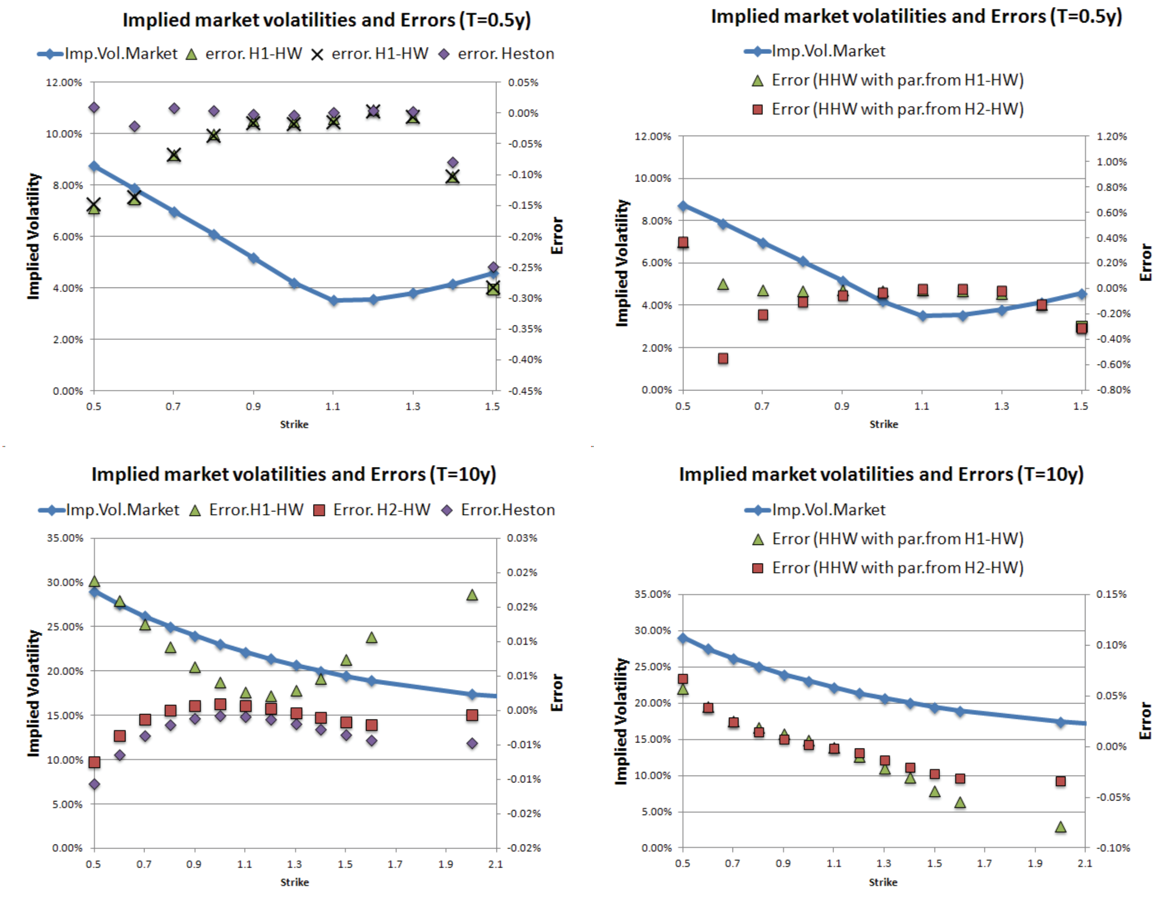


Figure 4. *For $\tau = 0.5$ and $\tau = 10$, $\rho_{x,r} = 20\%$, the implied Black–Scholes volatilities for Heston hybrid models are compared to the pure Heston model and a reference implied volatility curve. The left-hand graphs present the implied volatilities and errors for H1-HW and H2-HW. The implied volatilities for the full-scale HHW model, with the parameters from H1-HW and H2-HW, are in the right-hand figures.*

By setting $A := A(u,t,T)$, $B := B(u,t,T)$, $C := C(u,t,T)$, and $D := D(u,t,T)$, we find the following partial derivatives:

$$\frac{\partial\phi}{\partial t} = \phi\left(\frac{\partial A}{\partial t} + x(t)\frac{\partial B}{\partial t} + r(t)\frac{\partial C}{\partial t} + v(t)\frac{\partial D}{\partial t}\right), \tag{A.2}$$

$$\frac{\partial\phi}{\partial x} = B\phi, \quad \frac{\partial^2\phi}{\partial x^2} = B^2\phi, \quad \frac{\partial^2\phi}{\partial x\partial v} = BD\phi, \quad \frac{\partial^2\phi}{\partial x\partial r} = BC\phi, \tag{A.3}$$

$$\frac{\partial\phi}{\partial r} = C\phi, \quad \frac{\partial^2\phi}{\partial r^2} = C^2\phi, \tag{A.4}$$

$$\frac{\partial\phi}{\partial v} = D\phi, \quad \frac{\partial^2\phi}{\partial v^2} = D^2\phi. \tag{A.5}$$

By substitution, PDE (A.1) becomes

$$
\begin{aligned}
0 = {} & \frac{\partial A}{\partial t} + x\frac{\partial B}{\partial t} + r\frac{\partial C}{\partial t} + v\frac{\partial D}{\partial t} + \left(r - \frac{1}{2}v\right)B + \kappa(\bar{v} - v)D + \lambda(\theta(t) - r)C \\
& + \frac{1}{2}vB^2 + \frac{1}{2}\gamma^2 vD^2 + \frac{1}{2}\eta^2 C^2 + \rho_{x,v}\gamma vBD + \eta\rho_{x,r}\mathbb{E}(\sqrt{v(t)})BC - r.
\end{aligned} \tag{A.6}
$$

Now, by collecting the terms for $x(t)$, $r(t)$, and $v(t)$, we find the following set of ODEs:

$$\frac{\partial B}{\partial t} = 0, \tag{A.7}$$

$$\frac{\partial C}{\partial t} = -B + \lambda C + 1, \tag{A.8}$$

$$\frac{\partial D}{\partial t} = \frac{1}{2}B + \kappa D - \frac{1}{2}\gamma^2 D^2 - \rho_{x,v}\gamma BD - \frac{1}{2}B^2, \tag{A.9}$$

$$\frac{\partial A}{\partial t} = -\kappa\bar{v}D - \lambda\theta C - \frac{1}{2}\eta^2 C^2 - \rho_{x,r}\eta\mathbb{E}(\sqrt{v(t)})BC. \tag{A.10}$$

By setting $\tau = T - t$, the proof is finished. ■

A.2. Proof of Lemma 3.6.

Proof. Obviously, due to the final condition, $B(u,0) = iu$, we have $B(u,\tau) = iu$. For the second ODE, by multiplying both sides with $\mathrm{e}^{\lambda\tau}$, we get

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\mathrm{e}^{\lambda\tau}C(u,\tau)\right) = (iu-1)\mathrm{e}^{\lambda\tau}; \tag{A.11}$$

by integrating both sides and using the final condition, $C(u,0) = 0$, we find

$$C(u,\tau) = (iu-1)\lambda^{-1}\left(1 - \mathrm{e}^{-\lambda\tau}\right).$$

By setting $a = -\frac{1}{2}(u^2 + iu)$, $b = \gamma\rho_{x,v}iu - \kappa$, and $c = \frac{1}{2}\gamma^2$, the ODEs for $D(u,\tau)$ and $I_2(\tau)$ are given by the following Riccati-type equation:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}D(u,\tau) = a + bD(u,\tau) + cD^2(u,\tau), \quad D(u,0) = 0, \tag{A.12}$$

$$I_2(\tau) = \kappa\bar{v}\int_0^\tau D(u,s)\mathrm{d}s. \tag{A.13}$$

Equations (A.12) and (A.13) are of the same form as those in [22]. Their solutions are given by

$$D(u,\tau) = \frac{-b - D_1}{2c(1 - G\mathrm{e}^{-D_1\tau})}(1 - \mathrm{e}^{-D_1\tau}), \tag{A.14}$$

$$I_2(\tau) = \frac{1}{2c}\left((-b - D_1)\tau - 2\log\left(\frac{1 - G\mathrm{e}^{-D_1\tau}}{1 - G}\right)\right), \tag{A.15}$$

with $D_1 = \sqrt{b^2 - 4ac}$, $G = \frac{-b-D_1}{-b+D_1}$.

The evaluation of the integrals $I_1(\tau)$, $I_3(\tau)$, and $I_4(\tau)$ is straightforward. The proof is finished by appropriate substitutions. ∎

Appendix B. Hybrid model with full matrix of correlations. Similar to the approximation of the nonaffine terms in the instantaneous covariance matrix of the Heston hybrid model presented in section 3.1, we discuss here the inclusion of the additional correlation, $\rho_{r,v}$, between the interest rate, $r(t)$, and the stochastic variance, $v(t)$. We call the resulting model the HHW hybrid model-3 and denote it by H3-HW. For the state vector $\mathbf{X}(t) = [x(t), v(t), r(t)]^{\mathrm{T}}$, the H3-HW model has the following symmetric instantaneous covariance matrix:

$$\mathbf{\Sigma} := \sigma(\mathbf{X}(t))\sigma(\mathbf{X}(t))^{\mathrm{T}} = \begin{bmatrix} v(t) & \rho_{x,v}\gamma v(t) & \rho_{x,r}\eta\sqrt{v(t)} \\ * & \gamma^2 v(t) & \rho_{r,v}\gamma\eta\sqrt{v(t)} \\ * & * & \eta^2 \end{bmatrix}_{(3\times3)}. \tag{B.1}$$

The affinity issue arises in two terms of matrix (B.1), namely, in elements $(1,3)$ and $(2,3)$:

$$\mathbf{\Sigma}_{(1,3)} = \rho_{x,r}\eta\sqrt{v(t)}, \quad \mathbf{\Sigma}_{(2,3)} = \rho_{r,v}\gamma\eta\sqrt{v(t)}.$$

For completeness, we also present the associated Kolmogorov backward equation, which is now given by

$$\begin{aligned} 0 = {} & \frac{\partial\phi}{\partial t} + \left(r - \frac{1}{2}v\right)\frac{\partial\phi}{\partial x} + \kappa(\bar{v} - v)\frac{\partial\phi}{\partial v} + \lambda(\theta(t) - r)\frac{\partial\phi}{\partial r} + \frac{1}{2}v\frac{\partial^2\phi}{\partial x^2} + \frac{1}{2}\gamma^2 v\frac{\partial^2\phi}{\partial v^2} \\ & + \frac{1}{2}\eta^2\frac{\partial^2\phi}{\partial r^2} + \rho_{x,v}\gamma v\frac{\partial^2\phi}{\partial x\partial v} + \mathbf{\Sigma}_{(1,3)}\frac{\partial^2\phi}{\partial x\partial r} + \mathbf{\Sigma}_{(2,3)}\frac{\partial^2\phi}{\partial r\partial v} - r\phi, \end{aligned} \tag{B.2}$$

with the final condition equal to

$$\phi(u, \mathbf{X}(T), T, T) = \exp(iux(T)).$$

With $\rho_{r,v} = 0$ the H3-HW model with a full matrix of correlations collapses to the setup in section 3.1.

As before, we can use the deterministic approximations $\mathbf{\Sigma}_{(1,3)} \approx \rho_{x,r}\eta\mathbb{E}(\sqrt{v(t)})$ and $\mathbf{\Sigma}_{(2,3)} \approx \rho_{r,v}\gamma\eta\mathbb{E}(\sqrt{v(t)})$, for which Result 3.3 can be used.

The representations of the HHW model in (2.9) and the model in (2.3) with $\rho_{r,v} \neq 0$ for $p = 0$ are closely related. The following lemma specifies the relation in terms of the coefficients of the corresponding ChF.

Lemma B.1 (ChF for the H3-HW model with a full matrix of correlations). *The discounted ChF for the H3-HW model is of the following form:*

$$\phi_{H3\text{-}HW}(u, \mathbf{X}(t), \tau) = \exp\left(\hat{A}(u,\tau) + \hat{B}(u,\tau)x(t) + \hat{C}(u,\tau)r(t) + \hat{D}(u,\tau)v(t)\right),$$

with the functions $\hat{A}(u,\tau)$, $\hat{B}(u,\tau)$, $\hat{C}(u,\tau)$, and $\hat{D}(u,\tau)$ given by

$$\hat{B}(u,\tau) = B(u,\tau), \quad \hat{C}(u,\tau) = C(u,\tau), \quad \hat{D}(u,\tau) = D(u,\tau), \tag{B.3}$$

with $B(u,\tau)$ *in* (3.23)*,* $C(u,\tau)$ *in* (3.24)*, and* $D(u,\tau)$ *in* (3.25)*. For* $\hat{A}(u,\tau)$ *we have*

$$\hat{A}(u,\tau) = A(u,\tau) + \rho_{r,v}\gamma\eta \int_0^\tau \mathbb{E}(\sqrt{v(T-s)})\hat{C}(u,s)\hat{D}(u,s)\mathrm{d}s, \tag{B.4}$$

where $A(u,\tau)$ *is given in* (3.26)*.*

The accuracy of the HHW approximations with a full matrix of correlations is discussed in Appendix C.

Appendix C. Comparison to Markov projection method. In this appendix we compare our results to the Markovian projection (MP) method [5]. We check the results of three different approximation schemes: the MP method, Approx 1, i.e., the approximation with $\sqrt{v(t)} \approx \mathbb{E}(\sqrt{v(t)})$ (section 3.1), and Approx 2, i.e., the method with $\sqrt{v(t)} \approx \mathcal{N}(\cdot)$ (section 4.1).

In the experiment, taken directly from [4], we price an equity option with continuous dividend. The model parameters for the HHW model are given by $\kappa = 0.25$, $\bar{v} = v(0) = 0.0625$, $\gamma = 0.625$, $\lambda = 0.05$, and $\eta = 0.01$, and a zero-coupon bond is given by $P(0,T) = \mathrm{e}^{-0.05T}$ and a continuous dividend of 2%. A full matrix of correlations, as in [4], is given by

$$\mathbf{C} = \begin{bmatrix} 1 & \rho_{x,v} & \rho_{x,r} \\ \rho_{x,v} & 1 & \rho_{v,r} \\ \rho_{x,r} & \rho_{v,r} & 1 \end{bmatrix} = \begin{bmatrix} 100\% & -40\% & 30\% \\ -40\% & 100\% & 15\% \\ 30\% & 15\% & 100\% \end{bmatrix}. \tag{C.1}$$

The MC reference for the implied volatilities, the corresponding standard deviations, and the results for the MP method are all taken from [4].

In order to incorporate a continuous dividend in the equity model, one can model foreign-exchange (FX), in which the volatility of the foreign interest rates is set to zero. In such a setup, the forward, $F(t)$, is defined as

$$F(t) = S(t)\frac{P_f(t,T)}{P_d(t,T)} \quad \text{and} \quad F(0) = S(0)\frac{\mathrm{e}^{-0.02T}}{\mathrm{e}^{-0.05T}},$$

where $P_f(t,T)$ and $P_d(t,T)$ are the foreign and domestic zero-coupon bonds, respectively, paying €1 at the maturity T. By switching from the spot risk-neutral measure, $\mathbb{Q}$, to the T-forward measure, $\mathbb{Q}^T$, discounting will be decoupled from taking the expectation, i.e.,

$$\mathbb{E}\left(\frac{1}{B(T)}\max(S(T)-K,0)|\mathcal{F}(0)\right) = P_d(0,T)\mathbb{E}^T\left(\max(F(T)-K,0)|\mathcal{F}(0)\right).$$

Moreover, the forward, $F(t)$, is a martingale with dynamics given by

$$\begin{aligned} \mathrm{d}F(t)/F(t) &= \sqrt{v(t)}\mathrm{d}W_x^T(t) - \eta B_r(t,T)\mathrm{d}W_r^T(t), \\ \mathrm{d}v(t) &= \left(\kappa(\bar{v}-v(t)) + \gamma\rho_{v,r}\eta B_r(t,T)\sqrt{v(t)}\right)\mathrm{d}t + \gamma\sqrt{v(t)}\mathrm{d}W_v^T(t), \end{aligned} \tag{C.2}$$

where $B_r(t,T) = \frac{1}{\lambda}\left(\mathrm{e}^{-\lambda(T-t)} - 1\right)$, and the full correlation structure given in (C.1).

Table 6

The error for a deterministic (Approx 1) and a stochastic approximation (Approx 2) of the HHW model compared to the MP method. The MP and MC results with the corresponding standard deviations were taken from [4]. The error is defined as a difference between the reference implied volatilities and the approximation.

T	Strike	Imp. vol [%]	MP	Approx 1	Approx 2	Err. MP	Err. 1	Err. 2
	86.07	24.45	24.49	24.48	24.48	−0.04	−0.03	−0.03
	92.77	22.25	22.27	22.27	22.25	−0.02	−0.02	0.00
1y	100.00	20.36	20.32	20.35	20.30	0.04	0.01	0.06
	107.79	19.42	19.34	19.38	19.34	0.08	0.04	0.08
	116.18	19.67	19.64	19.62	19.64	0.03	0.05	0.03
	77.12	22.61	22.65	22.61	22.63	−0.04	0.00	−0.02
	87.82	20.05	20.05	20.09	20.06	0.00	−0.04	−0.01
3y	100.00	17.95	17.91	18.09	17.90	0.04	−0.14	0.05
	113.87	17.23	17.14	17.32	17.15	0.09	−0.09	0.08
	129.67	18.02	17.92	17.93	18.00	0.10	0.09	0.02
	71.50	21.89	21.94	21.90	21.95	−0.05	−0.01	−0.06
	84.56	19.43	19.45	19.52	19.48	−0.02	−0.09	−0.05
5y	100.00	17.49	17.44	17.71	17.45	0.05	−0.22	0.04
	118.26	16.83	16.72	17.01	16.76	0.11	−0.18	0.07
	139.85	17.55	17.42	17.49	17.57	0.13	0.06	−0.02
	62.23	21.55	21.61	21.57	21.68	−0.06	−0.02	−0.13
	78.89	19.52	19.51	19.67	19.61	0.01	−0.15	−0.09
10y	100.00	18.01	17.91	18.31	17.97	0.10	−0.30	−0.04
	126.77	17.41	17.22	17.67	17.30	0.19	−0.26	0.11
	160.70	17.75	17.51	17.79	17.78	0.24	−0.04	−0.03
	51.13	22.28	22.32	22.37	22.47	−0.04	−0.09	−0.19
	71.50	20.91	20.86	21.14	21.03	0.05	−0.23	−0.12
20y	100.00	19.94	19.77	20.27	19.91	0.17	−0.33	0.03
	139.85	19.44	19.16	19.77	19.32	0.28	−0.33	0.12
	195.58	19.40	19.05	19.63	19.39	0.35	−0.23	0.01

Under the log-transform, $x(t) = \log F(t)$, the Kolmogorov backward PDE reads as

$$
\begin{aligned}
-\frac{\partial \phi}{\partial t} &= \kappa(\bar{v}-v)\frac{\partial \phi}{\partial v} + \left(\frac{1}{2}v - \rho_{x,r}\eta B_r(t,T)\sqrt{v} - \frac{1}{2}\eta^2 B_r^2(t,T)\right)\left(\frac{\partial^2\phi}{\partial x^2} - \frac{\partial \phi}{\partial x}\right) \\
&\quad + \left(\rho_{x,v}\gamma v - \rho_{v,r}\gamma\eta\sqrt{v}B_r(t,T)\right)\frac{\partial^2 \phi}{\partial x \partial v} + \frac{1}{2}\gamma^2 v\frac{\partial^2\phi}{\partial v^2} + \rho_{v,r}\gamma\eta\sqrt{v}\frac{\partial \phi}{\partial v},
\end{aligned} \tag{C.3}
$$

with the final condition $\phi(u, \mathbf{X}(T), T, T) = \mathrm{e}^{iux(T)}$.

We linearize PDE (C.3) in two ways: by the deterministic approach described in section 3.1 and Appendix B, and by the stochastic approach as in section 4.1. Both approximations result in affine approximations of PDE (C.3).[6]

The results of the experiments performed, presented in Table 6, show a highly satisfactory accuracy of the HHW approximations introduced in this paper. When comparing to the MP method, we see that the MP method is more accurate for low strike values, whereas our proxies perform favorably for larger strike values, especially when large maturities are considered.

[6] Since the moments of the square-root process under the T-forward measure are difficult to find, we first project $\sqrt{v(t)}$ on a normal process, under measure $\mathbb{Q}$, and then change measures.

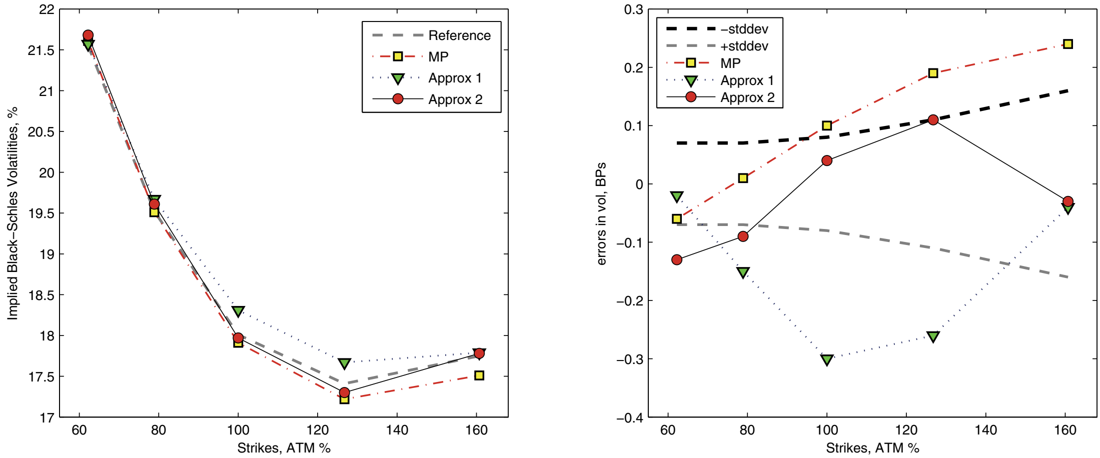


Figure 5. *Left: Implied volatilities for a maturity of* 10 *years. Right: The error for the different approximations. (MP stands for Markovian projection, Approx* 1 *is the deterministic approach, and Approx* 2 *corresponds to the approximation with* $\sqrt{v(t)} \approx \mathcal{N}(\cdot)$.*)*

In Figure 5 the error results for $T = 10$ are presented. In this experiment, the stochastic approximation, Approx 2, performed somewhat better than the deterministic approach, Approx 1.

In the case of the deterministic approach, pricing of European options is done in a split second (the corresponding ChF is analytic when the Feller condition is satisfied; one integration step is required otherwise). In the case of the stochastic approach a numerical routine for solving the ODEs is employed. This, however, can also be done highly efficiently, as already presented in Table 3.

Acknowledgments. The authors would like to thank the anonymous referees for valuable suggestions. Moreover, the authors thank Natalia Borovykh and Sacha van Weeren from Rabobank International for fruitful discussions and helpful comments.

REFERENCES

[1] S. Amstrup, L. MacDonald, and B. Manly, *Handbook of Capture-Recapture Analysis*, Princeton University Press, Princeton, NJ, 2006.

[2] L. Andersen, *Simple and efficient simulation of the Heston stochastic volatility model*, J. Comput. Finance, 11 (2008), pp. 1–42.

[3] J. Andreasen, *Closed Form Pricing of FX Options under Stochastic Rates and Volatility*, presentation at Global Derivatives Conference, Paris, 2006.

[4] A. Antonov, *Effective Approximation of FX/EQ Options for the Hybrid Models: Heston and Correlated Gaussian Interest Rates*, presentation at MathFinance 2007; available online from http://conference.mathfinance.com/2008/papers/antonov/slides.

[5] A. Antonov, M. Arneguy, and N. Audet, *Markovian Projection to a Displaced Volatility Heston Model*, SSRN working paper, 2008; available online from http://ssrn.com/abstract=1106223.

[6] E. Benhamou, A. Rivoira, and A. Gruz, *Stochastic Interest Rates for Local Volatility Hybrid Models*, SSRN working paper, 2008; available online from http://ssrn.com/abstract=1107711.

[7] F. Black and M. Scholes, *The pricing of options and corporate liabilities*, J. Polit. Econ., 81 (1973), pp. 637–654.

[8] D. Brigo and F. Mercurio, *Interest Rate Models—Theory and Practice: With Smile, Inflation and Credit*, 2nd ed., Springer Finance, Springer-Verlag, Berlin, 2007.

[9] M. Broadie and Y. Yamamoto, *Application of the fast Gauss transform to option pricing*, Management Sci., 49 (2003), pp. 1071–1088.

[10] P. P. Carr and D. B. Madan, *Option valuation using the fast Fourier transform*, J. Comput. Finance, 2 (1999), pp. 61–73.

[11] J. C. Cox, J. E. Ingersoll, and S. A. Ross, *A theory of the term structure of interest rates*, Econometrica, 53 (1985), pp. 385–407.

[12] E. Derman and I. Kani, *Stochastic implied trees: Arbitrage pricing with stochastic term and strike structure of volatility*, Int. J. Theor. Appl. Finance, 1 (1998), pp. 61–110.

[13] D. Duffie, J. Pan, and K. Singleton, *Transform analysis and asset pricing for affine jump-diffusions*, Econometrica, 68 (2000), pp. 1343–1376.

[14] D. Dufresne, *The Integrated Square-Root Process*, University of Montreal working paper, 2001; available online from http://repository.unimelb.edu.au/10187/1413.

[15] B. Dupire, *Pricing with a smile*, Risk, 7 (1994), pp. 18–20.

[16] F. Fang and C. W. Oosterlee, *A novel pricing method for European options based on Fourier-cosine series expansions*, SIAM J. Sci. Comput., 31 (2008), pp. 826–848.

[17] R. A. Fisher, *On the interpretation of χ^2 from contingency tables and calculations of P*, J. R. Statist. Soc., 85 (1922), pp. 87–94.

[18] G. Forsythe, M. Malcolm, and C. Moler, *Computer Methods for Mathematical Computations*, Prentice–Hall, Englewood Cliffs, NJ, 1977.

[19] A. Giese, *On the Pricing of Auto-callable Equity Securities in the Presence of Stochastic Volatility and Stochastic Interest Rates*, presentation at MathFinance Workshop: Derivatives and Risk Management in Theory and Practice, Frankfurt, 2006; available online from http://www.mathfinance.de/workshop/2006/papers/giese/slides.pdf.

[20] L. A. Grzelak, C. W. Oosterlee, and S. van Weeren, *Extension of stochastic volatility equity models with Hull-White interest rate process*, Quant. Finance, 2009, pp. 1469–7696.

[21] A. van Haastrecht, R. Lord, A. Pelsser, and D. Schrager, *Pricing long-maturity equity and FX derivatives with stochastic interest rates and stochastic volatility*, Insur. Math. Econ., 45 (2009), pp. 436–448.

[22] S. Heston, *A closed-form solution for options with stochastic volatility with applications to bond and currency options*, Rev. Financ. Stud., 6 (1993), pp. 327–343.

[23] J. Hull and A. White, *Using Hull-White interest rate trees*, J. Derivatives, 4 (1996), pp. 26–36.

[24] C. Hunter, *Hybrid derivatives*, in The Euromoney Derivatives and Risk Management Handbook, Euromoney Institutional Investor, London, 2005.

[25] P. Jäckel, *Stochastic volatility models: Past, present and future*, in The Best of Wilmott I: Incorporating the Quantitative Finance Review, John Wiley and Sons, Chichester, 2004, pp. 379–390.

[26] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, Vol. 2, 2nd ed., Wiley, New York, 1994.

[27] D. Kahaner, C. Moler, and S. Nash, *Numerical Methods and Software*, Prentice–Hall, Englewood Cliffs, NJ, 1989.

[28] E. E. Kummer, *Über die Hypergeometrische Reihe $F(a;b;x)$*, J. Reine Angew. Math., 15 (1936), pp. 39–83.

[29] R. Lee, *Option pricing by transform methods: Extensions, unification, and error control*, J. Comput. Finance, 7 (2004), pp. 51–86.

[30] A. Lewis, *Option Valuation under Stochastic Volatility*, Finance Press, Newport Beach, CA, 2001.

[31] M. Muskulus, K. in't Hout, J. Bierkens, A. P. C. van der Ploeg, J. in't Panhuis, F. Fang, B. Janssens, and C. W. Oosterlee, *The ING problem—a problem from financial industry; three papers on the Heston-Hull-White model*, in Proceedings of the 58th European Study Group Mathematics with Industry, Utrecht, The Netherlands, 2007.

[32] G. W. Oehlert, *A note on the delta method*, Amer. Statist., 46 (1992), pp. 27–29.

[33] B. Øksendal, *Stochastic Differential Equations*, 5th ed., Springer-Verlag, Berlin, 2000.

[34] P. B. Patnaik, *The non-central χ^2 and F-distributions and their applications*, Biometrika, 36 (1949), pp. 202–232.

[35] R. Schöbel and J. Zhu, *Stochastic volatility with an Ornstein-Uhlenbeck process: An extension*, Eur. Financ. Rev., 3 (1999), pp. 23–46.

[36] O. A. Vašiček, *An equilibrium characterization of the term structure*, J. Financ. Econ., 5 (1977), pp. 177–188.

[37] J. Zhu, *Modular Pricing of Options*, Springer-Verlag, Berlin, 2000.