MATHEMATICAL FINANCE
An International Journal of Mathematics, Statistics and Financial Economics

Mathematical Finance, Vol. 24, No. 4 (October 2014), 627–650

LIQUIDATION IN LIMIT ORDER BOOKS WITH CONTROLLED INTENSITY

ERHAN BAYRAKTAR

Department of Mathematics, University of Michigan

MICHAEL LUDKOVSKI

Department of Statistics and Applied Probability, University of California Santa Barbara

We consider a framework for solving optimal liquidation problems in limit order books. In particular, order arrivals are modeled as a point process whose intensity depends on the liquidation price. We set up a stochastic control problem in which the goal is to maximize the expected revenue from liquidating the entire position held. We solve this optimal liquidation problem for power-law and exponential-decay order book models explicitly and discuss several extensions. We also consider the continuous selling (or fluid) limit when the trading units are ever smaller and the intensity is ever larger. This limit provides an analytical approximation to the value function and the optimal solution. Using techniques from viscosity solutions we show that the discrete state problem and its optimal solution converge to the corresponding quantities in the continuous selling limit uniformly on compacts.

KEY WORDS: Limit order books, controlled intensity, optimal control of point processes, optimal control of queueing networks, fluid limit.

1. INTRODUCTION

Liquidation of large securities positions has emerged as an important problem in financial mathematics, linking together models of market microstructure and control theory. In this paper, we consider an investor who liquidates a position through limit orders placed in a limit order book (LOB). The investor does so by choosing the price of the limit order; the higher the price of the limit order, the smaller the probability that it would be filled. The objective of the investor is to come up with an optimal limit order strategy that maximizes her expected revenue by date T.

Our model for the above problem is based on a point-process view of LOBs, which treats liquidation as a sequence of discrete events, i.e., order matches. More precisely, we assume that the investor effectively controls the frequency of her trades by choosing the spread s above the current bid price P_t. The trade intensity is controlled as $\Lambda(s)$

We thank the associate editor and anonymous referees for many useful comments, including the suggestion to analyze the expected execution curves. We also thank Olivier Guéant, Sebastian Jaimungal, Etienne Chevalier, Charles Albert LeHalle, Vathana Ly Vath, and Huyên Pham for helpful discussions at various stages of our writing. E. Bayraktar is supported in part by the National Science Foundation under an applied mathematics research grant and a Career grant, DMS-0906257 and DMS-0955463, respectively, and in part by the Susan M. Smith Professorship.

Manuscript received May 2011; final revision received February 2012.

Address correspondence to Michael Ludkovski, Department of Statistics and Applied Probability, University of California Santa Barbara, CA 93106-3110; e-mail: ludkovski@pstat.ucsb.edu.

DOI: 10.1111/j.1467-9965.2012.00529.x

and when a trade occurs, the investor generates a liquidation profit of s. Similar setups have been proposed in Avellaneda and Stoikov (2008), Cont, Stoikov, and Talreja (2010), Cont and De Larrard (2010) and rely essentially on a queueing system representation of LOBs.

A crucial modeling difference is whether execution takes place through market or limit orders. If investor trades via market orders, she necessarily encounters price impact through "eating away" a portion of the LOB. The precise price impact depends on the *shape* of the LOB, as well as its resilience. Conversely, there is no transactions or fill risk as market orders execute instantaneously. This point of view is taken in, for example, Obizhaeva and Wang (2005), Alfonsi, Fruth, and Schied (2010), and Alfonsi and Schied (2010). On the other hand, if the investor trades through limit orders, liquidation depends on being "lifted" by a sufficiently large market order, leading to substantial fill risk that again depends on the shape and depth of the LOB. The fill risk is related to the concept of virtual price impact (Weber and Rosenow 2005) and is the focus of our model here. Related approaches to trading via limit orders can be found in Avellaneda and Stoikov (2008), Cartea and Jaimungal (2010), Guéant, Lehalle, and Tapia (2011), and Guilbaud and Pham (2011). Also, in our previous work Bayraktar and Ludkovski (2011), we considered the same LOB as here but with an uncontrolled trade intensity and temporary price impact from order size.

Ideally, a fully specified model will reconcile the two approaches above, as well as consider the underlying risk preferences of the investor. This is especially important for dealing with simultaneous trading on multiple exchanges; see the very recent preprints Kratz and Schoeneborn (2010), Klöck, Schied, and Sun (2011), as well as Section 5.5. A full treatment of this problem will be the subject of a separate paper.

Our starting point is a discrete-state problem for an investor holding n shares of an illiquid asset, $n \in \mathbb{N}$. Since practically speaking n is often large (on the order of hundreds of thousands), we also investigate the fluid limit of our setup. On a technical level, the fluid limit provides asymptotic results for the discrete-state problem (see Remark 3.6), which is the main focus of our paper.

On a formal level, our control problem is equivalent to a controlled death process and is closely related to fluid approximations of some queueing problems. We refer to Bäuerle (2000, 2001, 2002), Day (2011), Piunovskiy (2009), Piunovskiy and Zhang (2011) and references therein for the most relevant strand of this rich literature. In contrast with the previous literature, which uses probabilistic arguments, we utilize viscosity techniques to show convergence (both of the value functions and the corresponding optimal controls) from the discrete- to the continuous-state problems.

An investor holds n shares of an asset. Let $(P_t)_{t\geq 0}$ be the bid price process for the underlying asset. Let $r \geq 0$ be the risk-free rate. We assume that $e^{-rt}P_t$ is a martingale with respect to the optimization measure $\mathbb{P}$ on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{G}_t))$. This assumption is consistent with standard market microstructure models; see, for example, Alfonsi and Schied (2010). Let Λ_t be the (controlled) intensity of order fill, and let $s_t \geq 0$ be the spread between the bid price and the limit order of the investor. Denote by N_t the $\mathcal{G}$-adapted counting process of order fills and τ_k the corresponding arrival times,

$$N_t = \sum_k 1_{\{\tau_k \leq t\}}.$$

Then $N_t - \int_0^t \Lambda_s \, ds$ is a martingale and expected revenue is

$$\mathbb{E}\left[\sum_{i=1}^{n} e^{-r\tau_i}(P_{\tau_i} + s_{\tau_i} 1_{\{\tau_i \le T\}})\right]. \tag{1.1}$$

We assume that the investor has a deadline date $T \le +\infty$ by which all trades must be completed. Remaining shares are liquidated at zero profit at T.

To introduce the liquidation control, we assume that $\Lambda_t = \Lambda(s_t)$, so that the intensity of order fills is a function of the offered spread above the bid price. Moreover, we assume that the bid price P is unaffected by the limit orders created via (s_t). Since $e^{-rt}P_t$ is a martingale, the first term in (1.1) is independent of τ_i. Indeed, $\mathbb{E}[\sum_{i=1}^{n} e^{-r\tau_i} P_{\tau_i}] = nP_0$ and we may ignore P in the subsequent analysis.

We define

$$V(n, T) := \sup_{(s_t)\in\mathcal{S}_T} \mathbb{E}\left[\sum_{i=1}^{n} e^{-r\tau_i} s_{\tau_i} 1_{\{\tau_i \le T\}}\right] \tag{1.2}$$

$$= \sup_{(s_t)\in\mathcal{S}_T} \mathbb{E}\left[\int_0^{T\wedge\tau(X)} e^{-rt} s_t\, dN_t\right] = \sup_{(s_t)\in\mathcal{S}_T} \mathbb{E}\left[\int_0^{T\wedge\tau(X)} e^{-rt} s_t \Lambda(s_t)\, dt\right], \tag{1.3}$$

where

$$\tau(X) := \inf\{t \ge 0 : X_t = 0\}$$

is the time of liquidation. Here, $X_t \coloneqq X_0 - N_t$, with $X_0 = n$, is a “death” (or inventory) process with intensity $\Lambda(s_t)$. Note that T in (1.2) represents *time-to-maturity* and $\mathcal{S}_T$ is the collection of $\mathcal{F}$-adapted controls, $s_t \ge 0$ with $\mathcal{F}_t := \sigma(N_s : s \le t)$. The boundary conditions on V are $V(n, 0) = 0\,\forall n$ (terminal condition in time) and $V(0, T) = 0\,\forall T$ (exhaustion).

REMARK 1.1. Our model is related to the LOB setup of Avellaneda and Stoikov (2008), which assumes that limit orders are “lifted” through sufficiently large *market* buy orders. Namely, a market buy order of size q, hits all limit sell orders that are within $I(q)$ of the best bid. Assuming that buy market orders arrive in the form of a Poisson random measure on $\mathbb{R}_+ \times \mathbb{R}_+$ with arrival intensity $\bar{\lambda}dt$ and volume (mark) distribution $f(dq)$, $q \ge 0$, a sell limit order at a given spread u is lifted with probability $\mathbb{P}(I(q) > u)$. By the thinning lemma on Poisson processes, such matching buy orders form a Poisson process with intensity $\bar{\lambda} \int_{I^{-1}(u)}^{\infty} f(dq)$. Empirical studies suggest a power-law depth function $f(dq) \propto q^{-1-a}dq$ (Avellaneda and Stoikov 2008) and therefore if $\Lambda(0) < \infty$, we can view our model within the Avellaneda and Stoikov (2008) framework, $\Lambda(s) \propto [I^{-1}(s)]^{-a}$, with I^{-1} the virtual price impact function (Weber and Rosenow 2005).

As in Bäuerle and Rieder (2009), the above control problem can be transformed into a discrete-time Markov decision problem and the classical results from Bertsekas and Shreve (1978, ch. 8) can be used to prove a dynamic programming principle. Using the latter result one can show that the value function is a viscosity solution of

$$-V_T + \sup_{s\ge 0} \Lambda(s)\big[V(n-1, T) - V(n, T) + s\big] - r\,V(n, T) = 0, \tag{1.4}$$

with boundary conditions $V(0, T) = V(n, 0) = 0$ and V_T denoting partial derivative with respect to time-to-expiration. Standard results also imply that an optimal control can be taken of Markov feedback type, $s_t^* = s(X_t^*, T - t)$. However, in most of the examples below we will obtain explicit solutions to this dynamic programming equation. Then a *verification lemma* can be used to justify that the solution of (1.4) is indeed the value function.

The optimization problem described in (1.2) is simplified but highly tractable. In most of the examples below, we are able to obtain closed-form solutions which provide direct insight into the relationship between the LOB model and its depth function and the investor's liquidation strategy. In Section 2, we give an explicit solution for (1.2) in the case of a power-law intensity control $\Lambda(s)$. Section 3 then studies convergence of the discrete problem (1.4) to its continuous-state fluid limit. Our key Theorem 3.2, complemented by Proposition 3.4 and Corollary 3.5, gives a full account of this convergence using techniques from viscosity solutions of nonlinear partial differential equations (PDE). In Section 4 we return to (1.2) for the case where $\Lambda(s)$ is of exponential shape; we are again able to provide several closed-form solutions. Finally, Section 5 considers several extensions and numerical illustrations of (1.2), including generic $\Lambda(s)$, which shed additional light on the problem structure.

2. POWER-LAW LIMIT ORDER BOOKS

In this section, we assume that incoming buy orders have a power-law distribution for the spread, $\Lambda(s) = \frac{\lambda}{s^\alpha}$ for some $\alpha > 1$. It can be observed from the computations below that if $\alpha \leq 1$, then no optimal control exists. Similar assumption was made (and justified empirically) by Avellaneda and Stoikov (2008) who write that in realistic markets $\alpha \in [1.5, 3]$.

PROPOSITION 2.1. *Assume that $\Lambda(s) = \lambda s^{-\alpha}$ with boundary conditions $V(0, T) = V(n, 0) = 0$ for all n. Then the solution of (1.4) and the optimal spread are, respectively*

$$V(n, T) = c_n(1 - e^{-r\alpha T})^{1/\alpha}, \qquad s^*(n, T) = \left(\frac{\lambda}{\alpha r c_n}\right)^{1/(\alpha-1)} \cdot (1 - e^{-r\alpha T})^{1/\alpha}, \tag{2.1}$$

with c_n satisfying the recursion

$$r c_n = A_\alpha \lambda (c_n - c_{n-1})^{1-\alpha}, \quad n \geq 1, \quad c_0 = 0, \tag{2.2}$$

where

$$A_\alpha := \frac{(\alpha - 1)^{\alpha-1}}{\alpha^\alpha}. \tag{2.3}$$

REMARK 2.2. Note that V is "concave" in n in the sense that

$$V(n+1, T) - V(n, T) \leq V(n, T) - V(n-1, T),$$

i.e., its linear interpolation in n is concave in the usual sense. This follows immediately from (2.2) since

$$c_{n+1} - c_n = \left(\frac{r c_{n+1}}{\lambda A_\alpha}\right)^{1/(1-\alpha)} \leq \left(\frac{r c_n}{\lambda A_\alpha}\right)^{1/(1-\alpha)} = c_n - c_{n-1},$$

where the inequality follows from the fact that c_n (or $V(n, T)$) is increasing in n. The latter follows directly from (1.2).

Also observe from (2.1) and (2.2) that

$$s^*(n,T) = \frac{\alpha}{\alpha-1}(V(n,T) - V(n-1,T)),$$

which implies that $n \mapsto s^*(n, T)$ is a decreasing function, because $n \mapsto V(n, T)$ is "concave." One can also think of s^* as the derivative of the linear interpolation of V in n.

Proof. With the power law assumption (1.4) reduces to

$$-V_T + \sup_s \frac{\lambda}{s^\alpha}(V(n-1,T) - V(n,T) + s) - rV = 0, \tag{2.4}$$

and therefore the candidate optimal policy is $t \mapsto s^*(X_t, T-t)$ in which $s^*(n,T) = \frac{\alpha}{\alpha-1}(V(n,T) - V(n-1,T))$. To begin solving this equation, we start with $n = 1$. Since $V(0, T) = 0$ for all T, we obtain for $V = V(1, T)$

$$-V_T + A_\alpha \lambda V^{1-\alpha} - rV = 0.$$

This is a separable ordinary differential equation (ODE) which simplifies to

$$T + C = \int^{\cdot} \frac{V^{\alpha-1}}{A_\alpha\lambda - rV^\alpha} dV.$$

Using the boundary condition $V(1, 0) = 0$ we integrate to obtain

$$\log(A_\alpha\lambda - rV^\alpha) = -r\alpha(T+C) \quad \Longleftrightarrow \quad V(1,T) = \left\{\frac{A_\alpha\lambda}{r}(1-e^{-r\alpha T})\right\}^{1/\alpha}.$$

Considering the equation for general $n > 1$ we therefore make the ansatz $V(n, T) = c_n(1 - e^{-r\alpha T})^{1/\alpha}$ and plugging into (2.4) the relation (2.2) follows. □

REMARK 2.3. If no discounting is present $r = 0$, one can verify that the solution of (1.4) is $V(n, T) = d_n T^{1/\alpha}$, where the sequence (d_n) satisfies the recursion

$$d_n = \lambda\left(\frac{\alpha-1}{\alpha}\right)^{\alpha-1}(d_n - d_{n-1})^{1-\alpha}, \qquad d_0 = 0.$$

This result can also be obtained by taking the limit $r \to 0$ in (2.1), (2.2).

Fixing $X_0 = n$, the inter-trade intervals $\sigma_i \coloneqq \tau_i - \tau_{i-1}$, $i \le n$ have survival functions given by

$$\mathbb{P}(\sigma_i > t | \tau_{i-1}) = \exp\left(-\int_0^t \Lambda(s^*(n-i+1, T-\tau_{i-1}-s))\, ds\right).$$

Noting that $\Lambda(s^*(n,T)) = \frac{C(n)}{1-e^{-r\alpha T}}$ for some constant $C(n)$, it follows that $\int_0^\varepsilon \Lambda(s^*(n,T))\, dT = +\infty$ for all n and ε and therefore $\mathbb{P}(\sigma_i \le T - \tau_{i-1}) = 1$ for all $i \le n$. We conclude that even though there is no direct penalty if some orders remain at T, with probability one, the full inventory is liquidated by T, $X_T^* = 0$ $\mathbb{P}$-a.s. In particular, the problem with a *hard* liquidation constraint $V(x, 0) = -M1_{\{x \ge 0\}}$ for any liquidation penalty $M \ge 0$ will have the same solution as in Proposition 2.1.

2.1. Infinite Horizon

As the execution horizon T grows, the investor faces a weaker liquidation constraint. Nevertheless, she still prefers to sell earlier than later due to the discount parameter r that incentivizes faster liquidation. For the limit $T \to \infty$ we obtain an infinite-horizon model whereby strategies are time-homogenous.

Taking $T \to \infty$ in (2.1) we find that $V(n) = c_n$ and $s^*(n) = \lambda^{1/(\alpha-1)}(\alpha r c_n)^{1/(1-\alpha)}$. To understand how quickly execution takes place let us introduce *expected time to liquidate* $S(n)$ which is defined to be $S(n) := \mathbb{E}[\tau(X^*) | X_0^* = n]$, in which X^* is the death process whose intensity at time t is $\Lambda(s^*(X_t^*))$ (representing optimally controlled inventory at t). When the inventory is $X_t^* = n$, liquidation occurs at rate $\Lambda(s^*(n))$, so that the interval until the next trade has an exponential distribution with mean $1/\Lambda(s^*(n))$. It follows that

$$S(n) = \sum_{j=1}^{n} \Lambda(s^*(j))^{-1} = \lambda^{1/(\alpha-1)} \sum_{j=1}^{n} (\alpha r c_j)^{-\alpha/(\alpha-1)}. \tag{2.5}$$

3. CONTINUOUS SELLING LIMIT

To better understand the results of Proposition 2.1 we consider a limiting continuous model. Let us denote the number of shares initially held by x.

We first introduce a sequence of discrete control problems that converge to the continuous selling limit. For $0 < \Delta \leq 1$, consider the problem where shares are sold at Δ increments and the intensity of order fills is $\Lambda^{\Delta}(s) \coloneqq \Lambda(s)/\Delta$. We will denote by X^{Δ} the "death" process with this intensity and decrements of size Δ. Then the resulting value function

$$\begin{aligned} V^{\Delta}(x, T) &:= \sup_{(s_t)\in\mathcal{S}_T} \mathbb{E}\left[\sum_{i=1}^{x/\Delta} e^{-r\tau_i} \Delta \cdot s_{\tau_i} 1_{\{\tau_i \leq T\}}\right] \\ &= \sup_{(s_t)\in\mathcal{S}_T} \mathbb{E}\left[\int_0^{T\wedge\tau(X^{\Delta})} e^{-rt} s_t \Lambda(s_t)\, dt\right], \end{aligned} \tag{3.1}$$

$x \in \{0, \Delta, 2\Delta, \dots\}$, $T \in \mathbb{R}_+$ would satisfy

$$-V_T^{\Delta} + \sup_{s\geq 0} \frac{\lambda}{s^{\alpha}\Delta}(V^{\Delta}(x-\Delta, T) - V^{\Delta}(x, T) + s\Delta) - r V^{\Delta} = 0 \tag{3.2}$$

in viscosity sense.

Let us consider the first-order PDE

$$-v_T + \sup_{s\geq 0} \lambda \frac{s - v_x}{s^{\alpha}} - rv = 0, \tag{3.3}$$

which can be written as

$$-v_T + A_{\alpha} \lambda v_x^{1-\alpha} - rv = 0,$$

with boundary conditions $v(x,0)=v(0,T)=0$. The solution of (3.3) has the following deterministic control representation

$$v(x,T)=\sup_{(s_t)\in\mathcal{S}_T}\int_0^{T\wedge\tau(X^{(0),x})}\frac{\lambda}{s_t^{\alpha-1}}e^{-rt}dt, \tag{3.4}$$

where $dX_t^{(0),x}=-\lambda s_t^{-\alpha}\,dt$, $X_0^{(0),x}=x$. In fact, the solution of (3.3) is explicitly given by

$$v(x,T)=\left(\frac{\lambda}{r\alpha}\right)^{1/\alpha}x^{(\alpha-1)/\alpha}(1-e^{-r\alpha T})^{1/\alpha}. \tag{3.5}$$

We denote the optimizer in (3.3) by $s^{(0)}(x,T)$, which is explicitly given by

$$s^{(0)}(x,T)=\left(\frac{\lambda}{\alpha r}\right)^{1/\alpha}\frac{1}{x^{1/\alpha}}(1-e^{-r\alpha T})^{1/\alpha}.$$

REMARK 3.1. Plugging the optimizer back into the dynamics for $X^{(0),x}$ we obtain that

$$dX_t^{(0),x}=-\frac{\alpha r\,X_t^{(0),x}}{1-e^{-r\alpha(T-t)}}dt,$$

which can be explicitly solved as

$$X_t^{(0),x}=x\exp\left(-\int_0^t\frac{\alpha r}{1-e^{-\alpha r(T-u)}}du\right). \tag{3.6}$$

Observe that the function $t\mapsto X_t^{(0),x}$ is strictly decreasing and strictly convex with $X_T^{(0),x}=0$.

Let $x\in\mathbb{R}_+$ be fixed and let us consider all the collections $\{0,\Delta,2\Delta,\dots\}$ of grids that contain x as an element. In the next result, we will show that as $\Delta\to0$ then $V^\Delta(x)\to v(x)$. In fact, the next result shows that this convergence is uniform on compacts.

THEOREM 3.2. *As $\Delta\to0$, $V^\Delta\to v$ uniformly on compact sets.*

Proof. Let us consider the regularized stochastic control problem

$$V^{\Delta,k}(x,T):=\sup_{(s_t)\in\mathcal{S}_T^k}\mathbb{E}\left[\sum_{i=1}^{x/\Delta}e^{-r\tau_i}\Delta\cdot s_{\tau_i}1_{\{\tau_i\le T\}}\right],\quad x\in\{0,\Delta,2\Delta,\dots\}, \tag{3.7}$$

where $\mathcal{S}_T^k:=\{s\in\mathcal{S}_T:s_t\in[1/k,k]\}$, $k>1$. Using a representation similar to the one in (3.1) and using the lower bound on the controls $s\in\mathcal{S}_T^k$, it can be seen that

$$V^{\Delta,k}(x,T)\le\frac{\lambda}{r}k^{\alpha-1}. \tag{3.8}$$

We will follow the arguments of Barles and Souganidis (1991) in the proof of their Theorem 2.1 (also see theorem 4.1 on page 334 of Fleming and Soner 2006) to show that $V^{\Delta,k}$ converges uniformly on compacts to the unique viscosity solution of

$$-v_T^k+\sup_{s\in[1/k,k]}\lambda\frac{s-v_x^k}{s^\alpha}-rv^k=0,\quad v^k(x,0)=0. \tag{3.9}$$

Let $\bar{v}^k$ and $\underline{v}^k$ be defined by:

$$\bar{v}^k(x,T) := \limsup_{\delta\to 0}\limsup_{\Delta\to 0}\sup\{V^{\Delta,k}(y,S) : |x-y|+|T-S|\leq\delta,\ y\in\{0,\Delta,\ldots\}\},$$

$$\underline{v}^k(x,T) := \liminf_{\delta\to 0}\liminf_{\Delta\to 0}\inf\{V^{\Delta,k}(y,S) : |x-y|+|T-S|\leq\delta,\ y\in\{0,\Delta,\ldots\}\}.$$

By definition we have that $\underline{v}^k \leq v^k \leq \bar{v}^k$ and that $\underline{v}^k$ is lower semi-continuous, and $\bar{v}^k$ is upper semi-continuous; see for example, proposition 5.2.1 of Bardi and Capuzzo-Dolcetta (1997). We will show that $\bar{v}^k$ is a subsolution and that $\underline{v}^k$ is a supersolution of (3.9). It follows from theorem 5.4.20 in Bardi and Capuzzo-Dolcetta (1997) that a comparison result holds for this PDE (the compactness of the control space is required in order to apply this result). This comparison theorem would then imply that $\bar{v}^k \leq \underline{v}^k$. As a result, $v^k = \bar{v}^k = \underline{v}^k$ is the unique continuous viscosity solution of (3.9). This fact together with the way the functions $\bar{v}^k$ and $\underline{v}^k$ are defined also imply the local uniform convergence of $V^{\Delta,k}$ to v^k. (For a similar argument see page 35 of Crandall, Ishii, and Lions (1992).)

We now prove that $\bar{v}^k$ is a viscosity subsolution of (3.9); the fact that $\underline{v}^k$ is a viscosity supersolution follows similarly. Let (x_0, T_0) be a local maximum of $\bar{v}^k - \phi$ for some test function $\phi \in C^{1,1}$. Without loss of generality, we will assume that (x_0, T_0) is a strict local maximum and that $\bar{v}^k(x_0, T_0) = \phi(x_0, T_0)$, and $\phi \geq 2\frac{\lambda}{r}k^{\alpha-1}$ outside the ball $\mathrm{B}(x_0, T_0; \mathfrak{r})$, where $\mathfrak{r} > 0$ is chosen so that (x_0, T_0) is the maximum of $\bar{v}^k - \phi$ on $B(x_0, T_0; \mathfrak{r})$. Thanks to the choice of the test function outside this ball, (x_0, T_0) is in fact a global maximum of the function $\bar{v}^k - \phi$ and it is attained on $B(x_0, T_0; \mathfrak{r})$. (This is where the uniform boundedness assumption in (3.8) is used.)

Let $(x^\Delta, T^\Delta) \in \{0, \Delta, \ldots\} \times \mathbb{R}_+$ be a point at which $V^{\Delta,k} - \phi$ attains its (global) maximum. It follows from the definition of $\bar{v}^k$ and the fact that (x_0, T_0) is a strict global maximum of $\bar{v}^k - \phi$ that there exists a sequence $\Delta_n \to 0$ such that $(x^{\Delta_n}, T^{\Delta_n}) \to (x_0, T_0)$, $V^{\Delta_n,k} - \phi$ attains its global maximum at that point and $V^{\Delta_n,k}(x^{\Delta_n}, T^{\Delta_n}) \to \bar{v}^k(x_0, T_0)$. From the global maximality

$$V^{\Delta_n,k}(x,T) - V^{\Delta_n,k}(x^{\Delta_n}, T^{\Delta_n}) \leq \phi(x,T) - \phi(x^{\Delta_n}, T^{\Delta_n}).$$

Moreover, it can be argued as in Bäuerle and Rieder (2009) using the discrete dynamic programming principle (see Bertsekas and Shreve 1978) that $V^{\Delta_n,k}$ satisfies

$$-V_T^{\Delta_n,k} + \sup_{s\in[1/k,k]}\frac{\lambda}{s^\alpha\Delta_n}(V^{\Delta_n,k}(x^{\Delta_n}-\Delta_n, T^{\Delta_n}) - V^{\Delta_n,k}(x^{\Delta_n}, T^{\Delta_n}) + s\Delta_n) - rV^{\Delta_n,k} = 0$$

in the viscosity sense. Then

$$-\phi_T + \sup_{s\in[1/k,k]}\frac{\lambda}{s^\alpha\Delta_n}(\phi(x^{\Delta_n}-\Delta_n, T^{\Delta_n}) - \phi(x^{\Delta_n}, T^{\Delta_n}) + s\Delta_n) - r\phi + r(\phi - V^{\Delta_n,k}) \geq 0.$$

Taking the limit as $\Delta_n \to 0$ we obtain from this equation that

$$-\phi_T(x_0,T_0) + \sup_{s\in[1/k,k]}\frac{\lambda}{s^\alpha}(s - \phi_x(x_0,T_0)) - r\phi(x_0,T_0) \geq 0,$$

which proves the subsolution property of $\bar{v}^k$. Here, we exchange the limit in Δ_n and the supremum with respect to s using proposition 7.32 in Bertsekas and Shreve (1978) which we can apply thanks to the compactness of the control space.

It follows again from theorem 5.4.20 in Bardi and Capuzzo-Dolcetta (1997) that the unique solution of (3.9) is given by

$$v^k(x, T) = \sup_{(s_t)\in\mathcal{S}_T} \int_0^{T\wedge\tau(X^{(k),x})} \frac{\lambda e^{-rt}}{((s_t \vee 1/k)\wedge k)^{\alpha-1}} dt, \tag{3.10}$$

where $dX_t^{(k),x} = -\lambda/((s_t \vee 1/k)\wedge k)^{\alpha} dt$, $X_0^{(k),x} = x$. We will show that v^k converges pointwise to v:

$$\begin{aligned}\lim_{k\to\infty} v^k(x, T) &= \sup_k \sup_{(s_t)\in\mathcal{S}_T} \int_0^{T\wedge\tau(X^{(k),x})} \frac{\lambda e^{-rt}}{((s_t \vee 1/k)\wedge k)^{\alpha-1}} dt \\ &= \sup_{(s_t)\in\mathcal{S}_T} \sup_k \int_0^{T\wedge\tau(X^{(k),x})} \frac{\lambda e^{-rt}}{((s_t \vee 1/k)\wedge k)^{\alpha-1}} dt \\ &\geq \sup_{(s_t)\in\mathcal{S}_T} \int_0^{T\wedge\tau(X)} \frac{\lambda e^{-rt}}{s_t^{\alpha-1}} dt = v(x, T),\end{aligned}$$

where the inequality follows from the lower semi-continuity of the map $X \mapsto \tau(X)$; see lemma 5 in Day (2011). On the other hand, since v is a supersolution of (3.9), the comparison result theorem 5.4.20 in Bardi and Capuzzo-Dolcetta (1997) implies that $v^k \leq v$ for each k, and as a result

$$\lim_{k\to\infty} v^k(x, T) \leq v(x, T).$$

Combining the last two inequalities, we obtain the pointwise convergence of v^k to v. Pointwise convergence, on the other hand, implies uniform convergence on compacts due to Dini's theorem, since we already know that v is a continuous function of its arguments, and that v^k is an increasing sequence of functions. The latter fact follows from the fact that v^{k+1} is a supersolution of the PDE v^k satisfies. □

REMARK 3.3. Results somewhat similar to Theorem 3.2 appeared in Bäuerle (2000, 2001, 2002); Day (2011); Piunovskiy (2009); Piunovskiy and Zhang (2011) which are on the optimal control of queueing networks. (Among these papers only Day 2011 considered optimal time-to-empty queueing control problems.) To prove Theorem 3.2 we used a completely different approach than the above literature, which had relied on probabilistic arguments. Our approach relies in contrast on the analytical approximation ideas of Barles and Souganidis (1991). We see the prelimit control problem as the discretization (only in the space variable but not in the time variable) of the "fluid limit" first-order nonlinear PDE (3.3) and rely on convergence of the approximation schemes to the viscosity solutions of such nonlinear PDEs. This approach could be fruitful in general in proving "fluid limit" results associated to controlled queueing networks.

The following is a strengthening of Theorem 3.2 which is an interesting result in its own right.

PROPOSITION 3.4. *For any sequence* (Δ_k) *with* $\Delta_k = \delta 2^{-k}$, *we have* $V^{\Delta_k} \uparrow v$ *as* $k \to \infty$.

Proof. We show that for any $\Delta > 0$, $V^{2\Delta} \leq V^{\Delta}$. Due to the factoring of T and x in Proposition 2.1, it suffices to establish this result on the infinite horizon where strategies are constant between trading times.

Fix $\varepsilon > 0$ and let $s^{2\Delta}$ be an ε-optimal strategy for $V^{2\Delta}$. This policy is defined over $x \in \{0, 2\Delta, 4\Delta, \dots\}$. We will recursively construct a policy s^{Δ} over the domain $x \in \{0, \Delta, 2\Delta, \dots\}$ that outperforms $s^{2\Delta}$. The dynamic programming principle implies that

$$\begin{aligned} V^{2\Delta}(2n\Delta) &\leq \mathbb{E}[e^{-r\tau_1}[2s^{2\Delta}(2n\Delta)\Delta + V^{2\Delta}((2n-2)\Delta)] + \varepsilon \\ &= \frac{\Lambda(s^{2\Delta}(2n\Delta))}{\Lambda(s^{2\Delta}(2n\Delta)) + r}\{2s^{2\Delta}(2n\Delta)\Delta + V^{2\Delta}((2n-2)\Delta)\} + \varepsilon. \end{aligned}$$

Similarly, given the liquidation strategy s^{Δ} and corresponding trading times $\tilde{\tau}_i$, the resulting expected profits denoted as $\tilde{V}^{\Delta}(x)$, $x \in \{0, \Delta, 2\Delta, \dots\}$ satisfy for $y = 2n\Delta$,

$$\begin{aligned} \tilde{V}^{\Delta}(y) &= \mathbb{E}[e^{-r\tilde{\tau}_1}s^{\Delta}(y)\Delta + e^{-r\tilde{\tau}_2}\{s^{\Delta}(y-\Delta)\Delta + \tilde{V}^{\Delta}(y-2\Delta)\}] \\ &= \frac{2\Lambda(s^{\Delta}(y))}{2\Lambda(s^{\Delta}(y)) + r}\left\{s^{\Delta}(y)\Delta + \frac{2\Lambda(s^{\Delta}(y-\Delta))}{2\Lambda(s^{\Delta}(y-\Delta)) + r}[s^{\Delta}(y-\Delta)\Delta + \tilde{V}^{\Delta}(y-2\Delta)]\right\}. \end{aligned}$$

Given $s_{2n} \equiv s^{2\Delta}(2n\Delta)$ we prove below that there exists $u \in \mathbb{R}_+$, such that

$$\frac{\Lambda(s_{2n})}{\Lambda(s_{2n}) + r}(2s_{2n}\Delta + V) \leq \frac{2\Lambda(u)}{2\Lambda(u) + r}\left\{u\Delta + \frac{2\Lambda(u)}{2\Lambda(u) + r}(u\Delta + V)\right\}, \tag{3.11}$$

for any $V \geq 0$. This would establish $V^{\Delta}(2n\Delta) \geq \tilde{V}^{\Delta}(2n\Delta) \geq V^{2\Delta}(2n\Delta) - \varepsilon$ by induction on n after setting $s^{\Delta}(2n\Delta) = s^{\Delta}((2n-1)\Delta) = u$. Since ε is arbitrary, the statement of the proposition would then follow. Note that in the above construction, the Δ-investor trading in smaller increments and twice as much, uses the *same* spread u to trade when her inventory is $2n\Delta$ or $(2n-1)\Delta$.

Let $z^2 := \frac{\Lambda(s_{2n})}{\Lambda(s_{2n})+r}$ and define u implicitly through $\frac{2\Lambda(u)}{2\Lambda(u)+r} := z < 1$. Solving for s_{2n} and u in terms of z and using $\Lambda(s) = \lambda s^{-\alpha}$ we obtain

$$s_{2n} = \left(\frac{\lambda(1-z^2)}{rz^2}\right)^{\alpha^{-1}} > \left(\frac{2\lambda(1-z)}{rz}\right)^{\alpha^{-1}} = u.$$

Observe that by construction

$$\frac{\Lambda(s_{2n})}{\Lambda(s_{2n}) + r} = \frac{4\Lambda(u)^2}{(2\Lambda(u) + r)^2},$$

so that the Laplace transform at r of the duration to execute two trades by the Δ-investor is equal to the Laplace transform at r of the duration to execute one trade by the 2Δ-investor. Using this fact, (3.11) is equivalent to

$$\begin{aligned} &\frac{2\Lambda(s_{2n})}{\Lambda(s_{2n}) + r}s_{2n} \leq \frac{2\Lambda(u)}{2\Lambda(u) + r}\left\{u + \frac{2\Lambda(u)}{2\Lambda(u) + r}u\right\} \\ &\iff 2z^2 s_{2n} \leq z(1+z)u. \end{aligned}$$

Since $\alpha > 1$ and the terms on both sides of the above inequality are positive, we may raise both sides to the α-power and plug-in the expressions for s_{2n} and u to find

$$\begin{aligned}(2z^2)^\alpha s_{2n}^\alpha - z^\alpha(1+z)^\alpha u^\alpha &= (2z^2)^\alpha \frac{\lambda(1-z^2)}{rz^2} - z^\alpha(1+z)^\alpha \frac{2\lambda(1-z)}{rz}\\ &= 2\lambda r^{-1}(1-z)(1+z)z^{\alpha-1}[(2z)^{\alpha-1} - (1+z)^{\alpha-1}] < 0,\end{aligned}$$

where the last inequality follows since $z < 1$ and $\alpha > 1$. This shows that (3.11) holds and concludes the proof of the proposition. □

Next, we show that the strategies also converge thanks to the concavity of all the functions involved.

COROLLARY 3.5. *Let us denote by $s^{(\Delta)}$ the pointwise optimizer in (3.2). Then we have that $s^{(\Delta)}(x, T) \to s^{(0)}(x, T)$. (Here, x is fixed and we take the limit over the grids that pass through x.)*

Proof. On the one hand, as in Remark 2.2 we can think of $s^{(\Delta)}$ as something proportional to the left derivative of the linear interpolation of V^Δ, denoted by $\hat{V}^\Delta$, which is increasing and concave. On the other hand, $s^{(0)}$ is proportional to the derivative of the concave differentiable function v. By theorem 24.5 on page 233 of Rockafellar (1997) it however follows that for any $x > 0$,

$$|D_x^- \hat{V}^\Delta(x, T) - v_x(x, T)| \le \varepsilon$$

for small enough Δ, where D_x^- denotes the left derivative operator with respect to x. □

REMARK 3.6. Theorem 3.2 tells us about the asymptotics of c_n in (2.2):

$$c_n \sim \left(\frac{\lambda}{r\alpha}\right)^{1/\alpha} n^{(\alpha-1)/\alpha} \quad \text{as} \quad n \to \infty.$$

Corollary 3.5 can be used to find out the marginal price asymptotics in Proposition 2.1:

$$s^*(n, T) \sim \left(\frac{\lambda}{\alpha r}\right)^{1/\alpha} \frac{1}{n^{1/\alpha}} \quad \text{as} \quad n \to \infty. \tag{3.12}$$

Clearly, the spread will go to zero as $n \to \infty$. But here we are able to obtain the rate of convergence to zero as a function of the remaining inventory.

For time to execution on infinite horizon we have for $\tau_1 = \inf\{t : X_t \le x_1\}$ and $S(x_1, x_2) := \mathbb{E}[\tau_1 | X_0 = x_2]$ that

$$S(x_1, x_2) = \int_{x_1}^{x_2} \frac{1}{\Lambda(s^{(0)}(u))}\, du, \tag{3.13}$$

since intuitively when inventory is of size u, the expected time to liquidate an infinitesimal quantity du is inversely proportional to the current trading rate $\Lambda(s^{(0)}(u))$. Plugging in $s^{(0)}(u) = (\frac{\lambda}{\alpha r u})^{1/\alpha}$ we obtain $\Lambda(s^{(0)}(u)) = \alpha r u$ or $S(x_1, x_2) = \frac{1}{\alpha r}\log(\frac{x_2}{x_1})$ which shows that orders are filled in logarithmic time (as $x_1 \to 0$ the remainder is executed arbitrarily slow).

4. EXPONENTIAL-DECAY ORDER BOOKS

The power-law order book implies that trades can be made arbitrarily quickly as the spread goes to zero: $\lim_{s\to 0}\Lambda(s) = +\infty$. Also, it gives a relatively good chance of executing trades deep in the book, i.e., when s is large. For less liquid markets, both of these features might not be realistic. Accordingly, we consider an exponential-decay LOB, with

$$\Lambda(s) = \lambda e^{-\kappa s}, \quad \kappa > 0, \tag{4.1}$$

where κ controls the exponential depth of the book and $\lambda = \Lambda(0)$ is the order intensity at the bid price. The optimization problem for the spread is now of the form

$$\sup_{s\geq 0} \lambda\, e^{-\kappa s}(V(n-1) - V(n) + s),$$

which leads to the candidate optimizer $s^*(n) = \frac{1}{\kappa} + (V(n) - V(n-1))$. We observe that s^* is bounded away from zero so no trades are ever placed close to the bid.

4.1. Finite Horizon

With a finite horizon and no discounting we obtain the following closed-form solutions to the execution problem.

PROPOSITION 4.1. *Consider again $V^\Delta(x, T)$ defined in (3.1) with boundary condition $V^\Delta(x, 0) = V^\Delta(0, T) = 0$, $r = 0$ and $\Lambda(s)$ given in (4.1). Then for $x = n\Delta$,*

$$V^\Delta(x, T) = \frac{\Delta}{\kappa}\log\left(\sum_{j=0}^{n}\frac{1}{j!}\left(\frac{\lambda T}{\Delta e}\right)^j\right), \tag{4.2}$$

and

$$s^*(n\Delta, T) = \frac{1}{\kappa}\left(1 + \log\left(1 + \frac{\dfrac{(\lambda T)^n}{(\Delta e)^n n!}}{\displaystyle\sum_{j=0}^{n-1}\frac{(\lambda T)^j}{(\Delta e)^j j!}}\right)\right).$$

As $\Delta \searrow 0$, $V^\Delta(n\Delta, T) \to v(x, T)$ uniformly on compacts, where $v(x, T)$ solves the nonlinear first-order PDE

$$v_T(x, T) = \frac{\lambda}{\kappa}e^{-1-\kappa v_x(x,T)}, \tag{4.3}$$

with boundary conditions $v(0, T) = v(x, 0) = 0$. The solution to this PDE satisfies

$$\frac{x}{\kappa}\log\left(\frac{\lambda}{x}T\right) \leq v(x, T) \leq \frac{\lambda}{\kappa e}T. \tag{4.4}$$

REMARK 4.2. In the above notation x is the number of shares, which is fixed across the problems. When we are taking the continuous liquidation limit, we let $\Delta\downarrow 0$, the size of trading units, while taking the number of units n as $n\Delta = x$, for x constant.

Proof. $V^{\Delta}(n\Delta, T)$ satisfies the HJB equation

$$-\partial_T V^{\Delta}(n\Delta, T) + \sup\nolimits_{s\geq 0} \lambda\Delta^{-1}e^{-\kappa s}(V^{\Delta}((n-1)\Delta, T) - V^{\Delta}(n, T) + s\Delta) = 0,$$

which can be written as

$$\partial_T V^{\Delta}(n\Delta, T) = \frac{\lambda}{\kappa\Delta}\exp\left(-1 + \Delta^{-1}\kappa(V^{\Delta}((n-1)\Delta, T) - V^{\Delta}(n, T))\right). \tag{4.5}$$

Letting $B := \frac{\lambda}{\kappa e}$, integrating and using $V^{\Delta}(\cdot, 0) \equiv 0$ we find for $n = 1$ that

$$V^{\Delta}(\Delta, T) = \frac{\Delta}{\kappa}\log(1 + B\kappa\Delta^{-1}T). \tag{4.6}$$

Iterating over n the separable ODE for $V^{\Delta}(n, \cdot)$ in (4.5) we obtain (4.2). The expression for the optimal spread follows from $s^*(n\Delta, T) = \frac{1}{\kappa} + \frac{V^{\Delta}(n\Delta) - V^{\Delta}((n-1)\Delta)}{\Delta}$.

The proof that as $\Delta \downarrow 0$, $V^{\Delta}(n\Delta, T) \to v(x, T)$ uniformly on compacts can be proven as in Theorem 3.2.

Observe that both bounds in (4.4) satisfy (4.3). To prove the lower bound, let us introduce the following function:

$$\tilde{V}^{\Delta}(x, T) = \frac{\Delta}{\kappa}\log\left(\frac{1}{n!}\left(\frac{\lambda T}{\Delta e}\right)^n\right).$$

Clearly, $\tilde{V}^{\Delta} \leq V^{\Delta}$. From Stirling's formula we know that

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n,$$

where we use $\sim$ to indicate that the ratio of the left- to the right-hand side converges to 1 as $n \to \infty$. As a result, $\tilde{V}^{\Delta}(x, T) \sim x/\kappa \log(\lambda T/x)$, recalling that $x = n\Delta$. Now the lower bound in (4.4) follows since $v(x, T) \geq \lim_{\Delta\downarrow 0} \tilde{V}^{\Delta}$. We could have provided an alternative proof using a comparison theorem for the first-order nonlinear PDE (4.3) since in fact $x/\kappa \log(\lambda T/x)$ also satisfies this PDE with a smaller boundary value at $T = 0$. We preferred to be more constructive in our proof.

The fact that $\lambda T/(\kappa e)$ is an upper bound on v follows directly from the observation that $v_T \leq \lambda/(\kappa e)$ (recalling that v is increasing in x) and that $v(x, 0) = 0$. □

REMARK 4.3. Since the trading rate is bounded $\Lambda(s^*) \leq \lambda e^{-1}$, for $x > \lambda e^{-1}T$ the full inventory cannot be liquidated by horizon T. Therefore, in the region $\mathcal{D} := \{(x, T) : x > \lambda e^{-1}T\}$, v is independent of x and the upper bound is tight: $v(x, T) = \frac{\lambda}{\kappa e}T$ on $\mathcal{D}$.

REMARK 4.4. Here, we will determine the shape of $t \to X_t^{(0),x}$ in Remark 3.1 for the exponential order books. First,

$$dX_t^{(0),x} = -\Lambda\big(s\big(X_t^{(0),x}, T - t\big)\big)dt, \tag{4.7}$$

where Λ is given by (4.1). On the other hand, $s(X_t^{(0),x}, T - t) = \frac{1}{\kappa} + v_x(X_t^{(0),x}, T - t)$. Using this relationship, along with (4.3), which implies that $t \mapsto s(X_t^{(0),x}, T - t)$ is a constant function (let us denote that value by s^*), it follows from (4.7) that

$$\frac{d^2 X_t^{(0),x}}{dt^2} = 0, \quad t \in [0, T],$$

i.e., $t \mapsto X_t^{(0),x}$, $t \in [0, T]$, is a strictly decreasing linear function. In fact, one can compute s^* by maximizing the value function (3.4) (after replacing power rate with exponential) over constant spreads (since the optimal spread is known to be a constant). This yields that $s^* = \frac{1}{\kappa}\log\left(\frac{\lambda T}{x}\right)$ if $\lambda T/x \geq e$. Otherwise $s^* = 1/\kappa$. The expression for the optimal spread and (4.7) in turn imply that if $x \leq \lambda e^{-1}T$, $X_t^{(0),x} = x(1 - \frac{t}{T})$ (which is 0 at T) and if $x > \lambda e^{-1}T$ we have that $X_t^{(0),x} = x - \lambda t/e$ (which remains strictly positive at T).

4.2. Infinite Horizon

We also have closed-form expressions for the infinite horizon case.

PROPOSITION 4.5. *For exponential-decay LOB with $T = +\infty$ and discounting rate $r > 0$ we have*

$$V^{\Delta}(x) = \frac{\Delta}{\kappa}\mathbf{W}\left(\lambda r^{-1}\Delta^{-1}\exp\left(\kappa\frac{V^{\Delta}(x-\Delta)}{\Delta} - 1\right)\right), \quad x \in \{0, \Delta, \ldots\}, \tag{4.8}$$

with $V^{\Delta}(0) = 0$, where $\mathbf{W}$ is the Lambert-W function (or the double-log function defined as $z = \mathbf{W}(y)$ for $ze^z = y$).

As $\Delta \downarrow 0$, $V^{\Delta}(x) \to v(x)$ uniformly on compacts where

$$li\left(\frac{e\kappa r v(x)}{\lambda}\right) = -\frac{erx}{\lambda}, \tag{4.9}$$

and $li(y) := \int_0^y \frac{1}{\log t}\,dt$ is the logarithmic integral function.

Proof. The HJB equation for $V^{\Delta}(x)$ is

$$-r\,V^{\Delta}(x) + \sup_{s\geq 0}\lambda\Delta^{-1}e^{-\kappa s}\left(V^{\Delta}(x-\Delta) - V^{\Delta}(x) + s\Delta\right) = 0.$$

Using the optimizer $s^{(\Delta)} = \frac{1}{\kappa} + \frac{V^{\Delta}(x) - V^{\Delta}(x-\Delta)}{\Delta}$ we reduce to

$$V^{\Delta}(x) = \frac{\lambda\Delta}{\kappa r}\exp\left(-1 + \kappa\frac{V^{\Delta}(x-\Delta) - V^{\Delta}(x)}{\Delta}\right),$$

which has closed-form solution given by equation (4.8). Arguments similar to Theorem 3.2 imply that $V^{\Delta} \to v$ uniformly on compacts and the continuous inventory limit satisfies

$$-rv(x) + \sup_{s\geq 0}\lambda e^{-\kappa s}(s - v'(x)) = 0. \tag{4.10}$$

Solving for v' we obtain

$$v'(x) = -\frac{1}{\kappa}\left\{1 + \log\left(\frac{\kappa r v(x)}{\lambda}\right)\right\}.$$

The last nonlinear first-order ODE has closed-form solution given in (4.9). Asymptotically $\lim_{x\to\infty} v(x) = \frac{\lambda}{\kappa r e}$ and the optimal spread is

$$s^{(0)}(x) = \frac{1}{\kappa}\left\{\log\left(\frac{\lambda}{\kappa r v(x)}\right)\right\}. \tag{4.11}$$

□

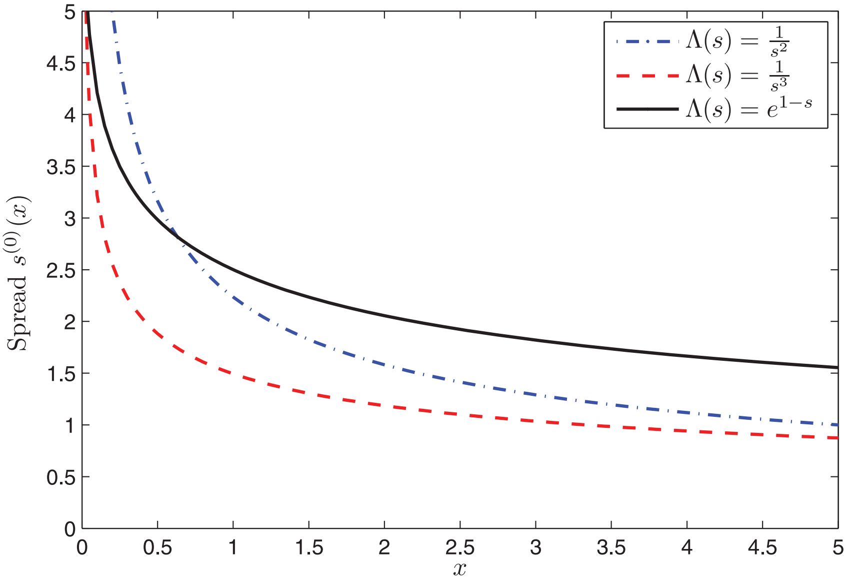

FIGURE 4.1. Optimal controls for power-law and exponential-decay order books. We take $r = 0.1$ and depth functions $\Lambda(s) \in \{s^{-2}, s^{-3}, e^{1-s}\}$, which have been normalized such that $\Lambda(1) = 1$ in all three cases. The plot shows the resulting fluid limit spreads $s^{(0)}(x)$.

We note that since v is increasing in x, $x \mapsto s^{(0)}(x)$ in (4.11) is decreasing and so v is concave. As before, $\lim_{x\to 0} s^{(0)}(x) = +\infty$ so the control space remains unbounded, however the pay-off rate $s^{(0)}\Lambda(s^{(0)})$ is bounded. Moreover, a direct check verifies that V^Δ and v are inversely proportional to the exponential depth parameter κ, i.e., doubling κ (making the order book more shallow) halves V^Δ and v, and correspondingly halves the optimal spreads $s^{(\Delta)}$ and $s^{(0)}$.

Figure 4.1 graphically illustrates the difference between exponential-decay and power-law LOBs. As observed, for an exponential LOB, s^* is bounded away from zero, while $\lim_{x\to\infty} s^*(x) = 0$ in power LOBs. Moreover, while $\lim_{x\downarrow 0} s^{(0)}(x) = +\infty$ in any LOB, the rate is much slower in an exponential LOB (due to thinner tail for large spreads) compared to power-LOB.

5. DISCUSSION AND FURTHER EXTENSIONS

5.1. Numerical Example: Convergence to the Fluid Limit

To illustrate the convergence to the fluid limit consider the problem of selling up to $x = 5$ blocks of shares in a power-law LOB with $\Lambda(s) = s^{-2}$. We suppose that a block corresponds to 100 shares and that the minimal trading unit is either 5 or 1 shares, i.e., $\Delta = 0.05$ and $\Delta = 0.01$, respectively. For a fixed Δ, we can easily compute $V^\Delta(x = n\Delta)$ or $v(x)$ using the results in Section 2. Figure 5.1 illustrates the percent difference between V^Δ and the fluid limit v. As shown in Proposition 3.4, V^Δ is decreasing in Δ and $\lim_{\Delta\downarrow 0} V^\Delta = v$. We observe that the convergence is quite rapid in x.

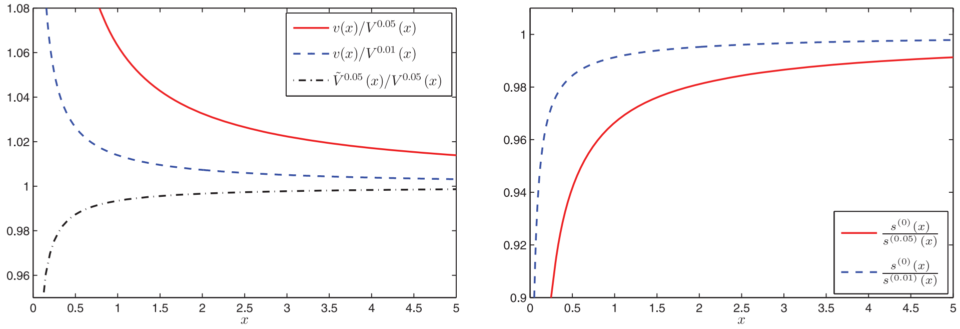

FIGURE 5.1. Convergence to the fluid limit. Left panel: the ratio between discrete and continuous $V^{\Delta}(x)/v(x)$ for $\Delta = 0.05$ and $\Delta = 0.01$. Additionally, we plot $\tilde{V}^{\Delta}/V^{\Delta}$ as defined in equation (5.1). Right panel: ratio of the fluid limit optimal control $s^{(0)}(x)$ to the discrete $s^{(\Delta)}(x)$ for $\Delta \in \{0.01, 0.05\}$.

The right panel of Figure 5.1 shows that the controls themselves are also very close. We observe that $s^{(\Delta)} \searrow s^{(0)}$. Note that here we are essentially comparing $s^{(0)}$ with its right-sided Riemann-sum approximation, since $s^{(\Delta)}(x)$ corresponds to the spread charged for all shares in $[x, x - \Delta)$ while $s^{(0)}(x)$ corresponds to the marginal spread *at* x. Accordingly, better approximations, such as $\check{s}^{(\Delta)}(x) := \frac{1}{\Delta}\int_{x-\Delta}^{x} s^{(0)}(u)\,du$, would make the discrete and fluid controls even closer.

Given the simple expression for the fluid limit control $s^{(0)}(x)$, a useful approximation is to use a discretized version of $s^{(0)}$ as an approximately optimal control for V^{Δ}. Let

$$\tilde{V}^{\Delta}(x) := \mathbb{E}\left[\sum_{i=1}^{x/\Delta} e^{-r\tau_i} s^{(0)}(x - i\Delta)\right], \qquad x \in \{0, \Delta, 2\Delta, \dots\}, \tag{5.1}$$

represent the expected gains from a discrete strategy which uses a spread of $s^{(0)}((n-i)\Delta)$ for the ith trade of size Δ. In the left panel of Figure 5.1 we see that this approximation is excellent for V^{Δ} even for moderate values of x (less than 1% difference for $\Delta = 0.01$ and $x > 1$).

5.2. Execution Curves

A popular way of describing a trade execution algorithm is through the execution curve $t \mapsto X_t/X_0$, see, e.g., Almgren (2000, 2003); Guéant et al. (2011). In our model with execution risk, X_t is a random variable, and we will therefore consider the natural analogue of average execution curve $E(x,t) := \mathbb{E}[X_t^{*,x}]$, where $X^{*,x}$ is the remaining inventory at $t \leq T$ starting with initial condition $X_0^{*,x} = x$. The baseline case where $\bar{E}(x,t) = x(1 - t/T)$ is linear, corresponds to "linear price impact" or zero-risk-aversion in Almgren (2000) and implies that the average trading rate is constant.

For notational convenience we temporarily fix $\Delta = 1$. Recall that

$$dX_t^{*,x} = -\Lambda\big(s^*\big(X_t^{*,x}, T - t\big)\big)\,dt + dM_t,$$

where (M_t) is a martingale (the compensated order *departure* process), which implies by an application of Itô's formula that $(E(x,t))_{x=0}^{\infty}$ satisfies the system of inhomogenous linear ODEs

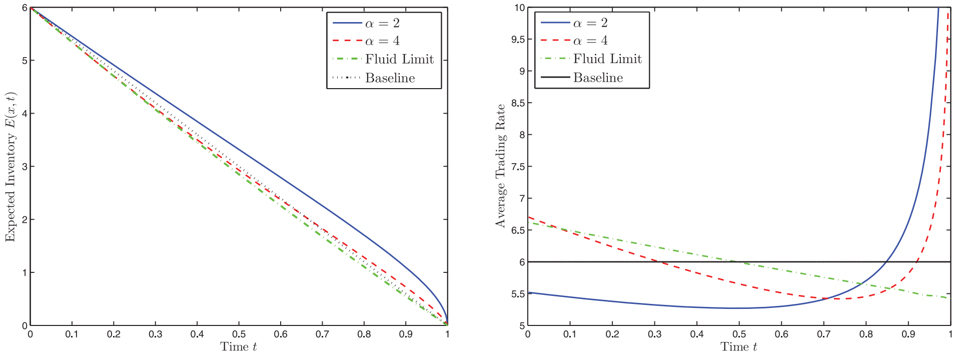

FIGURE 5.2. Execution curves for different power-law books. We take $X_0 = 6$, $r = 0.1$, $T = 1$, and depth functions $\Lambda(s) \in \{s^{-2}, s^{-4}\}$. Left panel: average inventory $E(x, t)$ as a function of time t. Right panel: average trading rate as a function of time.

$$\frac{dE(x,t)}{dt} = \Lambda(s^*(x, T-t))(E(x-1,t) - E(x,t)), \quad E(x,0) = x,$$

with $E(0, t) \equiv 0$. Any finite collection of these ODEs can be solved analytically using integrating factors, or numerically with any standard solver.

Thus, whenever we have an explicit formula for the optimal spread $s^*(n, T)$, $E(x, t)$ is also available analytically (or more practically as a solution of an ODE). For the case of power-law order books, we obtain from Proposition 2.1 that

$$\Lambda(s^*(k, T-t)) = \frac{\lambda^{-\frac{1}{\alpha-1}}(\alpha r c_k)^{\alpha/(\alpha-1)}}{1 - \exp(-\alpha r(T-t))}.$$

We observe that depending on the parameter values (in particular initial inventory k vis-à-vis order shape α), $\Lambda(s^*(k, T))$ could be smaller or bigger than k/T, i.e., the ordering between the initial trading rate and constant trading is ambiguous. At the other end, as $t \to T$, the spread $s^*(k, t)$ goes to zero and consequently $\lim_{t\to T} dE(x, t)/dt = -\infty$.

Figure 5.2 shows that as α increases, the shape of $t \mapsto E(x, t)$ changes substantially. In particular for large α (corresponding to "thinner" power laws), the execution curve has an S-shape, with trading rate high in the beginning and end of the time interval. On the other hand, for α small, the execution curve lies entirely above the baseline, i.e., the limit order trader consistently executes slower. This occurs due to the two competing effects of trying to extract profit (which slows execution) and the time decay, i.e., the need to make the deadline which speeds up trading. We observe that when the LOB is deep (small α), the profit effect dominates; this is a new phenomenon compared to most existing models, such as Guéant et al. (2011).

Finally, as a comparison, Figure 5.2 also shows the deterministic case $t \mapsto X_t^{(0),x}$, see (3.6). In the latter case, $t \mapsto X_t^{(0),x}$ is strictly convex, which contrasts strongly with the pre-limit situation. As $\Delta \to 0$, execution risk vanishes and the discounting effect takes over, making the investor sell more in the beginning. Indeed, in the fluid limit the investor can smoothly drive $X^{(0),x}$ to zero, while for $\Delta > 0$, $\mathbb{P}(X_t^{*,x} > 0) > 0$ for any $t < T$, but $X_T^{*,x} = 0$ since $dE(x,t)/dt|_{t=T-} = -\infty$.

In the exponential order book case with no discounting, we observe that the trading rate $\Lambda(s^*(n, T-t))$ is *independent* of the depth parameter κ. Moreover, numerical experiments suggest that $E(x, t)$ is (slightly) convex in t, i.e., the trading rate is monotonically decreasing in time. This agrees with the classical results of Almgren (2000). Finally, we recall that Remark 4.4 shows that in the fluid limit, the execution rate is constant over time, and $t \mapsto X^{(0),x}$ is linear. This occurs because $(X_t^{(0),x}, t)$ is the characteristic curve of the PDE given by (4.3). This phenomenon resembles the constant trading rate in Alfonsi et al. (2010) who studied (continuous) trading through market orders only, with the LOB depth function driving the price impact mechanism. Again, we find a sharp dichotomy between the deterministic limit where $X_T^{(0),x} = 0$ for $x \leq \lambda e^{-1} T$ and the stochastic version where $E(x, T) > 0$ strictly for all $x > 0$.

5.3. General Order Book Depth Functions

Our basic setting can be readily extended to allow for more sophisticated or complex models. Below we review several such extensions; for ease of presentation we treat them in the stationary infinite-horizon setting.

Let us revisit the optimal execution problem for a generic order book depth function $\Lambda(s)$. In general, there are no closed-form expressions for $V(n)$ and the continuous fluid limit $v(x)$ becomes a useful analytic tool to understand the solution structure. In that regard, both Theorem 3.2 and Corollary 3.5 continue to hold under some reasonable assumptions on the intensity function Λ.

THEOREM 5.1. *Consider the optimal liquidation problem on infinite horizon with a general intensity depth function Λ. Then the statements of Theorem 3.2 hold, and if we further assume that the function $x \mapsto \Lambda(x)$ is decreasing and that*

$$\frac{\Lambda(x)\Lambda''(x)}{(\Lambda'(x))^2} < 2, \quad \forall x \in \mathbb{R}_+, \tag{5.2}$$

then both V^Δ and v are concave, the corresponding controls $s^{(\Delta)}$ and $s^{(0)}$ are decreasing and the conclusion of Corollary 3.5 still holds.

REMARK 5.2. For power-law LOBs, condition (5.2) holds precisely when $\alpha > 1$, while for exponential LOBs it always holds. Both of these order books have decreasing intensity functions.

Proof. The proof of Theorem 3.2 can be done without much change since we did not make use of any special properties of $\Lambda(s)$ there. We will prove the stated concavity and monotonicity properties from which the statement of Corollary 3.5 follows immediately as before.

By time-stationarity between trading dates the controls are constant and the dynamic programming principle until the first jump time for $V^\Delta(x)$ gives

$$\begin{aligned} V^\Delta(x) &= \sup_{s \geq 0} \int_0^\infty \frac{\Lambda(s)}{\Delta} e^{-(\Lambda(s)\Delta^{-1}+r)t} (s\Delta + V^\Delta(x-\Delta))\, dt \\ &= \sup_{s \geq 0} \frac{\Lambda(s)}{\Lambda(s) + r\Delta} (s\Delta + V^\Delta(x-\Delta)). \end{aligned}$$

Differentiating the right-hand side with respect to s, the first-order condition for $s^* \equiv s^{(\Delta)}(x)$ is

$$r\Lambda'(s^*)(s^*\Delta + V^\Delta(x-\Delta)) + \Lambda(s^*)(\Lambda(s^*) + r\Delta) = 0$$
$$\iff \quad rV^\Delta(x-\Delta) = -rs^*\Delta - \frac{\Lambda(s^*)}{\Lambda'(s^*)}(\Lambda(s^*) + r\Delta) := F(s^*).$$

V^Δ is non-decreasing; therefore, if the derivative of F is negative, then $s^{(\Delta)}(x)$ decreases in x. Explicitly,

$$(5.3) \quad F'(s^*) = -(r\Delta + \Lambda(s^*))\left[2 - \frac{\Lambda\Lambda''}{(\Lambda')^2}(s^*)\right] < 0 \quad \iff \quad 2 > \frac{\Lambda(s^*)\Lambda''(s^*)}{(\Lambda'(s^*))^2}.$$

Thus, (5.2) is sufficient for $x \mapsto s^*(x)$ to be decreasing. Under this assumption and the assumption that Λ is decreasing we would have that $x \mapsto \frac{\Lambda(s^*(x))}{\Lambda(s^*(x))+r\Delta}$ is increasing, and as a result

$$\begin{aligned}
V^\Delta(x) - V^\Delta(x-\Delta) &= \frac{\Lambda(s^*(x))}{\Lambda(s^*(x)) + r\Delta}(s^*(x)\Delta + V^\Delta(x-\Delta)) \\
&\quad - \frac{\Lambda(s^*(x-\Delta))}{\Lambda(s^*(x-\Delta)) + r\Delta}(s^*(x-\Delta)\Delta + V^\Delta(x-2\Delta)) \\
&\leq \frac{\Lambda(s^*(x))}{\Lambda(s^*(x)) + r\Delta}\{s^*(x-\Delta)\Delta + V^\Delta(x-\Delta) - s^*(x-\Delta)\Delta - V^\Delta(x-2\Delta)\} \\
&\leq V^\Delta(x-\Delta) - V^\Delta(x-2\Delta),
\end{aligned}$$

so that V^Δ is concave.

Concavity of V^Δ on the other hand implies the concavity of v. This is thanks to the first assertion of the theorem from which we know that V^Δ converges to v uniformly on compacts. Next we will show that s is a decreasing function. The value function v satisfies the first-order PDE:

$$\sup_{s\geq 0} \Lambda(s)(s - v') - rv = 0, \quad v(0) = 0.$$

Optimizing over s yields

$$v'(x) = s^{(0)}(x) + \frac{\Lambda(s^{(0)}(x))}{\Lambda'(s^{(0)}(x))}.$$

Hence v is concave if and only if the right-hand side above is a decreasing function of x. However, it follows from (5.2) that the function

$$y \mapsto y + \frac{\Lambda(y)}{\Lambda'(y)}, \quad y \in \mathbb{R}_+,$$

is increasing. As a result the concavity of v, which we have already shown, is equivalent to $x \mapsto s^{(0)}(x)$ decreasing. □

5.4. Regime Switching Market Liquidity

Empirical evidence suggests that market liquidity is not constant (Cartea and Jaimungal 2010). As a first step toward capturing more complex liquidity behavior, we consider a simple regime-switching model for market activity level in the power-law LOBs. More precisely, suppose that arrival rates are modulated by a two-state Markov chain M with states $\{0, 1\}$, 0 representing an active market and 1 representing a slow market. We will denote the transition rate from 0 to 1 by θ_0, and the transition rate from 1 to 0 by θ_1. Under regime 0, the arrival rates of the orders are $\Lambda_0(s) = \lambda_0/s^\alpha$, and under regime 1, the arrival rates of the orders are $\Lambda_1(s) = \lambda_1/s^\alpha$. We will take $\lambda_0 > \lambda_1$. We assume that M is observed and known by market participants.

Denote by $U(n)$ (respectively $W(n)$) the infinite-horizon value function for an inventory of n shares under the active (resp. slow) market regime. The value functions satisfy the following system of equations:

$$A_\alpha\lambda_0[U(n) - U(n-1)]^{1-\alpha} - rU(n) + \theta_0[W(n) - U(n)] = 0,$$
$$A_\alpha\lambda_1[W(n) - W(n-1)]^{1-\alpha} - rW(n) + \theta_1[U(n) - W(n)] = 0,$$

with terminal condition $U(0) = W(0) = 0$. The continuous selling approximation of these functions, which we denote by u and w, respectively satisfy the following system of ODEs:

$$\begin{cases} A_\alpha\lambda_0 u_x^{1-\alpha} - (r+\theta_0)u(x) + \theta_0 w(x) = 0, \\ A_\alpha\lambda_1 w_x^{1-\alpha} - (r+\theta_1)w(x) + \theta_1 u(x) = 0, \end{cases} \tag{5.4}$$

with $u(0) = w(0) = 0$.

PROPOSITION 5.3. *The solutions to (5.4) are $u(x) = c_0^* x^p$ and $w(x) = c_1^* x^p$ where $p = (\alpha - 1)/\alpha$ and*

$$\left(\frac{\lambda_1}{r\alpha}\right)^{1/\alpha} < c_1^* < c_0^* < \left(\frac{\lambda_0}{r\alpha}\right)^{1/\alpha}. \tag{5.5}$$

Observe that the bounds in (5.5) correspond to the single-regime solutions given in (3.5) with $T = +\infty$.

Proof. We begin with an ansatz of $u(x) = c_0 x^p$ and $w(x) = c_1 x^p$ with p given in the statement of the proposition. Comparing with (5.4), the coefficients c_0 and c_1 need to satisfy

$$A_\alpha\lambda_0 p^{1-\alpha} c_0^{1-\alpha} - (r+\theta_0)c_0 + \theta_0 c_1 = 0,$$
$$A_\alpha\lambda_1 p^{1-\alpha} c_1^{1-\alpha} - (r+\theta_1)c_1 + \theta_1 c_0 = 0.$$

Re-writing as

$$\begin{aligned} c_0 &= \frac{r+\theta_1}{\theta_1}c_1 - \frac{\lambda_1}{\alpha\theta_1}c_1^{1-\alpha}, \\ c_1 &= \frac{r+\theta_0}{\theta_0}c_0 - \frac{\lambda_0}{\alpha\theta_0}c_0^{1-\alpha}, \end{aligned} \tag{5.6}$$

it easily follows that this system of equations has a unique solution (c_0^*, c_1^*). Indeed, c_0 as a function of c_1 is strictly increasing and goes from $-\infty$ at $c_1 = 0$ to ∞ and $c_1 = \infty$. Similarly, c_1 as a function of c_0 is also strictly increasing from $-\infty$ at $c_0 = 0$ to ∞ at $c_0 = \infty$. It directly follows from these facts that these two curves intersect and do so at only one point. Moreover, the identity function $c_0 = c_1$ intersects the first function (5.6) first,

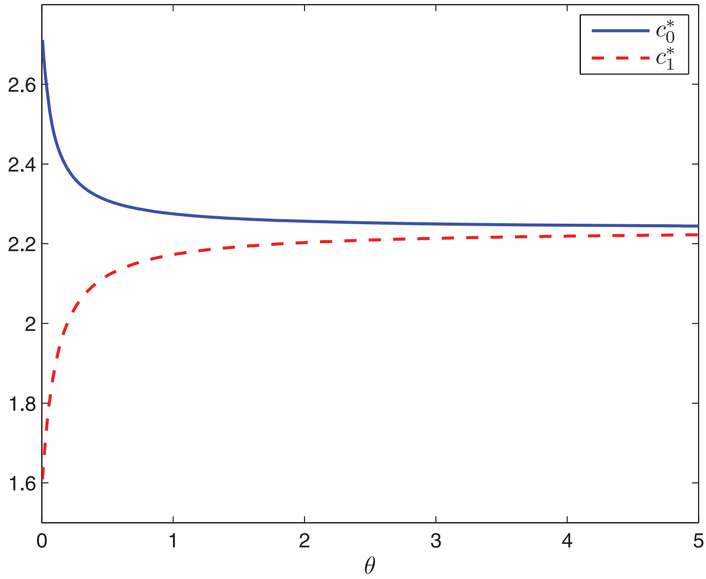

FIGURE 5.3. Regime switching model. We take $\alpha_0 = 2$, $\lambda_0 = 1.5$, $\lambda_1 = 0.5$, and $r = 0.1$. The regime-switching rates are equal $\theta_0 = \theta_1 = \theta$.

and the second function in the same equation last. This proves the ordering in (5.5) and concludes the proof. □

Proposition 5.3 shows that the asking spread will always be higher under the active market regime when the order book is deeper.

Figure 5.3 illustrates the impact of multiple liquidity regimes. We take $\lambda_0 = 1.5$, $\lambda_1 = 0.5$ so that trade intensity is tripled in the active regime. We plot c_i^* as a function of $\theta_0 = \theta_1 = \theta$ for $r = 0.1$ and $\alpha = 2$. As $\theta \to 0$, we have $c_i^* \to \sqrt{0.5\lambda_i r^{-1}}$, while as $\theta \to \infty$, $c_i^* \to \sqrt{\frac{1}{2r}\frac{\lambda_0+\lambda_1}{2}}$ the fast-switching limit.

5.5. Two-Exchange Multi-Scale Model

Another possibility is to consider an investor trading on multiple venues. For example, suppose the investor can liquidate her holdings through two different exchanges, with each exchange possessing its own LOB. To distinguish the two exchanges, we suppose that on exchange C(ontinuous) the orders are infinitesimally small, but on exchange L(arge) they are of *large but finite size* relative to the total order size. More precisely, we assume that in the continuous limit, the exchange C orders are infinitesimal, but exchange L orders are of size δ. If the remaining inventory is less than x we assume that the next trade on exchange L will liquidate the entire x. In other words, actual trades on exchange L are of size $\min(\delta, x)$.

To keep the model tractable, we assume that each exchange has power-law depth with identical depth parameter $\alpha > 1$. The resulting time-stationary value function $v(x)$ solves

$$\sup_{s_0 \geq 0} \lambda_0 \frac{s_0 - v_x}{s_0^\alpha} + \sup_{s_1 \geq 0} \lambda_1 \frac{(\delta \wedge x)s_1 - (v(x) - v((x-\delta)_+))}{s_1^\alpha} - rv = 0, \quad v(0) = 0. \tag{5.7}$$

Plugging in the first-order optimizers leads to

$$(5.8)\quad A_\alpha \lambda_0 v'(x)^{1-\alpha} + A_\alpha \lambda_1 (x \wedge \delta)^\alpha (v(x) - v((x-\delta)_+))^{1-\alpha} - rv = 0, \quad v(0) = 0.$$

LEMMA 5.4. *There is a unique solution to (5.8).*

Proof. Equation (5.8) is a first-order nonlinear delay ODE and can be solved by successive patching. Namely, first solve the ODE

$$(5.9)\qquad A_\alpha \lambda_0 v'_{(0)}(x)^{1-\alpha} + A_\alpha \lambda_1 x^\alpha (v_{(0)}(x))^{1-\alpha} - r v_{(0)} = 0,$$

with $v_{(0)}(0) = 0$ on $[0, \delta]$. We then solve

$$(5.10)\qquad A_\alpha \lambda_0 v'_{(1)}(x)^{1-\alpha} + A_\alpha \lambda_1 \delta^\alpha (v_{(1)}(x) - v_{(0)}(x-\delta))^{1-\alpha} - r v_{(1)} = 0$$

on $[\delta, 2\delta]$ with initial condition $v_{(1)}(\delta) = v_{(0)}(\delta)$. In (5.10) we treat $v_{(0)}$ as a source term, observing that $v_{(0)}(x-\delta)$ with $x \in [\delta, 2\delta]$ has already been computed before. Proceeding in this fashion, we finally set $v(x) = v_{(n)}(x)$ for $x \in [n\delta, (n+1)\delta]$ to recover the global solution. On each of the intervals $[n\delta, (n+1)\delta]$ the corresponding ODE has a locally Lipschitz driver so classical results give existence/uniqueness of solution $v_{(n)}$. □

REMARK 5.5. Numerical computation of the solution of (5.9) should be handled with care since $v'_{(0)}(0) = \infty$. We get around this singularity using the following observation: For x small enough, the benefit of large orders is negligible since the probability of getting a large order is very small. Therefore, close to zero, $v(x) \simeq v_0(x) = (\frac{\lambda_0}{\alpha r})^{1/\alpha} x^{\frac{\alpha-1}{\alpha}}$ from (3.5).

We also remark that the solution to (5.9) is in general no longer concave, with concavity likely to fail around the knots $\delta, 2\delta, \ldots$, where the derivative v' does not exist.

Typically, trading intensity on the small-order exchange is several magnitudes larger than via the big trades (done through, e.g., a proprietary dark pool, see, e.g., Klöck et al. (2011) where a single large dark pool trade liquidates the entire position), so $\lambda_0 \gg \lambda_1$. Fixing the time-scale as $\lambda_0 = 1$, we are therefore led to consider an asymptotic expansion in small λ_1. Formally, let $\lambda_1 = \bar{\lambda}\varepsilon$ for ε small and consider a power series expansion in ε,

$$v(x) = v_0(x) + \varepsilon v_1(x) + \varepsilon^2 v_2(x) + \cdots,$$

Plugging into (5.8) and matching powers of ε we find that $v_0(x)$ solves the 1-exchange problem of (3.5), so that $v_0(x) = (\frac{\lambda_0}{\alpha r})^{1/\alpha} x^{\frac{\alpha-1}{\alpha}}$. Next,

$$A_\alpha \alpha(1-\alpha) r x v_1'(x) + A_\alpha \bar{\lambda} (\delta \wedge x)^\alpha \left(\frac{\lambda_0}{\alpha r}\right)^{\frac{1-\alpha}{\alpha}} \left(x^{\frac{\alpha-1}{\alpha}} - (x-\delta)_+^{\frac{\alpha-1}{\alpha}}\right)^{1-\alpha} - r v_1(x) = 0.$$

This is a first-order linear ODE with nonconstant coefficients and therefore $v_1(x)$ can be expressed in closed form using integrating factors as

$$(5.11)\qquad v_1(x) = \begin{cases} C_1 x^{2-\alpha^{-1}}, & x \le \delta, \\ x^{-B_\alpha} \cdot \displaystyle\int_0^x C_2 y^{B_\alpha + \alpha - 1} \left(y^{\frac{\alpha-1}{\alpha}} - (y-\delta)^{\frac{\alpha-1}{\alpha}}\right)^{1-\alpha} dy, & x > \delta, \end{cases}$$

with $C_1 = \frac{\bar{\lambda}\alpha r}{(1-\alpha)} (\frac{\lambda_0}{\alpha r})^{(1-\alpha)/\alpha} \frac{1}{B_\alpha - 2 + \alpha^{-1}}$, $C_2 = A_\alpha \bar{\lambda} \delta^\alpha B_\alpha (\frac{\lambda_0}{\alpha r})^{(1-\alpha)/\alpha}$ and $B_\alpha = \frac{\alpha^{\alpha-1}}{(\alpha-1)^\alpha}$. The latter integral only involves powers of y and can be easily computed numerically. Similarly, the equations for higher-order terms are again first-order linear ODEs and so v_2, etc., can be written iteratively in closed form.

REFERENCES

ALFONSI, A., and A. SCHIED (2010): Optimal Trade Execution and Absence of Price Manipulations in Limit Order Book Models, *SIAM J. Financial Math.* 1, 490–522.

ALFONSI, A., A. FRUTH, and A. SCHIED (2010): Optimal Execution Strategies in Limit Order Books with General Shape Functions, *Quant. Finance* 10(2), 143–157.

ALMGREN, R. (2003): Optimal Execution with Nonlinear Impact Functions and Trading-Enhanced Risk, *Appl. Math. Finance* 10, 1–18.

ALMGREN, R. (2000): Optimal Execution of Portfolio Transactions, *Journal of Risk* 3, 5–39.

AVELLANEDA, M., and S. STOIKOV (2008): High-Frequency Trading in a Limit Order Book, *Quant. Finance* 8(3), 217–224.

BARDI, M., and I. CAPUZZO-DOLCETTA (1997): *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Systems & Control: Foundations & Applications. Boston, MA: Birkhäuser.

BARLES, G., and P. E. SOUGANIDIS (1991): Convergence of Approximation Schemes for Fully Nonlinear Second Order Equations, *Asymptotic Anal.* 4(3), 271–283.

BÄUERLE, N. (2000): Asymptotic Optimality of Tracking Policies in Stochastic Networks, *Ann. Appl. Probab.* 10(4), 1065–1083.

BÄUERLE, N. (2001): Discounted Stochastic Fluid Programs, *Math. Oper. Res.* 26(2), 401–420.

BÄUERLE, N. (2002): Optimal Control of Queueing Networks: An Approach via Fluid Models, *Adv. Appl. Probab.* 34(2), 313–328.

BÄUERLE, N., and U. RIEDER (2009): MDP Algorithms for Portfolio Optimization Problems in Pure Jump Markets, *Finance Stoch.* 13(4), 591–611.

BAYRAKTAR, E., and M. LUDKOVSKI (2011): Optimal Trade Execution in Illiquid Financial Markets, *Math. Finance* 21(4), 681–701.

BERTSEKAS, D. P., and S. E. SHREVE (1978): Stochastic Optimal Control, in *Mathematics in Science and Engineering*. Vol. 139, New York: Academic Press Inc.

CARTEA, A., and S. JAIMUNGAL (2010): Modeling Asset Prices for Algorithmic and High Frequency Trading. Available at SSRN: http://ssrn.com/abstract=1722202. Accessed April 19, 2012.

CONT, R., S. STOIKOV, and R. TALREJA (2010): A Stochastic Model for Order Book Dynamics, *Operations Research* 58, 549–563.

CONT, R., and A. DE LARRARD (2010): Price Dynamics in a Markovian Limit Order Market, Available at SSRN: http://ssrn.com/abstract=1735338. Accessed April 19, 2012.

CRANDALL, M. G., H. ISHII, and P.-L. LIONS (1992): User's Guide to Viscosity Solutions of Second Order Partial Differential Equations, *Bull. Amer. Math. Soc. (N.S.)* 27(1), 1–67.

DAY, M. V. (2011): Weak Convergence and Fluid Limits in Optimal Time-To-Empty Queueing Control Problems, *Appl. Math. Optim.* 64(3), 339–362.

FLEMING, W. H., and H. M. SONER (2006): Controlled Markov Processes and Viscosity Solutions, *Stochastic Modelling and Applied Probability*. Vol. 25, 2nd ed., New York: Springer.

GUÉANT, O., C.-A. LEHALLE, and J. F. TAPIA (2011): Optimal Portfolio Liquidation with Limit Orders. Available as Arxiv preprint arXiv:1106.3279.

GUILBAUD, F., and H. PHAM (2011): Optimal High Frequency Trading with Limit and Market Orders. Available at SSRN: http://ssrn.com/abstract=1871969. Accessed April 19, 2012.

KLÖCK, F., A. SCHIED, and Y. SUN (2011): Existence and Absence of Price Manipulation in a Market Impact Model with Dark Pool. Available at SSRN: http://ssrn.com/abstract=1785409. Accessed April 19, 2012.

KRATZ, P., and T. SCHOENEBORN (2010): Optimal Liquidation in Dark Pools. Available at SSRN: http://ssrn.com/abstract=1344583. Accessed April 19, 2012.

OBIZHAEVA, A. A., and J. WANG (2005): Optimal Trading Strategy and Supply/Demand Dynamics. Available at SSRN: http://ssrn.com/abstract=686168. Accessed April 19, 2012.

PIUNOVSKIY, A. B. (2009): Random Walk, Birth-and-Death Process and Their Fluid Approximations: Absorbing Case, *Math. Methods Oper. Res.* 70(2), 285–312.

PIUNOVSKIY, A. B., and Y. ZHANG (2011): Accuracy of Fluid Approximations to Controlled Birth-and-Death Processes: Absorbing Case, *Math. Methods Oper. Res.* 73, 159–187.

ROCKAFELLAR, R. T. (1997): *Convex Analysis*, Princeton Landmarks in Mathematics. Princeton, NJ: Princeton University Press.

WEBER, P., and B. ROSENOW (2005): Order Book Approach to Price Impact, *Quant. Finance* 5(4), 357–364.

MATHEMATICAL FINANCE

An International Journal of Mathematics, Statistics and Financial Economics

Mathematical Finance, Vol. 24, No. 4 (October 2014), 651–695

OPTIMAL TRADE EXECUTION AND PRICE MANIPULATION IN ORDER BOOKS WITH TIME-VARYING LIQUIDITY

ANTJE FRUTH

Technische Universität Berlin

TORSTEN SCHÖNEBORN

Deutsche Bank AG, London

MIKHAIL URUSOV

University of Duisburg-Essen

In financial markets, liquidity is not constant over time but exhibits strong seasonal patterns. In this paper, we consider a limit order book model that allows for time-dependent, deterministic depth and resilience of the book and determine optimal portfolio liquidation strategies. In a first model variant, we propose a trading-dependent spread that increases when market orders are matched against the order book. In this model, no price manipulation occurs and the optimal strategy is of the wait region/buy region type often encountered in singular control problems. In a second model, we assume that there is no spread in the order book. Under this assumption, we find that price manipulation can occur, depending on the model parameters. Even in the absence of classical price manipulation, there may be transaction triggered price manipulation. In specific cases, we can state the optimal strategy in closed form.

KEY WORDS: market impact model, optimal order execution, limit order book, resilience, time-varying liquidity, price manipulation, transaction-triggered price manipulation.

1. INTRODUCTION

Empirical investigations have demonstrated that liquidity varies over time. In particular, deterministic time-of-day and day-of-week liquidity patterns have been found in most markets, see, e.g., Chordia, Roll, and Subrahmanyam (2001), Kempf and Mayston (2008), and Lorenz and Osterrieder (2009). In spite of these findings, the academic literature on optimal trade execution usually assumes constant liquidity during the trading time horizon. In this paper, we relax this assumption and analyze the effects of deterministically[1] varying liquidity on optimal trade execution for a risk-neutral investor. We characterize optimal strategies in terms of a trade region and a wait region and find that optimal

We would like to thank Peter Bank and two anonymous referees for valuable suggestions that helped improve our paper.

Manuscript received September 2011; final revision received September 2012.

Address correspondence to Antje Fruth, Technische Universität Berlin, Institut für Mathematik, Sekr. MA 7-1, Strasse des 17. Juni 136, 10623 Berlin, Germany; e-mail: mail@antje-fruth.de.

[1] Not all changes in liquidity are deterministic; an additional stochastic component has been investigated empirically by, e.g., Esser and Mönch (2003) and Steinmann (2005). See, e.g., Fruth (2011) for an analysis of the implications of such stochastic liquidity on optimal trade execution.

DOI: 10.1111/mafi.12022

trading strategies depend on the expected pattern of time-dependent liquidity. In the case of extreme changes in liquidity, it can even be optimal to entirely refrain from trading in periods of low liquidity. Incorporating such patterns in trade execution models can hence lower transaction costs.

Time-dependent liquidity can potentially lead to price manipulation. In periods of low liquidity, a trader could buy the asset and push market prices up significantly; in a subsequent period of higher liquidity, he might be able to unwind this long position without depressing market prices to their original level, leaving the trader with a profit after such a round-trip trade. In reality, such round-trip trades are often not profitable due to the bid-ask spread: once the trader starts buying the asset in large quantities, the spread widens, resulting in a large cost for the trader when unwinding the position. We propose a model with trading-dependent spread and demonstrate that price manipulation does not exist in this model in spite of time-dependent liquidity. In a similar model with fixed zero spread, we find that price manipulation or transaction-triggered price manipulation (a term recently coined by Alfonsi, Schied, and Slynko 2011 and Gatheral, Schied, and Slynko 2012) can be a consequence of time-dependent liquidity. Phenomena of such type, i.e., existence of "illusory arbitrages," which disappear when bid-ask spread is taken into account, are also observed in different modeling approaches (see, e.g., section 5.1 in Madan and Schoutens 2011).

Our liquidity model is based on the limit order book market model of Obizhaeva and Wang (2006), which models both depth and resilience of the order book explicitly. The instantaneously available liquidity in the order book is described by the depth. Market orders issued by the large investor are matched with this liquidity, which increases the spread. Over time, incoming limit orders replenish the order book and reduce the spread; the speed of this process is determined by the resilience. In our paper, we generalize the model of Obizhaeva and Wang (2006) in that both depth and resilience can be independently time-dependent. In relation to the problem of optimal trade execution, we show that there is a time-dependent optimal ratio of remaining order size to bid-ask spread: If the actual ratio is larger than the optimal ratio, then the trader is in the "trade region" and it is optimal to reduce the ratio by executing a part of the total order. If the actual ratio is smaller than the optimal ratio, then the trader is in the "wait region" and it is optimal to wait for the spread to be reduced by future incoming limit orders before continuing to trade. We will see that allowing for time-varying liquidity parameters brings qualitatively new phenomena into the picture. For instance, it can happen that it is optimal to wait regardless of how big the remaining position is, while this cannot happen in the framework of Obizhaeva and Wang (2006).

Building on empirical investigations of the market impact of large transactions, a number of theoretical models of illiquid markets have emerged. One part of these market microstructure models focuses on the underlying mechanisms for illiquidity effects, e.g., Kyle (1985) and Easley and O'Hara (1987). We follow a second line that takes the liquidity effects as given and derives optimal trading strategies within such a stylized model market. Two broad types of market models have been proposed for this purpose. First, several models assume an instantaneous price impact, e.g., Bertsimas and Lo (1998), Almgren and Chriss (2001), and Almgren (2003). The instantaneous price impact typically combines depth and resilience of the market into one stylized quantity. Time-dependent liquidity in this setting then leads to executing the constant liquidity strategy in volume time or liquidity time, and no qualitatively new features occur. In a second group of models, resilience is finite and depth and resilience are separately modeled, e.g., Bouchaud et al. (2004), Obizhaeva and Wang (2006), Alfonsi, Fruth, and Schied

(2010), and Predoiu, Shaikhet, and Shreve (2011). Our model falls into this last group. Allowing for independently time-dependent depth and resilience leads to higher technical complexity, but allows us to capture a wider range of real-world phenomena.

The remainder of this paper is structured as follows. In the next section, we introduce the market model and formulate an optimization problem. In Section 3, we show that this model is free of price manipulation, which allows us to simplify the model setup and the optimization problem in Section 4. Before we state our main results on existence, uniqueness, and characterization of the optimal trading strategy in Sections 6–7, we first provide some elementary properties, like the dimension reduction of our control problem, in Section 5. Section 6 discusses the case where trading is constrained to discrete time and Section 7 contains the continuous time case. In Section 8, we investigate under which conditions price manipulation occurs in a zero spread model. In some special cases, we can calculate optimal strategies in closed form for our main model as well as for the zero spread model of Section 8; we provide some examples in Section 9. Section 10 concludes.

2. MODEL DESCRIPTION

In order to attack the problem of optimal trade execution under time-varying liquidity, we first need to specify a price impact model in Section 2.1. Our model is based on the work of Obizhaeva and Wang (2006), but allows for time-varying order book depth and resilience. Furthermore, we explicitly model both sides of the limit order book and thus can allow for strategies that buy and sell at different points in time. After having introduced the limit order book model, we specify the trader's objectives in Section 2.2.

2.1. Limit Order Book and Price Impact

Trading at most public exchanges is executed through a limit order book, which is a collection of the limit orders of all market participants in an electronic market. Each limit order has the number of shares that the market participant wants to trade, and a price per share attached to it. The price represents a minimal price in case of a sell and a maximal price in case of a buy order. Compared to a limit order, a market order does not have an attached price per share, but instead is executed immediately against the best limit orders waiting in the book. Thus, there is a trade-off between price saving and immediacy when using limit and market orders. We refer the reader to Cont, Stoikov, and Talreja (2010) for a more comprehensive introduction to limit order books.

In this paper, we consider a one-asset model that derives its price dynamics from a limit order book that is exposed to repeated market orders of a large investor (sometimes referred to as the trader). The goal of the investor is to use market orders[2] in order to purchase a large amount x of shares within a certain time period $[0, T]$, where T typically ranges from a few hours up to a few trading days. Without loss of generality, we assume that the investor needs to purchase the asset (the sell case is symmetrical), and hence first describe how buy market orders interact with the ask side of the order book (i.e., with the sell limit orders contained in the limit order book). Subsequently, we turn to the impact

[2] On this macroscopic time scale, the restriction to market orders is not severe. A subsequent consideration of small time windows including limit order trading is common practice in banks. See Naujokat and Westray (2011) for a discussion of a large investor execution problem where both market and limit orders are allowed.

of buy market orders on the bid side and of sell market orders on both sides of the limit order book.

Suppose first that the trader is not active. We assume that the corresponding unaffected best ask price A^u (i.e., the lowest ask price in the limit order book) is a càdlàg martingale on a given filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{P})$ satisfying the usual conditions. This unaffected price process is capturing all external price changes including those due to news as well as due to trading of noise traders and informed traders. Our model includes in particular the case of the Bachelier model $A^u_t = A^u_0 + \sigma W^A_t$ with a $(\mathcal{F}_t)$-Brownian motion W^A, as considered in Obizhaeva and Wang (2006). It also includes the driftless geometric Brownian motion $A^u_t = A^u_0 \exp(\sigma W^A_t - \frac{1}{2}\sigma^2 t)$, which avoids the counterintuitive negative prices of the Bachelier model. Moreover, we can allow for jumps in the dynamics of A^u.

We now describe the shape of the limit order book, i.e., the pattern of ask prices in the order book. We follow Obizhaeva and Wang (2006) and assume a block-shaped order book: The number of shares offered at prices in the interval $[A^u_t, A^u_t + \Delta A]$ is given by $q_t \cdot \Delta A$ with $q_t > 0$ being the order book height (see Figure 2.1 for a graphical illustration). Alfonsi et al. (2010) and Predoiu et al. (2011) consider order books that are not block-shaped and conclude that the optimal execution strategy of the investor is robust with respect to the order book shape. In our model, we allow the order book depth q_t to be time-dependent. As mentioned above, various empirical studies have demonstrated the time-varying features of liquidity, including order book depth. In theoretical models, however, liquidity is still usually assumed to be constant in time. To our knowledge, first attempts to nonconstant liquidity in portfolio liquidation problems have only been considered so far in extensions of the Almgren and Chriss (2001) model such as Kim and Boyd (2008) and Almgren (2009). In this modeling framework, price impact is purely temporary and several of the aspects of this paper do not surface.

Let us now turn to the interaction of the investor's trading with the order book. At time t, the best ask A_t might differ from the unaffected best ask A^u_t due to previous trades of the investor. Define $D_t := A_t - A^u_t$ as the *price impact* or *extra spread* caused by the past actions of the trader. Suppose that the trader places a buy market order of $\xi_t > 0$ shares. This market order consumes all the shares offered at prices between the ask price A_t just prior to order execution and A_{t+} immediately after order execution. A_{t+} is given by $(A_{t+} - A_t) \cdot q_t = \xi_t$ and we obtain

$$D_{t+} = D_t + \xi_t / q_t.$$

See Figure 2.1 for a graphical illustration.

It is a well-established empirical fact that the price impact D exhibits resilience over time. We assume that the immediate impact ξ_t/q_t can be split into a temporary impact component $K_t\xi_t$ that decays to zero and a permanent impact component $\gamma\xi_t$ with

$$\gamma + K_t = q_t^{-1}.$$

We assume that the temporary impact decays exponentially with a fixed time-dependent, deterministic recovery rate $\rho_t > 0$. The price impact at time $s \geq t$ of a buy market order $\xi_t > 0$ placed at time t is assumed to be

$$\gamma\xi_t + K_t e^{-\int_t^s \rho_u \, du} \xi_t.$$

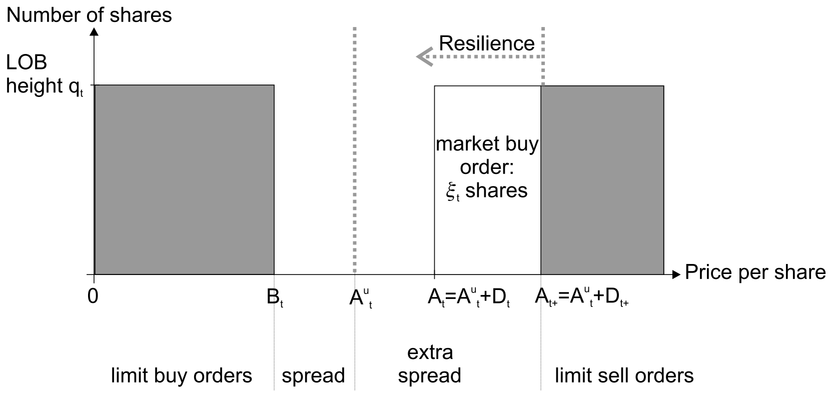

FIGURE 2.1. Snapshot of the block-shaped order book model at time t.

Notice that this temporary impact model is different from the one which is used, e.g., in Almgren and Chriss (2001) and Almgren (2003). It slowly decays to zero instead of vanishing immediately and thus prices depend on previous trades. Obizhaeva and Wang (2006) limit their analysis to a constant decay rate $\rho_t \equiv \rho$, but suggest the extension to time-dependent ρ_t. Weiss (2010) considers exponential resilience and shows that the results of Alfonsi et al. (2010) and in particular Obizhaeva and Wang (2006) can be adapted when the recovery rate depends on the extra spread D caused by the large investor. Gatheral (2010) considers more general deterministic decay functions than the exponential one in a model with a potentially nonlinear price impact and discusses which combinations of decay function and price impact yield "no arbitrage," i.e., nonnegative expected costs of a round trip. Alfonsi et al. (2011) study the optimal execution problem for more general deterministic decay functions than the exponential one in a model with constant order book height. For the calibration of resilience, see Large (2007) and for a discussion of a stochastic recovery rate ρ, we refer to Fruth (2011).

Let us now discuss the impact of market buy orders on the bid side of the limit order book. According to the mechanics of the limit order book, a single market buy order ξ_t directly influences the best ask A_{t+}, but does not influence the best bid price $B_{t+} = B_t$ immediately. The best ask A_{t+} recovers over time (in the absence of any other trading from the investor) on average to $A_t + \gamma\xi_t$. In reality, market orders only lead to a temporary widening of the spread. In order to close the spread, B_t needs to move up by $\gamma\xi_t$ over time and converge to $B_t + \gamma\xi_t$, i.e., the buy market order ξ_t influences the future evolution of B. We assume that B converges to this new level exponentially with the same rate ρ_t. The price impact on the best bid B_s at time $s \geq t$ of a buy market order $\xi_t > 0$ placed at time t is hence

$$\gamma\left(1 - e^{-\int_t^s \rho_u \, du}\right)\xi_t.$$

We assume that the impact of sell market orders is symmetrical to that of buy market orders. It should be noted that our model deviates from the existing literature by explicitly modeling both sides of the order book with a trading-dependent spread. For example, Obizhaeva and Wang (2006) only model one side of the order book and restrict trading to this side of the book.

Alfonsi et al. (2011), Gatheral et al. (2012), and Gatheral, Schied, and Slynko (2011), on the other hand, allow for trading on both sides of the order book, but assume that there is no spread, i.e., they assume $A^u = B^u$ for unaffected best ask and best bid prices, and that the best bid moves up instantaneously when a market buy order is matched with the ask side of the book. They find that under this assumption, the model parameters (for example, the decay kernel) need to fulfill certain conditions, otherwise price manipulation arises. We will revisit this topic in Sections 3 and 8.

We can now summarize the dynamics of the best ask A_t and best bid B_t for general trading strategies in continuous time. Let Θ and $\tilde{\Theta}$ be increasing processes that describe the number of shares which the investor bought, respectively, sold from time 0 until time t. We then have

$$\begin{aligned} A_t &= A_t^u + D_t, \\ B_t &= B_t^u - E_t, \end{aligned}$$

where

$$\begin{aligned} D_t = D_0 e^{-\int_0^t \rho_s ds} &+ \int_{[0,t)} \left(\gamma + K_s e^{-\int_s^t \rho_u du}\right) d\Theta_s \\ &- \int_{[0,t)} \gamma\left(1 - e^{-\int_s^t \rho_u du}\right) d\tilde{\Theta}_s, \quad t \in [0, T+], \end{aligned} \tag{2.1}$$

$$\begin{aligned} E_t = E_0 e^{-\int_0^t \rho_s ds} &+ \int_{[0,t)} \left(\gamma + K_s e^{-\int_s^t \rho_u du}\right) d\tilde{\Theta}_s \\ &- \int_{[0,t)} \gamma\left(1 - e^{-\int_s^t \rho_u du}\right) d\Theta_s, \quad t \in [0, T+], \end{aligned} \tag{2.2}$$

with some given nonnegative initial price impacts $D_0 \geq 0$ and $E_0 \geq 0$.

ASSUMPTION 2.1 (Basic assumptions on Θ, $\tilde{\Theta}$, A^u, B^u, K, and ρ). *Throughout this paper, we assume the following:*

- *The set of* admissible strategies *is given as*

$$\begin{aligned} \tilde{\mathcal{A}}_0 := \{(\Theta, \tilde{\Theta}) : \Omega \times [0, T+] \to [0, \infty)^2 \mid &\Theta \textit{ and } \tilde{\Theta} \textit{ are } (\mathcal{F}_t)-\textit{adapted} \\ &\textit{nondecreasing bounded càglàd processes with } (\Theta_0, \tilde{\Theta}_0) = (0, 0)\}. \end{aligned}$$

  Note that $(\Theta, \tilde{\Theta})$ *may have jumps. In particular, trading in rates and impulse trades are allowed.*
- *The unaffected best ask price process A^u is a càdlàg $\mathcal{H}^1$-martingale with a deterministic starting point A_0^u, i.e.,*

$$\mathbb{E}\sqrt{[A^u, A^u]_T} < \infty, \quad \textit{or, equivalently,} \quad \mathbb{E} \sup_{t \in [0, T]} \left|A_t^u\right| < \infty.$$

  The same condition holds for the unaffected best bid price B^u . Furthermore, $B_t^u \leq A_t^u$ for all $t \in [0, T]$.
- *The price impact coefficient K: $[0, T] \to (0, \infty)$ is a deterministic strictly positive bounded Borel function.*
- *The resilience speed $\rho : [0, T] \to (0, \infty)$ is a deterministic strictly positive Lebesgue integrable function.*

REMARK 2.2.

(i) The purchasing component Θ of a strategy from $\tilde{\mathcal{A}}_0$ consists of a left continuous nondecreasing process $(\Theta_t)_{t\in[0,T]}$ and an additional random variable Θ_{T+} with $\Delta\Theta_T := \Theta_{T+} - \Theta_T \geq 0$ being the last purchase of the strategy. Similarly, for $t \in [0, T]$, we use the notation $\Delta\Theta_t := \Theta_{t+} - \Theta_t$. The same conventions apply for the selling component $\tilde{\Theta}$.

(ii) The processes D and E depend on $(\Theta, \tilde{\Theta})$, although this is not explicitly marked in their notation.

(iii) As it is often done in the literature on optimal portfolio execution, Θ, $\tilde{\Theta}$, D, and E are assumed to be càglàd processes. In (2.1), the possibility $t = T+$ is by convention understood as

$$D_{T+} = D_0 e^{-\int_0^T \rho_s ds} + \int_{[0,T]} \big(\gamma + K_s e^{-\int_s^T \rho_u du}\big) d\Theta_s - \int_{[0,T]} \gamma e^{-\int_s^T \rho_u du} d\tilde{\Theta}_s.$$

A similar convention applies to all other formulas of such type. Furthermore, the integrals of the form

$$\int_{[0,t)} K_s d\Theta_s \quad \text{or} \quad \int_{[0,t]} K_s d\Theta_s$$

are understood as pathwise Lebesgue–Stieltjes integrals, i.e., Lebesgue integrals with respect to the measure with the distribution function $s \mapsto \Theta_{s+}$.

(iv) In the sequel, we need to apply stochastic analysis (e.g., integration by parts or Ito's formula) to càglàd processes of finite variation and/or standard semimartingales. This will always be done as follows: if U is a càglàd process of finite variation, we first consider the process U^+ defined by $U_t^+ := U_{t+}$ and then apply standard formulas from stochastic analysis to it. An example (which will be often used in proofs) is provided in Appendix A.

2.2. Optimization Problem

Let us go ahead by describing the cost-minimization problem of the trader. When placing a single buy market order of size $\xi_t \geq 0$ at time t, he purchases at prices $A_t^u + d$, with d ranging from D_t to D_{t+}, see Figure 2.1. Due to the block-shaped limit order book, the total costs of the buy market order amount to

$$\big(A_t^u + D_t\big)\xi_t + \frac{D_{t+} - D_t}{2}\xi_t = \big(A_t^u + D_t\big)\xi_t + \frac{\xi_t^2}{2q_t} = \xi_t\left(A_t + \frac{\xi_t}{2q_t}\right).$$

Thus, the total costs of the buy market order are the number of shares ξ_t times the average price per share $(A_t + \frac{\xi_t}{2q_t})$. More generally, the total costs of a strategy $(\Theta, \tilde{\Theta}) \in \tilde{\mathcal{A}}_0$ are given by the formula

$$\mathcal{C}(\Theta, \tilde{\Theta}) := \int_{[0,T]} \left(A_t + \frac{\Delta\Theta_t}{2q_t}\right) d\Theta_t - \int_{[0,T]} \left(B_t - \frac{\Delta\tilde{\Theta}_t}{2q_t}\right) d\tilde{\Theta}_t$$

(recall Remark 2.2.i) for the definitions of $\Delta\Theta$ and $\Delta\tilde{\Theta}$).

We now collect all admissible strategies that build up a position of $x \in [0, \infty)$ shares until time T in the set

$$\tilde{\mathcal{A}}_0(x) := \{(\Theta, \tilde{\Theta}) \in \tilde{\mathcal{A}}_0 \mid \Theta_{T+} - \tilde{\Theta}_{T+} = x \text{ a.s.}\}.$$

Our aim is to minimize the expected execution costs

$$\inf_{(\Theta,\tilde{\Theta})\in\tilde{\mathcal{A}}_0(x)} \mathbb{E}\,\mathcal{C}(\Theta, \tilde{\Theta}). \tag{2.3}$$

We hence consider the large investor to be risk-neutral and explicitly allow for his optimal strategy to consist of both buy and sell orders. In the next section, we will see that in our model, it is never optimal to submit sell orders when the overall goal is the purchase of $x > 0$ shares.

Let us finally note that problem (2.3) with $x \in (-\infty, 0]$ is the problem of maximizing the expected proceeds from liquidation of $|x|$ shares and, due to symmetry in modeling ask and bid sides, can be considered similarly to problem (2.3) with $x \in [0, \infty)$.

3. MARKET MANIPULATION

Market manipulation has been a concern for price impact models for some time. We now define the counterparts in our model of the notions of *price manipulation* in the sense of Huberman and Stanzl (2004)[3] and of *transaction-triggered price manipulation* in the sense of Alfonsi et al. (2011) and Gatheral et al. (2012). Note that in defining these notions in our model, we explicitly account for the possibility of D_0 and E_0 being nonzero.

DEFINITION 3.1. *A* round trip *is a strategy from* $\tilde{\mathcal{A}}_0(0)$. *A price manipulation strategy is a round trip* $(\Theta, \tilde{\Theta}) \in \tilde{\mathcal{A}}_0(0)$ *with strictly negative expected execution costs* $\mathbb{E}\,\mathcal{C}(\Theta, \tilde{\Theta}) < 0$. *A market impact model (represented by* A^u, B^u, K, *and* ρ*) admits* price manipulation *if there exist* $D_0 \geq 0$, $E_0 \geq 0$, *and* $(\Theta, \tilde{\Theta}) \in \tilde{\mathcal{A}}_0(0)$ *with* $\mathbb{E}\,\mathcal{C}(\Theta, \tilde{\Theta}) < 0$.

DEFINITION 3.2. *A market impact model (represented by* A^u, B^u, K, *and* ρ*) admits transaction-triggered price manipulation if the expected execution costs of a buy (or sell) program can be decreased by intermediate sell (respectively, buy) trades. More precisely, this means that there exist* $x \in [0, \infty)$, $D_0 \geq 0$, $E_0 \geq 0$, *and* $(\Theta^0, \tilde{\Theta}^0) \in \tilde{\mathcal{A}}_0(x)$ *with*

$$\mathbb{E}\,\mathcal{C}(\Theta^0, \tilde{\Theta}^0) < \inf\{\mathbb{E}\,\mathcal{C}(\Theta, 0) \mid (\Theta, 0) \in \tilde{\mathcal{A}}_0(x)\}, \tag{3.1}$$

or there exist $x \in (-\infty, 0]$, $D_0 \geq 0$, $E_0 \geq 0$, *and* $(\Theta^0, \tilde{\Theta}^0) \in \tilde{\mathcal{A}}_0(x)$ *with*

$$\mathbb{E}\,\mathcal{C}(\Theta^0, \tilde{\Theta}^0) < \inf\{\mathbb{E}\,\mathcal{C}(0, \tilde{\Theta}) \mid (0, \tilde{\Theta}) \in \tilde{\mathcal{A}}_0(x)\}. \tag{3.2}$$

Clearly, if a model admits price manipulation, then it admits transaction-triggered price manipulation. But transaction-triggered price manipulation can be present even if price manipulation does not exist in a model. This situation has been demonstrated in limit order book models with zero bid-ask spread by Schöneborn (2008) (chapter 9) in a multiagent setting and by Alfonsi et al. (2011) in a setting with nonexponential decay of price impact. In this section, we will show that the limit order book model introduced

[3] This definition should not be confused with other definitions of price manipulation such as the one in Kyle and Viswanathan (2008).

in Section 2 is free from both classical and transaction-triggered price manipulation. In Section 8, we will revisit this topic for a different (but related) limit order book model.

Before attacking the main question of price manipulation in Proposition 3.4, we consider the expected execution costs of a pure purchasing strategy and verify in Proposition 3.3 that the costs resulting from changes in the unaffected best ask price are zero and that the costs due to permanent impact are the same for all strategies.

PROPOSITION 3.3. *Let* $(\Theta, 0) \in \tilde{\mathcal{A}}_0(x)$ *(i.e.,* $\tilde{\Theta} \equiv 0$*) with* $x \in [0, \infty)$*. Then*

$$\begin{aligned}(3.3)\qquad \mathbb{E}\mathcal{C}(\Theta, 0) &= \mathbb{E}\left[\int_{[0,T]} \left(A_t + \frac{\Delta\Theta_t}{2q_t}\right) d\Theta_t\right]\\ &= A_0^u x + \frac{\gamma}{2}x^2 + \mathbb{E}\left[\int_{[0,T]} \left(D_t^{\gamma=0} + \frac{K_t}{2}\Delta\Theta_t\right) d\Theta_t\right],\end{aligned}$$

with

$$(3.4)\qquad D_t^{\gamma=0} := D_0 e^{-\int_0^t \rho_s ds} + \int_{[0,t)} K_s e^{-\int_s^t \rho_u du} d\Theta_s, \quad t \in [0, T+].$$

Proof. We start by looking at the expected costs caused by the unaffected best ask price martingale. Using (A.1) with $U := \Theta, Z := A^u$, the facts that Θ is bounded and that A^u is an $\mathcal{H}^1$-martingale yield

$$(3.5)\qquad \mathbb{E}\left[\int_{[0,T]} A_t^u d\Theta_t\right] = \mathbb{E}\left[A_T^u \Theta_{T+} - A_0^u \Theta_0\right] = A_0^u x.$$

Let us now turn to the simplification of our optimization problem due to permanent impact. To this end, we differentiate between the temporary price impact $D_t^{\gamma=0}$ and the total price impact $D_t = D_t^{\gamma=0} + \gamma\Theta_t$ that we get by adding the permanent impact. Notice that $\tilde{\Theta} \equiv 0$. We can then write

$$\begin{aligned}&\mathbb{E}\left[\int_{[0,T]} \left(A_t^u + D_t + \frac{\Delta\Theta_t}{2q_t}\right) d\Theta_t\right]\\ &\quad= A_0^u x + \mathbb{E}\left[\int_{[0,T]} \left(D_t^{\gamma=0} + \gamma\Theta_t + \frac{\gamma + K_t}{2}\Delta\Theta_t\right) d\Theta_t\right]\\ &\quad= A_0^u x + \mathbb{E}\left[\int_{[0,T]} \left(D_t^{\gamma=0} + \frac{K_t}{2}\Delta\Theta_t\right) d\Theta_t\right] + \gamma\mathbb{E}\left[\int_{[0,T]} \left(\Theta_t + \frac{\Delta\Theta_t}{2}\right) d\Theta_t\right].\end{aligned}$$

The assertion follows, since integration by parts for càglàd processes (see (A.2) with $U = V := \Theta$) and $\Theta_0 = 0$, $\Theta_{T+} = x$ yield

$$(3.6)\qquad \int_{[0,T]} \left(\Theta_t + \frac{\Delta\Theta_t}{2}\right) d\Theta_t = \frac{\Theta_{T+}^2 - \Theta_0^2}{2} = \frac{x^2}{2}.$$

□

We can now proceed to prove that our model is free of price manipulation and transaction-triggered price manipulation.

PROPOSITION 3.4 (Absence of transaction-triggered price manipulation). *In the model of Section 2, there is no transaction-triggered price manipulation. In particular, there is no price manipulation.*

Proof. Consider $x \in [0, \infty)$ and $(\Theta, \tilde{\Theta}) \in \tilde{\mathcal{A}}_0(x)$. Making use of

$$B_t = B_t^u - E_t \le A_t^u - E_t \le A_t^u + \gamma(\Theta_t - \tilde{\Theta}_t)$$

yields

$$\begin{aligned}
&\mathbb{E}\left[\int_{[0,T]} \left(A_t + \frac{\Delta\Theta_t}{2q_t}\right) d\Theta_t\right] - \mathbb{E}\left[\int_{[0,T]} \left(B_t - \frac{\Delta\tilde{\Theta}_t}{2q_t}\right) d\tilde{\Theta}_t\right] \\
&\ge \mathbb{E}\left[\int_{[0,T]} \left(A_t^u + \gamma\Theta_t + D_t^{\gamma=0} - \gamma\tilde{\Theta}_t + \frac{\gamma}{2}\Delta\Theta_t + \frac{K_t}{2}\Delta\Theta_t\right) d\Theta_t\right] \\
&\quad -\mathbb{E}\left[\int_{[0,T]} \left(A_t^u + \gamma\Theta_t - \gamma\tilde{\Theta}_t - \frac{\gamma}{2}\Delta\tilde{\Theta}_t - \frac{K_t}{2}\Delta\tilde{\Theta}_t\right) d\tilde{\Theta}_t\right] \\
&\ge \mathbb{E}\left[\int_{[0,T]} A_t^u d(\Theta_t - \tilde{\Theta}_t)\right] \\
&\quad + \gamma\mathbb{E}\left[\int_{[0,T]} \left(\Theta_t - \tilde{\Theta}_t + \frac{\Delta\Theta_t}{2}\right) d\Theta_t + \int_{[0,T]} \left(\tilde{\Theta}_t - \Theta_t + \frac{\Delta\tilde{\Theta}_t}{2}\right) d\tilde{\Theta}_t\right] \\
&\quad + \mathbb{E}\left[\int_{[0,T]} \left(D_t^{\gamma=0} + \frac{K_t}{2}\Delta\Theta_t\right) d\Theta_t\right].
\end{aligned}$$

Analogously to (3.5), the first of these terms equals $A_0^u x$ since $\Theta, \tilde{\Theta}$ are bounded and A^u is an $\mathcal{H}^1$-martingale. For the second one, we do integration by parts (use (A.2) three times) to deduce

$$\begin{aligned}
&\int_{[0,T]} (2\Theta_t + \Delta\Theta_t) d\Theta_t + \int_{[0,T]} (2\tilde{\Theta}_t + \Delta\tilde{\Theta}_t) d\tilde{\Theta}_t - 2\int_{[0,T]} \tilde{\Theta}_t d\Theta_t - 2\int_{[0,T]} \Theta_t d\tilde{\Theta}_t \\
&\quad = \Theta_{T+}^2 + \tilde{\Theta}_{T+}^2 - 2\Theta_{T+}\tilde{\Theta}_{T+} - 2\int_{[0,T]} \tilde{\Theta}_t d\Theta_t + 2\int_{[0,T]} \tilde{\Theta}_{t+} d\Theta_t \ge (\Theta_{T+} - \tilde{\Theta}_{T+})^2 = x^2.
\end{aligned}$$

Summarizing, we get

$$\begin{aligned}
\mathbb{E}\mathcal{C}(\Theta, \tilde{\Theta}) &= \mathbb{E}\left[\int_{[0,T]} \left(A_t + \frac{\Delta\Theta_t}{2q_t}\right) d\Theta_t - \int_{[0,T]} \left(B_t - \frac{\Delta\tilde{\Theta}_t}{2q_t}\right) d\tilde{\Theta}_t\right] \\
&\ge A_0^u x + \frac{\gamma}{2}x^2 + \mathbb{E}\left[\int_{[0,T]} \left(D_t^{\gamma=0} + \frac{K_t}{2}\Delta\Theta_t\right) d\Theta_t\right] \\
&\ge A_0^u x + \frac{\gamma}{2}x^2 + \mathbb{E}\left[\int_{[0,T]} \left(D_t^{\gamma=0} + \frac{K_t}{2}\Delta\check{\Theta}_t\right) d\check{\Theta}_t\right] \\
&= \mathbb{E}\mathcal{C}(\check{\Theta}, 0),
\end{aligned}$$

where we considered the pure purchasing strategy $(\check{\Theta}, 0) \in \tilde{\mathcal{A}}_0(x)$ defined by

$$\check{\Theta}_t := \left\{\begin{matrix} \Theta_t & \text{if } \ \Theta_t \le x \\ x & \text{otherwise} \end{matrix}\right\}$$

(note that the last equality is due to Proposition 3.3). Thus, (3.1) does not occur. By a similar reasoning, (3.2) does not occur as well. □

The central economic insight captured in the previous proposition is that price manipulation strategies can be severely penalized by a widening spread. This idea can easily be applied to different variations of our model, for example, to nonexponential decay kernels as in Gatheral et al. (2012).

4. REDUCTION OF THE OPTIMIZATION PROBLEM

Due to Propositions 3.3 and 3.4, we can significantly simplify the optimization problem (2.3). Let us fix $x \in [0, \infty)$. Then, it is enough to minimize the expectation in the right-hand side of (3.3) over the pure purchasing strategies that build up the position of x shares until time T. That is to say, the problem in general reduces to that with $A^u \equiv 0$, $\gamma = 0$, $\tilde{\Theta} \equiv 0$. Moreover, due to (3.3) and (3.4) and the fact that K and ρ are deterministic functions, it is enough to minimize over deterministic purchasing strategies. We are going to formulate the simplified optimization problem, where we now consider a general initial time $t \in [0, T]$ because we will use dynamic programming afterward.

Let us define the following simplified control sets only containing deterministic purchasing strategies:

$$\begin{aligned}\mathcal{A}_t &:= \{\Theta : [t, T+] \to [0, \infty) \mid \Theta \text{ is a deterministic} \\ &\qquad \text{nondecreasing càglàd function with } \Theta_t = 0\},\\ \mathcal{A}_t(x) &:= \{\Theta \in \mathcal{A}_t \mid \Theta_{T+} = x\}.\end{aligned}$$

As above, a strategy from $\mathcal{A}_t$ consists of a left continuous nondecreasing function $(\Theta_s)_{s \in [t,T]}$ and an additional value $\Theta_{T+} \in [0, \infty)$ with $\Delta\Theta_T := \Theta_{T+} - \Theta_T \geq 0$ being the last purchase of the strategy. For any fixed $t \in [0, T]$ and $\delta \in [0, \infty)$, we define the *cost function* $J(t, \delta, \cdot) : \mathcal{A}_t \to [0, \infty)$ as

$$J(\Theta) := J(t, \delta, \Theta) := \int_{[t,T]} \left(D_s + \frac{K_s}{2} \Delta\Theta_s \right) d\Theta_s, \tag{4.1}$$

where

$$D_s := \delta e^{-\int_t^s \rho_u du} + \int_{[t,s)} K_u e^{-\int_u^s \rho_r dr} d\Theta_u, \quad s \in [t, T+]. \tag{4.2}$$

The cost function J represents the total temporary impact costs of the strategy Θ on the time interval $[t, T]$ when the initial price impact $D_t = \delta$. Observe that J is well-defined and finite due to Assumption 2.1.

Let us now define the *value function for continuous trading time* U: $[0, T] \times [0, \infty)^2 \to [0, \infty)$ as

$$U(t, \delta, x) := \inf_{\Theta \in \mathcal{A}_t(x)} J(t, \delta, \Theta). \tag{4.3}$$

We also want to discuss *discrete trading time*, i.e., when trading is only allowed at given times

$$0 = t_0 < t_1 < \cdots < t_N = T.$$

Define $\tilde{n}(t) := \inf\{n = 0, \dots, N | t_n \geq t\}$. We then have to constrain our strategy sets to

$$\begin{aligned}\mathcal{A}_t^N &:= \{\Theta \in \mathcal{A}_t \mid \Theta_s = 0 \quad \text{on} \quad [t, t_{\tilde{n}(t)}],\\ &\qquad \Theta_s = \Theta_{t_n+} \quad \text{on} \quad (t_n, t_{n+1}] \quad \text{for} \quad n = \tilde{n}(t), \dots, N-1\} \subset \mathcal{A}_t,\\ \mathcal{A}_t^N(x) &:= \left\{\Theta \in \mathcal{A}_t^N \middle| \Theta_{T+} = x\right\} \subset \mathcal{A}_t(x),\end{aligned}$$

and the *value function for discrete trading time* becomes

$$U^N(t, \delta, x) := \inf_{\Theta \in \mathcal{A}_t^N(x)} J(t, \delta, \Theta) \geq U(t, \delta, x). \tag{4.4}$$

Note that the optimization problems in continuous time (4.3) and in discrete time (4.4) only refer to the ask side of the limit order book. The results for optimal trading strategies that we derive in the following sections are hence applicable not only to the specific limit order book model introduced in Section 2, but also to any model that excludes transaction-triggered price manipulation and where the ask price evolution for pure buying strategies is identical to the ask price evolution in our model. This includes, for example, models with different depth of the bid and ask sides of the limit order book, or different resiliences of the two sides of the book.

We close this section with the following simple result, which shows that our problem is economically sensible.

LEMMA 4.1 (Splitting argument). *Doing two separate trades $\xi_\alpha, \xi_\beta > 0$ at the same times has the same effect as trading at once $\xi := \xi_\alpha + \xi_\beta$, i.e., both alternatives incur the same impact costs and the same impact D_{s+}.*

Proof. The impact costs are in both cases

$$\begin{aligned}\left(D_s + \frac{K_s}{2}\xi\right)\xi &= D_s(\xi_\alpha + \xi_\beta) + \frac{K_s}{2}\left(\xi_\alpha^2 + 2\xi_\alpha\xi_\beta + \xi_\beta^2\right)\\ &= \left(D_s + \frac{K_s}{2}\xi_\alpha\right)\xi_\alpha + \left(D_s + K_s\xi_\alpha + \frac{K_s}{2}\xi_\beta\right)\xi_\beta,\end{aligned}$$

and the impact $D_{s+} = D_s + K_s(\xi_\alpha + \xi_\beta)$ after the trade is the same in both cases as well. □

5. PREPARATIONS

In this section, we first show that in our model, optimal strategies are linear in (δ, x), which allows us to reduce the dimensionality of our problem from three dimensions to two dimensions. Thereafter, we introduce the concept of WR-BR structure in Section 5.2, which appropriately describes the value function and optimal execution strategies in our model as we will see in Sections 6 and 7. Finally, we establish some elementary properties of the value function and optimal strategies in Section 5.3.

In this entire section, we usually refer only to the continuous time setting, for example, to the value function U. We refer to the discrete-time setting only when there is something there to be added explicitly. But all of the statements in this section hold both in continuous time (i.e., for U) and in discrete time (i.e., for U^N), and we will later use them in both situations.

5.1. Dimension Reduction of the Value Function

In this section, we prove a scaling property of the value function that helps us to reduce the dimension of our optimization problem. Our approach exploits both the block shape of the limit order book and the exponential decay of price impact, and hence does not generalize easily to more general dynamics of D as, e.g., in Predoiu et al. (2011). We formulate the result for continuous time, although it also holds for discrete time.

LEMMA 5.1 (Optimal strategies scale linearly). *For all $a \in [0, \infty)$, we have*

$$U(t, a\delta, ax) = a^2 U(t, \delta, x). \tag{5.1}$$

Furthermore, if $\Theta^* \in \mathcal{A}_t(x)$ is optimal for $U(t, \delta, x)$, then $a\Theta^* \in \mathcal{A}_t(ax)$ is optimal for $U(t, a\delta, ax)$.

Proof. The assertion is clear for $a = 0$. For any $a \in (0, \infty)$ and $\Theta \in \mathcal{A}_t$, we get from (4.1) and (4.2) that

$$J(t, a\delta, a\Theta) = a^2 J(t, \delta, \Theta). \tag{5.2}$$

Let $\Theta^* \in \mathcal{A}_t(x)$ be optimal for $U(t, \delta, x)$ and $\bar{\Theta} \in \mathcal{A}_t(ax)$ be optimal for $U(t, a\delta, ax)$. If no such optimal strategies exist, the same arguments can be performed with minimizing sequences of strategies. Using (5.2) two times and the optimality of Θ^*, $\bar{\Theta}$, we get

$$J(t, a\delta, \bar{\Theta}) \le J(t, a\delta, a\Theta^*) = a^2 J(t, \delta, \Theta^*) \le a^2 J\left(t, \delta, \frac{1}{a}\bar{\Theta}\right) = J(t, a\delta, \bar{\Theta}).$$

Hence, all inequalities are equalities. Therefore, $a\Theta^*$ is optimal for $U(t, a\delta, ax)$ and (5.1) holds. □

For $\delta > 0$, we can take $a = \frac{1}{\delta}$ and apply Lemma 5.1 to get

$$\begin{aligned} U(t, \delta, x) &= \delta^2 U\left(t, 1, \frac{x}{\delta}\right) = \delta^2 V(t, y) \quad \text{with} \\ y &:= \frac{x}{\delta}, \\ V(t, y) &:= U(t, 1, y), \quad V(T, y) = y + \frac{K_T}{2} y^2, \quad V(t, 0) \equiv 0. \end{aligned} \tag{5.3}$$

In this way, we are able to reduce our three-dimensional value function U defined in (4.3) to a two-dimensional function V. That is $U(t, \delta_{fix}, x)$ for some $\delta_{fix} > 0$ or $U(t, \delta, x_{fix})$ for some $x_{fix} > 0$ already determines the entire value function.[4] Instead of keeping track of the values x and δ separately, only the ratio of them is important. It should be noted, however, that the function V itself is not necessarily the value function of a modified optimization problem. In a similar way, we define the function V^N through the function U^N.

[4] In the following, we will often analyze the function V in order to derive properties of U. Technically, this does not directly allow us to draw conclusions for $U(t, 0, x)$, where $\delta = 0$, since in this case, $y = x/\delta$ is not defined. The extension of our proofs to the possibility $\delta = 0$, however, is straightforward using continuity arguments (see Proposition 5.5 below) or alternatively by analyzing $\tilde{V}(t, \tilde{y}) := U(t, \tilde{y}, 1)$.

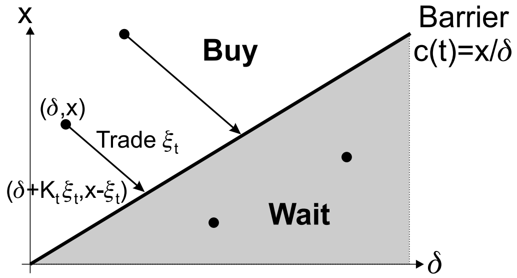

FIGURE 5.1. The δ-x-plane for fixed time t.

5.2. Introduction to Buy and Wait Regions

Let us consider an investor who at time t needs to purchase a position of $x > 0$ in the remaining time until T and is facing a limit order book dislocated by $D_t = \delta \geq 0$. Any trade ξ_t at time t is decreasing the number of shares that are still to be bought, but is increasing D at the same time (see Figure 5.1 for a graphical representation). In the δ-x plane, the investor can hence move downward and to the right. Note that due to the absence of transaction-triggered price manipulation (as shown in Proposition 3.4), any intermediate sell orders are suboptimal and hence will not be considered.

Intuitively, one might expect the large investor to behave as follows: If there are many shares x left to be bought and the price deviation δ is small, then the large investor would buy some shares immediately. In the opposite situation, i.e., small x and large δ, he would defer trading and wait for a decrease of the price deviation due to resilience. We might hence conjecture that the δ-x-plane is divided by a time-dependent barrier into one *buy* region above and one *wait* region below the barrier. Based on the linear scaling of optimal strategies (Lemma 5.1), we know that if (δ, x) is in the buy region at time t, then, for any $a > 0$, $(a\delta, ax)$ is also in the buy region. The barrier between the buy and wait regions therefore has to be a straight line through the origin and the buy and sell region can be characterized in terms of the ratio $y = \frac{x}{\delta}$. In this section, we formally introduce the buy and wait regions and the barrier function. In Sections 6 and 7, we prove that such a barrier exists for discrete and continuous trading time, respectively. In contrast to the case of a time-varying but deterministic illiquidity K considered in this paper, for *stochastic* K, this barrier conjecture holds true in many, but not all cases, see Fruth (2011).

We first define the buy and wait regions and subsequently define the barrier function. Based on the above scaling argument, we can limit our attention to points $(1, y)$ where $\delta = 1$, since for a point (δ, x) with $\delta > 0$, we can instead consider the point $(1, x/\delta)$.

DEFINITION 5.2 (Buy and wait region). For any $t \in [0, T]$, we define the *inner buy region* as

$$Br_t := \left\{ y \in (0, \infty) \,|\, \exists \xi \in (0, y) : U(t, 1, y) = U(t, 1 + K_t\xi, y - \xi) + \left(1 + \frac{K_t}{2}\xi\right)\xi \right\},$$

and call the following sets the *buy region* and *wait region* at time t:

$$BR_t := \overline{Br_t}, \qquad WR_t := [0, \infty) \setminus Br_t$$

(the bar means closure in $\mathbb{R}$).

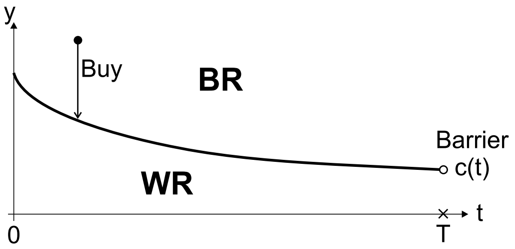

FIGURE 5.2. Schematic illustration of the buy and wait regions in continuous time.

The inner buy region at time t hence consists of all values y such that immediate buying at the state $(1, y)$ is value preserving. The wait region, on the other hand, contains all values y such that any nonzero purchase at $(1, y)$ destroys value. Let us note that $Br_T = (0, \infty)$, $BR_T = [0, \infty)$, and $WR_T = \{0\}$.

Regarding Definition 5.2, the following comment is in order. We do not claim in this definition that Br_t is an open set. A priori one might imagine, say, the set $(10, 20]$ as the inner buy region at some time point. But what we can say from the outset is that, due to the splitting argument (see Lemma 4.1), Br_t is in any case a union of (not necessarily open) intervals or the empty set.

The wait region/buy region conjecture can now be formalized as follows.

DEFINITION 5.3 (WR-BR structure). The value function U has *WR-BR structure* if there exists a *barrier function*

$$c : [0, T] \to [0, \infty]$$

such that for all $t \in [0, T]$,

$$Br_t = (c(t), \infty)$$

with the convention $(\infty, \infty) := \emptyset$. For the value function U^N in discrete time to have WR-BR structure, we only consider $t \in \{t_0, \dots, t_N\}$ and set $c^N(t) = \infty$ for $t \notin \{t_0, \dots, t_N\}$.

Let us note that we always have $c(T) = 0$. Below, we will see that it is indeed possible to have $c(t) = \infty$, i.e., at time t, any strictly positive trade is suboptimal no matter at which state we start. For $c(t) < \infty$, having WR-BR structure means that $BR_t \cap WR_t = \{c(t)\}$. Figure 5.2 illustrates the situation in continuous time.

Thus, up to now we have the following intuition. An optimal strategy is suggested by the barrier function whenever the value function has WR-BR structure. If the position of the large investor at time t satisfies $\frac{x}{\delta} > c(t)$, then the portfolio is in the buy region. We then expect that it is optimal to execute the largest discrete trade $\xi \in (0, x)$ such that the new ratio of remaining shares over price deviation $\frac{x-\xi}{\delta+K_t\xi}$ is still in the buy region, i.e., the optimal trade is

$$\xi^* = \frac{x - c(t)\delta}{1 + K_t c(t)},$$

which is equivalent to

$$c(t) = \frac{x - \xi^*}{\delta + K_t \xi^*}.$$

Notice that the ratio term $\frac{x-\xi}{\delta+K_t\xi}$ is strictly decreasing in ξ. Consequently, trades have the effect of reducing the ratio as indicated in Figure 5.2, while the resilience effect increases it. That is, one trades just enough shares to keep the ratio y below the barrier.[5]

In Figure 5.2, we demonstrate an intuitive case where the barrier decreases over time, i.e., buying becomes more aggressive as the investor runs out of time. This intuitive feature, however, does not need to hold for all possible evolutions of K and ρ as we will see, e.g., in Figure 6.1.

Below, we will see that the intuition presented above always works in discrete time: namely, the value function U^N always has WR-BR structure, there exists a unique optimal strategy, which is of the type "trade to the barrier when the ratio is in the buy region, do not trade when it is in the wait region" (see Section 6). In continuous time, the situation is more delicate. It may happen, for example, that the value function U has WR-BR structure, but the strategy consisting in trading toward the barrier is not optimal (see the example in the beginning of Section 7, where an optimal strategy does not exist). However, if the illiquidity K is continuous, there exists an optimal strategy, and, under additional technical assumptions, it is unique (see Section 7). Moreover, if K and ρ are smooth and satisfy some further technical conditions, we have explicit formulas for the barrier and the optimal strategy (see Section 9).

5.3. Some Properties of the Value Function and Buy and Wait Regions

We first state comparative statics satisfied by both the continuous and the discrete time value function. The value function is increasing in t, δ, x, and the price impact coefficient K as well as decreasing with respect to the resilience speed function.

PROPOSITION 5.4 (Comparative statics for the value function).

(a) The value function is nondecreasing in t, δ, x.

(b) Fix $t \in [0, T]$. Assume that $0 < \check{K}_s \leq \hat{K}_s$ for all $s \in [t, T]$. Then, the value function corresponding to $\check{K}$ is less than or equal to the one corresponding to $\hat{K}$.

(c) Fix $t \in [0, T]$. Assume that $0 < \check{\rho}_s \leq \hat{\rho}_s$ for all $s \in [t, T]$. Then, the value function corresponding to $\hat{\rho}$ is less than or equal to the one corresponding to $\check{\rho}$.

Proof. The proof is straightforward.

PROPOSITION 5.5 (Continuity of the value function). *For each $t \in [0, T]$, the functions*

$$U(t,\cdot,\cdot) : [0,\infty)^2 \to [0,\infty) \quad \textit{and} \quad V(t,\cdot) : [0,\infty) \to [0,\infty)$$

are continuous.

Proof. Due to Lemma 5.1, it is enough to prove that the function $U(t,\cdot,\cdot)$ is continuous. Let us fix $t \in [0, T]$, $x \geq 0$, $0 \leq \delta_1 < \delta_2$, $\epsilon > 0$ and take a strategy $\Theta^\epsilon \in \mathcal{A}_t(x)$ such that

$$J(t, \delta_1, \Theta^\epsilon) < U(t, \delta_1, x) + \epsilon.$$

[5] Intuitively, this implies that apart from a possible initial and final impulse trade, optimal buying occurs in infinitesimal amounts, provided that c is continuous in t on $[0, T)$. For diffusive K as in Fruth (2011), this would lead to singular optimal controls.

For $i = 1, 2$, we define

$$D_s^i := \delta_i e^{-\int_t^s \rho_u du} + \int_{[t,s)} K_u e^{-\int_u^s \rho_r dr} d\Theta_u^\epsilon, \quad s \in [t, T+].$$

Using Proposition 5.4, we get

$$\begin{aligned} U(t, \delta_1, x) \le U(t, \delta_2, x) &\le \int_{[t,T]} \left(D_s^2 + \frac{K_s}{2} \Delta\Theta_s^\epsilon \right) d\Theta_s^\epsilon \\ &\le \int_{[t,T]} \left(D_s^1 + \frac{K_s}{2} \Delta\Theta_s^\epsilon \right) d\Theta_s^\epsilon + (\delta_2 - \delta_1)x \\ &= J(t, \delta_1, \Theta^\epsilon) + (\delta_2 - \delta_1)x < U(t, \delta_1, x) + \epsilon + (\delta_2 - \delta_1)x. \end{aligned}$$

Thus, for each fixed $t \in [0, T]$ and $x \ge 0$, the function $U(t, \cdot, x)$ is continuous on $[0, \infty)$. For $t \in [0, T]$, $\delta \ge 0$ and $x > 0$, by Lemma 5.1, we have

$$U(t, \delta, x) = x^2 U(t, \delta/x, 1),$$

hence, the function $U(t, \cdot, \cdot)$ is continuous on $[0, \infty) \times (0, \infty)$. Considering the strategy of buying the whole position x at time t, we get

$$U(t, \delta, x) \le \left(\delta + \frac{K_t}{2} x \right) x \underset{x \searrow 0}{\rightarrow} 0 = U(t, \delta, 0),$$

i.e., the function $U(t, \cdot, \cdot)$ is also continuous on $[0, \infty) \times \{0\}$. This concludes the proof. □

PROPOSITION 5.6 (Trading never completes early). *For all $t \in [0, T)$, $\delta \in [0, \infty)$ and $x \in (0, \infty)$, the value function satisfies*

$$U(t, \delta, x) < \left(\delta + \frac{K_t}{2} x \right) x,$$

i.e., it is never optimal to buy the whole remaining position at any time $t \in [0, T)$.

Proof. For $\epsilon \in [0, x]$, define the strategies $\Theta^\epsilon \in \mathcal{A}_t(x)$ that buy $(x - \epsilon)$ shares at t and ϵ shares at T. The corresponding costs are

$$J(t, \delta, \Theta^\epsilon) = \left(\delta + \frac{K_t}{2}(x - \epsilon) \right)(x - \epsilon) + \left((\delta + K_t[x - \epsilon]) e^{-\int_t^T \rho_s ds} + \frac{K_T}{2} \epsilon \right) \epsilon.$$

Clearly,

$$U(t, \delta, x) \le J(t, \delta, \Theta^0) = \left(\delta + \frac{K_t}{2} x \right) x,$$

but we never have equality since

$$\frac{\partial}{\partial \epsilon} J(t, \delta, \Theta^\epsilon) \bigg|_{\epsilon=0} = -\left(1 - e^{-\int_t^T \rho_s ds}\right)(K_t x + \delta) < 0.$$

□

As discussed above, we always have $Br_T = (0, \infty)$ and $WR_T = \{0\}$. In two following propositions, we discuss Br_t (equivalently, WR_t) for $t \in [0, T)$.

PROPOSITION 5.7 (Wait region near 0). *Assume that the value function U has WR-BR structure with the barrier c. Then, for any $t \in [0, T)$, $c(t) \in (0, \infty]$(equivalently, there exists $\epsilon > 0$ such that $[0, \epsilon) \subset WR_t$).*

Proof. We need to exclude the possibility $c(t) = 0$, i.e., $Br_t = (0, \infty)$. But if $Br_t = (0, \infty)$, we get by Proposition 5.5 that for any $y > 0$,

$$V(t, y) = \left(1 + \frac{K_t}{2} y\right) y,$$

which contradicts Proposition 5.6. □

The following result illustrates that the barrier can be infinite.

PROPOSITION 5.8 (Infinite barrier). *Assume that there exists $0 \leq t_1 < t_2 \leq T$ such that*

$$K_s e^{-\int_s^{t_2} \rho_u du} > K_{t_2} \quad \textit{for all} \quad s \in [t_1, t_2).$$

Then, $Br_s = \emptyset$ for $s \in [t_1, t_2)$.

In particular, if the assumption of Proposition 5.8 holds with $t_1 = 0$ and $t_2 = T$, then the value function has WR-BR structure and the barrier is infinite except at terminal time T.

Proof. For any $s \in [t_1, t_2)$, $\delta \in [0, \infty)$, $x \in (0, \infty)$, and $\Theta \in \mathcal{A}_s(x)$ with $\Theta_{t_2} > 0$, we get the following by applying (4.2), the assumption of the proposition, monotonicity of J in δ, and integration by parts as in (3.6)

$$\begin{aligned} J(s, \delta, \Theta) &= \int_{[s,t_2)} \left(D_u + \frac{K_u}{2} \Delta\Theta_u\right) d\Theta_u + J(t_2, D_{t_2}, (\Theta_u - \Theta_{t_2})_{u \in [t_2, T+]}) \\ &\geq \int_{[s,t_2)} \left(\delta e^{-\int_s^u \rho_r dr} + \int_{[s,u)} K_r e^{-\int_r^u \rho_w dw} d\Theta_r + \frac{K_u}{2} \Delta\Theta_u\right) d\Theta_u \\ &\quad + J\big(t_2, \delta e^{-\int_s^{t_2} \rho_u du} + K_{t_2}\Theta_{t_2}, (\Theta_u - \Theta_{t_2})_{u \in [t_2, T+]}\big) \\ &> \left(\delta e^{-\int_s^{t_2} \rho_u du} + \frac{K_{t_2}}{2} \Theta_{t_2}\right) \Theta_{t_2} + J\big(t_2, \delta e^{-\int_s^{t_2} \rho_u du} + K_{t_2}\Theta_{t_2}, (\Theta_u - \Theta_{t_2})_{u \in [t_2, T+]}\big). \end{aligned}$$

That is, it is strictly suboptimal to trade on $[s, t_2)$. In particular, $Br_s = \emptyset$. □

Proposition 5.8 can be extended in the following way.

PROPOSITION 5.9 (Infinite barrier, extended version). *Let K be continuous and assume that there exists $0 \leq t_1 < t_2 \leq T$ such that*

$$K_{t_1} e^{-\int_{t_1}^{t_2} \rho_u du} > K_{t_2}.$$

Then, $Br_{t_1} = \emptyset$.

Proof. Define $\tilde{t}$ as the minimal value of the set

$$\underset{t \in [t_1, t_2]}{\operatorname{argmin}}\, K_t e^{\int_0^t \rho_u du}$$

with $\tilde{t}$ being well-defined due to the continuity of K. Then, we know that $\tilde{t} > t_1$. By definition of $\tilde{t}$, we have that for all $t \in [t_1, \tilde{t})$

$$K_t e^{\int_0^t \rho_u du} > K_{\tilde{t}} e^{\int_0^{\tilde{t}} \rho_u du},$$

and hence,

$$K_t e^{-\int_t^{\tilde{t}} \rho_u du} > K_{\tilde{t}}.$$

By Proposition 5.8, we can conclude that $Br_t = \emptyset$ for all $t \in [t_1, \tilde{t})$ and hence in particular for $t = t_1$. □

6. DISCRETE TIME

In this section, we show that the optimal execution problem in discrete time has WR-BR structure. Let us first rephrase the problem in the discrete time setting and define $K_n := K_{t_n}$, $D_n := D_{t_n}$, and $\xi_n := \Delta\Theta_{t_n}$ for $n = 0, \ldots, N$. The optimization problem (4.3) can then be expressed as

$$U^N(t_n, \delta, x) = \inf_{\substack{\xi_j \in [0,x] \\ \sum \xi_j = x}} \sum_{j=n}^{N} \left(D_j + \frac{K_j}{2}\xi_j \right) \xi_j, \tag{6.1}$$

with $D_n = \delta$ and $D_{j+1} = (D_j + K_j \xi_j) a_j$, where

$$a_j := \exp\left(- \int_{t_j}^{t_{j+1}} \rho_s ds \right). \tag{6.2}$$

Recall the dimension reduction from Lemma 5.1

$$U^N(t_n, \delta, x) = \delta^2 V^N\left(t_n, \frac{x}{\delta}\right) \quad \text{with} \quad V^N(t_n, y) := U^N(t_n, 1, y).$$

The following theorem establishes the WR-BR structure in discrete time.

THEOREM 6.1 (Discrete time: WR-BR structure). *In discrete time, there exists a unique optimal strategy, and the value function U^N has WR-BR structure with some barrier function c^N. Furthermore, $V^N(t_n, \cdot) : [0, \infty) \to [0, \infty)$ has the following properties for $n = 0, \ldots, N$.*

(i) *It is* **continuously differentiable**.
(ii) *It is* **piecewise quadratic**, *i.e., there exists $M \in \mathbb{N}$, constants $(\alpha_i, \beta_i, \gamma_i)_{i=1,\ldots,M}$ and $0 < y_1 < y_2 < \cdots < y_M = \infty$ such that*

$$V^N(t_n, y) = \alpha_{m(y)} y^2 + \beta_{m(y)} y + \gamma_{m(y)},$$

for the index function $m : [0, \infty) \to \{1, \ldots, M\}$ with $m(y) := \min\{i \mid y \le y_i\}$.
(iii) *The coefficients $(\alpha_i, \beta_i, \gamma_i)_{i=1,\ldots,M}$ from (ii) satisfy the* **inequalities**

$$\begin{aligned} \alpha_i, \beta_i &> 0, \\ 4\alpha_i \gamma_i + \beta_i - \beta_i^2 &\ge 0, \\ y_{i-1}\beta_i + 2\gamma_i &\ge 0. \end{aligned} \tag{6.3}$$

The properties (i)–(iii) of V^N included in the above theorem will be exploited in the backward induction proof of the WR-BR structure. The piecewise quadratic nature of the value function occurs, since the price impact D is affined in the trade size and is multiplied by the trade size in the value function (6.1). Let us, however, note that the value function in continuous time is no longer piecewise quadratic.

We prove Theorem 6.1 by backward induction. As a preparation, we investigate the relationship of the function V^N at times t_n and t_{n+1}. They are linked by the dynamic programming principle:

$$
\begin{aligned}
(6.4)\quad V^N(t_n, y) &= \min_{\xi\in[0,y]} \left\{ \left(1+\frac{K_n}{2}\xi\right)\xi + U^N(t_{n+1}, (1+K_n\xi)a_n, y-\xi) \right\} \\
&= \min_{\xi\in[0,y]} \left\{ \left(1+\frac{K_n}{2}\xi\right)\xi + (1+K_n\xi)^2 a_n^2 V^N\left(t_{n+1}, \frac{y-\xi}{(1+K_n\xi)a_n}\right) \right\}.
\end{aligned}
$$

Instead of focusing on the optimal trade ξ, we can alternatively look for the optimal new ratio $\eta(\xi) := \frac{y-\xi}{1+K_n\xi}$ of remaining shares over price deviation. Note that η is decreasing in the trade size ξ and bounded between zero (if the entire position is traded at once) and the current ratio y (if nothing is traded). A straightforward calculation confirms that (6.4) is equivalent to

$$
(6.5)\qquad V^N(t_n, y) = \frac{1}{2K_n}\Big[(1+K_n y)^2 \min_{\eta\in[0,y]} L^N(t_n, \eta) - 1\Big],
$$

where

$$
(6.6)\qquad L^N(t_n, \eta) := \frac{1 + 2K_n a_n^2 V^N\big(t_{n+1}, \eta a_n^{-1}\big)}{(1+K_n\eta)^2}.
$$

Note that in (6.5), the minimization is taken over η instead of ξ. Furthermore, the function L^N depends on η, but not on y or ξ separately. In the sequel, the function L^N will be essential in several arguments. The following lemma will be used in the proof of Theorem 6.1.

LEMMA 6.2. *Let $a \in (0, 1)$, $\kappa > 0$ and let the function $v : [0, \infty) \to [0, \infty)$ satisfy (i), (ii), and (iii) given in Theorem 6.1. Then, the following statements hold true.*

(a) There exists $c^* \in [0, \infty]$ such that

$$
L(y) := \frac{1 + 2\kappa a^2 v(ya^{-1})}{(1+\kappa y)^2}, \quad y \in [0, \infty),
$$

is strictly decreasing for $y \in [0, c^*)$ and strictly increasing for $y \in (c^*, \infty)$.

(b) The function

$$
\tilde{v}(y) := \left\{ \begin{array}{ll} \frac{1}{2\kappa}[(1+\kappa y)^2 L(c^*) - 1] & \text{if } \; y > c^* \\ a^2 v(ya^{-1}) & \text{otherwise} \end{array} \right\}
$$

again satisfies (i), (ii), and (iii) with possibly different coefficients.

Proof of Theorem 6.1. We proceed by backward induction. Notice that $V^N(t_N, y) = (1 + \frac{K_N}{2} y)y$ fulfills (i), (ii), and (iii) with $M = 1$, $\alpha_1 = \frac{K_N}{2}$, $\beta_1 = 1$, $\gamma_1 = 0$. Let us consider the induction step from t_{n+1} to t_n. We are going to use Lemma 6.2 for $a = a_n$, $\kappa = K_n$, $v = V^N(t_{n+1}, \cdot)$. We then have that $L = L^N(t_n, \cdot)$ and we obtain c^* as the unique minimum of $L^N(t_n, \cdot)$ from Lemma 6.2(a). From (6.5), we see that the unique optimal value for η is given by

$$\eta^* := \underset{\eta\in[0,y]}{\operatorname{argmin}} \frac{1}{2K_n}[(1 + K_n y)^2 L^N(t_n, \eta) - 1] = \min\{y, c^*\},$$

and accordingly that the unique optimal trade is given by

$$\xi^* := \xi(\eta^*) = \max\left\{0, \frac{y - c_n}{1 + K_n c_n}\right\}.$$

Therefore, we have a unique optimal strategy and the value function has WR-BR structure with $c^N(t_n) := c^*$. Plugging ξ^* into (6.4) and applying the definition of V^N yields $V^N(t_n, y) = \tilde{v}(y)$. Lemma 6.2(b) now concludes the induction step. □

Proof of Lemma 6.2. (a) The function L is continuously differentiable with

$$\begin{aligned} L'(y) &= \frac{2\kappa}{(1 + \kappa y)^3} l(y), \\ l(y) &:= y(2\alpha_{m(ya^{-1})} - \kappa\beta_{m(ya^{-1})}a) + (\beta_{m(ya^{-1})}a - 2\kappa\gamma_{m(ya^{-1})}a^2 - 1). \end{aligned} \tag{6.7}$$

First of all, we show that there is no interval where L is constant. Assume that there would be an interval where l is zero, i.e., there exists $i \in \{1, \dots, M\}$ such that $(2\alpha_i - \kappa\beta_i a) = 0$ and $(\beta_i a - 2\kappa\gamma_i a^2 - 1) = 0$. Solving these equations for α, respectively, γ yields

$$4\alpha_i\gamma_i + a^{-1}\beta_i - \beta_i^2 = 0.$$

This is a contradiction to (6.3).

Let us assume $l(\check{y}) > 0$ for some $\check{y} \in [0, \infty)$ with $j := m(\check{y}a^{-1})$. We are done if we can conclude $l(\hat{y}) > 0$ for all $\hat{y} \in [\check{y}, \infty)$. Because of the continuity of l, it is sufficient to show that L keeps increasing on $[\check{y}, y_j]$, i.e., we need to show $l(\hat{y}) > 0$ for all $\hat{y} \in [\check{y}, y_j]$. Due to the form of l, this is guaranteed when $2\alpha_j - \kappa\beta_j a > 0$. Let us suppose that this term would be negative, which is equivalent to $2\alpha_j\beta_j^{-1}a^{-1} \le \kappa$. Together with the inequalities from (6.3), one gets

$$\begin{aligned} al(\check{y}) &= -\kappa\, a(\check{y}a^{-1}\beta_j + 2\gamma_j) + (2\check{y}a^{-1}\alpha_j + \beta_j - a^{-1}) \\ &\le -2\alpha_j\beta_j^{-1}(\check{y}a^{-1}\beta_j + 2\gamma_j) + (2\check{y}a^{-1}\alpha_j + \beta_j - a^{-1}) \\ &= -\frac{1}{\beta_j}\left(4\alpha_j\gamma_j + \beta_j a^{-1} - \beta_j^2\right) < 0. \end{aligned}$$

This is a contradiction to $l(\check{y}) > 0$.

(b) If c^* is finite, the function $\tilde{v}$ is continuously differentiable at c^* since a brief calculation shows that $\tilde{v}'(c^*-) = \tilde{v}'(c^*+)$ is equivalent to $l(c^*) = 0$. We have

$$\tilde{v}(y) = \tilde{\alpha}_{\tilde{m}(y)}y^2 + \tilde{\beta}_{\tilde{m}(y)}y + \tilde{\gamma}_{\tilde{m}(y)},$$

i.e., $\tilde{v}$ is piecewise quadratic with $\tilde{M} = 1 + m(c^* a^{-1})$, $\tilde{y}_{\tilde{M}-1} := c^*$, $\tilde{y}_i := y_i a$ for $i = 1, \dots, \tilde{M} - 2$ and

$$\begin{aligned} &\tilde{\alpha}_{\tilde{M}} = \frac{\kappa}{2} L(c^*) > 0, \quad \tilde{\beta}_{\tilde{M}} = L(c^*) > 0, \quad \tilde{\gamma}_{\tilde{M}} = \frac{L(c^*) - 1}{2\kappa}, \\ &\tilde{\alpha}_i = \alpha_i > 0, \quad \tilde{\beta}_i = a\beta_i > 0, \quad \tilde{\gamma}_i = a^2 \gamma_i \quad \text{for } \ i = 1, \dots, \tilde{M} - 1. \end{aligned} \tag{6.8}$$

We therefore get

$$4\tilde{\alpha}_i \tilde{\gamma}_i + \tilde{\beta}_i - \tilde{\beta}_i^2 = \left\{ \begin{matrix} 0 & \text{if} \quad i = \tilde{M} \\ a^2 \big(4\alpha_i \gamma_i + a^{-1}\beta_i - \beta_i^2\big) & \text{otherwise} \end{matrix} \right\} \geq 0.$$

It remains to show that $\tilde{v}$ also inherits the last inequality in (6.3) from v. For $y \leq c^*$,

$$y\tilde{\beta}_{\tilde{m}(y)} + 2\tilde{\gamma}_{\tilde{m}(y)} = a^2 (ya^{-1} \beta_{m(ya^{-1})} + 2\gamma_{m(ya^{-1})}) \geq 0.$$

Due to $\tilde{v}$ being continuously differentiable in c^*, we get

$$\begin{aligned} \tilde{\alpha}_{\tilde{M}} (c^*)^2 + \tilde{\beta}_{\tilde{M}} c^* + \tilde{\gamma}_{\tilde{M}} &= \tilde{\alpha}_{\tilde{M}-1} (c^*)^2 + \tilde{\beta}_{\tilde{M}-1} c^* + \tilde{\gamma}_{\tilde{M}-1}, \\ 2\tilde{\alpha}_{\tilde{M}} c^* + \tilde{\beta}_{\tilde{M}} &= 2\tilde{\alpha}_{\tilde{M}-1} c^* + \tilde{\beta}_{\tilde{M}-1}. \end{aligned}$$

Taking two times the first equation and subtracting c^* times, the second equation yields

$$c^* \tilde{\beta}_{\tilde{M}} + 2\tilde{\gamma}_{\tilde{M}} = c^* \tilde{\beta}_{\tilde{M}-1} + 2\tilde{\gamma}_{\tilde{M}-1}.$$

Since we already know that the right-hand side is positive, also $y\tilde{\beta}_{\tilde{M}} + 2\tilde{\gamma}_{\tilde{M}} \geq 0$ for all $y > c^*$. □

We need the following lemma as a preparation for the WR-BR proof in continuous time.

LEMMA 6.3. *Let K be continuous. Then, at least one of two following statements is true:*

- *The function* $y \mapsto L^N(0, y)$ *is convex on* $[0, c^N(0))$.
- *The continuous time buy region is simply* $Br_0 = \emptyset$, *i.e.,* $c(0) = \infty$.

We stress that the first statement in this lemma concerns discrete time, while the second one concerns the continuous time optimization problem.

Proof. Recall that the definition of $L^N(0, \cdot)$ from (6.6) contains $V^N(t_1, \cdot)$ which is continuously differentiable and piecewise quadratic with coefficients $(\alpha_i, \beta_i, \gamma_i)$. Analogously to (6.7), it turns out that

$$\begin{aligned} \frac{\partial}{\partial y} L^N(0, y) = \frac{2K_0}{(1 + K_0 y)^3} \Bigg[y \left(2\alpha_{m(ye^{\int_0^{t_1} \rho_s ds})} - K_0 \beta_{m(ye^{\int_0^{t_1} \rho_s ds})} e^{-\int_0^{t_1} \rho_s ds} \right) \\ + \left(\beta_{m(ye^{\int_0^{t_1} \rho_s ds})} e^{-\int_0^{t_1} \rho_s ds} + 2K_0 \gamma_{m(ye^{\int_0^{t_1} \rho_s ds})} e^{-2\int_0^{t_1} \rho_s ds} - 1 \right) \Bigg]. \end{aligned}$$

We distinguish between two cases. First, assume that all i satisfy $(2\alpha_i - K_0 \beta_i e^{-\int_0^{t_1} \rho_s ds}) \geq 0$. Then, $\frac{\partial}{\partial y} L^N(0, \cdot)$ must be increasing on $[0, c^N(0))$ as desired, since $L^N(0, \cdot)$ is decreasing on this interval as we know from Lemma 6.2.

Assume to the contrary that there exists i such that $(2\alpha_i - K_0 \beta_i e^{-\int_0^{t_1} \rho_s ds}) < 0$. Recall how α_i and β_i are actually computed in the backward induction of Theorem 6.1. In each induction step, Lemma 6.2 is used and the coefficients $\tilde{\alpha}_{\tilde{M}}, \tilde{\beta}_{\tilde{M}}$ get updated in (6.8). It gets clear that there exists $n \in \{1, \ldots, N\}$ such that

$$2\alpha_i - K_0\beta_i e^{-\int_0^{t_1} \rho_s ds} = \left(K_{t_n} - K_0 e^{-\int_0^{t_n} \rho_s ds}\right) L^N(t_n, c^N(t_n)).$$

We get the resilience multiplier $e^{-\int_0^{t_n} \rho_s ds}$ due to the adjustment $\tilde{\beta}_i = a\beta_i$ from the second line of (6.8). Due to L^N being positive, it follows that

$$K_{t_n} < K_0\, e^{-\int_0^{t_n} \rho_s ds}.$$

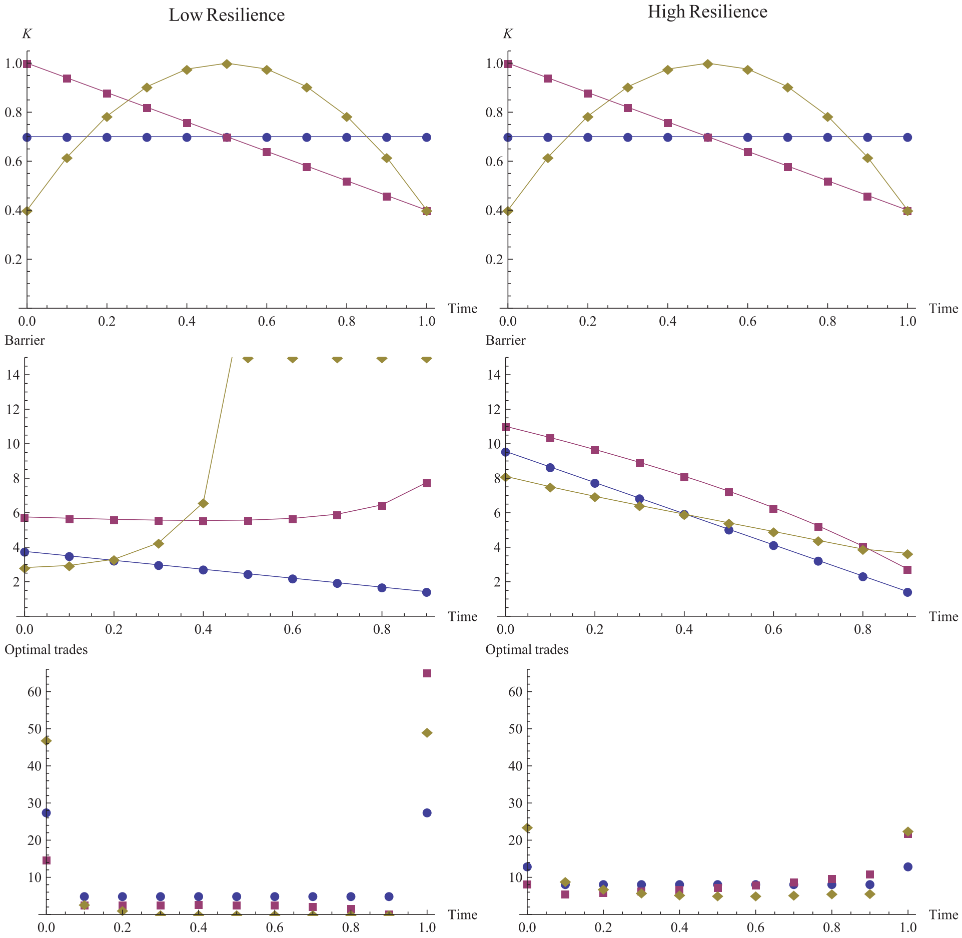

FIGURE 6.1. Illustration of the numerically computed barrier $(c^N(t_n))_{n=0,\ldots,N}$ and the corresponding optimal strategy $(\Delta\Theta_{t_n})_{n=0,\ldots,N}$ in discrete time for $T = 1$, $N = 10$, $x = 100$, $\delta = 0$, $\rho = 2$ (left-hand side), and $\rho = 10$ (right-hand side). We used $K_t^1 \equiv 0.7$, $K_t^2 = 1 - 0.6t$, and $K_t^3 = 1 - 2.4(t - 0.5)^2$ as the given evolution of the illiquidity.

That is, for this choice of K, it cannot be optimal to trade at $t = 0$ as we see from Proposition 5.9. Hence, the buy region at $t = 0$ is the empty set for both discrete and continuous time. □

The proof of Theorem 6.1 is constructive. It not only establishes the existence of a unique barrier, but also provides means to calculate the barrier numerically through the following recursive algorithm.

```
Initialize value function $V^N(t_N, y) = \left(1 + \frac{K_N}{2} y\right) y$
For $n = N-1, \ldots, 0$
    Set $L^N(t_n, y) := \frac{1 + 2K_n a_n^2 V^N\left(t_{n+1}, y a_n^{-1}\right)}{(1 + K_n y)^2}$
    Compute $c^N(t_n) := c_n := \operatorname*{argmin}_{y \geq 0} L^N(t_n, y)$
    Set $V^N(t_n, y) := \left\{\begin{matrix} \frac{1}{2K_n}[(1 + K_n y)^2 L^N(t_n, c_n) - 1] & \text{if} \quad y > c_n \\ a_n^2 V^N(t_{n+1}, y a_n^{-1}) & \text{otherwise} \end{matrix}\right\}$
```

We close this section with a numerical example. Figure 6.1 was generated using the above numerical scheme and illustrates the optimal barrier and trading strategy for several example definitions of K and ρ. For constant K, we recover the Obizhaeva and Wang (2006) "bathtub" strategy with impulse trades of the same size at the beginning and end of the trading horizon and trading with constant speed in between. The corresponding barrier is a decreasing straight line as we will explicitly see for continuous time in Example 9.5. For *high* values of the resilience ρ, the barriers have the typical decreasing shape, i.e., the buy region increases if less time to maturity remains. For *low* values of the resilience ρ, the barrier must not be decreasing and can even be infinite, i.e., the buy region is the empty set, as illustrated for K^3 with less liquidity in the middle than in the beginning and the end of the trading horizon.

7. CONTINUOUS TIME

We now turn to the continuous time setting. In Section 7.1, we discuss existence of optimal strategies using Helly's compactness theorem and a uniqueness result using convexity of the value function. Thereafter, in Section 7.2, we prove that the WR-BR result from Section 6 carries over to continuous time.

7.1. Existence of an Optimal Strategy

In continuous time, existence of an optimal strategy is not guaranteed in general. For instance, consider a constant resilience $\rho_t \equiv \rho > 0$ and the price impact parameter K following the Dirichlet-type function

$$K_t = \left\{\begin{matrix} 1 & \text{for } t \text{ rational} \\ 2 & \text{for } t \text{ irrational} \end{matrix}\right\}. \tag{7.1}$$

In order to analyze model (7.1), let us first recall that in the model with a constant price impact $K_t \equiv \kappa > 0$, there exists a unique optimal strategy, which has a nontrivial absolutely continuous component (see Obizhaeva and Wang 2006 or Example 9.5 below for explicit formulas). Approximating this strategy by strategies trading only at rational time points, we get that the value function in model (7.1) coincides with the value function for the price impact $K_t \equiv 1$. But there is no strategy in model (7.1) attaining this value because the nontrivial absolutely continuous component of the unique optimal strategy for $K_t \equiv 1$ will count with price impact 2 instead of 1 in the total costs. Thus, there is no optimal strategy in model (7.1).[6]

We can therefore hope to prove existence of optimal strategies only under additional conditions on the model parameters. In all of Section 7, we will assume that K is continuous; the following theorem asserts that this is a sufficient condition for existence of an optimal strategy.

THEOREM 7.1 (Continuous time: existence). *Let $K : [0, T] \to (0, \infty)$ be continuous. Then, there exists an optimal strategy $\Theta^* \in \mathcal{A}_t(x)$, i.e.,*

$$J(t, \delta, \Theta^*) = \inf_{\Theta \in \mathcal{A}_t(x)} J(t, \delta, \Theta).$$

In the proof, we construct an optimal strategy as the limit of a sequence of (possibly suboptimal) strategies. Before we can turn to the proof itself, we need to establish that strategy convergence leads to cost convergence.

PROPOSITION 7.2 (Costs are continuous in the strategy, K continuous). *Let $K : [0, T] \to (0, \infty)$ be continuous and let $\bar{\Theta}, (\Theta^n)$ be strategies in $\mathcal{A}_t(x)$ with $\Theta^n \overset{w}{\to} \bar{\Theta}$, i.e., $\lim_{n\to\infty} \Theta^n_s = \bar{\Theta}_s$ for every point $s \in [t, T]$ of continuity of $\bar{\Theta}$ (i.e., Θ^nconverges weakly to $\bar{\Theta}$). Then,*

$$|J(t, \delta, \bar{\Theta}) - J(t, \delta, \Theta^n)| \xrightarrow[n\to\infty]{} 0.$$

Note that Proposition 7.2 does not hold when K has a jump. To prove Proposition 7.2, we first show in Lemma 7.3 that the convergence of the price impact processes follows from the weak convergence of the corresponding strategies. We then conclude in Lemma 7.5 that Proposition 7.2 holds for absolutely continuous K. This finally leads to Proposition 7.2 covering all continuous K.

LEMMA 7.3 (Price impact process is continuous in the strategy). *Let K: $[0, T] \to (0, \infty)$ be continuous and let $\bar{\Theta}, (\Theta^n)$ be strategies in $\mathcal{A}_t(x)$ with $\Theta^n \overset{w}{\to} \bar{\Theta}$.*

Then, $\lim_{n\to\infty} D^n_s = \bar{D}_s$ for $s = T+$ and for every point $s \in [t, T]$ of continuity of $\bar{\Theta}$.

Proof. Recall equation (4.2)

$$D_s = \int_{[t,s)} K_u e^{-\int_u^s \rho_r dr} d\Theta_u + \delta e^{-\int_t^s \rho_u du},$$

which holds for $s = T+$ and $s \in [t, T]$. Due to the weak convergence (note that the total mass is preserved, i.e., $\bar{\Theta}_{T+} = \Theta^n_{T+} = x$, since $\bar{\Theta}, \Theta^n \in \mathcal{A}_t(x)$) and the integrand being

[6] Let us, however, note that the value function here has WR-BR structure with the barrier from Example 9.5 with $\kappa = 1$.

continuous in u, the assertion follows for $s = T+$. Due to the weak convergence, we also have that for all $s \in [t, T]$ with $\Delta\bar{\Theta}_s = 0$ and $f_s(u) := K_u e^{-\int_u^s \rho_r dr} I_{[t,s)}(u)$ (i.e., f_s is continuous $d\bar{\Theta}$-a.e.)

$$D_s^n = \int_{[t,T]} f_s(u) d\Theta_u^n + \delta e^{-\int_t^s \rho_u du} \xrightarrow[n\to\infty]{} \int_{[t,T]} f_s(u) d\bar{\Theta}_u + \delta e^{-\int_t^s \rho_u du} = \bar{D}_s.$$

□

LEMMA 7.4 (Costs rewritten in terms of the price impact process). *Let $K : [0, T] \to (0, \infty)$ be absolutely continuous, i.e., $K_s = K_0 + \int_0^s \mu_u \, du$. Then,*

$$J(t, \delta, \Theta) = \frac{1}{2}\left[\frac{D_{T+}^2}{K_T} - \frac{\delta^2}{K_t} + \int_{[t,T]} \left(\frac{2\rho_s}{K_s} + \frac{\mu_s}{K_s^2}\right) D_s^2 ds\right]. \tag{7.2}$$

Proof. Applying

$$d\Theta_s = \frac{dD_s + \rho_s D_s ds}{K_s}, \quad \Delta\Theta_s = \frac{\Delta D_s}{K_s}$$

yields

$$\begin{aligned} J(t, \delta, \Theta) &= \int_{[t,T]} \left(D_s + \frac{K_s}{2}\Delta\Theta_s\right) d\Theta_s \\ &= \int_{[t,T]} \frac{D_s + \frac{1}{2}\Delta D_s}{K_s} dD_s + \int_{[t,T]} \frac{\rho_s D_s^2}{K_s} ds + \int_{[t,T]} \frac{\frac{1}{2}\Delta D_s \rho_s D_s}{K_s} ds. \end{aligned}$$

In this expression, the last term is zero since D has only countably many jumps. Using integration by parts for càglàd processes, namely (A.2) with $U := D$, $V := \frac{D}{K}$, and $d(\frac{D_s}{K_s}) = \frac{1}{K_s} dD_s + D_s d(\frac{1}{K_s})$, we can write

$$\int_{[t,T]} \frac{D_s}{K_s} dD_s = \frac{1}{2}\left[\frac{D_{T+}^2}{K_T} - \frac{\delta^2}{K_t} - \int_{[t,T]} D_s^2 d\left(\frac{1}{K_s}\right) - \sum_{s\in[t,T]} \frac{(\Delta D_s)^2}{K_s}\right].$$

Plugging in $d(\frac{1}{K_s}) = -\frac{\mu_s}{K_s^2} ds$ yields (7.2) as desired. □

The following result is a direct consequence of Lemma 7.3 and Lemma 7.4.

LEMMA 7.5 (Costs are continuous in the strategy, K absolutely continuous). *Let $K : [0, T] \to (0, \infty)$ be absolutely continuous and $\bar{\Theta}, (\Theta^n)$ be strategies in $\mathcal{A}_t(x)$ with $\Theta^n \overset{w}{\to} \bar{\Theta}$. Then*

$$|J(t, \delta, \bar{\Theta}) - J(t, \delta, \Theta^n)| \xrightarrow[n\to\infty]{} 0.$$

Proof of Proposition 7.2. We use a proof by contradiction and suppose there exists a subsequence $(n_j) \subset \mathbb{N}$ such that

$$\lim_{j\to\infty} \int_{[t,T]} \left(D_s^{n_j} + \frac{K_s}{2}\Delta\Theta_s^{n_j}\right) d\Theta_s^{n_j} \neq \int_{[t,T]} \left(\bar{D}_s + \frac{K_s}{2}\Delta\bar{\Theta}_s\right) d\bar{\Theta}_s,$$

where the limit on the left-hand side exists. Without loss of generality, assume

$$\lim_{j\to\infty}\int_{[t,T]}\left(D_s^{n_j}+\frac{K_s}{2}\Delta\Theta_s^{n_j}\right)d\Theta_s^{n_j}<\int_{[t,T]}\left(\bar{D}_s+\frac{K_s}{2}\Delta\bar{\Theta}_s\right)d\bar{\Theta}_s. \tag{7.3}$$

We now want to bring Lemma 7.5 into play. For $\epsilon>0$, we denote by $K^\epsilon\colon[t,T]\to(0,\infty)$ an absolutely continuous function such that $\max_{s\in[t,T]}\left|K_s^\epsilon-K_s\right|\le\epsilon$. For $\Theta\in\mathcal{A}_t(x)$,

$$\begin{aligned}&\left|\int_{[t,T]}\left(D_s^\epsilon+\frac{K_s^\epsilon}{2}\Delta\Theta_s\right)d\Theta_s-\int_{[t,T]}\left(D_s+\frac{K_s}{2}\Delta\Theta_s\right)d\Theta_s\right|\\&\qquad\le\int_{[t,T]}\left(|D_s^\epsilon-D_s|+\frac{1}{2}|K_s^\epsilon-K_s|\Delta\Theta_s\right)d\Theta_s\le\frac{3}{2}x^2\epsilon.\end{aligned}$$

We therefore get from (7.3) that there exists $\epsilon>0$ such that

$$\limsup_{j\to\infty}\int_{[t,T]}\left(D_s^{n_j,\epsilon}+\frac{K_s^\epsilon}{2}\Delta\Theta_s^{n_j}\right)d\Theta_s^{n_j}<\int_{[t,T]}\left(\bar{D}_s^\epsilon+\frac{K_s^\epsilon}{2}\Delta\bar{\Theta}_s\right)d\bar{\Theta}_s.$$

This is a contradiction to Lemma 7.5. □

We can now conclude the proof of the existence Theorem 7.1.

Proof of Theorem 7.1. Let $(\Theta^n)\subset\mathcal{A}_t(x)$ be a minimizing sequence. Due to the monotonicity of the considered strategies, we can use *Helly's Theorem* in the form of theorem 2, section 2, chapter III of Shiryaev (1995), which also holds for left continuous processes and on $[t,T]$ instead of $(-\infty,\infty)$. It guarantees the existence of a deterministic $\bar{\Theta}\in\mathcal{A}_t(x)$ and a subsequence $(n_j)\subset\mathbb{N}$ such that (Θ^{n_j}) converges weakly to $\bar{\Theta}$. Note that we can always force $\bar{\Theta}_{T+}$ to be x, since weak convergence does not require that $\Theta_T^{n_j}$ converges to $\bar{\Theta}_T$ whenever $\bar{\Theta}$ has a jump at T. Due to Proposition 7.2, we can conclude that

$$U(t,\delta,x)=\lim_{j\to\infty}J(t,\delta,\Theta^{n_j})=J(t,\delta,\bar{\Theta}).$$ □

The price impact process D is affine in the corresponding strategy Θ. That is, in the case when K is not decreasing too quickly, Lemma 7.4 guarantees that the cost term J is strictly convex in the strategy Θ. Therefore, we get the following uniqueness result.

THEOREM 7.6. (Continuous time: uniqueness). *Let $K:[0,T]\to(0,\infty)$ be absolutely continuous, i.e., $K_s=K_0+\int_0^s\mu_u du$, and additionally*

$$\mu_s+2\rho_sK_s>0\textit{ a.e. on }[0,T]\textit{ with respect to the Lebesgue measure.}$$

Then, there exists a unique optimal strategy.

7.2. WR-BR Structure

For continuous K, we have now established existence and (under additional conditions) uniqueness of the optimal strategy. Let us now turn to the value function in continuous time and demonstrate that it has WR-BR structure, consistent with our findings in discrete time.

THEOREM 7.7 (Continuous time: WR-BR structure). *Let $K : [0, T] \to (0, \infty)$ be continuous. Then, the value function has WR-BR structure.*

We are going to deduce the structural result for the continuous time setting by using our discrete-time result. First, we show that the discrete time value function converges to the continuous time value function. Without loss of generality, we set $t = 0$.

LEMMA 7.8 (The discrete time value function converges to the continuous time one). *Let $K : [0, T] \to (0, \infty)$ be continuous and consider an equidistant time grid with N trading intervals. Then*

$$\lim_{N\to\infty} V^N(0, y) = V(0, y).$$

Proof. Due to Theorem 7.1, there exists a continuous time optimal strategy $\Theta^* \in \mathcal{A}_0(y)$. Approximate it suitably via step functions $\Theta^N \in \mathcal{A}_0^N(y)$. Then

$$V(0, y) = J(0, 1, \Theta^*) = \lim_{N\to\infty} J(0, 1, \Theta^N) \geq \limsup_{N\to\infty} V^N(0, y).$$

The inequality $V(0, y) \leq \liminf_{N\to\infty} V^N(0, y)$ is immediate. □

Proof of Theorem 7.7. By the same change of variable from ξ to η that was used in Section 6, we can transform the optimal trade equation

$$V(0, y) = \min_{\xi\in[0,y]} \left\{ \left(1 + \frac{K_0}{2}\xi\right)\xi + (1 + K_0\xi)^2 V\left(0, \frac{y-\xi}{1 + K_0\xi}\right)\right\}$$

into the optimal barrier equation

$$V(0, y) = \frac{1}{2K_0}\Big[(1 + K_0 y)^2 \min_{\eta\in[0,y]} L(0, \eta) - 1\Big], \tag{7.4}$$

where

$$L(0, y) := L(y) := \frac{1 + 2K_0 V(0, y)}{(1 + K_0 y)^2}. \tag{7.5}$$

Now it follows from (7.4) and (7.5) that

$$\min_{\eta\in[0,y]} L(\eta) = L(y),$$

in particular, the function L is nonincreasing in y. Define

$$\tilde{L}^N(y) := \min_{\eta\in[0,y]} L^N(0, \eta),$$

which is a nonincreasing positive function. If for some N, the function $y \mapsto L^N(0, y)$ is not convex on $[0, c^N(0))$, then the second alternative in Lemma 6.3 holds, i.e., we have WR-BR structure with $c(0) = \infty$. Thus, below, we assume that for any N, the function $y \mapsto L^N(0, y)$ is convex on $[0, c^N(0))$; hence, by Lemma 6.2 (a) and Theorem 6.1, the function $\tilde{L}^N$ is convex on $[0, \infty)$. Moreover, by rearranging (6.5), we obtain that

$$\tilde{L}^N(y) = \frac{1 + 2K_0 V^N(0, y)}{(1 + K_0 y)^2}.$$

Hence, $\tilde{L}^N$ converges pointwise to L as $N \to \infty$ by Lemma 7.8 and (7.5). Therefore, L is also convex.

Due to L being nonincreasing and convex, there exists a unique $c^* \in [0, \infty]$ such that L is strictly decreasing for $y \in [0, c^*)$ and constant for $y \in (c^*, \infty)$. One can now conclude that for all $y > c^*$ and $\eta \in (c^*, y)$, setting $\xi := \frac{y-\eta}{1+K_0\eta}$, i.e., $\eta = \frac{y-\xi}{1+K_0\xi}$, and using (7.5) and $L(y) = L(\eta)$, we have

$$\begin{aligned} V(0, y) &= \frac{1}{2K_0}[(1 + K_0 y)^2 L(y) - 1] \\ &= \frac{1}{2K_0}[(1 + K_0 y)^2 L(\eta) - 1] \\ &= \frac{1}{2K_0}[(1 + K_0 y)^2 L\left(\frac{y-\xi}{1+K_0\xi}\right) - 1]. \end{aligned}$$

We now use the definition of L from (7.5) once again to get

$$V(0, y) = \left(1 + \frac{K_0}{2}\xi\right)\xi + (1 + K_0\xi)^2 V\left(0, \frac{y-\xi}{1+K_0\xi}\right). \tag{7.6}$$

Therefore, $(c^*, \infty) \subset Br_0$. In case of $c^* > 0$, consider $y \le c^*$, take any $\eta \in [0, y)$, and set $\xi := \frac{y-\eta}{1+K_0\eta}$. Then, a similar calculation using that $L(y) < L(\eta)$ shows that $V(0, y)$ is strictly smaller than the right-hand side of (7.6). Hence, $Br_0 = (c^*, \infty)$. Thus, we get WR-BR structure with $c(0) = c^* \in [0, \infty]$ as desired. □

In Section 9, we will investigate the value function, barrier function, and optimal trading strategies for several example specifications of K and ρ.

8. ZERO SPREAD AND PRICE MANIPULATION

In the model introduced in Section 2, we assumed a trading-dependent spread between the best ask A_t and best bid B_t. This has allowed us to exclude both forms of price manipulation in Section 3. An alternative assumption that is often made in limit order book models is to disregard the bid-ask spread and to assume $A_t = B_t$, see, for example, Huberman and Stanzl (2004), Gatheral (2010), Alfonsi et al. (2011), and Gatheral et al. (2012). The canonical extension of these models to our framework including time-varying liquidity is the following.

ASSUMPTION 8.1. *In the* zero spread model*, we have the* unaffected price S^u*, which is a càdlàg* $\mathcal{H}^1$*-martingale with a deterministic starting point* S_0^u*, and assume that the best bid and ask are equal and given by* $A_t^\updownarrow = B_t^\updownarrow = S_t^u + D_t^\updownarrow$ *with*

$$D_t^\updownarrow = D_0^\updownarrow e^{-\int_0^t \rho_s\, ds} + \int_{[0,t)} K_s e^{-\int_s^t \rho_u\, du}(d\Theta_s - d\tilde{\Theta}_s), \quad t \in [0, T+], \tag{8.1}$$

where $D_0^\updownarrow \in \mathbb{R}$ *is the initial value for the price impact. For convenience, we will furthermore assume that* $K : [0, T] \to (0, \infty)$ *is twice continuously differentiable and* ρ: $[0, T] \to (0, \infty)$ *is continuously differentiable.*

As opposed to this zero spread model, the model introduced in Section 2 will be referred to as the *dynamic spread model* in the sequel. In this section, we study price manipulation and optimal execution in the zero spread model. In particular, we provide explicit formulas for optimal strategies. This, in turn, will be used in the next section to study explicitly several examples both in the dynamic spread and in the zero spread model.

We have excluded permanent impact from the definition above ($\gamma = 0$). It can easily be included, but, like in the dynamic spread model, proves to be irrelevant for optimal strategies and price manipulation. Note that for pure buying strategies ($\tilde{\Theta} \equiv 0$), the zero spread model is identical to the model introduced in Section 2. The difference between the two models is that if sell orders occur, then they are executed at the same price as the ask price. Furthermore, buy and sell orders impact this price symmetrically. We can hence consider the net trading strategy $\Theta^\updownarrow := \Theta - \tilde{\Theta}$ instead of buy orders Θ and sell orders $\tilde{\Theta}$ separately. The simplification of the stochastic optimization problem of Section 2.2 to a deterministic problem in Section 4 applies similarly to the zero spread model defined in Assumption 8.1. Thus, for any fixed $x \in \mathbb{R}$, we define the sets of strategies

$$\begin{aligned}\mathcal{A}^\updownarrow &:= \{\Theta^\updownarrow : [0, T+] \to \mathbb{R} \mid \Theta^\updownarrow \text{ is a deterministic càglàd} \\ &\qquad \text{function of finite variation with } \Theta_0^\updownarrow = 0\}, \\ \mathcal{A}^\updownarrow(x) &:= \left\{\Theta^\updownarrow \in \mathcal{A}^\updownarrow \middle| \, \Theta_{T+}^\updownarrow = x\right\}.\end{aligned}$$

Strategies from $\mathcal{A}^\updownarrow(x)$ allow buying and selling and build up the position of x shares until time T. We further define the cost function $J^\updownarrow : \mathbb{R} \times \mathcal{A}^\updownarrow \to \mathbb{R}$ as

$$J^\updownarrow(\Theta^\updownarrow) := J(\delta, \Theta^\updownarrow) := \int_{[0,T]} \left(D_s^\updownarrow + \frac{K_s}{2}\Delta\Theta_s^\updownarrow\right) d\Theta_s^\updownarrow,$$

where $D^\updownarrow$ is given by (8.1) with $D_0^\updownarrow = \delta$. The function $J^\updownarrow$ represents the total temporary impact costs[7] in the zero spread model of the strategy $\Theta^\updownarrow$ on the time interval $[0, T]$ when the initial price impact $D_0^\updownarrow = \delta$. Observe that $J^\updownarrow$ is well-defined and finite because K is bounded, which, in turn, follows from Assumption 8.1. The value function $U^\updownarrow : \mathbb{R}^2 \to \mathbb{R}$ is then given as

$$U^\updownarrow(\delta, x) := \inf_{\Theta^\updownarrow \in \mathcal{A}^\updownarrow(x)} J^\updownarrow(\delta, \Theta^\updownarrow). \tag{8.2}$$

The zero spread model admits *price manipulation* if, for $D_0^\updownarrow = 0$, there is a profitable round trip, i.e., there is $\Theta^\updownarrow \in \mathcal{A}^\updownarrow(0)$ with $J^\updownarrow(0, \Theta^\updownarrow) < 0$. The zero spread model admits *transaction-triggered price manipulation* if, for $D_0^\updownarrow = 0$, the execution costs of a buy (or sell) program can be decreased by intermediate sell (respectively, buy) trades (more precisely, this should be formulated like in Definition 3.2).

REMARK 8.2. The conceptual difference with Section 3 is that we require $D_0^\updownarrow = 0$ in these definitions. The reason is that even in "sensible" zero spread models (that do not admit both types of price manipulation according to definitions above), we typically have profitable round trips whenever $D_0^\updownarrow \neq 0$. In the zero spread model, the case $D_0^\updownarrow \neq 0$ can be interpreted as that the market price is not in its equilibrium state in the beginning. In the absence of trading, the process $(D_t^\updownarrow)$ approaches zero due to the resilience; hence, both best ask and best bid price processes $(A_t^\updownarrow)$ and $(B_t^\updownarrow)$ (which are equal) approach their

[7] In the case of liquidation of shares (i.e., $\Theta_{T+}^\updownarrow < 0$), the word "costs" should be understood as "minus proceeds from the liquidation."

evolution in the equilibrium (S^u_t). The knowledge of this "direction of deviation from S^u" plus the fact that both buy and sell orders are executed at the same price[8] clearly allow us to construct profitable round trips. For instance, in the Obizhaeva–Wang-type model with a constant price impact $K_t \equiv \kappa > 0$, the strategy

$$\Theta^{\updownarrow}_s := -\frac{D^{\updownarrow}_0}{2\kappa} I_{(0,\epsilon]}(s), \quad s \in [0, T+],$$

where $\epsilon \in (0, T]$ is a profitable round trip whenever $D^{\updownarrow}_0 \neq 0$, as can be checked by a straightforward calculation.

Let us first discuss classical price manipulation in the zero spread model. If the liquidity in the order book rises too fast (K falls too quickly), then a simple pump and dump strategy becomes attractive. In the initial low liquidity regime (high K), buying a large amount of shares increases the price significantly. Quickly thereafter liquidity increases. Then, the position can be liquidated with little impact at this elevated price, leaving the trader with a profit. The following result formalizes this line of thought.

PROPOSITION 8.3 (Price manipulation in the zero spread model). *Assume the zero spread model of Assumption 8.1 and that*

$$K'_t + 2\rho_t K_t < 0 \quad \textit{for some} \quad t \in [0, T).$$

Then price manipulation occurs and, for any δ, $x \in \mathbb{R}$, there is no optimal strategy in problem (8.2).

Proof. By the assumption of the theorem,

$$K'_t = \lim_{\epsilon \searrow 0} \frac{K_{t+\epsilon} - K_t}{\epsilon} < -2\rho_t K_t = \lim_{\epsilon \searrow 0} \frac{K_t\big(2e^{-\int_t^{t+\epsilon} \rho_u\,du} - 1\big) - K_t}{\epsilon},$$

hence, for a sufficiently small $\epsilon > 0$, we have

$$K_{t+\epsilon} < K_t\big(2e^{-\int_t^{t+\epsilon} \rho_u\,du} - 1\big). \tag{8.3}$$

Let us consider the round trip $\Theta^{\updownarrow m} \in \mathcal{A}^{\updownarrow}(0)$, which buys 1 share at time t and sells it at time $t + \epsilon$, i.e.,

$$\Theta^{\updownarrow m}_s := I_{(t,t+\epsilon]}(s), \quad s \in [0, T+].$$

A straightforward computation shows that, for $D^{\updownarrow}_0 = 0$, the cost of such a round trip is

$$J^{\updownarrow}(0, \Theta^{\updownarrow m}) = \frac{K_t + K_{t+\epsilon}}{2} - K_t e^{-\int_t^{t+\epsilon} \rho_u\,du}.$$

Due to (8.3), $J^{\updownarrow}(0, \Theta^{\updownarrow m}) < 0$. Thus, price manipulation occurs.

Let us fix δ, $x \in \mathbb{R}$, consider a strategy $\Theta^{\updownarrow} \in \mathcal{A}^{\updownarrow}(x)$, and, for any $z \in \mathbb{R}$, define

$$\Theta^{\updownarrow z} := \Theta^{\updownarrow} + z\Theta^{\updownarrow m}.$$

Then, $\Theta^{\updownarrow z} \in \mathcal{A}^{\updownarrow}(x)$ and we have

[8] Let us observe that this does not apply to the dynamic spread model of Section 2, where we have different processes D and E for the deviations of the best ask and best bid prices from the unaffected ones due to the previous trades.

$$J^{\ddagger}(\delta, \Theta^{\ddagger z}) = c_0 z^2 + c_1 z + c_2$$

with $c_0 = J^{\ddagger}(0, \Theta^{\ddagger m}) < 0$ and some constants c_1 and c_2. Since z is arbitrary, we get $U^{\ddagger}(\delta, x) = -\infty$. An optimal strategy in this situation would be a strategy from $\mathcal{A}^{\ddagger}(x)$ with the cost $-\infty$. But for any strategy $\Theta^{\ddagger}$, its cost $J^{\ddagger}(\delta, \Theta^{\ddagger})$ is finite as discussed above; hence, there is no optimal strategy in problem (8.2). □

Interestingly, the condition $K_t' + 2\rho_t K_t < 0$ for some t in Proposition 8.3 is not symmetric; quickly falling K leads to price manipulation, but quickly rising K does not.

If $K_t' + 2\rho_t K_t \geq 0$ holds at all points in time, then the situation remains unclear so far. In their model, Alfonsi et al. (2011) and Gatheral et al. (2012) have shown that even in the absence of profitable round-trip trades, we might still be facing transaction-triggered price manipulation. This can happen also in our zero spread model. The following theorem provides explicit formulas for optimal strategies and leads to a characterization of transaction-triggered price manipulation.

THEOREM 8.4 (Optimal strategies in the zero spread model). *Assume the zero spread model of Assumption 8.1 and that* $K_t' + 2\rho_t K_t > 0$ *on* $[0, T]$. *Define*

$$f_t := \frac{K_t' + \rho_t K_t}{K_t' + 2\rho_t K_t}, \quad t \in [0, T]. \tag{8.4}$$

Then, for any $\delta, x \in \mathbb{R}$, *the strategy* $\Theta^{\ddagger *}$ *given by the formulas*

$$\Delta\Theta_0^{\ddagger *} = \delta^{\ddagger}\frac{f_0}{K_0} - \frac{\delta}{K_0}, \qquad d\Theta_t^{\ddagger *} = \delta^{\ddagger}\frac{f_t' + \rho_t f_t}{K_t}dt, \qquad \Delta\Theta_T^{\ddagger *} = \delta^{\ddagger}\frac{1 - f_T}{K_T}, \tag{8.5}$$

with $\delta^{\ddagger} := \frac{1}{c}(x + \frac{\delta}{K_0})$ *and*

$$c := \int_0^T \frac{f_t' + \rho_t f_t}{K_t}dt + \frac{f_0}{K_0} + \frac{1 - f_T}{K_T} > 0 \tag{8.6}$$

is the unique optimal strategy in problem (8.2). Furthermore, we have

$$U^{\ddagger}(\delta, x) = J^{\ddagger}(\delta, \Theta^{\ddagger *}) = (\delta^{\ddagger})^2\left(\int_0^T (K_t' + 2\rho_t K_t)\frac{f_t^2}{2K_t^2}\,dt + \frac{1}{2K_T}\right) - \frac{\delta^2}{2K_0}. \tag{8.7}$$

COROLLARY 8.5 (Transaction-triggered price manipulation in the zero spread model). *Under the assumptions of Theorem 8.4, price manipulation does not occur. Furthermore, transaction-triggered price manipulation occurs if and only if* $f_0 < 0$ *or* $f_t' + \rho_t f_t < 0$ *for some* $t \in [0, T]$.

Proof. Using (8.5) with $\delta = 0$, we immediately get that price manipulation does not occur. Noting further that $f_T < 1$, we obtain that transaction-triggered price manipulation occurs if and only if either $f_0 < 0$ or $f_t' + \rho_t f_t < 0$ for some $t \in [0, T]$. □

We can summarize Proposition 8.3 and Corollary 8.5 as follows. If $K_t' + 2\rho_t K_t < 0$ for some t, i.e., liquidity grows very rapidly, then price manipulation (and hence transaction-triggered price manipulation) occurs. If $K_t' + 2\rho_t K_t > 0$ everywhere, but $f_0 < 0$ or $f_t' + \rho_t f_t < 0$ for some t, i.e., liquidity grows fast but not quite as fast, then price manipulation does not occur, but transaction-triggered price manipulation occurs. If $K_t' + 2\rho_t K_t > 0$

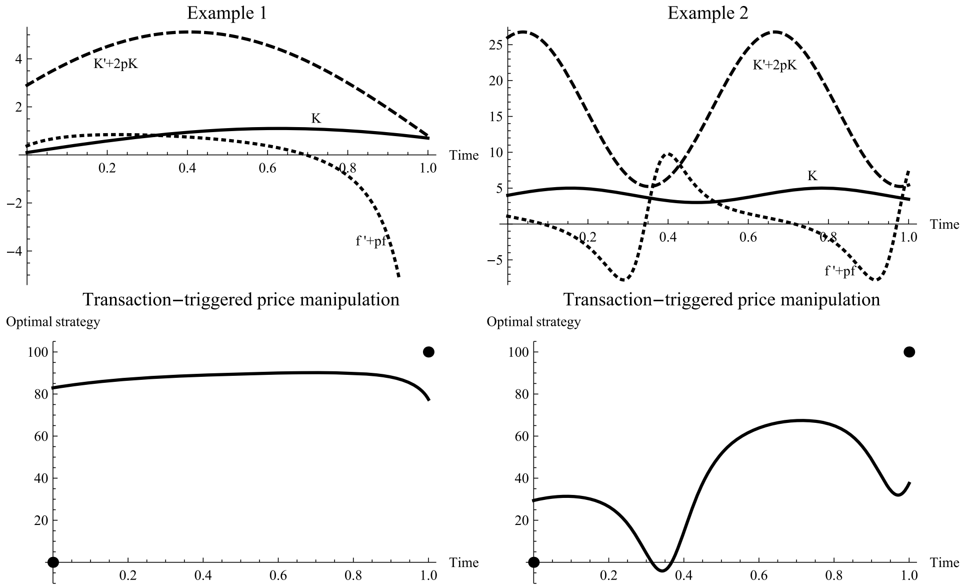

FIGURE 8.1. In Example 1, we consider $K_t = \sin(2.5t) + 0.1$ and $K_t = \sin(10t) + 4$ in Example 2. The other parameters are $T = 1$, $\rho = 2$, and $x = 100$, $\delta = 0$. The plots at the bottom illustrate the corresponding optimal strategies $\Theta^{\ddagger *}$ from (8.5).

everywhere, $f_0 \geq 0$ and $f_t' + \rho_t f_t \geq 0$ everywhere, i.e., liquidity never grows too fast, then neither form of price manipulation occurs and an investor wishing to purchase should only submit buy orders to the market. Figure 8.1 illustrates optimal transaction-triggered price manipulation strategies. In Example 1, the liquidity $\frac{1}{K}$ is slightly growing at the end of the trading horizon, which makes the optimal strategy $\Theta^{\ddagger *}$ nonmonotonic. As can be seen in Example 2, the number of shares held by the large investor can become negative although the overall goal is to buy a positive amount of shares.

In the proof of Theorem 8.4, we are going to exploit the fact that there is a one-to-one correspondence between $\Theta^\ddagger$ and $D^\ddagger$. We rewrite the cost term, which is essentially $\int_0^T D_t^\ddagger d\Theta_t^\ddagger$, in terms of the deviation process $D^\ddagger$ by applying

$$d\Theta_t^\ddagger = \frac{dD_t^\ddagger + \rho_t D_t^\ddagger dt}{K_t}, \tag{8.8}$$

and get the following result.

LEMMA 8.6 (Costs rewritten in terms of the price impact process). *Under Assumption 8.1, for any $\delta \in \mathbb{R}$ and $\Theta^\ddagger \in \mathcal{A}^\ddagger$, we have*

$$J^\ddagger(\delta, \Theta^\ddagger) = \frac{1}{2}\left[\frac{\left(D_{T+}^\ddagger\right)^2}{K_T} - \frac{\delta^2}{K_0} + \int_{[0,T]} (K_t' + 2\rho_t K_t)\frac{\left(D_t^\ddagger\right)^2}{K_t^2}\,dt\right]. \tag{8.9}$$

The formal proof, where one needs to take into account possible jumps of $\Theta^\ddagger$, is similar to that of Lemma 7.4.

Similar to Bank and Becherer (2009) and as explained in Gregory and Lin (1996), we can now use the Euler–Lagrange formalism to find necessary conditions on the optimal

$D^{\updownarrow}$. Under our assumptions, these conditions turn out to be sufficient and the optimal $D^{\updownarrow}$ directly gives us an optimal $\Theta^{\updownarrow}$. We obtained the formulas in Theorem 8.4 by pretending that $D^{\updownarrow}$ is smooth and then solving the Euler–Lagrange equation for $D^{\updownarrow}$. While there is no solution in a strict sense (because the optimal $D^{\updownarrow}$ has jumps at times 0 and T), we obtained our formulas by approximating a function $D^{\updownarrow}$ with jumps by smooth functions. Although we used the Euler–Lagrange approach to obtain the formulas in Theorem 8.4, we do not use it in the proof; instead, we prove by direct verification that the strategy $\Theta^{\updownarrow *}$ is optimal in the whole class $\mathcal{A}^{\updownarrow}(x)$.

First, we need an approximation argument. For $x \in \mathbb{R}$, let us define the set of strategies $\mathcal{A}^{\updownarrow c}(x) \subset \mathcal{A}^{\updownarrow}(x)$ with impulse trades at $t = 0$ and $t = T$ only:

$$\mathcal{A}^{\updownarrow c}(x) := \{\Theta^{\updownarrow} \in \mathcal{A}^{\updownarrow}(x) \mid \Theta^{\updownarrow} \text{ is continuous on } (0, T)\}.$$

We will also need a notation for a similar set of monotonic strategies, i.e., for $y \in [0, \infty)$, we define

$$\mathcal{A}_0^c(y) := \{\Theta \in \mathcal{A}_0(y) \mid \Theta \text{ is continuous on } (0, T)\}.$$

LEMMA 8.7 (Approximation by continuous strategies). *Assume the zero spread model of Assumption 8.1. Then, for any δ, $x \in \mathbb{R}$,*

$$U^{\updownarrow}(\delta, x) := \inf_{\Theta^{\updownarrow} \in \mathcal{A}^{\updownarrow}(x)} J^{\updownarrow}(\delta, \Theta^{\updownarrow}) = \inf_{\Theta^{\updownarrow} \in \mathcal{A}^{\updownarrow c}(x)} J^{\updownarrow}(\delta, \Theta^{\updownarrow}). \tag{8.10}$$

Proof. Let us take any $\Theta^{\updownarrow} \in \mathcal{A}^{\updownarrow}(x)$ and find $\Theta, \tilde{\Theta} \in \mathcal{A}_0$ such that $\Theta^{\updownarrow} = \Theta - \tilde{\Theta}$. We set $y := \Theta_{T+} \in [0, \infty)$, $\tilde{y} := \tilde{\Theta}_{T+} \in [0, \infty)$ so that $x = y - \tilde{y}$. Below, we will show that

$$\exists\, \Theta^n \in \mathcal{A}_0^c(y), \quad \tilde{\Theta}^n \in \mathcal{A}_0^c(\tilde{y}) \ \text{ such that } \ \Theta^n \xrightarrow{w} \Theta, \quad \tilde{\Theta}^n \xrightarrow{w} \tilde{\Theta}. \tag{8.11}$$

Let us define $\Theta^{\updownarrow n} := \Theta^n - \tilde{\Theta}^n \in \mathcal{A}^{\updownarrow c}(x)$. It follows from (8.1) and the weak convergence of the strategies that the price impact $D_t^{\updownarrow n}$ corresponding to $\Theta^{\updownarrow n}$ converges to the price impact $D_t^{\updownarrow}$ corresponding to $\Theta^{\updownarrow}$ for $t = T+$ and for every point $t \in [0, T]$, where both Θ and $\tilde{\Theta}$ are continuous (i.e., the convergence of price impact functions holds at $T+$ and everywhere on $[0, T]$ except at most a countable set). By (8.9), we get $J^{\updownarrow}(\delta, \Theta^{\updownarrow n}) \to J^{\updownarrow}(\delta, \Theta^{\updownarrow})$ as $n \to \infty$. Since $\Theta^{\updownarrow} \in \mathcal{A}^{\updownarrow}(x)$ was arbitrary, we obtain (8.10).

It remains to prove (8.11). Clearly, it is enough to consider some $\Theta \in \mathcal{A}_0(y)$ and to construct $\Theta^n \in \mathcal{A}_0^c(y)$ weakly convergent to Θ. Let $\mathcal{P}$ denote the class of all probability measures P on $([0, T], \mathcal{B}([0, T]))$ and

$$\mathcal{P}^c = \{P \in \mathcal{P} | P(\{s\}) = 0 \quad \text{for all} \quad s \in (0, T)\}.$$

The formula $P([0, s)) := \frac{\Theta_s}{y}$, $s \in [0, T]$, with $\Theta \in \mathcal{A}_0(y)$, provides a one-to-one correspondence between $\mathcal{A}_0(y)$ and $\mathcal{P}$, where $\mathcal{A}_0^c(y)$ is mapped on $\mathcal{P}^c$. Thus, it is enough to show that any probability measure $P \in \mathcal{P}$ can be weakly approximated by probability measures from $\mathcal{P}^c$. To this end, let us consider independent random variables ψ and ζ such that $\text{Law}(\psi) = P$ and $\text{Law}(\zeta)$ is continuous. For any $n \in \mathbb{N}$, we define

$$\psi_n := \left(\left(\psi + \frac{\zeta}{n}\right) \vee 0\right) \wedge T.$$

Then

$$Q_n := \mathrm{Law}(\psi_n) \in \mathcal{P}^c$$

and $Q_n \overset{w}{\to} P$ as $n \to \infty$ because $\psi_n \to \psi$ a.s. This concludes the proof. □

LEMMA 8.8. *Assume $K_t' + 2\rho_t K_t > 0$ on $[0, T]$ and define*

$$\chi(t) := \int_0^t \frac{f_s' + \rho_s f_s}{K_s} dt + \frac{f_0}{K_0} + \frac{1 - f_t}{K_t}.$$

Then, $\chi(t) > 0$ for all $t \in [0, T]$. In particular, $c = \chi(T) > 0$.

Proof. We have

$$\chi(0) = \frac{1}{K_0} > 0.$$

Furthermore,

$$\chi'(t) = \frac{f_t' + \rho_t f_t}{K_t} + \frac{-f_t' K_t - (1 - f_t) K_t'}{K_t^2} = \frac{\rho_t^2}{K_t' + 2\rho_t K_t} > 0.$$

□

Proof of Theorem 8.4 We first note that c from (8.6) is strictly positive by Lemma 8.8. Also, note that if an optimal strategy in (8.2) exists, then it is unique in the class $\mathcal{A}^{\updownarrow}(x)$ because the function $\Theta^{\updownarrow} \mapsto J^{\updownarrow}(\delta, \Theta^{\updownarrow})$ is strictly convex on $\mathcal{A}^{\updownarrow}$ (this is due to (8.9) and the assumption $K_t' + 2\rho_t K_t > 0$ on $[0, T]$).

For the strategy $\Theta^{\updownarrow *}$ given in (8.5), we have $\Theta_{T+}^{\updownarrow *} = x$ as desired. This follows from the formula for $\delta^{\updownarrow}$. Let us further observe that $\Theta^{\updownarrow *}$ corresponds to the deviation process

$$D_0^{\updownarrow *} = \delta, \quad D_t^{\updownarrow *} = \delta^{\updownarrow} f_t \text{ on } (0, T], \quad D_{T+}^{\updownarrow *} = \delta^{\updownarrow}, \tag{8.12}$$

which immediately follows from (8.8) (direct computation using (8.1) is somewhat longer). A straightforward calculation gives

$$J^{\updownarrow}(\delta, \Theta^{\updownarrow *}) = (\delta^{\updownarrow})^2 \left(\int_0^T f_t \frac{f_t' + \rho_t f_t}{K_t} dt + \frac{f_0^2}{2K_0} + \frac{1 - f_T^2}{2K_T} \right) - \frac{\delta^2}{2K_0}. \tag{8.13}$$

Using integration by parts, we get

$$\int_0^T \frac{f_t f_t'}{K_t} dt = \frac{1}{2} \left[\frac{f_T^2}{K_T} - \frac{f_0^2}{K_0} + \int_0^T \frac{f_t^2}{K_t^2} K_t' \, dt \right].$$

Substituting this into (8.13), we get that $J^{\updownarrow}(\delta, \Theta^{\updownarrow *})$ equals the right-hand side of (8.7).

It remains to prove optimality of $\Theta^{\updownarrow *}$. Due to Lemma 8.7, it is enough to prove that $\Theta^{\updownarrow *}$ is optimal in the class $\mathcal{A}^{\updownarrow c}(x)$, which we do below. In terms of $D^{\updownarrow *}$, the corresponding trading costs are

$$\begin{aligned} J^{\updownarrow}(\delta, \Theta^{\updownarrow *}) &= \int_{(0,T)} D_t^{\updownarrow *} d\Theta_t^{\updownarrow *} + \left(\delta + \frac{K_0}{2} \Delta\Theta_0^{\updownarrow *} \right) \Delta\Theta_0^{\updownarrow *} + \left(D_T^{\updownarrow *} + \frac{K_T}{2} \Delta\Theta_T^{\updownarrow *} \right) \Delta\Theta_T^{\updownarrow *} \\ &= \int_{(0,T)} \frac{D_t^{\updownarrow *}}{K_t} dD_t^{\updownarrow *} + \int_{(0,T)} \frac{\rho_t \left(D_t^{\updownarrow *}\right)^2}{K_t} dt + \frac{\left(D_{0+}^{\updownarrow *}\right)^2 - \delta^2}{2K_0} + \frac{\left(D_{T+}^{\updownarrow *}\right)^2 - \left(D_T^{\updownarrow *}\right)^2}{2K_T}. \end{aligned}$$

Let us now look at alternative strategies $\hat{\Theta} \in \mathcal{A}_0^{\ddagger c}(x)$ with corresponding $\hat{D} = D^{\ddagger *} + h$ and show that these alternative strategies cause higher trading costs than $\Theta^{\ddagger *}$. That is, in the following, we work with functions $h : [0, T+] \to \mathbb{R}$, which are of bounded variation and continuous on $(0, T)$ with $h_0 = 0$, $h_T = \lim_{t \nearrow T} h_t$ and a finite limit h_{0+} (so that there are possibly jumps $(h_{0+} - h_0), (h_{T+} - h_T) \in \mathbb{R}$). Using

$$\Delta\hat{\Theta}_0 = \Delta\Theta_0^{\ddagger *} + \frac{h_{0+}}{K_0}, \qquad d\hat{\Theta}_t = d\Theta_t^{\ddagger *} + \frac{dh_t + \rho_t h_t dt}{K_t}, \qquad \Delta\hat{\Theta}_T = \Delta\Theta_T^{\ddagger *} + \frac{h_{T+} - h_T}{K_T},$$

a straightforward calculation yields

$$\begin{aligned} J^{\ddagger}(\delta, \hat{\Theta}) &= \int_{(0,T)} \hat{D}_t d\hat{\Theta}_t + \left(\delta + \frac{K_0}{2}\Delta\hat{\Theta}_0\right)\Delta\hat{\Theta}_0 + \left(\hat{D}_T + \frac{K_T}{2}\Delta\hat{\Theta}_T\right)\Delta\hat{\Theta}_T \\ &= J^{\ddagger}(\delta, \Theta^{\ddagger *}) + \Delta J_1 + \Delta J_2, \\ \Delta J_1 &:= \int_{(0,T)} \frac{2\rho_t D_t^{\ddagger *} h_t}{K_t} dt + \int_{(0,T)} \frac{h_t}{K_t} dD_t^{\ddagger *} + \int_{(0,T)} \frac{D_t^{\ddagger *}}{K_t} dh_t \\ &\quad + \frac{D_{0+}^{\ddagger *} h_{0+}}{K_0} + \frac{D_{T+}^{\ddagger *} h_{T+} - D_T^{\ddagger *} h_T}{K_T}, \\ \Delta J_2 &:= \int_{(0,T)} \frac{\rho_t h_t^2}{K_t} dt + \int_{(0,T)} \frac{h_t}{K_t} dh_t + \frac{h_{0+}^2}{2K_0} + \frac{h_{T+}^2 - h_T^2}{2K_T}. \end{aligned}$$

Notice that we collect all terms containing $D^{\ddagger *}$ in ΔJ_1. We are now going to finish the proof by showing that $\Delta J_1 = 0$ and $\Delta J_2 > 0$ if h does not vanish.

Let us first rewrite ΔJ_1 exploiting the fact that $D_t^{\ddagger *} = \delta^{\ddagger} f_t$, use integration by parts, the definition of f and again integration by parts to get

$$\begin{aligned} \Delta J_1 &= \delta^{\ddagger} \left\{ \int_{(0,T)} \frac{2\rho_t f_t h_t}{K_t} dt + \int_{(0,T)} \frac{h_t}{K_t} df_t + \int_{(0,T)} \frac{f_t}{K_t} dh_t + \frac{f_0 h_{0+}}{K_0} + \frac{h_{T+} - f_T h_T}{K_T} \right\} \\ &= \delta^{\ddagger} \left\{ \int_{(0,T)} \frac{2\rho_t K_t + K_t'}{K_t^2} f_t h_t dt + \frac{h_{T+}}{K_T} \right\} \\ &= \delta^{\ddagger} \left\{ \int_{(0,T)} \frac{\rho_t h_t}{K_t} dt + \frac{h_{T+}}{K_T} + \int_{(0,T)} \frac{K_t'}{K_t} h_t dt \right\} \\ &= \delta^{\ddagger} \left\{ \int_{(0,T)} \frac{\rho_t h_t}{K_t} dt + \int_{(0,T)} \frac{1}{K_t} dh_t + \frac{h_{0+}}{K_0} + \frac{h_{T+} - h_T}{K_T} \right\}. \end{aligned}$$

Clearly, $\Delta J_1 = 0$ whenever $\delta^{\ddagger} = 0$. If $\delta^{\ddagger} \neq 0$, we have

$$\begin{aligned} x &= \int_{(0,T)} d\hat{\Theta}_t + \Delta\hat{\Theta}_0 + \Delta\hat{\Theta}_T \\ &= \left(\int_{(0,T)} d\Theta_t^{\ddagger *} + \Delta\Theta_0^{\ddagger *} + \Delta\Theta_T^{\ddagger *} \right) + \left(\int_{(0,T)} \frac{\rho_t h_t}{K_t} dt + \int_{(0,T)} \frac{1}{K_t} dh_t + \frac{h_{0+}}{K_0} + \frac{h_{T+} - h_T}{K_T} \right) \\ &= x + \frac{\Delta J_1}{\delta^{\ddagger}}. \end{aligned}$$

Therefore, $\Delta J_1 = 0$. Hence, $J^\updownarrow(\delta, \hat{\Theta}) - J^\updownarrow(\delta, \Theta^{\updownarrow *}) = \Delta J_2$. Applying integration by parts to the dh_t integral yields

$$\Delta J_2 = \int_{(0,T)} \frac{h_t^2}{2K_t}\left(2\rho_t + \frac{K_t'}{K_t}\right) dt + \frac{h_{T+}^2}{2K_T}.$$

Due to the assumption $K_t' + 2\rho_t K_t > 0$ on $[0, T]$, we get that ΔJ_2 is positive as desired. □

REMARK 8.9. At this point, it is natural to discuss the connection between our zero spread model and the zero spread model of Alfonsi et al. (2011) and Gatheral et al. (2012) (the former paper deals with discrete time, the latter one with continuous time). In that model, the price impact at time $t > s$ of the trade ξ_s at time s equals $\xi_s G(t - s)$, where G is called the *decay kernel*. In what follows we abbreviate this modeling approach by AGSS.

In our zero spread model, the price impact at time $t > s$ of the trade ξ_s at time s equals $\xi_s K_s e^{-\int_s^t \rho_u \, du}$ (when $D_0^\updownarrow = 0$). Here, we excluded permanent impact like in (8.1). If we had constant price impact coefficient $K_t \equiv \kappa$ and constant resilience $\rho_t \equiv \rho$ (Obizhaeva–Wang with zero spread), then our model would be a particular case of AGSS with $G(t - s) = \kappa e^{-\rho(t-s)}$. But since our liquidity parameters are time-varying, our model is not a particular case of AGSS.

To summarize: AGSS and our model can be viewed as generalizations of the Obizhaeva–Wang model in different directions. AGSS study general (not only exponential) decay kernels in a time-homogeneous framework, while we study a time-inhomogeneous framework (with exponential resilience). AGSS focus on optimal execution with general decay kernels and study that decay kernels give us viable models, while we study implications of intraday liquidity patterns on optimal execution and price manipulation.

9. EXAMPLES

Let us now turn to explicit examples of dynamics of the price impact parameter K and the resilience ρ. We can use the formulas derived in the previous section to calculate optimal trading strategies in problem (8.2) in the zero spread model. We also want to investigate optimal strategies in problem (4.3) in the dynamic spread model introduced in Section 2. In (4.3), we considered a general initial time $t \in [0, T]$. Without loss of generality, below, we will consider initial time 0 for both models, e.g., we will mean the function $U(0, \cdot, \cdot)$ when speaking about the value function in the dynamic spread model. Further, in the dynamic spread model, we had a nonnegative initial value δ for the deviation of the best ask price from its unaffected level and considered strategies with the overall goal to buy a nonnegative number of shares x. That is, we will consider $\delta, x \in [0, \infty)$ in this section when speaking about either model. It is clear that strategy (8.5) is optimal also in the dynamic spread model whenever it does not contain selling. Thus, Theorem 8.4, applied with $\delta, x \in [0, \infty)$, provides us with formulas for the value function and optimal strategy also in the dynamic spread model whenever there is no transaction-triggered price manipulation in the zero spread model (see Corollary 8.5) and δ is sufficiently close to 0 (so that $\Delta\Theta_0^{\updownarrow *}$ given by the first formula in (8.5) is still nonnegative). Furthermore, in this case, we get an explicit formula for the barrier function of Definition 5.3.

PROPOSITION 9.1 (Closed-form optimal barrier in the dynamic spread model). *Assume the dynamic spread model of Section 2 and that* $K : [0, T] \to (0, \infty)$ *is twice continuously differentiable and* ρ: $[0, T] \to (0, \infty)$ *is continuously differentiable. Let*

$$K_t' + 2\rho_t K_t > 0 \text{ on } [0, T], \quad f_0 \geq 0 \quad \text{and} \quad f_t' + \rho_t f_t \geq 0 \text{ on } [0, T], \tag{9.1}$$

where f *is defined in (8.4). Then, the barrier function of Definition 5.3 is explicitly given by*

$$c(t) = \frac{1}{f_t}\left(\int_t^T \frac{f_s' + \rho_s f_s}{K_s} ds + \frac{1 - f_T}{K_T}\right), \quad t \in [0, T), \quad c(T) = 0. \tag{9.2}$$

Furthermore, for any $x \in [0, \infty)$ *and* $\delta \in [0, \frac{x}{c(0)}]$, *there is a unique optimal strategy in the problem* $U(0, \delta, x)$ *(see (4.3)) and it is given by formula (8.5) in Theorem 8.4, and the value function* $U(0, \delta, x)$ *equals the right-hand side of (8.7).*

REMARK 9.2. (Comments to (9.2))

(i) First, let us note that (9.1) implies $f_t \geq 0$ on $[0, T]$ (see Lemma 9.3 below). Hence, the right-hand side of (9.2) has the form a/b with $a > 0$ (note that $f_T < 1$) and $b \geq 0$, i.e., $c(t) \in (0, \infty]$ for $t \in [0, T)$. The case $c(t) = \infty$ can occur (see, e.g., Example 9.6 with $\nu = -1$).

(ii) Let us further observe that

$$\lim_{t \nearrow T} c(t) = \frac{1 - f_T}{f_T K_T} \in (0, \infty],$$

i.e., the barrier always jumps at T.

Proof of Proposition 9.1. Let us first notice that c from (8.6) is strictly positive by Lemma 8.8 so that Theorem 8.4 applies. Further, it follows from (9.1) that in the zero spread model with such functions K and ρ, there is no transaction-triggered price manipulation. Hence, for any $x > 0$, the optimal strategy $\Theta^{\updownarrow *}$ from (8.5) with $\delta = 0$ in the problem $U^{\updownarrow}(0, x)$ will also be optimal in the problem $U(0, 0, x)$. Let us recall that the value $c(0)$ of the barrier is the ratio $\frac{x - \Delta\Theta_0^{\updownarrow *}}{D_{0+}^{\updownarrow *}}$ for the optimal strategy $\Theta^{\updownarrow *}$ in the problem $U(0, 0, x)$ and the corresponding $D^{\updownarrow *}$ (with $D_0^{\updownarrow *} = 0$). Thus, we get

$$c(0) = \frac{x - \Delta\Theta_0^{\updownarrow *}}{K_0 \Delta\Theta_0^{\updownarrow *}} = \frac{1}{f_0}\left(\int_0^T \frac{f_s' + \rho_s f_s}{K_s} ds + \frac{1 - f_T}{K_T}\right).$$

A similar reasoning applies to an arbitrary $t \in [0, T)$. Recall that we always have $c(T) = 0$. Finally, for $\delta > 0$, under condition (9.1), formula (8.5) for the zero spread model will give the optimal strategy in the problem $U(0, \delta, x)$ (i.e., for the dynamic spread model) if and only if $\Delta\Theta_0^{\updownarrow *} \geq 0$. Solving this inequality with respect to δ, we get $\delta \leq \frac{x}{c(0)}$ (note that $\delta^{\updownarrow}$ from (8.5) also depends on δ). □

Condition (9.1) ensures the applicability of Theorem 8.4 and additionally excludes transaction-triggered price manipulation in the zero spread model (see Corollary 8.5). The following result provides an equivalent form for this condition, which we will use below when studying specific examples.

LEMMA 9.3 (An equivalent form for condition (9.1)). *Assume that $K : [0, T] \to (0, \infty)$ is twice continuously differentiable and $\rho : [0, T] \to (0, \infty)$ is continuously differentiable. Then, condition (9.1) is equivalent to*

$$K_t' + \rho_t K_t \geq 0 \text{ on } [0, T] \quad \text{and} \quad f_t' + \rho_t f_t \geq 0 \text{ on } [0, T]. \tag{9.3}$$

Proof. Clearly, (9.3) implies (9.1). Let us prove the converse. Suppose (9.1) is satisfied and $K_s' + \rho_s K_s < 0$ for some $s \in [0, T]$. Then, there exists $[u, v] \subset [0, T]$ such that $u < v$, $f_u = 0$ and $f_t < 0$ on $(u, v]$. By the mean value theorem, there exists $w \in (u, v)$ such that $f_w' = (f_v - f_u)/(v - u)$. Thus, we get $f_w' < 0$ and $f_w < 0$, which contradicts the condition $f_t' + \rho_t f_t \geq 0$ on $[0, T]$. □

When we have transaction-triggered price manipulation in the zero spread model, optimal strategies in the dynamic spread model are different from the ones given in Theorem 8.4. The following proposition deals with the case of $K_t' + \rho_t K_t < 0$ for some t (cf. with (9.3)).

PROPOSITION 9.4 (Wait if decrease of K outweighs resilience). *Assume the dynamic spread model of Section 2. Let, for some $t \in [0, T)$, K be continuously differentiable at t and ρ continuous at t with $K_t' + \rho_t K_t < 0$. Then, $Br_t = \emptyset$, i.e., $c(t) = \infty$.*

Proof. Since $K' + \rho K$ is continuous at t, we have $K_s' + \rho_s K_s < 0$ on an interval around t. Then, there exists $\epsilon > 0$ such that $K_s e^{-\int_s^{t+\epsilon} \rho_u du} > K_{t+\epsilon}$ for all $s \in [t, t+\epsilon)$. By Proposition 5.8, it is not optimal to trade at t. □

Let us finally illustrate our results by discussing several examples. For simplicity, take constant resilience $\rho > 0$. Then, condition (9.3) takes the form

$$K_t' + \rho K_t \geq 0 \text{ on } [0, T] \quad \text{and} \quad K_t'' + 3\rho K_t' + 2\rho^2 K_t \geq 0 \text{ on } [0, T]. \tag{9.4}$$

A sufficient condition for (9.4), which is sometimes convenient (e.g., in Example 9.6 below), is

$$K_t' + \rho K_t \geq 0 \text{ on } [0, T] \quad \text{and} \quad K_t'' + \rho K_t' \geq 0 \text{ on } [0, T].$$

In all examples below, we consider $\delta = 0$ and $x \in [0, \infty)$.

EXAMPLE 9.5. (Constant price impact $K_t \equiv \kappa > 0$). Assume that the price impact $K_t \equiv \kappa > 0$ is constant. Clearly, condition (9.4) is satisfied, so we can use formula (8.5) to get the optimal strategy in both models. We have $f_t \equiv \frac{1}{2}$ and $\delta^{\ddagger} = \frac{2\kappa x}{\rho T+2}$. The optimal strategy in both the dynamic and zero spread models is given by the formula

$$\Delta\Theta_0 = \Delta\Theta_T = \frac{x}{\rho T + 2}, \quad d\Theta_t = \frac{x\rho}{\rho T + 2} dt,$$

which recovers the results from Obizhaeva and Wang (2006). The large investor trades with constant speed on $(0, T)$ and consumes all fresh limit sell orders entering the book due to resilience in such a way that the corresponding deviation process D_t is constant on $(0, T]$ (see (8.12) and note that f_t is constant). The barrier is linearly decreasing in time (see (9.2)):

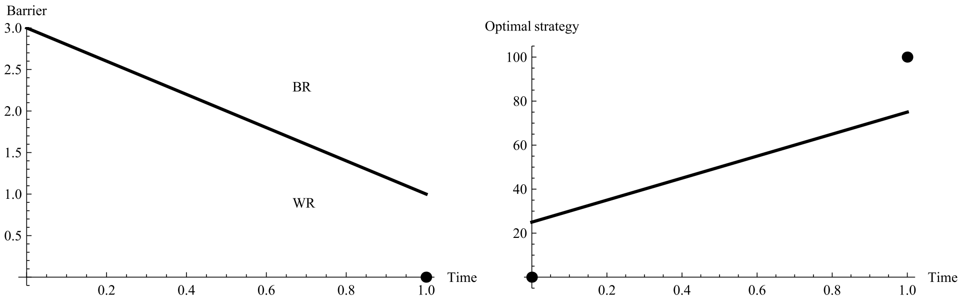

FIGURE 9.1. Constant price impact ($T = 1$, $\rho = 2$, $\kappa = 1$, $x = 100$, $\delta = 0$).

$$c(t) = \frac{1 + \rho(T - t)}{\kappa}, \quad t \in [0, T), \quad c(T) = 0.$$

Let us finally note that the optimal strategy does not depend on κ, while the barrier depends on κ. See Figure 9.1 for an illustration.

EXAMPLE 9.6 (Exponential price impact $K_t = \kappa e^{\nu\rho t}$, $\kappa > 0$, $\nu \in \mathbb{R} \setminus \{0\}$). Assume that the price impact $K_t = \kappa e^{\nu\rho t}$ is growing or falling exponentially with $\nu \in \mathbb{R} \setminus \{0\}$ being the slope of the exponential price impact relative to the resilience. The case $\nu = 0$ was studied in the previous example. We exclude this case here because some expressions below will take the form $0/0$ when $\nu = 0$ (however, the limits of these expressions as $\nu \to 0$ will recover the corresponding formulas from the previous example). Condition (9.4) is satisfied if and only if $\nu \geq -1$. We first consider the case $\nu \geq -1$. We have

$$f_t \equiv \frac{\nu + 1}{\nu + 2} \quad \text{and} \quad \delta^{\updownarrow} = \frac{x\kappa\nu(\nu + 2)}{(\nu + 1)^2 - e^{-\nu\rho T}}.$$

In particular, like in the previous example, the large investor trades in such a way that the deviation process D_t is constant on $(0, T]$. The optimal strategy in both the dynamic and zero spread models is given by the formula

$$\Delta\Theta_0 = \frac{x\nu(\nu + 1)}{(\nu + 1)^2 - e^{-\nu\rho T}}, \quad d\Theta_t = \frac{x\nu(\nu + 1)}{(\nu + 1)^2 - e^{-\nu\rho T}} \rho e^{-\nu\rho t}\, dt,$$
$$\Delta\Theta_T = \frac{x\nu}{(\nu + 1)^2 - e^{-\nu\rho T}} e^{-\nu\rho T}.$$

We see that, for $\nu = -1$, it is optimal to buy the entire order at T. Vice versa, the initial trade $\Delta\Theta_0$ approaches x as $\nu \nearrow \infty$. The barrier is given by the formula

$$c(t) = \frac{(\nu + 1)e^{-\nu\rho t} - e^{-\nu\rho T}}{\kappa\nu(\nu + 1)}, \quad t \in [0, T), \quad c(T) = 0$$

(in particular, $c(t) = \infty$ for $t \in [0, T)$ if $\nu = -1$ and the barrier is finite everywhere if $\nu > -1$). For each $\nu > -1$, the barrier is decreasing in t, i.e., buying becomes more aggressive as the investor runs out of time. Furthermore, one can check that for each $t \in [0, T)$, the barrier is decreasing in ν. That is, the greater is ν, the larger is the buy region since it is less attractive to wait. Like in the previous example, the optimal strategy does not depend on κ, while the barrier depends on κ.

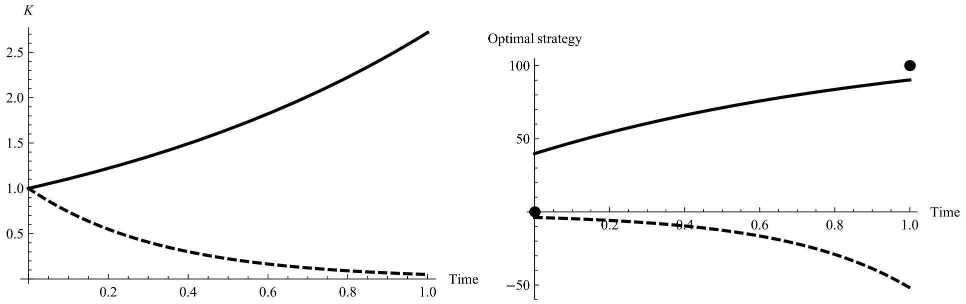

FIGURE 9.2. Exponential price impact ($T = 1$, $\rho = 2$, $\kappa = 1$, $x = 100$, $\delta = 0$, $\nu = 0.5$, and -1.5 (dashed)).

Let us now consider the case $\nu < -1$. In the zero spread model, transaction-triggered price manipulation occurs for $\nu \in (-2, -1)$ (one checks that the assumptions of Theorem 8.4 are satisfied) and classical price manipulation occurs for $\nu < -2$ (see Proposition 8.3). In the dynamic spread model, for $\nu < -1$, it is optimal to trade the entire order at T because $K_t e^{-\rho(T-t)} > K_T$ for all $t \in [0, T)$ (see Proposition 5.8). Thus, in the case $\nu < -1$, we have $c(t) = \infty$ for $t \in [0, T)$.

See Figure 9.2 for an illustration.

EXAMPLE 9.7 (Straight-line price impact $K_t = \kappa + mt$, $\kappa > 0$, $m > -\frac{\kappa}{T}$). Assume that the price impact $K_t = \kappa + mt$ changes linearly over time. The condition $m > -\frac{\kappa}{T}$ ensures that K is everywhere strictly positive. Condition (9.4) is satisfied if and only if $m \geq -\frac{2\rho\kappa}{3+2\rho T}$. Note that $-\frac{2\rho\kappa}{3+2\rho T} > -\frac{\kappa}{T}$. Let us first assume that $m \geq -\frac{2\rho\kappa}{3+2\rho T}$. In this case, the optimal strategy in both the dynamic and zero spread models is given by the formulas

$$\Delta\Theta_0 = \frac{2m(m+\kappa\rho)x}{(m+2\kappa\rho)\tilde{m}}, \qquad d\Theta_t = \frac{2m\kappa\rho^2\,(2\kappa\rho + m(3+2\rho t))\,x}{(m+2\kappa\rho+2m\rho t)^2\,\tilde{m}}dt,$$

$$\Delta\Theta_T = \frac{2m\kappa\rho x}{(m+2\kappa\rho+2m\rho T)\tilde{m}} \qquad \text{with } \tilde{m} := 2m + \kappa\rho\log\left(\frac{m+2\kappa\rho+2m\rho T}{m+2\kappa\rho}\right).$$

The barrier is given by the formula

$$c(t) = \rho\frac{2m - (m+2\kappa\rho+2m\rho t)\log\left(\dfrac{m+2\kappa\rho+2m\rho t}{m+2\kappa\rho+2m\rho T}\right)}{2m(m+\kappa\rho+m\rho t)}.$$

In the zero spread model, transaction-triggered price manipulation occurs for $m \in (-\frac{2\rho\kappa}{1+2\rho T}, -\frac{2\rho\kappa}{3+2\rho T})$ (see Theorem 8.4) and classical price manipulation occurs for $m \in (-\frac{\kappa}{T}, -\frac{2\rho\kappa}{1+2\rho T})$ (see Proposition 8.3). In the dynamic spread model, we can check by Proposition 5.8 that it is optimal to trade the entire order at T for

$$m \in \left(-\frac{\kappa}{T}, -\frac{\kappa}{T}(1 - e^{-\rho T})\right)$$

(see Lemma B.1). We observe that $-\frac{\kappa}{T}(1 - e^{-\rho T}) < -\frac{2\rho\kappa}{3+2\rho T}$ (see Lemma B.2). Let us finally note that the presented methods do not allow us to calculate the optimal strategy in closed form in the dynamic spread model for $m \in [-\frac{\kappa}{T}(1 - e^{-\rho T}), -\frac{2\rho\kappa}{3+2\rho T})$, but we

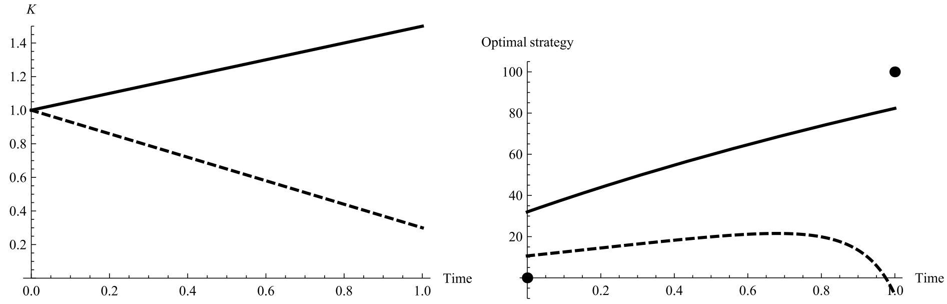

FIGURE 9.3. Straight-line price impact ($T = 1$, $\rho = 2$, $\kappa = 1$, $x = 100$, $\delta = 0$, $m = 0.5$, and -0.7 (dashed)).

can approximate it numerically in discrete time (see, e.g., the case with $K_t = 1 - 0.6\,t$, $\rho = 2$, and $T = 1$ in Figure 6.1).

See Figure 9.3 for an illustration.

10. CONCLUSION

Time-varying liquidity is a fundamental property of financial markets. Its implication for optimal liquidation in limit order book markets is the focus of this paper. We find that a model with a dynamic, trading influenced spread is very robust and free of two types of price manipulation. We prove that value functions and optimal liquidation strategies in this model are of wait-region/buy-region type, which is often encountered in problems of singular control. In the literature on optimal trade execution in limit order books, the spread is often assumed to be zero. Under this assumption, we show that time-varying liquidity can lead to classical as well as transaction-triggered price manipulation. For both dynamic and zero spread assumptions, we derive closed-form solutions for optimal strategies and provide several examples.

APPENDIX A: INTEGRATION BY PARTS FOR CÀGLÀD PROCESSES

In various proofs in this paper, we need to apply stochastic analysis (e.g., integration by parts or Ito's formula) to càglàd processes of finite variation and/or standard semimartingales. As noted in Section 2, this is always done as follows: if U is a càglàd process of finite variation, we first consider the process U^+ defined by $U_t^+ := U_{t+}$ and then apply standard formulas from stochastic analysis to it. As an example of such a procedure, we provide the following lemma, which is often applied in the proofs in this paper.

LEMMA A.1 (Integration by parts). *Let $U = (U_t)_{t\in[0,T+]}$ and $V = (V_t)_{t\in[0,T+]}$ be càglàd processes of finite variation and Z a semimartingale (in particular càdlàg), which may have a jump at* 0. *For $t \in [0, T]$, we have*

$$U_{t+}Z_t = U_0 Z_{0-} + \int_{[0,t]} U_s dZ_s + \int_{[0,t]} Z_s dU_s, \tag{A.1}$$

$$U_{t+}V_{t+} = U_0 V_0 + \int_{[0,t]} U_s dV_s + \int_{[0,t]} V_{s+} dU_s. \tag{A.2}$$

Proof. Let X and Y be càdlàg processes (possibly having a jump at 0) with X being a semimartingale and Y a finite variation process. By proposition I.4.49(a) in Jacod and Shiryaev (2003), which is a variant of integration by parts for the case where one of the semimartingales is of finite variation,

$$X_t Y_t = X_{0-} Y_{0-} + \int_{[0,t]} Y_{s-} d X_s + \int_{[0,t]} X_s d Y_s, \quad t \in [0, T]. \tag{A.3}$$

Equation (A.1) is a particular case of (A.3) applied to $X := Z$, $Y := U^+$ and equation (A.2) is a particular case of (A.3) applied to $X := V^+, Y := U^+$, where $U_t^+ := U_{t+}$ and $V_t^+ := V_{t+}$. □

APPENDIX B: TECHNICAL LEMMAS USED IN EXAMPLE 9.7

Below, we use the notation of Example 9.7.

LEMMA B.1. *For $m \in (-\frac{\kappa}{T}, -\frac{\kappa}{T}(1 - e^{-\rho T}))$, we have*

$$(\kappa + mt)e^{-\rho(T-t)} > \kappa + mT, \quad t \in [0, T), \tag{B.1}$$

i.e., Proposition 5.8 applies.

Proof. Inequality (B.1) is equivalent to

$$m < -\frac{\kappa(1 - e^{-\rho(T-t)})}{T - te^{-\rho(T-t)}}.$$

The assertion now follows from our assumption on m. To see this, we need to show that

$$-\frac{\kappa}{T}(1 - e^{-\rho T}) \le -\frac{\kappa(1 - e^{-\rho(T-t)})}{T - te^{-\rho(T-t)}},$$

which, in turn, is equivalent to

$$g(t) := \frac{1 - e^{-\rho(T-t)}}{T - te^{-\rho(T-t)}} \le \frac{1 - e^{-\rho T}}{T} = g(0).$$

This is a true statement since $1 - x \le e^{-x}$ for all $x \in \mathbb{R}$ and therefore

$$g'(t) = \frac{1 - \rho(T - t) - e^{-\rho(T-t)}}{e^{\rho(T-t)}(T - te^{-\rho(T-t)})^2} \le 0.$$

□

LEMMA B.2. *We have $-\frac{\kappa}{T}(1 - e^{-\rho T}) < -\frac{2\rho\kappa}{3+2\rho T}$.*

Proof. The statement reduces to proving that $\frac{2\rho T}{3+2\rho T} < 1 - e^{-\rho T}$. Setting $x := \rho T > 0$, we see that it is enough to establish that $e^{-x} < \frac{3}{3+2x}$, which is true as, clearly, $e^x > 1 + \frac{2}{3}x$. □

REFERENCES

ALFONSI, A., A. FRUTH, and A. SCHIED (2010): Optimal Execution Strategies in Limit Order Books with General Shape Functions, *Quant. Finance* 10, 143–157.

ALFONSI, A., A. SCHIED, and A. SLYNKO (2011): Order Book Resilience, Price Manipulation, and the Positive Portfolio Problem, Preprint.

ALMGREN, R. F. (2003): Optimal Execution with Nonlinear Impact Functions and Trading-Enhanced Risk, *Appl. Math. Finance* 10, 1–18.

ALMGREN, R. F. (2009): Optimal Trading in a Dynamic Market, Preprint.

ALMGREN, R. F., and N. CHRISS (2001): Optimal Execution of Portfolio Transactions, *J. Risk* 3, 5–40.

BANK, P., and D. BECHERER (2009): Talk: Optimal Portfolio Liquidation with Resilient Asset Prices, *Liquidity-Modelling Conference*, Oxford.

BERTSIMAS, D., and A. LO (1998): Optimal Control of Execution Costs, *J. Financ. Markets* 1, 1–50.

BOUCHAUD, J. P., Y. GEFEN, M. POTTERS, and M. WYART (2004): Fluctuations and Response in Financial Markets: The Subtle Nature of “Random” Price Changes, *Quant. Finance* 4, 176–190.

CHORDIA, T., R. ROLL, and A. SUBRAHMANYAM (2001): Market Liquidity and Trading Activity, *J. Finance* 56, 501–530.

CONT, R., S. STOIKOV, and R. TALREJA (2010): A Stochastic Model for Order Book Dynamics, *Oper. Res.* 58, 217–224.

EASLEY, D., and M. O’HARA (1987): Price, Trade Size, and Information in Securities Markets, *J. Financ. Econ.* 19, 69–90.

ESSER, A., and B. MÖNCH (2003): Modeling Feedback Effects with Stochastic Liquidity, Preprint.

FRUTH, A. (2011): *Optimal Order Execution with Stochastic Liquidity*, PhD thesis, TU Berlin.

GATHERAL, J. (2010): No-Dynamic-Arbitrage and Market Impact, *Quant. Finance* 10, 749–759.

GATHERAL, J., A. SCHIED, and A. SLYNKO (2011): Exponential Resilience and Decay of Market Impact, in *Econophysics of Order Driven Markets*, F. ABERGEL, B. K. CHAKRABARTI, A. CHAKRABORTI, and M. MITRA, eds. Milan: Springer-Verlag, pp. 225–236.

GATHERAL, J., A. SCHIED, and A. SLYNKO (2012): Transient Linear Price Impact and Fredholm Integral Equations, *Math. Finance* 22, 445–474.

GREGORY, J., and C. LIN (1996): *Constrained Optimization in the Calculus of Variations and Optimal Control Theory*, London: Springer.

HUBERMAN, G., and W. STANZL (2004): Price Manipulation and Quasi-Arbitrage, *Econometrica* 72, 1247–1275.

JACOD, J., and A. SHIRYAEV (2003): *Limit Theorems for Stochastic Processes*, 2nd ed., Berlin: Springer.

KEMPF, A., and D. MAYSTON (2008): Commonalities in the Liquidity of a Limit Order Book, *J. Financ. Res.* 31, 25–40.

KIM, S. J., and S. BOYD (2008): Optimal Execution under Time Inhomogeneous Price Impact and Volatility, Preprint.

KYLE, A. S. (1985): Continuous Auctions and Insider Trading, *Econometrica* 53, 1315–1335.

KYLE, A. S., and S. VISWANATHAN (2008): How to Define Illegal Price Manipulation, *Am. Econ. Rev.* 98, 274–279.

LARGE, J. (2007): Measuring the Resiliency of an Electronic Limit Order Book, *J. Financ. Markets* 10, 1–25.

LORENZ, J., and J. OSTERRIEDER (2009): Simulation of a Limit Order Driven Market, *J. Trading* 4, 23–30.

MADAN, D. B., and W. SCHOUTENS (2011): Tenor Specific Pricing, Preprint.

NAUJOKAT, F., and N. WESTRAY (2011): Curve Following in Illiquid Markets, *Math. Financ. Econ.* 4, 1–37.

OBIZHAEVA, A., and J. WANG (2006): Optimal Trading Strategy and Supply/Demand Dynamics, Preprint.

PREDOIU, S., G. SHAIKHET, and S. E. SHREVE (2011): Optimal Execution in a General One-Sided Limit-Order Book, *SIAM J. Financ. Math.* 2, 183–212.

SCHÖNEBORN, T. (2008): *Trade Execution in Illiquid Markets*, PhD thesis, TU Berlin.

SHIRYAEV, A. (1995): *Probability*, 2nd ed., New York: Springer.

STEINMANN, G. (2005): *Order Book Dynamics and Stochastic Liquidity in Risk-Management*, Master's thesis, ETH Zurich and University of Zurich.

WEISS, A. (2010): Executing Large Orders in a Microscopic Market Model, Preprint.

MATHEMATICAL FINANCE

An International Journal of Mathematics, Statistics and Financial Economics

Mathematical Finance, Vol. 24, No. 4 (October 2014), 696–727

OPTIMAL LIQUIDATION IN A LIMIT ORDER BOOK FOR A RISK-AVERSE INVESTOR

ARNE LØKKA

London School of Economics

In a limit order book model with exponential resilience, general shape function, and an unaffected stock price following the Bachelier model, we consider the problem of optimal liquidation for an investor with constant absolute risk aversion. We show that the problem can be reduced to a two-dimensional deterministic problem which involves no buy orders. We derive an explicit expression for the value function and the optimal liquidation strategy. The analysis is complicated by the fact that the intervention boundary, which determines the optimal liquidation strategy, is discontinuous if there are levels in the limit order book with relatively little market depth. Despite this complication, the equation for the intervention boundary is fairly simple. We show that the optimal liquidation strategy possesses the natural properties one would expect, and provide an explicit example for the case where the limit order book has a constant shape function.

KEY WORDS: limit order book, optimal liquidation, optimal execution, CARA utility, singular control, discontinuous intervention boundary.

1. INTRODUCTION

The growing popularity of algorithmic execution has resulted in an increasing interest in asset price models incorporating illiquidity and optimal execution of large orders. A brief history of the growth of algorithmic execution can be found in Aldridge (2010), which also provides an overview of current market practice and common models. These models tend to be based on the investor's trading rate, like the Almgren and Chriss (1999) model, or to be variations of the limit order book model of Obizhaeva and Wang (2005).

Bertsimas and Lo (1998) introduced a class of discrete time models for asset prices incorporating illiquidity effects, and used dynamic programming to derive the strategy which minimizes the expected cost of trading. Almgren and Chriss (1999, 2001) addressed the problem of maximizing the expected revenue of trading, taking into account its variance. Almgren (2003) generalized the model of Almgren and Chriss (1999, 2001) to incorporate nonlinear impact functions. Gatheral (2010) further generalized the Almgren model (2003) to incorporate a decay kernel and provide conditions for existence and absence of price manipulation strategies. Based on the model of Almgren (2003) with a linear impact factor, Schied and Schöneborn (2009) study the problem of optimal liquidation of a large stock position for an investor aiming to maximize the expected utility of their cash position at the end of time, and compare the qualitative behavior of the optimal strategy for different utility functions in case of increasing and decreasing

The author would like to thank an anonymous referee for helpful comments and suggestions.

Manuscript received August 2011; final revision received December 2012.

Address correspondence to Arne Løkka, Department of Mathematics, Columbia House, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom; e-mail: a.lokka@lse.ac.uk.

DOI: 10.1111/mafi.12033

prices. They relate their results to the strategies corresponding to the various criteria discussed in Kissell and Malamut (2005), which involves different benchmark price targets. The models listed above have in common that the underlying price process is based on the Bachelier model, with an additional impact depending on the transactions made by a large trader who wants to purchase or sell a large number of shares. There is a permanent effect depending on the size of the trade, and a temporary effect on the stock price depending on the time derivative of the large trader's position in the stock. The problem of finding the optimal liquidation or purchase strategy typically takes the form as the solution to an Euler–Lagrange equation.

Obizhaeva and Wang (2005) introduced a limit order book model, which specifies the stochastic dynamics of limit orders and the best offered bid and ask price. The cost of purchasing or selling shares thus depends on the limit orders and the behavior of the best offered bid and ask price as orders eat into the limit order book, as well as the resilience of the limit order book, which specifies how the best offered bid and ask prices recover after orders have been executed. Alfonsi, Fruth, and Schied (2010) generalize the limit order book model of Obizhaeva and Wang to allow more general shape functions, and provide two slightly different ways to model resilience. Predoiu, Shaikhet, and Shreve (2011) introduce a one-sided limit order book model where the shape of the limit orders are given in terms of a measure. Alfonsi and Schied (2010) and Alfonsi, Schied, and Slynko (2012) provide conditions for the absence of price manipulation strategies in limit order books and models related to that of Gatheral (2010). Gatheral, Schied, and Slynko (2011) introduce a limit order book with a general resilience function and provide a comparison between such models and that of Gatheral (2010). Schied, Schöneborn, and Tehranchi (2010) provide a connection between the maximization criterion of Almgren and Chriss (1999, 2001) and maximization of expected constant absolute risk aversion (CARA) utility. They also show that if the risk of the stock price is given in terms of a Lévy process and if there is a certain additive structure between the risk and the trading cost, then the optimal strategy for an investor with CARA utility is deterministic.

In this paper, we adopt a limit order book model with a general shape function and exponential resilience rate as in Alfonsi et al. (2010), where the stochastic dynamics of the nonaffected stock price follows the Bachelier model, and consider a large investor with constant absolute risk aversion who wants to liquidate his share position without time constraints. This model for the dynamics of the risky asset can be viewed as the limit order book equivalent to the models of Almgren and Chriss (1999), Almgren (2003), and Gatheral (2010). We do not explicitly model the ask orders and best ask price, but assume that the best bid price and bid orders are unaffected by the investor's buy orders and that the unaffected bid price process provides a lower bound for the best offered ask price. The utility maximization problem that we consider can be viewed as a limit order book version of the problem considered by Schied and Schöneborn (2009), provided that the investor has constant absolute risk aversion. Based on the ideas in Schied et al. (2010), we prove that the optimal strategy is deterministic and involves no buy orders. Moreover, the maximum expected utility is strictly negative and can be expressed in terms of the value function of a certain two-dimensional optimal control problem.

Under fairly mild condition on the shape function of the limit order book, we derive an explicit solution to the optimal liquidation problem for a general shape function. The optimal strategy turns out to differ from the corresponding optimal strategy in the Almgren (2003) model. Depending on the investor's past trading history, the optimal liquidation strategy involves an initial block trade and then continuously sell shares, or first wait for a certain amount of time to let the limit order book recover, and then continuously sell shares. If there are levels in the limit order book with relatively little market depth, there are periods where it is optimal to wait in order for the best bid price to recover from the level with relatively little depth. To the best of the author's knowledge, this Markovian dependence on the investor's past trading history and corresponding state of the limit order book has not previously been explicitly explored, and in the Almgren model (2004) there is no such dependence since the returns in this case do not depend on the investor's past trading history. Also, the optimal liquidation strategy in models based on the trading rate, like the Almgren (2003) model, does not involve any block trades. It is therefore interesting to note that the optimal liquidation strategy for the large investor in a limit order book model typically consists of an initial block trade.

In order to derive the optimal liquidation strategy for the large investor, we derive an explicit solution to the Hamilton–Jacobi–Bellman equation corresponding to the value function of the reduced deterministic optimization problem. The Hamilton–Jacobi–Bellman equation takes the form of a free boundary problem, where it turns out that the boundary can be discontinuous. The discontinuities happen when there are levels in the limit order book with relatively little market depth. While the discontinuities of the intervention boundary complicates the analysis and proofs, the equation for the boundary takes a fairly simple form. The optimal strategy then essentially consists of trading in such a way that the state process, which consists of the number of shares held and the current state of the limit order book, remains on the boundary at all times. We also show that the optimal strategy possesses the natural properties one would expect, e.g., that an increased depth implies faster liquidation, increased volatility of the unaffected stock price implies faster liquidation, and an increased risk aversion implies faster liquidation.

At the end, we provide an example of a limit order book with a constant shape function, and compare the optimal liquidation strategy to that provided in Schied and Schöneborn (2009) for the Almgren model with a linear impact function.

2. MODEL SPECIFICATION AND PROBLEM FORMULATION

We adopt the limit order book model of Alfonsi et al. (2010) with an unaffected best bid price process following the Bachelier model. However, instead of modeling the full limit order book, we explicitly model the bid order book and assume that the unaffected bid price provides a lower bound for the best ask price and that the bid prices are unaffected by the large investor's buy orders. These assumptions are satisfied in the full limit order book model. We show that under these assumptions, the optimal liquidation strategy does not involve any buy orders.

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ be a complete filtered probability space satisfying the usual conditions and supporting a one-dimensional Brownian motion W, and set

$$\mathbb{R}^+ = [0, \infty) \quad \text{and} \quad \mathbb{R}^- = (-\infty, 0].$$

We consider the following set of admissible liquidation strategies for the large investor.

DEFINITION 2.1. For $y \in \mathbb{R}^+$, let $\mathcal{A}(y)$ denote the set of all pairs (X, Y), where X and Y are $(\mathcal{F}_t)$-adapted, càdlàg processes, X is nondecreasing and Y is nonincreasing, $X_{0-} = 0$ and $Y_{0-} = y$, and

$$\int_0^\infty \| X_t + Y_t \|^2_{L^\infty(\Omega)} \, dt < \infty. \tag{2.1}$$

The nonnegative quantity X_t represents the number of shares bought over the time interval $[0, t]$, $Y_t - y$ represents the number of shares sold over the time interval $[0, t]$, and $X_t + Y_t$ is the net position in shares held at time t. Denote by $\mathcal{A}_D^-(y)$ the set of all deterministic strategies $(X, Y) \in \mathcal{A}(y)$ with $X = 0$. Thus, we can identify $\mathcal{A}_D^-(y)$ with the set of deterministic càdlàg, nonincreasing processes with values in $[0, y]$ satisfying $Y_{0-} = y$ and

$$\int_0^\infty |Y_t|^2 \, dt < \infty. \tag{2.2}$$

The unaffected bid process B^0 is assumed to follow the Bachelier model, that is

$$B_t^0 = b + \sigma W_t, \qquad t \geq 0,$$

where $b > 0$ is the bid price at time 0 and $\sigma > 0$ is the volatility. The interpretation of the unaffected bid process B^0 is that if the large investor makes no trades, then the best offered bid price at time t is B_t^0. The Bachelier model may seem simplistic, but the Bachelier model is widely used in the optimal liquidation literature (see, e.g., Almgren and Criss 1999; Kissell and Malamut 2005; Schied and Schöneborn 2009; and Gatheral 2010). We do not explicitly model the ask prices, but make the following assumption:

ASSUMPTION 2.2. The bid prices are unaffected by the large investor's buy orders and the best unaffected bid price provides a lower bound for the best ask price.

In particular, we note that the full limit order book in Obizhaeva and Wang (2005), Alfonsi et al. (2010), and Gatheral et al. (2011), all satisfy this assumption.

There are two other components which together specify the limit order book model. One is the shape function, which we denote by ϕ, and the other is the resilience rate, which describes how the market recovers. The shape function ϕ is static, and the connection between the shape function ϕ and the bid prices in the limit order book is that, at time t, the number of bids at price $B_t + x$ is equal to $\phi(x)\, dx$, where $x \leq 0$, provided the investor has not made any large trades before time t. We impose the following condition on the shape function:

ASSUMPTION 2.3. The shape function $\phi : \mathbb{R}^- \to (0, \phi_{\max}]$ is continuous.

Let B_t^Y denote the best bid price offered at time t if the large investor follows a liquidation strategy $(X, Y) \in \mathcal{A}(y)$. As the notation suggests, the best offered bid price depends on the past history of the strategy Y, but according to Assumption 2.2, does not depend on the buy strategy X. We assume that B^Y is a càdlàg process. Denote by D^Y the spread process given by

$$D_t^Y = B_t^Y - B_t^0, \qquad t \geq 0,$$

i.e., the spread between the best offered bid price and the unaffected bid price if the large investor adopts a liquidation strategy Y. If the large trader adopts a strategy Y which consists of selling a number $\triangle Y_t = Y_t - Y_{t-}$ of shares at time t, then the effect of this on the best offered bid price is that the new spread changes from D_{t-}^Y to D_t^Y, where D_t^Y satisfies

$$\int_{D_{t-}^Y}^{D_t^Y} \phi(u)\, du = \triangle Y_t.$$

This corresponds to the best bid orders being executed in order to match the large trader's sales order of $-\triangle Y_t$ number of shares. In order to ease notation, introduce the functions $\Phi : \mathbb{R}^- \to \mathbb{R}^-$ and $\psi : \mathbb{R}^- \to \mathbb{R}^-$ by

$$\Phi(x) = \int_0^x \phi(u)\,du \quad \text{and} \quad \psi(z) = \Phi^{-1}(z). \tag{2.3}$$

The inverse of Φ is well defined since Φ is strictly increasing, due to the assumption that ϕ takes strictly positive values. From the assumptions made on the properties of ϕ in Assumption 2.3, it follows that $\psi : \mathbb{R}^- \to \mathbb{R}^-$ is an increasing $C^1(\mathbb{R}^-)$ function satisfying

$$\psi(0) = 0, \tag{2.4}$$

$$\text{there exists } \delta > 0 \text{ such that } \psi'(z) \geq \delta, \text{ for all } z \in \mathbb{R}^-, \tag{2.5}$$

$$\text{there exists } C > 0 \text{ and } \epsilon > 0 \text{ such that } \psi'(z) \leq C, \text{ for all } z \in (-\epsilon, 0]. \tag{2.6}$$

As in Alfonsi et al. (2010) and Obizhaeva and Wang (2005), we assume that the limit order book has an exponential resilience rate, which means that the limit order book recovers at an exponential rate. Introduce the process Z^Y given by

$$Z_t^Y = z\mathrm{e}^{-\lambda t} + \int_0^t \mathrm{e}^{-\lambda(t-s)}\,dY_s, \qquad t \geq 0, \tag{2.7}$$

where $z \leq 0$ is the initial value of Z^Y at time 0 and $\lambda > 0$ is the resilience speed. For future reference, note that Z^Y is the unique càdlàg solution to

$$dZ_t^Y = -\lambda Z_{t-}^Y\,dt + dY_t, \qquad Z_0^Y = z \in \mathbb{R}^-. \tag{2.8}$$

The process Z^Y captures how the large investor's implementation of the liquidation strategy Y affects the best offered bid price and how this recovers over time through the relation

$$Z_t^Y = \Phi\big(D_t^Y\big),$$

where D_t^Y is the spread at time t. The initial state z therefore provides the initial state of the limit order book, which takes into account the past trading history of the large investor. In the literature, the letter E is often used rather than Z to denote the process Z^Y, since it represent the part of the order book which is "eaten up." With reference to the definition of the spread process D^Y and (2.3), the best offered bid price B_t^Y, at time t, is given by

$$B_t^Y = B_t^0 + \psi\big(Z_t^Y\big), \tag{2.9}$$

if the large investor use the liquidation strategy Y.

So far we have described how the large investor's trading affects the best offered bid price, but not how this affects the large investor's cash position. Suppose that the large investor's initial cash position is c and that he implements a strategy $(X, Y) \in \mathcal{A}(y)$ which consists of a number of block sales, i.e., Y is a decreasing step function and X is zero. Then, the large investor's cash position at time $T > 0$ is

$$C_T(X, Y) = c - \sum_{0 \leq t \leq T} \int_0^{\triangle Y_t} \{B_t^0 + \psi(Z_{t-}^Y + x)\}\, dx, \tag{2.10}$$

which corresponds to the best bids offered at all times being executed first so as to match the large trader's sales orders. Let Y^c denote the continuous part of Y, as defined in, e.g., Protter (1990). If the large trader implements a continuous sales strategy $Y = Y^c$ with no buy orders, then his cash position at time $T > 0$ is given by the Lebesgue–Stieltjes integral

$$\int_0^T B_{t-}^Y\, dY_t. \tag{2.11}$$

By assumption the best ask prices are greater than or equal to the unaffected bid prices, from which it follows that the cost of purchasing the shares at the best offered ask prices is greater than or equal to purchasing the shares at the unaffected bid prices. In view of (2.10) and (2.11), we therefore conclude that if the large investor's initial cash position is c and he uses a liquidation strategy $(X, Y) \in \mathcal{A}(y)$, his cash position at time $T > 0$ satisfies

$$\begin{aligned} C_T(X, Y) \leq c - \int_0^T B_{t-}^Y\, dY_t^c - \sum_{0 \leq t \leq T} \int_0^{\triangle Y_t} \{B_t^0 + \psi(Z_{t-}^Y + x)\}\, dx \\ - \int_0^T B_t^0\, dX_t, \end{aligned} \tag{2.12}$$

where we have equality in (2.12) if $X = 0$. We note that (2.12) can be viewed as a version of proposition 2.22 in Alfonsi and Schied (2010).

We assume that the large trader has constant absolute risk aversion, an initial cash position c, and an initial position in the stock consisting of y number of shares. We further assume that the large trader wants to maximize the expected utility of his cash position at the end of time. In mathematical terms, the large investor's optimal liquidation problem is

$$\sup_{(X, Y) \in \mathcal{A}(y)} \mathbb{E}[U(C_\infty(X, Y))], \tag{2.13}$$

where the utility function U is given by

$$U(c) = -\mathrm{e}^{-Ac}, \qquad A > 0.$$

This can be seen as the limit order book equivalent formulation of the optimal liquidation problem studied by Schied and Schöneborn (2009), but restricted to the case of large investors with constant absolute risk aversion.

3. PRELIMINARY OBSERVATIONS AND PROBLEM SIMPLIFICATION

Our approach to solving the utility maximization problem (2.13) is to first show that the problem can be reduced to a deterministic optimization problem involving only liquidation strategies in $\mathcal{A}_D^-(y)$. This reduction of the problem is based on the ideas in Schied et al. (2010), who proved that if the market has a certain structure and the investor has a constant absolute risk aversion, the optimal strategy is deterministic.

Let $(X, Y) \in \mathcal{A}(y)$ be an admissible liquidation strategy. Then, it follows from (2.12) that

$$C_T(X, Y) \le c + by - B_T^0(X_T + Y_T) + \int_0^T (X_{t-} + Y_{t-})\sigma\, dW_t - F_T(Y), \tag{3.1}$$

where F is given by

$$F_T(Y) = \int_0^T \psi\big(Z_{t-}^Y\big)\, dY_t^c + \sum_{0\le t\le T} \int_0^{\triangle Y_t} \psi\big(Z_{t-}^Y + x\big)\, dx, \tag{3.2}$$

and where we have equality in (3.1) if $X = 0$. It follows from (2.1) that $B_T^0(X_T + Y_T)$ tends to 0 in $L^1(\mathbb{P})$ as $T \to \infty$, and that

$$\int_0^\infty (X_{t-} + Y_{t-})\sigma\, dW_t$$

is a well-defined random variable with expectation 0 and finite variance. Also note that $F_T(Y)$ is an increasing function of T, and therefore $F_\infty(Y)$ is well defined, possibly being equal to $+\infty$. Thus, F_∞ is a function from the set of càdlàg, nonincreasing functions into the extended nonnegative real numbers. We conclude that

$$C_\infty(X, Y) \le c + by + \int_0^\infty (X_{t-} + Y_{t-})\sigma\, dW_t - F_\infty(Y), \tag{3.3}$$

where we have equality in (3.3) if $X = 0$. Also note that

$$C_\infty(X, Y) \le c + by + \int_0^\infty (X_{t-} + Y_{t-})\sigma\, dW_t,$$

from which it follows that the market is arbitrage-free. In particular, there does not exist any price manipulation strategies (see Huberman and Stanzl 2004; Alfonsi and Schied 2010; or Gatheral 2010). We have the following monotonicity result for the function F and deterministic strategies:

LEMMA 3.1. *Let F be given by (3.2) and $(X, Y) \in \mathcal{A}(y)$ be any deterministic liquidation strategy. Then there exists a strategy $\widetilde{Y} \in \mathcal{A}_D^-(y)$ such that*

$$\int_0^\infty (X_{t-} + Y_{t-})^2\, dt \ge \int_0^\infty \widetilde{Y}_{t-}^2\, dt \quad \textit{and} \quad F_\infty(Y) \ge F_\infty(\widetilde{Y}).$$

Proof. Let ξ be a càdlàg, nonincreasing function satisfying $\xi_{0-} = 0$. Then,

$$Z_t^{Y+\xi} - Z_t^Y = -\lambda \int_0^t \big(Z_{u-}^{Y+\xi} - Z_{u-}^Y\big)\, du + \xi_t,$$

from which it follows that $Z_t^{Y+\xi} \leq Z_t^Y$, for all $t \geq 0$. Therefore,

$$\begin{aligned} F_\infty(Y+\xi) &= \int_0^\infty \psi\big(Z_{t-}^{Y+\xi}\big)\, dY_t^c + \int_0^\infty \psi\big(Z_{t-}^{Y+\xi}\big)\, d\xi_t^c \\ &\quad + \sum_{t\geq 0} \int_0^{\triangle Y_t} \psi\big(Z_{t-}^{Y+\xi} + x\big)\, dx + \sum_{t\geq 0} \int_{\triangle Y_t}^{\triangle Y_t + \triangle \xi_t} \psi\big(Z_{t-}^{Y+\xi} + x\big)\, dx \\ &\geq \int_0^\infty \psi\big(Z_{t-}^Y\big)\, dY_t^c + \sum_{t\geq 0} \int_0^{\triangle Y_t} \psi\big(Z_{t-}^Y + x\big)\, dx \\ &= F_\infty(Y), \end{aligned}$$

by the monotonicity of ψ. Thus, if $Y \leq \widetilde{Y}$ then $F_\infty(Y) \geq F_\infty(\widetilde{Y})$. Let $(X, Y) \in \mathcal{A}(y)$ be a deterministic strategy, and define $\widetilde{Y}_t = \max\{0, Y_t\}$ and $\widetilde{X} = 0$. Then, the strategy $(\widetilde{X}, \widetilde{Y})$ is in $\mathcal{A}_D^-(y)$, and since $Y \leq \widetilde{Y}$ and

$$\int_0^\infty (X_{t-} + Y_{t-})^2\, dt \geq \int_0^\infty \widetilde{Y}_{t-}^2\, dt,$$

the result follows. □

Let $(X, Y) \in \mathcal{A}(y)$ and define the process M by

$$M_t = \exp\left(-\sigma A \int_0^t (X_{s-} + Y_{s-})\, dW_s - \frac{1}{2}\sigma^2 A^2 \int_0^t (X_{s-} + Y_{s-})^2\, ds\right).$$

From the assumption (2.1), it follows that M is a martingale closed by M_∞ (see, e.g., Protter 1990). We can therefore define a probability measure $\widetilde{\mathbb{P}}$ by

$$\frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}} = M_\infty.$$

Based on the ideas of theorem 2.8 in Schied et al. (2010), (3.3), and Lemma 3.1, we calculate

$$\begin{aligned} &\sup_{(X,Y)\in\mathcal{A}(y)} \mathbb{E}[U(C_\infty(X, Y))] \\ &\overset{(*)}{\leq} -\mathrm{e}^{-A(c+by)} \inf_{(X,Y)\in\mathcal{A}(y)} \mathbb{E}\left[\exp\left(-A\int_0^\infty (X_{t-} + Y_{t-})\sigma\, dW_t + AF_\infty(Y)\right)\right] \\ &= -\mathrm{e}^{-A(c+by)} \inf_{(X,Y)\in\mathcal{A}(y)} \mathbb{E}\left[M_\infty \exp\left(\frac{1}{2}\sigma^2 A^2 \int_0^\infty (X_{t-} + Y_{t-})^2\, dt + AF_\infty(Y)\right)\right] \\ &= -\mathrm{e}^{-A(c+by)} \inf_{(X,Y)\in\mathcal{A}(y)} \widetilde{\mathbb{E}}\left[\exp\left(\frac{1}{2}\sigma^2 A^2 \int_0^\infty (X_{t-} + Y_{t-})^2\, dt + AF_\infty(Y)\right)\right] \\ &= -\mathrm{e}^{-A(c+by)} \exp\left(\inf_{Y\in\mathcal{A}_D^-(y)} \left\{\frac{1}{2}\sigma^2 A^2 \int_0^\infty (X_{t-} + Y_{t-})^2\, dt + AF_\infty(Y)\right\}\right), \end{aligned} \tag{3.4}$$

where $(*)$ holds with equality for strategies $(X, Y) \in \mathcal{A}(y)$ with $X = 0$. In particular, $(*)$ in (3.4) holds with equality for strategies $Y \in \mathcal{A}_D^-(y)$. The last step in (3.4) follows from Jensen's inequality and Lemma 3.1. We have therefore reduced the utility maximization problem (2.13) to an optimization problem involving only deterministic liquidation strategies which involve no buy orders.

In order to derive a Hamilton–Jacobi–Bellman equation for the value function of the optimization problem, we want to obtain an expression for F_∞ which is more convenient for this purpose. As a first step toward this, the next result expresses the cash position at a time $T > 0$ corresponding to a liquidation strategy that involves no buy orders.

LEMMA 3.2. *For every initial cash position c and liquidation strategy $(X, Y) \in \mathcal{A}(y)$ satisfying $X = 0$, the large investor's cash position at time $T > 0$ is given by*

$$\begin{aligned} C_T(Y) = c + B^0_{0-} Y_{0-} - \int_{Z^Y_{0-}}^{Z^Y_{0-} - Y_{0-}} \psi(s)\, ds - B^0_T Y_T + \int_{Z^Y_T}^{Z^Y_T - Y_T} \psi(s)\, ds \\ + \int_0^T Y_{t-}\, dB^0_t + \int_0^T \lambda Z^Y_{t-} \{\psi\left(Z^Y_{t-} - Y_{t-}\right) - \psi\left(Z^Y_{t-}\right)\} \, dt. \end{aligned} \tag{3.5}$$

The expression for the cash position at time T in Lemma 3.2 should be of independent interest. It can be used to derive the Hamilton–Jacobi–Bellman equation for the optimal liquidation problem with a finite time-horizon and a general utility function. Our idea is similar to the idea in Schied and Schöneborn (2009), which consists of rewriting the expression for the cash position at the end of time in such a way that you make use of the assumption that the stock position at the end of time is zero. The next result makes use of Lemma 3.2 and provides an expression for the cash position at the end of time in accordance with this idea.

LEMMA 3.3. *Assume that the large investor's initial cash position is c and that the large investor uses a liquidation strategy $(X, Y) \in \mathcal{A}(y)$, with $X = 0$. Then, the large investor's cash position at the end of time is*

$$\begin{aligned} C_\infty(Y) = c + by - \int_z^{z-y} \psi(s)\, ds \\ + \int_0^\infty \sigma Y_{t-}\, dW_t + \int_0^\infty \lambda Z^Y_{t-} \{\psi(Z^Y_{t-} - Y_{t-}) - \psi(Z^Y_{t-})\}\, dt, \end{aligned} \tag{3.6}$$

where $b = B^0_0$ and $z = Z^Y_{0-}$. Moreover,

$$0 \le \int_0^\infty \lambda Z^Y_{t-} \{\psi(Z^Y_{t-} - Y_{t-}) - \psi(Z^Y_{t-})\}\, dt \le \int_z^{z-y} \psi(s)\, ds. \tag{3.7}$$

The various terms in (3.6), for the cash position at the end of time, all have economic interpretations. The term

$$c + by - \int_z^{z-y} \psi(s)\, ds$$

corresponds to the cash position of the large investor after immediate liquidation of the entire position in risky assets. The term

$$\int_0^\infty \sigma Y_{t-}\, dW_t$$

represents the risk of the large investor's cash position at the end of time if the stock position is not liquidated immediately, and

$$\int_0^\infty \lambda Z_{t-}^Y \{\psi(Z_{t-}^Y - Y_{t-}) - \psi(Z_{t-}^Y)\}\, dt$$

represents the gain to the large investor's cash position for not liquidating immediately.

REMARK 3.4. From the bounds established in the proof of Lemma (3.6), it follows that if Y^n is a sequence of liquidation strategies in $\mathcal{A}_D^-(y)$ such that Y^n converges to $Y \in \mathcal{A}_D^-(y)$ in total variation, then $C_\infty(Y^n)$ converges to $C_\infty(Y)$ in $L^1(\mathbb{P})$. This shows that the limit order book model exhibits a certain robustness. For instance, given a sequence of absolutely continuous liquidation strategies converging to a liquidation strategy consisting of block trades, the corresponding cash position of the absolutely continuous strategy will converge to the cash position corresponding to that consisting of the block trades. This differs completely from the situation in the supply curve models of Cetin, Jarrow, and Protter (2004) and Bank and Baum (2004).

For strategies $Y \in \mathcal{A}_D^-(y)$, (3.3) takes the form

$$C_\infty(Y) = c + by + \int_0^\infty \sigma Y_{t-}\, dW_t - F_\infty(Y),$$

and by comparing this expression for $C_\infty(Y)$ with the expression for $C_\infty(Y)$ in Lemma 3.3, we conclude that

$$F_\infty(Y) = \int_z^{z-y} \psi(s)\, ds + \int_0^\infty \lambda Z_{t-}^Y \{\psi(Z_{t-}^Y) - \psi(Z_{t-}^Y - Y_{t-})\}\, dt. \tag{3.8}$$

From the calculations in (3.4) and equation (3.8), it follows that the optimization problem (2.13) takes the form

$$\sup_{(X,Y)\in\mathcal{A}(y)} \mathbb{E}[U(C_\infty(X, Y))] = -\exp\left(-A(c+by) + A\int_z^{z-y} \psi(s)\, ds\right) \exp(AV(y, z)), \tag{3.9}$$

where

$$V(y, z) = \inf_{Y\in\mathcal{A}_D^-(y)} \int_0^\infty \lambda Y_{t-} \left(aY_{t-} + Z_{t-}^Y \frac{\psi(Z_{t-}^Y) - \psi(Z_{t-}^Y - Y_{t-})}{Y_{t-}}\right) dt, \tag{3.10}$$

with $a = \frac{\sigma^2 A}{2\lambda}$ and $z = Z_{0-}^Y$. Thus, we have reduced the original utility optimization problem to (3.10). Moreover, from assumption (2.2) and Lemma 3.3, it follows that V given by (3.10) is well defined and real valued.

REMARK 3.5. Observe that from Lemma 3.3, it follows that for a liquidation strategy $Y \in \mathcal{A}_D^-(y)$,

$$\mathbb{E}[C_\infty(Y)] = c + by - \int_z^{z-y} \psi(s)\, ds + \int_0^\infty \lambda Z_{t-}^Y \{\psi(Z_{t-}^Y - Y_{t-}) - \psi(Z_{t-}^Y)\}\, dt$$

and

$$\mathrm{Var}(C_\infty(Y)) = \int_0^\infty \sigma^2 Y_{t-}^2 \, dt.$$

Therefore, the right-hand side of (3.10), with $a = \frac{\sigma^2 A}{2\lambda}$, can be written

$$c + by - \int_z^{z-y} \psi(s)\, ds + \inf_{Y \in \mathcal{A}_D^-(y)} \left\{ \frac{A}{2} \mathrm{Var}(C_\infty(Y)) - \mathbb{E}[C_\infty(Y)] \right\},$$

which corresponds to the Almgren and Chriss (1999) criterion. Schied et al. (2010) proved that this relationship between the Almgren and Chriss criterion and CARA utility holds also in the more general setting where the unaffected price process follows a Lévy process.

4. THE SOLUTION TO THE OPTIMIZATION PROBLEM

Our next aim is to derive an explicit solution to the optimization problem (3.10), which will be based on the principle of dynamic programming. With reference to the general theory of optimal control (see, e.g., Fleming and Soner 1993), the Hamilton–Jacobi–Bellman equation corresponding to V given by (3.10) takes the form

(4.1)
$$\max\{z v_z(y, z) - ay^2 - z(\psi(z) - \psi(z - y)), \max_{0 \le \triangle \le y} v(y, z) - v(y - \triangle, z - \triangle)\} = 0,$$

with associated boundary condition $v(0, z) = 0$, for all $z \le 0$. Formally, we can obtain equation (4.1) as follows. As we are optimizing over deterministic selling strategies (no buy orders), there are only two choices; to sell a number $\triangle > 0$ of shares or to wait. Given a state (y, z), it may or may not be optimal to sell a number $\triangle$ of shares, hence

$$v(y, z) \le v(y - \triangle, z - \triangle) \tag{4.2}$$

as the sale of $\triangle$ number of shares decrease the number of shares held from y to $y - \triangle$ and the state of the limit order book from z to $z - \triangle$. The inequality (4.2) should hold for all $0 \le \triangle \le y$, which implies that

$$\max_{0 \le \triangle \le y} \{v(y, z) - v(y - \triangle, z - \triangle)\} \le 0. \tag{4.3}$$

On the other hand, it may or may not be optimal to wait for a period of time $\triangle t > 0$, hence

(4.4)
$$\begin{aligned} v(y, z) &\le v(y, Z_{\triangle t}) + \int_0^{\triangle t} \lambda y \left(ay + Z_{u-} \frac{\psi(Z_{u-}) - \psi(Z_{u-} - y)}{y} \right) du \\ &= v(y, z) + \int_0^{\triangle t} \left\{ \lambda y \left(ay + Z_{u-} \frac{\psi(Z_{u-}) - \psi(Z_{u-} - y)}{y} \right) - v_z(y, Z_{u-}) \lambda Z_{u-} \right\} du, \end{aligned}$$

where $Z_u = ze^{-\lambda u}$, for $0 \le u \le \triangle t$. By multiplying inequality (4.4) by $(\triangle t)^{-1}$ and letting $\triangle t$ tend to 0, we obtain

$$zv_z(y,z) - ay^2 - z\{\psi(z) - \psi(z-y)\} \le 0. \tag{4.5}$$

Since it is optimal to either sell a certain number of shares or wait, we must have equality in either (4.3) or (4.5), and we obtain (4.1).

For this type of singular optimal control problem, the optimal strategy can be characterized by two disjoint sets $\mathcal{D}_s$ and $\mathcal{D}_w$, where the union of $\mathcal{D}_s$ and $\mathcal{D}_w$ is equal to the state space $\mathbb{R}^+ \times \mathbb{R}^-$. These sets satisfy

$$\begin{aligned} &\max_{0\le\triangle\le y} \{v(y,z) - v(y-\triangle, z-\triangle)\} = 0, && \text{for } (y,z) \in \mathcal{D}_s,\\ &zv_z(y,z) - ay^2 - z\{\psi(z) - \psi(z-y)\} = 0, && \text{for } (y,z) \in \mathcal{D}_w. \end{aligned} \tag{4.6}$$

If $(y,z) \in \mathcal{D}_s$, then the optimal strategy consists of making an immediate sale of $\triangle$ number of shares, where $\triangle$ is such that $(y-\triangle, z-\triangle)$ is on the nearest boundary between $\mathcal{D}_s$ and $\mathcal{D}_w$ (or the line $y = 0$). If $(y,z) \in \mathcal{D}_w$, then the optimal strategy consists of waiting until the first time $(y, ze^{-\lambda t})$ is on the boundary between $\mathcal{D}_s$ and $\mathcal{D}_w$, as the limit order book recovers at an exponential rate λ. If (y,z) is on the boundary between $\mathcal{D}_s$ and $\mathcal{D}_w$, then the optimal strategy consists of taking minimal action to ensure that the state process (Y_t, Z_t^Y) remains on the boundary.

The simplest case is when $\mathcal{D}_s$ and $\mathcal{D}_w$ are separated by a function h, i.e., $(y,z) \in \mathcal{D}_s$ if $z \ge h(y)$ and $(y,z) \in \mathcal{D}_w$ if $z < h(y)$ (or vice versa). Since

$$\lambda y \left(ay + z\frac{\psi(z) - \psi(z-y)}{y} \right) \tag{4.7}$$

is positive if z is large (i.e., close to zero) compared to y, and increasingly negative for small values of z, it is natural that $(y,z) \in \mathcal{D}_s$ if $z \ge h(y)$ and $(y,z) \in \mathcal{D}_w$ if $z < h(y)$. It turns out that this indeed is the case, and that the boundary between $\mathcal{D}_s$ and $\mathcal{D}_w$ can be described by a strictly decreasing càdlàg function $h : \mathbb{R}^+ \to \mathbb{R}^-$ satisfying $h(0) = 0$ and $\lim_{y\to\infty} h(y) = -\infty$, and that we should be looking for a function v satisfying

$$v_y(y,z) + v_z(y,z) = 0, \quad \text{for } z \ge h(y), \tag{4.8}$$

$$zv_z(y,z) - ay^2 - z(\psi(z) - \psi(z-y)) \le 0, \quad \text{for } z > h(y), \tag{4.9}$$

and

$$zv_z(y,z) - ay^2 - z(\psi(z) - \psi(z-y)) = 0, \quad \text{for } z \le h(y), \tag{4.10}$$

$$D_y^+ v(y,z) + v_z(y,z) \le 0, \quad \text{for } z < h(y), \tag{4.11}$$

where

$$D_y^+ v(y,z) = \lim_{\epsilon\to 0^+} \frac{1}{\epsilon}(v(y+\epsilon, z) - v(y,z)). \tag{4.12}$$

Let us examine in more detail the strategy corresponding to the case where $\mathcal{D}_s$ and $\mathcal{D}_w$ are described by an intervention boundary function h. Given a strictly decreasing càdlàg

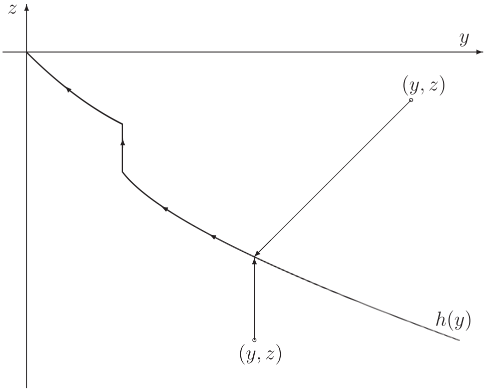

FIGURE 4.1. Illustration of the strategy Y^h corresponding to h.

function $h:\mathbb{R}^+\to\mathbb{R}^-$ satisfying $h(0)=0$ and $\lim_{y\to\infty}h(y)=-\infty$, and corresponding sets $\mathcal{D}_s$ and $\mathcal{D}_w$ given by $(y,z)\in\mathcal{D}_s$ if $z\geq h(y)$ and $(y,z)\in\mathcal{D}_w$ if $z<h(y)$, we might ask whether the corresponding liquidation strategy, denoted by Y^h, exists. For future reference, introduce the following functions related to h:

$$\gamma_h(y)=h(y)-y,\qquad \text{for } y\in\mathbb{R}^+, \tag{4.13}$$

$$\rho_h(z)=z-h^{-1}(z),\qquad \text{for } z\in\mathbb{R}^-, \tag{4.14}$$

$$h^{-1}(z)=\sup\{y\in\mathbb{R}^+ : h(y)\geq z\},\qquad \text{for } z\in\mathbb{R}^-, \tag{4.15}$$

$$\gamma_h^{-1}(z)=\sup\{y\in\mathbb{R}^+ : \gamma_h(y)\geq z\},\qquad \text{for } z\in\mathbb{R}^-, \tag{4.16}$$

and let ρ_h^{-1} denote the inverse of ρ_h. We can then observe that if $(y,z)\in\mathcal{D}_s$, then the strategy Y^h corresponding to the intervention boundary described by h, will consist of making an initial sale of $\triangle$ number of share such that $(y-\triangle,z-\triangle)$ is on the graph of h (see Figure 4.1). Hence, we want $(y-\triangle,z-\triangle)=(h^{-1}(z-\triangle),z-\triangle)$. With $Z_{0-}^{Y^h}=z$ and $Z_0^{Y^h}=z-\triangle$, we see that this equation is equivalent to

$$\rho_h\big(Z_0^{Y^h}\big)=Z_0^{Y^h}-h^{-1}\big(Z_0^{Y^h}\big)=z-y,$$

from which it follows that $Z_0^{Y^h}=\rho_h^{-1}(z-y)$ and $\triangle=z-\rho_h^{-1}(z-y)$. For future reference, we can also observe that $\rho_h^{-1}(x)=x+\gamma_h^{-1}(x)$, for $x\in\mathbb{R}^-$. Hence, the number $\triangle$ of shares can also be expressed by $\triangle=y-\gamma_h^{-1}(z-y)$. If $(y,z)\in\mathcal{D}_w$, i.e., $z<h(y)$, then the strategy Y^h consists in waiting until the state process $(Y_t^h,Z_t^{Y^h})$ is on the graph of h (see Figure 4.1). While no action is taken, $Y_t^h=y$ and $Z_t^{Y^h}=ze^{-\lambda t}$, from which it follows that the first time t_w that the state process is on the graph of h is given by the equation $ze^{-\lambda t_w}=h(y)$. Once the state process (Y^h,Z^{Y^h}) is on the graph of h, the strategy Y^h consists of taking minimal action such that the state process remains on the graph of h (see Figure 4.1). This implies that $(Y_t^h,Z_t^{Y^h})=(h^{-1}(Z_t^{Y^h}),Z_t^{Y^h})$. With reference to (2.8),

this implies that Z^{Y^h} should solve

$$dZ_t^{Y^h} = -\lambda Z_{t-}^{Y^h}\,dt + dh^{-1}\big(Z_t^{Y^h}\big),$$

which is equivalent to

$$d\rho_h\big(Z_t^{Y^h}\big) = -\lambda Z_{t-}^{Y^h}\,dt.$$

The following result establishes the existence and uniqueness of such a strategy Y^h for a given intervention boundary function h.

LEMMA 4.1. *Let $(y, z) \in \mathbb{R}^+ \times \mathbb{R}^-$ and let $h : \mathbb{R}^+ \to \mathbb{R}^-$ be a strictly decreasing càdlàg function satisfying $h(0) = 0$ and $\lim_{y\to\infty} h(y) = -\infty$, let h^{-1} and γ_h^{-1} be given by (4.15) and (4.16). Set*

$$t_w = \begin{cases} 0, & \text{if } z \geq h(y), \\ \lambda^{-1}\{\ln(-z) - \ln(-h(y))\}, & \text{if } z < h(y), \end{cases} \tag{4.17}$$

and let Y^h denote the decreasing càdlàg liquidation strategy with the following description:

- *If $z \geq h(y)$, then immediately sell $y - \gamma_h^{-1}(z - y)$ number of shares. This block trade ensures that $Y_0^h = h^{-1}(Z_0^{Y^h})$. Then continuously sell shares so that $Y_t^h = h^{-1}(Z_t^{Y^h})$, for all $t \geq 0$.*
- *If $z < h(y)$, then do nothing until time t_w. The time t_w has the property that $y = h^{-1}(Z_{t_w}^{Y^h})$. Then, continuously sell shares so that $Y_t^h = h^{-1}(Z_t^{Y^h})$, for all $t \geq t_w$.*

Such a strategy Y^h exists and is unique. In particular,

$$Y_t^h = h^{-1}(Z_t^{Y^h}), \quad \text{for } t \geq t_w, \tag{4.18}$$

and Z^{Y^h} is the unique solution to

$$Z_t^{Y^h} = Z_{t_w}^{Y^h} - h^{-1}(Z_{t_w}^{Y^h}) - \int_{t_w}^{t} \lambda Z_u^{Y^h}\,du + h^{-1}\left(Z_t^{Y^h}\right), \tag{4.19}$$

where

$$Z_{t_w}^{Y_h} = h(y), \text{ if } z < h(y), \quad \text{and} \quad Z_{t_w}^{Y^h} = z - y + \gamma_h^{-1}(z - y), \text{ if } z \geq h(y). \tag{4.20}$$

If $t_w > 0$, then $Y_t^h = y$ and $Z_t^{Y^h} = z\mathrm{e}^{-\lambda t}$, for $0 \leq t \leq t_w$.

The next result concerns the relative speed of liquidation corresponding to two different intervention boundaries. In particular, we will need this result later in order to prove that our candidate for the optimal strategy is admissible.

LEMMA 4.2. *Let $h_1, h_2 : \mathbb{R}^+ \to \mathbb{R}^-$ be strictly decreasing càdlàg functions satisfying $h_1(0) = h_2(0) = 0$ and $\lim_{y\to\infty} h_1(y) = \lim_{y\to\infty} h_1(y) = -\infty$. Denote by Y^{h_1} and Y^{h_2} the strategies defined in Lemma 4.1 by (4.18)–(4.19) corresponding to h_1 and h_2, respectively. If $h_1 \leq h_2$ then $Y^{h_1} \leq Y^{h_2}$. In particular, if there exists $C > 0$ and $\epsilon > 0$ such that $h(y) \leq -Cy$ for all $y < \epsilon$, then $Y^h \in \mathcal{A}_D^-(y)$.*

In order to obtain an explicit expression for the value function of our problem, we progress by deriving an explicit expression for the performance associated with a strategy

Y^h, given an arbitrary intervention boundary function h. For an initial state (y, z) and strategy Y^h, with associated bid order book state process Z^{Y^h}, define the performance function J_h by

$$J_h(y, z) = \int_0^\infty \lambda Y^h_{t-} \left(a Y^h_{t-} + Z^{Y^h}_{t-} \frac{\psi(Z^{Y^h}_{t-}) - \psi(Z^{Y^h}_{t-} - Y^h_{t-})}{Y^h_{t-}} \right) dt, \tag{4.21}$$

where $Y^h_{0-} = y$ and $Z^{Y^h}_{0-} = z$. Also, define

$$I_h(z) = J_h(h^{-1}(z), z), \qquad \text{for } z \in \mathbb{R}^-,$$

which corresponds to the performance of the strategy Y^h if the initial state is on the graph of h. If the initial state (y, z) is such that $z \geq h(y)$ then the strategy Y^h consists of an initial sale of $y - \gamma_h^{-1}(z - y) = z - \rho_h^{-1}(z - y)$ number of shares, and the state after the block sale is $(Y^h_0, Z^{Y^h}_0) = (h^{-1}(\rho_h^{-1}(z - y)), \rho_h^{-1}(z - y))$, where γ_h^{-1} and ρ_h are given by (4.16) and (4.14), respectively. Hence,

$$J_h(y, z) = I_h(\rho_h^{-1}(z - y)), \qquad \text{for } z \geq h(y). \tag{4.22}$$

Thus, if we obtain an explicit expression for I_h, we have an explicit expression for $J_h(y, z)$, for all $z \geq h(y)$. For an initial state $(y, z) = (h^{-1}(z), z)$, the liquidation strategy Y^h satisfies $(Y^h_t, Z^{Y^h}_t) = (h^{-1}(Z^{Y^h}_t), Z^{Y^h}_t)$, for all $t \geq 0$. Therefore,

$$I_h(z) = \int_0^\infty \left(a\lambda\left(h^{-1}(Z^{Y^h}_{t-})\right)^2 + \lambda Z^{Y^h}_{t-}\left\{\psi(Z^{Y^h}_{t-}) - \psi\left(\rho_h(Z^{Y^h}_{t-})\right)\right\}\right) dt, \qquad Z^{Y^h}_{0-} = z.$$

With reference to (2.8), we note that formally,

$$dt = -\frac{d\rho_h(Z^{Y^h}_t)}{\lambda Z^{Y^h}_{t-}},$$

and hence

$$\begin{aligned}
I_h(z) &= -\int_0^\infty \frac{a\left(h^{-1}(Z^{Y^h}_{t-})\right)^2}{Z^{Y^h}_{t-}} \, d\rho_h(Z^{Y^h}_t) + \int_0^\infty \psi\left(\rho_h(Z^{Y^h}_{t-})\right) d\rho_h(Z^{Y^h}_t) \\
&\quad - \int_0^\infty \psi(Z^{Y^h}_{t-}) \, d\rho_h(Z^{Y^h}_t) \\
&= -\int_0^\infty \frac{a\left(\gamma_h^{-1}\left(\rho_h(Z^{Y^h}_{t-})\right)\right)^2}{\rho_h^{-1}\left(\rho_h(Z^{Y^h}_{t-})\right)} \, d\rho_h(Z^{Y^h}_t) + \int_0^\infty \psi\left(\rho_h(Z^{Y^h}_{t-})\right) d\rho_h(Z^{Y^h}_t) \\
&\quad - \int_0^\infty \psi\left(\rho_h^{-1}\left(\rho_h(Z^{Y^h}_{t-})\right)\right) d\rho_h(Z^{Y^h}_t) \\
&= \int_0^{\rho_h(z)} \left(\frac{a\left(\gamma_h^{-1}(u)\right)^2}{\rho_h^{-1}(u)} + \psi\left(\rho_h^{-1}(u)\right) - \psi(u) \right) du,
\end{aligned}$$

where we have used that $\gamma_h^{-1}(\rho_h^{-1}(x)) = h^{-1}(x)$, for $x \in \mathbb{R}^-$. With reference to (4.22) we conclude that

$$J_h(y,z) = \int_0^{z-y} \left(\frac{a\left(\gamma_h^{-1}(u)\right)^2}{\rho_h^{-1}(u)} + \psi\left(\rho_h^{-1}(u)\right) - \psi(u) \right) du, \qquad \text{for } z \geq h(y). \tag{4.23}$$

In order to obtain an expression for $J_h(y,z)$ for $z < h(y)$, we note that the strategy Y^h consists of waiting until the first time t_w for which $(Y^h_{t_w}, Z^{Y^h}_{t_w})$ is on the graph of h, where $Y^h_t = y$ and $Z^{Y^h}_t = z\mathrm{e}^{-\lambda t}$, for $0 \leq t \leq t_w$. Therefore, $t_w = \lambda^{-1}\ln(z/h(y))$, and

$$\begin{aligned} J_h(y,z) &= \int_0^{t_w} \lambda y \left(ay + z\mathrm{e}^{-\lambda t} \frac{\psi(z\mathrm{e}^{-\lambda t}) - \psi(z\mathrm{e}^{-\lambda t} - y)}{y} \right) dt + I_h(h(y)) \\ &= ay^2 \ln\left(\frac{z}{h(y)}\right) + \int_{z-y}^{h(y)-y} \psi(u)\,du - \int_z^{h(y)} \psi(u)\,du \\ &\quad + \int_0^{h(y)-y} \left(\frac{a\left(\gamma_h^{-1}(u)\right)^2}{\rho_h^{-1}(u)} + \psi\left(\rho_h^{-1}(u)\right) - \psi(u) \right) du, \qquad \text{for } z < h(y). \end{aligned}$$

While this provides an explicit expression for $J_h(y,z)$, it is not obvious from this expression that it is continuous in y (and has a one-sided derivative with respect to y), as h is only a càdlàg function. However, we can calculate that

$$\begin{aligned} \int_0^{\gamma_h(y)} \left(\frac{a\left(\gamma_h^{-1}(u)\right)^2}{\rho_h^{-1}(u)} + \psi\left(\rho_h^{-1}(u)\right) \right) du &= \int_0^y \left(\frac{au^2}{h(u-)} + \psi(h(u-)) \right) d\gamma_h^c(u) \\ &\quad + \sum_{0 \leq u \leq y} au^2 \ln\left(\frac{h(u)}{h(u-)}\right) + \sum_{0 \leq u \leq y} \int_{h(u-)}^{h(u)} \psi(s)\,ds. \end{aligned}$$

From this expression, as well as

$$\int_0^{h(y)} \psi(u)\,du = \int_0^y \psi(u)\,dh^c(u) + \sum_{0 \leq u \leq y} \int_{h(u-)}^{h(u)} \psi(s)\,ds,$$

and

$$ay^2 \ln(-h(y)) = \int_0^y 2au \ln(h(u-))\,du + \int_0^y \frac{au^2}{h(u-)}\,dh^c(u) + \sum_{0 \leq u \leq y} au^2 \ln\left(\frac{h(u)}{h(u-)}\right),$$

it follows that the performance function $J_h(y,z)$ admits the expression

$$\begin{aligned} J_h(y,z) &= ay^2 \ln(-z) + \int_{z-y}^z \psi(u)\,du \\ &\quad - \int_0^y \left(\frac{au^2}{h(u)} + \psi(h(u)) + 2au \ln(-h(u)) \right) du. \qquad \text{for } z < h(y). \end{aligned}$$

The next result provides an explicit equation for the intervention boundary function h and the corresponding value function which solves equation (4.1) with associated boundary condition $v(0,z) = 0$, for all $z \in \mathbb{R}^-$.

PROPOSITION 4.3. *For $y \in \mathbb{R}^+$, define the function $\Gamma(\cdot; y) : \mathbb{R}^- \to \mathbb{R}$ by*

$$\Gamma(x; y) = \psi(x) + \frac{ay^2}{x} + 2ay\ln(-x),$$

and let $h = h(y)$ be the smallest $h \in \mathbb{R}^-$ satisfying

$$\max_{x \le 0} \Gamma(x; y) = \Gamma(h(y); y). \tag{4.24}$$

This defines a unique strictly decreasing càdlàg function $h : \mathbb{R}^+ \to \mathbb{R}^-$ satisfying $h(0) = 0$ and $\lim_{y\to\infty} h(y) = -\infty$, and in particular

$$\psi'(h(y))h(y)^2 + 2ayh(y) - ay^2 = 0, \qquad \text{for all } y \in \mathbb{R}^+. \tag{4.25}$$

Let γ_h^{-1}, ρ_h, and ρ_h^{-1} be the functions defined in (4.16) and (4.14). Then, $v : \mathbb{R}^+ \times \mathbb{R}^- \to \mathbb{R}^-$ given by

$$v(y, z) = \int_0^{z-y} \left(\frac{a\big(\gamma_h^{-1}(s)\big)^2}{\rho_h^{-1}(s)} + \psi\big(\rho_h^{-1}(s)\big) - \psi(s) \right) ds, \qquad \text{for } z \ge h(y), \tag{4.26}$$

and

$$\begin{aligned} v(y, z) &= ay^2\ln(-z) + \int_{z-y}^{z} \psi(s)\,ds \\ &\quad - \int_0^y \left(\frac{as^2}{h(s)} + \psi(h(s)) + 2as\ln(-h(s)) \right) ds, \qquad \text{for } z < h(y) \end{aligned} \tag{4.27}$$

is a $C^{0,1}(\mathbb{R}^+ \times \mathbb{R}^-)$ function which solves equation (4.1) with the boundary condition $v(0, z) = 0$, for all $z \in \mathbb{R}^-$. In particular, v satisfies (4.8)–(4.11). Moreover, $D_y^+ v : \mathbb{R}^+ \times \mathbb{R}^- \to \mathbb{R}$ is continuous in z and càdlàg in y, and v is continuously differentiable with respect to y, for $z \ge h(y)$.

Based on the expression for the performance function $J_h(y, z)$ and the principle of smooth fit, it follows that the intervention boundary function h should satisfy (4.25). However, (4.25) does not necessarily have a unique solution and the value function does not in general satisfy the smoothness principle. Based on the observation that

$$D_y^+ v(y, z) + v_z(y, z) = \Gamma(z; y) - \Gamma(h(y); y), \qquad \text{for } z < h(y),$$

and with reference to (4.11), equation (4.24) is a natural candidate.

The following result states that the function v given by (4.26) and (4.27) is equal to the value function V given by (3.10), that the strategy Y^h corresponding to h given by (4.24) is an optimal liquidation strategy, and hence provides the solution to the utility maximization problem (2.13).

THEOREM 4.4. *Let the large investor's risk aversion be A, the volatility of the nonaffected asset price be σ and let the resilience rate be λ. Set $a = \frac{\sigma^2 A}{2\lambda}$ and let h denote the smallest solution to (4.24), let v be given by (4.26) and (4.27), and let V be given by (3.10). Then, $v = V$ and*

$$\sup_{(X,Y)\in\mathcal{A}(y)} \mathbb{E}[U(C_\infty(X, Y))] = -\exp\left(-A(c + by) + A\int_z^{z-y} \psi(s)\,ds \right) \exp(Av(y, z)),$$

where $z = Z^Y_{0-}$ is the initial state of the bid order book. The optimal strategy Y^* is equal to $Y^h \in \mathcal{A}^-_D(y)$, where Y^h is the strategy given by (4.18)–(4.20) in Lemma 4.1 corresponding to h, with $Y^h_{0-} = y$.

As a corollary, we obtain that the optimal liquidation strategy possesses all the properties one would expect, like increased limit order depth implies faster liquidation, increased volatility of the unaffected stock price implies faster liquidation, and increased risk aversion implies faster liquidation.

COROLLARY 4.5. *Let $Y^*_1 \in \mathcal{A}^-_D(y)$ be the strategy which attains the optimality in (2.13) given a shape function ϕ_1, and let $Y^*_2 \in \mathcal{A}^-_D(y)$ denote the strategy which attains the optimality in (2.13) given a shape function ϕ_2. Then, $\phi_1 \le \phi_2$ implies $Y^*_1 \ge Y^*_2$, provided the volatility σ, the resilience rate λ, and the large investor's risk aversion A is the same.*

Let $Y^*_1 \in \mathcal{A}^-_D(y)$ be the strategy which attains the optimality in (2.13) given volatility σ_1 and risk aversion A_1, and let $Y^*_2 \in \mathcal{A}^-_D(y)$ denote the strategy which attains the optimality in (2.13) given volatility σ_2 and risk aversion A_2. Then, $\sigma_1^2 A_1 \le \sigma_2^2 A_2$ implies that $Y^*_1 \ge Y^*_2$, provided that the shape function ϕ and the resilience rate λ is the same.

Proof. Observe that $\phi_1 \le \phi_2$ implies that $\psi_1' \ge \psi_2'$, and hence $\psi_1 \le \psi_2$. It follows that $h_1 \ge h_2$, where h_1 and h_2 denote smallest solution to (4.24) corresponding to ψ_1 and ψ_2, respectively. The result then follows from Lemma 4.2.

Notice that $a_1 \le a_2$ implies that $h_1 \ge h_2$, where h_1 and h_2 denote the smallest solution to (4.24) corresponding to a_1 and a_2, respectively. The result then follows from Lemma 4.2. □

The next example shows that if the shape function ϕ is constant, and hence ψ is a linear function, the solution to equation (4.1) is a linear function, and the corresponding optimal strategy takes an even simpler form. In this case, the impact of the large trader's strategy is linear, so it is natural to compare the results with the corresponding strategy for the Almgren and Chriss (1999) model with an infinite horizon as in Schied and Schöneborn (2009).

EXAMPLE. Suppose that the shape function ϕ of the limit order book is constant, i.e., $\phi = c$, for some $c > 0$. Let the large investor's risk aversion be A, the volatility of the unaffected stock price be σ, let the resilience rate be λ and set $a = \frac{\sigma^2 A}{2\lambda}$. Observe that for all $y \in \mathbb{R}^+$, equation (4.25) has a unique solution $h = h(y)$ given by

$$h(y) = -\kappa y,$$

where $\kappa = ac + \sqrt{a^2c^2 + ac}$. It follows that this function h is the unique solution to equation (4.24), and therefore defines the optimal intervention boundary. Moreover, we can observe that

$$h^{-1}(z) = -\frac{1}{\kappa} z, \qquad \gamma_h^{-1}(z) = -\frac{1}{\kappa+1} z, \qquad \text{and} \qquad \rho_h^{-1}(z) = \frac{\kappa}{\kappa+1} z.$$

With reference to (4.26) and (4.27), it follows from Proposition 4.3 that

$$v(y,z) = \frac{ac-\kappa}{2c\kappa(\kappa+1)}(y-z)^2, \qquad \text{if } z \ge -\kappa y,$$

and

$$v(y,z) = \frac{ac(\kappa+1)+\kappa(\kappa-1)}{2c\kappa}y^2 + \frac{zy}{c} + ay^2 \ln\left(\frac{-z}{\kappa y}\right), \qquad \text{if } z < -\kappa y$$

is a $C^{1,1}(\mathbb{R}^+ \times \mathbb{R}^-)$ solution to equation (4.1) with boundary condition $v(0,z)=0$, for all $z \le 0$. Equation (4.19) takes the form

$$Z_t^{Y^h} = Z_{t_w}^{Y^h} + \frac{1}{\kappa} Z_{t_w}^{Y^h} - \int_{t_w}^{t} \lambda Z_{u-}\,du - \frac{1}{\kappa} Z_t^{Y^h}, \qquad \text{for } t \ge t_w,$$

from which it follows that

$$Z_t^{Y^h} = Z_{t_w}^{Y^h} \exp\left(-t\frac{\lambda\kappa}{\kappa+1}\right), \qquad \text{for } t \ge t_w.$$

Therefore, the strategy $Y^* \in \mathcal{A}_D^-(y)$ which attains the optimality in (2.13) is as follows:

(a) if $z \ge -\kappa y$, then immediately sell $\frac{\kappa y+z}{1+\kappa}$ number of shares, i.e., $Y_0^* - y = -\frac{\kappa y+z}{1+\kappa}$, and then continuously sell shares according to

$$Y_t^* = \frac{y-z}{\kappa+1}\exp\left(-t\frac{\lambda\kappa}{1+\kappa}\right), \qquad \text{for } t \ge 0;$$

(b) if $z < -\kappa y$, then do nothing until time $t_w = \lambda^{-1}\{\ln(-z) - \ln(\kappa y)\}$, and then continuously sell shares according to

$$Y_t^* = y\exp\left(-(t-t_w)\frac{\lambda\kappa}{1+\kappa}\right), \qquad \text{for } t \ge t_w.$$

It is natural to compare our result for the limit order book with constant shape function with the optimal liquidation strategy in the Almgren and Chriss (1999) model. In this model, the stock price dynamics are

$$P_t = P_0 + \sigma W_t + \alpha(Y_t - Y_0) + \beta \dot{Y}_t, \tag{4.28}$$

where Y_t denotes the number of shares held by the large investor at time t, and where the process Y is absolutely continuous with density $\dot{Y}$, i.e.,

$$Y_t = y + \int_0^t \dot{Y}_u\,du.$$

The parameter $\alpha \ge 0$ is a parameter for the level of permanent impact of the large investor's trading, and the parameter $\beta \ge 0$ describes the temporary impact of the large investor's trading. The optimal liquidation strategy for a large investor with an initial position of y number of shares is

$$Y_t^* = y\exp\left(-t\sqrt{\frac{\sigma^2 A}{2\beta}}\right), \qquad t \ge 0$$

(see Schied and Schöneborn 2009), if the large investor has constant absolute risk aversion A and aims to maximize his cash position at the end of time. We can observe that the optimal strategy in the limit order book model and the optimal strategy in the Almgren and Chriss (1999) model look similar, as in both models liquidation follows an exponential function. Yet, there are two aspects which make the strategies different. In the limit order book model, the optimal strategy depends on the past history of the large investor, while in the Almgren model there is no such dependence since future returns are unaffected by the large investor's past trades. Also, in the limit order book model, the optimal strategy typically consist of an initial block trade, while in the Almgren and Chriss model the optimal strategy is absolutely continuous. However, as pointed out by an anonymous referee, we can recover the optimal liquidation strategy for the Almgren and Chriss model by taking the limit as $\lambda \to \infty$ with $c = c^{(\lambda)} = \frac{1}{\beta\lambda}$. Thus, as the resilience rate tends to infinity, the quantity available in the limit order book decreases such that the cost of trading has a finite limit. More specifically, since

$$\lim_{x\to 0^+} \frac{x + \sqrt{x^2 + x}}{\sqrt{x}} = 1,$$

we calculate that

$$\lim_{\lambda\to\infty} \lambda\kappa^{(\lambda)} = \lim_{\lambda\to\infty} \lambda\sqrt{\frac{\sigma^2 A}{2\lambda} c^{(\lambda)}} = \sqrt{\frac{\sigma^2 A}{2\beta}}.$$

In the limit as $\lambda \to \infty$, the state process Z is identically equal to 0. Therefore, if we let $Y^{*,\lambda}$ denote the optimal strategy described in (a) and (b) corresponding to λ, we conclude that

$$\lim_{\lambda\to\infty} Y_t^{*,\lambda} = \lim_{\lambda\to\infty} \frac{y}{\kappa^{(\lambda)} + 1} \exp\left(-t\frac{\lambda\kappa^{(\lambda)}}{1 + \kappa^{(\lambda)}}\right) = y \exp\left(-t\sqrt{\frac{\sigma^2 A}{2\beta}}\right),$$

which is the optimal strategy in the Almgren and Chriss model. To explain this result, we can note that if Y is an admissible liquidation strategy which has the form

$$\dot{Y}_t = \sum_{n=1}^{m} q_n \mathbf{1}_{[\tau_n, \tau_{n+1})}(t)$$

for stopping times $0 \le \tau_1 < \tau_2 \cdots < \tau_{m+1}$ and random variables $q_1, \ldots, q_{m+1}$, where q_n is $\mathcal{F}_{\tau_n}$-measurable, then

$$\begin{aligned}
\lim_{\lambda\to\infty} B_t^Y &= B_t^0 + \lim_{\lambda\to\infty} \frac{Z_t^{Y,\lambda}}{c^{(\lambda)}} \\
&= B_t^0 + \lim_{\lambda\to\infty} \beta\lambda \mathrm{e}^{-\lambda t} \int_0^t \mathrm{e}^{\lambda s} \dot{Y}_s \, ds \\
&= B_t^0 + \sum_{n=1}^{m} \beta q_n \lim_{\lambda\to\infty} \left(\mathrm{e}^{\lambda(\min\{\tau_{n+1}, t\} - t)} - \mathrm{e}^{\lambda(\min\{\tau_n, t\} - t)}\right) \\
&= B_t^0 + \beta \dot{Y}_t,
\end{aligned}$$

which is the Almgren and Chriss model with permanent impact factor α equal to zero.

5. PROOFS OF RESULTS

Proof of Lemma 3.2. With reference to equation (2.12) we calculate

$$
\begin{aligned}
C_T(Y) &= c - \int_0^T B_{t-}^Y \, dY_t^c - \sum_{0\le t\le T} \int_0^{\triangle Y_t} \{B_t^0 + \psi(Z_{t-}^Y + x)\}\, dx \\
&= c - \int_0^T B_{t-}^Y \, dY_t^c - \sum_{0\le t\le T} B_{t-}^Y \triangle Y_t - \sum_{0\le t\le T} \int_0^{\triangle Y_t} \{\psi(Z_{t-}^Y + x) - \psi(Z_{t-}^Y)\}\, dx \qquad (5.1)\\
&= c - \int_0^T B_{t-}^Y \, dY_t - \sum_{0\le t\le T} \int_0^{\triangle Y_t} \{\psi(Z_{t-}^Y + x) - \psi(Z_{t-}^Y)\}\, dx.
\end{aligned}
$$

Since Y is a càdlàg process of finite variation, it follows from theorems II.26 and II.28 in Protter (1990) that the quadratic covariation $[B^Y, Y]$ between B^Y and Y is

$$
[B^Y, Y]_T = \sum_{0\le t\le T} \{\psi(Z_{t-}^Y + \triangle Y_t) - \psi(Z_{t-}^Y)\}\triangle Y_t.
$$

Hence,

$$
\begin{aligned}
B_T^Y Y_T - B_{0-}^Y Y_{0-} = \int_0^T B_{t-}^Y \, dY_t + \int_0^T Y_{t-} \, dB_t^Y \qquad (5.2)\\
+ \sum_{0\le t\le T} \{\psi(Z_{t-}^Y + \triangle Y_t) - \psi(Z_{t-}^Y)\}\triangle Y_t.
\end{aligned}
$$

With reference to the dynamics of B^Y given by (2.9),

$$
\begin{aligned}
\int_0^T Y_{t-}\, dB_t^Y = \int_0^T \sigma Y_{t-}\, dW_t - \int_0^T \lambda Z_{t-}^Y \psi'(Z_{t-}^Y) Y_{t-}\, dt \qquad (5.3)\\
+ \int_0^T \psi'(Z_{t-}^Y) Y_{t-}\, dY_t^c + \sum_{0\le t\le T} Y_{t-}\{\psi(Z_{t-}^Y + \triangle Y_t) - \psi(Z_{t-}^Y)\}.
\end{aligned}
$$

Equation (5.2) provides an expression for $\int_0^T B_{t-}^Y \, dY_t$, which combined with (5.1) and (5.3) imply

$$
\begin{aligned}
C_T(Y) = c - B_T^Y Y_T + B_{0-}^Y Y_{0-} + \int_0^T \sigma Y_{t-}\, dW_t - \int_0^T \lambda Z_{t-}^Y \psi'(Z_{t-}^Y) Y_{t-}\, dt \\
+ \int_0^T \psi'(Z_{t-}^Y) Y_{t-}\, dY_t^c + \sum_{0\le t\le T} Y_{t-}\{\psi(Z_{t-}^Y + \triangle Y_t) - \psi(Z_{t-}^Y)\} \qquad (5.4)\\
- \sum_{0\le t\le T} \int_0^{\triangle Y_t} \{\psi(Z_{t-}^Y + x) - \psi(Z_{t-}^Y)\}\, dx \\
+ \sum_{0\le t\le T} \{\psi(Z_{t-}^Y + \triangle Y_t) - \psi(Z_{t-}^Y)\}\triangle Y_t.
\end{aligned}
$$

Define the function $f : \mathbb{R}^+ \times \mathbb{R}^- \to \mathbb{R}^2$ by

$$
f(y, z) = y\psi(z) + \int_z^{z-y} \psi(s)\, ds.
$$

Then by Itô's formula,

$$
\begin{aligned}
f\big(Y_T, Z_T^Y\big) = f\big(Y_{0-}, Z_{0-}^Y\big) &+ \int_0^T \psi'\big(Z_{t-}^Y\big)\, Y_{t-}\, dY_t^c \\
&- \int_0^T \lambda Z_{t-}^Y \psi'\big(Z_{t-}^Y\big)\, Y_{t-}\, dt - \int_0^T \lambda Z_{t-}^Y \big\{\psi\big(Z_{t-}^Y - Y_{t-}\big) - \psi\big(Z_{t-}^Y\big)\big\}\, dt \\
&+ \sum_{0\le t\le T} Y_{t-}\big\{\psi\big(Z_{t-}^Y + \triangle Y_t\big) - \psi\big(Z_{t-}^Y\big)\big\} + \sum_{0\le t\le T} \psi\big(Z_{t-}^Y + \triangle Y_t\big)\triangle Y_t \\
&+ \sum_{0\le t\le T} \int_{Z_{t-}^Y + \triangle Y_t}^{Z_{t-}^Y} \psi(s)\, ds.
\end{aligned}
$$

This provides an expression for $\int_0^T \psi'(Z_{t-}^Y)\, Y_t\, dY_t^c$, which inserted in (5.4) implies that

$$
\begin{aligned}
C_T(Y) = c - B_T^Y Y_T + B_{0-}^Y Y_{0-} &+ f\big(Y_T, Z_T^Y\big) - f\big(Y_{0-}, Z_{0-}^Y\big) \\
&+ \int_0^T \sigma Y_{t-}\, dW_t + \int_0^T \lambda Z_{t-}^Y \big\{\psi\big(Z_{t-}^Y - Y_{t-}\big) - \psi\big(Z_{t-}^Y\big)\big\}\, dt,
\end{aligned}
$$

from which the result follows. □

Proof of Lemma 3.3. With reference to the expression for the large investor's cash position at time $T > 0$ obtained in Lemma 3.2, we first note that $z - y \le Z_t^Y \le 0$, for all $t \ge 0$. Therefore,

$$
\lim_{T\to\infty} \left(Y_T Z_T^Y + \int_{Z_T^Y}^{Z_T^Y - Y_T} \psi(s)\, ds \right) = 0
$$

almost surely and in $L^1(\mathbb{P})$. By the Cauchy–Schwartz inequality, we calculate that

$$
\lim_{T\to\infty} \mathbb{E}\big[\big|B_T^Y Y_T\big|\big] \le \sigma \left(\lim_{T\to\infty} \mathbb{E}\big[T Y_T^2\big] \right)^{1/2} = 0, \tag{5.5}
$$

since (2.2) implies

$$
\lim_{t\to\infty} \mathbb{E}\big[t Y_t^2\big] = 0. \tag{5.6}
$$

The convergence of $\int_0^\infty \sigma Y_{t-}\, dW_t$ follows from (2.2) and the Itô isometry. In order to establish the inequalities (3.7), observe that $B^0 \ge B^Y$, for every $(X, Y) \in \mathcal{A}(y)$. Since Y is decreasing, it follows that

$$
C_T(Y) \le c - \int_0^T B_t^0\, dY_t = c + B_{0-}^0 Y_{0-} - B_T^0 Y_T + \int_0^T \sigma Y_{t-}\, dW_t.
$$

By similar arguments as in (5.5), $B_T^0 Y_T$ converges to 0 in $L^1(\mathbb{P})$ as $T \to \infty$. Since

$$
\lambda Z_{t-}^Y \big\{\psi\big(Z_{t-}^Y - Y_{t-}\big) - \psi\big(Z_{t-}^Y\big)\big\} \ge 0, \qquad t \ge 0,
$$

we conclude that the inequalities (3.7) hold and that the integral

$$
\int_0^\infty \lambda Z_{t-}^Y \big\{\psi\big(Z_{t-}^Y - Y_{t-}\big) - \psi\big(Z_{t-}^Y\big)\big\}\, dt
$$

is almost surely convergent. □

Proof of Lemma 4.1. Let $\mathbb{Z}(z)$ be the set of continuous functions $Z : \mathbb{R}^+ \to [z - h^{-1}(z), 0]$ satisfying $Z_0 = z$ and $\lim_{t\to\infty} Z_t = 0$, where $z \in \mathbb{R}^-$. We consider $\mathbb{Z}(z)$ as a subset of $C_0(\mathbb{R})$, i.e., the space of all real-valued continuous functions vanishing at infinity, equipped with the uniform topology. Observe that $\mathbb{Z}(z)$ is a convex set, and by the Ascoli–Arzelà theorem (see, e.g., Folland 1984) it follows that $\mathbb{Z}(z)$ is compact. Introduce the function Ψ_z given by

$$(\Psi_z(Z))_t = \begin{cases} z - h^{-1}(z) - \int_0^t \lambda Z_u\,du + h^{-1}(Z_t), & \text{for } 0 \le t \le \bar{t}, \\ 0, & \text{for } t > \bar{t}, \end{cases}$$

where

$$\bar{t} = \inf\left\{ t \ge 0 \,:\, z - h^{-1}(z) - \int_0^t \lambda Z_u\,du + h^{-1}(Z_t) = 0 \right\}.$$

Since h^{-1} is continuous and decreasing, it follows that $\Psi_z : \mathbb{Z}(z) \to \mathbb{Z}(z)$ is continuous. The Schauder–Tychonoff fixed point theorem (see, e.g., Rudin 1991) therefore guarantees the existence of a $Z \in \mathbb{Z}(z)$ such that $Z = \Psi_z(Z)$. We want to show that such a Z is unique. Assume that $Z^{(1)} = \Psi_z(Z^{(1)})$ and $Z^{(2)} = \Psi_z(Z^{(2)})$, where $Z^{(1)}, Z^{(2)} \in \mathbb{Z}(z)$ and $Z_t^{(1)} = Z_t^{(2)}$ for $0 \le t \le t_1$, and $Z_t^{(1)} < Z_t^{(2)}$ for $t_1 < t < t_2$. Then for $t_1 < t < t_2$,

$$\begin{aligned} Z_t^{(1)} &= z - h^{-1}(z) - \int_0^t \lambda Z_u^{(1)} + h^{-1}\big(Z_t^{(1)}\big) \\ &> z - h^{-1}(z) - \int_0^t \lambda Z_u^{(2)} + h^{-1}\big(Z_t^{(2)}\big) \\ &= Z_t^{(2)}, \end{aligned}$$

which contradicts the assumption that $Z_t^{(1)} < Z_t^{(2)}$ for $t_1 < t < t_2$. We conclude that there exists a unique $Z \in \mathbb{Z}(z)$ such that $Z = \Psi_z(Z)$. Moreover, since the function $z \mapsto z - h^{-1}(z)$ is strictly increasing for $z \in \mathbb{R}^-$ and $t \mapsto z - h^{-1}(z) - \int_0^t \lambda Z_u\,du$ is strictly increasing as long as $Z_t < 0$, it follows that the solution to $Z = \Psi_z(Z)$ is strictly increasing while $Z_t < 0$.

Suppose that $z \ge h(y)$, and let Z^{Y^h} be given by (4.19) and (4.20). The existence and uniqueness of such a Z^{Y^h} follows from the previous part of the proof. We calculate that

$$h^{-1}\big(\xi + \gamma_h^{-1}(\xi)\big) = \gamma_h^{-1}(\xi), \qquad \text{for all } \xi \in \mathbb{R}^-,$$

and therefore

$$h^{-1}\big(Z_0^{Y^h}\big) = h^{-1}\big(z - y + \gamma_h^{-1}(z-y)\big) = \gamma_h^{-1}(z-y) = Y_0^h,$$

as required. With reference to (4.18) and (4.19), we have that

$$\begin{aligned} Z_t^{Y^h} &= Z_0^{Y^h} - h^{-1}\big(Z_0^{Y^h}\big) - \int_0^t \lambda Z_u^{Y^h}\,du + h^{-1}\big(Z_t^{Y^h}\big) \\ &= Z_0^{Y^h} - \int_0^t \lambda Z_u^{Y^h}\,du + Y_t^h - Y_0^h, \end{aligned} \tag{5.7}$$

since $z \geq h(Y)$ is equivalent to $t_w = 0$, and $Y_t^h = h^{-1}(Z_t^{Y^h})$. Equation (5.7) shows that Z^{Y^h} satisfies (2.8), and we conclude that Y^h is the unique process with the property that $Y_t^h = h^{-1}(Z_t^{Y^h})$, for all $t \geq 0$. Also note that from the first part of the proof, $t \mapsto Z_t^{Y^h}$ is continuous and decreasing for $t > 0$, which combined with the monotonicity and continuity of h^{-1} imply that Y^h is càdlàg and decreasing. Hence, Y^h is the unique decreasing càdlàg process satisfying the description given in part (a).

Suppose that $z < h(y)$, and let Y^h be given by (4.18)–(4.20). Then

$$Z_t^{Y^h} = z\mathrm{e}^{-\lambda t}, \qquad \text{for } 0 \leq t \leq t_w, \tag{5.8}$$

and $Z_{t_w}^{Y^h} = h(y) = \lim_{t \to t_w} Z_t^{Y^h}$, which correspond to the description given in part (b). With reference to (4.18)–(4.20), we have that

$$\begin{aligned} Z_t^{Y^h} &= h(y) - y - \int_{t_w}^t \lambda Z_u^{Y^h}\, du + h^{-1}(Z_t^{Y^h}) \\ &= h(y) - \int_{t_w}^t \lambda Z_u^{Y^h}\, du + Y_t^h - Y_0^h \end{aligned}$$

for $t \geq t_w$. Since $Y_t^h = y = h^{-1}(h(y))$ and Z^{Y^h} is given by (5.8) for $0 \leq t < t_w$, it follows that

$$Z_t^{Y^h} = z - \int_0^t \lambda Z_u^{Y^h}\, du + Y_t^h - Y_0^h$$

for $t \geq 0$, which verifies that Z^{Y^h} satisfies (2.8), and $Y_t^h = h^{-1}(Z_t^{Y^h})$, for all $t \geq t_w$. We conclude that Y^h is the unique decreasing càdlàg process as described in part (b). □

Proof of Lemma 4.2. Let $t_w^{h_1}$ and $t_w^{h_2}$ be given by (4.17), corresponding to h_1 and h_2, respectively. If $z < h_1(y)$ then $Y_t^{h_1} = Y_t^{h_2}$ for $0 \leq t \leq t_w^{h_1}$ and $Y_t^{h_1} \leq Y_t^{h_2}$ for $t_w^{h_1} \leq t \leq t_w^{h_2}$. If $h_1(y) \leq z < h_2(y)$ then $Y_t^{h_1} \leq Y_t^{h_2}$ for $0 = t_w^{h_1} \leq t \leq t_w^{h_2}$. Also, if $z \geq h_2(y)$, then $t_w^{h_1} = t_w^{h_2} = 0$ and $Y_0^{h_1} \leq Y_0^{h_2}$. We want to show that $\{t \geq 0 \,:\, Y_t^{h_1} > Y_t^{h_2}\} = \emptyset$. In order to get a contradiction, suppose that

$$t_1 = \inf\left\{t \geq 0 \,:\, Y_t^{h_1} > Y_t^{h_2}\right\} < \infty \quad \text{ and define } t_2 = \inf\left\{t \wedge \infty \geq t_1 \,:\, Y_t^{h_1} \leq Y_t^{h_2}\right\}.$$

By the previous observations $t_1 > t_w$, and by continuity of $Y_t^{h_1}$ and $Y_t^{h_2}$, for $t > 0$, it follows that $t_1 < t_2$. The monotonicity of h_1 and h_2 imply that if $Y_t^{h_1} > Y_t^{h_2}$ then

$$Z_t^{Y^{h_1}} \leq h_1\big(Y_t^{h_1}-\big) < h_1\big(Y_t^{h_2}\big) \leq h_2\big(Y_t^{h_2}\big) \leq Z_t^{Y^{h_2}}, \tag{5.9}$$

and if $Z_t^{Y^{h_1}} < Z_t^{Y^{h_2}}$ then

$$Y_t^{h_1} = h_1^{-1}\big(Z_t^{Y^{h_1}}\big) > h_1^{-1}\big(Z_t^{Y^{h_2}}\big) \geq h_2^{-1}\big(Z_t^{Y^{h_2}}\big) = Y_t^{h_2}. \tag{5.10}$$

With reference to (4.18) and (4.19), we have that for $t_1 < t < t_2$,

$$\begin{aligned} Z_t^{Y^{h_1}} &= Z_{t_1}^{Y^{h_1}} - Y_{t_1}^{h_1} - \int_{t_1}^{t} \lambda Z_u^{Y^{h_1}}\,du + h_1^{-1}\big(Z_t^{Y^{h_1}}\big) \\ &\geq Z_{t_1}^{Y^{h_1}} - Y_{t_1}^{h_1} - \int_{t_1}^{t} \lambda Z_u^{Y^{h_1}}\,du + h_2^{-1}\big(Z_t^{Y^{h_1}}\big) \\ &> Z_{t_1}^{Y^{h_1}} - Y_{t_1}^{h_1} - \int_{t_1}^{t} \lambda Z_u^{Y^{h_2}}\,du + h_2^{-1}\big(Z_t^{Y^{h_2}}\big) \\ &\geq Z_{t_1}^{Y^{h_2}} - Y_{t_1}^{h_2} - \int_{t_1}^{t} \lambda Z_u^{Y^{h_2}}\,du + h_2^{-1}\big(Z_t^{Y^{h_2}}\big) \\ &= Z_t^{Y^{h_2}}, \end{aligned} \tag{5.11}$$

where the last equality is due to $Y_{t_1}^{h_1} = Y_{t_1}^{h_2}$, (5.9) and (5.10), which imply that $Z_{t_1}^{Y^{h_1}} \geq Z_{t_1}^{Y^{h_2}}$. However, in view of (5.9) and (5.10), the inequality (5.11) contradicts the definition of t_1. Thus, we conclude that $Y^{h_1} \leq Y^{h_2}$.

Let $\bar{h} : \mathbb{R}^+ \to \mathbb{R}^-$ be given by $\bar{h}(y) = -Cy$, for $C > 0$, and let $Y^{\bar{h}}$ denote the corresponding strategy given by (4.18)–(4.20) in Lemma 4.1. Then (4.19) takes the form

$$Z_t^{Y^{\bar{h}}} = Z_{t_w}^{Y^{\bar{h}}} + \frac{1}{C} Z_{t_w}^{Y^{\bar{h}}} - \int_{t_w}^{t} \lambda Z_u^{Y^{\bar{h}}}\,du - \frac{1}{C} Z_t^{Y^{\bar{h}}},$$

which has a unique solution

$$Z_t^{Y^{\bar{h}}} = Z_{t_w}^{Y^{\bar{h}}} \exp\left(-t \frac{\lambda C}{1+C}\right), \qquad \text{for } t \geq t_w.$$

Also, (4.20) takes the form

$$Z_{t_w}^{Y^{\bar{h}}} = -Cy, \text{ if } z < -Cy \quad \text{and} \quad Z_{t_w}^{Y^{\bar{h}}} = (z-y)\frac{C}{1+C}, \text{ if } z \geq -Cy.$$

Therefore, the strategy $Y^{\bar{h}}$ is given by

$$Y_t^{\bar{h}} = -\frac{z-y}{1+C} \exp\left(-t \frac{\lambda C}{1+C}\right), \quad \text{for } t \geq 0, \quad \text{if } z \geq -Cy, \tag{5.12}$$

and

$$Y_t^{\bar{h}} = y \exp\left(-(t-t_w)\frac{\lambda C}{1+C}\right), \quad \text{for } t \geq t_w, \quad \text{if } z < -Cy, \tag{5.13}$$

where $t_w = \lambda^{-1}\{\ln(-z) - \ln(Cy)\}$, if $z < -Cy$. Since $t_w < \infty$, for all $(y, z) \in \mathbb{R}^+ \times \mathbb{R}^-$ and the right-hand side of (5.12) and (5.13) are square integrable, it follows that $Y^{\bar{h}} \in \mathcal{A}_D^-(y)$, for all initial positions $Y_{0-}^{\bar{h}} = y$.

Assume that there exists $C > 0$ and $\epsilon > 0$ such that $h(y) \leq -Cy$, for $0 \leq y < \epsilon$. Then, for every $y_0 \in \mathbb{R}^+$, there exists $C_{y_0} > 0$ such that $h(y) \leq -C_{y_0} y$, for all $0 \leq y \leq y_0$. Also observe that the strategy Y^h given by (4.18)–(4.20) in Lemma 4.1, with initial position $Y_{0-}^h = y_0 \in \mathbb{R}^+$, is completely determined by the values of $h(y)$ for $0 \leq y \leq y_0$. Therefore, if h_1 and h_2 are two functions satisfying $h_1(y) = h_2(y)$, for $0 \leq y \leq y_0$, then $Y^{h_1} = Y^{h_2}$ if the initial position $Y_{0-}^{h_1} = Y_{0-}^{h_2} = y$ is less than or equal to y_0. With reference to the previous parts of the proof, we therefore conclude that $Y^h \in \mathcal{A}_D^-(y)$ for all initial positions $Y_{0-}^h = y$. □

Proof of Proposition 4.3. First note that for $y > 0$, the properties of ψ given in (2.4)–(2.6) imply that

$$\lim_{x\to-\infty} \Gamma(x; y) = -\infty \quad \text{and} \quad \lim_{x\to 0} \Gamma(x; y) = -\infty.$$

Moreover, $\Gamma(x; y)$ is continuously differentiable in x and y, for all $x < 0$ and $y > 0$. We conclude that $h = h(y)$ defined as the smallest solution to (4.24) is well defined and that h must satisfy (4.25). Moreover, ψ is strictly increasing and $\Gamma(x; 0) = \psi(x)$, which implies that $h(0) = 0$. Let $\bar{h} = \bar{h}(y)$ denote the largest solution to (4.25), and define $L : \mathbb{R}^+ \times \mathbb{R}^- \to \mathbb{R}^+$ by $L(h, y) = ay^2 - 2ayh$ and $H : \mathbb{R}^- \to \mathbb{R}^+$ by $H(h) = \psi'(h)h^2$. Since H is continuous and $\lim_{y\to\infty} L(h, y) = \infty$, for all $h \in \mathbb{R}^-$, it follows that $\lim_{y\to\infty} \bar{h}(y) = -\infty$. Since $h \le \bar{h}$, we conclude that $\lim_{y\to\infty} h(y) = -\infty$.

For $\triangle > 0$, $y \in \mathbb{R}^+$ and $x \in \mathbb{R}^-$, we calculate that

$$\Gamma(x; y + \triangle) - \Gamma(x; y) = \int_y^{y+\triangle} 2a\left\{\frac{u}{x} + \ln(-x)\right\} du,$$

and

$$\frac{d}{dx}[\Gamma(x; y + \triangle) - \Gamma(x; y)] = \int_y^{y+\triangle} 2a\left\{\frac{1}{x} - \frac{u}{x^2}\right\} du < 0. \tag{5.14}$$

We want to show that $h(y)$ is strictly decreasing as a function of y. In order to get a contradiction, suppose that there exists $y \in \mathbb{R}^+$ and $\triangle > 0$ such that $h(y + \triangle) \ge h(y)$. With reference to (5.14), this implies that

$$\Gamma(h(y + \triangle); y + \triangle) - \Gamma(h(y + \triangle); y) < \Gamma(h(y); y + \triangle) - \Gamma(h(y); y).$$

However, this contradicts the definition of h, which implies that

$$\Gamma(h(y + \triangle); y + \triangle) \ge \Gamma(h(y); y + \triangle) \quad \text{and} \quad \Gamma(h(y); y) \ge \Gamma(h(y + \triangle); y).$$

We conclude that h is strictly decreasing. The definition of $h = h(y)$ as the smallest solution to (4.24) implies that h is càdlàg.

Introduce the function $Q : \mathbb{R}^+ \times \mathbb{R}^- \to \mathbb{R}$ given by

$$\begin{aligned} Q(y, z) = &\int_0^{z-y} \left(\frac{a\big(\gamma_h^{-1}(s)\big)^2}{\rho_h^{-1}(s)} + \psi\big(\rho_h^{-1}(s)\big) - \psi(s)\right) ds \\ &- \left\{ay^2 \ln(-z) + \int_{z-y}^{z} \psi(s)\, ds - \int_0^y \left(\frac{as^2}{h(s)} + \psi(h(s)) + 2as\ln(-h(s))\right) ds\right\}, \end{aligned}$$

which is the difference between the expression for v given by (4.26) and the expression for v given by (4.27). We calculate that

$$Q_z(y, z) = \frac{a\big(\gamma_h^{-1}(z-y)\big)^2}{\rho_h^{-1}(z-y)} + \psi\big(\rho_h^{-1}(z-y)\big) - \psi(z-y) - \left\{\frac{ay^2}{z} + \psi(z) - \psi(z-y)\right\},$$

which we observe is a continuous function of (y, z). Moreover, in view of the observation

$$\gamma_h^{-1}(z - h^{-1}(z)) = \sup\{y \geq 0 : \gamma_h(y) \geq z - h^{-1}(z)\} = h^{-1}(z),$$

it follows that $Q_z(h^{-1}(z), z) = 0$, for all $z \in \mathbb{R}^-$. We further calculate that

$$\begin{aligned} D_y^+ Q(y, z) = &-\frac{a\big(\gamma_h^{-1}(z - y)\big)^2}{\rho_h^{-1}(z - y)} - \psi\big(\rho_h^{-1}(z - y)\big) + \psi(z - y) \\ &- \left\{-\frac{ay^2}{h(y)} + \psi(z - y) - \psi(h(y)) + 2ay\{\ln(-z) - \ln(-h(y))\}\right\}, \end{aligned}$$

and observe that the function $D_y^+ Q(y, z)$ is continuous in z and càdlàg in y. Since $\rho_h^{-1}(h(y) - y) = h(y)$, it follows that $D_y^+ Q(y, h(y)) = 0$, for all $y \in \mathbb{R}^+$. Moreover, $Q(0, 0) = 0$ and hence

$$Q(y, h(y)) = \int_0^y Q_z(u, h(u))\, dh(u) + \int_0^y D_y^+ Q(u, h(u))\, du = 0, \qquad \text{for all } y \in \mathbb{R}^+,$$

and we conclude that

$$Q\big(h^{-1}(z), z\big) = 0, \qquad \text{for all } z \in \mathbb{R}^-,$$

since $Q_z(h^{-1}(z), z) = 0$, for all $z \in \mathbb{R}^-$ implies that $Q_z(y, h(y)) = 0$, for all $y \in \mathbb{R}^+$. From the properties of the function Q given above, we conclude that $v \in C^{0,1}(\mathbb{R}^+ \times \mathbb{R}^-)$ and that $D_y^+ v(y, z)$ is continuous in z and càdlàg in y. Moreover, straightforward calculations show that v is continuously differentiable with respect to y for $z \geq h(y)$.

Standard calculations show that v satisfies (4.8) and (4.10). In order to verify that v satisfies (4.9), we calculate that

$$\begin{aligned} &zv_z(y, z) - ay^2 - z\{\psi(z) - \psi(z - y)\} \\ &\quad = -ay^2 + z\left\{\frac{a\big(\gamma_h^{-1}(z - y)\big)^2}{\rho_h^{-1}(z - y))} + \psi\big(\rho_h^{-1}(z - y)\big) - \psi(z)\right\} \end{aligned} \tag{5.15}$$

for $z \geq h(y)$. Set $s = \rho_h^{-1}(z - y)$, which is equivalent to $z - y = \rho_h(s)$. As ρ_h^{-1} is an increasing function and $y \geq h^{-1}(z)$ is equivalent to $z - y \leq \rho_h(z)$, it follows that $y \geq h^{-1}(z)$ if and only if $s \leq z$. Since $\gamma_h^{-1}(z - y) = \gamma_h^{-1}(\rho_h(s)) = h^{-1}(s)$, it follows from (5.15) that

$$\sup_{y \geq h^{-1}(z)} \{zv_z(y, z) - ay^2 - z\{\psi(z) - \psi(z - y)\}\} = z \inf_{s \leq z} G(s; z), \tag{5.16}$$

where

$$G(s; z) = \frac{a(h^{-1}(s))^2}{s} - \frac{a(z - \rho_h(s))^2}{z} + \psi(s) - \psi(z).$$

In particular, observe that $G(z;z) = 0$. We calculate that

$$\begin{aligned}
0 &= G(s;z) + \int_s^z -\frac{a(h^{-1}(u))^2}{u^2}\,du + \int_s^z \frac{2ah^{-1}(u)}{u}\,dh^{-1}(u)\\
&\quad - \int_s^z \frac{2a}{z}\{z-u+h^{-1}(u)\}\,d(h^{-1}(u)-u) - \int_s^z \psi'(u)\,du\\
&= G(s;z) - \int_s^z \left(\psi'(u) + \frac{a(h^{-1}(u))^2}{u^2} - \frac{2ah^{-1}(u)}{u}\right) du\\
&\quad + \int_s^z 2a\left\{h^{-1}(u)\left(\frac{1}{u}-\frac{1}{z}\right) + \left(\frac{u}{z}-1\right)\right\} d(h^{-1}(u)-u)\\
&= G(s;z) - \int_s^{h(h^{-1}(s)-)} \left(\psi'(u) + \frac{a(h^{-1}(s))^2}{u^2} - \frac{2ah^{-1}(s)}{u}\right) du\\
&\quad - \int_{h(h^{-1}(z))}^z \left(\psi'(u) + \frac{a(h^{-1}(z))^2}{u^2} - \frac{2ah^{-1}(z)}{u}\right) du\\
&\quad + \int_s^z 2a\left\{h^{-1}(u)\left(\frac{1}{u}-\frac{1}{z}\right) + \left(\frac{u}{z}-1\right)\right\} d(h^{-1}(u)-u),
\end{aligned}\tag{5.17}$$

since $h^{-1}(u)$ is constant for $h(h^{-1}(\xi)) \le u \le h(h^{-1}(\xi)-)$, for any $\xi \in \mathbb{R}^-$ and

$$\int_{h(h^{-1}(s)-)}^{h(h^{-1}(z))} \left(\psi'(u) + \frac{a(h^{-1}(s))^2}{u^2} - \frac{2ah^{-1}(s)}{u}\right) du = 0$$

for any $s \le z \le 0$, by the definition of h and the continuity of Γ, which implies that $\Gamma(h(h^{-1}(s)-); h^{-1}(s)) = \Gamma(h(h^{-1}(s)); h^{-1}(s))$, for every $s \in \mathbb{R}^-$. Also observe that

$$-\int_s^{h(h^{-1}(s)-)} \left(\psi'(u) + \frac{a(h^{-1}(s))^2}{u^2} - \frac{2ah^{-1}(s)}{u}\right) du = \Gamma(s; h^{-1}(s)) - \Gamma\big(h(h^{-1}(s)); h^{-1}(s)\big),$$

which is negative by the optimality of h. Since $s \le z$ and $u \mapsto h^{-1}(u) - u$ is strictly decreasing, it follows from (5.17) that

$$\begin{aligned}
0 &\le G(s;z) - \int_{h(h^{-1}(z))}^z \left(\psi'(u) + \frac{a(h^{-1}(z))^2}{u^2} - \frac{2ah^{-1}(z)}{u}\right) du\\
&\quad + \int_{h(h^{-1}(z))}^z 2a\left\{h^{-1}(u)\left(\frac{1}{u}-\frac{1}{z}\right) + \left(\frac{u}{z}-1\right)\right\} d(h^{-1}(u)-u)\\
&= G(s;z) + \psi(h(h^{-1}(z))) - \psi(z)\\
&\quad + \int_{h(h^{-1}(z))}^z \left\{2a\left(\frac{h^{-1}(z)}{z}+1\right) - \frac{a(h^{-1}(z))^2}{u^2} - \frac{2au}{z}\right\} du\\
&\le G(s;z),
\end{aligned}\tag{5.18}$$

since ψ is strictly increasing, $h(h^{-1}(z)) \le z$ and

$$2a\left(\frac{h^{-1}(z)}{z}+1\right) - \frac{a(h^{-1}(z))^2}{u^2} - \frac{2au}{z} \le \frac{2ah^{-1}(z)}{z} < 0, \qquad \text{for } u \le z.$$

With reference to (5.16) and (5.18), we conclude that v satisfies (4.9). Finally, we need to show that v satisfies (4.11). We calculate that

$$\begin{aligned} D_y^+ v(y,z) + v_z(y,z) &= ay^2\left(\frac{1}{z} - \frac{1}{h(y)}\right) + \psi(z) - \psi(h(y)) + 2ay\ln\left(\frac{z}{h(y)}\right) \\ &= \Gamma(z;y) - \Gamma(h(y);y) \\ &\leq 0 \end{aligned} \tag{5.19}$$

by the definition of h. □

Proof of Theorem 4.4. Let δ be a nonnegative $C^\infty(\mathbb{R})$ function with support in [0, 1] satisfying $\int_0^1 \delta(x)\,dx = 1$, and define a sequence of functions $\{\delta_n\}_{n=1}^\infty$ by

$$\delta_n(s) = n\,\delta(ns), \quad s \geq 0.$$

We modify v to obtain a sequence of function $\{v^{(n)}\}_{n=1}^\infty$, given by

$$v^{(n)}(y,z) = \int_0^1 v(y+s,z)\,\delta_n(s)\,ds.$$

Then, $v^{(n)} \in C^{1,1}(\mathbb{R}^+ \times \mathbb{R}^-)$, for all $n \in \mathbb{N}$, and

$$\begin{aligned} v(y,z) &= \lim_{n\to\infty} v^{(n)}(y,z), \\ v_z(y,z) &= \lim_{n\to\infty} v_z^{(n)}(y,z), \\ D_y^+ v(y,z) &= \lim_{n\to\infty} v_y^{(n)}(y,z), \end{aligned}$$

where the last equality is due to $D_y^+ v(y,z)$ being càdlàg in y. Moreover, for every $(y_0, z_0) \in \mathbb{R}^+ \times \mathbb{R}^-$ there exists a $K > 0$ such that

$$|v^{(n)}(y,z)| \leq K, \quad 0 \leq y \leq y_0,\ z_0 - y_0 \leq z \leq 0 \text{ and } n \in \mathbb{N}, \tag{5.20}$$

$$|v_z^{(n)}(y,z)| \leq K, \quad 0 \leq y \leq y_0,\ z_0 - y_0 \leq z \leq 0 \text{ and } n \in \mathbb{N}, \tag{5.21}$$

$$|v_y^{(n)}(y,z)| \leq K, \quad 0 \leq y \leq y_0,\ z_0 - y_0 \leq z \leq 0 \text{ and } n \in \mathbb{N}. \tag{5.22}$$

Then

$$\begin{aligned} & v^{(n)}\big(Y_T, Z_T^Y\big) + \int_0^T \big\{\lambda a Y_{t-}^2 + \lambda Z_{t-}^Y\big\{\psi\big(Z_{t-}^Y\big) - \psi\big(Z_{t-}^Y - Y_{t-}\big)\big\}\big\}\,dt \\ &\quad = v^{(n)}(y,z) + \int_0^T \big\{v_y^{(n)}\big(Y_{t-}, Z_{t-}^Y\big) + v_z^{(n)}\big(Y_{t-}, Z_{t-}^Y\big)\big\}\,dY_t^c \\ &\qquad + \sum_{0\leq t\leq T} \big\{v^{(n)}\big(Y_{t-} + \triangle Y_t, Z_{t-}^Y + \triangle Y_t\big) - v^{(n)}\big(Y_{t-}, Z_{t-}^Y\big)\big\} \\ &\qquad -\lambda \int_0^T \big\{Z_{t-}^Y v_z^{(n)}\big(Y_{t-}, Z_{t-}^Y\big) - a Y_{t-}^2 - Z_{t-}^Y\big\{\psi\big(Z_{t-}^Y\big) - \psi\big(Z_{t-}^Y - Y_{t-}\big)\big\}\big\}\,dt \end{aligned} \tag{5.23}$$

for all $Y \in \mathcal{A}_D^-(y)$. For every $\epsilon > 0$ and $Y \in \mathcal{A}_D^-(y)$ there exists $t_0 \in \mathbb{R}^+$ such that $Y_t \leq \epsilon$, for all $t \geq t_0$. Therefore,

$$|Z_t^Y| \leq \left|Z_{t_0}^Y \mathrm{e}^{-\lambda(t-t_0)}\right| + \left|\int_{t_0}^{t} \mathrm{e}^{-\lambda(t-s)}\, dY_s\right| \leq |Z_{t_0}^Y|\, \mathrm{e}^{-\lambda(t-t_0)} + \epsilon,$$

from which it follows that Z_t^Y tends to 0 as $t \to \infty$. With reference to (2.8), we therefore conclude that

$$\int_0^\infty |Z_t^Y|\, dt = \frac{|z-y|}{\lambda}. \tag{5.24}$$

With reference to (5.21), (5.22), and (5.24), we calculate that

$$\int_0^\infty \sup_{n\in\mathbb{N}} \left|Z_{t-}^Y\left(v_z^{(n)}\left(Y_{t-}, Z_{t-}^Y\right) - \left\{\psi\left(Z_{t-}^Y\right) - \psi\left(Z_{t-}^Y - Y_{t-}\right)\right\}\right)\right| dt \leq \frac{K + C_1}{\lambda}|z-y|$$

for some constant $C_1 > 0$ which may depend on the initial conditions y and z. Similarly,

$$\int_0^\infty \sup_{n\in\mathbb{N}} \left|v_y^{(n)}\left(Y_{t-}, Z_{t-}^Y\right) + v_z^{(n)}\left(Y_{t-}, Z_{t-}^Y\right)\right| d(-Y_t^c) \leq 2Ky$$

and

$$\sum_{0\leq t\leq\infty} \sup_{n\in\mathbb{N}} \left|v^{(n)}\left(Y_{t-} + \triangle Y_t, Z_{t-}^Y + \triangle Y_t\right) - v^{(n)}\left(Y_{t-}, Z_{t-}^Y\right)\right| \leq 2Ky.$$

Hence, by (5.23) and the dominated convergence theorem, we obtain that

$$\begin{aligned} &\int_0^\infty \left\{\lambda a Y_{t-}^2 + \lambda Z_{t-}^Y\left\{\psi\left(Z_{t-}^Y\right) - \psi\left(Z_{t-}^Y - Y_{t-}\right)\right\}\right\} dt \\ &\quad = v(y,z) + \int_0^\infty \left\{D_y^+ v\left(Y_{t-}, Z_{t-}^Y\right) + v_z\left(Y_{t-}, Z_{t-}^Y\right)\right\} dY_t^c \\ &\qquad + \sum_{t\geq 0} \left\{v\left(Y_{t-} + \triangle Y_t, Z_{t-}^Y + \triangle Y_t\right) - v\left(Y_{t-}, Z_{t-}^Y\right)\right\} \\ &\qquad - \lambda \int_0^\infty \left\{Z_{t-}^Y v_z\left(Y_{t-}, Z_{t-}^Y\right) - a Y_{t-}^2 - Z_{t-}^Y\left\{\psi\left(Z_{t-}^Y\right) - \psi\left(Z_{t-}^Y - Y_{t-}\right)\right\}\right\} dt \end{aligned} \tag{5.25}$$

for any $Y \in \mathcal{A}_D^-(y)$, by taking the limits as $n \to \infty$ and $T \to \infty$, and noting that $v(Y_T, Z_T^Y)$ tends to 0 as $T \to \infty$ due to the boundary condition $v(0,z) = 0$. Since according to Proposition 4.3, v satisfies (4.8)–(4.11), it follows that

$$\int_0^\infty \left\{\lambda a Y_{t-}^2 + \lambda Z_{t-}^Y\left\{\psi\left(Z_{t-}^Y\right) - \psi\left(Z_{t-}^Y - Y_{t-}\right)\right\}\right\} dt \geq v(y,z), \tag{5.26}$$

and thus $V \geq v$.

With reference to Assumption 2.3 and (2.6), it follows that there exists $C > 0$ and $\epsilon > 0$ such that $h(y) \geq -Cy$, for $-\epsilon \leq y \leq 0$, where h denotes the smallest solution to (4.24). According to Lemma 4.2, we therefore have $Y^* = Y^h \in \mathcal{A}_D^-(y)$, for all initial positions $Y_{0-}^* = y$. We want to show that (5.26) holds with equality for Y^*. Observe that $\triangle Y_t^* < 0$ only if $t = 0$ and $z > h(y)$, in which case $\triangle Y_0^*$ is such that after the jump, $Y_0^* = h^{-1}(Z_0^{Y^*})$.

With reference to (4.8) and Proposition 4.3, we have that $v_y(y,z)+v_z(y,z)=0$, for $z\geq h(y)$. Therefore,

$$\sum_{t\geq 0}\big\{v\big(Y^*_{t-}+\triangle Y^*_t,\, Z^{Y^*}_{t-}+\triangle Y^*_t\big)-v\big(Y^*_{t-},\, Z^{Y^*}_{t-}\big)\big\}=0. \tag{5.27}$$

If $z<h(y)$ then $Y^*_t=y$, for $0\leq t\leq t_w$, where t_w is as defined in Lemma 4.1. Also

$$Z^{Y^*}_t = z\mathrm{e}^{-\lambda t}, \qquad \text{for } 0\leq t\leq t_w.$$

Moreover $Z^{Y^*}_t<h(y)$, for $t<t_w$, and $Z^{Y^*}_{t_w}=h(y)$. With reference to (4.10) and Proposition 4.3, it follows that

$$\int_0^{t_w}\big\{Z^{Y^*}_{t-}v_z\big(Y^*_{t-},\, Z^{Y^*}_{t-}\big)-a(Y^*_{t-})^2-Z^{Y^*}_{t-}\big\{\psi\big(Z^{Y^*}_{t-}\big)-\psi\big(Z^{Y^*}_{t-}-Y^*_{t-}\big)\big\}\big\}\,dt=0.$$

By definition $Y^*_t=h^{-1}(Z^{Y^*}_t)$, for $t\geq t_w$. According to Proposition 4.3, v satisfies (4.8), therefore

$$\int_{t_w}^{\infty}\big\{D^+_y v\big(Y^*_{t-},\, Z^{Y^*}_{t-}\big)+v_z\big(Y^*_{t-},\, Z^{Y^*}_{t-}\big)\big\}\,d(Y^*_t)^c=0.$$

Moreover, since v satisfies (4.10), therefore

$$\int_{t_w}^{\infty}\big\{Z^{Y^*}_{t-}v_z\big(Y^*_{t-},\, Z^{Y^*}_{t-}\big)-a(Y^*_{t-})^2-Z^{Y^*}_{t-}\big\{\psi\big(Z^{Y^*}_{t-}\big)-\psi\big(Z^{Y^*}_{t-}-Y^*_{t-}\big)\big\}\big\}\,dt=0.$$

With reference to (5.25), we therefore conclude that $v=V$ and that $Y^*=Y^h\in\mathcal{A}^-_D(y)$ is an admissible optimal liquidation strategy for the optimization problem (3.10). The result then follows from (3.9). □

REFERENCES

ALDRIDGE, I. (2010): *High-Frequency Trading*, Hoboken, NJ: John Wiley & Sons.

ALFONSI, A., A. FRUTH, and A. SCHIED (2010): Optimal Execution Strategies in Limit Order Books with General Shape Functions, *Quant. Finance* 10, 143–157.

ALFONSI, A., and A. SCHIED (2010): Optimal Trade Execution and Absence of Price Manipulations in Limit Order Book Models, *SIAM J. Financ. Math.* 1, 490–522.

ALFONSI, A., A. SCHIED, and A. SLYNKO (2012): Order Book Resilience, Price Manipulation, and the Positive Portfolio Problem, *SIAM J. Financ. Math.* 3(1), 511–533.

ALMGREN, R., and N. CHRISS (1999): Value under Liquidation, *Risk* 12(12), 61–63.

ALMGREN, R., and N. CHRISS (2001): Optimal Execution of Portfolio Transactions, *J. Risk* 3(2), 5–39.

ALMGREN, R. F. (2003): Optimal Execution with Nonlinear Impact Functions and Trading-Enhanced Risk, *Appl. Math. Finance* 10, 1–18.

BANK, P., and D. BAUM (2004): Hedging and Portfolio Optimization in Financial Markets with a Large Trader, *Math. Finance* 14(1), 1–18.

BERTSIMAS, D., and A. W. LO (1998): Optimal Control of Execution Costs, *J. Financ. Market* 1, 1–50.

CETIN, U., R. A. JARROW, and P. PROTTER (2004): Liquidity Risk and Arbitrage Pricing Theory, *Financ. Stoch.* 8, 1–31.

FLEMING, W. H., and H. M. SONER (1993): *Controlled Markov Processes and Viscosity Solutions*, New York: Springer-Verlag.

FOLLAND, G. B. (1984): *Real Analysis*, New York: John Wiley & Sons.

GATHERAL, J. (2010): No-Dynamic-Arbitrage and Market Impact, *Quant. Finance* 10(7), 749–759.

GATHERAL, J., A. SCHIED, and A. SLYNKO (2011): Exponential Resilience and Decay of Market Impact, in *Econophysics of Order-Driven Markets*, F. Abergel, B. K. Chakrabarti, A. Chakraborti, and M. Mitra, eds., Milan: Springer, pp. 225–236.

HUBERMAN, G., and W. STANZL (2004): Price Manipulation and Quasi-Arbitrage, *Econometrica* 72(4), 1247–1275.

KISSELL, R., and R. MALAMUT (2005): Understanding the Profit and Loss Distribution of Trading Algorithms, Institutional Investor, Guide to Algorithmic Trading, Spring 2005.

OBIZHAEVA, A., and J. WANG (2005): Optimal Trading Strategy and Supply/Demand Dynamics, Working Paper.

PREDOIU, S., G. SHAIKHET, and S. SHREVE (2011): Optimal Execution in a General One-Sided Limit-Order Book, *SIAM J. Financ. Math.* 2, 183–212.

PROTTER, P. (1990): *Stochastic Integration and Differential Equations*, Berlin, Heidelberg, New York: Springer.

RUDIN, W. (1991): *Functional Analysis*, New York: McGraw-Hill.

SCHIED, A., and T. SCHÖNEBORN (2009): Risk Aversion and the Dynamics of Optimal Liquidation Strategies in Illiquid Markets, *Financ. Stoch.* 13, 181–204.

SCHIED, A., T. SCHÖNEBORN, and M. TEHRANCHI (2010): Optimal Basket Liquidation for CARA Investors Is Deterministic, *Appl. Math. Finance* 17(6), 471–489.

MATHEMATICAL FINANCE

An International Journal of Mathematics, Statistics and Financial Economics

Mathematical Finance, Vol. 24, No. 4 (October 2014), 728–761

ADMISSIBILITY OF GENERIC MARKET MODELS OF FORWARD SWAP RATES

LIBO LI AND MAREK RUTKOWSKI

University of Sydney

Our main goal is to re-examine and extend certain results from the papers by Galluccio et al. and Pietersz and van Regenmortel. We establish several results providing alternate necessary and sufficient conditions for admissibility of a family of forward swaps, that is, the property that it is supported by a (positive) family of bonds associated with the underlying tenor structure. We also derive the generic expression for the joint dynamics of a family of forward swap rates under a single probability measure and we show that these dynamics are uniquely determined by a selection of volatility processes with respect to the set of driving martingales.

KEY WORDS: forward swap rate, market model, Libor, admissibility.

1. INTRODUCTION

Our main goal is to re-examine the *admissibility problem* for an arbitrary family of forward swaps. This issue was previously studied in Galluccio et al. (2007) (see also Pietersz and Regenmortel 2006 for the related study of a special case) who attempted to provide necessary and sufficient conditions for a family of forward swaps to have the ability to underpin an arbitrage-free model of the term structure of interest rates with a finite tenor structure. Additionally, we will also examine an extension of this problem to the situation when the payments dates of forward swaps are chosen to be only some of the tenor structure dates between the start date of a swap and its maturity date. This generalization is motivated by the case where some forward swaps (typically, of shorter maturities) have quarterly payments, whereas for some other swaps (with longer maturities) the payments are exchanged according to the semiannual schedule.

Let $\mathcal{T} = \{T_0, \dots, T_n\}$ with $0 < T_0 < T_1 < \cdots < T_{n-1} < T_n$ be a fixed sequence of dates representing the complete *tenor structure* for all forward swaps under consideration. For every $i = 1, \dots, n$, we write $a_i = T_i - T_{i-1}$ to denote the length of the ith accrual period. Let $B(t, T_i)$ stand for the price of the *unit zero-coupon bond* maturing at T_i. Unless stated otherwise, it is assumed that for every $i = 0, \dots, n$ the bond price $B(t, T_i), t \in [0, T_i]$, is governed by a positive stochastic process with the obvious terminal condition $B(T_i, T_i) = 1$.

Let $\mathcal{S} = \{S_1, \dots, S_l\}$ be an arbitrary family of l distinct fixed-for-floating forward swaps associated with the tenor structure $\mathcal{T}$, in the sense that any reset or settlement date for any swap S_j in $\mathcal{S}$ belongs to $\mathcal{T}$. For concreteness, we assume throughout that

We thank an anonymous referee for the careful reading of our work and valuable comments. The research of Marek Rutkowski was supported under Australian Research Council's Discovery Projects funding scheme (DP0881460).

Manuscript received April 2010; final revision received February 2012.

Address correspondence to Marek Rutkowski, School of Mathematics and Statistics FO7, University of Sydney, NSW 2006, Australia; e-mail: marek.rutkowski@sydney.edu.au.

DOI: 10.1111/mafi.12001

the frequencies of fixed and floating payments are the same. As commonly assumed, we postulate that the floating leg payments are specified by the level of the LIBOR at each reset date and payments are exchanged at the next settlement date. The *forward swap rate* κ^j for the standard forward swap S_j, which starts at T_{s_j} and matures at T_{m_j}, is well known to be given by the following expression (see, for instance, section 13.1 in Musiela and Rutkowski 2005)

$$\kappa_t^j = \kappa_t^{s_j,m_j} := \frac{B(t, T_{s_j}) - B(t, T_{m_j})}{\sum_{i=s_j+1}^{m_j} a_i B(t, T_i)} = \frac{P_t^{s_j,m_j}}{A_t^{s_j,m_j}}, \quad \forall\, t \in [0, T_{s_j}], \tag{1.1}$$

where we denote

$$P_t^{s_j,m_j} := B(t, T_{s_j}) - B(t, T_{m_j}), \quad \forall\, t \in [0, T_{s_j}], \tag{1.2}$$

and

$$A_t^{s_j,m_j} := \sum_{i=s_j+1}^{m_j} a_i B(t, T_i), \quad \forall\, t \in [0, T_{s_j+1}]. \tag{1.3}$$

The process A^{s_j,m_j} is usually called the *swap annuity* or the *swap numéraire* for the forward swap S_j, although some authors prefer the term *level process*. In practical applications, it is also referred to as the present value of the basis point of the swap. The dates $T_{s_j}, T_{s_j+1}, \dots, T_{m_j-1}$ are then the *reset* dates for the forward swap S_j, whereas the dates $T_{s_j+1}, T_{s_j+2}, \dots, T_{m_j}$ are *settlement* dates. Therefore, the collection of dates $T_{s_j}, T_{s_j+1}, \dots, T_{m_j}$ represents the *reset/settlement* dates in the forward swap S_j. It is a simple observation that these dates are the only *relevant dates* for the (standard) forward swap S_j.

REMARK 1.1. In what follows, we will consider a more general set-up, in which the start date T_{s_j} is the first reset date and the maturity date T_{m_j} is the last settlement date in the forward swap S_j. Other reset/settlement dates in a forward swap S_j can be chosen freely from the set of dates $T_i \in \mathcal{T}$ that satisfy $T_{s_j} < T_i < T_{m_j}$. Then all reset/settlement dates in the forward swap S_j will be simply referred to as the *relevant dates* in S_j. They can also be seen as the tenor structure *generated* by the forward swap S_j; this observation justifies the use of the symbol $\mathcal{T}(S_j)$. In the general case, we write

$$\kappa_t^j = \kappa_t^{s_j,m_j} = \frac{B(t, T_{s_j}) - B(t, T_{m_j})}{\sum_{i\in\mathcal{A}_j} \widetilde{a}_i B(t, T_i)} = \frac{P_t^{s_j,m_j}}{A_t^{s_j,m_j}}, \quad \forall\, t \in [0, T_{s_j}], \tag{1.4}$$

where $\mathcal{A}_j$ is the set of indices of all settlement dates in the swap S_j and $\widetilde{a}_i$ represents the ith adjusted accrual period. For instance, if a forward swap S_j has only one settlement date, specifically, its maturity date, then $\mathcal{T}(S_j) = \{T_{s_j}, T_{m_j}\}$ and $\mathcal{A}_j = \{m_j\}$. Furthermore, equation (1.4) becomes

$$\kappa_t^j = \frac{B(t, T_{s_j}) - B(t, T_{m_j})}{\widetilde{a}_{m_j} B(t, T_{m_j})},$$

where $\widetilde{a}_{m_j} = T_{m_j} - T_{s_j} = \sum_{i=s_j+1}^{m_j} a_i$; this corresponds to the forward LIBOR (see equation (1.10)) If all the dates from $\mathcal{T}$ between the start and maturity dates are the settlement dates for a given forward swap, the contract is referred to as a *standard* forward swap.

Let us select the bond maturing at some date $T_b \in \mathcal{T}$ as a *bond numéraire* and let us denote by $\mathcal{B}^b$ the family of the *deflated bond prices*

$$\mathcal{B}^b := \{ B^b(t, T_i) = B(t, T_i)/B(t, T_b),\ i = 0, \dots, n \}.$$

In terms of the deflated bond prices $B^b(t, T_i)$, for the standard forward swap we obtain

$$\kappa_t^{s_j,m_j} = \frac{B^b(t, T_{s_j}) - B^b(t, T_{m_j})}{\sum_{i=s_j+1}^{m_j} a_i B^b(t, T_i)} = \frac{P_t^{b,s_j,m_j}}{A_t^{b,s_j,m_j}}, \quad \forall\, t \in [0, T_{s_j} \wedge T_b], \tag{1.5}$$

where we write

$$P_t^{b,s_j,m_j} = B^b(t, T_{s_j}) - B^b(t, T_{m_j}), \quad \forall\, t \in [0, T_{s_j} \wedge T_b], \tag{1.6}$$

and

$$A_t^{b,s_j,m_j} = \sum_{i=s_j+1}^{m_j} a_i B^b(t, T_i), \quad \forall\, t \in [0, T_{s_j+1} \wedge T_b]. \tag{1.7}$$

We call the process A^{b,s_j,m_j} the *deflated swap annuity* or *deflated swap numéraire*. As we shall see in what follows, the (deflated) swap numéraires are crucial objects in the specification of the Radon–Nikodým densities for *swap measures* under which forward swap rates are (local) martingales.

REMARK 1.2. Note that the choice of a bond numéraire is arbitrary, in the sense that one may take the bond price $B(t, T_b)$ for any maturity date $T_b \in \mathcal{T}$ to play this role. It will be rather clear that the results presented in the sequel do not depend on a particular choice of a bond numéraire, but only provided that all bond prices are either assumed or shown to follow positive processes.

We are in a position to describe the main problems considered in this work. We are interested in deriving the joint dynamics of forward swap rates associated with a family $\mathcal{S} = \{S_1, \dots, S_l\}$ of forward swaps, as well as in checking whether these dynamics are supported by the existence of a (unique) family of deflated bond prices $\mathcal{B}^b = \{B^b(t, T_i),\ i = 1, \dots, n\}$ for some choice (or all choices) of a bond numéraire.

We will thus deal with two related problems, which can be informally described as follows:

- Under which assumptions on a family $\mathcal{S}$ can a model of forward swap rates be supported by the existence of an arbitrage-free term structure model consistent with swap rates, where by a term structure model we mean the joint dynamics of positive deflated bond prices?
- Under which assumptions on a family $\mathcal{S}$ are the joint dynamics of a family of forward swap rates uniquely specified under a single probability measure in terms of respective volatility processes?

The issues related to the above two questions are studied in Sections 2 and 3, respectively. In both cases, the method hinges on a formulation of a suitable "inverse problem" and a thorough examination of its well-posedness. Specifically, to provide an answer to the first question, we investigate the existence of a unique and nonzero (or strictly positive) solution to the inverse problem for deflated bond prices when the forward swap rates are treated as "observables". This is called hereafter the *inverse problem* (IP.1) and the well-posedness of the problem (IP.1) is referred to as the (weak) $\mathcal{T}$-admissibility of a family $\mathcal{S}$. Several alternative necessary and/or sufficient conditions for the well-posedness of the problem (IP.1) are provided in Theorem 2.33 and Propositions 2.36 and 2.37.

It is worthwhile to stress that our conditions differ from these provided in Galluccio et al. (2007), who focused on the existence of "cycles" in a given family of forward swaps, whereas our necessary and/or sufficient conditions have a completely different character. It should be acknowledged, however, that a detailed analysis of their work was the original motivation for this research. In fact, we show by means of counter-examples that the main result in Galluccio et al. (2007) is defective and thus the present paper could also be seen as a partial correction to their result.

For the second above-mentioned question, we introduce and examine the corresponding inverse problem for swap annuities, termed the *inverse problem* (IP.2), and the related concept of the (weak) $\mathcal{A}$-admissibility (see Definition 3.2). It will be shown that the (weak) $\mathcal{T}$-admissibility implies the (weak) $\mathcal{A}$-admissibility (see Lemma 3.3), but the converse does not hold, as we demonstrate by means of a counterexample (see Example 3.4).

Finally, we describe in Section 3.2 a general methodology for the explicit computation of the joint dynamics of forward swap rates, which is aimed to cover all potentially interesting cases of $\mathcal{A}$-admissible families of forward swaps. For the sake of generality, we examine the set-up of a model driven by semimartingales; for most practical purposes this abstract framework can be easily specified to the case of a model driven by a multi-dimensional correlated Brownian motions (or jump-diffusion processes). We illustrate this approach by means of Example 3.11.

It should be acknowledged that the research presented in this work is by no means complete and in fact some related problems remain still open. Firstly, various market models similar to forward swap rate models can also be formally derived for certain families of forward CDS spreads and LIBORs. For preliminary studies of extended market models of this kind, the interested reader is referred to Brigo (2008) and Li and Rutkowski (2011). To the best of our knowledge, the issues of either $\mathcal{T}$- or $\mathcal{A}$-admissibility of a joint family of forward swap rates and forward CDS spreads were not systematically examined in the existing literature, however, and thus these issues remain largely open.

Secondly, and more importantly, the recent credit crisis has dramatically changed the way in which practitioners discount future payoffs. The revised approach to the discounting convention resulted also in an essential change in the way in which valuation of interest rate derivatives is performed and, as a consequence, it also reignited interest in term structure modeling. The standard formulae considered in this paper can be extended to cover the new perception of the market, as explained in recent papers by Bianchetti (2009, 2010), Bianchetti and Carlicchi (2011), Fujii et al. (2009a,b), Kijima et al. (2009), Mercurio (2009), Moreni and Pallavicini (2010), and Pallavicini and Tarenghi (2009). According to the new pricing paradigm of fixed-for-floating interest rate swaps and other related contracts such as swaptions, a single discounting curve used in the traditional approach should be replaced by (at least) two yield curves: the first one is used to specify the level of future floating cash flows, whereas the second one (a proxy for the risk-free interest rate) is used for discounting of fixed and floating cash flows. According to this extended swap valuation methodology, formula (1.4) should be modified as follows (see, for instance, formula (9) in Kijima et al. 2009 or page 15 in Mercurio 2009)

$$\widehat{\kappa}_t^j = \widehat{\kappa}_t^{s_j,m_j} := \frac{\sum_{i\in\mathcal{A}_j} \widetilde{a}_i \widehat{B}(t, T_i)\widetilde{L}_i(t)}{\sum_{i\in\mathcal{A}_j} \widetilde{a}_i \widehat{B}(t, T_i)}, \tag{1.8}$$

where $\widetilde{L}_i(t)$ represents the forward LIBOR over the relevant accrual period, as computed from the risky LIBOR curve, and $\widehat{B}(t, T_i)$ is a judiciously chosen market risk-free yield curve, such as the OIS (Overnight Index Swap) yield curve. The OIS is a fixed-for-floating interest rate swap which pays periodically the daily compounded overnight rate, instead of the LIBOR. Since the overnight (collateral) rate can be seen as risk-free, the forward swap rate corresponding to the OIS is still given by expression (1.4), where $B(t, T_i)$ is interpreted as the price of a risk-free (for instance, collateralized) zero-coupon bond. For more details on the choice of relevant market rates for each currency, the interested reader may consult, for instance, Fujii et al. (2009a,b) or Moreni and Pallavicini (2010). Regarding the LIBOR, one can make the modeling postulate that the forward LIBOR $\widetilde{L}_i$, as seen at time t for the period $[T_{i-1}, T_i]$, can be formally given by the following expression:

$$\widetilde{L}_i(t) = \frac{\widetilde{B}(t, T_{i-1}) - \widetilde{B}(t, T_i)}{\widetilde{a}_i \widetilde{B}(t, T_i)}, \tag{1.9}$$

where the quantities $\widetilde{B}(t, T_i)$, $i = 0, \dots, n$, are aimed to represent the implicit LIBOR yield curve. It is worth noting that (1.4) can also be represented as follows:

$$\kappa_t^j = \kappa_t^{s_j,m_j} = \frac{\sum_{i\in\mathcal{A}_j} \widetilde{a}_i B(t, T_i)L_i(t)}{\sum_{i\in\mathcal{A}_j} \widetilde{a}_i B(t, T_i)},$$

where the forward LIBOR $L_i(t)$ is given by the traditional pre-crisis formula, that is,

$$L_i(t) = \frac{B(t, T_{i-1}) - B(t, T_i)}{\widetilde{a}_i B(t, T_i)}. \tag{1.10}$$

A comparison of (1.4) and (1.8) makes it clear that the two expressions coincide when the equality $\widehat{B}(t, T_i) = B(t, T_i)$ holds and the forward LIBOR is given by (1.10), so that, as expected, the new approach reduces to the classic one when all rates are nonrisky. We do not deal with default-risky rates here, so the issues related to admissible families of forward swap rates given by the generic expression (1.8) are left for future research.

2. ADMISSIBLE FAMILIES OF FORWARD SWAP RATES

In this section, we address the issue whether a given family of forward swap rates is supported by the existence of the implied family of deflated bond prices. This corresponds to the inverse problem examined by Galluccio et al. (2007), which can be informally stated as follows:

- Provide necessary and sufficient conditions for a given family of forward swap rates $\mathcal{S} = \{S_1, \dots, S_l\}$ and the corresponding (diffusion-type) swap rate processes $\{\kappa^1, \dots, \kappa^l\}$ to uniquely specify a family of nonzero deflated bond prices $\mathcal{B}^b = \{B^b(t, T_0), \dots, B^b(t, T_n)\}$ for any choice of $b \in \{0, \dots, n\}$.

2.1. Linear Systems Associated with Forward Swaps

Let us note that for a given family $\mathcal{B}^n = \{B^n(t, T_i),\ i = 0, \dots, n\}$ of stochastic processes representing the deflated bond prices when $B(t, T_n)$ is the numéraire bond, we can uniquely determine the forward swap rate κ^{s_j, m_j} for any start T_{s_j} and maturity T_{m_j} dates from the tenor structure $\mathcal{T}$. This simple observation paves the way for the most commonly used *direct approach* to modeling of forward swap rates, in which one starts by choosing a term structure model and subsequently derives the joint dynamics of a family of forward swap rates of interest.

It is also clear that each forward swap rate κ^{s_j, m_j} generates a linear equation in which deflated bond prices can be seen as "unknowns." Specifically, one obtains the following *swap equation* associated with the forward swap S_j and the numéraire bond $B(t, T_b)$

$$-B^b(t, T_{s_j}) + \sum_{i=s_j+1}^{m_j-1} \kappa_t^{s_j, m_j} a_i B^b(t, T_i) + \left(1 + \kappa_t^{s_j, m_j}\right) a_{m_j} B^b(t, T_{m_j}) = 0. \tag{2.1}$$

In this context, the following *inverse problem* arises in a natural way: describe all families of forward swaps associated with the tenor structure $\mathcal{T}$ such that the corresponding family of forward swap rates uniquely specifies the associated family of nonzero (and preferably strictly positive) deflated bond prices.

To formally address the above-mentioned inverse problem, we identify a forward swap S_j with the corresponding swap equation, in which a generic value κ_j of the forward swap rate κ^{s_j, m_j} plays the role of a parameter. Let us now introduce some notation. For a fixed b, let x_i stand for a generic value of the deflated bond price $B^b(t, T_i)$ and let κ_j be a generic value of the forward swap rate $\kappa_t^j := \kappa_t^{s_j, m_j}$. Since x_i and κ_j are aimed to represent generic values of the corresponding processes in a yet unspecified stochastic model, we have that $(x_0, \dots, x_n) \in \mathbb{R}^{n+1}$ and $(\kappa_1, \dots, \kappa_l) \in \mathbb{R}^l$, in general. Since the bond $B(t, T_b)$ is chosen to be the numéraire asset, it is clear, by the definition of the deflated bond price, that the variable x_b satisfies $x_b = 1$ and thus it is more adequate to consider a generic value $(x_0, \dots, x_{b-1}, x_{b+1}, \dots, x_n) \in \mathbb{R}^n$. Using this notation, we may represent equation (1.5) as follows, for every $j = 1, \dots, l$,

$$\kappa_j = \frac{x_{s_j} - x_{m_j}}{\sum_{i=s_j+1}^{m_j} a_i x_i}. \tag{2.2}$$

For brevity, we write $c_{j,i} = \kappa_j a_i$ and $\widetilde{c}_{j,m_j} = (1 + \kappa_j) a_{m_j}$, so that equation (2.1) becomes

$$-x_{s_j} + \sum_{i=s_j+1}^{m_j-1} c_{j,i} x_i + \widetilde{c}_{j,m_j} x_{m_j} = 0. \tag{2.3}$$

DEFINITION 2.1. For a given family $\mathcal{S} = \{S_1, \dots, S_l\}$ of forward swaps and any fixed $b \in \{0, \dots, n\}$, the associated linear system (2.3) parametrized by a vector $(\kappa_1, \dots, \kappa_l) \in \mathbb{R}^l$, is denoted as

$$C^b(\kappa_1, \dots, \kappa_l)\overline{x}^b = \overline{e}^b(\kappa_1, \dots, \kappa_l),$$

where $\overline{x}^b = (x_0, \dots, x_{b-1}, x_{b+1}, \dots, x_n)$ is the vector of unknowns and the vector $\overline{e}^b(\kappa_1, \dots, \kappa_l)$ belongs to $\mathbb{R}^l$.

The matrix $C^b(\kappa_1, \dots, \kappa_l)$ and the vector $\overline{e}^b(\kappa_1, \dots, \kappa_l)$ depend on the vector $(\kappa_1, \dots, \kappa_l) \in \mathbb{R}^l$. To alleviate notation, we shall frequently suppress the variables $\kappa_1, \dots, \kappa_l$, however, and we will simply write C^b and $\overline{e}^b$.

As already mentioned in Remark 1.1, we will also consider more general forward swaps for which the actual payment dates constitute merely a subset of the collection of all dates between the start and maturity dates. In that case, the sum in the denominator of (2.2) is taken over some dates between T_{s_j+1} and T_{m_j} only. Furthermore, the coefficients a_i are suitably adjusted to represent the length of times between the consecutive payment dates, and a modified version of equation (2.3) is easily derived, so that Definition 2.1 covers this general set-up as well. For instance, if a forward swap S_j settles at its maturity date only, then equation (2.2) becomes

$$\kappa_j = \frac{x_{s_j} - x_{m_j}}{\widetilde{a}_{m_j} x_{m_j}},$$

where $\widetilde{a}_j = T_{m_j} - T_{s_j} = \sum_{i=s_j+1}^{m_j} a_i$. The following definition is also tailored to cover the whole spectrum of cases considered in this work.

DEFINITION 2.2. A date $T_i \in \mathcal{T}$ is a relevant date for the swap S_j whenever $T_i = T_{s_j}$, $T_i = T_{m_j}$, or the term $c_{j,i}$ is nonzero. We write $\mathcal{T}(S_j)$ to denote the set of all relevant dates for the swap S_j and we set $\mathcal{T}_0(S_j) = \{T_{s_j}, T_{m_j}\}$.

The following notation will be useful:

$\mathcal{T}_0(\mathcal{S})$ – the set of all start/maturity dates for the forward swaps from a family $\mathcal{S}$, that is,

$$\mathcal{T}_0(\mathcal{S}) = \bigcup_{j=1}^{l} \mathcal{T}_0(S_j) = \bigcup_{j=1}^{l} \{T_{s_j}, T_{m_j}\}.$$

$\mathcal{T}(\mathcal{S})$ – the set of all relevant dates for the forward swaps from a family $\mathcal{S}$, that is,

$$\mathcal{T}(\mathcal{S}) = \bigcup_{j=1}^{l} \mathcal{T}(S_j).$$

It is clear that a date T_i does not belong to $\mathcal{T}(\mathcal{S})$ whenever the corresponding column of the matrix C^b has all entries equal to zero. Manifestly, the nonuniqueness of a solution to the inverse problem holds in that case, since the deflated bond price $B^b(t, T_i)$ cannot be retrieved. This is indeed trivial since the variable x_i does not appear in any equation in the linear system $C^b\overline{x}^b = \overline{e}^b$. In what follows, we only consider families $\mathcal{S}$ that generate the tenor structure $\mathcal{T}$, in the sense that $\mathcal{T} = \mathcal{T}(\mathcal{S})$.

2.2. Graph Theory Terminology

Before proceeding to the detailed analysis of the inverse problem, we need first to recall some basic graph theory terminology, which will be employed in what follows. We also provide some preliminary results concerning the existence of a cycle in a graph, which is formally associated with a family of forward swap rates. For an in-depth introduction to the graph theory, the interested reader may consult the monograph by Bollobás (1979).

DEFINITION 2.3. A graph G is an ordered pair of disjoint sets (V, E) such that E is a subset of the set of unordered pairs of V. The set V is called the vertex set and E is called the edge set.

DEFINITION 2.4.

(i) Two vertices v_1 and v_2 are said to be *adjacent* if there exists an edge $e \in E$ such that e is the unordered pair $\{v_1, v_2\}$.
(ii) A *path* is a finite sequence of vertices such that each of the vertices is adjacent to the next vertex in the sequence.
(iii) A *cycle* is a path such that the start vertex and end vertex are the same.
(iv) A graph G is said to be *connected* if there exists a path between any two vertices.
(v) A graph G is *acyclic* if no cycles exist.

The next result is borrowed from Bollobás (1979); it is in fact produced by combining theorem 4 on page 7 and exercise 9 on page 23 therein.

THEOREM 2.5. *The following statements are equivalent for a graph G with m vertices:*

(i) G is a minimal connected graph, that is, G is connected and if $\{x, y\} \in E$ *then* $G - \{x, y\}$ *is disconnected,*
(ii) G is a maximal acyclic graph, that is, G is acyclic and if x and y any nonadjacent vertices of G then the graph $G + \{x, y\}$ *contain a cycle,*
(iii) G is connected and has $m - 1$ *edges,*
(iv) G is acyclic and has $m - 1$ *edges.*

Let $\mathcal{S}$ be a family consisting of l swaps on the complete tenor structure $\mathcal{T} = \{T_0, T_1, \ldots, T_n\}$. As previously introduced, the set $\mathcal{T}_0(\mathcal{S})$ is the set of all start or maturity dates of swaps in $\mathcal{S}$. To translate our set-up into graph theory terminology, we regard the set $\mathcal{T}_0(\mathcal{S})$ as a vertex set. A swap $S_j \in \mathcal{S}$ which start at T_{s_j} and matures at T_{m_j} is characterized by the pair of elements $\{T_{s_j}, T_{m_j}\}$. This characterization allows us to regard $\mathcal{S}$ as the edge set. Note that we only need to deal here with graphs that have a finite number of vertices.

DEFINITION 2.6. The *graph of* $\mathcal{S}$ is the vertex and edge pair given as $(\mathcal{T}_0(\mathcal{S}), \mathcal{S})$. It is clear that the graph of $\mathcal{S}$ has l edges and at most $n + 1$ vertices.

As will be elaborated on in Section 2.4 (see, in particular, Remark 2.17), the (non)existence of a cycle in the graph of $\mathcal{S}$ does not seem to be of primary interest in the study of $\mathcal{T}$-admissibility of a family $\mathcal{S}$. However, in order to re-examine proposition 2.1 in Galluccio et al. (2007), we will first present some basic properties of the graph of $\mathcal{S}$ directly related to the (non)existence of a cycle in this graph. The first lemma in this vein is a straightforward consequence of Theorem 2.5.

LEMMA 2.7. *Let $\mathcal{S} = \{S_1, \dots, S_l\}$ be any family of swaps with $\mathcal{T}(\mathcal{S}) = \mathcal{T}$. If $l > n$ then there exists a cycle in the graph of $\mathcal{S}$.*

The next auxiliary result deals with the acyclic case.

LEMMA 2.8. *Let $\mathcal{S} = \{S_1, \dots, S_n\}$ be a family of n swaps with $\mathcal{T}(\mathcal{S}) = \mathcal{T}$. If there are no cycles in the graph of $\mathcal{S}$ then each date from the tenor structure $\mathcal{T}$ is either the start date or the maturity date of some swap S_j from $\mathcal{S}$, that is, then the equality $\mathcal{T}_0(\mathcal{S}) = \mathcal{T}$ holds.*

Proof. Let us assume that there exists a date $T_k \in \mathcal{T}$ which does not belong to $\mathcal{T}_0(\mathcal{S})$. Then there exists a family $\widetilde{\mathcal{S}}$ comprising of n swaps and such that $\mathcal{T}_0(\widetilde{\mathcal{S}}) = \{T_0 < \dots < T_{k-1} < T_{k+1} < \dots < T_n\}$. Indeed, to obtain $\widetilde{\mathcal{S}}$ from $\mathcal{S}$, it suffices to delete the date T_k from all swap equations. Hence, by Lemma 2.7, a cycle necessarily exists in the family $\widetilde{\mathcal{S}}$ and thus also in $\mathcal{S}$. □

The last auxiliary result furnishes a simple, but useful, observation.

LEMMA 2.9. *Assume that $l = n$ and the graph $\mathcal{S}$ is connected. Then $\mathcal{T}_0(\mathcal{S}) = \mathcal{T}$ and thus there are no cycles in $\mathcal{S}$.*

2.3. Inverse Problem for Deflated Bonds

We are in a position to formalize the inverse problem for the deflated bond prices and the related concepts of admissibility of a family $\mathcal{S}$. Note that the term "almost every" refers throughout to the Lebesgue measure on the respective space $\mathbb{R}^l$ (or $\mathbb{R}^l_+$).

Inverse Problem (IP.1): A family $\mathcal{S} = \{S_1, \dots, S_l\}$ admits a solution to the *inverse problem* (IP.1) if for any $b \in \{0, \dots, n\}$ and for almost every $(\kappa_1, \dots, \kappa_l) \in \mathbb{R}^l$ there exists a solution $\overline{x}^b \in \mathbb{R}^n$ to the linear system $C^b(\kappa_1, \dots, \kappa_l)\overline{x}^b = \overline{e}^b(\kappa_1, \dots, \kappa_l)$ associated with $\mathcal{S}$.

Galluccio et al. (2007) prefer to deal directly with the stochastic linear system $\mathcal{C}^b\mathcal{B}^b = \overline{e}^b$ where $\mathcal{C}^b$ is the matrix of stochastic processes obtained by replacing the variables $(\kappa_1, \dots, \kappa_l)$ by diffusion-type forward swap rate processes $\kappa^j = \kappa^{s_j, m_j}$, $j = 1, \dots, l$ and the unknowns $(x_0, \dots, x_{b-1}, x_{b+1}, \dots, x_n) \in \mathbb{R}^n$ are replaced by the (unknown) deflated bond price processes $B^b(t, T_i), i = 0, \dots, b-1, b+1, \dots, n$. To the best of our understanding, they work under the tacit assumption that for any $t \in (0, T_0]$ the random variable $(\kappa^1_t, \dots, \kappa^l_t)$ has a strictly positive joint probability density function on $(\mathbb{R}^l, \mathcal{B}(\mathbb{R}^l))$, so that the probability distribution of $(\kappa^1_t, \dots, \kappa^l_t)$ has the full support $\mathbb{R}^l$. In these circumstances, the inverse problem (IP.1) introduced above appears to be equivalent to the inverse problem studied by Galluccio et al. (2007).

The following definition reflects the idea of admissibility of a family of forward swaps, as proposed by Galluccio et al. (2007). It is based on the inverse problem (IP.1), but it also refers to some additional features of a solution to this problem, such as its uniqueness and positivity.

DEFINITION 2.10. We say that a family $\mathcal{S}$ of forward swaps associated with $\mathcal{T}$ is weakly $\mathcal{T}$-admissible if for any choice of $b \in \{0, \dots, n\}$ the following property holds: for almost every $(\kappa_1, \dots, \kappa_l) \in \mathbb{R}^l$ there exists a unique nonzero solution $\overline{x}^b \in \mathbb{R}^n$ to the linear system

$C^b(\kappa_1, \dots, \kappa_l)\overline{x}^b = \overline{e}^b(\kappa_1, \dots, \kappa_l)$ associated with $\mathcal{S}$. If, in addition, the solution is strictly positive for almost every $(\kappa_1, \dots, \kappa_l) \in \mathbb{R}^l_+$ then we say that a family $\mathcal{S}$ is $\mathcal{T}$-admissible.

We write $\mathfrak{A}$ to denote the class of all families of forward swaps that are weakly $\mathcal{T}$-admissible, whereas $\overline{\mathfrak{A}}$ stands for the class of all families that are $\mathcal{T}$-admissible.

REMARK 2.11. The property of $\mathcal{T}$-admissibility is the strongest form of invertibility, which appears to be the most convenient from the viewpoint of further applications. If it holds then the Radon–Nikodým densities for various swap measures can be computed from the family $\mathcal{B}^b$ and expressed in terms of the underlying forward swap rates. Consequently, in that case it will be easy to derive the joint dynamics of forward swap rates under a single probability measure. In addition, these dynamics will be supported by an arbitrage-free model of positive deflated bond prices.

Before proceeding, let us comment on the approach presented in the paper by Galluccio et al. (2007). In Galluccio et al. (2007), the authors formalize the concept of admissibility of a family $\mathcal{S}$ through the following definition that corresponds to Definition 2.2 in Galluccio et al. (2007). For an overview of the relevant terminology from the graph theory, we refer to Section 2.2.

DEFINITION 2.12. A family $\mathcal{S}$ of forward swaps associated with $\mathcal{T}$ is admissible if the following conditions are satisfied:

(i) the number of forward swaps in $\mathcal{S}$ equals n, that is, $l = n$,
(ii) any date $T_i \in \mathcal{T}$ coincides with the reset/settlement date of at least one forward swap from $\mathcal{S}$,
(iii) there are no cycles (also called *degenerate subsets* in Galluccio et al. 2007) in $\mathcal{S}$.

The main theoretical result in Galluccio et al. (2007) states that the admissibility of $\mathcal{S}$, in the sense of Definition 2.12, is a necessary and sufficient condition for the following property: a set of nonzero deflated bond prices associated with $\mathcal{S}$ exists and is unique, $\mathbb{P}$-a.s., for any choice of the bond numéraire and any generic process $(\kappa^1, \dots, \kappa^l)$ for forward swap rates. In other words, the main result in Galluccio et al. (2007) (see proposition 2.1 therein) establishes the equivalence between the admissibility of $\mathcal{S}$ and the existence of a unique solution to their inverse problem, $\mathbb{P}$-a.s.

As we shall argue in what follows, however, this result is not true, in general. As a consequence, we observe that Definition 2.12 of admissibility of a family $\mathcal{S}$ could be misleading, since it implicitly hinges on the validity of proposition 2.1 in Galluccio et al. (2007). It is our understanding that the method of proof of proposition 2.1 in Galluccio et al. (2007) hinges on the following conjectures:

- If the number l of forward swaps in $\mathcal{S}$ differs from n then either there is no solution to the linear system $C^b\overline{x}^b = \overline{e}^b$ or the nonuniqueness of a solution holds. Therefore, the necessary requirement is that C^b should be a square matrix, so that $l = n$.
- If $l = n$, but a *cycle* exists in $\mathcal{S}$ (see Section 2.2), then the rows in the matrix C^b are linearly dependent, and thus the rank of C^b is less than n, at least for a certain choice of $b \in \{0, \dots, n\}$. Consequently, there is no guarantee that a solution to the linear system $C^b\overline{x}^b = \overline{e}^b$ exists, for almost all realizations of forward swap rates and for an arbitrary choice of $B(t, T_b)$ as a numéraire bond.

- Otherwise, that is, when $l = n$ and there are no cycles in $\mathcal{S}$ then, for any choice of $b \in \{0, \dots, n\}$, the rows in the matrix C^b are linearly independent and thus the matrix C^b is nonsingular, for almost all realizations of forward swap rates. Consequently, the unique solution to $C^b \overline{x}^b = \overline{e}^b$ exists with probability one for any diffusion-type dynamics of forward swap rates. Moreover, for almost all realizations of forward swap rates, the solution $\overline{x}^b$ is nonzero, meaning that $x_i \neq 0$ for every $i = 0, \dots, n$ (recall that $x_b = 1$).

We note that the proof of Proposition 2.1 in Galluccio et al. (2007) suffers from the following shortcomings:

- First, the proof of the sufficiency clause is based on the argument that if the first column has only one nonzero entry then the row vectors are linearly independent. This is, of course, not true in general (it seems that the authors assume that the coefficients in the linear combination should all be nonzero, whereas in fact it suffices that not all of them vanish). A similar argument is used when there a several nonzero entries.
- Second, the proof of the necessity clause is based on the observation that the sum of equations corresponding to the upper and lower paths in a cycle yields an equation of a special shape. This fact does not prove, however, that the row vectors are linearly dependent, as was claimed in Galluccio et al. (2007). To illustrate this point, we present below an example of a family $\mathcal{S}$ with a cycle for which for any choice of $b \in \{0, \dots, n\}$ the corresponding matrix C^b is nonsingular, for almost all $(\kappa_1, \dots, \kappa_l) \in \mathbb{R}^l$.

EXAMPLE 2.13. The following simple counter-example shows that a family $\mathcal{S}$ with a cycle can be weakly $\mathcal{T}$-admissible. Let $n = 3$ and let $(s_j, m_j), j = 1, 2, 3$ be given as (0, 2), (2, 3), and (0, 3), respectively. The swaps S_1 and S_2 yield the path (T_0, T_3) and this path is also given by the swap S_3, so that a cycle (T_0, T_0) is present. For $b = 0$, we obtain the following linear system

$$C^0 \overline{x}^0 = \begin{bmatrix} c_{1,1} & \widetilde{c}_{1,2} & 0 \\ 0 & -1 & \widetilde{c}_{2,3} \\ c_{3,1} & c_{3,2} & \widetilde{c}_{3,3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \overline{e}^0.$$

One can check by direct computations that, for any choice of $b \in \{0, 1, 2, 3\}$, the following properties hold:

(i) for almost all $(\kappa_1, \kappa_2, \kappa_3) \in \mathbb{R}^3$, the matrix C^b is nonsingular and
(ii) for almost all $(\kappa_1, \kappa_2, \kappa_3) \in \mathbb{R}^3$, the unique solution $\overline{x}^b$ has all entries nonzero.

We conclude that the considered family $\mathcal{S}$ is not admissible, in the sense of Definition 2.12, but it satisfies Definition 2.10 of the weak $\mathcal{T}$-admissibility. This example makes it clear that the necessity clause in proposition 2.1 in Galluccio et al. (2007) is incorrect. We will later comment on the sufficiency clause in proposition 2.1 in that paper.

Let us conclude this subsection by mentioning that although Galluccio et al. (2007) examine the admissibility of a family of forward swaps with respect to the tenor structure $\mathcal{T}$ in the context of the construction of swap market models, they remain silent on the joint dynamics of forward swaps, and they focus instead on the dynamics of each forward swap rate under the corresponding swap martingale measure. However, one can infer from the discussion in Section 2 in Galluccio et al. (2007) that when a family $\mathcal{S}$ is admissible then one can specify the model by choosing arbitrarily the volatilities in formula (2.2) in Galluccio et al. (2007). The implicit goal is to obtain a generic arbitrage-free model supported by the unique family of nonzero deflated bond prices. We will return to this issue in Section 3.2.

2.4. Weak Admissibility of Forward Swaps

The purpose of this subsection is to provide necessary and sufficient conditions for a family of forward swaps to be weakly $\mathcal{T}$-admissible. We postulate throughout that a family $\mathcal{S}$ of forward swaps is such that the equality $\mathcal{T}(\mathcal{S}) = \mathcal{T}$ holds. Equivalently, for an arbitrary choice of $b \in \{0, \dots, n\}$ there is no null column in the matrix C^b associated with $\mathcal{S}$.

2.4.1. Preliminary Results. Before examining various forms of admissibility, let us quickly note that the following property holds.

LEMMA 2.14. *Let us fix $(\kappa_1, \dots, \kappa_l) \in \mathbb{R}^l$. The following properties are equivalent:*

(i) the system $C^b(\kappa_1, \dots, \kappa_l)\overline{x}^b = \overline{e}^b(\kappa_1, \dots, \kappa_l)$ has a nonzero solution for some $b \in \{0, \dots, n\}$,
(ii) the system $C^b(\kappa_1, \dots, \kappa_l)\overline{x}^b = \overline{e}^b(\kappa_1, \dots, \kappa_l)$ has a nonzero solution for all $b \in \{0, \dots, n\}$.

Proof. (i) ⇒ (ii) If a nonzero solution exists for a particular b, then for any $\widetilde{b} \in \{0, \dots n\}$ the solution to the system $C^{\widetilde{b}}\overline{x}^{\widetilde{b}} = \overline{e}^{\widetilde{b}}$ can be obtained by taking the ratios of the solution to the system $C^b\overline{x}^b = \overline{e}^b$. This argument corresponds to the simple observation that to obtain the deflated bond prices for any other choice of a numéraire bond, one may use the equality

$$B^{\widetilde{b}}(t, T_i) = \frac{B^b(t, T_i)}{B^b(t, T_{\widetilde{b}})}, \quad \forall \widetilde{b} \neq b.$$

The implication (ii) ⇒ (i) is obvious. □

As we have already seen from Example 2.13, the nonexistence of a cycle is not a necessary condition for the weak $\mathcal{T}$-admissibility of a family of swaps. We therefore introduce the concept of $(\mathcal{T}, b)$-inadmissibility, which will prove useful in the sequel.

DEFINITION 2.15. Let us fix $b \in \{0, \dots, n\}$. A subset $\widetilde{\mathcal{S}}$ of a family $\mathcal{S}$ of swaps is said to be $(\mathcal{T}, b)$-inadmissible if the number of swaps in $\widetilde{\mathcal{S}}$ is strictly greater than the number of dates in $\mathcal{T}(\widetilde{\mathcal{S}}) \setminus \{T_b\}$. We denote by $\mathfrak{M}_0$ the class of all families of swaps that do not contain a $(\mathcal{T}, b)$-inadmissible subset for every $b \in \{0, \dots, n\}$.

Note that the number of dates in $\mathcal{T}(\widetilde{\mathcal{S}}) \setminus \{T_b\}$ is equal to the number of all variables from the set $x_0, \dots, x_{b-1}, x_{b+1}, \dots, x_n$ that are associated with a given subset $\widetilde{\mathcal{S}}$.

The following properties can be checked by inspection:

(a) if a subset $\widetilde{\mathcal{S}}$ is $(\mathcal{T}, b)$-inadmissible for some b such that $T_b \in \mathcal{T}(\widetilde{\mathcal{S}})$ then it is also $(\mathcal{T}, \widetilde{b})$-inadmissible for any $\widetilde{b}$ such that $T_{\widetilde{b}} \in \mathcal{T}(\widetilde{\mathcal{S}})$.

(b) if a subset $\widetilde{\mathcal{S}}$ is $(\mathcal{T}, b)$-inadmissible for some b such that $T_b \notin \mathcal{T}(\widetilde{\mathcal{S}})$ then it is also $(\mathcal{T}, \widetilde{b})$-inadmissible for any $\widetilde{b} \in \{0, \dots, n\}$.

EXAMPLE 2.16. The family $\mathcal{S}$ introduced in Example 2.13 represents a cycle, denoted as $\mathcal{S}_c$, which is not a $(\mathcal{T}, b)$-inadmissible family for any choice of $b \in \{0, \dots, 3\}$. Indeed, we have here $\mathcal{S}_c = \mathcal{S}$ and $\mathcal{T}(\mathcal{S}_c) = \mathcal{T}(\mathcal{S}) = \mathcal{T}$. Hence, for any $b \in \{0, \dots, 3\}$, the number of dates in the set $\mathcal{T}(\mathcal{S}_c) \setminus \{T_b\}$ is equal to 3 and thus it always coincides with the number of forward swaps in $\mathcal{S}_c$.

REMARK 2.17. The existence a $(\mathcal{T}, b)$-inadmissible cycle $\mathcal{S}_c$ for some $b \in \{0, \dots, n\}$ implies the existence of a $(\mathcal{T}, \widetilde{b})$-inadmissible subset $\widetilde{\mathcal{S}}$ for some $\widetilde{b} \in \{0, \dots, n\}$ (simply take $b = \widetilde{b}$). Note, however, that the existence of a $(\mathcal{T}, \widetilde{b})$-inadmissible subset $\widetilde{\mathcal{S}}$ for some $\widetilde{b} \in \{0, \dots, n\}$ does not imply the existence of a $(\mathcal{T}, b)$-inadmissible cycle $\mathcal{S}_c$, in general. This feature can be illustrated through the following counter-example.

EXAMPLE 2.18. The forward swaps that we will consider here are standard forward swaps, so that all dates between the starting date and maturity date are settlement dates. Let $\widetilde{\mathcal{S}}$ be equal to $\mathcal{S}_c^1 \cup \mathcal{S}_c^2$, where $\mathcal{S}_c^1 = \{S_0 = \{T_0, T_3\}, S_1 = \{T_0, T_4\}, S_2 = \{T_3, T_4\}\}$ and $S_c^2 = \{S_3 = \{T_1, T_2\}, S_4 = \{T_2, T_5\}, S_5 = \{T_1, T_5\}\}$. It is easy to see that $\widetilde{\mathcal{S}}$ is a $(\mathcal{T}, b)$-inadmissible subset for every $b = \{0, \dots, 5\}$. However, the only cycles $\mathcal{S}_c^1$ and $\mathcal{S}_c^2$ in the family $\widetilde{\mathcal{S}}$ are clearly not $(\mathcal{T}, b)$-inadmissible cycles for any $b = \{0, \dots, 5\}$.

We denote by $\mathfrak{C}$ ($\mathfrak{N}_c$, resp.) the class of all families of forward swaps that contain (do not contain, resp.) a cycle.

PROPOSITION 2.19. *The existence of a* $(\mathcal{T}, b)$*-inadmissible subset for some b implies the existence of a cycle. In particular, the strict inclusion* $\mathfrak{N}_c \subsetneq \mathfrak{M}_0$ *holds.*

Proof. By definition, there exists a subset of swaps $\widetilde{\mathcal{S}}$ such that the number of relevant dates (variables) is strictly less than the number of swaps in the $\widetilde{\mathcal{S}}$. Now, let us consider the graph $(\mathcal{T}_0(\widetilde{\mathcal{S}}), \widetilde{\mathcal{S}})$. Since $\widetilde{\mathcal{S}}$ is a $(\mathcal{T}, b)$-inadmissible subset, the number of edges is strictly greater than the number of vertices. By Theorem 2.5, there must exist a cycle (any other relevant date that is not in $\mathcal{T}(\widetilde{\mathcal{S}})$, but falls between the smallest and largest date in $\mathcal{T}(\widetilde{\mathcal{S}})$, manifestly stretches a cycle). We conclude that the inclusion $\mathfrak{N}_c \subset \mathfrak{M}_0$ holds. However, from Example 2.13, we deduce that the inclusion is in fact strict. □

In view of Proposition 2.19, the condition of the nonexistence of a cycle in $\mathcal{S}$ is stronger than the assumption of the nonexistence of a $(\mathcal{T}, b)$-inadmissible subset, which will be shown to be a necessary condition for the weak $\mathcal{T}$-admissibility (see Proposition 2.21). Recall also that Example 2.13 describes a weakly $\mathcal{T}$-admissible family $\mathcal{S}$ that contains a cycle. We thus see that the nonexistence of a cycle is not a necessary condition for the weak $\mathcal{T}$-admissibility, that is, the class $\mathfrak{A}$ is not included in $\mathfrak{N}_c$.

The question whether the nonexistence of a cycle is a sufficient condition for the weak $\mathcal{T}$-admissibility, that is, whether $\mathfrak{N}_c \subsetneq \mathfrak{A}$ (note that Example 2.13 shows that $\mathfrak{N}_c \neq \mathfrak{A}$) remains open (see Question 2.34).

2.4.2. Necessary Conditions for the Weak $\mathcal{T}$-Admissibility. Our goal is to show that the property that a family $\mathcal{S}$ belongs to the class $\mathfrak{M}_0$ is a necessary condition for the weak $\mathcal{T}$-admissibility of a family $\mathcal{S}$. Unfortunately, it appears that this condition is not sufficient for the weak $\mathcal{T}$-admissibility and thus further studies will be needed.

We continue the study of the class $\mathfrak{M}_0$. By a *vertical block* in a matrix, we mean any collection of its rows in which the number of nonnull columns is less than the number of rows. According to the standard terminology of linear algebra, a "vertical block" represents an *overdetermined subsystem*. However, for brevity, we prefer to use a more intuitive term vertical block. By a *permutation* of the matrix C^b we mean a finite number of either row or column swaps (or both) performed on the matrix C^b.

LEMMA 2.20. *Let us fix some $b \in \{0, \dots, n\}$. The following properties are equivalent:*

(i) there exists a $(\mathcal{T}, b)$-inadmissible subset $\widetilde{\mathcal{S}}$,
(ii) there exists a vertical block in the matrix C^b after some permutation.

Proof. It is clear that any $(\mathcal{T}, b)$-inadmissible subset $\mathcal{S}$ corresponds, after a suitable permutation, to a vertical block, so that (i) implies (ii). The converse follows immediately from the definition of a $(\mathcal{T}, b)$-inadmissible subset. □

The following result shows that the existence of a $(\mathcal{T}, b)$-inadmissible subset in $\mathcal{S}$ implies the nonexistence of a unique solution to $C^b\overline{x}^b = \overline{e}^b$, at least for a certain $b \in \{0, \dots, n\}$.

PROPOSITION 2.21. *If a family $\mathcal{S}$ contains a $(\mathcal{T}, b)$-inadmissible subset $\widetilde{\mathcal{S}}$ for some $b \in \{0, \dots, n\}$ then $\mathcal{S}$ is not weakly $\mathcal{T}$-admissible. Consequently, the inclusion $\mathfrak{A} \subset \mathfrak{M}_0$ is valid.*

Proof. It suffices to focus on the variables corresponding to the dates from the set $\mathcal{T}(\widetilde{\mathcal{S}}) \setminus T_b$, where b is such that the subset $\widetilde{\mathcal{S}}$ is $(\mathcal{T}, b)$-inadmissible.

If $T_b \notin \mathcal{T}(\widetilde{\mathcal{S}})$ then these variables need to satisfy an overdetermined homogeneous linear system. We then deal with the following two subcases:

(i) if the determinant of a square subsystem is a nonzero rational function then the null solution is the unique solution to the particular subset of unknowns,
(ii) the determinant of a square subsystem is zero and thus the solution either fails to exist or is not unique.

Therefore, we see that in both cases the family $\mathcal{S}$ is not weakly $\mathcal{T}$-admissible.

If, on the contrary, $T_b \in \mathcal{T}(\widetilde{\mathcal{S}})$ then these variables need to satisfy a nonhomogeneous linear system in which the number of equations is larger than the number of variables. Then by changing the date T_b to some date $T_{\widetilde{b}} \notin \mathcal{T}(\widetilde{\mathcal{S}})$, we obtain once again a homogeneous linear system, in which the number of equations is at least equal to the number of variables. Therefore, we are back in either case (i) or case (ii) (note that if all dates are already in $\mathcal{T}(\widetilde{\mathcal{S}})$, that is, when $\mathcal{T} = \mathcal{T}(\widetilde{\mathcal{S}})$, we can add one more date to $\mathcal{T}$ in order to perform this step). □

It is important to observe that the inclusion $\mathfrak{A} \subset \mathfrak{M}_0$ is in fact strict, that is, $\mathfrak{A} \subsetneq \mathfrak{M}_0$, in general. Specifically, for any $n \geq 2$ there exists a family $\mathcal{S}$ of forward swaps such that $\mathcal{S} \in \mathfrak{A}^c \cap \mathfrak{M}_0$ (e.g., it is fairly easy to produce an example of a family with a cycle, which is not weakly $\mathcal{T}$-admissible). We thus conclude that the property that a family $\mathcal{S}$ belongs to the class $\mathfrak{M}_0$ is a necessary, but not sufficient, condition for the weak $\mathcal{T}$-admissibility of a family $\mathcal{S}$.

As a preliminary step toward identification of a sufficient condition for the weak $\mathcal{T}$-admissibility of a family $\mathcal{S}$, we establish a simple result furnishing a stronger necessary condition.

LEMMA 2.22. *If for some $b \in \{0, \dots, n\}$, the linear system $C^b \overline{x}^b = \overline{e}^b$ associated with $\mathcal{S}$ has a vertical block after a finite number of row operations then the family of forward swaps $\mathcal{S}$ is not weakly $\mathcal{T}$-admissible.*

Proof. Suppose that after a finite number of row operations there exists a vertical block and either:

(i) the vertical block forms a homogeneous overdetermined subsystem or
(ii) the vertical block forms a nonhomogeneous overdetermined subsystem.

In case (i), if any square subsystem has nonzero determinant then the null solution is the unique solution to the subsystem. If, on the contrary, any square subsystem has zero determinant then the original matrix is not invertible.

In case (ii), one can adjust the choice of a numéraire bond. By picking $B(t, T_{\tilde{b}})$ such that x_b is not in the overdetermined subsystem, the nonhomogeneous overdetermined subsystem becomes a homogeneous overdetermined subsystem and we are back in case (i) (similar argument applies if the subsystem is square) and we conclude that the family $\mathcal{S}$ is not weakly $\mathcal{T}$-admissible. □

An $l \times n$ matrix ($l \leq n$) is said to be *diagonalizable* if there exists a permutation such that the diagonal entries of an $l \times l$ sub-matrix are nonzero. In our set-up, we may also formulate the following definition.

DEFINITION 2.23. For a fixed $b \in \{0, \dots, n\}$, we say that the matrix C^b is diagonalizable if there exists a permutation of C^b such that all entries on the first diagonal of C^b are nonzero. We denote by $\mathfrak{D}$ the class of all families of forward swaps such that the matrix C^b is diagonalizable, for all $b \in \{0, \dots, n\}$.

The proof of the following result is deferred to the Appendix.

PROPOSITION 2.24. *The following properties are equivalent, for any fixed $b \in \{0, \dots, n\}$:*

(i) the matrix C^b is diagonalizable,
(ii) there is no $(\mathcal{T}, b)$-inadmissible subset $\widetilde{\mathcal{S}}$ in $\mathcal{S}$.

Consequently, the equality $\mathfrak{D} = \mathfrak{M}_0$ holds.

Recall that, by Lemma 2.20, the existence of a vertical block in the matrix C^b is equivalent to the existence of a $(\mathcal{T}, b)$-inadmissible subset $\widetilde{\mathcal{S}}$. Let us then suppose that for some b there exists a vertical block in C^b after some permutation of variables. It is worth noting that then it may still be true that the matrix C^b can be diagonalized for another choice of b.

EXAMPLE 2.25. Unfortunately, it may also happen that the matrix C^b is diagonalizable for every $b \in \{0, \dots, n\}$, but it fails to have a nonzero determinant for some $b \in \{0, \dots, n\}$.

To illustrate this claim, let us consider the following family of standard forward swaps:

$$\begin{aligned}\mathcal{S} = \{&S_1 = \{T_0, T_4\}, S_2 = \{T_0, T_5\}, S_3 = \{T_0, T_6\}, S_4 = \{T_4, T_5\},\\ &S_5 = \{T_5, T_6\}, S_6 = \{T_1, T_3\}\}.\end{aligned}$$

Then, for every $b \in \{0, 1, 2, 3, 4, 5, 6\}$, the matrix C^b associated with the family $\mathcal{S}$ can be checked to be diagonalizable. However, the determinant of the matrix C^6 vanishes identically, since we have that

$$C^6 = \begin{bmatrix} -1 & a_1\kappa_1 & a_2\kappa_1 & a_3\kappa_1 & 1 + a_4\kappa_1 & 0 \\ -1 & a_1\kappa_2 & a_2\kappa_2 & a_3\kappa_2 & a_4\kappa_2 & 1 + a_5\kappa_2 \\ -1 & a_1\kappa_3 & a_2\kappa_3 & a_3\kappa_3 & a_4\kappa_3 & a_5\kappa_3 \\ 0 & 0 & 0 & 0 & -1 & 1 + a_5\kappa_4 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & a_2\kappa_6 & 1 + a_3\kappa_6 & 0 & 0 \end{bmatrix}, \quad e^6 = \begin{bmatrix} 0 \\ 0 \\ 1 + a_6\kappa_3 \\ 0 \\ 1 + a_6\kappa_5 \\ 0 \end{bmatrix}.$$

It is easy to see that the fourth column can be written as linear combinations of the second and third columns. This implies, of course, that the determinant of C^6 vanishes identically, and thus we conclude that the diagonalizability of the matrix C^b for every b is not a sufficient condition for the weak $\mathcal{T}$-admissibility of a family of forward swaps. Put another way, the property that $\mathcal{S}$ belongs to $\mathfrak{D}$ does not imply that $\mathcal{S}$ is in $\mathfrak{A}$. Since we already know that $\mathfrak{A} \subset \mathfrak{M}_0$ (see Proposition 2.21) and $\mathfrak{D} = \mathfrak{M}_0$ (from Proposition 2.24), we conclude that $\mathfrak{A} \subsetneq \mathfrak{M}_0$.

In what follows, by an *internal date* of a forward swap we mean a relevant date that is neither the start nor the maturity date in a given swap.

DEFINITION 2.26. A family of forward swaps $\mathcal{S}$ is said to have an inadmissible subset if it has the following characteristics: there exist k dates for some $k \geq 2$ such that

(i) they are internal dates to at least two swaps and
(ii) they are relevant to at most $k - 2$ other swaps.

We denote by $\mathfrak{M}_1$ the class of all families of forward swaps that do not contain an inadmissible subset.

The next result furnishes a necessary condition for the weak $\mathcal{T}$-admissibility of a family $\mathcal{S}$.

PROPOSITION 2.27. *The nonexistence of inadmissible subsets is necessary for a family of swaps to be weakly $\mathcal{T}$-admissible, so that $\mathfrak{A} \subseteq \mathfrak{M}_1$.*

Proof. Suppose a family of forward swaps $\mathcal{S}$ has an inadmissible subset. To show that $\mathcal{S}$ is not weakly $\mathcal{T}$-admissible, we pick b outside of those k internal dates and concentrate on those k internal dates (as usual, represented by the columns in C^b). Without loss of generality, we may pick one column as the pivot column and perform column operations to eliminate other columns. As a consequence, we will end up with the subsystem without the pivot column, which consists of $k - 1$ columns and at most $k - 2$ nonzero rows.

This gives the linear dependence of the columns. The following example will make this argument more explicit; a moment's reflection will show the generality of our reasoning. To illustrate the method, we focus on columns number two, three, and four in the matrix in Example 2.25

$$\begin{bmatrix} a_1\kappa_1 & a_2\kappa_1 & a_3\kappa_1 \\ a_1\kappa_2 & a_2\kappa_2 & a_3\kappa_2 \\ a_1\kappa_3 & a_2\kappa_3 & a_3\kappa_3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & a_2\kappa_6 & 1 + a_3\kappa_6 \end{bmatrix} \xrightarrow[c_3 = \frac{a_3}{a_1}c_1 - c_3]{c_2 = \frac{a_2}{a_1}c_1 - c_2} \begin{bmatrix} a_1\kappa_1 & 0 & 0 \\ a_1\kappa_2 & 0 & 0 \\ a_1\kappa_3 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & a_2\kappa_6 - \frac{a_2}{a_1} & 1 + a_3\kappa_6 - \frac{a_3}{a_1} \end{bmatrix}.$$

It is now clear that the columns are linearly dependent. By definition, this implies this family of swap is not weakly $\mathcal{T}$-admissible. □

PROPOSITION 2.28. *The existence of an inadmissible subset implies the existence of a cycle, so that $\mathfrak{M}_1^c \subset \mathfrak{C}$. Moreover, the strict inclusion $\mathfrak{N}_c \subsetneq \mathfrak{M}_1$ holds, in general.*

Proof. By definition of the class $\mathfrak{M}_1^c$, there exist k internal dates such that at most $k - 2$ swaps either start or mature on those k internal dates. Therefore, there must be at least $n - k + 2$ swaps starting and maturing outside the k internal dates. If one consider the graph of $\mathcal{S}$ without those k internal dates, it is a graph with $n - k$ vertices and at least $n - k + 2$ edges and thus, by Theorem 2.5, a cycle necessarily exists in $\mathcal{S}$. Hence, the inclusion $\mathfrak{M}_1^c \subset \mathfrak{C}$ is valid. For the second inclusion, we also observe that Proposition 2.27 yields $\mathfrak{A} \subseteq \mathfrak{M}_1$, whereas Example 2.13 describes a family $\mathcal{S}$ with $n = 3$ belonging to $\mathfrak{A} \cap \mathfrak{C}$. This also means that $\mathfrak{M}_1^c \subsetneq \mathfrak{C}$, in general. □

2.4.3. Sufficient Conditions for the Weak $\mathcal{T}$-Admissibility. Throughout this section, we set $l = n$ and we consider an arbitrary family $\mathcal{S} = \{S_1, \dots, S_n\}$ of forward swaps. Our next goal is to provide a sufficient condition for the weak $\mathcal{T}$-admissibility of a family $\mathcal{S}$, that is, the existence and uniqueness of a nonzero solution to the linear system $C^b\overline{x}^b = \overline{e}^b$, for an arbitrary choice of $b \in \{0, \dots, n\}$, for almost every $(\kappa_1, \dots, \kappa_n) \in \mathbb{R}^n$.

DEFINITION 2.29. We denote by $\overline{\mathfrak{D}}$ the class of all diagonalizable families $\mathcal{S}$ of n forward swaps such that the diagonalizability property of the matrix C^b is preserved under elementary row operations, for all $b \in \{0, \dots, n\}$.

REMARK 2.30. From Example 2.25, we see that the diagonalizability of the matrix C^b is not preserved under column operations. Since the row rank and column rank of any matrix are the same, for simplicity, we decided to work with row operations in the above definition.

It is fair to acknowledge that although the class $\overline{\mathfrak{D}}$ is suitable to give a necessary and sufficient condition for the weak $\mathcal{T}$-admissibility, it may be difficult to check whether a given family $\mathcal{S}$ does indeed belong to the class $\overline{\mathfrak{D}}$ or not.

PROPOSITION 2.31.

(i) If a family $\mathcal{S}$ belongs to $\overline{\mathfrak{D}}$ then $\mathcal{S}$ is weakly $\mathcal{T}$-admissible and thus the inclusion $\overline{\mathfrak{D}} \subset \mathfrak{A}$ is valid.

(ii) The converse inclusion $\mathfrak{A} \subset \overline{\mathfrak{D}}$ holds as well.

Consequently, the equality $\mathfrak{A} = \overline{\mathfrak{D}}$ is true.

Proof. We first prove (i), that is, we establish a sufficient condition for the weak $\mathcal{T}$-admissibility of $\mathcal{S}$. By assumption, a family $\mathcal{S}$ is in $\mathfrak{D}$ and thus the matrix C^b is diagonalizable for every $b \in \{0, \ldots, n\}$. Therefore, for a fixed $b \in \{0, \ldots, n\}$, the matrix C^b can be diagonalized. Next, we take the first row as the pivot row and we eliminate the entries below the first diagonal and, by the assumption that $\mathcal{S}$ belongs to $\overline{\mathfrak{D}}$, the sub-matrix is again diagonalizable. The procedure then continues with the second row becoming the pivot row, so that the entries below the second diagonal are eliminated. At each stage, the assumption that diagonalizability is preserved is used to diagonalize the matrix after each row operation. After completing this procedure, we obtain a matrix in row echelon form with the diagonals consisting of nonzero rational functions in variables $\kappa_1, \ldots, \kappa_n$. It is then easy to see that the determinant is nonzero, for almost all generic values of $(\kappa_1, \ldots, \kappa_n) \in \mathbb{R}^n$. This ends the proof of the inclusion $\overline{\mathfrak{D}} \subset \mathfrak{A}$.

For part (ii), we need to show that $\mathfrak{A} \subset \overline{\mathfrak{D}}$. We already know from Propositions 2.21 and 2.24 that the inclusion $\mathfrak{A} \subset \mathfrak{D}$ holds. It thus suffices to show that $\overline{\mathfrak{D}}^c \cap \mathfrak{A} = \emptyset$. To this end, we will argue by contradiction. Let us then assume that the diagonalizability property is not preserved under elementary row operations for all $b \in \{0, \ldots, n\}$, but the linear system $C^b \overline{x}^b = \overline{e}^b$ associated with $\mathcal{S}$ has a unique nonzero solution for any choice of b. The first condition implies that for some $\widetilde{b} \in \{0, \ldots, n\}$, the matrix $C^{\widetilde{b}}$ is no longer diagonalizable (that is, there exists a vertical block) after a finite number of elementary row operations. Using Lemma 2.22, we conclude that the family $\mathcal{S}$ is not weakly $\mathcal{T}$-admissible. However, we also know that the solution of the linear system is invariant under elementary row operations and thus we arrive at the contradiction. □

REMARK 2.32. We emphasize once again that it is important to assume that, for all $b \in \{0, \ldots, n\}$, the matrix C^b belongs to $\overline{\mathfrak{D}}$, that is, the diagonalization property is invariant under elementary row operations. The weaker assumption that for all $b \in \{0, \ldots, n\}$, the matrix C^b is diagonalizable is not sufficient for the weak $\mathcal{T}$-admissibility, as was illustrated through Example 2.25.

We will now summarize the results regarding the weak $\mathcal{T}$-admissibility of a family $\mathcal{S}$ obtained so far. For the reader's convenience, we recall the notation for the relevant classes of families $\mathcal{S}$ of forward swaps:

$\mathfrak{A}$ – weakly $\mathcal{T}$-admissible families;
$\mathfrak{C}$ – families that contain a cycle;
$\mathfrak{N}_c$ – families that do not contain a cycle;
$\mathfrak{D}$ – families such that the matrix C^b is diagonalizable, for all $b \in \{0, \ldots, n\}$;
$\mathfrak{M}_0$ – families that do not contain a $(\mathcal{T}, b)$-inadmissible subset, for all $b \in \{0, \ldots, n\}$;
$\mathfrak{M}_1$ – families that do not contain an inadmissible subset;
$\overline{\mathfrak{D}}$ – families from $\mathfrak{D}$ for which the diagonalization property of C^b is preserved under elementary row operations, for all $b \in \{0, \ldots, n\}$.

THEOREM 2.33. *Let $\mathcal{T} = \{T_0, \dots, T_n\}$ be a fixed tenor structure. We consider the class of all families $\mathcal{S} = \{S_1, \dots, S_n\}$ of forward swaps associated with $\mathcal{T}$. Then the following properties hold:*

(i) $\mathfrak{A} = \overline{\mathfrak{D}} \subsetneq \mathfrak{D} = \mathfrak{M}_0$;
(ii) $\mathfrak{A} \subset \mathfrak{D} \cap \mathfrak{M}_1 \subset \mathfrak{D} = \mathfrak{M}_0$;
(iii) $\mathfrak{N}_c \subsetneq \mathfrak{D} = \mathfrak{M}_0$ and $\mathfrak{N}_c \subsetneq \mathfrak{M}_1$;
(iv) $\mathfrak{A} \cap \mathfrak{C} \neq \emptyset$.

Proof. It suffices to put together the partial results established thus far:

(i) it suffices to combine Propositions 2.21, 2.24, and 2.31 with Example 2.25;
(ii) the first inclusion follows from Propositions 2.21 and 2.27; the second is obvious;
(iii) Propositions 2.19 and 2.24 yield the first part in (iii); the second part follows from Proposition 2.28;
(iv) Example 2.13 describes a family $\mathcal{S}$ with $n = 3$ belonging to $\mathfrak{A} \cap \mathfrak{C}$. □

Recall that we have argued that can (non)existence of a cycle is not of primary interest, as opposed to the (non)existence of a $(\mathcal{T}, b)$-admissible subset. As already mentioned, when analyzing the necessity clause, Galluccio et al. (2007) failed to realize that not every cycle is a $(\mathcal{T}, b)$-inadmissible subset. Unfortunately, the sufficiency clause in the main result in Galluccio et al. (2007) remains still unclear to us, since we were unable to either establish the inclusion $\mathfrak{N}_c \subsetneq \mathfrak{A}$ or invalidate it by means of a counterexample. This motivates us to formulate the following problem which, to the best of our knowledge, remains open.

QUESTION 2.34. Is the nonexistence of a cycle in a family of forward swaps $\mathcal{S}$ a sufficient condition for the weak $\mathcal{T}$-admissibility of $\mathcal{S}$, i.e., does the inclusion $\mathfrak{N}_c \subsetneq \mathfrak{A}$ hold true?

Alternatively, one may attempt to answer the following question.

QUESTION 2.35. Is the diagonalizability property of a family of forward swaps $\mathcal{S}$, when combined with the nonexistence of an inadmissible subset, a sufficient condition for the weak $\mathcal{T}$-admissibility of $\mathcal{S}$, i.e., does the inclusion $\mathfrak{D} \cap \mathfrak{M}_1 \subset \mathfrak{A}$ hold true?

In view of part (ii) in Theorem 2.33, we see that if the answer to Question 2.35 is positive, then one obtains in fact the equality $\mathfrak{A} = \mathfrak{D} \cap \mathfrak{M}_1$. Hence, in view of part (iii) in Theorem 2.33, the answer to Question 2.34 would be positive as well.

2.5. $\mathcal{T}$-Admissibility of Forward Swaps

We continue working under the assumption that $l = n$. Recall that, by Definition 2.10, a family of forward swaps is $\mathcal{T}$-admissible, if it is weakly $\mathcal{T}$-admissible and the solution to the linear system associated with that family is strictly positive for all generic values of $(\kappa_1, \dots, \kappa_n) \in \mathbb{R}^n_+$. In this section, we will give conditions for which a weakly $\mathcal{T}$-admissible family is $\mathcal{T}$-admissible.

The proof of Proposition 2.36 follows rather closely the proof of the sufficiency clause in Theorem 1 of Pietersz and Regenmortel (2006). It is important to point out, however, that their result furnishes the sufficient and necessary conditions for the existence of deflated bonds when one is given a family of swaps that enjoys the property that no two swaps start on the same date. The set-up examined in Pietersz and Regenmortel (2006) is thus different from ours and the positivity of the deflated bond prices in their paper was simply obtained as a by-product of the existence of deflated bond prices up to the bonds' maturity dates. One can observe that in fact the assumption of positivity of the deflated bond prices was not used at all in the proof of the necessity clause in theorem 1 in Pietersz and Regenmortel (2006). For this reason, we need to re-examine the necessity clause. For the reader's convenience, we first give the proof of the sufficiency clause.

PROPOSITION 2.36. *Assume that a family $\mathcal{S}$ of forward swaps is weakly $\mathcal{T}$-admissible. If no two swaps start on the same date then a family $\mathcal{S}$ is $\mathcal{T}$-admissible.*

Proof. It is sufficient to demonstrate that a positive solution exists for $b = n$. Assuming that no two forward swaps start on the same date, one can arrange the rows of the matrix C^n so that it becomes an upper triangular matrix with nonzero entries on the diagonal. It is thus clear that the family $\mathcal{S}$ is weakly $\mathcal{T}$-admissible. We will establish the positivity of solution by induction and back substitution. By picking $b = n$, we have $B^n(t, T_n) = 1$, which is strictly positive and greater or equal to 1. Let us assume that $B^n(t, T_j) \geq 1$ for all $j > i$. We would like to show $B^n(t, T_i) \geq 1$ is positive for all generic values of $(\kappa_1, \dots \kappa_n) \in \mathbb{R}^n_+$. By the back substitution algorithm, we obtain

$$B^n(t, T_i) = \kappa_i \sum_{j=i+1}^{m_i} a_j B^n(t, T_j) + B^n(t, T_{m_i}), \quad a_j \geq 0, \quad \forall j = \{i, \dots, m_j\}. \tag{2.4}$$

Next, by the induction hypothesis and the assumption that $(\kappa_1, \dots \kappa_n) \in \mathbb{R}^n_+$, the following holds

$$\kappa_i \sum_{j=i+1}^{m_i} a_j B^n(t, T_j) \geq 0, \qquad B^n(t, T_{m_i}) \geq 1.$$

It now follows easily from (2.4) that $B^n(t, T_i) \geq 1$ for almost all generic values of $(\kappa_1, \dots, \kappa_n) \in \mathbb{R}^n_+$. Note that it was important to assume that the matrix was upper-triangular with nonzero diagonal, since otherwise the back substitution algorithm would not apply. □

Our next goal is to establish a partial converse of Proposition 2.36. For this purpose, we need to introduce an additional property of a family $\mathcal{S}$.

Property (A). A family of forward swaps $\mathcal{S}$ is said to satisfy *Property (A)* if for any two forward swaps S_i and S_k from $\mathcal{S}$ such that $T_{s_i} = T_{s_k}$ and $T_{m_k} < T_{m_i}$, we have that $\mathcal{T}(S_k) \subset \mathcal{T}(S_i)$.

The proof of the next result is relegated to the Appendix.

PROPOSITION 2.37. *If a family $\mathcal{S}$ is $\mathcal{T}$-admissible and satisfies Property (A) then no two swaps in $\mathcal{S}$ starting on the same date exist.*

Let us point out that Property (A) is crucial for the proof of the above proposition. Of course, this assumption is satisfied in the case of standard forward swaps. We now give an example of a family with two forward swaps starting at the same date, for which the solution is strictly positive.

EXAMPLE 2.38. Consider the family $\mathcal{S} = \{S_1 = \{T_0, T_2\}, S_2 = \{T_1, T_3\}, S_3 = \{T_0, T_3\}\}$. Then the associated linear system for $b = 0$ is given by

$$C^0\overline{x}^0 = \begin{bmatrix} 0 & 1+\widetilde{a}_2\kappa_1 & 0 \\ -1 & 0 & 1+\widetilde{a}_3\kappa_2 \\ 0 & 0 & 1+\widetilde{a}_3\kappa_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \overline{e}^0.$$

It is easy to see that this family of forward swaps does not satisfy Property (A). By solving the above system, we obtain the unique solution

$$\overline{x}^0 = \begin{bmatrix} \dfrac{1+\widetilde{a}_3\kappa_2}{1+\widetilde{a}_3\kappa_3} \\ (1+\widetilde{a}_2\kappa_1)^{-1} \\ (1+\widetilde{a}_3\kappa_3)^{-1} \end{bmatrix}.$$

From the form of the solution, it is also readily seen that the deflated bonds prices are strictly positive for all generic values of $(\kappa_1, \kappa_2, \kappa_3) \in \mathbb{R}^3_+$.

3. ADMISSIBLE MARKET MODELS OF FORWARD SWAP RATES

In this section, we will examine the second issue mentioned in the Introduction, that is, the problem whether the joint dynamics of a given family of forward swap rates is uniquely determined. To this end, we will focus on the Radon–Nikodým densities of the corresponding family of *forward swap measures*. The crucial object in the study of forward swap measures is in turn the *swap numéraire*, that is, the denominator appearing in the definition of the forward swap rate. For a detailed analysis of some special cases of models considered in this section, such as the LIBOR market model and the market model for co-terminal swaps, we refer to Brace et al. (1997), Davis and Mataix-Pastor (2007), Jamshidian (1997, 1999, 2008), Musiela and Rutkowski (1997), Rutkowski (1999) (see also, chapters 12 and 13 in the monograph by Musiela and Rutkowski 2005 and the references therein). Due to space limitations, we will not deal here either with any particular examples of market models of forward swap rates, or with other related questions, such as the issue of positivity of rates whose dynamics are not directly postulated, but are implicitly given by a specification of a market model for a certain family $\mathcal{S}$ of forward swaps (that is, the out-of-sample swap rates).

3.1. Inverse Problem for Swap Annuities

Recall that by a forward swap, we mean the start date T_{s_j}, the maturity date T_{m_j} as well as a pair $P_t^{s_j,m_j}$, A^{s_j,m_j} of processes that are given by (1.2) and (1.3), respectively. The strictly positive process A^{s_j,m_j} is called the *swap numéraire* for the jth forward swap. The forward swap rate κ^j is defined as follows:

$$\kappa_t^j = \kappa_t^{s_j,m_j} = \frac{P_t^{s_j,m_j}}{A_t^{s_j,m_j}}, \quad \forall\, t \in [0, T_{s_j}]. \tag{3.1}$$

As before, we consider a finite family $\mathcal{S} = \{S_1, \dots, S_l\}$ of forward swaps.

DEFINITION 3.1. For any $j, d \in \{1, \dots, l\}$, the annuity deflated swap numéraire is given by

$$\widetilde{A}_t^{d,j} := \frac{A_t^{s_j,m_j}}{A_t^{s_d,m_d}}, \quad \forall\, t \in [0, T_{s_j} \wedge T_{s_d}]. \tag{3.2}$$

One should take note here that the *annuity deflated swap numéraire* process $\widetilde{A}^{d,j}$ is in fact the process of our primary interest, since it will later on act as the Radon–Nikodým density process in the specification of the joint dynamics of forward swap rates $\kappa^1, \dots, \kappa^l$ under a single probability measure. This observation motivates us to formulate the following inverse problem for annuity deflated swap numéraires: provide conditions under which the family of processes $\widetilde{A}^{d,j}$, as specified by equations (1.2), (1.3), (3.1), and (3.2), admit a unique representation in terms of forward swap rates $\kappa^1, \dots, \kappa^l$ only. In essence, we would like to eliminate bond prices from (1.2), (1.3), (3.1), and (3.2) to express $\widetilde{A}^{d,j}$ in terms of fixed, but arbitrary, parameters $\kappa^1, \dots, \kappa^l$.

Recall that the *deflated swap annuity* is given by (see (1.7))

$$A_t^{b,s_j,m_j} := \frac{A_t^{s_j,m_j}}{B(t, T_b)} = \sum_{i \in \mathcal{A}_j} \widetilde{a}_i B^b(t, T_i), \tag{3.3}$$

where $\mathcal{A}_j$ is the set of all indices corresponding to settlement dates in the forward swap S_j. Hence (3.2) can also be represented as follows:

$$\widetilde{A}_t^{d,j} = \frac{\sum_{i \in \mathcal{A}_j} \widetilde{a}_i B^b(t, T_i)}{\sum_{k \in \mathcal{A}_d} \widetilde{a}_k B^b(t, T_k)} = \frac{A_t^{b,s_j,m_j}}{A_t^{b,s_d,m_d}}, \quad \forall\, t \in [0, T_b \wedge T_{s_j} \wedge T_{s_d}], \tag{3.4}$$

where the second equality is clear since the *deflated swap annuity* A_t^{b,s_j,m_j} is defined by the equality $A_t^{b,s_j,m_j} := A_t^{s_j,m_j} / B(t, T_b)$, for any choice of $b \in \{0, \dots, n\}$. In addition, we write

$$\widetilde{P}_t^{d,j} := \frac{P_t^{s_j,m_j}}{A_t^{s_d,m_d}} = \frac{B(t, T_{s_j}) - B(t, T_{m_j})}{\sum_{k \in \mathcal{A}_d} \widetilde{a}_k B(t, T_k)} = \frac{B^b(t, T_{s_j}) - B^b(t, T_{m_j})}{\sum_{k \in \mathcal{A}_d} \widetilde{a}_k B^b(t, T_k)},$$

so that (3.1) yields, for every $t \in [0, T_b \wedge T_{s_j} \wedge T_{s_d}]$,

$$\kappa_t^j = \kappa_t^{s_j,m_j} = \frac{\widetilde{P}_t^{d,j}}{\widetilde{A}_t^{d,j}} = \frac{B^b(t, T_{s_j}) - B^b(t, T_{m_j})}{\widetilde{A}_t^{d,j} \sum_{k \in \mathcal{A}_d} \widetilde{a}_k B^b(t, T_k)}.$$

Using the notation for generic values of stochastic processes appearing in the formula above, we get

$$\kappa_j = \frac{x_{s_j} - x_{m_j}}{\alpha_{d,j} \sum_{k \in \mathcal{A}_d} \widetilde{a}_k x_k}, \quad j = 1, \ldots, l, \tag{3.5}$$

and

$$\alpha_{d,j} = \frac{\sum_{i \in \mathcal{A}_j} \widetilde{a}_i x_i}{\sum_{k \in \mathcal{A}_d} \widetilde{a}_k x_k}, \quad d, j = 1, \ldots, l. \tag{3.6}$$

Inverse Problem (IP.2): We say that a family $\mathcal{S} = \{S_1, \ldots, S_l\}$ admits a solution to the *inverse problem* (IP.2) if, every $d, j \in \{1, \ldots, l\}$, the annuity deflated swap numéraire $\widetilde{A}_t^{d,j}$ can be uniquely expressed as a function of forward swap rates $\kappa_t^1, \ldots, \kappa_t^l$. More formally, for any fixed $d \in \{1, \ldots, l\}$ and for almost all $(\kappa_1, \ldots, \kappa_l) \in \mathbb{R}^l$, the system of equations (3.5) and (3.6) can be solved for $(\alpha_{d,1}, \ldots, \alpha_{d,l})$ and a unique solution is given in terms of $(\kappa_1, \ldots, \kappa_l)$ only, i.e., it does not depend explicitly on variables $x_0, \ldots, x_n$.

The following definition describes the class of families of forward swaps possessing desirable properties from the viewpoint of dynamical properties of a market model.

DEFINITION 3.2. We say that a family $\mathcal{S}$ of forward swaps associated with $\mathcal{T}$ is *weakly* $\mathcal{A}$-admissible if for any choice of $d \in \{1, \ldots, l\}$ the following property holds: for almost every $(\kappa_1, \ldots, \kappa_l) \in \mathbb{R}^l$ there exists a unique nonzero solution $(\alpha_{d,1}, \ldots, \alpha_{d,l}) \in \mathbb{R}^l$ to the system of equations (3.5) associated with $\mathcal{S}$ and it is given in terms of $(\kappa_1, \ldots, \kappa_l)$ only. If, in addition, the solution is strictly positive for almost every $(\kappa_1, \ldots, \kappa_l) \in \mathbb{R}^l_+$ then we say that a family $\mathcal{S}$ is $\mathcal{A}$-admissible.

We will now examine the relationship between the concepts of the (weak) $\mathcal{T}$-admissibility and (weak) $\mathcal{A}$-admissibility. Lemma 3.3 shows that the (weak) $\mathcal{T}$-admissibility of a family $\mathcal{S}$ implies its (weak) $\mathcal{A}$-admissibility; the converse does not hold, however, as will be demonstrated by means of a counter-example (see Example 3.4).

LEMMA 3.3. *The (weak) $\mathcal{T}$-admissibility of a family $\mathcal{S}$ implies its (weak) $\mathcal{A}$-admissibility.*

Proof. If the family $\mathcal{S}$ is weakly $\mathcal{T}$-admissible then the uniqueness of a solution to the inverse problem (IP.1) holds and the unique vector of nonzero deflated bond prices $\bar{x}^b$ admits a representation in terms of $(\kappa_1, \ldots, \kappa_l) \in \mathbb{R}^l$. Consequently, in view of (3.4), we conclude that $\mathcal{S}$ is weakly $\mathcal{A}$-admissible. Furthermore, if $\mathcal{S}$ is $\mathcal{T}$-admissible then for almost every $(\kappa_1, \ldots, \kappa_l) \in \mathbb{R}^l_+$ the solution to the inverse problem (IP.1) is strictly positive and thus $\alpha_{d,j} = \alpha_{d,j}(\kappa_1, \ldots, \kappa_l)$ is strictly positive as well. □

EXAMPLE 3.4. It is worth stressing that the weak $\mathcal{T}$-admissibility is not a necessary condition for the weak $\mathcal{A}$-admissibility of a family $\mathcal{S}$, since it is possible to produce an example of a family of forward swaps, which is weakly $\mathcal{A}$-admissible, but fails to be weakly

$\mathcal{T}$-admissible. For this purpose, we set $n = 3$ and we consider the standard forward swaps $S_1 = \{T_0, T_4\}$, $S_2 = \{T_0, T_3\}$, and $S_3 = \{T_2, T_3\}$. The linear system associated with the family $\mathcal{S} = \{S_1, S_2, S_3\}$ for $b = 0$ reads

$$\begin{bmatrix} a_1\kappa_1 & a_2\kappa_1 & a_3\kappa_1 & 1 + a_4\kappa_1 \\ a_1\kappa_2 & a_2\kappa_2 & 1 + a_3\kappa_2 & 0 \\ 0 & 0 & -1 & 1 + a_4\kappa_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}.$$

A general solution to the corresponding inverse problem (IP.1) is thus given by

$$\overline{x}^0 = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -\dfrac{(-a_2\kappa_1\gamma - a_2\kappa_1a_4\kappa_3\gamma + a_2\kappa_2\gamma + a_2\kappa_1\kappa_2a_4\gamma) + a_4\kappa_3 - a_3\kappa_1 + a_3\kappa_2 - a_3\kappa_1a_4\kappa_3 + a_3a_4\kappa_3\kappa_2 - a_4\kappa_1}{(-\kappa_1 - \kappa_1a_4\kappa_3 + \kappa_2 + \kappa_2a_4\kappa_1)\,a_1} \\ \gamma \\ -\dfrac{(1 + a_4\kappa_3)(\kappa_1 - \kappa_2)}{-\kappa_1 - \kappa_1a_4\kappa_3 + \kappa_2 + \kappa_2a_4\kappa_1} \\ -\dfrac{\kappa_1 - \kappa_2}{-\kappa_1 - \kappa_1a_4\kappa_3 + \kappa_2 + \kappa_2a_4\kappa_1} \end{bmatrix},$$

where γ is a generic value of the deflated bond price $B^0(t, T_2)$. It is thus clear that the family $\mathcal{S}$ is not weakly $\mathcal{T}$-admissible, since the required uniqueness of a solution to the system $C^0\overline{x}^0 = \overline{e}^0$ does not hold.

We will now show that the family $\mathcal{S}$ is weakly $\mathcal{A}$-admissible. To this end, we substitute a solution into the deflated swap annuity equation (1.7) (or, equivalently, (3.3)), to obtain the following expressions for generic values α_{0,s_j,m_j}, $j = 1, 2, 3$ of A_t^{0,s_j,m_j}, $j = 1, 2, 3$

$$\alpha_{0,s_1,m_1} = \frac{a_4(-\kappa_3 + \kappa_2)}{-\kappa_1 - \kappa_1a_4\kappa_3 + \kappa_2 + \kappa_2a_4\kappa_1},$$

$$\alpha_{0,s_2,m_2} = \frac{a_4(-\kappa_3 + \kappa_1)}{-\kappa_1 - \kappa_1a_4\kappa_3 + \kappa_2 + \kappa_2a_4\kappa_1},$$

$$\alpha_{0,s_3,m_3} = -\frac{\kappa_1 - \kappa_2}{-\kappa_1 - \kappa_1a_4\kappa_3 + \kappa_2 + \kappa_2a_4\kappa_1}.$$

We conclude that the deflated swap annuities can be uniquely expressed as functions of forward swaps rates $(\kappa_1, \kappa_2, \kappa_3) \in \mathbb{R}^3$ and thus, in view of (3.4), it is now easy to see that the considered family of forward swaps is indeed weakly $\mathcal{A}$-admissible.

Let us now consider a family of forward swaps $\mathcal{S} = \{S_1, \ldots, S_l\}$, which admits a nonzero (but possibly nonunique) solution to the inverse problem (IP.1) for some b. This means that the following equalities hold for some functions $g^{b,k}$ and $g_j^{b,k}$:

$$x_k = g^{b,k}(\kappa_1, \ldots, \kappa_l) + \sum_{j=1}^{i} g_j^{b,k}(\kappa_1, \ldots, \kappa_l)x_{n_j}, \quad \forall\, T_k \in \mathcal{T}, \tag{3.7}$$

where the variables $x_{n_1}, \ldots, x_{n_i}$ correspond to the sub-family of deflated bond prices which parameterize the solution to the linear system associated with $\mathcal{S}$. By substituting equation (3.7) into equation (3.3), we arrive at the following equalities:

$$\alpha^{b,s_j,m_j} = \widetilde{g}^{b,j}(\kappa_1,\dots,\kappa_l) + \sum_{k=1}^{i} \widetilde{g}_k^{b,j}(\kappa_1,\dots,\kappa_l)x_{n_k} \tag{3.8}$$

for certain functions $\widetilde{g}^{b,j}$ and $\widetilde{g}_k^{b,j}$. It is obvious that, independently of the choice of b, the annuity deflated swap numeráire can now be expressed as follows:

$$\alpha_{d,j} = \frac{\widetilde{g}^{b,j}(\kappa_1,\dots,\kappa_l) + \sum_{k=1}^{i} \widetilde{g}_k^{b,j}(\kappa_1,\dots,\kappa_l)x_{n_k}}{\widetilde{g}^{b,d}(\kappa_1,\dots,\kappa_l) + \sum_{k=1}^{i} \widetilde{g}_k^{b,d}(\kappa_1,\dots,\kappa_l)x_{n_k}}. \tag{3.9}$$

We observe that the family $\mathcal{S}$ is weakly $\mathcal{A}$-admissible, provided that equation (3.9) can be reduced to obtain

$$\alpha_{d,j} = G^{d,j}(\kappa_1,\dots,\kappa_l) \tag{3.10}$$

for some nonzero rational functions $G^{d,j} : \mathbb{R}^l \to \mathbb{R}$. However, if the parameterizing family of bond prices cannot be completely eliminated from (3.9) for some annuity deflated swap numéraire $\alpha_{d,j}$ then the family of annuity deflated swap numéraires $\widetilde{A}^{d,j}$, $d, j = 1,\dots,l$ is not uniquely determined by the forward swap rates, which in turn means the family $\mathcal{S}$ of forward swaps fails to be weakly $\mathcal{A}$-admissible.

3.2. Dynamics of Forward Swap Rates

We consider here a family of forward swap rate processes $\{\kappa^1,\dots,\kappa^l\}$. Our goal is to show that, under mild conditions imposed on a family of forward swap rate processes, their joint dynamics are uniquely specified by a family of volatility processes with respect to some spanning martingales. We first recall some results from stochastic calculus, which will be used in what follows.

3.2.1. Girsanov's Transforms. Throughout this section, the multidimensional Itô integral should be interpreted as the vector stochastic integral (see, for instance, Shiryaev and Cherny 2002). Let us first quote a general version of the Girsanov theorem (see, for instance, Brémaud and Yor 1978 or theorem 9.4.4.1 in Jeanblanc et al. 2009).

PROPOSITION 3.5. *Let $\widetilde{\mathbb{P}}$ and $\mathbb{P}$ be equivalent probability measures on $(\Omega, \mathcal{F}_T)$ with the Radon–Nikodým density process*

$$Z_t = \frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}}\bigg|_{\mathcal{F}_t}, \quad \forall\, t \in [0, T]. \tag{3.11}$$

Suppose that M is a $(\mathbb{P}, \mathbb{F})$-local martingale. Then the process

$$\widetilde{M}_t = M_t - \int_{(0,t]} \frac{1}{Z_s}\, d[Z, M]_s$$

is a $(\widetilde{\mathbb{P}}, \mathbb{F})$-local martingale.

Let $\mathcal{M}_{loc}(\mathbb{P})$ ($\mathcal{M}(\mathbb{P})$, resp.) stand for the class of all $(\mathbb{P},\mathbb{F})$-local martingales ($(\mathbb{P},\mathbb{F})$-martingales, resp.). Assume that Z is a strictly positive $(\mathbb{P},\mathbb{F})$-local martingale such that $Z_0 = 1$. Also let an equivalent probability measure $\widetilde{\mathbb{P}}$ be given by (3.11). Then the linear map $\Psi_Z : \mathcal{M}_{loc}(\mathbb{P}) \to \mathcal{M}_{loc}(\widetilde{\mathbb{P}})$ given by the formula

$$\Psi_Z(M) = M_t - \int_{(0,t]} \frac{1}{Z_s}\, d[Z, M]_s, \quad \forall\, M \in \mathcal{M}_{loc}(\mathbb{P}), \tag{3.12}$$

is called the *Girsanov transform* associated with the Radon–Nikodým density process Z. By the symmetry of the problem, the process Z^{-1} is the Radon–Nikodým density of $\mathbb{P}$ with respect to $\widetilde{\mathbb{P}}$. The corresponding Girsanov transform $\Psi_{Z^{-1}} : \mathcal{M}_{loc}(\widetilde{\mathbb{P}}) \to \mathcal{M}_{loc}(\mathbb{P})$ associated with Z^{-1} is thus given by the formula

$$\Psi_{Z^{-1}}(\widetilde{M}) = \widetilde{M}_t - \int_{(0,t]} Z_s\, d[Z^{-1}, \widetilde{M}]_s, \quad \forall\, \widetilde{M} \in \mathcal{M}_{loc}(\widetilde{\mathbb{P}}).$$

PROPOSITION 3.6. *Let $\widetilde{\mathbb{P}}$ be a probability measure equivalent to $\mathbb{P}$ on $(\Omega, \mathcal{F}_T)$ with the Radon–Nikodým density process Z. Then for any $(\widetilde{\mathbb{P}},\mathbb{F})$-local martingale $\widetilde{N}$ there exists a $(\mathbb{P},\mathbb{F})$-local martingale $\widehat{N}$ such that*

$$\widetilde{N}_t = \widehat{N}_t - \int_{(0,t]} \frac{1}{Z_s}\, d[Z, \widehat{N}]_s. \tag{3.13}$$

The process $\widehat{N}$ is given by the formula

$$\widehat{N}_t = \widetilde{N}_0 + \int_{(0,t]} \frac{1}{Z_{s-}}\, dL_s - \int_{(0,t]} \frac{L_{s-}}{Z_{s-}^2}\, dZ_s, \tag{3.14}$$

where we denote $L = \widetilde{N}Z$.

From Proposition 3.6, it follows immediately that the process $\widehat{N}$ given by (3.14) belongs to the set $(\Psi_Z)^{-1}(\widetilde{N})$. In fact, we have that $\widehat{N} = (\Psi_Z)^{-1}(\widetilde{N})$, as the following well-known result shows.

LEMMA 3.7. *Let $\widetilde{N}$ be any $(\widetilde{\mathbb{P}},\mathbb{F})$-local martingale and let the process $\widehat{N}$ be given by formula (3.14) with $L = \widetilde{N}Z$. Then the process $\widehat{N}$ is also given by the following expression:*

$$\widehat{N}_t = \widetilde{N}_t - \int_{(0,t]} Z_s\, d[Z^{-1}, \widetilde{N}]_s. \tag{3.15}$$

The linear map $\Psi_Z : \mathcal{M}_{loc}(\mathbb{P}) \to \mathcal{M}_{loc}(\widetilde{\mathbb{P}})$ is bijective and the inverse map $(\Psi_Z)^{-1} : \mathcal{M}_{loc}(\widetilde{\mathbb{P}}) \to \mathcal{M}_{loc}(\mathbb{P})$ satisfies $(\Psi_Z)^{-1} = \Psi_{Z^{-1}}$.

Let us finally recall the following standard lemma.

LEMMA 3.8. *Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a filtered probability space and let for any $j = 1, \dots, l$ the process Z^j be a strictly positive $(\mathbb{P},\mathbb{F})$-martingale with $\mathbb{E}_{\mathbb{P}}(Z_0) = 1$. Then for any fixed $T > 0$ and for each $j = 1, \dots, l$, there exists a probability measure $\mathbb{P}^j$ equivalent to $\mathbb{P}$ on $(\Omega, \mathcal{F}_T)$ with the Radon–Nikodým density process given by*

$$\frac{d\mathbb{P}^j}{d\mathbb{P}}\bigg|_{\mathcal{F}_t} = Z^j_t, \quad \forall\, t \in [0, T]. \tag{3.16}$$

3.2.2. Construction of a Generic Market Model. We are in a position to examine a general framework under which it is possible to derive the joint dynamics of a family of forward swap rates. In the financial interpretation of the next condition, the processes $\kappa^1, \dots, \kappa^l$ are aimed to represent forward swap rates for a family of forward swaps $\mathcal{S}$, whereas the processes $Z^1, \dots, Z^l$ will play the role of Radon–Nikodým densities of swap martingale measures with respect to the underlying probability measure $\mathbb{P}$ (in a practical implementation, $\mathbb{P}$ is typically chosen to be one of the swap martingale measures).

ASSUMPTION 3.9. *We are given a filtered probability space* $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$. *We postulate that the* $\mathbb{F}$*-adapted processes* $\kappa^1, \dots, \kappa^l$ *and* $Z^1, \dots, Z^l$ *satisfy the following conditions, for every* $j = 1, \dots, l$,

(i) the process Z^j *is a strictly positive* $(\mathbb{P}, \mathbb{F})$*-martingale with* $\mathbb{E}_{\mathbb{P}}(Z^j_0) = 1$,
(ii) the process $\kappa^j Z^j$ *is a* $(\mathbb{P}, \mathbb{F})$*-martingale,*
(iii) the process Z^j *is given as a function of some subset of the family* $\kappa^1, \dots, \kappa^l$*; specifically, there exists a subset* $\{\kappa^{n_1}, \dots, \kappa^{n_{l_j}}\}$ *of the full collection* $\{\kappa^1, \dots, \kappa^l\}$ *of forward swap rates and a function* $f_j : \mathbb{R}^{l_j} \to \mathbb{R}$ *of class* C^2 *such that* $Z^j = f_j(\kappa^{n_1}, \dots, \kappa^{n_{l_j}})$.

Parts (i) and (ii) in Assumption 3.9, when combined with Lemma 3.8, yield the existence of a family of probability measures $\mathbb{P}^1, \dots, \mathbb{P}^l$, equivalent to $\mathbb{P}$ on $(\Omega, \mathcal{F}_T)$ for some fixed $T > 0$, such that the process κ^j is a $(\mathbb{P}^j, \mathbb{F})$-martingale for every $j = 1, \dots, l$, and this in turn implies that κ^j is a $(\mathbb{P}, \mathbb{F})$-semimartingale. Let us remark that the probability measure $\mathbb{P}^j$ is called the jth *swap martingale measure.*

We deduce from part (iii) that the continuous martingale part of Z^j, denoted by $Z^{j,c}$, admits the following integral representation:

$$Z^{j,c}_t = Z^j_0 + \sum_{i=1}^{l_j} \int_0^t \frac{\partial f_j}{\partial x_i}\big(\kappa^{n_1}_s, \dots, \kappa^{n_{l_j}}_s\big)\, d\kappa^{n_i,c}_s, \tag{3.17}$$

where $\kappa^{i,c}$ stands for the continuous martingale part of κ^i. To establish equality (3.17), it suffices to apply the Itô formula and use the properties of the stochastic integral.

We will argue that, under the standing Assumption 3.9, the semimartingale decomposition of κ^j can be uniquely specified under $\mathbb{P}$ by the choice of the initial values, the volatility processes and the driving martingale, which is henceforth denoted by M. For the purpose of an explicit construction of the model for processes $\kappa^1, \dots, \kappa^l$, we thus start by selecting an $\mathbb{R}^k$-valued $(\mathbb{P}, \mathbb{F})$-martingale $M = (M^1, \dots, M^k)$ and we define the process κ^j under $\mathbb{P}^j$ as follows, for every $j = 1, \dots, l$,

$$\kappa^j_t = \kappa^j_0 + \int_{(0,t]} \kappa^j_{s-} \sigma^j_s \cdot d\Psi^j(M)_s, \tag{3.18}$$

where σ^j is the $\mathbb{R}^k$-valued volatility process and the $(\mathbb{P}^j, \mathbb{F})$-martingale $\Psi^j(M)$ equals (see (3.12))

$$\Psi^j(M)_t = M_t - \int_{(0,t]} \frac{1}{Z^j_s}\, d[Z^{j,c}, M^c]_s - \sum_{0<s\le t} \frac{1}{Z^j_s} \Delta Z^j_s \Delta M_s. \tag{3.19}$$

PROPOSITION 3.10. *Under Assumption 3.9, if the processes $\kappa^1, \dots, \kappa^l$ satisfy (3.18) and (3.19) then for every $j = 1, \dots, l$ the dynamics of κ^j are*

$$d\kappa_t^j = \sum_{v=1}^{k} \kappa_{t-}^j \sigma_t^{j,v}\, dM_t^v - \frac{1}{f_j\big(\kappa_t^{n_1}, \dots, \kappa_t^{n_{l_j}}\big)} \sum_{i=1}^{l_j} \frac{\partial f_j}{\partial x_i}\big(\kappa_t^{n_1}, \dots, \kappa_t^{n_{l_j}}\big) \times \sum_{v,m=1}^{k} \kappa_t^j \kappa_t^{n_i} \sigma_t^{j,v} \sigma_t^{n_i,m}\, d[M^{v,c}, M^{m,c}]_t - \frac{\kappa_{t-}^j}{Z_t^j} \Delta Z_t^j \sum_{v=1}^{k} \sigma_t^{j,v} \Delta M_t^v. \tag{3.20}$$

Proof. In view of formulae (3.18) and (3.19), it is clear that the continuous martingale part of κ^j can be expressed as follows:

$$\kappa_t^{j,c} = \kappa_0^j + \int_{(0,t]} \kappa_s^j \sigma_s^j \cdot dM_s^c,$$

and thus we infer from (3.17) that the continuous martingale part $Z^{j,c}$ of Z^j admits the following representation:

$$Z_t^{j,c} = Z_0^j + \sum_{i=1}^{l_j} \int_{(0,t]} \frac{\partial f_j}{\partial x_i}\big(\kappa_s^{n_1}, \dots, \kappa_s^{n_{l_j}}\big) \kappa_s^{n_i} \sigma_s^{n_i} \cdot dM_s^c. \tag{3.21}$$

This allows us to express the process $\Psi^j(M)$ in terms of M, the processes $\kappa^1, \dots, \kappa^n$, their volatilities, and the matrix $[M^c]$. To be more specific, we deduce from (3.19) and (3.21) that (note that $[M^{m,c}, M^c]$ is an $\mathbb{R}^k$-valued process)

$$\Psi^j(M)_t = M_t - \int_{(0,t]} \frac{1}{f_j\big(\kappa_s^{n_1}, \dots, \kappa_s^{n_{l_j}}\big)} \sum_{i=1}^{l_j} \frac{\partial f_j}{\partial x_i}\big(\kappa_s^{n_1}, \dots, \kappa_s^{n_{l_j}}\big) \sum_{m=1}^{k} \kappa_s^{n_i} \sigma_s^{n_i,m}\, d[M^{m,c}, M^c]_s - \sum_{0<s\leq t} \frac{1}{Z_s^j} \Delta Z_s^j \Delta M_s, \tag{3.22}$$

where $Z_s^j = f_j(\kappa_s^{n_1}, \dots, \kappa_s^{n_{l_j}})$ and

$$\Delta Z_s^j = f_j\big(\kappa_s^{n_1}, \dots, \kappa_s^{n_{l_j}}\big) - f_j\big(\kappa_{s-}^{n_1}, \dots, \kappa_{s-}^{n_{l_j}}\big).$$

By combining (3.18) with (3.22), we arrive at formula (3.20). We thus conclude that, under Assumption 3.9, by choosing the driving $(\mathbb{P}, \mathbb{F})$-martingale M and the volatility processes $\sigma^1, \dots, \sigma^l$ in formula (3.18), we completely specify the joint dynamics of processes $\kappa^1, \dots, \kappa^l$ under $\mathbb{P}$. □

Note that when using Proposition 3.10 we do not postulate that a model of forward swap rates is already given; in fact, the goal of this result was to describe a general approach to constructing such a model starting from some multi-dimensional driving martingale $M = (M^1, \dots, M^k)$ and arbitrarily chosen $\mathbb{R}^k$-valued volatility processes $\sigma^1, \dots, \sigma^l$ (of course, we should ensure that the stochastic integrals appearing in (3.20) are well defined). Typically, a probability measure $\mathbb{P}$ is selected to be one of the *swap martingale measures* $\mathbb{P}^j$. Therefore, condition (iii) in Assumption 3.9 is a consequence of the $\mathcal{A}$-admissibility of a family $\mathcal{S}$. This observation explains why we focused on the inverse problem (IP.2) and the resulting concept of $\mathcal{A}$-admissibility. Intuitively, an $\mathcal{A}$-admissible family of forward swaps is self-consistent, meaning that the joint dynamics of forward swap rate processes $\kappa^1, \dots, \kappa^l$ are entirely determined by the choice of volatilities and driving martingales.

EXAMPLE 3.11. We consider the set-up introduced in Example 2.38. Our goal is to specify the joint dynamics of the family of forward swap rates $(\kappa^1, \kappa^2, \kappa^3)$ under a single probability measure. To this end, we assume that we are given a family of volatility processes $\sigma^{i,j}_t$ for $i, j = 1, 2, 3$ and a three-dimensional Brownian motion (W^1, W^2, W^3) defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, with the correlation structure given by $d\langle W^i, W^j\rangle_t = \rho^{i,j}_t \, dt$. We first recall that the family $\mathcal{S}$ considered in Example 2.38 was shown to be $\mathcal{T}$-admissible and thus, by Lemma 3.3, it is also $\mathcal{A}$-admissible.

The swap annuities in this example are given by: $A^{0,2}_t = \widetilde{a}_2 B(t, T_2)$, $A^{1,3}_t = \widetilde{a}_3 B(t, T_3)$ and $A^{0,3}_t = \widetilde{a}_3 B(t, T_3)$. Hence the annuity deflated swap numéraires for $d = 3$ are given by $\widetilde{A}^{3,2}_t = \widetilde{A}^{3,3}_t = 1$ and

$$\widetilde{A}^{3,1}_t = \frac{\widetilde{a}_2 B(t, T_2)}{\widetilde{a}_3 B(t, T_3)} = \frac{\widetilde{a}_2 B^0(t, T_2)}{\widetilde{a}_3 B^0(t, T_3)}.$$

To derive the joint dynamics of $(\kappa^1, \kappa^2, \kappa^3)$, we postulate that the probability measure $\mathbb{P}$ is such that the process κ^3 is a $(\mathbb{P}, \mathbb{F})$-martingale, so that $\mathbb{P} = \mathbb{P}^3$, and we define the probability measures $d\mathbb{P}^i = Z^i \, d\mathbb{P}$ for $i = 1, 2$, where the Radon–Nikodým density $Z^2 = 1$ and the Radon–Nikodým density process Z^1 is given by

$$Z^1_t = \frac{1}{c_1} \widetilde{A}^{3,1}_t = \frac{1}{c_1} \frac{A^{0,2}}{A^{1,3}} = \frac{1}{c_1} \frac{\widetilde{a}_2 B^0(t, T_2)}{\widetilde{a}_3 B^0(t, T_3)},$$

where c_1 is a normalizing constant. Using the computations of Example 2.38, we see that

$$Z^1_t = f_1\big(\kappa^1_t, \kappa^2_t, \kappa^3_t\big) = \frac{1}{c_1} \frac{\widetilde{a}_2\big(1 + \widetilde{a}_3 \kappa^3_t\big)}{\widetilde{a}_3\big(1 + \widetilde{a}_2 \kappa^1_t\big)},$$

and thus Z^1 is a strictly positive process provided that the forward swap rates κ^1 and κ^3 are strictly positive (the latter property is not obvious a priori, but it will follow from equations (3.23) and (3.24)). To apply equation (3.20), we first compute the first-order partial derivatives

$$\frac{\partial f_1}{\partial x_1}\big(\kappa^1_t, \kappa^2_t, \kappa^3_t\big) = \frac{1}{c_1} \frac{\widetilde{a}_2^2\big(1 + \widetilde{a}_3 \kappa^3_t\big)}{\widetilde{a}_3\big(1 + \widetilde{a}_2 \kappa^1_t\big)^2}, \qquad \frac{\partial f_1}{\partial x_3}\big(\kappa^1_t, \kappa^2_t, \kappa^3_t\big) = \frac{1}{c_1} \frac{\widetilde{a}_2}{\big(1 + \widetilde{a}_2 \kappa^1_t\big)}.$$

Using equation (3.20), we conclude that the dynamics of the process κ^1 under the forward swap measure $\mathbb{P}^3$ are given by

$$\begin{aligned} d\kappa^1_t = {} & \sum_{i=1}^{3} \kappa^1_t \sigma^{1,i}_t \, dW^i_t + \frac{\widetilde{a}_3\big(1 + \widetilde{a}_2 \kappa^1_t\big)}{\widetilde{a}_2\big(1 + \widetilde{a}_3 \kappa^3_t\big)} \frac{\widetilde{a}_2^2\big(1 + \widetilde{a}_3 \kappa^3_t\big)}{\widetilde{a}_3\big(1 + \widetilde{a}_2 \kappa^1_t\big)^2} \sum_{i,j=1}^{3} \big(\kappa^1_t\big)^2 \sigma^{1,i}_t \sigma^{1,j}_t \, d\langle W^i, W^j\rangle_t \\ & - \frac{\widetilde{a}_3\big(1 + \widetilde{a}_2 \kappa^1_t\big)}{\widetilde{a}_2\big(1 + \widetilde{a}_3 \kappa^3_t\big)} \frac{\widetilde{a}_2}{\big(1 + \widetilde{a}_2 \kappa^1_t\big)} \sum_{i,j=1}^{3} \kappa^1_t \kappa^3_t \sigma^{1,i}_t \sigma^{3,j}_t \, d\langle W^i, W^j\rangle_t, \end{aligned}$$

that is,

$$d\kappa_t^1 = \sum_{i=1}^{3} \kappa_t^1 \sigma_t^{1,i}\, dW_t^i + \frac{\widetilde{a}_2}{1+\widetilde{a}_2\kappa_t^1} \sum_{i,j=1}^{3} \left(\kappa_t^1\right)^2 \sigma_t^{1,i} \sigma_t^{1,j} \rho_t^{i,j}\, dt$$
$$- \frac{\widetilde{a}_3}{1+\widetilde{a}_3\kappa_t^3} \sum_{i,j=1}^{3} \kappa_t^1 \kappa_t^3 \sigma_t^{1,i} \sigma_t^{3,j} \rho_t^{i,j}\, dt,$$

while the processes κ^2 and κ^3 are governed under $\mathbb{P}^3$ by

$$d\kappa_t^2 = \sum_{i=1}^{3} \kappa_t^2 \sigma_t^{2,i}\, dW_t^i, \quad d\kappa_t^3 = \sum_{i=1}^{3} \kappa_t^3 \sigma_t^{3,i}\, dW_t^i. \tag{3.23}$$

The last formula holds, since the Radon–Nikodým density Z^2 is identically one. This peculiar feature comes from the fact that the swaps S_2 and S_3 in Example 3.4 only differ by the start date and both have a unique settlement date T_3. Moreover, one can check that the process Z^1 is indeed a strictly positive $(\mathbb{P}, \mathbb{F})$-martingale and under $\mathbb{P}^1$ we have that

$$d\kappa_t^1 = \sum_{i=1}^{3} \kappa_t^1 \sigma_t^{1,i}\, d\widetilde{W}_t^i, \tag{3.24}$$

where $(\widetilde{W}^1, \widetilde{W}^2, \widetilde{W}^3)$ is a Brownian motion under $\mathbb{P}^1$ with the same correlation structure as the underlying Brownian motion (W^1, W^2, W^3). We note that for $i = 1, 2, 3$ the processes κ^i are strictly positive under $\mathbb{P} = \mathbb{P}^3$, since the process κ^i is a stochastic exponential under $\mathbb{P}^i$ for each $i = 1, 2, 3$, and the probability measures $\mathbb{P}^i$ are equivalent to each other.

Let us conclude this section by pointing out that the weak $\mathcal{A}$-admissibility (see Example 3.4 for such family) is not sufficient for specification of the joint dynamics since, on the one hand, we cannot guarantee that the Radon–Nikodým density processes are strictly positive and, on the other hand, if one were able to derive some model of the joint dynamics of forward swap rates, it is likely to be inconsistent in the sense that, without additional assumptions, the swap annuities could fail to be positive. For more details and explicit examples of applications of Proposition 3.10 in the context of market models of CDS spreads, the interested reader is referred to Li and Rutkowski (2011).

APPENDIX

In the Appendix, we provide the proofs of Propositions 2.24 and 2.37.

A.1. Proof of Proposition 2.24

(i) ⇒ (ii). Let us first assume that C^b is diagonalizable. We need to show that there is no $(\mathcal{T}, b)$-inadmissible subset $\widetilde{\mathcal{S}}$. Let us consider the diagonalized version of C^b and let us take any subset $\widetilde{\mathcal{S}}$ of $\mathcal{S}$. If the subset has k elements then the number of variables associated with $\widetilde{\mathcal{S}}$ is at least equal to k, since the number of distinct nonzero terms corresponding to $\widetilde{\mathcal{S}}$ that come from the diagonal of C^b equals k. One can also argue that if a vertical block exists in C^b then, by inspection, the matrix cannot be diagonalized.

(ii) ⇒ (i). The proof of the implication (ii) ⇒ (i) in Proposition 2.24, which is presented here, hinges on the inductive argument. An alternative proof, based on a more explicit algorithm for a fixed number of forward swaps, is also available from the authors upon request.

Let us now assume that (ii) holds so that there is no $(\mathcal{T}, b)$-inadmissible subset $\widetilde{\mathcal{S}}$ of $\mathcal{S}$. We will proceed by induction. Suppose that the implication (ii) ⇒ (i) is true for any $(l-1) \times (n-1)$ matrix. We will argue by contradiction. Suppose that (ii) holds, but C^b cannot be diagonalized. This means that for every nonzero term $c_{1,1}, \dots, c_{1,d}$ in the first row, the $(l-1) \times (n-1)$ sub-matrix obtained from C^b by deleting the first row and the column corresponding to nonzero term cannot be diagonalized (without loss of generality, we assume here that the entries $c_{1,1}, \dots, c_{1,d}$ in C^b are nonzero and $c_{1,d+1} = \cdots = c_{1,n} = 0$ for some $d \in \{1, \dots, n\}$. Indeed, if for some nonzero $c_{1,j}$ the $(l-1) \times (n$ $-1)$ sub-matrix can be diagonalized then, obviously, the matrix can be diagonalized as well.)

Therefore, by the inductive assumption, for any $i = 1, \dots, d$ there exists a subset $\widetilde{\mathcal{S}}_i$ of "shorter" swaps (that is, forward swaps without the variable x_d) which is not admissible with respect to the reduced set of variables (dates). For each $i = 1, \dots, d$, we select the subset $\widetilde{\mathcal{S}}_i$ with the least number of elements. Let k_i be the number of swaps in $\widetilde{\mathcal{S}}_i$. Then the number of the corresponding variables n_i (dates) is less than k_i, that is $n_i \le k_i - 1$.

We will argue that by taking the union $\cup_{i=1}^{d} \widetilde{\mathcal{S}}_i$ of the corresponding "longer" swaps and by adding the first row (i.e., the first "longer" swap) we will produce a $(\mathcal{T}, b)$-inadmissible subset $\widetilde{\mathcal{S}}$ of $\mathcal{S}$. This will mean that a contradiction arises since we have assumed that (ii) holds.

Let us first examine the special case when the sets $\widetilde{\mathcal{S}}_i$ are pairwise disjoint. Then the above conclusion is rather obvious, since in the union $\cup_{i=1}^{d} \widetilde{\mathcal{S}}_i$ the number of forward swaps equals $\sum_{i=1}^{d} k_i$, whereas the number of variables is given by the following expression:

$$\sum_{i=1}^{d} n_i \le \sum_{i=1}^{d} k_i - d.$$

By supplementing the first row (i.e., the first swap) we obtain a family $\widetilde{\mathcal{S}}$ with $\sum_{i=1}^{d} k_i + 1$ swaps in which the number of variables is less or equal to $\sum_{i=1}^{d} k_i$ (we have to add d variables that come from the first swap).

Let us now consider the general case, where an overlap between $\widetilde{\mathcal{S}}_i$ and $\widetilde{\mathcal{S}}_j$ may occur. We will now argue by induction with respect to $i = 1, \dots, d$. Suppose that for some $v \le d-1$ the number of swaps in $\cup_{i=1}^{v} \widetilde{\mathcal{S}}_i$ is greater or equal to the number of variables corresponding to $\cup_{i=1}^{v} \widetilde{\mathcal{S}}_i$ plus v. We argue that the number of swaps in $\cup_{i=1}^{v+1} \widetilde{\mathcal{S}}_i$ is greater or equal to the number of variables corresponding to $\cup_{i=1}^{v+1} \widetilde{\mathcal{S}}_i$ plus $v+1$. If some swaps from $\widetilde{\mathcal{S}}_{v+1}$ are already in $\cup_{i=1}^{v} \widetilde{\mathcal{S}}_i$, they do not add any new variables (except for, perhaps, some of the variables $x_1, \dots, x_d$).

Let us then suppose that there are g swaps that are in $\widetilde{\mathcal{S}}_{v+1}$, but not in $\cup_{i=1}^{v} \widetilde{\mathcal{S}}_i$. The number of new variables corresponding to these swaps is no more than $g-1$; we use here the property that we have chosen the subset $\widetilde{\mathcal{S}}_{v+1}$ with the least number of elements, so that the number of variables corresponding to $k_{v+1} - g$ swaps that are already in $\cup_{i=1}^{v} \widetilde{\mathcal{S}}_i$ equals at least $k_{v+1} - g$ and the total number of variables corresponding to $\widetilde{\mathcal{S}}_{v+1}$ is no more than $k_{v+1} - 1$. Therefore, the number of swaps has grown by g and the number of variables by $g-1$, as desired (plus, perhaps, some of the variables $x_1, \dots, x_d$).

By induction with respect to $i = 1, \ldots, d$, we conclude that the number of swaps in $\cup_{i=1}^{d} \widetilde{\mathcal{S}}_i$ is greater or equal to the number of variables corresponding to $\cup_{i=1}^{d} \widetilde{\mathcal{S}}_i$ plus d. By supplementing the first row (first swap) we obtain a family $\widetilde{\mathcal{S}}$ in which the total number of swaps is strictly greater than the number of variables (dates). This completes the proof of the lemma. □

A.2. Proof of Proposition 2.37

We work here under the assumption that $l = n$ and we assume that the family $\mathcal{S}$ is $\mathcal{T}$-admissible and satisfies Property (A). This implies, in particular, that $\mathcal{S}$ is weakly $\mathcal{T}$-admissible and the solution to the linear system $C^b \overline{x}^b = \overline{e}^b$ is strictly positive for almost all generic values of $(\kappa_1, \ldots, \kappa_n) \in \mathbb{R}^n_+$. As usual, we denote by $\mathcal{A}_i$ the indices of settlement dates of the forward swap S_i and let $\widetilde{\mathcal{A}}_i := \mathcal{A}_i \setminus \{m_i\}$. Our goal is to show that no two forward swaps start on the same date. Suppose, on the contrary, that there exist two forward swaps, S_k and S_i say, starting on the same date, so that $T_{s_k} = T_{s_i} = T_s$ for some s.

By rearranging the forward swap equations generated by S_k and S_i, we obtain

$$B^n(t, T_s) = B^n(t, T_{m_i}) + \kappa_t^i \sum_{j \in \mathcal{A}_i} a_j B^n(t, T_j) \tag{A.1}$$

and

$$B^n(t, T_{m_k}) = \frac{B^n(t, T_s) - \kappa_t^k \sum_{j \in \widetilde{\mathcal{A}}_k} a_j B^n(t, T_j)}{\left(1 + a_{m_k} \kappa_t^k\right)}. \tag{A.2}$$

After substituting (A.2) into (A.1), we get

$$x_s = x_{m_i}(1 + a_{m_i}\kappa_i) + \kappa_i \sum_{j \in \widetilde{\mathcal{A}}_i \setminus \{m_k\}} a_j x_j + \frac{a_{m_k}\kappa_i x_s - a_{m_k}\kappa_k\kappa_i \sum_{j \in \widetilde{\mathcal{A}}_k} a_j x_j}{(1 + a_{m_k}\kappa_k)},$$

where we use the shorthand notation $\kappa_j = \kappa_t^j$ and $x_j = B^n(t, T_j)$. By rearranging the above equality, we obtain

$$(1 + a_{m_k}\kappa_k - a_{m_k}\kappa_i)x_s = (1 + a_{m_k}\kappa_k)(1 + a_{m_i}\kappa_i)x_{m_i} + (1 + a_{m_k}\kappa_k)\kappa_i \sum_{j \in \widetilde{\mathcal{A}}_i \setminus \{m_k\}} a_j x_j$$
$$- a_{m_k}\kappa_i\kappa_k \sum_{j \in \widetilde{\mathcal{A}}_k} a_j x_j.$$

Consequently,

$$
\begin{aligned}
(1 + a_{m_k}\kappa_k - a_{m_k}\kappa_i)x_s &= (1 + a_{m_i}\kappa_i)(1 + a_{m_k}\kappa_k)x_{m_i} + (1 + a_{m_k}\kappa_k)\kappa_i \sum_{j\in\widetilde{\mathcal{A}}_i\setminus\mathcal{A}_k} a_j x_j \\
&\quad + \kappa_i((1 + a_{m_k}\kappa_k) - a_{m_k}\kappa_k) \sum_{j\in\widetilde{\mathcal{A}}_i\cap\widetilde{\mathcal{A}}_k} a_j x_j - a_{m_k}\kappa_i\kappa_k \sum_{j\in\widetilde{\mathcal{A}}_k\setminus\widetilde{\mathcal{A}}_i} a_j x_j \\
&= (1 + a_{m_i}\kappa_i)(1 + a_{m_k}\kappa_k)x_{m_i} + (1 + a_{m_k}\kappa_k)\kappa_i \sum_{j\in\widetilde{\mathcal{A}}_i\setminus\mathcal{A}_k} a_j x_j \\
&\quad + \kappa_i \sum_{j\in\widetilde{\mathcal{A}}_i\cap\widetilde{\mathcal{A}}_k} a_j x_j - a_{m_k}\kappa_i\kappa_k \sum_{j\in\widetilde{\mathcal{A}}_k\setminus\widetilde{\mathcal{A}}_i} a_j x_j .
\end{aligned}
$$

Under the assumption that $\mathcal{T}(S^l) \subset \mathcal{T}(S^i)$, we have that $\widetilde{\mathcal{A}}_i \cap \widetilde{\mathcal{A}}_k = \widetilde{\mathcal{A}}_k$ and

$$
a_{m_k}\kappa_i\kappa_k \sum_{j\in\widetilde{\mathcal{A}}_k\setminus\widetilde{\mathcal{A}}_i} a_j x_j = 0.
$$

This in turn yields the following equality

$$
\begin{aligned}
(1 + a_{m_k}\kappa_k - a_{m_k}\kappa_i)x_s &= (1 + a_{m_i}\kappa_i)(1 + a_{m_k}\kappa_k)x_{m_i} + (1 + a_{m_k}\kappa_k)\kappa_i \sum_{j\in\widetilde{\mathcal{A}}_i\setminus\mathcal{A}_k} a_j x_j \\
&\quad + \kappa_i \sum_{j\in\widetilde{\mathcal{A}}_k} a_j x_j .
\end{aligned}
$$

We thus conclude that x_s is given by

$$
x_s = \frac{(1 + a_{m_i}\kappa_i)(1 + a_{m_k}\kappa_k)x_{m_i} + (1 + a_{m_k}\kappa_k)\kappa_i \sum_{j\in\widetilde{\mathcal{A}}_i\setminus\mathcal{A}_k} a_j x_j + \kappa_i \sum_{j\in\widetilde{\mathcal{A}}_i\cap\widetilde{\mathcal{A}}_k} a_j x_j}{1 + a_{m_k}\kappa_k - a_{m_k}\kappa_i}.
$$

Assume that $(\kappa_1, \dots, \kappa_n) \in \mathbb{R}^n_+$ and $x_i \in \mathbb{R}_+$ for every $i \neq s$. Then the deflated bond price $x_s = B^n(t, T_s)$ is strictly positive if and only if $1 + a_{m_k}\kappa_k - a_{m_k}\kappa_i > 0$. This contradicts the assumption that $\mathcal{S}$ is $\mathcal{T}(\mathcal{S})$-admissible, since it is not true that x_s is strictly positive for almost all generic values of $(\kappa_1, \dots, \kappa_n) \in \mathbb{R}^n_+$. □

REFERENCES

BIANCHETTI, M. (2009): Two Curves, One Price: Pricing and Hedging Interest Rate Derivatives, Decoupling Forwarding and Discounting Curves. Working Paper.

BIANCHETTI, M. (2010): Two Curves, One Price, *Risk Mag.* August, 74–80.

BIANCHETTI, M., and M. CARLICCHI (2011): Interest Rates after the Credit Crunch: Multiple-Curve Vanilla Derivatives and SABR. Working Paper.

BOLLOBÁS, B. (1979): *Graph Theory, An Introductory Course*, Berlin: Springer.

BRACE, A., D. GĄTAREK, and M. MUSIELA (1997): The Market Model of Interest Rate Dynamics, *Math. Finance* 7, 127–154.

BRÉMAUD, P., and M. YOR (1978): Changes of Filtrations and of Probability Measures, *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 45, 269–295.

BRIGO, D. (2008): CDS Options through Candidate Market Models and the CDS-Calibrated CIR++ Stochastic Intensity Model, in *Credit Risk: Models, Derivatives and Management*, Vol. 6, N. Wagner, ed., Boca Raton, London, New York: Chapman & Hall/CRC Financial Mathematics Series, pp. 393–426.

DAVIS, M., and V. MATAIX-PASTOR (2007): Negative Libor Rates in the Swap Market Model, *Finance Stoch.* 11, 181–193.

FUJII, M., Y. SHIMADA, and A. TAKAHASHI (2009a): A Note on Construction of Multiple Swap Curves with and without Collateral. Working Paper.

FUJII, M., Y. SHIMADA, and A. TAKAHASHI (2009b): A Market Model of Interest Rates with Dynamics Basis Spreads in the Presence of Collateral and Multiple Currencies. Working Paper.

GALLUCCIO, S., J.-M. LY, Z. HUANG, and O. SCAILLET (2007): Theory and Calibration of Swap Market Models, *Math. Finance* 17, 111–141.

JACOD, J., and M. YOR (1977): Étude des Solutions Extrémales et Représentation Intégrale des Solutions pour Certains Problèmes de Martingales, *Z. für Wahrscheinlichkeitstheorie verw. Gebiete* 38, 83–125.

JAMSHIDIAN, F. (1997): LIBOR and Swap Market Models and Measures, *Finance Stoch.* 1, 293–330.

JAMSHIDIAN, F. (1999): LIBOR Market Model with Semimartingales. Working Paper, NetAnalytic Limited.

JAMSHIDIAN, F. (2008): Bivariate Support of Forward Libor and Swap Rates, *Math. Finance* 18, 427–443.

JEANBLANC, M., M. YOR, and M. CHESNEY (2009): *Mathematical Methods for Financial Markets*, London: Springer.

KIJIMA, M., K. TANAKA, and T. WONG (2009): A Multi-Quality Model of Interest Rates, *Quant. Finance* 9, 133–145.

LI, L., and M. RUTKOWSKI (2011): Market Models of Forward CDS Spreads, in *Progress in Probability, Stochastic Analysis with Financial Applications*, A. Kohatsu-Higa, N. Privault, and S.-J. Sheu, eds., Basel: Springer, pp. 255–280.

MERCURIO, F. (2009): Interest Rates and the Credit Crunch: New Formulas and the Market Model. Working Paper.

MORENI, N., and A. PALLAVICINI (2010): Parsimonious HJM Modelling for Multiple Yield-Curve Dynamics. Working Paper.

MUSIELA, M., and M. RUTKOWSKI (1997): Continuous-Time Term Structure Models: Forward Measure Approach, *Finance Stoch.* 1, 261–291.

MUSIELA, M., and M. RUTKOWSKI (2005): *Martingale Methods in Financial Modelling*, 2nd ed., Berlin: Springer.

PALLAVICINI, A., and M. TARENGHI (2009): Interest-Rate Modeling with Multiple Yield Curves. Working Paper.

PIETERSZ, R., and M. VAN REGENMORTEL (2006): Generic Market Models, *Finance Stoch.* 10, 507–528.

RUTKOWSKI, M. (1999): Models of Forward Libor and Swap Rates, *Appl. Math. Finance* 6, 1–32.

SHIRYAEV, A. N., and A. S. CHERNY (2002): Vector Stochastic Integrals and the Fundamental Theorems of Asset Pricing, *Proc. Steklov Inst. Math.* 237, 6–49.

MATHEMATICAL FINANCE

An International Journal of Mathematics, Statistics and Financial Economics

Mathematical Finance, Vol. 24, No. 4 (October 2014), 762–789

PRICING SWAPTIONS UNDER MULTIFACTOR GAUSSIAN HJM MODELS

JOÃO PEDRO VIDAL NUNES

BRU-UNIDE and ISCTE-IUL Business School

PEDRO MIGUEL SILVA PRAZERES

Sociedade Gestora dos Fundos de Pensões do Banco de Portugal

Several approximations have been proposed in the literature for the pricing of European-style swaptions under multifactor term structure models. However, none of them provides an estimate for the inherent approximation error. Until now, only the Edgeworth expansion technique of Collin-Dufresne and Goldstein is able to characterize the order of the approximation error. Under a multifactor HJM Gaussian framework, this paper proposes a new approximation for European-style swaptions, which is able to set bounds on the magnitude of the approximation error and is based on the *conditioning approach* initiated by Curran and Rogers and Shi. All the proposed pricing bounds will arise as a simple by-product of the Nielsen and Sandmann setup, and will be shown to provide a better accuracy–efficiency trade-off than all the approximations already proposed in the literature.

KEY WORDS: Gaussian HJM multifactor models, European-style swaptions, conditioning approach, rank 1 approximation, lognormal approximation, stochastic duration, Edgeworth expansion, hyperplane approximation, low-variance martingale approximation.

1. INTRODUCTION

The main purpose of the present paper is to offer a fast and extremely accurate analytical approximation for European-style swaptions under a multifactor Gaussian Heath, Jarrow, and Morton (1992)—HJM, hereafter—framework.

European-style swaptions are essentially options on coupon-bearing bonds, that is, on a portfolio of pure discount bonds. Under several single-factor term structure models,[1] a European-style swaption can be valued analytically through its decomposition into a portfolio of options on zero-coupon bonds—see, for instance, Jamshidian (1989) under the Vasiček (1977) model, or Longstaff (1993) for the Cox, Ingersoll, and Ross (1985) setup. However, under a (more realistic) multifactor term structure framework, no exact

Financial support by FCT's grant number PTDC/EGE-ECO/099255/2008 is gratefully acknowledged. The authors also thank the co-editor and two anonymous referees, whose corrections have significantly improved this paper.

The analyses, opinions, and findings of this paper represent the views of the authors, and they are not necessarily those of the Sociedade Gestora dos Fundos de Pensões do Banco de Portugal, the Banco de Portugal, or the Eurosystem.

Manuscript received August 2011; final revision received August 2012.

Address correspondence to João Pedro Nunes, Department of Finance, ISCTE-IUL Business School, Complexo INDEG/ISCTE, Av. Prof Anibal Bettencourt, Lisboa 1600-189, Portugal, e-mail: joao.nunes@iscte.pt.

[1] As long as discount factors are monotonic functions of the single state variable, and if closed-form solutions exist for options on zero-coupon bonds.

DOI: 10.1111/mafi.12019

closed-form solution has ever been found for European-style swaptions, because the optimal exercise boundary involves a nonlinear function of several random variables, whose joint probability density is unknown.

Since European-style swaptions are among the most widely traded fixed-income derivatives, it is not surprising that several pricing approximations have been proposed in the literature. El Karoui and Rochet (1989) obtained an analytical approximation for European-style options on coupon-bearing bonds, under a multifactor Gaussian HJM model, by using a *proportionality assumption*. Such assumption enables the exercise boundary to be expressed as a monotonic function of a univariate normal random variable and is equivalent to the *rank 1* approximation suggested by Brace and Musiela (1994). Still under the same framework, Pang (1996) approximates the probability distribution of the underlying coupon-bearing bond by a lognormal distribution, which has the same first two moments. Although it is well known that the sum of lognormal random variables is not lognormally distributed, the price of a coupon bond weighs mostly its last pure discount bond price component (i.e., the one associated with the redemption of the bond's face value and with the payment of the last coupon). Therefore, the intuition behind the lognormal approximation proposed by Pang (1996) is that the probability distribution of the coupon-bearing bond price should essentially depend upon the probabilistic behavior of its last component, which is lognormally distributed for the Gaussian framework considered.

A completely different approach was undertaken by Wei (1997), for single-factor models, and developed by Munk (1999), for any multifactor term structure model. These authors approximate the price of a European-style option on a coupon-bearing bond by a multiple of the price of a European-style option on a zero-coupon bond with maturity equal to the *stochastic duration*[2] of the coupon-bearing bond (and with an adjusted strike price). A similar approach is also pursued by Schrager and Pelsser (2006), since these authors approximate the affine dynamics of the swap rate (under the relevant swap measure) by replacing some (low variance) martingales by their expectations.

Under the Duffie and Kan (1996) general affine class of interest rate models, Singleton and Umantsev (2002) approximate directly the optimal exercise boundary through a linear function of the model's factors, which enables all the relevant exercise probabilities to be computed through the Fourier transform method of Duffie, Pan, and Singleton (2000, Proposition 2).[3] The basic idea is to use a hyperplane to approximate only the segment of the concave exercise boundary where the density of the state variables is mostly concentrated. For an $\mathbb{A}_2(2)$ specification, these authors outperform the stochastic duration approach, both in terms of accuracy and speed.[4]

All the previous approximation schemes are uncontrolled in the sense that the order or magnitude of the approximation error is unknown. The only exception corresponds to the Edgeworth expansion technique, applied by Collin-Dufresne and Goldstein (2002) to the family of affine term structure models, and extended by Collin-Dufresne and

[2] The stochastic duration of a coupon-bearing bond can be defined as the maturity of a pure discount bond with the same instantaneous variance of relative price changes.

[3] Moreover, the numerical (and time-consuming) solution of the complex-valued ordinary differential equations stated in Duffie, Pan, and Singleton (2000, equations 2.5 and 2.6) can also be avoided through the Gaussian approximation and the Gauss–Hermite quadrature approach proposed by Joslin (2010, appendix B). However, under the Gaussian framework adopted in this paper, such approximation is unnecessary because the transition density function of the model's factors is known in closed form.

[4] Note that any n-factor affine term structure model can be cast (through an appropriate invariant transformation) into the $\mathbb{A}_m(n)$ canonical formulation of Dai and Singleton (2000, definition 1), where m ($\leq n$) is the number of state variables driving the factor's variances.

Goldstein (2003) to the HJM and to the random field *affine* frameworks. As long as the moments of the underlying coupon-bearing bond can be obtained analytically (under all the necessary forward measures), the corresponding probability density functions can be approximated through a (truncated) cumulant expansion, whose highest order term characterizes the order of the approximation error. Through Monte Carlo experiments, these authors have reported an accuracy level that is much higher than the one associated with earlier applications of Edgeworth series expansions to the pricing of Asian options—see, for example, Turnbull and Wakeman (1991). Since the Edgeworth expansion is a series expansion about the normal distribution, the authors argue that it seems natural that its accuracy increases for underlying assets characterized by lower volatility regimes, as it is the case for interest rates (when compared against the equity market).

The novel approximation for European-style swaptions proposed in this paper is based on the *conditioning approach* initiated by Curran (1994) and Rogers and Shi (1995) in the context of Asian option pricing, and extended by Nielsen and Sandmann (2002) to a stochastic interest rate setting. This new pricing approach is restricted to a multifactor HJM Gaussian setup, but should be faster to implement than the Edgeworth expansion technique, and will provide explicit (and tight) bounds for the approximation error.

The analytical tractability provided by the multifactor Gaussian—but not necessarily Markovian or time homogeneous—HJM term structure model proposed is obtained at the expense of an important theoretical drawback: interest rates are assumed to be normally distributed, and can therefore attain negative values with positive probability. But even though the no-arbitrage Gaussian setup adopted in this paper is more restrictive than, for instance, the more general affine framework used by Collin-Dufresne and Goldstein (2003), the extremely accurate pricing solutions to be proposed in this paper can always be used as control variates for more general diffusion pricing models. Moreover and following, for instance Nunes, Clewlow, and Hodges (1999, theorem 1) or Kristensen and Mele (2011, definition 1), the Gaussian pricing formulae offered in this paper can also be used as the (most accurate) zero-order term of the perturbed or Taylor series expansion pricing solution associated to a more general affine term structure model. However, the extension to stochastic volatility term structure models is outside the scope of the present paper, whose contribution is, nevertheless, the derivation of an extremely accurate and controlled approximation for European-style swaptions under a multifactor Gaussian HJM framework.[5]

The Gaussian framework adopted also offers a common ground for the comparison of all alternative pricing methods. Since the conditioning approach proposed will provide extremely tight bounds for the approximation error, it will be possible to compare rigorously the accuracy and efficiency of all the approximations already proposed in the literature for European-style swaptions. The alternative approximations have been compared in the literature against benchmark prices obtained through Monte Carlo studies that involve different levels of accuracy. For instance, Collin-Dufresne and Goldstein (2002, page 16) run 2×10^6 simulations, whereas Schrager and Pelsser (2006, page 689) simulate only 500,000 paths, all using standard variance reduction techniques. Based on a much more demanding setting—involving 10^9 simulations, coupled with antithetic variates—to produce the Monte Carlo proxy of the exact swaption price, this paper will

[5] Following Nunes et al. (1999, equation 14) or Kristensen and Mele (2011, equation 17), the higher order (although less significant) terms of the perturbed or Taylor series stochastic volatility expansion must be obtained through the nontrivial differentiation of the proposed Gaussian pricing solution with respect to the vector of state variables. Such extension is left for future research.

show that the conditioning approach significantly improves upon the existing literature in both speed and accuracy.

Next sections are organized as follows. Section 2 summarizes the multifactor Gaussian HJM model adopted. Section 3 uses the conditioning approach to derive explicit lower and upper bounds for the price of European-style swaptions. Section 4 runs several Monte Carlo experiments to compare the accuracy and efficiency of these explicit pricing solutions against all the approximations already proposed in the literature. The main conclusion, stated in Section 5, is that the conditioning approach offers the best accuracy–efficiency trade-off for the pricing of European-style swaptions.

2. MULTIFACTOR GAUSSIAN HJM MODEL

We consider a stochastic intertemporal economy defined on a finite trading interval $\mathcal{T} = [t_0, \tau]$, for some fixed time $\tau > t_0$. Uncertainty is represented by a filtered probability space $(\Omega, \mathcal{F}, \mathbb{Q})$, where all the information accruing to all the agents in the economy is described by the augmented, right continuous, and complete filtration $\mathbb{F} = \{\mathcal{F}_t : t \in \mathcal{T}\}$ generated through the standard Brownian motion $W^{\mathbb{Q}}(t) \in \mathbb{R}^n$, initialized at zero and defined under $\mathbb{Q}$. The probability measure $\mathbb{Q}$ represents the martingale measure obtained when the "money market account" is taken as the numéraire of the economy underlying the model under analysis.[6]

The Gaussian HJM model under use can be formulated in terms of pure discount bond prices, which are assumed to evolve through time (under measure $\mathbb{Q}$) according to the following stochastic differential equation:

$$\frac{dP(t,T)}{P(t,T)} = r(t)\,dt + \sigma(t,T)' \cdot dW^{\mathbb{Q}}(t), \tag{2.1}$$

where $P(t, T)$ represents the time-t price of a (unit face value) zero-coupon bond expiring at time T, for all $T \in \mathcal{T}$ and $t \in [t_0, T]$, $r(t)$ is the time-t instantaneous spot rate, and "$\cdot$" denotes the inner product in $\mathbb{R}^n$. The n-dimensional adapted volatility function $\sigma(\cdot, T) : [t_0, T] \to \mathbb{R}^n$ is assumed to be deterministic and to satisfy the usual mild measurability and integrability requirements—as stated, for instance, in Lamberton and Lapeyre (1996, theorem 3.5.5)—as well as the "pull-to-par" boundary condition $\sigma(u, u) = 0 \in \mathbb{R}^n$, $\forall u \in [t_0, T]$.

Using, for instance, Nunes (2004, proposition 2.2), it is well known that equation (2.1) yields the following solution, for any arbitrary *forward measure* $\mathbb{Q}^c$:

$$\begin{aligned} P(T_a, T_b) = \frac{P(t_0, T_b)}{P(t_0, T_a)} \exp\Bigg\{ & -\frac{1}{2} g(t_0, T_a, T_b) + l(t_0, T_a, T_b, T_c) \\ & + \int_{t_0}^{T_a} [\sigma(s, T_b) - \sigma(s, T_a)]' \cdot dW^{\mathbb{Q}_c}(s) \Bigg\}, \end{aligned} \tag{2.2}$$

for $t_0 \le T_a \le T_b$ and $T_c \ge t_0$, with[7]

$$g(t_0, T_a, T_b) := \int_{t_0}^{T_a} \|\sigma(s, T_b) - \sigma(s, T_a)\|^2 ds, \tag{2.3}$$

[6] Meaning that the relative prices of all assets with respect to the numéraire given by a "money market account" are $\mathbb{Q}$-martingales.

[7] $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^n$.

$$l(t_0, T_a, T_b, T_c) := \int_{t_0}^{T_a} [\sigma(s, T_b) - \sigma(s, T_a)]' \cdot [\sigma(s, T_c) - \sigma(s, T_a)]\, ds, \tag{2.4}$$

and where

$$dW^{\mathbb{Q}_c}(t) = dW^{\mathbb{Q}}(t) - \sigma(t, T_c)\, dt$$

is also a vector of standard Brownian motion increments in $\mathbb{R}^n$ —with the same standard filtration as $dW^{\mathbb{Q}}(t)$ —but defined under the $\mathbb{Q}^c$ forward measure, which arises if the numéraire is changed to a zero-coupon bond with maturity at time T_c, and is defined through the Radon–Nikodým derivative

$$\frac{d\mathbb{Q}_c}{d\mathbb{Q}}\bigg| \mathcal{F}_t = \exp\left[\int_{t_0}^{t} \sigma(s, T_c)' \cdot dW^{\mathbb{Q}}(s) - \frac{1}{2}\int_{t_0}^{t} \|\sigma(s, T_c)\|^2 ds\right].$$

3. CONDITIONING APPROACH

This section adapts the *conditioning approach* of Nielsen and Sandmann (2002) to the valuation of European-style receiver swaptions, that is European-style put options on a swap rate. More precisely,[8]

Definition 3.1. The terminal payoff of a European-style receiver swaption with strike rate C, maturity at date T_0 ($\geq t_0$), and on an interest rate swap (IRS) with a unit nominal value and reset dates $T_0 < \cdots < T_{N_0 - 1}$ is equal to

$$[C - y_{0,N_0}(T_0)]^+ \sum_{i=1}^{N_0} \tau_i P(T_0, T_i), \tag{3.1}$$

where τ_i is the year fraction between T_{i-1} and T_i (under some market daycount convention), and

$$y_{0,N_0}(T_0) = \frac{1 - P(T_0, T_{N_0})}{\sum_{i=1}^{N_0} \tau_i P(T_0, T_i)} \tag{3.2}$$

is the time-T_0 spot swap rate.

Definition 3.1 assumes, as is common market practice, that the swaption maturity coincides with the first reset date of the underlying IRS (i.e., time T_0). Therefore, combining expressions (3.1) and (3.2), and as shown, for instance, by Singleton and Umantsev (2002, equation 5.2), the terminal payoff of the $T_0 \times T_{N_0}$ receiver swaption can be rewritten as

$$\left[C \sum_{i=1}^{N_0} \tau_i P(T_0, T_i) + P(T_0, T_{N_0}) - 1\right]^+, \tag{3.3}$$

[8] European-style payer swaptions can be priced through the *payer–receiver swaption parity* stated, for instance, in Longstaff, Santa-Clara, and Schwartz (2001, page 2073): a long position in a European-style receiver swaption and a short position in a European-style payer swaption (with the same strike rate and tenor structure) is equivalent to a receiver forward swap with start date at the maturity date of the swaptions and fixed interest rate equal to the strike rate.

which corresponds to the terminal value of a European-style call with strike equal to the unit nominal value, with maturity at time T_0, and on a coupon-bearing bond promising N_0 cash flows of value $C\tau_i + \mathbb{1}_{\{i=N_0\}}$ at times T_i (with $i = 1, \ldots, N_0$), where $\mathbb{1}_{\{A\}}$ is the indicator function of set A.

For simplicity, the pricing bounds derived in the next lines will be specified for European-style calls on a coupon-bearing bond. Given the equivalence between expressions (3.1) and (3.3), the valuation formulas obtained are also applied, in Section 4, to the pricing of European-style receiver swaptions.

3.1. Lower Bound

Denote by $c_{t_0}[B(t_0); X; T_0]$ the time-t_0 fair price of a European-style call with strike X, expiry date at time T_0 ($\geq t_0$), and on a coupon-bearing bond with present value $B(t_0)$. Following Geman, El Karoui, and Rochet (1995), and since $\mathbb{Q}_0$ is assumed to be a martingale measure with respect to the numéraire $P(t_0, T_0)$, then[9]

$$c_{t_0}[B(t_0); X; T_0] = P(t_0, T_0)\mathbb{E}_{\mathbb{Q}_0}\{[B(t_0) - X]^+ | \mathcal{F}_{t_0}\}. \tag{3.4}$$

As in Rogers and Shi (1995), let $Z \in \mathbb{R}$ be any $\mathcal{F}_{T_0}$ -measurable random variable. From the law of iterative expectations and using Jensen's inequality,

$$\begin{aligned} c_{t_0}[B(t_0); X; T_0] &= P(t_0, T_0)\mathbb{E}_{\mathbb{Q}_0}\{\mathbb{E}_{\mathbb{Q}_0}[(B(t_0) - X)^+ | Z] | \mathcal{F}_{t_0}\} \\ &\geq c^l_{t_0}[B(t_0); X; T_0], \end{aligned} \tag{3.5}$$

where

$$c^l_{t_0}[B(t_0); X; T_0] := P(t_0, T_0)\mathbb{E}_{\mathbb{Q}_0}\{[\mathbb{E}_{\mathbb{Q}_0}(B(t_0)|Z) - X]^+ | \mathcal{F}_{t_0}\} \tag{3.6}$$

defines a lower bound for the true call option price. The next proposition provides an explicit solution for the conditional expectation on the right-hand side of equation (3.6), by assuming a standard normal distribution for the conditioning variable. Later—in Proposition 3.6—the conditioning variable Z will be completely defined in order to minimize the inherent approximation error.

PROPOSITION 3.2. *Under the Gaussian HJM model (2.1), the time-t_0 price of a European-style call with strike X, with maturity at time T_0 ($\geq t_0$), and on a coupon-bearing bond with present value $B(t_0)$ and generating N_0 cash flows of value k_i ($i = 1, \ldots, N_0$) at times $T_1 < \cdots < T_{N_0}$ (with $T_1 \geq T_0$), is bounded from below by*

$$\begin{aligned} &c^l_{t_0}[B(t_0); X; T_0] = P(t_0, T_0) \\ &\quad \times \mathbb{E}_{\mathbb{Q}_0}\left\{\left[\sum_{i=1}^{N_0} k_i \frac{P(t_0, T_i)}{P(t_0, T_0)} \exp\left(-\frac{1}{2}m(t_0, T_0, T_i)^2 + m(t_0, T_0, T_i)Z\right) - X\right]^+ \middle| \mathcal{F}_{t_0}\right\}, \end{aligned} \tag{3.7}$$

[9] $\mathbb{E}_{\mathbb{Q}_c}(Y|\mathcal{F}_{t_0})$ denotes the expected value of the random variable Y, conditional on $\mathcal{F}_{t_0}$, and computed under the equivalent martingale measure $\mathbb{Q}_c$.

where $Z \backsim N^1(0, 1)$ *and*[10]

$$(3.8) \qquad m(t_0, T_0, T_i) := \mathbb{E}_{\mathbb{Q}_0}\left\{ Z \int_{t_0}^{T_0} [\sigma(s, T_i) - \sigma(s, T_0)]' \cdot dW^{\mathbb{Q}_0}(s) \middle| \mathcal{F}_{t_0} \right\}.$$

Proof. Since N_0 cash flows k_i $(i = 1, \ldots, N_0)$ will be generated by the underlying coupon bond between the option's expiry date (T_0) and the bond's maturity date (T_{N_0}), then

$$(3.9) \qquad \mathbb{E}_{\mathbb{Q}_0}[B(t_0)|Z] = \mathbb{E}_{\mathbb{Q}_0}\left[\sum_{i=1}^{N_0} k_i P(T_0, T_i) \middle| Z \right]$$
$$= \sum_{i=1}^{N_0} k_i \mathbb{E}_{\mathbb{Q}_0}[P(T_0, T_i)|Z].$$

The conditional expectation of each discount factor follows from equation (2.2), with $T_a = T_c = T_0$ and $T_b = T_i$:

$$(3.10) \quad \mathbb{E}_{\mathbb{Q}_0}[P(T_0, T_i)|Z] = \frac{P(t_0, T_i)}{P(t_0, T_0)} \exp\left[-\frac{1}{2} g(t_0, T_0, T_i) \right]$$
$$\mathbb{E}_{\mathbb{Q}_0}\left\{ \exp\left[\int_{t_0}^{T_0} (\sigma(s, T_i) - \sigma(s, T_0))' \cdot dW^{\mathbb{Q}_0}(s) \right] \middle| Z \right\}.$$

Assuming that Z possesses a standard univariate normal distribution, since

$$\int_{t_0}^{T_0} [\sigma(s, T_i) - \sigma(s, T_0)]' \cdot dW^{\mathbb{Q}_0}(s) \backsim N^1(0, g(t_0, T_0, T_i)),$$

and following, for instance, Mood, Graybill, and Boes (1974, page 167), then

$$(3.11) \quad \int_{t_0}^{T_0} [\sigma(s, T_i) - \sigma(s, T_0)]' \cdot dW^{\mathbb{Q}_0}(s) \Big| Z \backsim N^1(m(t_0, T_0, T_i)Z, v(T_i, T_i)^2),$$

with

$$(3.12) \qquad v(T_i, T_i)^2 := g(t_0, T_0, T_i) - m(t_0, T_0, T_i)^2,$$

and where the deterministic function $m(t_0, T_0, T_i)$ is defined by the covariance (3.8).

Applying result (3.11) and attending to the definition of the moment generating function of a normal random variable, equation (3.10) becomes

$$(3.13) \quad \mathbb{E}_{\mathbb{Q}_0}[P(T_0, T_i)|Z] = \frac{P(t_0, T_i)}{P(t_0, T_0)} \exp\left[-\frac{1}{2} m(t_0, T_0, T_i)^2 + m(t_0, T_0, T_i)Z \right].$$

Combining equations (3.6), (3.9), and (3.13), the lower bound (3.7) follows immediately for the call price. □

Note that the quasi-analytical pricing solution (3.7) still involves a single integration over the domain of Z. In order to obtain an explicit solution for equation (3.7), and

[10] Hereafter, the notation $Y \backsim N^1(\mu, \sigma^2)$ is intended to mean that the one-dimensional random variable Y is normally distributed, with mean μ and variance σ^2.

following Nielsen and Sandmann (2002), we consider the following family $\{\mathcal{P}, \mathcal{N}, \mathcal{M}\}$ of disjoint sets such that $\mathcal{P} \cup \mathcal{N} \cup \mathcal{M} = \{1, \ldots, N_0\} \equiv \mathcal{D}$:

$$\mathcal{P} := \{i \in \mathcal{D} : m(t_0, T_0, T_i) > 0\}, \tag{3.14}$$

$$\mathcal{N} := \{i \in \mathcal{D} : m(t_0, T_0, T_i) < 0\}, \tag{3.15}$$

and

$$\mathcal{M} := \{i \in \mathcal{D} : m(t_0, T_0, T_i) = 0\}. \tag{3.16}$$

Hence, equation (3.7) can be rewritten as

$$c^l_{t_0}[B(t_0); X; T_0] = P(t_0, T_0)\mathbb{E}_{\mathbb{Q}_0}\left\{\left[\sum_{i\in\mathcal{P}\cup\mathcal{N}} k_i \frac{P(t_0, T_i)}{P(t_0, T_0)} f_i(Z) - \hat{X}\right]^+ \middle| \mathcal{F}_{t_0}\right\}, \tag{3.17}$$

where

$$\hat{X} := X - \sum_{i\in\mathcal{M}} k_i \frac{P(t_0, T_i)}{P(t_0, T_0)}, \tag{3.18}$$

and

$$f_i(Z) := \exp\left(-\frac{1}{2}m(t_0, T_0, T_i)^2 + m(t_0, T_0, T_i)Z\right). \tag{3.19}$$

Since k_i and $\frac{P(t_0,T_i)}{P(t_0,T_0)}$ are positive and because all functions $f_i(Z)$ are convex, for all values of i, then equation

$$\sum_{i\in\mathcal{P}\cup\mathcal{N}} k_i \frac{P(t_0, T_i)}{P(t_0, T_0)} f_i(Z) = \hat{X} \tag{3.20}$$

possesses either zero, one or two solutions in Z. Similarly to Nielsen and Sandmann (2002, Definition 1),

DEFINITION 3.3. Let the two possible solutions of equation (3.20) be represented by z_* and z^*, where $z_* \le z^*$.

- If $\mathcal{P} \ne \emptyset$ but $\mathcal{N} = \emptyset$, then define the unique solution by z^* and set $z_* = -\infty$.
- If $\mathcal{P} = \emptyset$ but $\mathcal{N} \ne \emptyset$, then define the unique solution by z_* and set $z^* = \infty$.
- If $\mathcal{P} \ne \emptyset$ and $\mathcal{N} \ne \emptyset$, then either two solutions z_* and z^* exist or no solution exists and $z_* = z^* = \infty$.

PROPOSITION 3.4. *Under the assumptions of Proposition 3.2,*

(3.21)
$$\begin{aligned} c^l_{t_0}[B(t_0); X; T_0] = &\sum_{i\in\mathcal{P}\cup\mathcal{N}} k_i P(t_0, T_i)\Phi[z_* - m(t_0, T_0, T_i)] - \hat{X}P(t_0, T_0)\Phi(z_*) \\ &+ \sum_{i\in\mathcal{P}\cup\mathcal{N}} k_i P(t_0, T_i)\Phi[m(t_0, T_0, T_i) - z^*] - \hat{X}P(t_0, T_0)\Phi(-z^*), \end{aligned}$$

where $\Phi(\cdot)$ represents the cumulative density function of the univariate standard normal distribution, $\hat{X}$ is defined by equation (3.18), while z_* and z^* are given by Definition 3.3.

Proof. Applying Definition 3.3, equation (3.17) can be restated as

$$
\begin{aligned}
c^l_{t_0}[B(t_0); X; T_0] = {} & P(t_0, T_0)\mathbb{E}_{\mathbb{Q}_0}\left\{\left[\sum_{i\in\mathcal{P}\cup\mathcal{N}} k_i \frac{P(t_0, T_i)}{P(t_0, T_0)} f_i(Z) - \hat{X}\right]\mathbb{1}_{\{Z\leq z_*\}}\middle|\mathcal{F}_{t_0}\right\} \\
& + P(t_0, T_0)\mathbb{E}_{\mathbb{Q}_0}\left\{\left[\sum_{i\in\mathcal{P}\cup\mathcal{N}} k_i \frac{P(t_0, T_i)}{P(t_0, T_0)} f_i(Z) - \hat{X}\right]\mathbb{1}_{\{Z\geq z^*\}}\middle|\mathcal{F}_{t_0}\right\}.
\end{aligned} \tag{3.22}
$$

Since $Z \backsim N^1(0, 1)$ and using equation (3.19), equation (3.22) yields

$$
\begin{aligned}
& c^l_{t_0}[B(t_0); X; T_0] \\
& = \sum_{i\in\mathcal{P}\cup\mathcal{N}} k_i P(t_0, T_i)\int_{-\infty}^{z_*}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}[z - m(t_0, T_0, T_i)]^2\right\}dz - \hat{X}P(t_0, T_0)\Phi(z_*) \\
& \quad + \sum_{i\in\mathcal{P}\cup\mathcal{N}} k_i P(t_0, T_i)\int_{z^*}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}[z - m(t_0, T_0, T_i)]^2\right\}dz - \hat{X}P(t_0, T_0)\Phi(-z^*),
\end{aligned}
$$

and equation (3.21) follows immediately. □

Instead of the method proposed by Nielsen and Sandmann (2002) that relies on the numerical solution of equation (3.20), one could also have solved equation (3.7) through Lord (2006, theorem 1). However, and if the covariance function $m(t_0, T_0, T_i)$ is not monotone in T_i, Lord (2006, theorem 1) would require the numerical and time-consuming minimization of the function $\mathbb{E}_{\mathbb{Q}_0}(B(t_0)|\, Z)$.

The pricing solution (3.21) will only become a completely explicit solution after the specification of the conditioning random variable Z, which will define the deterministic function $m(t_0, T_0, T_i)$. For that purpose, and following Rogers and Shi (1995) and Nielsen and Sandmann (2002), the minimization of the approximation error will be pursued.

3.2. Upper Bound

Following Rogers and Shi (1995, equation 3.5),

$$
\mathbb{E}_{\mathbb{Q}_0}\{\mathbb{E}_{\mathbb{Q}_0}[(B(t_0) - X)^+|Z] - [\mathbb{E}_{\mathbb{Q}_0}(B(t_0)|Z) - X]^+|\mathcal{F}_{t_0}\} \leq \frac{1}{2}\mathbb{E}_{\mathbb{Q}_0}\{\sqrt{\mathrm{var}[B(t_0)|Z]}|\mathcal{F}_{t_0}\}, \tag{3.23}
$$

where

$$
\mathrm{var}[B(t_0)|Z] := \mathbb{E}_{\mathbb{Q}_0}\{[B(t_0) - \mathbb{E}_{\mathbb{Q}_0}(B(t_0)|Z)]^2|Z\} \tag{3.24}
$$

represents the conditional variance of the time-T_0 underlying coupon bond price. Therefore,

Proposition 3.5. *Under the Gaussian HJM model (2.1), the time-t_0 price of a European-style call with strike X, with maturity at time T_0 ($\geq t_0$), and on a coupon-bearing bond with present value $B(t_0)$ and generating N_0 cash flows of value k_i ($i = 1, \ldots, N_0$) at times $T_1 < \cdots < T_{N_0}$ (with $T_1 \geq T_0$), is bounded from above by*

$$
c^u_{t_0}[B(t_0); X; T_0] = c^l_{t_0}[B(t_0); X; T_0] + \varepsilon_{t_0}[B(t_0); X; T_0], \tag{3.25}
$$

where the lower bound $c^l_{t_0}[B(t_0); X; T_0]$ is given by Proposition 3.4 and the implicit approximation error is defined by

$$\varepsilon_{t_0}[B(t_0); X; T_0] := \frac{1}{2}P(t_0, T_0)\sqrt{\mathbb{E}_{\mathbb{Q}_0}\{\mathrm{var}[B(t_0)|Z]|\mathcal{F}_{t_0}\}}, \tag{3.26}$$

with

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}_0}\{\mathrm{var}[B(t_0)|Z]|\mathcal{F}_{t_0}\} = \sum_{i=1}^{N_0}\sum_{j=1}^{N_0} k_i k_j \frac{P(t_0, T_i)P(t_0, T_j)}{P(t_0, T_0)^2}\\ \times \{\exp[l(t_0, T_0, T_i, T_j)] - \exp[m(t_0, T_0, T_i)m(t_0, T_0, T_j)]\}.\end{aligned} \tag{3.27}$$

Proof. Multiplying both sides of inequality (3.23) by the discount factor $P(t_0, T_0)$ and applying Cauchy–Schwarz inequality, as in Nielsen and Sandmann (2002), then

$$c_{t_0}[B(t_0); X; T_0] - c^l_{t_0}[B(t_0); X; T_0] \leq \frac{1}{2}P(t_0, T_0)\sqrt{\mathbb{E}_{\mathbb{Q}_0}\{\mathrm{var}[B(t_0)|Z]|\mathcal{F}_{t_0}\}},$$

and equations (3.25) and (3.26) follow.

Concerning the conditional variance, equations (2.2) and (3.13) yield

$$\begin{aligned}&B(t_0) - \mathbb{E}_{\mathbb{Q}_0}(B(t_0)|Z)\\ &= \sum_{i=1}^{N_0} k_i \frac{P(T_0, T_i)}{P(t_0, T_0)}\left\{\exp\left[-\frac{1}{2}g(t_0, T_0, T_i) + \int_{t_0}^{T_0}(\sigma(s, T_i) - \sigma(s, T_0))' \cdot dW^{\mathbb{Q}_0}(s)\right]\right.\\ &\qquad \left. - \exp\left[-\frac{1}{2}m(t_0, T_0, T_i)^2 + m(t_0, T_0, T_i)Z\right]\right\}.\end{aligned}$$

Applying definition (3.24) and using the conditional probability distribution (3.11), then

$$\begin{aligned}\mathrm{var}[B(t_0)|Z] = \sum_{i=1}^{N_0}\sum_{j=1}^{N_0} k_i k_j \frac{P(t_0, T_i)P(t_0, T_j)}{P(t_0, T_0)^2}\{\exp[v(T_i, T_j)^2] - 1\}\\ \times \exp\left\{[m(t_0, T_0, T_i) + m(t_0, T_0, T_j)]Z - \frac{1}{2}\left[m(t_0, T_0, T_i)^2 + m(t_0, T_0, T_j)^2\right]\right\},\end{aligned} \tag{3.28}$$

where

$$v(T_i, T_j)^2 := l(t_0, T_0, T_i, T_j) - m(t_0, T_0, T_i)m(t_0, T_0, T_j) \tag{3.29}$$

corresponds to the conditional covariance between $\int_{t_0}^{T_0}[\sigma(s, T_i) - \sigma(s, T_0)]' \cdot dW^{\mathbb{Q}_0}(s)$ and $\int_{t_0}^{T_0}[\sigma(s, T_j) - \sigma(s, T_0)]' \cdot dW^{\mathbb{Q}_0}(s)$. Taking expectations of both sides of equation (3.28), and since $Z \backsim N^1(0, 1)$, the analytical solution (3.27) arises. □

Proposition 3.5 shows that with the purpose of minimizing the option' approximation error, Z shall be chosen in order to reduce the quantity (3.27). The next proposition provides a first-order approximation for the (unknown) $\arg\min_Z \mathbb{E}_{\mathbb{Q}_0}\{\mathrm{var}[B(t_0)|Z]|\mathcal{F}_{t_0}\}$.

PROPOSITION 3.6. *Under the Gaussian HJM model (2.1), if*

$$Z := \frac{1}{\alpha}\sum_{i=1}^{N_0} k_i \frac{P(t_0, T_i)}{P(t_0, T_0)} \int_{t_0}^{T_0} [\sigma(s, T_i) - \sigma(s, T_0)]' \cdot dW^{\mathbb{Q}_0}(s), \tag{3.30}$$

with

$$\alpha^2 := \sum_{i=1}^{N_0}\sum_{j=1}^{N_0} k_i k_j \frac{P(t_0, T_i)P(t_0, T_j)}{P(t_0, T_0)^2} l(t_0, T_0, T_i, T_j), \tag{3.31}$$

then

$$Z \backsim N^1(0, 1), \tag{3.32}$$

$$m(t_0, T_0, T_i) = \frac{1}{\alpha}\sum_{j=1}^{N_0} k_j \frac{P(t_0, T_j)}{P(t_0, T_0)} l(t_0, T_0, T_j, T_i), \tag{3.33}$$

and

$$\mathbb{E}_{\mathbb{Q}_0}\{\text{var}[B(t_0)|Z]|\mathcal{F}_{t_0}\} \approx 0. \tag{3.34}$$

Proof. From equation (3.31), since $\alpha^2 = \mathbb{E}_{\mathbb{Q}_0}[(\alpha Z)^2|\mathcal{F}_{t_0}]$, with Z defined through equation (3.30), then condition (3.32) is verified.

Using definition (3.8), then

$$m(t_0, T_0, T_i) = \frac{1}{\alpha}\sum_{j=1}^{N_0} k_j \frac{P(t_0, T_j)}{P(t_0, T_0)} \mathbb{E}_{\mathbb{Q}_0}\left\{\int_{t_0}^{T_0} [\sigma(s, T_j) - \sigma(s, T_0)]' \cdot dW^{\mathbb{Q}_0}(s) \right.$$
$$\left. \times \int_{t_0}^{T_0} [\sigma(s, T_i) - \sigma(s, T_0)]' \cdot dW^{\mathbb{Q}_0}(s) \right| \mathcal{F}_{t_0}\Bigg\},$$

which yields equation (3.33) after considering definition (2.4).

Finally, applying the first-order approximation $\exp(x) \approx 1 + x$ to equation (3.27), then

$$\mathbb{E}_{\mathbb{Q}_0}\{\text{var}[B(t_0)|Z]|\mathcal{F}_{t_0}\} \approx \sum_{i=1}^{N_0}\sum_{j=1}^{N_0} k_i k_j \frac{P(t_0, T_i)P(t_0, T_j)}{P(t_0, T_0)^2} \times \{l(t_0, T_0, T_i, T_j) - m(t_0, T_0, T_i)m(t_0, T_0, T_j)\}. \tag{3.35}$$

Moreover and using the analytical solutions (3.31) and (3.33), it follows that

$$m(t_0, T_0, T_i)m(t_0, T_0, T_j) = \frac{\left[\sum_{p=1}^{N_0} k_p \frac{P(t_0, T_p)}{P(t_0, T_0)} l(t_0, T_0, T_p, T_i)\right]\left[\sum_{q=1}^{N_0} k_q \frac{P(t_0, T_q)}{P(t_0, T_0)} l(t_0, T_0, T_q, T_j)\right]}{\sum_{p=1}^{N_0}\sum_{q=1}^{N_0} k_p k_q \frac{P(t_0, T_p)P(t_0, T_q)}{P(t_0, T_0)^2} l(t_0, T_0, T_p, T_q)}.$$

Therefore,

$$\begin{aligned}
&\sum_{i=1}^{N_0}\sum_{j=1}^{N_0} k_i k_j \frac{P(t_0, T_i)P(t_0, T_j)}{P(t_0, T_0)^2} m(t_0, T_0, T_i) m(t_0, T_0, T_j)\\
&\quad = \frac{\displaystyle\sum_{p=1}^{N_0}\sum_{i=1}^{N_0} k_p k_i \frac{P(t_0, T_p)P(t_0, T_i)}{P(t_0, T_0)^2} l(t_0, T_0, T_p, T_i)}{\displaystyle\sum_{p=1}^{N_0}\sum_{q=1}^{N_0} k_p k_q \frac{P(t_0, T_p)P(t_0, T_q)}{P(t_0, T_0)^2} l(t_0, T_0, T_p, T_q)}\\
&\quad\quad \times \left[\sum_{q=1}^{N_0}\sum_{j=1}^{N_0} k_q k_j \frac{P(t_0, T_q)P(t_0, T_j)}{P(t_0, T_0)^2} l(t_0, T_0, T_q, T_j)\right]\\
&\quad = \sum_{i=1}^{N_0}\sum_{j=1}^{N_0} k_i k_j \frac{P(T_0, T_i)P(T_0, T_j)}{P(t_0, T_0)^2} l(t_0, T_0, T_i, T_j),
\end{aligned}$$

where the second equality follows by replacing i with q in the numerator of the second line and q with i in the third line. Consequently, the right-hand side of equation (3.35) becomes equal to zero, i.e., result (3.34) is obtained. □

4. NUMERICAL ANALYSIS

The purpose of the numerical experiments contained in this section is to compare the accuracy–efficiency performance of the conditioning approach against the rank 1 approximation of Brace and Musiela (1994), the lognormal approximation of Pang (1996), the stochastic duration approach of Munk (1999), the Edgeworth expansion technique of Collin-Dufresne and Goldstein (2002), the hyperplane approximation of Singleton and Umantsev (2002), and the low-variance martingale method of Schrager and Pelsser (2006).

Even though the pricing bounds proposed in Section 3 are applicable to the more general class of HJM Gaussian term structure models, and in order to compare all the approximations proposed in the literature for European-style swaptions, all numerical examples will be run under the nested n-factor affine $\mathbb{A}_0(n)$ specification that is required by both the hyperplane and the low-variance approximations. More precisely, a Gauss–Markov and time-inhomogeneous version of the Duffie and Kan (1996) model will be considered, which specifies the short-term interest rate $r(t)$ as an affine function of the model's factors:

$$r(t) = f + G' \cdot Y(t), \tag{4.1}$$

where $f \in \mathbb{R}$ and $G \in \mathbb{R}^n$ are the model's parameters, while $Y(t) \in \mathbb{R}^n$ denotes the time-t vector of state variables. Additionally, the state variables are assumed to follow a multivariate and time-inhomogeneous elastic random walk:

$$dY(t) = (a \cdot Y(t) + b)\, dt + \Sigma \cdot dW^{\mathbb{Q}}(t), \tag{4.2}$$

where $a, \Sigma \in \mathbb{R}^{n\times n}$, and $b \in \mathbb{R}^n$ are the model's parameters.

Given the affine specifications adopted for the drift and the instantaneous variance of the stochastic differential equation (4.2), it is easy to show—see, for instance, Langetieg (1980, equations 30, 32, and 33)—that pure discount bond prices are exponential-affine functions of the state variables:

$$P(t, T) = \exp[A(t, T) + B(t, T)' \cdot Y(t)], \tag{4.3}$$

where

$$B(t, T)' = G' \cdot a^{-1} \cdot [I_n - e^{a(T-t)}], \tag{4.4}$$

$$\begin{aligned} A(t, T) = {} & (T-t)(G' \cdot a^{-1} \cdot b - f) + B(t, T)' \cdot a^{-1} \cdot [b + \Sigma \cdot \Sigma' \cdot (a^{-1})' \cdot G] \\ & + \frac{1}{2} G' \cdot a^{-1} \cdot [(T-t)\Sigma \cdot \Sigma' + \Delta(t, T)] \cdot (a^{-1})' \cdot G, \end{aligned} \tag{4.5}$$

and

$$\Delta(t, T) := \int_t^T e^{a(T-s)} \cdot \Sigma \cdot \Sigma' \cdot e^{a'(T-s)} ds, \tag{4.6}$$

with $I_n \in \mathbb{R}^{n \times n}$ denoting an identity matrix. Note that equations (4.4) and (4.5) assume that matrix a is nonsingular. If this is not the case, functions $A(t, T)$ and $B(t, T)$ can always be obtained through the more general solutions described in Lund (1994, appendix A). Moreover, function $\Delta(t, T)$ can be computed explicitly from Langetieg (1980, footnote 23), as long as matrix a is diagonalizable; otherwise, it is always possible to evaluate $\Delta(t, T)$ numerically using Padé approximations with scaling and squaring, based on Van Loan (1978, Theorem 1).

Using the diffusion process (4.2), applying Itô's lemma to equation (4.3), and considering the stochastic differential equation (2.1), it is easy to show that the $\mathbb{A}_0(n)$ specification adopted can be cast into the more general Gaussian HJM model presented in Section 2, as long as two conditions are met:

$$\sigma(t, T) = \Sigma' \cdot B(t, T); \tag{4.7}$$

and the discount function initially "observed" in the market must be replaced by equation (4.3). Adopting these two conditions, the pricing solutions proposed in Section 3 will be used under the nested $\mathbb{A}_0(n)$ specification, and function $l(t_0, T_a, T_b, T_c)$ will be computed explicitly—see Appendix A.

Table 4.1 values at-the-money-forward (ATMF) European-style swaptions, using different valuation approaches, and under the three-factor Gaussian and affine model specified in Collin-Dufresne and Goldstein (2002, exhibit 1) or Schrager and Pelsser (2006, table 4.1), i.e., for $f = 6\%$, $G' = [1 \quad 1 \quad 1]$, $Y(t) = [1\% \quad 0.5\% \quad -2\%]$, $a = \text{diag}\{-1, -0.2, -0.5\}$, $b' = [0 \quad 0 \quad 0]$, and $\Sigma = \text{diag}\{1\%, 0.5\%, 0.2\%\} \cdot L$, where L is the lower triangular matrix obtained from the Cholesky decomposition of a (3×3) correlation matrix R, with $R_{12} = -0.2$, $R_{13} = -0.1$, and $R_{23} = 0.3$. Four different option maturities (of 1, 2 , 5, and 10 years) and eight different swap maturities (of 1, 2, 5, 10, 15, 20, 25, and 30 years) are considered, yielding a total of 32 swaptions.

Since there is no exact pricing solution in the literature for European-style swaptions under multifactor term structure models, the accuracy of each alternative valuation approach is measured by its percentage error with respect to a proxy of the exact swaption

Table 4.1
Prices of ATMF European-Style Swaptions on Plain-Vanilla Interest Rate Swaps with Semiannual Cash Flows and under a Three-Factor Gauss-Markov HJM Model

			Percentage Pricing Errors of Analytical Approximations							
Swaption ×	Monte Carlo		CA Bounds							
Swap	Price	% SE	Lower	Upper	LVA	HA	EE	SD	LA	R1A
1 × 1	0.002082	0.0024%	0.0040%	0.0045%	0.0041%	0.0040%	0.0036%	0.0250%	0.0040%	0.0117%
1 × 2	0.003312	0.0024%	0.0001%	0.0035%	0.0004%	0.0001%	−0.0010%	0.0768%	0.0001%	0.0600%
1 × 5	0.005331	0.0024%	−0.0024%	0.0189%	−0.0014%	−0.0026%	−0.0054%	0.0690%	−0.0018%	0.3130%
1 × 10	0.006558	0.0024%	−0.0014%	0.0435%	0.0003%	−0.0025%	−0.0064%	0.0404%	0.0003%	0.5292%
1 × 15	0.006893	0.0024%	0.0004%	0.0550%	0.0023%	−0.0013%	−0.0054%	0.0382%	0.0026%	0.5954%
1 × 20	0.006984	0.0024%	0.0000%	0.0578%	0.0020%	−0.0018%	−0.0060%	0.0370%	0.0024%	0.6141%
1 × 25	0.007009	0.0024%	0.0000%	0.0587%	0.0021%	−0.0019%	−0.0061%	0.0368%	0.0024%	0.6195%
1 × 30	0.007016	0.0024%	−0.0003%	0.0587%	0.0018%	−0.0022%	−0.0064%	0.0364%	0.0022%	0.6208%
2 × 1	0.002355	0.0024%	−0.0019%	−0.0013%	−0.0017%	−0.0019%	−0.0024%	0.0376%	−0.0019%	0.0055%
2 × 2	0.003843	0.0024%	0.0044%	0.0084%	0.0049%	0.0044%	0.0028%	0.1387%	0.0045%	0.0566%
2 × 5	0.006369	0.0024%	−0.0007%	0.0233%	0.0009%	−0.0011%	−0.0055%	0.1618%	0.0004%	0.2444%
2 × 10	0.007907	0.0025%	−0.0035%	0.0463%	−0.0007%	−0.0057%	−0.0117%	0.1340%	−0.0005%	0.3931%
2 × 15	0.008327	0.0025%	0.0002%	0.0603%	0.0034%	−0.0025%	−0.0094%	0.1340%	0.0039%	0.4414%
2 × 20	0.008442	0.0025%	0.0002%	0.0638%	0.0035%	−0.0028%	−0.0098%	0.1332%	0.0043%	0.4545%
2 × 25	0.008474	0.0025%	−0.0006%	0.0639%	0.0028%	−0.0037%	−0.0107%	0.1321%	0.0035%	0.4573%
2 × 30	0.008483	0.0025%	−0.0009%	0.0639%	0.0024%	−0.0041%	−0.0111%	0.1317%	0.0032%	0.4579%
5 × 1	0.002321	0.0024%	0.0018%	0.0025%	0.0021%	0.0018%	0.0010%	0.0521%	0.0018%	0.0082%
5 × 2	0.003872	0.0024%	−0.0004%	0.0041%	0.0004%	−0.0004%	−0.0026%	0.1726%	−0.0001%	0.0411%
5 × 5	0.006568	0.0025%	−0.0017%	0.0237%	0.0008%	−0.0022%	−0.0090%	0.2652%	0.0002%	0.1749%
5 × 10	0.008216	0.0025%	−0.0015%	0.0501%	0.0028%	−0.0046%	−0.0143%	0.2812%	0.0032%	0.2750%
5 × 15	0.008667	0.0025%	−0.0027%	0.0595%	0.0023%	−0.0072%	−0.0176%	0.2858%	0.0033%	0.3029%
5 × 20	0.008792	0.0025%	−0.0033%	0.0623%	0.0019%	−0.0083%	−0.0189%	0.2865%	0.0031%	0.3106%
5 × 25	0.008826	0.0025%	0.0004%	0.0670%	0.0056%	−0.0048%	−0.0154%	0.2904%	0.0069%	0.3166%

(*continued*)

TABLE 4.1
Continued

Swaption ×			Percentage Pricing Errors of Analytical Approximations							
	Monte Carlo		CA Bounds							
Swap	Price	% SE	Lower	Upper	LVA	HA	EE	SD	LA	R1A
5 × 30	0.008836	0.0025%	0.0007%	0.0676%	0.0059%	−0.0045%	−0.0151%	0.2907%	0.0072%	0.3175%
10 × 1	0.001800	0.0024%	−0.0006%	0.0001%	−0.0003%	−0.0006%	−0.0014%	0.0460%	−0.0005%	0.0054%
10 × 2	0.003020	0.0024%	0.0014%	0.0059%	0.0023%	0.0014%	−0.0011%	0.1740%	0.0017%	0.0395%
10 × 5	0.005153	0.0025%	−0.0011%	0.0243%	0.0016%	−0.0018%	−0.0094%	0.2200%	0.0009%	0.1573%
10 × 10	0.006459	0.0025%	−0.0009%	0.0507%	0.0039%	−0.0044%	−0.0154%	0.3214%	0.0044%	0.2458%
10 × 15	0.006816	0.0025%	−0.0017%	0.0605%	0.0039%	−0.0069%	−0.0186%	0.2602%	0.0050%	0.2707%
10 × 20	0.006914	0.0025%	−0.0018%	0.0639%	0.0040%	−0.0076%	−0.0194%	0.3596%	0.0055%	0.2782%
10 × 25	0.006941	0.0025%	−0.0006%	0.0662%	0.0053%	−0.0065%	−0.0184%	0.3617%	0.0068%	0.2815%
10 × 30	0.006948	0.0025%	0.0043%	0.0714%	0.0103%	−0.0016%	−0.0135%	0.3668%	0.0118%	0.2870%
MPE			−0.0003%	0.0409%	0.0025%	−0.0026%	−0.0087%	0.1687%	0.0028%	0.2871%
MAPE			0.0014%	0.0410%	0.0027%	0.0033%	0.0092%	0.1687%	0.0031%	0.2871%
MAE vol.			0.0001%	0.0010%	0.0001%	0.0001%	0.0002%	0.0046%	0.0001%	0.0077%
CPU (sec.)	180,821.50		5.91		0.29	16.21	1,903.50	0.55	3.17	0.59

This table values 32 ATMF European-style swaptions under the three-factor Gaussian and affine model specified in Collin-Dufresne and Goldstein (2002, exhibit 1) or Schrager and Pelsser (2006, table 4.1). The first column shows the maturity (in years) of the swaption contract and of its underlying swap. The second and third columns contain the Monte Carlo option price estimate and its percentage standard error (%SE)—i.e., the ratio between the standard error and the Monte Carlo price estimate—obtained using 10^9 paths, standard antithetic variables, and the exact probability distribution of the state variables at the maturity date of the swaption contract. The bounds provided by the conditioning approach (CA) in Propositions 3.4 and 3.5 are implemented in the fourth and fifth columns. The sixth column contains the low-variance martingale approximation (LVA) of Schrager and Pelsser (2006). The hyperplane approximation of Singleton and Umantsev (2002) is implemented at a 1% significance level in the seventh column. The eighth column presents the Edgeworth expansion (EE) method of Collin-Dufresne and Goldstein (2002) using a third-order approximation for the log-characteristic function of the terminal state vector. The last three columns contain the percentage pricing errors generated by the stochastic duration (SD) approach of Munk (1999), the lognormal approximation (LA) of Pang (1996), and the rank 1 approximation (R1A) of Brace and Musiela (1994). The last four lines report mean percentage pricing errors (MPE), mean absolute percentage pricing errors (MAPE), mean absolute volatility errors (MAE vol.), and computation times (in seconds). All percentage errors are computed against the Monte Carlo price estimate.

price. Such a proxy is obtained through Monte Carlo simulation, using the exact probability distribution of the state variables at the maturity date of the swaption contract—as described in Appendix B. All Monte Carlo experiments are run over 10^9 paths, and using standard antithetic variables. Consequently, for all the 32 swaptions valued, the ratio between the standard error and the Monte Carlo price estimate (labeled as the percentage standard error) is always below 0.3 basis points.

The Monte Carlo simulations were run under version 1.9.4.13 of GNU Pascal and on an Intel Xeon 3.33 GHz processor with 12 GB of RAM memory, whereas all the approximations tested (from the fourth to the last column of Table 4.1) were implemented through *Matlab* (R2010a). All nonlinear equations involved in the implementation of the conditioning approach, the rank 1 approximation, the stochastic duration approach, or the hyperplane approximation were solved through the built-in function "fsolve" of *Matlab*.[11]

The lower bound (3.21) provided by the conditioning approach proposed is the most accurate approximation tested: It yields a mean absolute percentage pricing error (MAPE, henceforth) of only 0.14 basis points, which is even smaller than the Monte Carlo percentage standard error. Moreover, the tight upper bound shown on the fifth column of Table 4.1 also provides a sharp forecast for the maximum approximation error attached to the conditioning approach: On average, the maximum absolute percentage pricing error associated to the lower bound approximation is equal to only 3.96 basis points (i.e., 0.0410%–0.0014%).

For ATMF swaptions, the low-variance martingale approach of Schrager and Pelsser (2006), the lognormal approximation, and the hyperplane approximation of Singleton and Umantsev (2002)—which is implemented at a 1% significance level—are almost as accurate as the conditioning approach.[12] Additionally, the low-variance martingale approach is also the most efficient approximation: The whole set of 32 swaptions contracts is priced under a CPU time of only 0.29 seconds.

The Edgeworth expansion method of Collin-Dufresne and Goldstein (2002) is the most time-consuming approximation tested: It takes more than 1,903 seconds to price all the 32 swaptions. Since the mth moment of the probability distribution of the model's factors $Y(T_0)$ requires the computation of $(N_0)^m$ terms (where N_0 represents the number of cash flow payment dates of the underlying swap), then the inefficiency of the Edgeworth expansion increases with the time to maturity of the underlying interest rate swap: For instance, the CPU times associated to the 10 × 1 and the 10 × 30 swaptions are equal to 0.033 and 239.37 seconds, respectively. Given the inefficiency of the Edgeworth expansion for long maturity swaps and following, for instance, Chu and Kwok (2007, page 382), the

[11] All programming codes and data used in this paper are available in the following URL: http://home.iscte.pt/~ jpvn/weblinks/CodesJPNandPP.zip.

[12] Note that equation (A.3) of Schrager and Pelsser (2006, page 692) contains a typo and has been replaced by

$$\sigma_{n,N} = \sqrt{\sum_{i=1}^{M} \widehat{\Sigma}_{(ii)}^{2} \left(\widetilde{C}_{n,N}^{(i)}\right)^{2} \left[\frac{e^{2A_{(ii)}T_n} - 1}{2A_{(ii)}}\right] + 2\sum_{i=1}^{M} \sum_{j=i+1}^{M} \rho_{ij} \widehat{\Sigma}_{(ii)} \widehat{\Sigma}_{(jj)} \widetilde{C}_{n,N}^{(i)} \widetilde{C}_{n,N}^{(j)} \left[\frac{e^{[A_{(ii)}+A_{(jj)}]T_n} - 1}{A_{(ii)} + A_{(jj)}}\right]},$$

using the notation of the original paper.

Taylor series expansion of the log-characteristic function of $Y(T_0)$ is truncated only up to the third order.[13]

Overall, the less accurate methods are the rank 1 approximation, and the stochastic duration approach: Their average absolute percentage errors are about 200 and 118 times higher, respectively, than the MAPE associated to the lower bound of the conditioning approach. Moreover, these two approximations systematically overvalue the ATMF swaption contracts under analysis since their reported (positive) mean percentage errors are consistently identical to the corresponding MAPE. Nevertheless, and since swaptions are usually quoted in flat yield volatilities,[14] the penultimate line of Table 4.1 recomputes the mean absolute errors of each analytical approximation in terms of the Black (1976) flat yield volatility implicit to each swaption price. As expected, the error differences among all pricing methods are now less pronounced: All mean absolute volatility errors are lower than 1 basis point, and hence, are also clearly within the typical bid-ask spreads observed in the swaptions market.[15]

The most challenging setup to test the accuracy of all the competing pricing approximations corresponds to the valuation of out-the-money-forward (OTMF) swaption contracts, i.e., option contracts with zero *intrinsic value*. For this purpose, Table 4.2 prices OTMF European-style swaptions under the three-factor Gaussian and affine model adopted in Table 4.1, and for two different strikes that are set at 85% and 90% of the current forward swap rate. Three different option maturities (of 1, 2, and 5 years) and six different swap maturities (of 1, 2, 5, 10, 20, and 30 years) are considered, yielding a total of 36 swaptions.

Percentage pricing errors are computed, under different analytical approximations, for only 27 out of the whole set of 36 swaption contracts. For nine of the swaption contracts considered, the Monte Carlo price estimate is so close to zero (and its standard error is so large) that the percentage pricing errors would not be meaningful for any of the analytical approximations tested. All the nine contracts with missing pricing errors in Table 4.2 possess Monte Carlo price estimates below 10^{-6} as well as percentage standard errors above 0.3%.

As before, the lower bound of the conditioning approach yields the most precise approximation for OTMF swaptions, with a MAPE of only 2.18 basis points. However, and specially for swaptions on long-dated swaps (with a time to maturity of 10 or more years), the upper bound is so loose that it can no longer serve as an indicator for the error of the approximation.

Similarly to Table 4.1, the low-variance approximation of Schrager and Pelsser (2006) is still the fastest pricing methodology (with a CPU time of only 0.27 seconds), but now

[13] Note that equations (26) and (30) of Collin-Dufresne and Goldstein (2002, page 16) contain a typo and have been replaced by

$$B_0(\tau) = -\delta\tau + \frac{1}{2}\sum_{i,j}\frac{\sigma_i\sigma_j\rho_{ij}}{\kappa_i\kappa_j}[\tau - B_{\kappa_i}(\tau) - B_{\kappa_j}(\tau) + B_{\kappa_i+\kappa_j}(\tau)],$$

and

$$M(\tau) = \sum_{i,j}\frac{\sigma_i\sigma_j\rho_{ij}}{\kappa_j}F_i\left[B_{\kappa_i}(\tau) - \frac{e^{-\kappa_j(W-T)} - e^{-\kappa_i\tau-\kappa_j(W-t)}}{\kappa_i+\kappa_j}\right] + \frac{1}{2}\sum_{i,j}\sigma_i\sigma_j\rho_{ij}F_iF_jB_{\kappa_i+\kappa_j}(\tau),$$

respectively, using the notation of the original paper.

[14] I.e., under the usual “market” assumption of lognormally distributed forward swap rates.

[15] For instance, in 2011, the average bid-ask spread of US ATMF European-style swaptions (for the maturities and tenors described in the first column of Table 4.1) ranged between 30 and 67 basis points—data collected daily from Bloomberg between January 01, 2011 and December 31, 2011.

TABLE 4.2
Prices of OTMF European-Style Receiver Swaptions on Plain-Vanilla Interest Rate Swaps with Semiannual Cash Flows and under a Three-Factor Gauss-Markov HJM Model

Swaption ×	Monte Carlo		Percentage Pricing Errors of Analytical Approximations							
			CA Bounds							
Swap	Price	% SE	Lower	Upper	LVA	HA	EE	SD	LA	R1A
			Panel A: Swaptions quoted at 85% of the forward swap rate							
1 × 1	0.000154	0.0111%	0.0197%	0.0255%	1.0749%	0.0199%	0.0075%	0.0913%	0.0117%	0.0488%
1 × 2	0.000103	0.0167%	−0.0009%	0.0941%	1.8874%	0.0070%	0.0603%	0.3109%	−0.0707%	0.2950%
1 × 5	0.000013	0.0543%	−0.1412%	7.3571%	6.9317%	0.3134%	4.3318%	−0.3886%	−0.9917%	2.6053%
1 × 10	0.000000	0.4928%	–	–	–	–	–	–	–	–
1 × 20	0.000000	72.5132%	–	–	–	–	–	–	–	–
1 × 30	0.000000	NA	–	–	–	–	–	–	–	–
2 × 1	0.000277	0.0089%	0.0009%	0.0047%	0.8982%	0.0012%	−0.0206%	0.1194%	−0.0065%	0.0242%
2 × 2	0.000254	0.0117%	0.0150%	0.0674%	1.4757%	0.0233%	−0.0242%	0.4851%	−0.0462%	0.2199%
2 × 5	0.000082	0.0252%	0.0115%	1.6128%	4.8410%	0.3809%	1.4967%	0.3586%	−0.6443%	1.5382%
2 × 10	0.000005	0.1010%	−0.0290%	65.5140%	21.0190%	3.0352%	24.1319%	−1.8662%	−3.2955%	4.2996%
2 × 20	0.000000	2.2576%	–	–	–	–	–	–	–	–
2 × 30	0.000000	23.4079%	–	–	–	–	–	–	–	–
5 × 1	0.000373	0.0076%	0.0057%	0.0091%	0.8107%	0.0060%	−0.0211%	0.1388%	−0.0016%	0.0235%
5 × 2	0.000422	0.0093%	−0.0062%	0.0286%	1.1962%	0.0025%	−0.0914%	0.5094%	−0.0625%	0.1300%
5 × 5	0.000240	0.0156%	0.0142%	0.6121%	3.5897%	0.3248%	0.2654%	0.7578%	−0.5167%	0.8544%
5 × 10	0.000040	0.0399%	−0.0153%	9.2530%	14.4558%	2.3050%	10.4057%	−0.2691%	−2.4444%	2.1306%
5 × 20	0.000001	0.3067%	–	–	–	–	–	–	–	–
5 × 30	0.000000	1.5701%	–	–	–	–	–	–	–	–

(*continued*)

TABLE 4.2
Continued

Swaption × Swap	Monte Carlo		Percentage Pricing Errors of Analytical Approximations							
			CA Bounds							
	Price	% SE	Lower	Upper	LVA	HA	EE	SD	LA	R1A
			Panel B: Swaptions quoted at 90% of the forward swap rate							
1 × 1	0.000437	0.0067%	−0.0039%	−0.0019%	0.4571%	−0.0037%	−0.0150%	0.0476%	−0.0075%	0.0164%
1 × 2	0.000432	0.0085%	−0.0064%	0.0175%	0.7705%	−0.0030%	−0.0445%	0.2124%	−0.0366%	0.1872%
1 × 5	0.000185	0.0159%	−0.0124%	0.5469%	2.5783%	0.1643%	0.1077%	0.0149%	−0.3427%	1.5981%
1 × 10	0.000023	0.0460%	−0.0519%	11.6615%	10.4281%	1.4425%	5.3939%	−1.3531%	−1.6902%	4.5859%
1 × 20	0.000000	0.4229%	–	–	–	–	–	–	–	–
1 × 30	0.000000	2.4067%	–	–	–	–	–	–	–	–
2 × 1	0.000642	0.0058%	−0.0020%	−0.0001%	0.4063%	−0.0018%	−0.0140%	0.0854%	−0.0055%	0.0149%
2 × 2	0.000758	0.0069%	0.0074%	0.0259%	0.6397%	0.0112%	−0.0387%	0.3438%	−0.0204%	0.1483%
2 × 5	0.000523	0.0106%	0.0165%	0.2826%	1.9030%	0.1684%	−0.1359%	0.3738%	−0.2524%	0.9671%
2 × 10	0.000143	0.0216%	−0.0183%	2.4887%	7.1171%	1.1153%	1.4861%	−0.4124%	−1.2304%	2.4083%
2 × 20	0.000007	0.0890%	0.0265%	67.5232%	32.4355%	4.4250%	24.2704%	−2.6396%	−4.0460%	5.0962%
2 × 30	0.000001	0.2661%	0.0630%	700.0488%	69.2668%	7.3914%	59.6131%	−4.7587%	−6.6443%	7.0860%
5 × 1	0.000756	0.0053%	−0.0038%	−0.0020%	0.3744%	−0.0036%	−0.0168%	0.0973%	−0.0074%	0.0095%
5 × 2	0.001010	0.0060%	0.0190%	0.0344%	0.5624%	0.0232%	−0.0314%	0.4012%	−0.0076%	0.1167%
5 × 5	0.000938	0.0082%	−0.0100%	0.1511%	1.4690%	0.1244%	−0.2617%	0.5970%	−0.2387%	0.5424%
5 × 10	0.000408	0.0136%	−0.0114%	0.9370%	5.2659%	0.8892%	−0.0271%	0.3407%	−0.9567%	1.2640%
5 × 20	0.000053	0.0360%	−0.0265%	9.9383%	22.0731%	3.2570%	9.9860%	−0.7644%	−3.0194%	2.4497%
5 × 30	0.000011	0.0753%	0.0511%	49.4814%	44.4240%	5.4328%	30.0070%	−1.8024%	−4.7930%	3.3877%

(continued)

TABLE 4.2
Continued

			Percentage Pricing Errors of Analytical Approximations							
Swaption ×	Monte Carlo		CA Bounds							
Swap	Price	% SE	Lower	Upper	LVA	HA	EE	SD	LA	R1A
Panel B: Swaptions quoted at 90% of the forward swap rate										
MPE			−0.0033%	34.3597%	9.5686%	1.1427%	6.3267%	−0.3322%	−1.1617%	1.5573%
MAPE			0.0218%	34.3600%	9.5686%	1.1436%	6.3817%	0.7237%	1.1626%	1.5573%
MAE vol.			0.0002%	0.0319%	0.0311%	0.0030%	0.0131%	0.0041%	0.0037%	0.0067%
CPU (sec.)	175,024.62		5.48		0.27	15.40	1,906.56	0.57	2.93	0.60

This table values 36 OTMF European-style swaptions under the three-factor Gaussian and affine model specified in Table 4.1. The first column shows the maturity (in years) of the swaption contract and of its underlying swap. The second and third columns contain the Monte Carlo option price estimate and its percentage standard error (%SE) obtained using 10^9 paths, standard antithetic variables, and the exact probability distribution of the state variables at the maturity date of the swaption contract. The %SE that are not available (NA) correspond to Monte Carlo price estimates that are equal to zero. The bounds provided by the conditioning approach (CA) in Propositions 3.4 and 3.5 are implemented in the fourth and fifth columns. The sixth column contains the low-variance martingale approximation (LVA) of Schrager and Pelsser (2006). The hyperplane approximation of Singleton and Umantsev (2002) is implemented at a 1% significance level in the seventh column. The eighth column presents the Edgeworth expansion (EE) method of Collin-Dufresne and Goldstein (2002) using a third-order approximation for the log-characteristic function of the terminal state vector. The last three columns contain the percentage pricing errors generated by the stochastic duration (SD) approach of Munk (1999), the lognormal approximation (LA) of Pang (1996), and the rank 1 approximation (R1A) of Brace and Musiela (1994). The last four lines report mean percentage pricing errors (MPE), mean absolute percentage pricing errors (MAPE), mean absolute volatility errors (MAE vol.), and computation times (in seconds). All percentage errors are computed against the Monte Carlo price estimate, except for the nine contracts with percentage standard errors above 0.3%.

this is also the most inaccurate approximation tested—yielding an average percentage error above 9.56%. The sixth column of Table 4.2 shows that the significant upward bias of the low-variance approximation is mainly due to the large pricing errors attached to the swaption contracts on 20 and 30 years' interest rate swaps, which were not considered in Schrager and Pelsser (2006, table 6.2). However, and since these swaption contracts with long-term tenors possess small dollar values, the penultimate line of Table 4.2 shows that the corresponding mean absolute Black (1976) flat yield volatility errors are only around 3 basis points.

Again, the Edgeworth expansion technique is the most time-consuming approximation tested as well as one of the less accurate pricing methods: It yields a CPU time of 1, 906.56 seconds and a MAPE of 6.38%. Overall, the low-variance approximation and the Edgeworth expansion technique are both dominated by the hyperplane approach, and even by the simpler rank 1, lognormal and stochastic duration approximations: For instance, the stochastic duration approach combines a CPU time of only 0.57 seconds with a MAPE of 0.72%. Nevertheless, note that the mean absolute error of the Munk (1999) approach is still 33 times higher than the MAPE reported by the lower bound of the conditioning method.

Table 4.3 shows that differences in accuracy among the alternative analytical approximations tested are much less pronounced for in-the-money-forward (ITMF) swaption contracts. Table 4.3 prices ITMF European-style swaptions under the same three-factor Gaussian and affine model already adopted in Tables 4.1 and 4.2, and for two different strikes that are set at 110% and 115% of the current forward swap rate. Similarly to Table 4.2, three different option maturities (of 1, 2, and 5 years) and six different swap maturities (of 1, 2, 5, 10, 20, and 30 years) are considered, yielding a total of 36 swaptions.

Once again, the lower bound of the conditioning approach is the most accurate valuation method with a MAPE of only 0.02 basis points. Moreover, the conditioning approach also offers tight error bounds since the upper bound on the swaption prices is even smaller than most of the price estimates produced by the alternative approximations tested: Its MAPE of 0.52 basis points is only above the average pricing errors associated to the hyperplane, the lognormal, and the Edgeworth approximations.

As before, the low-variance approximation—that is still the fastest valuation method—is also the less accurate approach, yielding an average percentage error of −2.12 basis points (that corresponds to a mean absolute volatility error of 4.34 basis points). However, and in contrast with Table 4.2, the low-variance approximation consistently underprices ITMF swaptions since its mean percentage error and MAPE are exactly symmetrical.

Finally, a word of caution must be said about the accuracy of the proxy used for the exact price of the swaption contracts. Tables 4.1, 4.2, and 4.3 show that some of the percentage errors computed for the lower (upper) bound of the conditioning approach are slightly positive (negative), meaning that the Monte Carlo price estimate can be, for some contracts, slightly below or above the exact lower or upper price bound, respectively. Nevertheless, and given the large number of simulations run, the average pricing errors (over all swaption contacts in Tables 4.1, 4.2, and 4.3) are always nonpositive (nonnegative) for the lower (upper) bound of the conditioning approach.[16]

[16] One way to avoid the noise introduced by the Monte Carlo estimates would be to consider a single-factor Gaussian term structure model, because an exact analytical pricing solution for European-style swaptions is provided by Jamshidian (1989). This approach is followed, for instance, by Schrager and Pelsser (2006, table 4.2). However, such a test would be redundant in our case, because it is easy to show—following El Karoui and Rochet (1989, page 22)—that the upper and lower bounds of the conditioning approach collapse

Table 4.3
Prices of ITMF European-Style Receiver Swaptions on Plain-Vanilla Interest Rate Swaps with Semiannual Cash Flows and under a Three-Factor Gauss-Markov HJM Model

Swaption ×	Monte Carlo		Percentage Pricing Errors of Analytical Approximations							
			CA Bounds							
Swap	Price	% SE	Lower	Upper	LVA	HA	EE	SD	LA	R1A
			Panel A: Swaptions quoted at 110% of the forward swap rate							
1 × 1	0.005634	0.0005%	0.0008%	0.0010%	−0.0349%	0.0008%	−0.0001%	0.0063%	0.0011%	0.0027%
1 × 2	0.010674	0.0004%	0.0001%	0.0012%	−0.0313%	−0.0001%	−0.0015%	0.0137%	0.0016%	0.0096%
1 × 5	0.024284	0.0001%	−0.0002%	0.0049%	−0.0201%	−0.0018%	0.0005%	0.0063%	0.0029%	0.0150%
1 × 10	0.042493	0.0000%	0.0000%	0.0075%	−0.0061%	−0.0011%	0.0032%	0.0014%	0.0011%	0.0034%
1 × 20	0.065979	0.0000%	0.0000%	0.0066%	−0.0002%	0.0000%	0.0006%	0.0001%	0.0000%	0.0001%
1 × 30	0.078861	0.0000%	0.0000%	0.0057%	0.0000%	0.0000%	0.0001%	0.0000%	0.0000%	0.0000%
2 × 1	0.005689	0.0007%	0.0002%	0.0004%	−0.0458%	0.0001%	−0.0012%	0.0131%	0.0006%	0.0025%
2 × 2	0.010659	0.0005%	0.0001%	0.0017%	−0.0447%	−0.0002%	−0.0031%	0.0338%	0.0026%	0.0123%
2 × 5	0.023572	0.0003%	0.0001%	0.0072%	−0.0418%	−0.0039%	−0.0036%	0.0243%	0.0073%	0.0259%
2 × 10	0.040481	0.0001%	0.0001%	0.0107%	−0.0259%	−0.0047%	0.0044%	0.0093%	0.0055%	0.0113%
2 × 20	0.062458	0.0000%	0.0000%	0.0093%	−0.0045%	−0.0008%	0.0048%	0.0011%	0.0008%	0.0011%
2 × 30	0.074583	0.0000%	0.0000%	0.0080%	−0.0010%	−0.0002%	0.0021%	0.0002%	0.0001%	0.0002%
5 × 1	0.005142	0.0008%	0.0002%	0.0005%	−0.0552%	0.0002%	−0.0017%	0.0196%	0.0009%	0.0026%
5 × 2	0.009537	0.0007%	−0.0005%	0.0015%	−0.0577%	−0.0010%	−0.0058%	0.0552%	0.0030%	0.0121%
5 × 5	0.020483	0.0004%	0.0003%	0.0092%	−0.0669%	−0.0071%	−0.0112%	0.0592%	0.0132%	0.0312%
5 × 10	0.034347	0.0002%	−0.0002%	0.0132%	−0.0636%	−0.0129%	−0.0027%	0.0321%	0.0137%	0.0190%
5 × 20	0.052436	0.0001%	0.0000%	0.0119%	−0.0248%	−0.0046%	0.0104%	0.0079%	0.0044%	0.0039%
5 × 30	0.062523	0.0000%	0.0000%	0.0104%	−0.0096%	−0.0016%	0.0090%	0.0027%	0.0016%	0.0012%

(*continued*)

TABLE 4.3
Continued

			Percentage Pricing Errors of Analytical Approximations							
Swaption ×	Monte Carlo		CA Bounds							
Swap	Price	% SE	Lower	Upper	LVA	HA	EE	SD	LA	R1A
			Panel B: Swaptions quoted at 115% of the forward swap rate							
1 × 1	0.007948	0.0002%	−0.0002%	−0.0001%	−0.0207%	−0.0002%	−0.0005%	0.0021%	0.0000%	0.0006%
1 × 2	0.015460	0.0001%	−0.0001%	0.0008%	−0.0127%	−0.0001%	0.0003%	0.0040%	0.0006%	0.0026%
1 × 5	0.036149	0.0000%	0.0000%	0.0036%	−0.0027%	−0.0002%	0.0018%	0.0007%	0.0004%	0.0014%
1 × 10	0.063698	0.0000%	0.0000%	0.0052%	−0.0001%	0.0000%	0.0004%	0.0000%	0.0000%	0.0000%
1 × 20	0.098967	0.0000%	0.0000%	0.0046%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%
1 × 30	0.118291	0.0000%	0.0000%	0.0040%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%
2 × 1	0.007845	0.0003%	−0.0001%	0.0001%	−0.0319%	−0.0001%	−0.0009%	0.0062%	0.0002%	0.0010%
2 × 2	0.015098	0.0002%	0.0002%	0.0014%	−0.0245%	0.0000%	−0.0006%	0.0135%	0.0016%	0.0049%
2 × 5	0.034636	0.0001%	0.0001%	0.0051%	−0.0117%	−0.0011%	0.0033%	0.0056%	0.0022%	0.0052%
2 × 10	0.060483	0.0000%	0.0000%	0.0073%	−0.0021%	−0.0004%	0.0034%	0.0007%	0.0004%	0.0006%
2 × 20	0.093669	0.0000%	0.0000%	0.0065%	0.0000%	0.0000%	0.0002%	0.0000%	0.0000%	0.0000%
2 × 30	0.111871	0.0000%	0.0000%	0.0056%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%
5 × 1	0.006950	0.0004%	−0.0005%	−0.0002%	−0.0437%	−0.0005%	−0.0019%	0.0102%	0.0000%	0.0008%
5 × 2	0.013205	0.0003%	0.0010%	0.0025%	−0.0374%	0.0007%	−0.0017%	0.0278%	0.0035%	0.0069%
5 × 5	0.029533	0.0002%	−0.0002%	0.0062%	−0.0295%	−0.0036%	0.0008%	0.0196%	0.0056%	0.0092%
5 × 10	0.050896	0.0001%	0.0000%	0.0093%	−0.0122%	−0.0026%	0.0087%	0.0053%	0.0028%	0.0026%
5 × 20	0.078540	0.0000%	0.0000%	0.0083%	−0.0009%	−0.0002%	0.0031%	0.0002%	0.0001%	0.0001%
5 × 30	0.093753	0.0000%	0.0000%	0.0071%	−0.0001%	−0.0001%	0.0008%	0.0000%	0.0000%	0.0000%

(*continued*)

TABLE 4.3
Continued

			Percentage Pricing Errors of Analytical Approximations							
Swaption ×	Monte Carlo		CA Bounds							
Swap	Price	% SE	Lower	Upper	LVA	HA	EE	SD	LA	R1A
Panel B: Swaptions quoted at 115% of the forward swap rate										
MPE			0.0000%	0.0052%	−0.0212%	−0.0013%	0.0006%	0.0106%	0.0022%	0.0053%
MAPE			0.0002%	0.0052%	0.0212%	0.0014%	0.0026%	0.0106%	0.0022%	0.0053%
MAE vol.			0.0076%	0.1455%	0.0434%	0.0124%	0.0387%	0.0138%	0.0093%	0.0129%
CPU (sec.)	174,856.55		5.46		0.28	15.47	1,903.70	0.56	2.96	0.60

This table values 36 ITMF European-style swaptions under the three-factor Gaussian and affine model specified in Tables 4.1 and 4.2. The first column shows the maturity (in years) of the swaption contract and its underlying swap. The second and third columns contain the Monte Carlo option price estimate and its percentage standard error (%SE) obtained using 10^9 paths, standard antithetic variables, and the exact probability distribution of the state variables at the maturity date of the swaption contract. The bounds provided by the conditioning approach (CA) in Propositions 3.4 and 3.5 are implemented in the fourth and fifth columns. The sixth column contains the LVA of Schrager and Pelsser (2006). The hyperplane approximation of Singleton and Umantsev (2002) is implemented at a 1% significance level in the seventh column. The eighth column presents the Edgeworth expansion (EE) method of Collin-Dufresne and Goldstein (2002) using a third-order approximation for the log-characteristic function of the terminal state vector. The last three columns contain the percentage pricing errors generated by the stochastic duration (SD) approach of Munk (1999), the lognormal approximation (LA) of Pang (1996), and the rank 1 approximation (R1A) of Brace and Musiela (1994). The last four lines report mean percentage pricing errors (MPE), mean absolute percentage pricing errors (MAPE), mean absolute volatility errors (MAE vol.), and computation times (in seconds). All percentage errors are computed against the Monte Carlo price estimate.

Note that this (small) bias occurs even though the Monte Carlo price estimates are produced using a large number of simulations (10^9 paths), antithetic variables, and the exact probability distribution of the state variables at the maturity date of the swaption contract, which is, to the authors knowledge, the most demanding setting already used in the literature on the pricing of swaption contracts. Therefore, one should expect this Monte Carlo bias to be even more relevant in the previous literature on the pricing of European-style swaptions; for instance, Schrager and Pelsser (2006, page 689) test their low-variance approximation using only 500,000 Monte Carlo simulations. The use, in this paper, of a more demanding Monte Carlo setting, and the inclusion of longer swaption contracts (with 5 and 10 years of time to maturity) on long-term interest rate swaps (namely on 20 and 30 years' swaps) explains the novel findings with respect to the previous literature concerning the inaccuracy of the low-variance, and of the Edgeworth approximations for OTMF contracts.

5. CONCLUSIONS

This paper offers two contributions to the literature on swaptions pricing. First, it derives a new analytical approximation for European-style swaptions under a multifactor Gauss–Markov framework, using the *conditioning approach* proposed by Curran (1994), Rogers and Shi (1995), and Nielsen and Sandmann (2002). Second, a comprehensive and rigorous Monte Carlo study is run to compare, in terms of efficiency and accuracy, all the approximations already proposed in the literature for European-style swaptions under multifactor term structure models.

The numerical results obtained show that the exact lower bound of the swaption price provided by the conditioning approach is the most accurate pricing method for ATMF, OTMF, and ITMF contracts. Moreover, the conditioning approach proposed in this paper also offers tight bounds for the approximation error, because the analytical lower and upper bounds proposed in Propositions 3.4 and 3.5 are usually very close to each other (except for some deep OTMF contracts).

By contrast, the low-variance martingale method of Schrager and Pelsser (2006) — which is the fastest pricing approach tested—and the Edgeworth expansion yield the highest pricing errors for OTMF swaption contracts. The latter approach proposed by Collin-Dufresne and Goldstein (2002) is also extremely time-consuming for swaption contracts on long-term swaps. The hyperplane approximation of Singleton and Umantsev (2002) is more accurate and faster than the Edgeworth expansion technique, but still less accurate and slower than the proposed conditioning approach. Finally, the simpler rank 1, lognormal, and stochastic duration approximations are also very fast to implement but, nevertheless, still much less accurate than the lower bound of the conditioning approach.

APPENDIX A: FUNCTION $l(t_0, T_a, T_b, T_c)$ UNDER THE $\mathbb{A}_0(n)$ SPECIFICATION

Under the nested $\mathbb{A}_0(n)$ specification, function $l(t_0, T_a, T_b, T_c)$ can be computed explicitly. For this purpose, equations (2.4) and (4.7) yield:

into the exact swaption price under any single-factor Gauss–Markov and time-homogeneous term structure model.

$$\text{(A.1)}\quad \begin{aligned} &l(t_0, T_a, T_b, T_c)\\ &\quad = G' \cdot a^{-1} \cdot \int_{t_0}^{T_a} [e^{a(T_a-s)} - e^{a(T_b-s)}]' \cdot \Sigma \cdot \Sigma' \cdot [e^{a'(T_a-s)} - e^{a'(T_c-s)}]\, ds \cdot (a^{-1})' \cdot G. \end{aligned}$$

Using definition (4.6), equation (A.1) can be restated as

$$\text{(A.2)}\quad \begin{aligned} l(t_0, T_a, T_b, T_c) &= G' \cdot a^{-1} \cdot [\Delta(t_0, T_a) - \Delta(t_0, T_a) \cdot e^{a'(T_c-T_a)} - e^{a(T_b-T_a)} \cdot \Delta(t_0, T_a)\\ &\qquad + e^{a(T_b-T_a)} \cdot \Delta(t_0, T_a) \cdot e^{a'(T_c-T_a)}] \cdot (a^{-1})' \cdot G\\ &= B(T_a, T_b)' \cdot \Delta(t_0, T_a) \cdot B(T_a, T_c), \end{aligned}$$

where the last line follows from equation (4.4).

APPENDIX B: MONTE CARLO SIMULATION

To implement equation (3.4) through Monte Carlo simulation, it is necessary to rewrite the stochastic differential equation (4.2) under the forward probability measure $\mathbb{Q}_0$. For this purpose, considering equation (4.7) and using, for instance, Nunes (2004, equation 2.9), it follows that

$$\text{(B.1)}\qquad dW^{\mathbb{Q}_0}(t) = dW^{\mathbb{Q}}(t) - \Sigma' \cdot B(t, T_0)\, dt$$

is also a vector of standard Brownian motion increments in $\mathbb{R}^n$ and under the forward measure $\mathbb{Q}_0$. Hence, equations (4.2) and (B.1) imply that

$$\text{(B.2)}\qquad dY(t) = [a \cdot Y(t) + b + \Sigma \cdot \Sigma' \cdot B(t, T_0)]\, dt + \Sigma \cdot dW^{\mathbb{Q}_0}(t).$$

Applying Itô's lemma to $e^{-at} \cdot Y(t)$ and using definition (4.4), the following strong solution is obtained for the stochastic differential equation (B.2):

$$\text{(B.3)}\qquad \begin{aligned} Y(t_0) = e^{a(T_0-t_0)} \cdot Y(t_0) + [e^{a(T_0-t_0)} - I_n] \cdot a^{-1} \cdot [b + \Sigma \cdot \Sigma' \cdot (a^{-1})' \cdot G]\\ - \Delta(t_0, T_0) \cdot (a^{-1})' \cdot G + \int_{t_0}^{T_0} e^{a(T_0-u)} \cdot \Sigma \cdot dW^{\mathbb{Q}_0}(u). \end{aligned}$$

Equation (B.3) yields the exact probability distribution of the state vector $Y(T_0)$ after noting that Arnold (1992, corollary 4.5.6) implies that the Itô's integral $\int_{t_0}^{T_0} e^{a(T_0-u)} \cdot \Sigma \cdot dW^{\mathbb{Q}_0}(u)$ possesses a n-dimensional normal distribution with zero mean and variance–covariance matrix equal to $\Delta(t_0, T_0)$. The Itô's integral is simulated by generating n normally distributed deviates with zero mean and unit variance—through routines "ran3" and "gasdev" of Press, Flannery, Teukolsky, and Vetterling (1994)—that are then correlated using the Cholesky decomposition of matrix $\Delta(t_0, T_0)$.

REFERENCES

ARNOLD, L. (1992): *Stochastic Differential Equations: Theory and Applications*, Malabar: Krieger Publishing Company.

BLACK, F. (1976): The Pricing of Commodity Contracts, *J. Financ. Econ.* 3, 167–179.

Brace, A., and M. Musiela (1994): A Multifactor Gauss Markov Implementation of Heath, Jarrow, and Morton, *Math. Finance* 4, 259–283.

Chu, C., and Y. Kwok (2007): Valuation of Guaranteed Annuity Options in Affine Term Structure Models, *Int. J. Theor. Appl. Finance* 10, 363–387.

Collin-Dufresne, P., and R. Goldstein (2002): Pricing Swaptions within an Affine Framework, *J. Derivat.* 10, 9–26.

Collin-Dufresne, P., and R. Goldstein (2003): Generalizing the Affine Framework to HJM and Random Field Models, Working Paper, Carnegie Mellon University and Washington University.

Cox, J., J. Ingersoll, and S. Ross (1985): A Theory of the Term Structure of Interest Rates, *Econometrica* 53, 385–407.

Curran, M. (1994): Valuing Asian and Portfolio Options by Conditioning on the Geometric Mean Price, *Manag. Sci.* 40, 1705–1711.

Dai, Q., and K. Singleton (2000): Specification Analysis of Affine Term Structure Models, *J. Finance* 55, 1943–1978.

Duffie, D., and R. Kan (1996): A Yield-Factor Model of Interest Rates, *Math. Finance* 6, 379–406.

Duffie, D., J. Pan, and K. Singleton (2000): Transform Analysis and Asset Pricing for Affine Jump-Diffusions, *Econometrica* 68, 1343–1376.

El Karoui, N., and J.-C. Rochet (1989): A Pricing Formula for Options on Coupon Bonds, Working Paper 72, SEEDS.

Geman, H., N. El Karoui, and J.-C. Rochet (1995): Changes of Numéraire, Changes of Probability Measure and Option Pricing, *J. Appl. Probab.* 32, 443–458.

Heath, D., R. Jarrow, and A. Morton (1992): Bond Pricing and the Term Structure of Interest Rates: A New Methodology for Contingent Claims Valuation, *Econometrica* 60, 77–105.

Jamshidian, F. (1989): An Exact Bond Option Pricing Formula, *J. Finance* 44, 205–209.

Joslin, S. (2010): Pricing and Hedging Volatility Risk in Fixed Income Markets, Working Paper, MIT Sloan School of Management.

Kristensen, D., and A. Mele (2011): Adding and Subtracting Black-Scholes: A New Approach to Approximating Derivative Prices in Continuous-Time Models, *J. Financ. Econ.* 102, 390–415.

Lamberton, D., and B. Lapeyre (1996): *Introduction to Stochastic Calculus Applied to Finance*, London: Chapman & Hall/CRC.

Langetieg, T. (1980): A Multivariate Model of the Term Structure, *J. Finance* 35, 71–97.

Longstaff, F. (1993): The Valuation of Options on Coupon Bonds, *J. Banking Finance* 17, 27–42.

Longstaff, F., P. Santa-Clara, and E. Schwartz (2001): The Relative Valuation of Caps and Swaptions: Theory and Empirical Evidence, *J. Finance* 56, 2067–2109.

Lord, R. (2006): Partially Exact and Bounded Approximations for Arithmetic Asian Options, *J. Comput. Finance* 10, 1–52.

Lund, J. (1994): Econometric Analysis of Continuous-Time Arbitrage-Free Models of the Term Structure of Interest Rates, Working Paper, The Aarhus School of Business.

Mood, A., F. Graybill, and D. Boes (1974): *Introduction to the Theory of Statistics*, 3rd ed., Singapore: McGraw-Hill.

Munk, C. (1999): Stochastic Duration and Fast Coupon Bond Option Pricing in Multi-Factor Models, *Rev. Derivat. Res.* 3, 157–181.

NIELSEN, J., and K. SANDMANN (2002): Pricing of Asian Exchange Rate Options under Stochastic Interest Rates as a Sum of Options, *Finance Stoch.* 6, 355–370.

NUNES, J. (2004): Multifactor Valuation of Floating Range Notes, *Math. Finance* 14, 79–97.

NUNES, J., L. CLEWLOW, and S. HODGES (1999): Interest Rate Derivatives in a Duffie and Kan Model with Stochastic Volatility: An Arrow-Debreu Pricing Approach, *Rev. Derivat. Res.* 3, 5–66.

PANG, K. (1996): Can We Price Caps and Swaptions Consistently? Working Paper, Financial Options Research Centre, University of Warwick.

PRESS, W., B. FLANNERY, S. TEUKOLSKY, and W. VETTERLING (1994): *Numerical Recipes in Pascal: The Art of Scientific Computing*, Cambridge: Cambridge University Press.

ROGERS, L., and Z. SHI (1995): The Value of an Asian Option, *J. Appl. Probab.* 32, 1077–1088.

SCHRAGER, D., and A. PELSSER (2006): Pricing Swaptions and Coupon Bond Options in Affine Term Structure Models, *Math. Finance* 16, 673–694.

SINGLETON, K., and L. UMANTSEV (2002): Pricing Coupon-Bond Options and Swaptions in Affine Term Structure Models, *Math. Finance* 12, 427–446.

TURNBULL, S., and L. WAKEMAN (1991): A Quick Algorithm for Pricing European Average Options, *J. Financ. Quant. Anal.* 26, 377–389.

VAN LOAN, C. (1978): Computing Integrals Involving the Matrix Exponential, *IEEE Trans. Autom. Control* 23, 395–404.

VASIČEK, O. (1977): An Equilibrium Characterization of the Term Structure, *J. Financ. Econ.* 5, 177–188.

WEI, J. (1997): A Simple Approach to Bond Option Pricing, *J. Futures Markets* 17, 131–160.

MATHEMATICAL FINANCE

An International Journal of Mathematics, Statistics and Financial Economics

Mathematical Finance, Vol. 24, No. 4 (October 2014), 790–820

SWAPTION PRICING IN AFFINE AND OTHER MODELS

DON H. KIM

Board of Governors of the Federal Reserve System

This paper shows that Singleton and Umantsev's method for swaption pricing in affine models can be simplified and extended to other models. Two alternative methods for approximating the option exercise boundary are introduced: one based on the multivariate Taylor series expansion, and the other based on duration-matched zero-coupon bond approximation. Applied to affine models and quadratic-Gaussian models, these methods are found to give accurate swaption prices.

KEY WORDS: swaptions, coupon bond options, affine models, quadratic-Gaussian models.

1. INTRODUCTION

In recent years there has been considerable interest in joint modeling of interest rates and interest rate derivatives within a time-consistent no-arbitrage term structure model.[1] The empirical investigation in this area has often used interest rate cap data (e.g., Li and Zhao 2006; Almeida, Graveline, and Joslin 2011), as popular time-consistent models like affine models lead to relatively simple pricing of zero-coupon bond options (caps are a collection of zero-coupon bond options). Swaptions are also a popular and important interest-rate derivative security, but they are seldom used,[2] most likely because swaptions, effectively being an option on a *coupon bond*, do not have a simple pricing formula in affine and other models in a multifactor setting, hence empirical study with swaptions is computationally more involved. For some questions of interest, such as the cap-swaption relative pricing puzzle,[3] one cannot avoid using swaption data.

In the case of affine models, several *approximate* methods for pricing swaptions are known, including the methods of Munk (1999),[4] Singleton and Umantsev (2002), Collin-Dufresne and Goldstein (2002), and Schrager and Pelsser (2006). Among these, the method of Singleton and Umantsev (2002, henceforth SU) has an advantage that it

Helpful comments from Young-Ho Eom, Ken Singleton, and two anonymous referees are gratefully acknowledged. The opinions expressed in this paper are those of the author and do not necessarily reflect the views of the Board of Governors of the Federal Reserve System.

Manuscript received December 2011; final revision received July 2012.

Address correspondence to Don H. Kim, Division of Monetary Affairs, Federal Reserve Board, Washington DC 20551, USA; e-mail: don.h.kim@frb.gov.

[1] By time-consistent models, we mean term structure models in which the model parameters do not change over time and the yield and volatility variations are captured by state variables. Compared to Heath, Jarrow, and Morton (1992)-type models (including the popular Libor Market Models) which are recalibrated every day, time-consistent models with small number of state variables may fit yields and interest rate derivative prices less accurately, but they can address questions that the HJM-type models cannot (due to the fact that the yield curve is an exogenous input in those models). Some recent efforts like Heidari and Wu (2009) aim at the middle ground, by retaining time-consistent features while capturing both bond yields and bond derivative prices well.

[2] Some exceptions include Jagannathan, Kaplin, and Sun (2003) and Joslin (2010).

[3] See, for example, Longstaff, Santa-Clara, and Schwartz (2002).

[4] Munk's method is a generalization of Wei's (1997) method to multifactor settings.

DOI: 10.1111/mafi.12014

yields accurate swaption prices for a wide range of strikes (including substantially out-of-money cases).[5]

However, as Schrager and Pelsser (2006, henceforth SP) have pointed out, the SU method also has some drawbacks, namely that (1) SU's procedure for obtaining the approximate boundary requires knowing the probability density function of the state variables, which is not available in closed-form for general affine models (though it is available for the 2-factor Cox–Ingersoll–Ross (CIR) model that SU used for illustration) and that (2) one needs as many Fourier inversions as there are cash flows in the swaption contract. These features can make the SU method much more time-consuming than other methods such as Munk (1999) and SP (2006).[6]

In this paper, we show that the "drawbacks" of the SU method can be overcome to a large extent. We retain SU's (2002) key idea of approximating the exercise boundary, but use a different method for obtaining the approximate boundary that does not require the knowledge of the probability density function. In addition, we show that it is not necessary to compute the Fourier inversion (T-forward measure probability) for all cash flow dates T. Since the T-forward measure probabilities are a smooth monotonic function of T, with the computation of T-forward measure probabilities at $5 \sim 6$ values of T one can evaluate accurately all the necessary probabilities by interpolation. With these modifications, the computation of swaption prices would be only about $2 \sim 3$ times more time-consuming than Munk's (1999) stochastic duration approach (for a substantial gain in accuracy).

To approximate the exercise boundary, we propose two alternative methods: one is to approximate the boundary using the multivariate Taylor series expansion of the log price around a suitably chosen state vector. Another is to use the boundary implied by a zero-coupon bond of a suitably chosen "effective maturity." With the latter approximation, there are several possible choices of the effective maturity; we explore using the stochastic duration of Cox, Ingersoll, and Ross (1979) as well as the simpler duration concepts due to Fisher and Weil (1979) and Macaulay. We show that the Taylor series approximation and the zero-coupon bond approximation with stochastic duration lead to accurate results for the entire range of strikes of practical interest, and provide a discussion of *why* the method of approximating the boundary works so well.

Another contribution of this paper is to show that the method of approximating the exercise boundary can be also applied to models beyond the affine class. Although it was not clear how SU's (2002) original approximation could be extended to other models, the Taylor series approximation and the zero-coupon bond approximation in this paper allow straightforward applications to other models (that have tractable pricing of zero-coupon bond options). We demonstrate this in the setting of quadratic-Gaussian (QG) models. QG models are a well-known class of term structure models, and at times have distinct advantages over affine models, but the existing literature has not studied swaption pricing within QG models.

The rest of the paper is organized as follows. In Section 2, after briefly reviewing the relevant literature on swaption pricing, we discuss our methods in general terms. In particular, the two methods for approximating the boundary—Taylor series method

[5] Most of the available methods generate out-of-the-money swaption prices that are not very accurate. For example, the method of Collin-Dufresne and Goldstein (2002) is based on the Edgeworth expansion which can work poorly in the tail regions of the distribution, and the method of Schrager and Pelsser (2006) also becomes notably less accurate for out-of-the-money options.

[6] As SP (2006) note, the method of Collin-Dufresne and Goldstein (2002) is also time-consuming in general affine models (that do not have a closed-form solution for bond prices).

(Approximation A) and the zero-coupon bond method (Approximation B)—are explained. In Section 3, we discuss the implementation of the methods in the case of affine models. A specific illustration is given in the setting of the 2-factor CIR model, and further test of the performance of the methods is done using the three affine model settings of Schrager and Pelsser (2006). In Section 4, we discuss the implementation in the case of QG models, and present an illustration with a 2-factor QG model. Section 5 concludes.

2. SWAPTION PRICING

2.1. Review of the Relevant Literature

A swaption is an option to get into an interest rate swap contract. To fix notation, let t_0 be the swaption expiration date, and let $t_1, t_2, \ldots, t_N$ be the cash flow dates (swap payment dates). The *tenor* (swap maturity at time t_0) is thus $t_N - t_0$, and the option maturity is $t_0 - t$ at time t. For simplicity, we shall assume that the intervals $t_1 - t_0, t_2 - t_1, \ldots,$ are the same, and denote the interval by δt.

Note that the time t value of the swap contract to the payer side with fixed rate K is given by

$$V^{pay}\left(t, \{t_i\}_{i=1}^N; K\right) = Z(t, t_0) - \left[K\delta t \sum_{i=1}^{N-1} Z(t, t_i) + (1 + K\delta t)Z(t, t_N)\right], \tag{2.1}$$

where $Z(t, T)$ is the time t price of a zero-coupon bond maturing at time T (with face value of 1), i.e., a zero-coupon bond with time-to-maturity $T - t$. (Note that $Z(t, t) = 1$). The fixed rate which makes the value of the swap zero (forward swap rate) is given by

$$K_f(t) = \frac{Z(t, t_0) - Z(t, t_N)}{\delta t \sum_{i=1}^{N} Z(t, t_i)}. \tag{2.2}$$

ATMF (at-the-money-forward) swaptions are swaptions with fixed rate given by $K = K_f(t)$.

The time t price of the payer swaption with fixed rate (strike rate) K is simply

$$S^{pay}(t; K) = \mathrm{E}_t\left[e^{-\int_t^{t_0} r(x_s)\,ds} \max[V^{pay}(t_0, \{t_i\}_{i=1}^N; K), 0]\right], \tag{2.3}$$

where E_t denotes the time-t conditional expectation *in the risk-neutral measure*, and $r_t = r(x_t)$ is the short rate at time t, where x_t denotes the n-dimensional vector of state variables in the term structure model (n-factor model). Different term structure models have different specification of $r(x_t)$ and/or different dynamics of the x_t vector. Note that the zero-coupon bond price $Z(t, T)$ is given by $Z(t, T) = \mathrm{E}_t[e^{-\int_t^T r(x_s)\,ds}]$.

Note that (2.3) can be written as

$$S^{pay}(t; K) = \mathrm{E}_t\left[e^{-\int_t^{t_0} r(x_s) ds} \max[1 - P(t_0, \{t_i\}_{i=1}^N; K), 0]\right], \tag{2.4}$$

where

$$
\begin{aligned}
(2.5)\qquad P\big(t_0, \{t_i\}_{i=1}^N; K\big) &= \sum_{i=1}^{N} c_i Z(t_0, t_i),\\
c_i &\equiv K\delta t, \qquad i = 1, \ldots, N-1\\
c_N &\equiv 1 + K\delta t.
\end{aligned}
$$

In other words, $P(t_0, \{t_i\}_{i=1}^N; K)$ is the time t_0 price of a coupon bond with cash flows $\{c_i\}_{i=1}^N$ at $\{t_i\}_{i=1}^N$.[7] Since $P(t_0, \{t_i\}_{i=1}^N; K)$ depends on the state vector at time t_0, we shall also denote it simply as $P(x_{t_0})$.

Using the T-forward measure (with Radon–Nikodym derivative $d\mathbb{P}^T/d\mathbb{P} = \exp(-\int_t^T r_s ds)/Z(t, T)$), it is straightforward to show that (2.4) can be transformed to

$$(2.6)\qquad S^{pay}(t; K) = Z(t, t_0)\mathrm{E}_t^{t_0}[\mathrm{I}_{P(x_{t_0})<1}] - \sum_{i=1}^{N} c_i Z(t, t_i)\mathrm{E}_t^{t_i}[\mathrm{I}_{P(x_{t_0})<1}],$$

where E_t^T denotes the time t conditional expectation under the T-forward measure, and $\mathrm{I}_{(\cdot)}$ is the indicator function. By the definition of the indicator function, (2.6) can be also written as

$$(2.7)\qquad S^{pay}(t; K) = Z(t, t_0)\mathbb{P}_t^{t_0}[P(x_{t_0}) < 1] - \sum_{i=1}^{N} c_i Z(t, t_i)\mathbb{P}_t^{t_i}[P(x_{t_0}) < 1],$$

where $\mathbb{P}_t^T(A)$ denotes the time t conditional probability in the T-forward measure of the event A happening.

For most multifactor models, the probability $\mathbb{P}_t^T[P(x_{t_0}) < 1]$ cannot be computed exactly, and some approximation is necessary. Equation (2.7) is the starting point of the approximation of SU (2002) as well as Collin-Dufresne and Goldstein (2002). SP (2006) on the other hand starts from an equivalent formulation

$$(2.8)\qquad S^{pay}(t; K) = \mathrm{E}_t^{\tilde{Q}(t_0, \{t_i\}_{i=1}^N)}\big[\max\big[V^{pay}(t_0, \{t_i\}_{i=1}^N; K), 0\big]\big],$$

where the superscript $\tilde{Q}(t_0, \{t_i\}_{i=1}^N)$ denotes the swap-measure; their approach tries to find a tractable approximate dynamics of the bond price under the swap measure.[8]

SU's (2002) approach is based on their observation that in the case of affine models the exercise boundary $P(x_{t_0}) - 1 = 0$ can be well approximated by a simpler boundary

$$(2.9)\qquad b \cdot x_{t_0} - a = 0,$$

therefore $\mathbb{P}_t^T[P(x_{t_0}) < 1]$ in (2.7) is approximated as $\mathbb{P}_t^T[b \cdot x_{t_0} < a]$, which can be computed via Fourier inversion.

The specific procedure SU (2002) gave for obtaining a and b is somewhat complicated. In their illustration with a 2-factor CIR model, they use $f(x_{t_0,2} \mid x_t)$, the conditional probability density function (pdf) of the 2nd element of the state vector, to find an

[7] As equation (2.5) makes it clear, a payer swaption is a *put option* on a coupon bond. The *call option* version is called a receiver swaption. Since call option pricing and put option pricing are analogous, we shall focus on payer swaptions in this paper.

[8] See also Heidari, Hirsa, and Madan (2007) for an application of this "approximate dynamics" approach.

appropriate lower bound $x_{t_0,2}^{\min}$ and upper bound $x_{t_0,2}^{\max}$ (between which the pdf has most of the weight), and then the equation $P(x_{t_0}) - 1$ is solved to find the corresponding $x_{t_0,1}$ elements. The straight line connecting $(x_{t_0,1}^{\min}, x_{t_0,2}^{\min})$ and $(x_{t_0,1}^{\max}, x_{t_0,2}^{\max})$ gives the coefficients a and b. For a model with more than 2-factors, SU remark that a similar procedure can be done with least squares fitting.

As SP (2006) note, SU's method for obtaining the exercise boundary uses information about the pdf of at least one element of the state vector x_{t_0}. This is available for the 2-factor CIR model that SU used for illustration, but not all affine models have a closed-form pdf. For example, an $A_2(2)$ model (according to Dai and Singleton's 2000 classification) with nonzero off-diagonal feedback matrix does not have a closed-form expression for the pdf. Furthermore, it is not clear how to carry over SU's method for obtaining the approximate boundary to the case of swaption pricing in other models like the QG model.

2.2. New Boundary Approximations

In this paper, we propose two simpler methods for exercise boundary approximation, which can be also straightforwardly extended to models beyond the affine class.

Approximation A (Taylor series approximation). Note that $P(x_{t_0})$ can be written as

$$P(x_{t_0}) = \left(\sum_{i=1}^{N} c_i\right) \sum_{i=1}^{N} w_i e^{-(t_i - t_0)Y(x_{t_0}, t_i - t_0)}, \tag{2.10}$$

where the weights w_i's are defined as $w_i = c_i/(\sum_{j=1}^{N} c_j)$, and $Y(x_{t_0}, t_i - t_0)$ is the time t_0 value of the yield on a zero-coupon bond with time-to-maturity of $t_i - t_0$. Thus,

$$\begin{aligned} \log P(x_{t_0}) &= const. + \log\left(\sum_{i=1}^{N} w_i e^{-(t_i - t_0)Y(x_{t_0}, t_i - t_0)}\right) \\ &= const. - \sum_{i=1}^{N} (t_i - t_0) w_i Y(x_{t_0}, t_i - t_0) + \text{convexity correction}. \end{aligned} \tag{2.11}$$

In the case of affine models, $Y(x_{t_0}, t_i - t_0)$ is an affine function of x_{t_0} for all $t_i - t_0$, thus $\log P(x_{t_0})$ is an approximately affine function. Therefore, first-order Taylor expansion of $\log P(x_{t_0})$ around a point on the true boundary, say x^*, can provide a good approximation. In other words, we can approximate the boundary condition $P(x_{t_0}) = 1$ (i.e., $\log P(x_{t_0}) = 0$) with

$$\partial_x \log P(x^*) \cdot (x_{t_0} - x^*) = 0 \tag{2.12}$$

(since $\log P(x^*) = 0$). This can be further simplified as

$$\partial_x P(x^*) \cdot (x_{t_0} - x^*) = 0, \tag{2.13}$$

in other words, we have

$$a = \partial_x P(x^*) \cdot x^*, \qquad b = \partial_x P(x^*) \tag{2.14}$$

in (2.9).

On the other hand, in the case of QG models, all zero-coupon yields are quadratic functions of the state vector, therefore, $\log P(x_{t_0})$ is an approximately quadratic function of x_{t_0}. Therefore, we can take the Taylor expansion of $\log P(x_{t_0})$ up to the second order:

$$\partial_x \log P(x^*) \cdot (x_{t_0} - x^*) + \frac{1}{2}(x_{t_0} - x^*)' \partial_x \partial_{x'} \log P(x^*)(x_{t_0} - x^*) = 0, \tag{2.15}$$

which can be simplified as

$$\partial_x P(x^*) \cdot (x_{t_0} - x^*) + \frac{1}{2}(x_{t_0} - x^*)' \left(\partial_x \partial_{x'} P(x^*) - \frac{\partial_x P(x^*) \partial_{x'} P(x^*)}{P(x^*)} \right) (x_{t_0} - x^*) = 0. \tag{2.16}$$

In other words, we can approximate the condition $P(x_{t_0}) - 1 = 0$ with

$$\mathrm{a} + \mathrm{b} \cdot x_{t_0} + x'_{t_0} \mathrm{C} x_{t_0} = 0, \tag{2.17}$$

where

$$\begin{aligned} \mathrm{a} &= -\partial_x P(x^*) \cdot x^* + \frac{1}{2} x^{*'} (\partial_x \partial_{x'} P(x^*) - P(x^*)^{-1} \partial_x P(x^*)(\partial_x P(x^*))') x^* \\ \mathrm{b} &= \partial_x P(x^*) - (\partial_x \partial_{x'} P(x^*) - P(x^*)^{-1} \partial_x P(x^*)(\partial_x P(x^*))') x^* \\ \mathrm{C} &= \frac{1}{2}(\partial_x \partial_{x'} P(x^*) - P(x^*)^{-1} \partial_x P(x^*)(\partial_x P(x^*))'). \end{aligned} \tag{2.18}$$

Note that the boundary of the form (2.17) leads to a tractable Fourier inversion computation of $\mathbb{P}_t^T[\cdot]$ for QG models, i.e., $\mathbb{P}_t^T[P(x_{t_0}) < 1] \approx \mathbb{P}_t^T[\mathrm{a} + \mathrm{b} \cdot x_{t_0} + x'_{t_o} \mathrm{C} x_{t_0} < 0]$.

A good choice of x^* for $\mathbb{P}_t^T[\cdot]$ would be the point on the line (surface) $P(x_{t_0}) - 1 = 0$ at which the value of the pdf $f^T(x_{t_0} \mid x_t)$ is the largest ($f^T(x_{t_0} \mid x_t)$ denotes the conditional probability density function of x_{t_0} in the T-forward measure). This is schematically illustrated in Figure 2.1 for a 2-factor affine model case. The ellipse-like objects denote the locus of x_1 and x_2 values which have the same pdf value; the smaller the object, the higher the probability density it represents. The solid line represents the boundary equation $P(x) = 1$. The * symbol on this line is x^*; this is the tangent point of an equi-pdf locus and the line $P(x) - 1 = 0$. The nice thing about the Taylor expansion around this x^* is that the approximation is very accurate in the part of the line $P(x) - 1 = 0$ where it matters most (in the sense that the density is highest).

In practice, except for special cases, the pdf $f^T(x_{t_0} \mid x_t)$ is not available in closed-form. To overcome this problem, we can define x^* as

$$x^* = \operatorname*{argmin}_x \left\{ \frac{1}{2}(x - \mathrm{E}^T[x_{t_0} \mid x_t])' (\mathrm{Var}^T[x_{t_0} \mid x_t])^{-1} (x - \mathrm{E}^T[x_{t_0} \mid x_t]) : P(x) - 1 = 0 \right\}. \tag{2.19}$$

In other words, we approximate $f^T(x_{t_0} \mid x_t)$ by a multivariate normal distribution with mean $\mathrm{E}^T[x_{t_0} \mid x_t]$ and variance–covariance matrix $\mathrm{Var}^T[x_{t_0} \mid x_t]$.

We can further simplify the procedure by approximating the T-forward measure moments with risk-neutral moments, i.e., $\mathrm{E}^T[x_{t_0} \mid x_t] \approx \mathrm{E}[x_{t_0} \mid x_t]$ and $\mathrm{Var}^T(x_{t_0} \mid x_t) \approx \mathrm{Var}(x_{t_0} \mid x_t)$. Thus we have

$$x^* = \operatorname*{argmin}_x \left\{ \frac{1}{2}(x - \mathrm{E}[x_{t_0} \mid x_t])' (\mathrm{Var}[x_{t_0} \mid x_t])^{-1} (x - \mathrm{E}[x_{t_0} \mid x_t]) : P(x) - 1 = 0 \right\}. \tag{2.20}$$

Note that for models like the affine class and QG class, $\mathrm{E}[x_{t_0} \mid x_t]$ and $\mathrm{Var}[x_{t_0} \mid x_t]$ are available in closed-form, whereas we would need to solve a system of ODEs (Riccati

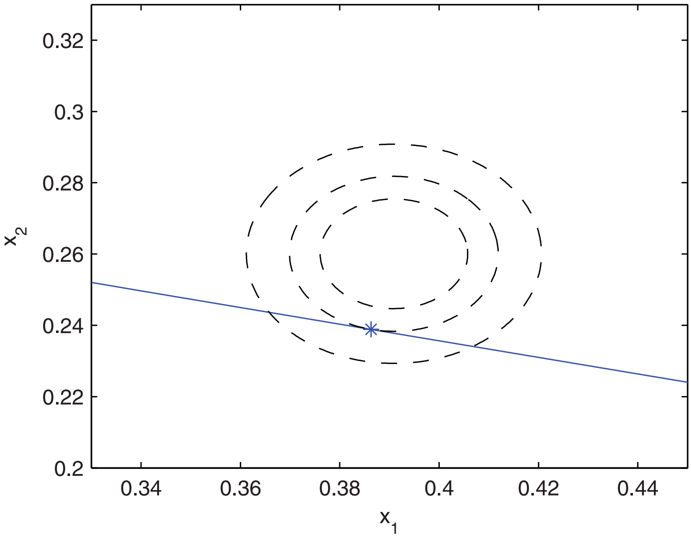

FIGURE 2.1. Schematic illustration of the x^* vector, the equi-pdf loci, and the affine model boundary.

system) to obtain $\mathrm{E}^T[x_{t_0} \mid x_t]$ and $\mathrm{Var}^T[x_{t_0} \mid x_t]$. Furthermore, with the formulation (2.20), the exercise boundary does not depend on T, therefore the same boundary can be used for evaluating all probabilities $\mathbb{P}_t^{t_i}[\cdot]\,(i = 0, 1, \ldots, N)$ in (2.7). In this paper, we shall thus use (2.20) to obtain x^*. To handle the constrained optimization in (2.20), one can solve the nonlinear equation

$$(2.21) \qquad \begin{bmatrix} x - \mathrm{E}[x_{t_0} \mid x_t] + \eta\,\mathrm{Var}[x_{t_0} \mid x_t]\partial_x P(x) \\ P(x) - 1 \end{bmatrix} = 0_{(n+1)\times 1},$$

where η is a Lagrange multiplier. This nonlinear system for (x, η) can be solved quickly with Newton's method.[9]

Approximation B (zero-coupon bond approximation). Another way to approximate the boundary is to use the boundary implied by a zero-coupon bond of a suitable maturity. The idea is to find at time t a zero-coupon bond that has risk characteristics similar to the coupon bond

$$(2.22) \qquad P\big(t, \{t_i\}_{i=1}^N; K\big) = \sum_{i=1}^N c_i\, Z(t, t_i),$$

where c_i's were given in (2.5). (Note that at time t_0, the price of this coupon bond becomes $P(t_0, \{t_i\}_{i=1}^N; K)$, the object considered in the above Taylor series approximation.)

An obvious choice for the zero-coupon bond maturity would be the stochastic duration of $P(t, \{t_i\}_{i=1}^N; K)$. The stochastic duration $D_s(t)$ is defined as the zero-coupon bond maturity that leads to the same relative volatility as the coupon bond, i.e., $D_s(t)$ is

[9] Newton's method with analytical Jacobian and with $(x = \mathrm{E}[x_{t_0} \mid x_t], \eta = 0)$ as the starting guess was found to converge to the solution within 5∼ 6 iterations.

implicitly given by the equation

$$\left(\frac{dZ(t, t + D_s(t))}{Z(t, t + D_s(t))}\right)^2 = \left(\frac{dP\big(t, \{t_i\}_{i=1}^N; K\big)}{P\big(t, \{t_i\}_{i=1}^N; K\big)}\right)^2. \tag{2.23}$$

Though the stochastic duration has a sounder theoretical basis, one can also consider other possibilities which are simpler. These include the Fisher–Weil duration

$$D_{FW}(t) = \left(\sum_{t=1}^{N}(t_i - t)c_i Z(t, t_i)\right) \bigg/ P\big(t, \{t_i\}_{i=1}^N; K\big), \tag{2.24}$$

and Macaulay's duration

$$D_M(t) = \left(\sum_{t=1}^{N}(t_i - t)c_i e^{-(t_i - t)y}\right) \bigg/ P\big(t, \{t_i\}_{i=1}^N; K\big), \tag{2.25}$$

where y is the solution of the equation $P(t, \{t_i\}_{i=1}^N; K) = \sum_{i=1}^N c_i e^{-(t_i-t)y}$.

With one of these duration measures (D_s, D_{FW}, D_M), we can approximate the time t_0 price of the coupon bond as

$$P\big(t_0, \{t_i\}_{i=1}^N; K\big) \approx \xi\, Z(t_0, t + D(t)), \tag{2.26}$$

where $\xi \equiv P(t, \{t_i\}_{i=1}^N; K)/Z(t, t + D(t))$. Therefore, the boundary equation $P(x_{t_0}) - 1 = 0$ can be approximated as $\xi Z(t_0, t + D(t)) - 1 = 0$, which is equivalent to

$$\log Z(t_0, t + D(t)) + \log(\xi) = 0. \tag{2.27}$$

Note that $\log Z(t_0, t + D(t))$ is an affine function of x_{t_0} in the case of affine models and a quadratic function of x_{t_0} in the case of QG models. Therefore, the "zero-coupon bond approximation" (2.27) gives the desired form of the boundary equation ((2.9) for affine models and (2.17) for QG models). In summary, we can approximate $\mathbb{P}_t^T[P(x_{t_0}) < 1]$ as $\mathbb{P}_t^T[\xi\, Z(t_0, t + D(t)) < 1]$.

2.3. The T-dependence of $\mathbb{P}^T$

To compute the swaption price in (2.7), one needs to evaluate the T-forward measure probability $\mathbb{P}_t^T[\cdot]$ at $T = t_0, t_1, \ldots, t_N$. For a swaption with a long tenor, this can entail quite a lot of evaluations ($1 + (t_N - t_0)/\delta t$ evaluations). These probabilities are typically evaluated with numerical Fourier inversion, which can be time-consuming.

One can however avoid performing the $1 + (t_N - t_0)/\delta t$ Fourier inversions. As we shall argue below, $\mathbb{P}_t^T[\cdot]$ is a smooth, monotonically decreasing function of T. Therefore, computing $\mathbb{P}_t^T[\cdot]$ at $5 \sim 6$ values of T in the interval $[t_0, t_N]$ would be sufficient to get an accurate interpolation formula for $\mathbb{P}_t^T[\cdot]$, from which all the necessary probabilities can be evaluated.

For concreteness, let us focus on Approximation B (zero-coupon bond approximation). In this case we are dealing with $\mathbb{P}_t^T[\xi\, Z(t_0, t + D(t)) < 1]$, which can be also written $\mathbb{P}_t^T[Y(x_{t_0}, \Delta) > k]$, with $\Delta = t + D(t) - t_0$ and $k = \log(\xi)/(t + D(t) - t_0)$; recall that $Y(x_{t_0}, \tau)$ is the zero-coupon yield at time t_0 with time-to-maturity τ. We can make a heuristic argument that $(\partial/\partial T)\mathbb{P}_t^T[Y(x_{t_0}, \Delta) > k] < 0$ for all T, as follows:

Note that $(\partial/\partial T)\mathbb{P}_t^T[Y(x_{t_0}, \Delta) > k]$ can be expressed as

(2.28)
$$
\begin{aligned}
\frac{\partial}{\partial T}\mathbb{P}_t^T[Y(x_{t_0}, \Delta) > k] &= \frac{\partial}{\partial T}\big(Z(t, T)^{-1}\mathrm{E}_t[e^{-\int_t^T r_s ds}\mathrm{I}_{Y(x_{t_0},\Delta)>k}]\big) \\
&= \frac{\mathrm{E}_t\big[e^{-\int_t^T r_s ds} r_T\big]\mathrm{E}_t\big[e^{-\int_t^T r_s ds}\mathrm{I}_{Y(x_{t_0},\Delta)>k}\big]}{Z(t, T)^2} - \frac{\mathrm{E}_t\big[e^{-\int_t^T r_s ds} r_T\,\mathrm{I}_{Y(x_{t_0},\Delta)>k}\big]}{Z(t, T)} \\
&= \mathrm{E}_t^T[r_T]\mathrm{E}_t^T[\mathrm{I}_{Y(x_{t_0},\Delta)>k}] - \mathrm{E}_t^T[r_T\,\mathrm{I}_{Y(x_{t_0},\Delta)>k}] \\
&= -\mathrm{cov}_t^T(r_T, \mathrm{I}_{Y(x_{t_0},\Delta)>k}).
\end{aligned}
$$

In general we would expect r_T and $Y(x_{t_0}, \Delta)$ to be positively correlated, thus $-\mathrm{cov}_t^T(r_T, \mathrm{I}_{Y(x_{t_0},\Delta)>k}) < 0$ as well, i.e., $\mathbb{P}_t^T[Y(x_{t_0}, \Delta) > k]$ is expected to be a decreasing function of T for all T. In view of the simple behavior of the function $\mathbb{P}_t^T[Y(x_{t_0}, \Delta) > k]$, its cubic-spline interpolation based on its values at a small number of T values will give accurate swaption prices. We also expect that the interpolation approach will work similarly well in the case of Approximation A (Taylor series approximation).

2.4. Comparison with Munk (1999)

Our Approximation B with $D_s(t)$ (stochastic duration) as $D(t)$ may appear similar to Munk's (1999) stochastic duration method. It is thus instructive to compare the two. Munk's approach completely equates one unit of coupon bond $P(t, \{t_i\}_{i=1}^N; K)$ with ξ units of zero coupon bond $Z(t, t + D_s(t))$. Thus, Munk's (1999) formula for the payer swaption is given by

$$
\begin{aligned}
S^{pay}(t; K) &= \xi\mathrm{E}_t\big[e^{-\int_t^{t_0} r_s ds}\max[\xi^{-1} - Z(t_0, t + D_s(t)), 0]\big] \\
&= Z(t, t_0)\,\mathbb{P}_t^{t_0}[\xi Z(t_0, t + D_s(t)) < 1] \\
&\quad - P\big(t, \{t_i\}_{i=1}^N; K\big)\,\mathbb{P}_t^{t+D_s(t)}[\xi Z(t_0, t + D_s(t)) < 1].
\end{aligned} \tag{2.29}
$$

On the other hand, our Approximation B uses stochastic duration (or another duration measure) only for approximating the boundary. This leads to the payer swaption price formula

$$
\begin{aligned}
S^{pay}(t; K) = Z(t, t_0)\,\mathbb{P}_t^{t_0}[\xi Z(t_0, t + D_s(t)) < 1] \\
- \sum_{i=1}^N c_i Z(t, t_i)\,\mathbb{P}_t^{t_i}[\xi Z(t_0, t + D_s(t)) < 1].
\end{aligned} \tag{2.30}
$$

It may not be clear that (2.30), which can have a lot of terms in the summation, gives a more accurate result than (2.29). But note that (2.30) can be also written as

$$
S^{pay}(t; K) = \mathrm{E}_t\big[e^{-\int_t^{t_0} r(x_s)ds}(1 - P(x_{t_0}))\mathrm{I}_{\xi Z(t_0,t+D_s(t))<1}\big], \tag{2.31}
$$

which differs from the "true" expression just by the argument in the indicator function (the true expression has $\mathrm{I}_{P(x_{t_0})<1}$ instead). Thus, the error in the approximate formula (2.30) is simply

$$
\mathrm{E}_t\big[e^{-\int_t^{t_0} r(x_s)ds}(1 - P(x_{t_0}))(\mathrm{I}_{\xi Z(t_0,t+D_s(t))<1} - \mathrm{I}_{P(x_{t_0})<1})\big]. \tag{2.32}
$$

The only nonzero contribution to this expression is from the part where $\mathrm{I}_{\xi Z(t_0,t+D_s(t))<1} \neq \mathrm{I}_{P(x_{t_0})<1}$. Geometrically, this corresponds to the region between the *true boundary* and the *approximate boundary*. But in this region, the factor $e^{-\int_t^{t_0} r_s ds}(1-P(x_{t_0}))$ must be fairly small, since $1-P(x_{t_0})=0$ on the true boundary, so the expression (2.32) can be quite small. Thus, even if the individual probabilities $\mathbb{P}_t^{t_i}$ were not very accurate (i.e., even if the errors $\mathbb{P}_t^{t_i}[\xi Z(t_0, t+D_s(t)) < 1] - \mathbb{P}_t^{t_i}[P(x_{t_0}) < 1]$ were not small in magnitude), equation (2.30) could still give an accurate value of the swaption, i.e., the errors in $\mathbb{P}_t^{t_i}$ cancel out to a large degree. The same logic applies to Approximation A (Taylor series approximation), and we can expect it to produce very accurate swaption prices.

Note also that (2.30) can be expressed as

$$
\begin{aligned}
S^{pay}(t;K) = {} & Z(t,t_0)\,\mathbb{P}_t^{t_0}[\xi Z(t_0, t+D_s(t)) < 1] \\
& - P\big(t, \{t_i\}_{i=1}^N; K\big) \sum_{i=1}^{N} \omega_i \mathbb{P}_t^{t_i}[\xi Z(t_0, t+D_s(t)) < 1],
\end{aligned} \tag{2.33}
$$

where $\omega_i = c_i Z(t, t_i)/(\sum_j c_j Z(t, t_j))$. This takes the same form as (2.29), except that $\mathbb{P}_t^{t+D_s(t)}[\xi Z(t_0, t+D_s(t)) < 1]$ is replaced by $\sum_{i=1}^N \omega_i \mathbb{P}_t^{t_i}[\xi Z(t_0, t+D_s(t)) < 1]$. In other words, Munk (1999) approximates the weighted *average of functions* ($\sum_i \omega_i \mathbb{P}^{t_i}$) as a function *of an average* ($\mathbb{P}^{t+D_s(t)}$).[10] Therefore, the difference between (2.29) and (2.30) can be viewed as a convexity effect. For out-of-money options, $\mathbb{P}_t^T$'s will be small (hence close to the zero-bound), thus the convexity feature in $\mathbb{P}_t^T$ (as a function of T) will be more pronounced. Therefore the (relative) discrepancy between Munk's (1999) method and Approximation B with stochastic duration is likely to be the largest in out-of-the-money options.

3. SWAPTION PRICING IN AFFINE MODELS

3.1. General Affine Models

In the general version of affine models, due to Duffie and Kan (1996), we have

$$
r(x_t) = a_r + b_r \cdot x_t, \tag{3.1}
$$

where a_r is a constant and b_r is an n-dimensional constant vector. The n-dimensional vector of state variables x_t has the risk-neutral dynamics given by[11]

$$
dx_t = \mathcal{K}(\theta - x_t)\,dt + \Sigma(x_t)\,dW_t, \tag{3.2}
$$

where $\mathcal{K}$ is $n \times n$ constant matrix, θ is a constant n-vector, W_t is an n-dimensional vector of standard Brownian motions, and $\Sigma(x_t)$ is an $n \times n$ matrix function of x_t such that

$$
(\Sigma(x_t)\Sigma(x_t)')_{ij} = h_{ij} + H_{ij}' x_t, \tag{3.3}
$$

where h_{ij} is a constant (scalar), and H_{ij} is a constant n-vector.

[10] Munk's (1999) approximation can be thus viewed as consisting of two approximations: approximating the exercise boundary, and approximating an average of functions as a function of an average. Note that when $D(t) = D_s(t)$, the two average concepts are different. But if the Fisher–Weil duration is used (i.e., $D(t) = D_{FW}(t)$), the two average concepts are the same, since $D_{FW}(t) = \sum_i \omega_i (t_i - t)$ (where ω_i is the same as the ω_i in (2.33)).

[11] Since the P-measure (physical measure) dynamics is not needed for pricing, in this paper we shall specify only the Q-measure (risk-neutral) measure dynamics. Therefore, all the model parameters are risk-neutral parameters, unless otherwise stated.

The conditional expectation and the variance–covariance matrix of the state vector (which are used in Approximation A) are available in closed-form in affine models. The Appendix provides an expression for $\mathrm{E}[x_T|x_t]$ and $\mathrm{Var}[x_T|x_t]$ ((A.11) and (A.16), or (A.17) and (A.20)) that is valid for all affine models (including the case of nondiagonal $\mathcal{K}$ matrix and the case in which $\mathcal{K}$ has complex eigenvalues).

It is well known that the zero-coupon bond price $Z(t, T)$ in this model is given by[12]

$$Z(t, T) = \exp(A(T-t) + B(T-t) \cdot x_t), \tag{3.4}$$

where $A(\tau)$, $B(\tau)$ are solutions of

$$\begin{aligned} \frac{dA(\tau)}{d\tau} &= -a_r - (\mathcal{K}\theta)' B(\tau) + \frac{1}{2}\sum_{i,j} B_i(\tau) h_{ij} B_j(\tau) \\ \frac{dB(\tau)}{d\tau} &= -b_r + \mathcal{K}' B(\tau) + \frac{1}{2}\sum_{i,j} B_i(\tau) H_{ij} B_j(\tau) \end{aligned} \tag{3.5}$$

with the boundary conditions $A(0) = 0$ and $B(0) = 0_{n\times 1}$.

Both in the case of the Taylor series approximation and the zero-coupon bond approximation, we need to evaluate the expression $\mathbb{P}_t^T[b \cdot x_{t_0} < a]$. In the case of Approximation A, the coefficients a, b are given by (2.14), which involves the functions $P(x)$, $\partial_x P(x)$, given by

$$\begin{aligned} P(x) &= \sum_{i=1}^{N} c_i\, e^{A(t_i - t_0) + B(t_i - t_0)' x}, \\ \partial_x P(x) &= \sum_{i=1}^{N} c_i\, B(t_i - t_0)\, e^{A(t_i - t_0) + B(t_i - t_0)' x}. \end{aligned} \tag{3.6}$$

In the case of Approximation B, we have, from (2.27),

$$\begin{aligned} a &= -A(t + D(t) - t_0) - \log(\xi) \\ b &= B(t + D(t) - t_0). \end{aligned} \tag{3.7}$$

The stochastic duration $D_s(t)$ for affine models is given by the solution of the nonlinear equation

$$B(D_s)\Sigma(x_t)\Sigma(x_t)' B(D_s)' = \left(\sum_{i=1}^{N} \omega_i B(t_i - t)\right)' \Sigma(x_t)\Sigma(x_t)' \left(\sum_{i=1}^{N} \omega_i B(t_i - t)\right), \tag{3.8}$$

where $\omega_i = c_i Z(t, t_i)/(\sum_j c_j Z(t, t_j))$. For general affine models (which do not have closed-form expression for B), the nonlinear equation (3.8) for D_s can be time-consuming to solve. One can speed up the computation by using an interpolation to compute the left-hand side based on precomputed values of $B(\tau)\Sigma(x_t)\Sigma(x_t)' B(\tau)'$ at $\tau = \{t_i - t\}_{i=1}^{N}$.

The T-forward measure probability $\mathbb{P}_t^T[b \cdot x_{t_0} < a]$ can be expressed as an inverse Fourier transform of the T-forward measure characteristic function of $b \cdot x_{t_0}$, which we

[12] See, e.g., Duffie and Kan (1996) and Duffie, Pan, and Singleton (2000).

shall refer to as $f_t^T(z)$. Using the Levy inversion formula we have

$$\mathbb{P}_t^T[b \cdot x_{t_0} < a] = \frac{1}{2} - \frac{1}{\pi}\int_0^\infty dz \frac{\mathrm{Im}(f_t^T(z)e^{-iza})}{z}. \tag{3.9}$$

Note that the characteristic function $f_t^T(z)$ is given by

$$\begin{aligned} f_t^T(z) &= \mathrm{E}_t^T\big[\exp(izb \cdot x_{t_0})\big] = Z(t,T)^{-1}\mathrm{E}_t\big[e^{-\int_t^T r_s ds} e^{izb \cdot x_{t_0}}\big] \\ &= Z(t,T)^{-1}\mathrm{E}_t\big[e^{-\int_t^{t_0} r_s ds} e^{A(T-t_0)+(izb+B(T-t_0))\cdot x_{t_0}}\big]. \end{aligned} \tag{3.10}$$

It is well known that the expectation in (3.10) is an exponential affine function of x_t (see, e.g., Duffie et al. 2000). In the end, we have

$$f_t^T(z) = \exp(\tilde{A}(t_0 - t; T - t_0, z) - A(T-t) + (\tilde{B}(t_0 - t; T - t_0, z) - B(T-t)) \cdot x_t), \tag{3.11}$$

where $\tilde{A}(t_0 - t; T - t, z)$ and $\tilde{B}(t_0 - t; T - t, z)$ are solutions of the same Riccati equation (3.5), now with boundary conditions $\tilde{A}(0; T - t_0, z) = A(T - t_0)$ and $\tilde{B}(0; T - t_0, z) = izb + B(T - t_0)$.

3.2. Two-Factor CIR Model Illustration

To illustrate the application of the method of Section 2 to affine models, let us use the 2-factor CIR model, as in SU (2002), Collin-Dufresne and Goldstein (2002), and SP (2006). Note that in multifactor CIR models the bond price (3.5) and the characteristic function (3.10) have closed-form expressions, given, for example, in Collin-Dufresne and Goldstein (2002).

To test the performance of the method in a realistic setting, we shall use the parameters obtained from an actual econometric estimation, specifically that of Duffie and Singleton (1997). These parameters are given by

$$a_r = -0.58,\ b_r = [1, 1]',\ \mathcal{K} = \begin{bmatrix} 0.5080 & 0 \\ 0 & -0.0010 \end{bmatrix},\ \theta = \begin{bmatrix} 0.4005 \\ -0.7740 \end{bmatrix} \tag{3.12}$$

$$h_{ij} = 0\ (\forall i, j),\ H_{11} = \begin{bmatrix} 0.023^2 \\ 0 \end{bmatrix},\ H_{22} = \begin{bmatrix} 0 \\ 0.019^2 \end{bmatrix},\ H_{12} = H_{21} = 0_{2\times 1}.$$

Note that $[\mathcal{K}]_{22}$ is negative, which implies a nonstationary (explosive) risk-neutral dynamics; hence the risk-neutral unconditional mean of x_t does not exist. We shall take $x_t = [0.374, 0.258]'$, which is the estimated *physical* unconditional mean.

Table 3.1 shows the swaption prices for tenors of 2, 5, 10 years and option maturities of 1, 2, 5 years. The strike rate K is chosen as (ATMF-1.5%, ATMF-0.75%, ATMF, ATMF+0.75%, ATMF+1.5%), where ATMF is the at-the-money forward swap rate (equation (2.2)). The interval δt is set to $1/2$.

We compare the results based on five methods: Munk's (1991) stochastic duration method ("Munk"), Approximation A ("A"), Approximation B with stochastic duration ("B-sd"), Approximation B with Fisher–Weil duration ("B-FW"),[13] and the Monte Carlo

[13] We have also examined the use of Macaulay duration. The results were similar to the Fisher–Weil duration, hence omitted for brevity.

TABLE 3.1

Upper Panel, Middle Panel, and Lower Panel Give the Swaption Prices (in basis points) for 1-Year, 2-Year, and 5-Year Option Maturities, Respectively, in a 2-Factor CIR Model

Method	Tenor	ATMF-1.5%	ATMF-0.75%	ATMF	ATMF+0.75%	ATMF+1.5%
MC	2	271.35 (0.03)	165.62 (0.05)	85.84 (0.06)	36.50 (0.05)	12.38 (0.03)
	5	598.05 (0.04)	352.29 (0.08)	169.27 (0.12)	63.15 (0.09)	17.59 (0.05)
	10	991.68 (0.04)	574.54 (0.12)	265.51 (0.18)	92.33 (0.14)	23.09 (0.07)
A	2	271.31	165.61	85.87	36.53	12.40
	5	598.02	352.31	169.42	63.24	17.66
	10	991.66	574.57	265.78	92.45	23.18
B-sd	2	271.31	165.60	85.87	36.52	12.40
	5	598.01	352.28	169.38	63.20	17.64
	10	991.62	574.48	265.65	92.36	23.13
B-FW	2	271.31	165.60	85.87	36.52	12.40
	5	598.01	352.29	169.39	63.21	17.63
	10	991.63	574.48	265.67	92.37	23.12
Munk	2	271.32	165.64	85.93	36.59	12.46
	5	597.97	352.36	169.72	63.68	18.00
	10	990.91	573.80	266.17	93.93	24.46
MC	2	267.33 (0.04)	177.16 (0.06)	106.68 (0.08)	57.64 (0.07)	27.64 (0.05)
	5	582.96 (0.07)	376.01 (0.12)	216.48 (0.15)	109.47 (0.13)	47.98 (0.09)
	10	960.77 (0.09)	611.74 (0.16)	344.45 (0.22)	168.40 (0.20)	70.48 (0.14)
A	2	267.38	177.30	106.89	57.82	27.77
	5	583.02	376.23	216.88	109.78	48.17
	10	960.79	611.98	344.95	168.78	70.69
B-sd	2	267.38	177.30	106.89	57.82	27.77
	5	583.00	376.20	216.83	109.74	48.14
	10	960.73	611.87	344.81	168.68	70.63
B-FW	2	267.38	177.30	106.89	57.82	27.77
	5	583.00	376.20	216.84	109.74	48.14
	10	960.73	611.87	344.83	168.70	70.63
Munk	2	267.43	177.39	107.02	57.97	27.90
	5	583.00	376.46	217.46	110.63	49.03
	10	959.17	611.01	345.75	171.29	73.71
MC	2	240.14 (0.05)	176.38 (0.07)	123.87 (0.08)	82.89 (0.08)	52.71 (0.07)
	5	521.94 (0.09)	378.55 (0.13)	261.40 (0.16)	171.21 (0.16)	106.06 (0.14)
	10	857.04 (0.11)	617.66 (0.18)	422.88 (0.23)	273.96 (0.24)	167.47 (0.21)
A	2	240.13	176.35	123.83	82.86	52.69
	5	521.97	378.53	261.36	171.18	106.05
	10	857.12	617.67	422.85	273.93	167.45
B-sd	2	240.13	176.35	123.83	82.86	52.69
	5	521.95	378.51	261.33	171.15	106.02
	10	857.04	617.55	422.71	273.82	167.39
B-FW	2	240.13	176.35	123.83	82.86	52.69
	5	521.95	378.51	261.33	171.15	106.02
	10	857.03	617.55	422.72	273.84	167.40
Munk	2	240.22	176.49	124.02	83.08	52.92
	5	521.93	378.87	262.16	172.40	107.54
	10	854.34	616.51	424.09	277.61	172.86

Note: The Monte Carlo errors are given in parentheses.

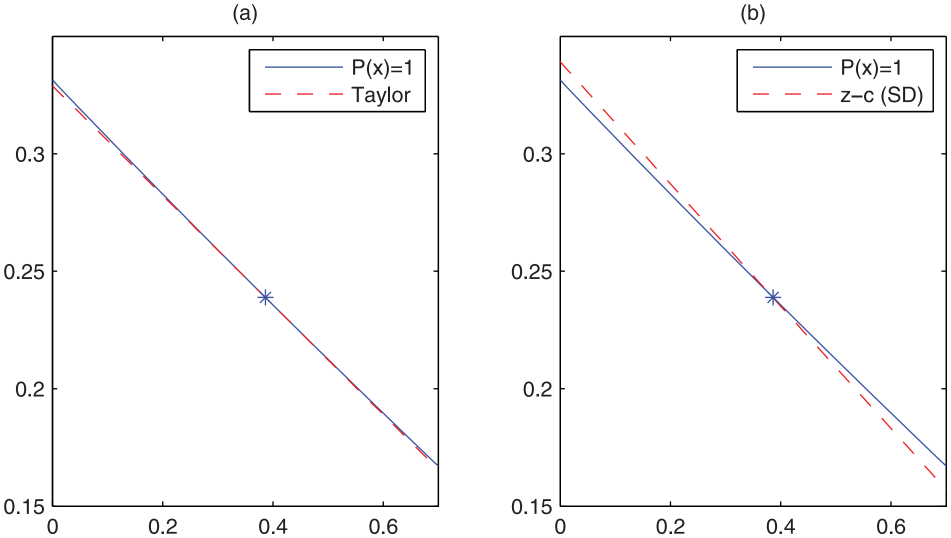

FIGURE 3.1. (a): The true boundary (solid line) and the boundary based on Approximation A (dashed line). The * symbol denotes the location of x^*. (b): The boundary based on Approximation B-sd (dashed line).

method ("MC"). Monte Carlo results are taken as a proxy for the true swaption price. The Monte Carlo values are computed using 1,000,000 simulations with standard variance reduction technique (antithetic variates), and dt is discretized as $1/250$.[14]

In Approximations A and B, for the 5-year and 10-year tenors we use the cubic-spline interpolation based on the evaluation of $\mathbb{P}_t^T$ at $T = t_0 + (j/m)(t_N - t_0)$, where $j = 0, 1, \ldots, m$ and $m = 4$ (as discussed in Section 2.3).

Both Approximation A (Taylor series approximation) and Approximation B (zero-coupon bond approximation) were found to give accurate swaption prices for the entire range of strikes considered. For example, for the 2-year (option-) maturity, 10-year-tenor swaption, Approximation A gives the price of 960.79 and 70.69 for ATMF-1.5% and ATMF+1.5%, respectively, while the MC values are 960.77 and 70.48. With Approximation B, both B-sd and B-FW gave similarly accurate results. These results (A, B-sd, B-FW) are notably more accurate than Munk's (1999) stochastic duration method, which gave 957.71 and 73.71.

We now take a closer look, by comparing the exercise boundaries implied by Approximation A and Approximation B with stochastic duration (B-sd). Figure 3.1(a) and (b) show the Approximation A boundary and the Approximation B-sd boundary, respectively, for 2-year-maturity, 10-year tenor, and strike rate $K = 6\%$. The true boundary is also plotted in both Figure 3.1(a) and (b) for comparison. Remarkably, the two boundaries (A boundary and B-sd boundary) differ significantly, despite the fact that both produce similarly accurate swaption prices as we have seen in Table 3.1: The A boundary is quite similar to the true boundary, while the B-sd boundary intersects the true boundary with a visible angle. Two things explain why the B-sd approximation still gives

[14] With this setting, the MC results had an excellent match with the exact results in the case of the 1-factor Vasicek model, which has a closed-form solution for swaption prices (see, e.g., Jamshidian 1989; Brigo and Mercurio 2001, p. 102).

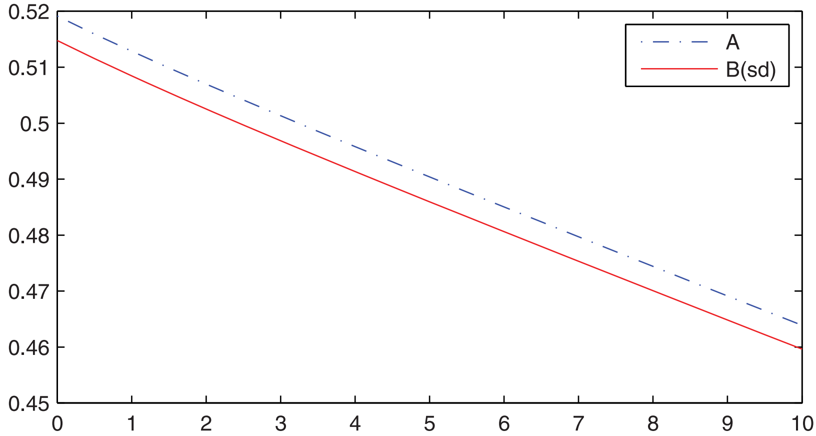

FIGURE 3.2. Plots of $\mathbb{P}_t^T(P(x_{t_0}) < 1)$ as a function of $T - t_0$, for Approximation A and Approximation B-sd.

accurate swaption prices: (1) it can be seen that the B-sd boundary intersects the true boundary at a point very close to x^*, therefore the most important part of the boundary is reasonably well approximated, and (2) more importantly, even if the individual probabilities $\mathbb{P}_t^{t_i}$ deviated significantly from the true values, the errors in $\mathbb{P}_t^{t_i}$ would cancel out to a large extent in the formula for the swaption price (equation (2.7)), as we have discussed in Section 2.4.

Figure 3.2 shows the probabilities $\mathbb{P}_t^T[P(x_{t_0}) < 1]$ as a function of $T - t_0$ in the case of the 2-year-maturity, 10-year-tenor ATMF swaption for both Approximation A and Approximation B-sd. The function $\mathbb{P}^T$ is indeed smooth and monotonically decreasing (as discussed in Section 2.3), but the gap between the $\mathbb{P}_t^T[\cdot]$ for A and the $\mathbb{P}_t^T[\cdot]$ for B-sd is also notable. The relative difference between the A and B-sd probabilities is about 1% (as can be seen from the figure), which is much larger than the relative difference between the A and B-sd swaption prices (344.95 vs. 344.81), indicating a substantial degree of cancellation of errors. This robustness (the fact that one can obtain an accurate swaption price with an inaccurate evaluation of probabilities) is an attractive feature of the exercise boundary approximation approach.

3.3. Comparison with Schrager and Pelsser (2006)

Here, we provide additional evidence on the quality of Approximations A and B applied to affine models. We use three parameter/state-vector settings examined by SP (2006): a 1-factor affine-Gaussian model (Vasicek model), a 3-factor affine-Gaussian model, and a 2-factor CIR model. In the notation of the present paper, they are given as follows.

One-factor affine-Gaussian model:

$$a_r = 0,\ b_r = 1,\ \mathcal{K} = 0.05,\ \theta = 0.05,\ h = 0.01^2,\ H = 0,\ x_t = 0.05. \tag{3.13}$$

Three-factor affine-Gaussian model:

$$a_r = 0.06,\ b_r = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},\ \mathcal{K} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.5 \end{bmatrix},\ \theta = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \tag{3.14}$$

$$h = 10^{-4} \begin{bmatrix} 1.00 & -0.10 & -0.02 \\ -0.10 & 0.25 & 0.03 \\ -0.02 & 0.03 & 0.04 \end{bmatrix},\ H_{ij} = 0_{2\times 1} (\forall i, j),\ x_t = \begin{bmatrix} 0.01 \\ 0.005 \\ -0.02 \end{bmatrix}.$$

Two-factor CIR model:

$$a_r = 0.02,\ b_r = \begin{bmatrix} 1 \\ 1 \end{bmatrix},\ \mathcal{K} = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix},\ \theta = \begin{bmatrix} 0.03 \\ 0.01 \end{bmatrix},\ h = 0_{2\times 2}, \tag{3.15}$$

$$H_{11} = \begin{bmatrix} 0.04^2 \\ 0 \end{bmatrix},\ H_{22} = \begin{bmatrix} 0 \\ 0.02^2 \end{bmatrix},\ H_{12} = H_{21} = 0_{2\times 1},\ x_t = \begin{bmatrix} 0.04 \\ 0.02 \end{bmatrix}.$$

We note incidentally that in the case of affine-Gaussian models (e.g., equations (3.13) and (3.14)), the object $\mathbb{P}_t^T[b \cdot x_{t_0} < a]$ needed for Approximations A and B is available in closed-form because the T-forward measure distribution of x_{t_0} is still normal. Specifically, we have

(3.16)

$$\begin{aligned} \mathbb{P}_t^T[b \cdot x_{t_0} < a] &= N((a - b'\mathrm{E}^T[x_{t_0} \mid x_t])/\sqrt{b'\mathrm{Var}[x_{t_0} \mid x_t]b}) \\ \mathrm{E}^T[x_{t_0} \mid x_t] &= \mathrm{E}[x_{t_0} \mid x_t] + (\mathrm{Var}[x_{t_0} \mid x_t]e^{-(T-t_0)\mathcal{K}'} - \mathcal{K}^{-1}(I_{n\times n} - e^{-(t_0-t)\mathcal{K}})h)\mathcal{K}^{-1'}b_r \\ \mathrm{E}[x_{t_0} \mid x_t] &= e^{-(t_0-t)\mathcal{K}}(x_t - \theta) + \theta \\ \mathrm{vec}(\mathrm{Var}[x_{t_0} \mid x_t]) &= ((I_{n\times n} \otimes \mathcal{K}) + (\mathcal{K} \otimes I_{n\times n}))^{-1}\mathrm{vec}(h - e^{-(t_0-t)\mathcal{K}}h\, e^{-(t_0-t)\mathcal{K}'}), \end{aligned}$$

where $N(\cdot)$ denotes the standard normal cdf, and the expression e^X with square matrix X denotes the matrix exponential, vec(X) denotes the vectorization of the matrix X, and the symbol $\otimes$ denotes the Kronecker product.

Tables 3.2–3.4 display the results for the 1-factor affine-Gaussian model (3.13), three-factor affine-Gaussian model (3.14), and the 2-factor CIR model (3.15), respectively. These tables give the payer swaption prices for option maturities of 1, 2, 5 years and tenors of 1, 2, 5, 10 years (same maturities and tenors as in SP 2006). As Schrager and Pelsser (2006) do, we examine the ATM case (strike=ATMF), as well as an ITM case (strike=0.85ATMF) and an OTM case (strike=1.15ATMF). The interval between cash flows (δt) is set to 1/2 (i.e., 6 months). For 5-year and 10-year tenors, the same interpolation of $\mathbb{P}^T$ as in Section 3.2 was used. Because a simple closed-form formula for swaption price is available in the case of the 1-factor affine-Gaussian model (recall footnote 14), the exact prices are given in Table 3.2 for comparison with Approximations A and B; on the other hand, in Tables 3.3 and 3.4 the Monte Carlo prices (with error estimates) are given for comparison.

As can be seen in Tables 3.2–3.4, the performance of Approximations A and B-sd is excellent in all three cases examined ((3.13), (3.14), and (3.15)) for all option maturities, tenors, and strikes (ATM, ITM, OTM). Note that in the case of the 1-factor

TABLE 3.2
Upper Panel, Middle Panel, and Lower Panel Give the Swaption Prices (in basis points) for 1-Year, 2-Year, and 5-Year Option Maturities, Respectively, in a 1-Factor Affine-Gaussian Model (Vasicek model) for Tenors of 1-Year, 2-Year, 5-Year, and 10-Year

Method	Strike	1-year	2-year	5-year	10-year
Exact	ITM	80.59	155.87	353.28	605.66
	ATM	35.67	67.95	147.65	238.27
	OTM	11.25	20.78	41.39	58.74
A	ITM	80.59	155.87	353.28	605.66
	ATM	35.67	67.95	147.65	238.27
	OTM	11.25	20.78	41.39	58.74
B-sd	ITM	80.59	155.87	353.28	605.66
	ATM	35.67	67.95	147.65	238.27
	OTM	11.25	20.78	41.39	58.74
B-FW	ITM	80.59	155.87	353.28	605.62
	ATM	35.67	67.95	147.65	238.27
	OTM	11.25	20.78	41.39	58.69
Exact	ITM	86.86	167.45	376.20	637.30
	ATM	46.84	89.23	193.96	313.24
	OTM	21.17	39.55	81.33	121.06
A	ITM	86.86	167.45	376.20	637.30
	ATM	46.84	89.23	193.96	313.24
	OTM	21.17	39.55	81.33	121.06
B-sd	ITM	86.86	167.45	376.20	637.30
	ATM	46.84	89.23	193.96	313.24
	OTM	21.17	39.55	81.33	121.06
B-FW	ITM	86.86	167.45	376.20	637.26
	ATM	46.84	89.23	193.96	313.24
	OTM	21.17	39.55	81.33	121.03
Exact	ITM	91.40	175.62	391.03	654.44
	ATM	59.50	113.41	246.88	399.67
	OTM	35.82	67.53	142.40	220.01
A	ITM	91.40	175.62	391.03	654.44
	ATM	59.50	113.41	246.88	399.67
	OTM	35.82	67.53	142.40	220.01
B-sd	ITM	91.40	175.62	391.03	654.43
	ATM	59.50	113.41	246.87	399.66
	OTM	35.82	67.53	142.40	220.00
B-FW	ITM	91.40	175.62	391.03	654.40
	ATM	59.50	113.41	246.87	399.67
	OTM	35.82	67.53	142.40	220.00

Note: "ATM" denotes strike rate set at ATMF, "ITM" denotes the strike rate of 0.85ATMF, and "OTM" denotes the strike rate of 1.15ATMF.

TABLE 3.3

Upper Panel, Middle Panel, and Lower Panel Give the Swaption Prices (in basis points) for 1-Year, 2-Year, and 5-Year Option Maturities, Respectively, in a 3-Factor Affine Gaussian Model for Tenors of 1-Year, 2-Year, 5-Year, and 10-Year

Method	Strike	1-year	2-year	5-year	10-year
MC	ITM	79.45 (0.01)	154.57 (0.00)	361.49 (0.00)	637.00 (0.00)
	ATM	20.83 (0.02)	33.14 (0.02)	53.35 (0.04)	65.63 (0.05)
	OTM	1.57 (0.01)	1.07 (0.01)	0.15 (0.00)	0.00 (0.00)
A	ITM	79.44	154.56	361.47	636.98
	ATM	20.82	33.12	53.31	65.54
	OTM	1.57	1.06	0.15	0.00
B-sd	ITM	79.44	154.56	361.47	636.98
	ATM	20.82	33.12	53.30	65.53
	OTM	1.57	1.06	0.15	0.00
B-FW	ITM	79.44	154.56	361.47	636.98
	ATM	20.82	33.12	53.31	65.54
	OTM	1.57	1.06	0.14	−0.00
MC	ITM	78.41 (0.01)	150.93 (0.01)	346.33 (0.01)	604.87 (0.01)
	ATM	23.54 (0.02)	38.42 (0.03)	63.67 (0.05)	79.05 (0.06)
	OTM	2.81 (0.01)	2.60 (0.01)	0.90 (0.01)	0.08 (0.00)
A	ITM	78.40	150.91	346.28	604.81
	ATM	23.55	38.43	63.68	79.03
	OTM	2.82	2.61	0.90	0.08
B-sd	ITM	78.40	150.91	346.27	604.81
	ATM	23.55	38.43	63.68	79.02
	OTM	2.82	2.61	0.90	0.07
B-FW	ITM	78.40	150.91	346.27	604.79
	ATM	23.55	38.43	63.68	79.03
	OTM	2.82	2.61	0.89	0.03
MC	ITM	69.45 (0.01)	131.96 (0.01)	295.18 (0.01)	508.86 (0.01)
	ATM	23.21 (0.02)	38.73 (0.03)	65.69 (0.05)	82.17 (0.06)
	OTM	3.79 (0.01)	4.31 (0.01)	2.56 (0.01)	0.51 (0.01)
A	ITM	69.44	131.95	295.16	508.84
	ATM	23.21	38.72	65.68	82.13
	OTM	3.79	4.32	2.57	0.51
B-sd	ITM	69.44	131.95	295.16	508.84
	ATM	23.21	38.72	65.68	82.13
	OTM	3.79	4.32	2.57	0.51
B-FW	ITM	69.44	131.95	295.15	508.77
	ATM	23.21	38.72	65.68	82.13
	OTM	3.79	4.32	2.54	0.39

Notes: "ATM" denotes strike rate set at ATMF, "ITM" denotes the strike rate of 0.85ATMF, and "OTM" denotes the strike rate of 1.15ATMF. Monte Carlo errors are given in parentheses.

TABLE 3.4
Upper Panel, Middle Panel, and Lower Panel Give the Swaption Prices (in basis points) for 1-Year, 2-Year, and 5-Year Option Maturities, Respectively, in a 2-Factor CIR Model for Tenors of 1-Year, 2-Year, 5-Year, and 10-Year

Method	Strike	1-year	2-year	5-year	10-year
MC	ITM	100.92 (0.01)	190.57 (0.01)	410.88 (0.01)	675.43 (0.01)
	ATM	24.90 (0.02)	43.83 (0.04)	77.71 (0.06)	98.30 (0.08)
	OTM	1.95 (0.01)	2.69 (0.01)	2.17 (0.01)	0.67 (0.01)
A	ITM	101.12	190.59	410.91	675.48
	ATM	24.75	43.82	77.70	98.28
	OTM	1.89	2.69	2.16	0.67
B-sd	ITM	101.12	190.59	410.91	675.48
	ATM	24.75	43.82	77.70	98.28
	OTM	1.89	2.69	2.16	0.67
B-FW	ITM	101.12	190.59	410.90	675.45
	ATM	24.75	43.82	77.70	98.28
	OTM	1.89	2.69	2.13	0.50
MC	ITM	92.40 (0.01)	174.12 (0.02)	374.67 (0.02)	618.30 (0.02)
	ATM	29.41 (0.02)	51.79 (0.04)	91.94 (0.07)	116.30 (0.09)
	OTM	5.06 (0.01)	7.55 (0.02)	7.84 (0.03)	3.88 (0.02)
A	ITM	92.43	174.14	374.71	618.35
	ATM	29.40	51.75	91.86	116.19
	OTM	5.02	7.56	7.86	3.88
B-sd	ITM	92.43	174.14	374.71	618.35
	ATM	29.40	51.75	91.86	116.19
	OTM	5.02	7.56	7.86	3.88
B-FW	ITM	92.43	174.14	374.69	618.21
	ATM	29.40	51.75	91.86	116.19
	OTM	5.02	7.55	7.81	3.55
MC	ITM	72.03 (0.02)	135.72 (0.03)	292.63 (0.03)	487.37 (0.03)
	ATM	28.44 (0.02)	50.18 (0.04)	89.29 (0.07)	112.95 (0.09)
	OTM	8.18 (0.02)	12.77 (0.03)	15.33 (0.04)	9.76 (0.03)
A	ITM	72.03	135.75	292.70	487.44
	ATM	28.47	50.19	89.30	112.96
	OTM	8.18	12.78	15.34	9.77
B-sd	ITM	72.03	135.75	292.70	487.44
	ATM	28.47	50.19	89.30	112.96
	OTM	8.18	12.78	15.34	9.77
B-FW	ITM	72.03	135.75	292.67	487.23
	ATM	28.47	50.19	89.30	112.96
	OTM	8.18	12.78	15.31	9.44

Notes: "ATM" denotes strike rate set at ATMF, "ITM" denotes the strike rate of 0.85ATMF, and "OTM" denotes the strike rate of 1.15ATMF. Monte Carlo errors are given in parentheses.

affine-Gaussian model, Approximation A leads to the exact value of $\mathbb{P}_t^T[P(x_{t_0}) < 1]$, since the exercise boundary is just a point in the 1-factor case (while the boundary is a line in the 2-factor case and a surface in the case of 3 or more factors). Therefore, the result of Approximation A is identical to the exact value for tenors of 1-year and 2-year (which do not involve the interpolation of $\mathbb{P}^T$). For tenors of 5-year and 10-year, the exact values and Approximation A values can differ due to an interpolation error, but as can be seen in Table 3.2, this error is so small that the numbers are identical up to the last significant digit used here (two digits below the decimal) for all option maturities (1-year, 2-year, and 5-year). As can be also seen in Table 3.2, Approximation B-sd also led to an excellent agreement with the exact values, while Approximation B-FW's results were slightly less accurate. (For example, the swaption with 2-year option maturity and 10-year tenor has the price of 121.03 according to Approximation B-FW, while the exact formula, Approximation A, and Approximation B-sd all give 121.06.) The examination of the 3-factor affine-Gaussian results (Table 3.3) and the 2-factor CIR results (Table 3.4) leads to the same qualitative conclusion about the performance of Approximations A, B-sd, and B-FW.

Comparing the accuracy of Approximations A and B-sd with the method of SP (2006), it can be seen that Approximations A and B-sd are more accurate. With Approximations A and B-sd, the absolute error magnitudes never exceed 0.2 basis point in *any* of the cases, and typically the errors are much smaller than that. SP's (2006) method leads to absolute errors whose magnitude often exceeds 0.2 basis point in the case of the 1-factor Vasicek model (SP 2006, table 4.2) for OTM and ITM options, reaching numbers as high as 1.99 basis points for OTM options and 2.71 basis points for ITM options. In the case of the 2-factor CIR model, SP (2006, table 6.1) generate error magnitude as large as 0.71 basis points (OTM, 5-year maturity and 10-year tenor). Qualitative differences in absolute errors between the methods of this paper (Approximations A, B-sd) and SP's (2006) method also translate to qualitative differences in the magnitude of relative pricing errors as well, particularly with OTM options.

As regards the computation speed, it is not easy to compare the speed of various proposed methods, since one may not be able to say that each method under comparison has been fully "optimally" programmed. However, here we can make a conservative statement about the speed of Approximation B-sd in comparison with Munk's (1999) stochastic duration method, since the same computational elements are involved: the determination of the stochastic duration (which amounts to the determination of the exercise boundary in Approximation B-sd), and the evaluation of terms of the form $\mathbb{P}_t^T[b \cdot x_{t_0} < a]$. In general, the evaluation of the $\mathbb{P}^T$ terms is going to be the most time-consuming step (though in the case of the 1-factor and 3-factor affine-Gaussian models considered in this subsection, there is a closed-form expression for it). In the case of Munk's stochastic duration method one needs two evaluations of $\mathbb{P}^T$ (at $T = t_0$ and $T = t + D_s$), while in the case of the Approximation B-sd used in this subsection, one needs at most five evaluations of $\mathbb{P}^T$; recall that an interpolation based on five values gave highly accurate results. Therefore, the computation time for Approximation B-sd would be *no more than 2.5 times* that of Munk's stochastic duration method. In the case of Approximation A, the procedure for determining the option exercise boundary is different, but the rest of the steps are the same as Approximation B-sd. As noted in footnote 9, the process for determining the exercise boundary for Approximation A is pretty fast, therefore we would expect Approximation A to be comparable to Approximation B-sd in speed.

According to SP's (2006) estimation of computational speed (in their table 6.3), the computation time for the SP method is about 2.3–2.5 times that of Munk's stochastic duration method, which would imply that Approximations A and B-sd in this paper are comparable to SP's (2006) method in terms of speed.

4. SWAPTION PRICING IN QG MODELS

4.1. General QG Models

To illustrate the application of the methods of this paper to swaption pricing in non-affine models, we shall use QG models. The general QG model is given by

$$r(x_t) = a_r + b_r \cdot x_t + x_t' C_r x_t, \tag{4.1}$$

where a_r is a constant, b_r is an n-dimensional vector of constants, and C_r is an $n \times n$ constant symmetric matrix. The n-dimensional state vector x_t has the risk-neutral dynamics

$$dx_t = \mathcal{K}(\theta - x_t)\,dt + \Sigma\,dW_t, \tag{4.2}$$

where θ is an n-dimensional constant vector, $\mathcal{K}$, Σ are $n \times n$ constant matrices, and W_t is an n-dimensional vector of standard Brownian motions.

Note that the x_t dynamics is a special case of (3.2), with $h_{ij} = (\Sigma\Sigma')_{ij}$, $H_{ij} = 0_{n\times 1}\ \forall\, i, j$. Therefore $\mathrm{E}[x_T|x_t]$ and $\mathrm{Var}[x_T|x_t]$ for the QG model are again available in closed-form. (See (A.11) and (A.16) in the Appendix; alternatively, one can use the expression for $\mathrm{E}[x_T|x_t]$ and $\mathrm{Var}[x_T|x_t]$ in (3.16).)

It is well known that the zero-coupon bond price $Z(t, T)$ in this model is given by[15]

$$Z(t, T) = \exp(A(T - t) + B(T - t) \cdot x_t + x_t' C(T - t) x_t), \tag{4.3}$$

where $A(\tau)$, $B(\tau)$, $C(\tau)$ are the solution of the Riccati equation

$$\begin{aligned} \frac{dA(\tau)}{d\tau} &= (\mathcal{K}\theta)' B(\tau) + \frac{1}{2}\mathrm{tr}((2C(\tau) + B(\tau)B(\tau)')\Sigma\Sigma') - a_r \\ \frac{dB(\tau)}{d\tau} &= -\mathcal{K}' B(\tau) + 2C(\tau)\mathcal{K}\theta + 2C(\tau)\Sigma\Sigma' B(\tau) - b_r \\ \frac{dC(\tau)}{d\tau} &= -(\mathcal{K}' C(\tau) + C(\tau)\mathcal{K}) + 2C(\tau)\Sigma\Sigma' C(\tau) - C_r, \end{aligned} \tag{4.4}$$

with boundary condition $A(0) = 0$, $B(0) = 0_{n\times 1}$, $C(0) = 0_{n\times n}$.

In the case of Approximation A, the coefficients a, b, C for the approximation $\mathbb{P}_t^T[P(x_{t_0}) < 1] \approx \mathbb{P}_t^T[\mathrm{a} + \mathrm{b} \cdot x_{t_0} + x_{t_0}' \mathrm{C}\, x_{t_0} < 0]$ is given by (2.18), which involves func-

[15] See, e.g., Ahn, Dittmar, and Gallant (2002) and Leippold and Wu (2002).

tions $P(x)$, $\partial_x P(x)$, $\partial_x \partial_{x'} P(x)$, given by

$$
\begin{aligned}
(4.5)\qquad P(x) &= \sum_{i=1}^{N} c_i\, e^{A(t_i-t_0)+B(t_i-t_0)\cdot x+x'C(t_i-t_0)x},\\
\partial_x P(x) &= \sum_{i=1}^{N} c_i\, e^{A(t_i-t_0)+B(t_i-t_0)\cdot x+x'C(t_i-t_0)x}(B(t_i-t_0)+2C(t_i-t_0)x),\\
\partial_x \partial_{x'} P(x) &= \sum_{i=1}^{N} c_i\, e^{A(t_i-t_0)+B(t_i-t_0)\cdot x+x'C(t_i-t_0)x}(2C(t_i-t_0)\\
&\quad +(B(t_i-t_0)+2C(t_i-t_0)x)(B(t_i-t_0)+2C(t_i-t_0)x)').
\end{aligned}
$$

In the case of Approximation B, we have

$$
\begin{aligned}
(4.6)\qquad \mathrm{a} &= A(t+D(t)-t_0)+\log(\xi)\\
\mathrm{b} &= B(t+D(t)-t_0)\\
\mathrm{C} &= C(t+D(t)-t_0).
\end{aligned}
$$

If the stochastic duration $D_s(t)$ is used ($D(t)=D_s(t)$), it can be computed as the solution of the equation

$$
\begin{aligned}
(4.7)\qquad &(B(D_s)+2C(D_s)x_t)'\Sigma\Sigma'(B(D_s)+2C(D_s)x_t) =\\
&\left(\sum_{i=1}^{N}\omega_i(B(t_i-t)+2C(t_i-t)x_t)\right)'\Sigma\Sigma'\left(\sum_{i=1}^{N}\omega_i(B(t_i-t)+2C(t_i-t)x_t)\right),
\end{aligned}
$$

where $\omega_i = c_i Z(t, t_i)/(\sum_j c_j Z(t, t_j))$. Again, as in the case of affine models (equation (3.8)), one can speed up the solution of (4.7) by computing the left-hand side using an interpolation formula based on precomputed values of $(B(\tau)+2C(\tau)x_t)'\Sigma\Sigma'(B(\tau)+2C(\tau)x_t)$ at $\tau = \{t_i - t\}_{i=1}^{N}$.

To compute $\mathbb{P}_t^T[\mathrm{a}+\mathrm{b}\cdot x_{t_0}+x_{t_0}'\mathrm{C}x_{t_0}<0]$, again we use the Levy inversion formula

$$
(4.8)\qquad \mathbb{P}_t^T[\mathrm{a}+\mathrm{b}\cdot x_{t_0}+x_{t_0}'\mathrm{C}x_{t_0}<0] = \frac{1}{2}-\frac{1}{\pi}\int_0^\infty dz\frac{\mathrm{Im}\big(f_t^T(z)\big)}{z}.
$$

The T-forward measure characteristic function $f_t^T(z)$ is given by

$$
\begin{aligned}
(4.9)\qquad f_t^T(z) &= \mathrm{E}_t^T\big[e^{iz(\mathrm{a}+\mathrm{b}\cdot x_{t_0}+x_{t_0}'\mathrm{C}x_{t_0})}\big]\\
&= Z(t,T)^{-1}\mathrm{E}_t\big[e^{-\int_t^T r_s ds}e^{iz(\mathrm{a}+\mathrm{b}\cdot x_{t_0}+x_{t_0}'\mathrm{C}x_{t_0})}\big]\\
&= Z(t,T)^{-1}\mathrm{E}_t\big[e^{-\int_t^{t_0} r_s ds}e^{(A(T-t_0)+iz\mathrm{a})+(B(T-t_0)+iz\mathrm{b})\cdot x_{t_0}+x_{t_0}'(C(T-t_0)+iz\mathrm{C})x_{t_0}}\big]\\
&= e^{\tilde{A}(t_0-t;T-t_0,z)-A(T-t)+(\tilde{B}(t_0-t;T-t_0,z)-B(T-t))\cdot x_t+x_t'(\tilde{C}(t_0-t;T-t_0,z)-C(T-t))x_t},
\end{aligned}
$$

where $\tilde{A}(\tau; T-t_0, z)$, $\tilde{B}(\tau; T-t_0, z)$, $\tilde{C}(\tau; T-t_0, z)$ are the solution of the Riccati equations (4.4), with boundary conditions $\tilde{A}(0; T-t_0, z) = A(T-t_0)+iz\mathrm{a}$, $\tilde{B}(0; T-t_0, z) = B(T-t_0)+iz\mathrm{b}$, $\tilde{C}(0; T-t_0, z) = C(T-t_0)+iz\mathrm{C}$.

4.2. Two-Factor QG Model Illustration

Let us now apply the methods to the swaption pricing in the general 2-factor QG model, with parameters from an estimation in Kim (2007):

$$a_r = 0.0444,\ b_r = \begin{bmatrix} 0 \\ 0 \end{bmatrix},\ C_r = \begin{bmatrix} 1 & 0.4412 \\ 0.4412 & -1 \end{bmatrix}, \tag{4.10}$$

$$\mathcal{K} = \begin{bmatrix} -0.0541 & 0.0361 \\ -1.2113 & 0.4376 \end{bmatrix},\ \theta = \begin{bmatrix} 0.1932 \\ 0.1421 \end{bmatrix},\ \Sigma = \begin{bmatrix} 0.0145 & 0 \\ 0 & 0.0236 \end{bmatrix}.$$

We set the state vector $x_t = [0.1690, -0.0501]'$, which is the estimated long-run physical mean.

Note that C_r is not a positive-definite matrix; more specifically it has one negative eigenvalue and one positive eigenvalue.[16] Thus, for low maturities, $C(\tau)$ (in equation (4.3)) is not positive-definite, but it turns out that $C(\tau)$ becomes positive definite for large enough τ.

These parameters therefore allow us to examine the swaption pricing for quite different boundary shapes: For short tenors, the boundary equation $P(x) - 1 = 0$ is approximately a hyperbola, while for long tenors it is approximately an ellipse. This is illustrated in Figure 4.1. The boundary curve $P(x) - 1 = 0$, with $K = 8\%$, is shown in Figure 4.1(a, b), (c, d), and (e, f) for tenors of 2-year, 5-year, and 10-year, respectively.

The approximate boundary based on Approximation A is shown in Figure 4.1(a), (c), and (e). The * point again denotes x^*, given by (2.20). The approximate boundary based on Approximation B with stochastic duration (B-sd) is shown in Figure 4.1(b), (d), and (f). It can be seen that both A and B-sd match the true boundary $P(x) - 1 = 0$ very well at least in the neighborhood of x^*. Away from x^*, the approximate boundaries can deviate from the true boundary a lot. A striking example is the case of the 5-year tenor in Figure 4.1(d): the true boundary is ellipse-like, but the boundary based on Approximation B-sd is a hyperbola. This, however, is not necessarily a problem, since the contribution of those parts with large mismatch is very small.

Table 4.1 shows the swaption prices for tenors of 2, 5, 10 years and option maturities of 1, 2, 5 years, based on five different methods: the Monte Carlo simulation, Approximation A, Approximation B with stochastic duration, Approximation B with Fisher–Weil duration, and Munk's (1999) stochastic duration approximation.

As in affine model illustrations in Section 3.2, δt was set to $1/2$, and for the 5-year and 10-year tenors we use the cubic-spline interpolation based on the evaluation of $\mathbb{P}_t^T$ at $T = t_0 + (j/m)(t_N - t_0)$, where $j = 0, 1, \ldots, m$ and $m = 4$ (for Approximations A and B).

Again, as in Section 3.2, the strike rates are chosen as (ATMF-1.5%, ATMF-0.75%, ATMF, ATMF+0.75%, ATMF+1.5%). Note that how much the strike is out-of-money (or in-the-money) depends on the volatility. The parameters for the QG model illustration (equation (4.10)) imply less volatile interest rates than the 2-factor CIR model parameters in Section 3.2 (equation (3.12));[17] therefore, the strike of ATMF+1.5% in this section is

[16] This makes nominal short rate no longer positive-definite, but one gains additional flexibility to fit interest rate dynamics (compared to the positive-definite case). This is analogous to the mixed-sign multifactor CIR model of Backus et al. (2001), who find that it fits data better than the standard n-factor CIR model.

[17] The estimation in Kim (2007) is based on a later period sample, which had less volatile interest rates than Duffie and Singleton (1997).

TABLE 4.1
Upper Panel, Middle Panel, and Lower Panel Give the Swaption Prices for 1-Year, 2-Year, and 5-Year Option Maturities, Respectively, in a 2-Factor QG Model

Method	Tenor	ATMF−1.5%	ATMF−0.75%	ATMF	ATMF+0.75%	ATMF+1.5%
MC	2	255.46 (0.01)	138.47 (0.02)	52.28 (0.04)	11.84 (0.02)	1.46 (0.01)
	5	567.72 (0.02)	304.40 (0.03)	106.36 (0.07)	17.86 (0.04)	1.02 (0.01)
	10	943.25 (0.04)	494.64 (0.03)	148.30 (0.09)	12.20 (0.03)	0.04 (0.00)
A	2	255.43	138.48	52.33	11.87	1.47
	5	567.67	304.42	106.45	17.92	1.02
	10	943.16	494.69	148.55	12.32	0.04
B-sd	2	255.43	138.48	52.33	11.87	1.47
	5	567.66	304.41	106.44	17.92	1.01
	10	943.11	494.66	148.49	12.32	−0.00
B-FW	2	255.43	138.48	52.33	11.87	1.47
	5	567.66	304.40	106.44	17.89	0.97
	10	943.10	494.49	148.53	11.89	−0.15
Munk	2	255.41	138.40	52.25	11.86	1.48
	5	567.38	303.59	106.02	18.37	1.21
	10	942.54	492.38	147.75	14.12	0.17
MC	2	240.36 (0.02)	140.50 (0.03)	65.81 (0.05)	23.12 (0.04)	5.81 (0.02)
	5	532.17 (0.03)	304.37 (0.04)	130.23 (0.08)	34.68 (0.05)	4.43 (0.02)
	10	880.43 (0.07)	488.21 (0.03)	182.48 (0.09)	28.13 (0.05)	0.30 (0.00)
A	2	240.34	140.44	65.71	23.06	5.78
	5	532.21	304.31	130.08	34.60	4.40
	10	880.59	488.27	182.41	28.14	0.30
B-sd	2	240.34	140.44	65.71	23.06	5.78
	5	532.20	304.31	130.06	34.59	4.39
	10	880.52	488.26	182.32	28.14	0.09
B-FW	2	240.34	140.44	65.71	23.06	5.78
	5	532.20	304.29	130.07	34.58	4.31
	10	880.48	488.09	182.38	27.84	−0.32
Munk	2	240.25	140.29	65.59	23.04	5.82
	5	531.25	302.89	129.45	35.36	5.10
	10	878.30	484.58	181.36	31.37	1.09
MC	2	198.10 (0.03)	124.42 (0.03)	65.92 (0.04)	27.59 (0.03)	8.59 (0.02)
	5	438.24 (0.06)	268.77 (0.04)	130.24 (0.05)	39.95 (0.05)	5.04 (0.02)
	10	724.44 (0.12)	432.84 (0.07)	190.77 (0.05)	36.27 (0.04)	0.24 (0.00)
A	2	198.09	124.40	65.91	27.58	8.58
	5	438.21	268.75	130.21	39.92	5.04
	10	724.40	432.83	190.72	36.24	0.24
B-sd	2	198.09	124.40	65.91	27.58	8.58
	5	438.20	268.74	130.18	39.90	4.99
	10	724.39	432.80	190.63	36.16	−0.07
B-FW	2	198.09	124.40	65.91	27.58	8.58
	5	438.20	268.73	130.19	39.91	4.93
	10	724.32	432.73	190.69	36.12	−0.34
Munk	2	197.94	124.25	65.84	27.65	8.71
	5	436.88	267.41	129.80	41.18	6.25
	10	721.63	430.23	190.44	40.04	0.99

Note: The Monte Carlo errors are given in parentheses.

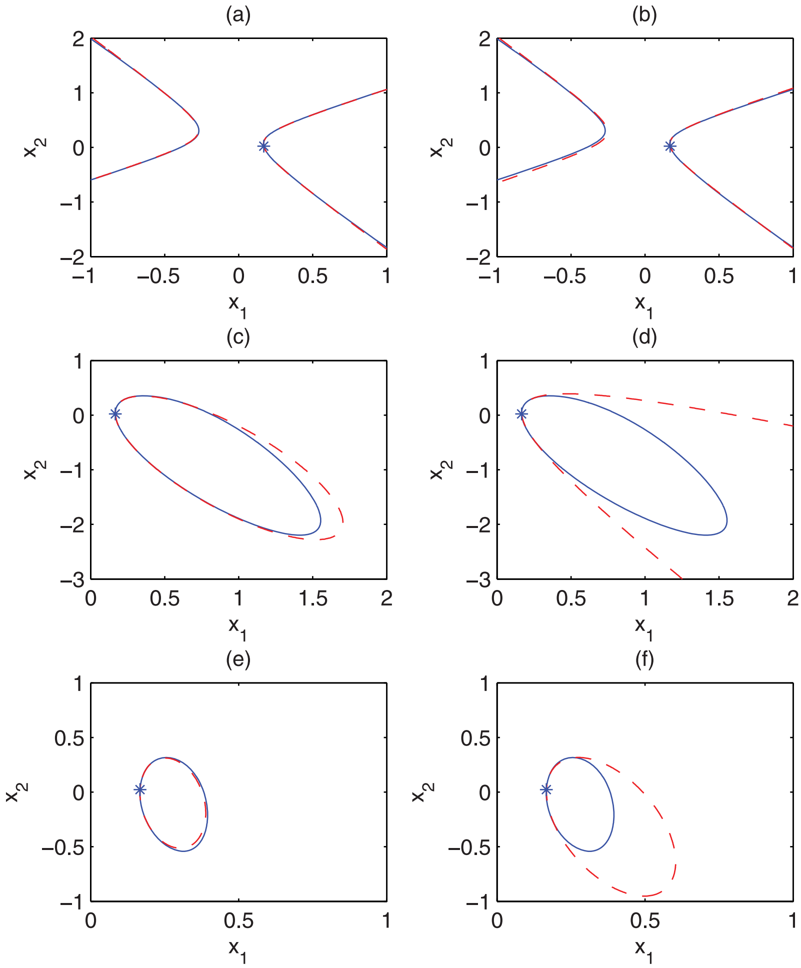

FIGURE 4.1. True exercise boundaries (solid line) and approximate boundaries (dashed line). (a) 2-year tenor, Approximation A. (b) 2-year tenor, Approximation B-sd. (c) 5-year tenor, Approximation A. (d) 5-year tenor, Approximation B-sd. (e) 10-year tenor, Approximation A. (f) 10-year tenor, Approximation B-sd.

more deeply out-of-money than ATMF+1.5% in Section 3.2. This can be also seen from the fact that the swaption prices for the strike of ATMF+1.5% are very small, e.g., for the 1-year-maturity, 10-year-tenor swaption, the price at ATMF+1.5% is only 0.04 basis points.

As can be seen in Table 4.1, the results based on Approximation A have an impressive match with those based on Monte Carlo simulation for the entire range of option maturities (1, 2, 5 years), tenors (2, 5, 10 years), and strikes (from ATMF-1.5% to ATMF+1.5%) considered in this paper. Approximation B with stochastic duration (B-sd) has a similarly good performance, except for the case of the 10-year tenor and the strike of ATMF+1.5%, where it generates negative option prices for the option maturities of 1-year and 5-year. This is not likely to be a serious problem, since these options are not likely to exist (trade) in practice; they are too deep out-of-money, as noted earlier.

Approximation B with Fisher–Weil duration (B-FW) also yields very accurate results. But in some cases (low-price out-of-money options), its results are not as accurate as Approximation B with stochastic duration. For example, the swaption price for the 10-year tenor and the strike of ATMF + 0.75% is (11.89, 27.84) for (1-year, 2-year) option maturity for Approximation B-FW. On the other hand, the corresponding numbers for Approximation B-sd are (12.32, 28.14), which is much closer to the Monte Carlo value of (12.20, 28.13). In any case, even Approximation B-FW is much more accurate than the Munk's (1990) stochastic duration method, which generates the corresponding price of (14.12, 31.37). Approximations A and B have been also tested with other parameters and state vectors; these tests led to similarly good conclusions about the performance of Approximations A and B.

5. CONCLUDING REMARKS

In this paper, we have explored several modifications of SU's (2002) method of approximating the exercise boundary so that (1) the method does not require the knowledge of the pdf, and (2) it is easily generalized and applied to models beyond the affine class. Among these modifications, we find that the Taylor series method (Approximation A) performs very well across the entire option maturities, tenors, and strikes considered. The zero-coupon bond approximation with stochastic duration (Approximation B-sd) also performed similarly well, except at moneyness levels that are so deeply out-of-the-money as to be practically irrelevant. The methods work very well partly because the errors in the evaluation of T-forward measure probabilities $\mathbb{P}_t^T[P(x_{t_0}) < 1]$ cancel out to a large extent. For applications beyond the affine class, we have focused on QG models and found that the methods of this paper (Approximations A and B-sd) work well for QG models too; we can expect the methods to have broad applicability to yet other models, as long as they have a tractable formula for the zero-coupon bond option price.

APPENDIX: CONDITIONAL MEAN AND VARIANCE IN AFFINE MODELS

In this appendix, we derive the expression for the conditional first and second moments of the state vector in affine models (needed for Approximation A). More specifically, we evaluate $\mathrm{E}[x_T|x_t]$ and $\mathrm{Var}[x_T|x_t]$ for the n-dimensional vector process

$$dx_t = (\mathcal{K}\theta - \mathcal{K}x_t)\,dt + \Sigma(x_t)\,dW_t, \tag{A.1}$$

where $(\Sigma(x_t)\Sigma(x_t)')_{ij} \equiv h_{ij} + H'_{ij}x_t$. ($W_t$ is an n-dimensional vector of independent Brownian motions, $\mathcal{K}$ is an $n \times n$ constant matrix, $\mathcal{K}\theta$ and H_{ij} $(i, j = 1, \ldots, n)$ are n-dimensional constant vectors, and h_{ij} $(i, j = 1, \ldots, n)$ is a scalar.) We allow for as general $\mathcal{K}$ matrix as possible, e.g., $\mathcal{K}$ could be noninvertible, nondiagonal, or nonstationary. It could also have complex eigenvalues.[18]

We use the fact that the matrix exponential $e^{\mathcal{K}t}$ $(\equiv I_{n\times n} + \mathcal{K}t + (1/2)\mathcal{K}^2t^2 + \cdots)$ has the property that $d(e^{\mathcal{K}t}) = \mathcal{K}e^{\mathcal{K}t}dt = e^{\mathcal{K}t}\mathcal{K}dt$. Using the Ito's lemma, we have

$$d(e^{\mathcal{K}t}x_t) = \mathcal{K}e^{\mathcal{K}t}x_t dt + e^{\mathcal{K}t}dx_t. \tag{A.2}$$

[18] An equivalent expression for $\mathrm{E}[x_T|x_t]$ and $\mathrm{Var}[x_T|x_t]$ has been derived by Fisher and Gilles (1996), with a different notation for the specification of the $\Sigma(x_t)\Sigma(x_t)'$ matrix.

Substituting in dx_t, we have

$$d(e^{\mathcal{K}t}x_t) = e^{\mathcal{K}t}\mathcal{K}\theta dt + e^{\mathcal{K}t}\Sigma(x_t)\,dW_t. \tag{A.3}$$

Integrating this expression from t to T, we have

$$e^{\mathcal{K}T}x_T - e^{\mathcal{K}t}x_t = \left(\int_t^T e^{\mathcal{K}s}ds\right)\mathcal{K}\theta + \int_t^T e^{\mathcal{K}s}\Sigma(x_s)\,dW_s. \tag{A.4}$$

Therefore,

$$x_T = e^{-\mathcal{K}(T-t)}x_t + \left(\int_t^T e^{-\mathcal{K}(T-s)}ds\right)\mathcal{K}\theta + \int_t^T e^{-\mathcal{K}(T-s)}\Sigma(x_s)\,dW_s. \tag{A.5}$$

From this, the conditional mean is immediately obtained:

$$\mathrm{E}[x_T \mid x_t] = e^{-\mathcal{K}(T-t)}x_t + \left(\int_t^T e^{-\mathcal{K}(T-s)}ds\right)\mathcal{K}\theta. \tag{A.6}$$

The conditional variance–covariance matrix is

$$\mathrm{Var}[x_T \mid x_t] = \int_t^T e^{-\mathcal{K}(T-s)}\mathrm{E}[\Sigma(x_s)\Sigma(x_s)' \mid x_t]e^{-\mathcal{K}'(T-s)}\,ds. \tag{A.7}$$

We shall now consider the case of invertible $\mathcal{K}$ matrix and the case of noninvertible $\mathcal{K}$ matrix separately.

Invertible $\mathcal{K}$ matrix: If $\mathcal{K}$ is an invertible matrix, the integral in (A.6) is easily done, and we have

$$\mathrm{E}[x_T \mid x_t] = e^{-\mathcal{K}(T-t)}(x_t - \theta) + \theta. \tag{A.8}$$

The matrix exponential in (A.8) can be evaluated by diagonalizing the $\mathcal{K}$ matrix,

$$\mathcal{K} = U\,\mathrm{D}([\gamma_1, \dots, \gamma_n]')\,U^{-1}, \tag{A.9}$$

where $\mathrm{D}(x)$ denotes a diagonal matrix whose diagonal elements are the vector x, γ_i's are the eigenvalues of $\mathcal{K}$, and U is a matrix whose columns are eigenvectors of $\mathcal{K}$. Using the fact that

$$\begin{aligned} e^{\mathcal{K}t} &= e^{U\,\mathrm{D}([\gamma_1,\dots,\gamma_n]')\,U^{-1}t} \\ &= I_{n\times n} + tU\,\mathrm{D}([\gamma_1, \dots, \gamma_n]')\,U^{-1} + \frac{1}{2}t^2U\,\mathrm{D}^2([\gamma_1, \dots, \gamma_n]')\,U^{-1} + \cdots \\ &= U\mathrm{D}([e^{\gamma_1 t}, \dots, e^{\gamma_n t}]')U^{-1}, \end{aligned} \tag{A.10}$$

we have

$$\mathrm{E}[x_T \mid x_t] = U\mathrm{D}\big([e^{-\gamma_1(T-t)}, \dots, e^{-\gamma_n(T-t)}]'\big)U^{-1}(x_t - \theta) + \theta. \tag{A.11}$$

To evaluate the conditional variance–covariance matrix, again we use the diagonalization of the $\mathcal{K}$ matrix (equation (A.9)). Straightforward calculation shows that the elements of the matrix $\mathrm{Var}[x_T|x_t]$ are given by

(A.12)
$$(\mathrm{Var}[x_T \mid x_t])_{ij} = \sum\nolimits_{p,q,l,m} \int_t^T ds\, U_{ip}\, e^{-\gamma_p(T-s)}(U^{-1})_{pl}(\mathrm{E}[\Sigma(x_s)\Sigma(x_s)' \mid x_t])_{lm}(U^{-1'})_{mq} e^{-\gamma_q(T-s)}(U')_{qj}.$$

Since

(A.13)
$$\begin{aligned}(\mathrm{E}[\Sigma(x_s)\Sigma(x_s)' \mid x_t])_{lm} &= h_{lm} + H_{lm}'\mathrm{E}[x_s \mid x_t] \\ &= h_{lm} + H_{lm}'\theta + H_{lm}'U\mathrm{D}(\left[e^{-\gamma_1(s-t)}, \ldots, e^{-\gamma_n(s-t)}\right]')U^{-1}(x_t - \theta),\end{aligned}$$

we have

$$\begin{aligned}(\mathrm{Var}[x_T \mid x_t])_{ij} = &\sum_{p,q,l,m} U_{ip}\Gamma_{pq}(U^{-1})_{pl}\Omega_{lm}(U^{-1'})_{mq}(U')_{qj} \\ &+ \sum_{p,q,l,m,k} U_{ip}(\Gamma_k)_{pq}(U^{-1})_{pl}(\Omega_k)_{lm}(U^{-1'})_{mq}(U')_{qj}\end{aligned} \tag{A.14}$$

where we have defined $n \times n$ symmetric matrices Ω, Ω_k, Γ, Γ_k ($k = 1, \ldots, n$) as

$$\begin{aligned}\Omega_{ij} &= h_{ij} + H_{ij}'\theta, \\ (\Omega_k)_{ij} &= (H_{ij}'U)_k(U^{-1}(x_t - \theta))_k, \\ \Gamma_{ij} &= \frac{1 - e^{-(\gamma_i+\gamma_j)(T-t)}}{\gamma_i + \gamma_j} \\ (\Gamma_k)_{ij} &= \frac{e^{-\gamma_k(T-t)} - e^{-(\gamma_i+\gamma_j)(T-t)}}{\gamma_i + \gamma_j - \gamma_k}.\end{aligned} \tag{A.15}$$

In the expression for Ω_k, the notation $(v)_k$ denotes the kth element of the vector v.

Equation (A.14) can be written more simply in matrix notation as

$$\mathrm{Var}[x_T \mid x_t] = U\left(((U^{-1}\Omega U^{-1'}) \circ \Gamma) + \sum_{k=1}^{n}((U^{-1}\Omega_k U^{-1'}) \circ \Gamma_k)\right)U', \tag{A.16}$$

where the circle $\circ$ denotes the Hadamard product (entry-wise multiplication). Note that this formula is valid regardless of whether $\mathcal{K}$ has all-real eigenvalues or not.[19]

Noninvertible $\mathcal{K}$ matrix: If the matrix $\mathcal{K}$ is not invertible,[20] at least one of its eigenvalues is zero, and the long-run mean θ is not defined. To obtain $\mathrm{E}[x_T|x_t]$ in this case, we start from (A.6), and evaluate the matrix integral using the diagonalized form of $\mathcal{K}$

[19] If $\mathcal{K}$ has complex eigenvalues, the matrix U is also complex. In this case, the formula (A.16) still gives the correct answer, but note that U' is the transpose matrix, not the Hermitian conjugate matrix. (In some packages like matlab, the notation U' gives the Hermitian conjugate matrix when U is complex.)

[20] A prominent example of a noninvertible $\mathcal{K}$ matrix is the affine stochastic volatility model of stock prices, as in Duffie et al. (2000).

(equation (A.9)). This gives

$$\text{(A.17)}\quad \mathrm{E}[x_T \mid x_t] = U\mathrm{D}([e^{-\gamma_1(T-t)}, \ldots, e^{-\gamma_n(T-t)}]')U^{-1}x_t + U\mathrm{D}\left(\left[\int_t^T e^{-\gamma_1(T-s)}ds, \ldots, \int_t^T e^{-\gamma_n(T-s)}ds\right]'\right) U^{-1}\mathcal{K}\theta.$$

The integral in (A.17) is

$$\text{(A.18)}\quad \int_t^T e^{-\gamma_i(T-s)}ds = \begin{cases} (T-t) & (\gamma_i = 0) \\ (1-e^{-\gamma_i(T-t)})/\gamma_i & (\gamma_i \neq 0). \end{cases}$$

To obtain $\mathrm{Var}[x_T|x_t]$, we proceed as in the case of the invertible $\mathcal{K}$ matrix, except that instead of (A.13), we have

$$\text{(A.19)}\quad (\mathrm{E}[\Sigma(x_s)\Sigma(x_s)' \mid x_t])_{lm} = h_{lm} + H'_{lm}U\mathrm{D}([e^{-\gamma_1(s-t)}, \ldots, e^{-\gamma_n(s-t)}]')U^{-1}x_t - H'_{lm}U\mathrm{D}\left(\left[\int_t^s e^{-\gamma_1(u-t)}du, \ldots, \int_t^s e^{-\gamma_n(u-t)}du\right]'\right) U^{-1}\mathcal{K}\theta.$$

It is then straightforward to show that

$$\text{(A.20)}\quad \mathrm{Var}[x_T \mid x_t] = U\left((U^{-1}\Omega^a U^{-1'}) \circ \Gamma^a + \sum_{k=1}^n ((U^{-1}\Omega_k^b U^{-1'}) \circ \Gamma_k^b - (U^{-1}\Omega_k^c U^{-1'}) \circ \Gamma_k^c)\right) U',$$

where the $n \times n$ symmetric matrices Ω^a, Ω_k^b, Ω_k^c, Γ^a, Γ_k^b, Γ_k^c $(k = 1, \ldots, n)$ are defined as

$$\text{(A.21)}\quad \begin{aligned} (\Omega^a)_{ij} &= h_{ij}, \\ (\Omega_k^b)_{ij} &= (H'_{ij}U)_k(U^{-1}x_t)_k, \\ (\Omega_k^c)_{ij} &= (H'_{ij}U)_k(U^{-1}\mathcal{K}\theta)_k, \\ (\Gamma^a)_{ij} &= \int_t^T e^{-(\gamma_i+\gamma_j)(T-s)}ds, \\ (\Gamma_k^b)_{ij} &= \int_t^T e^{-(\gamma_i+\gamma_j)(T-s)-\gamma_k(s-t)}ds, \\ (\Gamma_k^c)_{ij} &= \int_t^T e^{-(\gamma_i+\gamma_j)(T-s)}\left(\int_t^s e^{-\gamma_k(u-t)}du\right) ds. \end{aligned}$$

The integrals in (A.21) have different expressions, depending on whether $\gamma_k = 0$ or not, whether $\gamma_i + \gamma_j = 0$ or not, and whether $\gamma_k = \gamma_i + \gamma_j$ or not:[21]

$$\text{(A.22)}\quad (\Gamma^a)_{ij} = \begin{cases} T-t & (\gamma_i + \gamma_j = 0) \\ \dfrac{1-e^{-(\gamma_i+\gamma_j)(T-t)}}{\gamma_i+\gamma_j} & (\gamma_i + \gamma_j \neq 0) \end{cases}$$

[21] Note that the situation $\gamma_i + \gamma_j = 0$ would occur when $i = j$ and $\gamma_i = 0$, and the situation $\gamma_k = \gamma_i + \gamma_j$ would occur when $i = k$ and $\gamma_j = 0$ or $j = k$ and $\gamma_i = 0$.

$$(\text{A.23}) \qquad (\Gamma_k^b)_{ij} = \begin{cases} e^{-\gamma_k(T-t)}(T-t) & (\gamma_i+\gamma_j=\gamma_k) \\ \dfrac{e^{-\gamma_k(T-t)} - e^{-(\gamma_i+\gamma_j)(T-t)}}{\gamma_i+\gamma_j-\gamma_k} & (\gamma_i+\gamma_j \neq \gamma_k) \end{cases}$$

(A.24)

$$(\Gamma_k^c)_{ij} = \begin{cases} (T-t)^2/2 & (\gamma_k = 0, \gamma_i+\gamma_j = 0) \\ (e^{-(\gamma_i+\gamma_j)(T-t)} + (\gamma_i+\gamma_j)(T-t) - 1)/(\gamma_i+\gamma_j)^2 & (\gamma_k = 0, \gamma_i+\gamma_j \neq 0) \\ ((\Gamma^a)_{ij} - (\Gamma_k^b)_{ij})/\gamma_k & (\gamma_k \neq 0). \end{cases}$$

REFERENCES

AHN, D.-H., R. DITTMAR, and A. R. GALLANT (2002): Quadratic Term Structure Models: Theory and Evidence, *Rev. Finan. Stud.* 12, 243–288.

ALMEIDA, C., J. GRAVELINE, and S. JOSLIN (2011): Do Interest Rate Options Contain Information about Excess Returns?, *J. Econ.* 164, 35–44.

BACKUS, D., S. FORESI, A. MOZUMDAR, and L. WU (2001): Predictable Changes in Yields and Forward Rates, *J. Finan. Econ.* 59, 281–311.

BRIGO, D., and F. MERCURIO (2001): *Interest Rate Models Theory and Practice*, Berlin, Heidelberg: Springer-Verlag.

COLLIN-DUFRESNE, P., and R. GOLDSTEIN (2002): Swaption Pricing in an Affine Framework, *J. Derivatives* 10, 9–26.

COX, J. C., J. E. INGERSOLL, and S. A. ROSS (1979): Duration and the Measurement of Basis Risk, *J. Business* 52, 51–61.

DAI, Q., and K. J. SINGLETON (2000): Specification Analysis of Affine Term Structure Models, *J. Finance* 55, 1943–1978.

DUFFIE, D., and R. KAN (1996): A Yield-Factor Model of Interest Rates, *Math. Finance* 6, 379–406.

DUFFIE, D., J. PAN, and K. J. SINGLETON (2000): Transform Analysis and Asset Pricing for Affine Jump Diffusions, *Econometrica* 68, 1343–1376.

DUFFIE, D., and K. J. SINGLETON (1997): An Econometric Model of the Term Structure of Interest-Rate Swap Yields, *J. Finance* 52, 1287–1321.

FISHER, L., and R. L. WEIL (1979): Coping with the Risk of Interest-Rate Fluctuations, *J. Business* 44, 408–431.

FISHER, M., and C. GILLES (1996): Term Premia in Exponential-Affine Models of the Term Structure, Working Paper.

HEATH, D., R. JARROW, and A. MORTON (1992): Bond Pricing and the Term Structure of Interest Rates: A New Methodology for Contingent Claims Valuation, *Econometrica* 60, 77–105.

HEIDARI, M., A. HIRSA, and D. B. MADAN (2007): Pricing of Swaptions in Affine Term Structures with Stochastic Volatility, *Adv. Math. Finance*, Basel: Birkhäuser.

HEIDARI, M., and L. WU (2009): A Joint Framework for Consistently Pricing Interest Rates and Interest Rate Derivatives, *J. Finan. Quant. Anal.* 44, 517–550.

JAGANNATHAN, R., A. KAPLIN, and S. SUN (2003): An Evaluation of Multi-Factor CIR Models Using LIBOR, Swap Rates, and Cap and Swaption Prices, *J. Econ.* 116, 113–146.

JAMSHIDIAN, F. (1989): An Exact Option Pricing Formula, *J. Finance* 44, 205–209.

JOSLIN, S. (2010): Pricing and Hedging Volatility Risk in Fixed Income Markets, USC Working Paper.

KIM, D. H. (2007): Spanned Stochastic Volatility in Bond Markets: A Reexamination of the Relative Pricing between Bonds and Bond Options, BIS Working Paper.

LEIPPOLD, M., and L. WU (2002): Asset Pricing under the Quadratic Class, *J. Finan. Quant. Anal.* 37, 375–409.

LI, H., and F. ZHAO (2006): Unspanned Stochastic Volatility: Evidence from Hedging Interest Rate Derivatives, *J. Finance* 61, 341–378.

LONGSTAFF, F. A., P. SANTA-CLARA, and E. S. SCHWARTZ (2002): The Relative Valuation of Caps and Swaptions: Theory and Empirical Evidence, *J. Finance* 56, 2067–2109.

MUNK, C. (1999): Stochastic Duration and Fast Coupon Bond Option Pricing in Multi-Factor Models, *Rev. Derivat.* 3, 157–181.

SCHRAGER, D. F., and A. A. J. PELSSER (2006): Pricing Swaptions and Coupon Bond Options in Affine Term Structure Models, *Math. Finance* 16, 673–694.

SINGLETON, K. J., and L. UMANTSEV (2002): Pricing Coupon-Bond Options and Swaptions in Affine Term Structure Models, *Math. Finance* 12, 427–446.

WEI, J. Z. (1997): A Simple Approach to Bond Option Pricing, *J. Futures Markets* 17, 131–160.

MATHEMATICAL FINANCE
An International Journal of Mathematics, Statistics and Financial Economics

Mathematical Finance, Vol. 24, No. 4 (October 2014), 821–854

ARBITRAGE BOUNDS FOR PRICES OF WEIGHTED VARIANCE SWAPS

MARK DAVIS

Department of Mathematics, Imperial College London

JAN OBŁÓJ

Mathematical Institute, University of Oxford

VIMAL RAVAL

Department of Mathematics, Imperial College London

We develop a theory of robust pricing and hedging of a weighted variance swap given market prices for a finite number of co-maturing put options. We assume the put option prices do not admit arbitrage and deduce no-arbitrage bounds on the weighted variance swap along with super- and sub-replicating strategies that enforce them. We find that market quotes for variance swaps are surprisingly close to the model-free lower bounds we determine. We solve the problem by transforming it into an analogous question for a European option with a convex payoff. The lower bound becomes a problem in semi-infinite linear programming which we solve in detail. The upper bound is explicit. We work in a model-independent and probability-free setup. In particular, we use and extend Föllmer's pathwise stochastic calculus. Appropriate notions of arbitrage and admissibility are introduced. This allows us to establish the usual hedging relation between the variance swap and the "log contract" and similar connections for weighted variance swaps. Our results take the form of a FTAP: we show that the absence of (weak) arbitrage is equivalent to the existence of a classical model which reproduces the observed prices via risk-neutral expectations of discounted payoffs.

KEY WORDS: weighted variance swap, weak arbitrage, arbitrage conditions, model-independent bounds, pathwise Itô calculus, semi-infinite linear programming, fundamental theorem of asset pricing, model error.

1. INTRODUCTION

In the practice of quantitative finance, the risks of "model error" are by now universally appreciated. Different models, each perfectly calibrated to market prices of a set of liquidly traded instruments, may give widely different prices for contracts outside the calibration set. The implied hedging strategies, even for contracts within the calibration set, can vary from accurate to useless, depending on how well the model captures the sample

Jan Obłój gratefully acknowledges support from the Oxford-Man Institute of Quantitative Finance. The work of Vimal Raval was supported by EPSRC under a Doctoral Training Award. The authors thank an anonymous referee for several insightful comments which have improved the paper.

Manuscript received June 2011; final revision received September 2012.

Address correspondence to Jan Obłój, Mathematical Institute, University of Oxford, Oxford OX1 3LB, UK; e-mail: obloj@maths.ox.ac.uk.

DOI: 10.1111/mafi.12021

path behavior of the hedging instruments. This fundamental problem has motivated a large body of research ranging from asset pricing through portfolio optimization, risk management to macroeconomics (see, e.g., Cont 2006; Föllmer et al. 2009; Acciaio et al. 2011; Hansen and Sargent 2010 and the references therein). Another stream of literature, to which this paper is a contribution, develops a robust approach to mathematical finance, see Hobson (1998), Davis and Hobson (2007), Cox and Obłój (2011a). In contrast to the classical approach, no probabilistic setup is assumed. Instead we suppose we are given current market quotes for the underlying assets and some liquidly traded options. We are interested in no-arbitrage bounds on a price of an option implied by this market information. Further, we want to understand if, and how, such bounds may be enforced, in a model-independent way, through hedging.

In this paper, we consider robust pricing and hedging of a weighted variance swap when prices of a finite number of co-maturing put options are given. If $(S_t,\, t \in [0, T])$ denotes the price of a financial asset, the *weighted realized variance* is defined as

$$RV_T = \sum_{i=1}^{n} h(S_{t_i}) \left(\log \frac{S_{t_i}}{S_{t_{i-1}}}\right)^2, \tag{1.1}$$

where t_i is a pre-specified sequence of times $0 = t_0 < t_1 < \cdots < t_n = T$, in practice often daily sampling, and h is a given weight function. A *weighted variance swap* is a forward contract in which cash amounts equal to $A \times RV_T$ and $A \times P_T^{\mathrm{RV}}$ are exchanged at time T, where A is the dollar value of one variance point and P_T^{RV} is the variance swap "price," agreed at time 0. Three representative cases considered in this paper are the *plain vanilla variance swap* $h(s) \equiv 1$, the *corridor variance swap* $h(s) = \mathbf{1}_I(s)$ where I is a possibly semi-infinite interval in $\mathbb{R}^+$, and the *gamma swap* in which $h(s) = s$. The reader can consult Gatheral (2006) for information about variance swaps and more extended discussion of the basic facts presented below. In this section we restrict the discussion to the vanilla variance swap; we return to the other contracts in Section 4.

In the classical approach, if we model S_t under a risk-neutral measure $\mathbb{Q}$ as[1] $S_t = F_t \exp(X_t - \frac{1}{2}\langle X\rangle_t)$, where F_t is the forward price (assumed to be continuous and of bounded variation) and X_t is a continuous martingale with quadratic variation process $\langle X\rangle_t$, then S_t is a continuous semimartingale and

$$\log \frac{S_{t_i}}{S_{t_{i-1}}} = (g(t_i) - g(t_{i-1})) + \left(X_{t_i} - X_{t_{i-1}}\right),$$

with $g(t) = \log(F_{t_i}) - \frac{1}{2}\langle X\rangle_{t_i}$. For $m = 1, 2\ldots$ let $\{s_i^m,\, i = 0, \ldots, k_m\}$ be the ordered set of stopping times in $[0, T]$ containing the times t_i together with times $\tau_0^m = 0$, $\tau_k^m = \inf\{t > \tau_{k-1}^m : |X_t - X_{\tau_{k-1}^m}| > 2^{-m}\}$ wherever these are smaller than T. If RV_T^m denotes the realized variance computed as in (1.1) but using the times s_i^m, then $RV_T^m \to \langle X\rangle_T = \langle\log S\rangle_T$ almost surely, see Rogers and Williams (2000), theorem IV.30.1. For this reason, most of the pricing literature on realized variance studies the continuous-time limit $\langle X\rangle_T$ rather than the finite sum (1.1) and we continue the tradition in this paper.

The key insight into the analysis of variance derivatives in the continuous limit was provided by Neuberger (1994). We outline the arguments here, see Section 4 for all the details. With S_t as above and f a C^2 function, it follows from the general Itô formula

[1] For example, $S_t = (e^{(r-q)t} S_0)(e^{\sigma W_t - \frac{1}{2}\sigma^2 t})$ in the Black–Scholes model, with the conventional notation.

that $Y_t = f(S_t)$ is a continuous semimartingale with $\mathrm{d}\langle Y\rangle_t = (f'(S_t))^2 \mathrm{d}\langle S\rangle_t$. In particular, with the above notation

$$\mathrm{d}(\log S_t) = \frac{1}{S_t}\mathrm{d}S_t - \frac{1}{2S_t^2}\mathrm{d}\langle S\rangle_t = \frac{1}{S_t}\mathrm{d}S_t - \frac{1}{2}\mathrm{d}\langle \log S\rangle_t,$$

so that

$$\langle \log S\rangle_T = 2\int_0^T \frac{1}{S_t}\mathrm{d}S_t - 2\log(S_T/S_0) \tag{1.2}$$

This shows that the realized variation is replicated by a portfolio consisting of a self-financing trading strategy that invests a constant \$$2D_T$ in the underlying asset S together with a European option whose exercise value at time T is $\lambda(S_T) = -2\log(S_T/F_T)$. Here D_T is the time-T discount factor. Assuming that the stochastic integral is a martingale, (1.2) shows that the risk-neutral value of the variance swap rate P_T^{RV} is

$$P_T^{\mathrm{RV}} = \mathbb{E}[\langle \log S\rangle_T] = -2\mathbb{E}[\log(S_T/F_T)] = 2\frac{1}{D_T}P_{\log}, \tag{1.3}$$

that is, the variance swap rate is equal to the forward value $P_{\log}/D_T$ of two "log contracts"—European options with exercise value $-\log(S_T/F_T)$, a convex function of S_T. (1.3) gives us a way to evaluate P_T^{RV} in any given model. The next step is to rephrase $P_{\log}$ in terms of call and put prices only.

Recall that for a convex function $f : \mathbb{R}^+ \to \mathbb{R}$ the recipe $f''(a, b] = f'_+(b) - f'_+(a)$ defines a positive measure $f''(\mathrm{d}x)$ on $\mathcal{B}(\mathbb{R}^+)$, equal to $f''(x)\mathrm{d}x$ if f is C^2. We then have the Taylor formula

$$f(x) = f(x_0) + f'_+(x_0)(x - x_0) + \int_0^{x_0}(y - x)^+\nu(\mathrm{d}y) + \int_{x_0}^{\infty}(x - y)^+ f''(\mathrm{d}y). \tag{1.4}$$

Applying this formula with $x = S_T$, $x_0 = F_T$ and $f(s) = \log(s/S_0)$, and combining with (1.2) gives

$$\begin{aligned}\frac{1}{2}\langle \log S\rangle_T = &\int_0^T \frac{1}{S_t}\mathrm{d}S_t - \log(F_T/S_0) + 1 - S_T/F_T \\ &+ \int_0^{F_T}\frac{(K - S_T)^+}{K^2}\mathrm{d}K + \int_{F_T}^{\infty}\frac{(S_T - K)^+}{K^2}\,\mathrm{d}K.\end{aligned} \tag{1.5}$$

Assuming that puts and calls are available for all strikes $K \in \mathbb{R}^+$ at traded prices P_K, C_K, (1.5) provides a perfect hedge for the realized variance in terms of self-financing dynamic trading in the underlying (the first four terms) and a static portfolio of puts and calls and hence uniquely specifies the variance swap rate P_T^{RV} as

$$P_T^{\mathrm{RV}} = \frac{2}{D_T}\int_0^{\infty}\frac{1}{K^2}\left(P_K\mathbf{1}_{(K\le F_T)} + C_K\mathbf{1}_{(K>F_T)}\right)\mathrm{d}K. \tag{1.6}$$

Our objective in this paper is to investigate the situation where, as in reality, puts and calls are available at only a finite number of strikes. In this case we cannot expect to get a unique value for P_T^{RV} as in (1.6). However, from (1.2) we expect to obtain arbitrage bounds on P_T^{RV} from bounds on the value of the log contract—a European option whose exercise value is the convex function $-\log(S_T/F_T)$. It will turn out that the weighted variance swaps we consider are associated in a similar way with other convex functions.

In Davis and Hobson (2007) conditions were stated under which a given set of prices $(P_1, \dots, P_n)$ for put options with strikes $(K_1, \dots, K_n)$ all maturing at the same time T is consistent with absence of arbitrage. Thus the first essential question we have to answer is: given a set of put prices consistent with absence of arbitrage and a convex function λ_T, what is the range of prices P_λ for a European option with exercise value $\lambda_T(S_T)$ such that this option does not introduce arbitrage when traded in addition to the existing puts? We answer this question in Section 3 following a statement of the standing assumptions and a summary of the Davis and Hobson (2007) results in Section 2.

The basic technique in Section 3 is to apply the duality theory of semi-infinite linear programming to obtain the values of the most expensive sub-replicating portfolio and the cheapest super-replicating portfolio, where the portfolios in question contain static positions in the put options and dynamic trading in the underlying asset. The results for the lower and upper bounds are given in Propositions 3.1 and 3.5, respectively. (Often, in particular in the case of plain vanilla variance swaps, the upper bound is infinite.) For completeness, we state and prove the fundamental Karlin-Isii duality theorem in an appendix, Appendix A. With these results in hand, the arbitrage conditions are stated and proved in Theorem 3.6.

In this first part of the paper—Sections 2 and 3—we are concerned exclusively with European options whose exercise value depends only on the asset price S_T at the common exercise time T. In this case the sample path properties of $\{S_t, t \in [0, T]\}$ play no role and the only relevant feature of a "model" is the marginal distribution of S_T. For this reason we make no assumptions about the sample paths. When we come to the second part of the paper, Sections 4 and 5, analyzing weighted variance swaps, then of course the sample path properties are of fundamental importance. Our minimal assumption is that $\{S_t, t \in [0, T]\}$ is continuous in t. However, this is not enough to make sense out of the problem. The connection between the variance swap and the "log contract," given at (1.2), is based on stochastic calculus, assuming that S_t is a continuous semimartingale. In our case the starting point is simply a set of prices. No model is provided, but we have to define what we mean by the continuous-time limit of the realized variance and by continuous-time trading. An appropriate framework is the pathwise stochastic calculus of Föllmer (1981) where the quadratic variation and an Itô-like integral are defined for a certain subset $\mathcal{Q}$ of the space of continuous functions $C[0, T]$. Then (1.2) holds for $S(\cdot) \in \mathcal{Q}$ and we have the connection we sought between variance swap and log contract. For weighted variance swaps the same connection exists, replacing "$-\log$" by some other convex function λ_T, as long as the latter is C^2. In the case of the corridor variance swap, however, λ''_T is discontinuous and we need to restrict ourselves to a smaller class of paths $\mathcal{L}$ for which the Itô formula is valid for functions in the Sobolev space $\mathcal{W}_2$. All of this is consistent with models in which S_t is a continuous semimartingale, because then $\mathbb{P}\{\omega : S(\cdot, \omega) \in \mathcal{L}\} = 1$. We discuss these matters in a separate appendix, Appendix B.

With these preliminaries in hand we formulate and prove, in Section 4, the main result of the paper, Theorem 4.3, giving conditions under which the quoted price of a weighted variance swap is consistent with the prices of existing calls and puts in the market, assuming the latter are arbitrage-free. When the conditions fail there is a weak arbitrage, and we identify strategies that realize it.

In mathematical finance it is generally the case that option bounds based on super- and sub-replication are of limited value because the gap is too wide. However, here we know that as the number of put options increases then in the limit there is only one price, namely, for a vanilla variance swap, the number P_T^{RV} of (1.6), so we may expect that in a market containing a reasonable number of liquidly traded puts the bounds will be tight.

In Section 5, we present data from the S&P 500 index options market which shows that our computed lower bounds are surprisingly close to the quoted variance swap prices.

This paper leaves many related questions unanswered. In particular we would like to know what happens when there is more than one exercise time T, and how the results are affected if we allow jumps in the price path. The latter question is considered in the small time-to-expiry limit in the recent paper Keller-Ressel and Muhle-Karbe (2013). A paper more in the spirit of ours, but taking a complementary approach, is Hobson and Klimmek (2011) (HK). In our paper the variance swap payoff is defined by the continuous-time limit, and we assume that prices of a finite number of traded put and/or call prices are known. By contrast, HK deal with the real variance swap contract (i.e., discrete sampling) but assume that put and call prices are known for all strikes $K \in \mathbb{R}^+$, so in practice the results will depend on how the traded prices are interpolated and extrapolated to create the whole volatility surface. Importantly, HK allow for jumps in the price process; this is indeed the main focus of their work. The price bounds they obtain are sharp in the continuous-sampling limit. Combining our methods with those of HK would be an interesting, and possibly quite challenging, direction for future research, see Remark 4.6 below.

Exact pricing formulas for a wide variety of contracts on discretely sampled realized variance, in a general Lévy process setting, are provided in Crosby and Davis (2011).

2. PROBLEM FORMULATION

Let $(S_t, t \in [0, T])$ be the price of a traded financial asset, which is assumed nonnegative: $S_t \in \mathbb{R}^+ = [0, \infty)$. In addition to this underlying asset, various derivative securities as detailed below are also traded. All of these derivatives are European contracts maturing at the same time T. The present time is 0. We make the following standing assumptions as in Davis and Hobson (2007):

(i) The market is frictionless: assets can be traded in arbitrary amounts, short or long, without transaction costs, and the interest rates for borrowing and lending are the same.

(ii) There is no interest rate volatility. We denote by D_t the market discount factor for time t, i.e., the price at time 0 of a zero-coupon bond maturing at t.

(iii) There is a uniquely defined forward price F_t for delivery of one unit of the asset at time t. This will be the case if the asset pays no dividends or if, for example, it has a deterministic dividend yield. We let Γ_t denote the number of shares that will be owned at time t if dividend income is re-invested in shares; then $F_0 = S_0$ and $F_t = S_0/(D_t\Gamma_t)$ for $t > 0$.

We suppose that n put options with strikes $0 < K_1 < \cdots < K_n$ and maturity T are traded, at time-0 prices $P_1, \ldots, P_n$. In Davis and Hobson (2007), the arbitrage relations among these prices were investigated. The facts are as follows. A *static portfolio X* is simply a linear combination of the traded assets, with weights $\pi_1, \ldots, \pi_n$, ϕ, ψ on the options, on the underlying asset and on cash respectively, it being assumed that dividend income is re-invested in shares. The value of the portfolio at maturity is

$$X_T = \sum_{i=1}^{n} \pi_i [K_i - S_T]^+ + \phi \Gamma_T S_T + \psi D_T^{-1} \tag{2.1}$$

and the set-up cost at time 0 is

$$X_0 = \sum_{i=1}^{n} \pi_i P_i + \phi S_0 + \psi. \tag{2.2}$$

Note that it does not make any sense a priori to speak about X_t—the value of the portfolio at any intermediate time—as this is not determined by the market input and we do not assume the options are quoted in the market at intermediate dates $t \in (0, T)$. To make sense to statements like "X_T is nonnegative" we need to specify the universe of scenarios we consider. Later, in Section 4, this will mean the space of possible paths $(S_t\colon t \le T)$. However here, because we only allow static trading as in (2.1), we essentially look at a one-period model and we just need to specify the possible range of values for S_T and we suppose that S_T may take any value in $[0, \infty)$.

A *model* $\mathcal{M}$ is a filtered probability space $(\Omega, \mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{P})$ together with a positive $\mathbb{F}$-adapted process $(S_t)_{t\in[0,T]}$ such that S_0 coincides with the given time-0 asset price. Given market prices of options, we say that $\mathcal{M}$ is a *market model* for these options if $M_t = S_t/F_t$ is an $\mathbb{F}$-martingale (in particular, S_t is integrable) and the market prices equal to the $\mathbb{P}$–expectations of discounted payoffs. In particular, $\mathcal{M}$ is a market model for put options if $P_i = D_T\mathbb{E}[K_i - S_T]^+$ for $i = 1, \ldots, n$. We simply say that $\mathcal{M}$ is a market model if the set of market options with given prices is clear from the context. It follows that in a market model we have joint dynamics on $[0, T]$ of all assets' prices such that the initial prices agree with the market quotes and the discounted prices are martingales. By the (easy part of the) First Fundamental Theorem of Asset Pricing (Delbaen and Schachermayer 1994) the dynamics do not admit arbitrage.

Our main interest is in the existence of a market model and we want to characterize it in terms of the given market quoted prices. This requires notions of arbitrage in absence of a model. We say that there is *model-independent arbitrage* if one can form a portfolio with a negative setup cost and a nonnegative payoff. There is *weak arbitrage* if for any model there is a static portfolio $(\pi_1, \ldots, \pi_n, \phi, \psi)$ such that $X_0 \le 0$ but $\mathbb{P}(X_T \ge 0) = 1$ and $\mathbb{P}[X_T > 0] > 0$. In particular, a model independent arbitrage is a special case of weak arbitrage, as in Cox and Obłój (2011b). Throughout the paper, we say that prices are *consistent with absence of arbitrage* when they do not admit a weak arbitrage.

It is convenient at this point to move to normalized units. If for $i = 1, \ldots, n$ we define

$$p_i = \frac{P_i}{D_T F_T}, \qquad k_i = \frac{K_i}{F_T}, \tag{2.3}$$

then, in a market model, we have $P_i = D_T F_T \mathbb{E}[K_i/F_T - M_T]^+$ and hence

$$p_i = \mathbb{E}[k_i - M_T]^+.$$

Also, we can harmlessly introduce another put option with strike $k_0 = 0$ and price $p_0 = 0$. Let $\underline{n} = \max\{i : p_i = 0\}, \overline{n} = \inf\{i : p_i = k_i - 1\}(= +\infty \text{ if } \{\cdots\} = \emptyset)$ and $\mathbf{K} = [k_{\underline{n}}, k_{\overline{n}}]$, where it is hereafter understood as $\mathbf{K} = [k_{\underline{n}}, \infty)$ if $\overline{n} = \infty$.

The main result of Davis and Hobson (2007, thm. 3.1), which we now recall, says that there is no weak arbitrage if and only if there exists a market model. The conditions are illustrated in Figure 2.1.

PROPOSITION 2.1. *Let $r : [0, k_n] \to \mathbb{R}^+$ be the linear interpolant of the points (k_i, p_i), $i = 0, \ldots, n$. Then there exists a market model if and only if $r(\cdot)$ is a nonnegative, convex,*

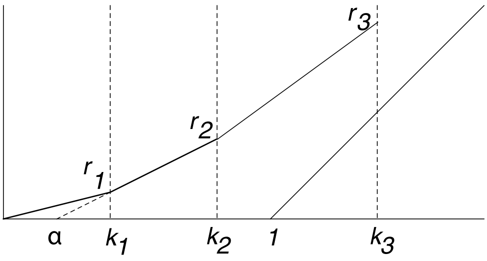

FIGURE 2.1. Normalized put prices (k_i, p_i) consistent with absence of arbitrage. An additional put with price $r = 0$ and strike $k \in [0, \alpha]$ does not introduce arbitrage.

increasing function such that $r(0) = 0$, $r(k) \geq [k - 1]^+$ *and* $r'_-(k_{\bar{n} \wedge n}) < 1$, *where* r'_- *denotes the left-hand derivative. If these conditions hold except that* $\bar{n} = \infty$ *and* $r'_-(k_n) = 1$ *then there is weak arbitrage. Otherwise there is a model-independent arbitrage. Furthermore, if a market model exists then one may choose it so that the distribution of* S_T *has finite support.*

REMARKS.

(i) In Davis and Hobson (2007), the case of call options was studied. The result stated here follows by put-call parity, valid in view of our frictionless markets assumption. The normalized call price is $c_i = p_i + 1 - k_i$.

(ii) Of course, no put option would be quoted at zero price, so in applications $\underline{n} = 0$ always. As will be seen below, it is useful for analysis to include the artificial case $\underline{n} > 0$.

Throughout the rest of the paper, we assume that *the put option prices* $(P_1, \ldots, P_n)$ *do not admit a weak arbitrage and hence there exists a market model consistent with the put prices*. Let $\mathcal{M}$ be a market model and let μ be the distribution of M_T in this model. Then μ satisfies

$$\int_{\mathbb{R}^+} 1\, \mu(\mathrm{d}x) = 1 \tag{2.4a}$$

$$\int_{\mathbb{R}^+} x\, \mu(\mathrm{d}x) = 1 \tag{2.4b}$$

$$\int_{\mathbb{R}^+} [k_i - x]^+ \mu(\mathrm{d}x) = p_i, \quad i = 1, \ldots, n. \tag{2.4c}$$

Conversely, given a probability measure μ on $\mathbb{R}^+$ which satisfies the above we can construct a market model $\mathcal{M}$ such that μ is the distribution of M_T. For example, let $(\Omega, \mathbb{F}, \mathbb{P}, (W_t)_{t \in \mathbb{R}^+})$ be the Wiener space. By the Skorokhod embedding theorem (cf. Obłój 2004), there is a stopping time τ such that $W_\tau \sim \mu$ and $(W_{\tau \wedge t})$ is a uniformly integrable martingale. It follows that we can put $M_t = 1 + W_{\tau \wedge (t/(T-t))}$ for $t \in [0, T)$. This argument shows that the search for a market model reduces to a search for a measure μ satisfying 2.4. We will denote by $\mathbb{M}_P$ the set of measures μ satisfying the conditions 2.4.

LEMMA 2.2. *For any* $\mu \in \mathbb{M}_P$, $\mu(\mathbb{R}^+ \backslash \mathbf{K}) = 0$.

Proof. That $\mu[0, k_{\underline{n}}) = 0$ when $\underline{n} > 0$ follows from (2.4c) with $i = \underline{n}$. When $\bar{n} \leq n$ we have $c_{\bar{n}} = 0$, i.e., there is a free call option with strike $k_{\bar{n}}$ and we conclude that $\mu(k_{\bar{n}}, \infty) = 0$. □

The question we wish to address is whether, when prices of additional options are quoted, consistency with absence of arbitrage is maintained. As discussed in Section 1, we start by considering the case where one extra option is included, a European option maturing at T with convex payoff.

3. HEDGING CONVEX PAYOFFS

Suppose that, in addition to the n put options, a European option is offered at price P_λ at time 0, with exercise value $\lambda_T(S_T)$ at time T, where λ_T is a convex function. We can obtain lower and upper bounds on the price of λ_T by constructing sub-replicating and super-replicating static portfolios in the other traded assets. These bounds are given in Sections 3.1 and 3.2, respectively and are combined in Section 3.3 to obtain the arbitrage conditions on the price P_λ.

We work in normalized units throughout, that is, the static portfolios have time-T values that are linear combinations of cash, underlying M_T and option exercise values $[k_i - M_T]^+$. The prices of units of these components at time 0 are D_T, D_T and $D_T p_i$, respectively, where a unit of cash is \$1. Indeed, to price M_T observe that \$1 invested in the underlying at time 0 yields $\Gamma_T S_T / S_0 = S_T / F_T D_T = M_T / D_T$ at time T. To achieve a consistent normalization for λ_T we define the convex function λ as

$$\lambda(x) = \frac{1}{F_T} \lambda_T(F_T x). \tag{3.1}$$

In a market model $\mathcal{M}$ we have $P_\lambda = D_T \mathbb{E}[\lambda_T(S_T)] = D_T F_T \mathbb{E}[\lambda(M_T)]$, so the normalized price is

$$p_\lambda = \frac{P_\lambda}{D_T F_T} = \mathbb{E}[\lambda(M_T)]$$

and the cost for delivering a payoff $\lambda(M_T)$ is $D_T p_\lambda$.

3.1. Lower Bound

A sub-replicating portfolio is a static portfolio formed at time 0 such that its value at time T is majorized by $\lambda(M_T)$ for all values of M_T. Obviously, a necessary condition for absence of model independent arbitrage is that $D_T p_\lambda$ be not less than the set-up cost of any sub-replicating portfolio. It turns out that the options k_i with $i \leq \underline{n}$ or $i \geq \bar{n}$ are redundant, so the assets in the portfolio are indexed by $k = 1, \ldots, m$ where

$$m = (n+1) \wedge \bar{n} - \underline{n} + 1$$

and the time-T values of these assets, as functions of $x = M_T$ are

$$\begin{aligned} a_1(x) &= 1 \quad \text{(Cash)} \\ a_2(x) &= x \quad \text{(Underlying)} \\ a_{i+2}(x) &= [k_{\underline{n}+i} - x]^+, \quad i = 1, \ldots, m-2 \quad \text{(Options)}. \end{aligned} \tag{3.2}$$

We let $\mathbf{a}(x)$ be the m-vector with components $a_k(x)$. Note that $a_m(x)$ is equal to $[k_{\bar{n}-1} - x]^+$ if $\bar{n} \leq n$ and to $[k_n - x]^+$ otherwise. The set-up costs for the components in (3.2), as observed above, are D_T, D_T and $D_T p_{\underline{n}+i}$, respectively. The corresponding *forward* prices are as in 2.4:

$$\begin{aligned} b_1 &= 1 \\ b_2 &= 1 \\ b_{i+2} &= p_{\underline{n}+i}, \quad i = 1, \dots, m-2. \end{aligned} \tag{3.3}$$

We let $\mathbf{b}$ denote the m-vector of the forward prices. A static portfolio is defined by a vector $\mathbf{y}$ whose kth component is the number of units of the kth asset in the portfolio. The forward set-up cost is $\mathbf{y}^T\mathbf{b}$ and the value at T is $\mathbf{y}^T\mathbf{a}(M_T)$.

With this notation, the problem of determining the most expensive sub-replicating portfolio is equivalent to solving the (primal) semi-infinite linear program

$$P_{\mathrm{LB}}: \quad \sup_{\mathbf{y}\in\mathbb{R}^m} \mathbf{y}^T\mathbf{b} \quad \text{subject to} \quad \mathbf{y}^T\mathbf{a}(x) \leq \lambda(x)\ \forall x \in \mathbf{K}.$$

The constraints are enforced only for $x \in \mathbf{K}$. If $\underline{n} > 0$ [$\bar{n} \leq n$] we have a free put with strike $k_{\underline{n}}$ [call with strike $k_{\bar{n}}$] and, because λ is convex, we can extend the sub-replicating portfolio to all of $\mathbb{R}^+$ at no cost.

The key result here is the basic duality theorem of semi-infinite linear programming, due to Isii (1960) and Karlin, see Karlin and Studden (1966). This theorem, stated as Theorem A.1, and its proof are given in Appendix A. The dual program corresponding to P_{LB} is

$$D_{\mathrm{LB}}: \quad \inf_{\mu\in\mathbb{M}} \int_{\mathbf{K}} \lambda(x)\mu(\mathrm{d}x) \quad \text{subject to} \quad \int_{\mathbf{K}} \mathbf{a}(x)\mu(\mathrm{d}x) = \mathbf{b}^0,$$

where $\mathbb{M}$ is the set of Borel measures such that each a_i is integrable. The constraints in D_{LB} can be expressed as μ satisfying 2.4 for $\underline{n} < i < \bar{n}$. This is simply equivalent to $\mu \in \mathbb{M}_P$ because, as shown in Lemma 2.2, any $\mu \in \mathbb{M}_P$ has support in $\mathbf{K}$. Let V_P^L and V_D^L be the values of the primal and dual problems, respectively. It is a general and easily proved fact that $V_P^L \leq V_D^L$. The "duality gap" is $V_D^L - V_P^L$. The Karlin–Isii theorem gives conditions under which there is no duality gap and we have existence in P_{LB}.

PROPOSITION 3.1. *We suppose as above that $\lambda(x)$ is a convex function on $\mathbb{R}^+$, finite for all $x > 0$, and that (k_i, p_i) is a set of normalized put option strike and price pairs which do not admit a weak arbitrage. If $\lambda(x)$ is unbounded as $x \to 0$ and $\underline{n} = 0$ then we further assume that $p_1/k_1 < p_2/k_2$. Then $V_D^L = V_P^L$ and there exists a maximizing vector $\hat{\mathbf{y}}$. The most expensive sub-replicating portfolio of a European option with payoff $\lambda_T(S_T)$ at maturity T is the static portfolio $X^\dagger$ as in (2.1) with weights $\psi^\dagger = F_T D_T \hat{\mathbf{y}}_1$, $\phi^\dagger = \hat{\mathbf{y}}_2/\Gamma_T$, $\pi^\dagger_{\underline{n}+i} = \hat{\mathbf{y}}_{2+i}$ for $i = 1, \dots, m-2$ and $\pi_i^\dagger = 0$ otherwise. For this portfolio, $X_0^\dagger = D_T F_T V_D^L$.*

If there is existence in the dual problem D_{LB} then there is an optimal measure $\mu^\dagger$ which is a finite linear combination of Dirac measures $\mu^\dagger = \sum_{j=1}^m w_j \delta_{x_j}(\mathrm{d}x)$ such that each interval $[k_j, k_{j+1})$ contains at most one point x_j. For this measure

$$\mu^\dagger(\{x\}) > 0 \;\Rightarrow\; X_T^\dagger\big|_{S_T = F_T x} = \sum_{i=1}^n \pi_i^\dagger [K_i - F_T x]^+ + \phi^\dagger \Gamma_T F_T x + \psi^\dagger D_T^{-1} = \lambda_T(F_T x). \tag{3.4}$$

Proof. The first part of the proposition is an application of Theorem A.1. The primal problem P_{LB} is feasible because any support line corresponds to a portfolio (containing no options). The functions $a_1, \dots, a_m$ are linearly independent. Recall from Proposition 2.1 that if $\{(k_i, p_i)\}$ do not admit weak arbitrage then there is a measure μ satisfying the conditions 2.4 and such that μ is a finite weighted sum of Dirac measures. It follows that $\int_{\mathbb{R}^+} |\lambda(x)|\mu(\mathrm{d}x) < \infty$ unless one of the Dirac measures is placed at $x = 0$ and λ is unbounded at zero. If $\underline{n} > 0$ then there is no mass on the interval $[0, \underline{n})$, hence none at zero. When $\underline{n} = 0$ we always have $p_1/k_1 \leq p_2/k_2$. If $p_1/k_1 = p_2/k_2$ then the payoff $[k_2 - M_T]^+ - k_2[k_1 - M_T]^+/k_1$ has null cost and is strictly positive on $(0, k_2)$. Because $p_1 > 0$ there must be some mass to the left of k_1, and this mass must be placed at 0, else there is an arbitrage opportunity. But then $\int \lambda d\mu = +\infty$ and $V_D^L = +\infty$. The condition in the proposition excludes this case. In every other case there is a realizing measure μ such that $\mu(\{0\}) = 0$. Indeed, if $p_1/k_1 < p_2/k_2$ then the extended set of put prices $\{(k, 0), (k_1, p_1), \dots, (k_n, p_n)\}$ is consistent with absence of arbitrage if $k \in [0, \alpha]$, where $\alpha = (k_1 p_2 - k_2 p_1)/(p_2 - p_1)$ (see Figure 2.1). Any model realizing these prices puts weight 0 on the interval $[0, k)$. Thus V_D^L is finite under the conditions we have stated. It remains to verify that the vector $\mathbf{b}$ belongs to the interior of the moment cone M_m defined at (A.1). For this, it suffices to note that for all i such that $k_i \in (k_{\underline{n}}, k_{\bar{n}})$ it holds that $[k_i - 1]^+ < p_i < k_i$, and so the condition is satisfied. We now conclude from Theorem A.1 that $V_P^L = V_D^L$ and that we have existence in the primal problem. The expressions for $\psi^\dagger$ etc. follow from the relationships (2.3) between normalized and un-normalized prices.

Assume now that the dual problem has a solution. Any optimal measure $\mu^\dagger \in \mathbb{M}_P$ satisfies

$$\int_{\mathbf{K}} \lambda(x)\mu^\dagger(\mathrm{d}x) = \inf_{\mu \in \mathbb{M}_P} \left\{ \int_{\mathbf{K}} \lambda(x)\mu(\mathrm{d}x) \right\}.$$

Recall $\mathbf{K} = [k_{\underline{n}}, k_{\bar{n}}]$ and partition $\mathbf{K}$ into intervals $I_{\underline{n}+1}, \dots, I_{\bar{n} \wedge n+1}$ defined by

$$I_i = [k_{i-1}, k_i) \quad \text{for} \quad i = \underline{n} + 1, \dots \bar{n} \wedge n \quad \text{and} \quad I_{\bar{n} \wedge n+1} = [k_{\bar{n} \wedge n}, k_{\bar{n}}),$$

so $I_{\bar{n} \wedge n+1} = \emptyset$ if $\bar{n} \leq n$. Lemma 3.2 below asserts that we may take $\mu^\dagger$ atomic with at most one atom in each of the intervals I_i. By definition the optimal subhedging portfolio $X^\dagger$ satisfies $X_T^\dagger \leq \lambda_T(S_T)$ while our duality result shows that the $\mu^\dagger$ expectations of these random variables coincide. Hence $X_T^\dagger = \lambda_T(S_T)$ a.s. for $M_T = S_T/F_T$ distributed according to $\mu^\dagger$, and (3.4) follows. □

LEMMA 3.2. *Let $\mu \in \mathbb{M}_P$ and suppose $\int_{\mathbf{K}} |\lambda(x)|\mu(\mathrm{d}x) < \infty$. Define*

$$\mathcal{I}_\mu = \{i \leq \bar{n} \wedge n + 1 | \mu(I_i) > 0\}. \tag{3.5}$$

Now let μ' be the measure

$$\mu' = \sum_{i \in \mathcal{I}_\mu} \mu(I_i)\delta_{x_i},$$

in which δ_x denotes the Dirac measure at x, and for an index $i \in \mathcal{I}$, $x_i = \frac{\int_{I_i} x d\mu(x)}{\mu(I_i)}$. Then $\mu' \in \mathbb{M}_P$ and

$$\int_{\mathbf{K}} \lambda(x)\mu'(\mathrm{d}x) \leq \int_{\mathbf{K}} \lambda(x)\mu(\mathrm{d}x). \tag{3.6}$$

Proof of lemma: The inequality (3.6) follows from the conditional Jensen inequality. A direct computation shows that μ' satisfies 2.4, so that $\mu' \in \mathbb{M}_P$. □

It remains now to understand when there is existence in the dual problem. We exclude the case when λ_T is affine on some $[z, \infty)$ which is tedious. We characterize the existence of a dual minimizer in terms of properties of the solution to the primal problem and also present a set of sufficient conditions. Of the conditions given, (i) would never be encountered in practice (it implies the existence of free call options) and (iii), (iv) depend only on the function λ_T and not on the put prices P_i. The examples presented in Section 3.1.1 show that if these conditions fail there may still be existence, but this will now depend on the P_i. Condition (ii) is closer to being necessary and sufficient, but is not stated in terms of the basic data of the problem.

PROPOSITION 3.3. *Assume λ_T is not affine on some half-line $[z, \infty)$. Then,in the setup of Proposition 3.1, the existence of a minimizer in the dual problem D_{LB} fails if and only if*

$$\bar{n} = \infty \quad \text{and} \quad X_T^\dagger\big|_{S_T=s} = \phi^\dagger \Gamma_T s + \psi^\dagger D_T^{-1} < \lambda_T(s), \quad \textit{for all } s \geq K_n. \tag{3.7}$$

In particular, each of the following is a sufficient condition for existence of a minimizer in D_{LB}:

(i) $\bar{n} < \infty$;

(ii) *we have*

$$X_T^\dagger\big|_{S_T=K_n} = \phi^\dagger \Gamma_T K_n + \psi^\dagger D_T^{-1} < \lambda_T(K_n) \quad \textit{and} \quad \lim_{s\to\infty} \lambda_T(s) - \phi^\dagger \Gamma_T s = \infty; \tag{3.8}$$

(iii) *for any $y < \lambda_T(K_n)$ there is some $x > K_n$ such that the point (K_n, y) lies on a support line to λ_T at x;*

(iv) *λ_T satisfies*

$$\int_0^\infty x\lambda_T''(\mathrm{d}x) = +\infty. \tag{3.9}$$

Proof. We consider two cases.

Case 1: $\bar{n} \leq n$, i.e., condition (i) holds. In this case the support of any measure $\mu \in \mathbb{M}_P$ is contained in the finite union $I_{\underline{n}+1} \cup \cdots \cup I_{\bar{n}}$ of bounded intervals, and $\sum_{i=\underline{n}+1}^{\bar{n}} \mu(I_i) = 1$. Further, $p_{\bar{n}-1} > k_{\bar{n}} - 1$ and (2.4b) together imply that $\mu(I_{\bar{n}}) > 0$. Let $\mu_j, j = 1, 2, \ldots$ be a sequence of measures such that

$$\int_{\mathbf{K}} \lambda(x)\mu_j(\mathrm{d}x) \to \inf_{\mu\in\mathbb{M}_P} \left\{ \int_{\mathbf{K}} \lambda(x)\mu(\mathrm{d}x) \right\} \quad \text{as } j \to \infty.$$

By Lemma 3.2, we may and do assume that each μ_j is atomic with at most one atom per interval. We denote $w_j^i = \mu_j(I_i)$ and let x_j^i denote the location of the atom in I_i. For definiteness, let $x_j^i = \Delta$, where Δ is some isolated point, if μ_j has no atom in I_i, i.e., if $w_j^i = 0$. Let A be the set of indices i such that x_j^i converges to Δ as $j \to \infty$, i.e., $x_j^i \neq \Delta$ for only finitely many j, and let B be the complementary set of indices. Then there exists j^* such that $\sum_{i\in B} w_j^i = 1$ for $j > j^*$. Because the w_j^i and x_j^i are contained in compact intervals there exists a subsequence $j_k, k = 1, 2 \ldots$ and points w_*^i, x_*^i such that $w_{j_k}^i \to w_*^i$

and $x^i_{j_k} \to x^i_*$ as $k \to \infty$. It is clear that $\sum_{i\in B} w^i_* = 1$ and that

$$V^L_D = \lim_{k\to\infty} \int_{\mathbf{K}} \lambda(x)\mu_{j_k}(\mathrm{d}x) = \int_{\mathbf{K}} \lambda(x)\mu^\dagger(\mathrm{d}x)$$

where $\mu^\dagger(\mathrm{d}x) = \sum_{i\in B} w^i_* \delta_{x^i_*}(\mathrm{d}x)$. Similarly, the integrals of 1, x and $[k_i - x]^+$ converge, so $\mu^\dagger \in \mathbb{M}_P$. Finally, because the intervals I_i are open on the right, it is possible that $x^i_* \in I_{i+1}$. If also $x^{i+1}_* \in I_{i+1}$ we can invoke Lemma 3.2 to conclude that this 2-point distribution in I_{i+1} can be replaced by a 1-point distribution without increasing the integral. Thus $\mu^\dagger$ retains the property of being an atomic measure with at most one atom per interval. We have existence in the dual problem and, by the arguments above, $X^\dagger_T = \lambda_T(S_T)$ $\mathrm{d}\mu^\dagger$-a.s.

Case 2: $\bar{n} = \infty$. Let $X^\dagger_T(s)$ denote the payoff of the portfolio $X^\dagger$ at time T when $S_T = s$. Suppose that a minimizer $\mu^\dagger$ exists in D_{LB}, which we may take as atomic. Because $\bar{n} = \infty$ there exists $x \in (k_n, \infty)$ such that $\mu^\dagger(\{x\}) > 0$ and hence, by (3.4), $X^\dagger_T(F_T x) = \lambda(F_T x)$. This shows that (3.7) fails because $F_T x > K_n$.

It remains to show the converse: that if $\bar{n} = \infty$ but (3.7) fails then a minimizer $\mu^\dagger$ exists. Given our assumption on λ_T, if (3.7) fails then $K^\dagger := \sup\{s \geq 0 : X^\dagger_T(s) = \lambda_T(s)\} \in [K_n, \infty)$. We continue with analysis similar to Case 1 above. Here we have intervals $I_{\underline{n}}, \ldots, I_n$ covering $\mathbf{K}_n = [k_{\underline{n}}, k_n)$, but also a further interval $I_{n+1} = [k_n, \infty)$ which does not necessarily have zero mass. Because I_{n+1} is unbounded, the argument above needs some modification. We take a minimizing sequence of point mass measures μ_j as in Case 1. If x^{n+1}_j converge (on a subsequence) to a finite x^{n+1}_* then we can restrict our attention to a compact $[0, x^{n+1}_* + 1]$ and everything works as in Case 1 above. Suppose to the contrary that $\liminf x^{n+1}_j = \infty$, as $j \to \infty$. We first apply the arguments in Case 1 to the sequence $\tilde{\mu}_j$ of restrictions of μ_j to $\mathbf{K}_n$. Everything is the same as above except that now $\sum_i w^i_j \leq 1$. A subsequence converges to a sub-probability measure $\tilde{\mu}^\dagger$ on $\mathbf{K}_n$, equal to a weighted sum of Dirac measures as above. Define $\mu^\dagger = \tilde{\mu}^\dagger$ if $\iota_1 \equiv \tilde{\mu}^\dagger(\mathbf{K}_n) = 1$ and otherwise $\mu^\dagger = \tilde{\mu}^\dagger + (1 - \iota_1)\delta_x$, where $x \in I_{n+1}$ is to be determined. Whatever the value of x, $\mu^\dagger$ satisfies conditions (2.4a) and (2.4c) (the put values depend only on $\tilde{\mu}^\dagger$).

Because $\mu_j \in \mathbb{M}_P$, $\sum_{i=1}^{n+1} w^i_j x^i_j = 1$ and in particular

$$1 - \sum_{i=1}^{n} w^i_j x^i_j \geq w^{n+1}_j k_n.$$

Taking the limit along the subsequence we conclude that

$$1 - \sum_{i=1}^{n} w^i_* x^i_* \geq k_n \left(1 - \sum_{i=1}^{n} w^i_*\right) = k_n(1 - \iota_1). \tag{3.10}$$

By the convergence argument of Case 1, $\iota_2 \equiv \int_0^{k_n} x\mu^\dagger(\mathrm{d}x) \leq 1$. If $\iota_1 < 1$ we have only to choose $x = (1 - \iota_2)/(1 - \iota_1)$ to ensure that the "forward" condition (2.4b) is also satisfied. The inequality (3.10) guarantees that $x \geq k_n$. Thus $\mu^\dagger \in \mathbb{M}_P$ and, because $\bar{n} = \infty$, $x > k_n$ and hence $K^\dagger > K_n$.

We now show that the complementary case $\iota_1 = 1$ contradicts $K^\dagger \in [K_n, \infty)$. Indeed, if $\iota_1 = 1$ we have, because $\bar{n} = \infty$,

$$k_n - 1 < p_n = \int_0^\infty [k_n - x]^+ \mu^\dagger(\mathrm{d}x) = \int_0^{k_n} (k_n - x)\mu^\dagger(\mathrm{d}x) = k_n - \sum_{i=1}^{n} w^i_* x^i_* .$$

Observe that then

$$(3.11)\qquad w_j^{n+1}x_j^{n+1} = 1 - \sum_{i=1}^{n} w_j^i x_j^i \xrightarrow[j\to\infty]{} 1 - \sum_{i=1}^{n} w_*^i x_*^i = p_n - k_n + 1 > 0.$$

Because $K^\dagger \in [K_n, \infty)$, taking $\gamma = \lambda_T'(K^\dagger + 1) - X_T^{\dagger\,'}(K^\dagger + 1) = \lambda_T'(K^\dagger + 1) - \phi^\dagger \Gamma_T > 0$, we have

$$(3.12)\qquad \lambda_T(s) - X_T^\dagger(s) \geq \gamma(s - K^\dagger + 1), \quad \forall s \geq K^\dagger + 1.$$

Define now a new function $\tilde{\lambda}_T$ by

$$\tilde{\lambda}_T(s) = \lambda_T(s)\mathbf{1}_{s\leq K^\dagger} + \left(\phi^\dagger \Gamma_T s + \psi^\dagger D_T^{-1}\right)\mathbf{1}_{s>K^\dagger}$$

so that we have $X^\dagger \leq \tilde{\lambda}_T(S_T) \leq \lambda_T(S_T)$ and, by definition, for $S_T > K^\dagger$ we have $X^\dagger = \tilde{\lambda}_T(S_T) < \lambda_T(S_T)$. It follows that $X^\dagger$ is also the most expensive subreplicating portfolio for $\tilde{\lambda}_T(s)$ and hence the primal and dual problems for λ and for $\tilde{\lambda}(x) = \frac{1}{F_T}\tilde{\lambda}_T(F_T x)$ have all the same value. Writing this explicitly and using (3.12) and(3.11) gives:

$$\begin{aligned} 0 &= \lim_{j\to\infty} \int_0^\infty (\lambda(x) - \tilde{\lambda}(x))\mu_j(\mathrm{d}x) = \lim_{j\to\infty} \left(\lambda\left(x_j^{n+1}\right) - \tilde{\lambda}\left(x_j^{n+1}\right)\right) w_j^{n+1} \\ &\geq \lim_{j\to\infty} \gamma\left(x_j^{n+1} - K^\dagger - 1\right) w_j^{n+1} = \gamma(p_n - k_n + 1) > 0, \end{aligned}$$

a contradiction.

We turn to showing that each of (*i*)–(*iv*) is sufficient for existence of a dual minimizer $\mu^\dagger$.[2] Obviously $\bar{n} < \infty$ is sufficient as observed above. We now show that either of (ii) or (iii) implies that $K^\dagger \in [K_n, \infty)$ and that (3.12) holds. In the light of the arguments above, this will be sufficient. $X_T^\dagger(s)$ is linear on $[K_n, \infty)$ and λ_T is convex hence the difference of the two either converges to a constant or diverges to infinity. In the latter case the difference grows quicker than a linear function, more precisely (3.12) holds. The former case is explicitly excluded in (ii) and is contradictory with (iii) as the line $\ell(s) := \phi^\dagger \Gamma_T s + \psi^\dagger D_T^{-1} + \lim_{u\to\infty}(\lambda_T(u) - X_T^\dagger(u))$ is asymptotically tangential to $\lambda_T(s)$ as $s \to \infty$ and hence the condition in (iii) is violated for any $y < \ell(K_n)$. In particular, it follows that $K^\dagger < K_n$ then we can add to $X^\dagger$ a positive number of call options with strike K_n and obtain a subreplicating portfolio with an initial cost strictly larger (recall that $\bar{n} = \infty$) than $X^\dagger$ which contradicts the optimality of $X^\dagger$. This completes the proof that either (ii) or (iii) is sufficient for existence.

Finally, we argue that (iii) and (iv) are equivalent. In (iii) the point x satisfies

$$\lambda_T(x) + \xi(K_n - x) = y, \qquad \text{for some} \qquad \xi \in [\lambda_T'(x-), \lambda_T'(x+)].$$

This equation has a solution x for all $y < \lambda_T(K_n)$ if and only if $\lim_{x\to\infty} \lambda_T(x) - x\lambda_T'(x) = -\infty$. The equivalence with (iv) follows integrating by parts

$$\int_{K_n}^\infty x\lambda_T''(\mathrm{d}x) = (x\lambda_T'(x) - \lambda_T(x))|_{K_n}^\infty.$$

□

[2] In fact for this part of the Proposition, we do not need to impose any additional conditions on λ_T apart from convexity.

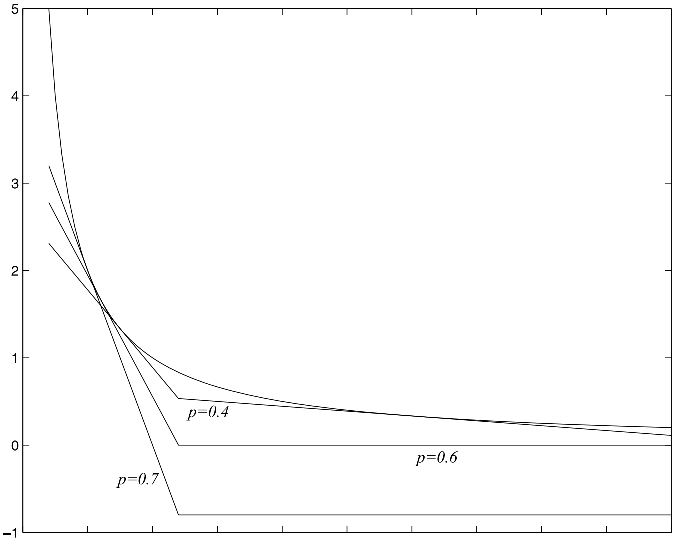

FIGURE 3.1. Most expensive subhedging portfolios for $\lambda(x) = 1/x$ given a single put option with strike 1.2 and prices $p = 0.4, 0.6, 0.7$.

TABLE 3.1
Data for Figure 2.1

p	x_0	x_1	value	ψ	ϕ	w_0	w_1
0.4	0.75	3	1.2222	0.6667	−0.1111	0.8889	0.1111
0.6	0.6	-	1.6667	0	0	1.00	-
0.7	0.5	-	2.00	−0.8	0	1.00	-

3.1.1. Examples with one put option. We consider now examples in which just one put option price is specified. We illustrate different cases when existence in the dual problem holds or fails. For simplicity assume all prices are normalized, i.e., $D_T = F_T = 1$, and we are given only a single put option with strike $k = 1.2$. The convex function takes the form $\lambda(x) = 1/x + ax^b$. Computation of the most expensive sub-hedging portfolio can be done by a simple search procedure.

Consider first the case $a = 0$, so that $\lambda(x) = 1/x$. The results are shown in Figure 3.1 with data shown in Table 3.1: p is the put price, x_0, x_1 the points of tangency, w_0, w_1 the implied probability weights on x_0, x_1 and ψ, ϕ the units of, respectively, cash and forward in the portfolio.

$p = 0.4$ is a “regular” case: we have tangent lines at x_0, x_1 and the solution to the dual problem puts weights 8/9, 1/9 respectively on these points. As p increases it is advantageous to include more puts in the portfolio, so x_1 increases. At $p = 0.6$ we reach a boundary case where $\psi = \phi = 0$, and the put is correctly priced by the Dirac measure with weight 1 at $x_0 = 0.6 = k/2$. Obviously, this measure does not correctly price the forward, but it does correctly price the portfolio because $\phi = 0$. When $p > 0.6$, the only way to increase the put component further is to take $\psi < 0$ (and then clearly

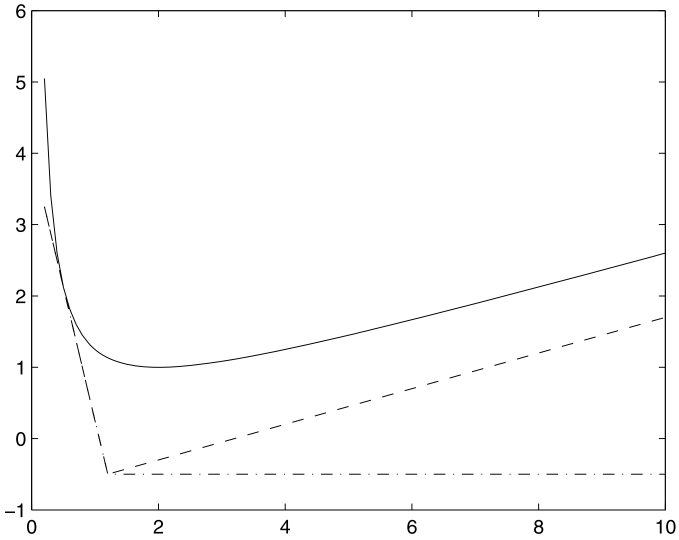

FIGURE 3.2. The most expensive subhedging portfolio (upper dashed line) for $\lambda(x) = 1/x + 0.25x$ given a single put option with strike 1.2 and price $p = 0.7$. The lower dashed line is the portfolio priced at $\int \lambda(x)\mu^{\dagger}(\mathrm{d}x)$ which is suboptimal.

the optimal value of ϕ is 0). When $p = 0.7$ the optimal value is $\psi = -0.8$ and we find that in this and every other such case the implied weight is $w_0 = 1$, as the general theory predicts.

Next, take $a = 0.25$, $b = 1$, so that $\lambda(x) = 1/x + 0.25x$, has its minimum at $x = 2$ and is asymptotically linear. We take $p = 0.7$. Taking the expectation with respect to the limiting measure $\mu^{\dagger}$ gives the value of the lower sub-hedge in Figure 3.2, but this is not optimal: we can add a maximum number 0.25 of call options, which have positive value, giving the upper sub-hedge in Figure 3.2. This is optimal, but does not correspond to any dual measure.

Finally, let $a = 0.0625, b = 2$, giving $\lambda(x) = 1/x + 0.0625x^2$. The minimum is still at $x = 2$ but λ is asymptotically quadratic. This function satisfies condition (iii) of Proposition 3.3 and there is dual existence for every arbitrage-free value of p. The optimal portfolio for $p = 0.7$ is shown in Figure 3.3.

3.2. Upper Bound

To compute the cheapest super-replicating portfolio we have to solve the linear program

$$P_{\mathrm{UB}} : \inf_{\mathbf{y}\in\mathbb{R}^m} \mathbf{y}^T\mathbf{b} \quad \text{subject to} \quad \mathbf{y}^T\mathbf{a}(x) \geq \lambda(x)\ \forall x \in \mathbf{K}.$$

The corresponding dual program is

$$D_{\mathrm{UB}} : \sup_{\mu\in\mathbb{M}} \int_{\mathbf{K}} \lambda(x)\mu(\mathrm{d}x) \quad \text{subject to} \quad \int_{\mathbf{K}} \mathbf{a}(x)\mu(\mathrm{d}x) = \mathbf{b},$$

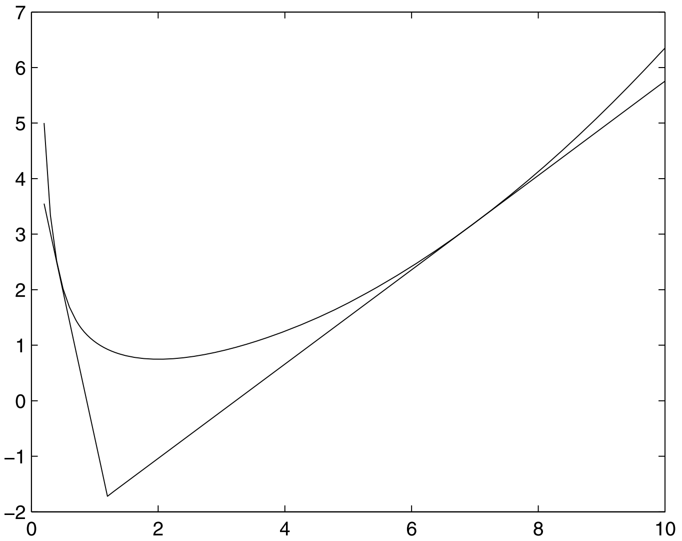

FIGURE 3.3. Most expensive subhedging portfolios for $\lambda(x) = 1/x + 0.0625x^2$ given a single put option with strike 1.2 and price $p = 0.7$.

where, by Lemma 2.2, we may replace $\mathbf{K}$ by $\mathbb{R}^+$. By (1.4) we have for $x \in \mathbb{R}^+$

$$\lambda(x) = \lambda(1) + \lambda'(1+)(x-1) + \int_{(0,1]} [k-x]^+ \lambda''(\mathrm{d}k) + \int_{(1,\infty)} [x-k]^+ \lambda''(\mathrm{d}k).$$

Consider $\mu \in \mathbb{M}_P$ and let $p_\mu(k) = \int [k - x]^+ \mu(\mathrm{d}x)$, $c_\mu(k) = \int [x - k]^+ \mu(\mathrm{d}k)$ be the (normalized) prices of puts and calls. Integrating the above against μ gives

$$\int \lambda(x)\mu(\mathrm{d}x) = \lambda(1) + \int_{(0,1]} p_\mu(k)\lambda''(\mathrm{d}k) + \int_{(1,\infty)} c_\mu(k)\lambda''(\mathrm{d}k). \tag{3.13}$$

Recall that $c_\mu(k) = p_\mu(k) + 1 - k$ and hence maximizing in $c_\mu(k)$ or in $p_\mu(k)$ is the same. $p_\mu(k)$ is a convex function dominated by the linear interpolation of points (k_i, p_i), $i = 0, 1, \dots, n$, extended to the right of (k_n, p_n) with slope 1. Because we assume the given put prices do not admit weak arbitrage, it follows from Davis and Hobson (2007) (see also Proposition 2.1 above) that this upper bound is attained either exactly or in the limit. More precisely, if $\bar{n} = \infty$ one can take $\mu_z \in \mathbb{M}_P$ supported on $\{k_{\underline{n}}, , \dots, , k_n, z\}$, for z large enough, which attain the upper bound on $[0, k_n]$ and asymptotically induce the upper bound on (k_n, ∞) as $z \to \infty$. It follows from (3.13) that the value of the dual problem is $V_D^U = \lim_{z\to\infty} \int_{\mathbf{K}} \lambda(x)\mu_z(\mathrm{d}x)$. If $\bar{n} \leq n$ one can take $z = k_{\bar{n}}$ and $\mu_{k_{\bar{n}}}$ attains the upper bound. It follows from (3.13) that then $V_D^U = \int_{\mathbf{K}} \lambda(x)\mu_{k_{\bar{n}}}(\mathrm{d}x)$.

From this observation, one expects that $\check{\mathbf{y}}$—the solution to the primal problem—will correspond to a (normalized) portfolio $\check{\mathbf{y}}^T\mathbf{a}(x)$ which linearly interpolates $(k_i, \lambda(k_i))$, $i = k_{\underline{n}}, \dots, n \wedge \bar{n}$ and (if $\bar{n} = \infty$) extends linearly to the right as to dominate $\lambda(k)$. The function $x \mapsto \mathbf{y}^T\mathbf{a}(x)$ is piecewise linear with a finite number of pieces and no such function can majorize the convex function λ over $\mathbb{R}^+$ unless $\lambda(0) < \infty$ and $\lambda'(\infty) = \gamma < \infty$.

In general we impose:

$$(3.14)\qquad \begin{aligned} &(a) \quad \underline{n} > 0 \text{ or } (\underline{n} = 0 \text{ and } \lambda(0) < \infty) \quad \textit{and} \\ &(b) \quad \bar{n} \leq n \text{ or } (\bar{n} = \infty \text{ and } \lambda'(\infty) < \infty)\,. \end{aligned}$$

We have the following result.

PROPOSITION 3.4. *If condition (3.14) holds then there exists a solution* $\check{\mathbf{y}}$ *to the linear program* P_{UB}. *The function* $\check{\mathbf{y}}^T\mathbf{a}(x)$ *is the linear interpolation of the points* $(k_{\underline{n}}, \lambda(k_{\underline{n}})), , \dots, (k_{n\wedge\bar{n}}, \lambda(k_{n\wedge\bar{n}}))$ *together with, if* $\bar{n} = \infty$, *the line* $l(x) = \check{\mathbf{y}}^T\mathbf{a}(k_n) + (x - k_n)\lambda'(\infty)$ *for* $x \geq k_n$. *Primal and dual problem have the same value* $V_U^P = V_U^D$ *and the existence of a maximizer in the dual problem fails if and only if* $\bar{n} = \infty$ *and* λ *is not affine on* $[k_n, \infty)$.

If the condition (3.14) is not satisfied, there is no feasible solution and $V_U^D = \infty$.

Proof. As argued above, (3.14) is a necessary condition for existence of a feasible solution. Suppose, for example, that $\bar{n} = \infty = \lambda'(\infty)$. Then $(z - k_n)\mu_z(\{z\}) = c_n$ and hence $z\mu_z(\{z\}) \to c_n$ as $z \to \infty$. Together with $\lambda'(\infty) = \infty$ this implies that $V_U^D = \infty$. Other cases are similar.

Suppose (3.14) holds and first consider the case when $\bar{n} = \infty$. λ is bounded on $[k_{\underline{n}}, k_n]$ and the linear interpolation is well defined as is the extension beyond k_n. Further, there exists some constant δ such that $\check{\mathbf{y}}^T\mathbf{a}(x) - \lambda(x) \leq \delta$ for all $x \in \mathbb{R}^+$. The weight $\check{\mathbf{y}}_{2+j}$ on the jth put option is the change in slope at $k_{\underline{n}+j}$, the "underlying" weight $\mathbf{y}_2$ is equal to γ, and at $x = k_n$ we have $\lambda_n = \check{\mathbf{y}}_1 + \check{\mathbf{y}}_2 k_n$, so the "cash" weight is $\check{\mathbf{y}}_1 = \lambda_n - k_n\gamma$. The value of the objective function is, by definition,

$$\check{\mathbf{y}}^T\mathbf{b} = \int_{\mathbb{R}^+} \check{\mathbf{y}}^T\mathbf{a}(x)\mu(\mathrm{d}x),$$

for any $\mu \in \mathbb{M}_P$. In particular, because $\check{\mathbf{y}}^T\mathbf{a}(k_i) = \lambda(k_i)$, taking z large enough, we have

$$(3.15)\qquad \begin{aligned} \check{\mathbf{y}}^T\mathbf{b}^0 &= \int_{\mathbb{R}^+} \check{\mathbf{y}}^T\mathbf{a}(x)\mu_z(\mathrm{d}x) = \sum_{i=\underline{n}}^{\bar{n}} \lambda(k_i)\mu_z(\{k_i\}) + \check{\mathbf{y}}^T\mathbf{a}(z)\mu_z(\{z\}) \\ &= \int_{\mathbb{R}^+} \lambda(x)\mu_z(\mathrm{d}x) + (\check{\mathbf{y}}^T\mathbf{a}(z) - \lambda(z))\mu_z(\{z\}) \leq \int_{\mathbb{R}^+} \lambda(x)\mu_z(\mathrm{d}x) + \delta\mu(\{z\}). \end{aligned}$$

Recall that $\int x\mu_z(\mathrm{d}x) = 1$ and in particular $\mu(\{z\}) \to 0$ as $z \to \infty$. Taking the limit as $z \to \infty$ in the above, we conclude that $\check{\mathbf{y}}^T\mathbf{b}^0 \leq V_D^U$. The basic inequality $V_D^U \leq V_P^U$ between the primal and dual values then implies that $\check{\mathbf{y}}$ is optimal for P_{UB} and $V_D^U = V_P^U$. The existence of the solution to D_{UB} fails unless there exists $z \geq k_n$ with $\check{\mathbf{y}}^T\mathbf{a}(z) = \lambda(z)$, which happens if and only if $\lambda(z)$ is affine on $[k_n, \infty)$.

When $\bar{n} \leq n$ the arguments are analogous, except that now we need to ensure $\mathbf{y}^T\mathbf{a}(x) \geq \lambda(x)$ only for $x \in \mathbf{K} = [k_{\underline{n}}, k_{\bar{n}}]$. The dual problem has a maximizer as observed in the remarks above the Proposition. The primal problem also has a solution $\check{\mathbf{y}}$ but it is not unique. Indeed, let $\mathbf{y}^0 = (-k_n, 1, 0, \dots, 0, 1)$ and observe that ${\mathbf{y}^0}^T\mathbf{b} = 0$ and ${\mathbf{y}^0}^T\mathbf{a}(x) \equiv 0$ for $x \in \mathbf{K}$. In consequence, we can add to $\check{\mathbf{y}}$ multiples of $\mathbf{y}^0$ without affecting its performance for P_{UB}. □

We can now summarize the results for the cheapest super-replicating portfolio as in (2.1), (2.2). The difference with the above is that we need to ensure super-replication for all possible values of S_T and not only for $S_T \in [K_{\underline{n}}, K_{\bar{n}}]$. Because the payoff X_T, as a function of S_T, is piecewise linear with a finite number of pieces it is necessary that λ_T satisfies

$$\lambda_T(0) < \infty \quad \text{and} \quad \lambda'_T(\infty) = \gamma < \infty. \tag{3.16}$$

Under this condition, the above results show that the cheapest super-replicating portfolio has a payoff which linearly interpolates $(K_i, \lambda_T(K_i))$, $i = 0, \ldots, n$ and extends to the right of K_n with slope $\lambda'_T(\infty)$.

PROPOSITION 3.5. *If (3.16) holds then there is a cheapest super-replicating portfolio* $(\psi^*, \phi^*, \pi_i^*)$ *of the European option with payoff* $\lambda_T(S_T)$ *whose initial price is*

$$X_0^* = \sup_z D_T \int_{\mathbb{R}^+} \lambda_T(F_T x)\mu_z(\mathrm{d}x).$$

The underlying component is $\phi^* = \gamma\Gamma_T$*, the cash component is* $\psi^* = D_T(\lambda_T(K_n) - \gamma K_n)$*, and the option components are*

$$\pi_i^* = \frac{\lambda_T(K_{j+1}) - \lambda_T(K_j)}{K_{j+1} - K_j} - \frac{\lambda_T(K_j) - \lambda_T(K_{j-1})}{K_j - K_{j-1}}.$$

If condition (3.16) is not satisfied, there is no super-replicating portfolio.

3.3. Arbitrage Conditions

With the above results in hand we can state the arbitrage relationships when a European option whose exercise value at T is a convex function $\lambda_T(S_T)$ can be traded at time 0 at price P_λ in a market where there already exist traded put options, whose prices P_i are in themselves consistent with absence of arbitrage. Recalling the notation of Propositions 3.1 and 3.5, $X_0^\dagger$ and X_0^* are respectively the setup costs of the most expensive sub-replicating and cheapest super-replicating portfolios, with $X_0^* = +\infty$ when no super-replicating portfolio exists.

THEOREM 3.6. *Assume the put prices do not admit a weak arbitrage. Consider a convex function* λ_T *and suppose that if* λ_T *is affine on some half–line* $[z, \infty)$ *then it is strictly convex on* $[0, z)$*. The following are equivalent:*

1. *The prices* $P_\lambda, P_1, \ldots, P_n$ *do not admit a weak arbitrage.*
2. *There exists a market model for put options in which* $P_\lambda = D_T\mathbb{E}[\lambda_T(S_T)]$.
3. *The following condition (3.17) holds and either* $P_\lambda \in (X_0^\dagger, X_0^*)$, or $P_\lambda = X_0^\dagger$ *and existence holds in* D_{LB}*, or* $P_\lambda = X_0^* < \infty$ *and existence holds in* D_{UB}.

$$P_2 > \frac{K_2}{K_1} P_1 \text{ if } \underline{n} = 0 \text{ and } \lambda \text{ is unbounded at the origin.} \tag{3.17}$$

If (3.17) holds and $P_\lambda \notin [X_0^\dagger, X_0^*]$ *then there is a model-independent arbitrage. If (3.17) holds and* $P_\lambda = X_0^\dagger$ *or* X_0^* *and existence fails in* D_{LB} *or in* D_{UB}*, respectively, or if (3.17) fails, then there is a weak arbitrage.*

REMARK 3.7. We note that the robust pricing and hedging problem solved above is essentially invariant if λ_T is modified by an affine factor. More precisely, if we consider a European option with payoff $\lambda^1_T(S_T) = \lambda_T(S_T) + \phi S_T + \psi$ and let P_{λ^1} denote its price then the prices $P_1, \dots, P_n$, P_λ are consistent with absence of arbitrage if and only if $P_1, \dots, P_n$, $P_{\lambda^1} = P_\lambda + \phi S_0/\Gamma_T + \psi D_T$ are.

Proof. Suppose first that condition (3.17) holds. We saw in the proof of Proposition 3.1 that this condition (under its equivalent form $p_2 > (k_2/k_1)p_1$) guarantees the existence of a sub-replicating portfolio with value $X^\dagger_0$. If $P_\lambda \in (X^\dagger_0, X^*_0)$ then there exists $\epsilon > 0$ such that $P_\lambda \in (X^\dagger_0 + \epsilon, X^*_0 - \epsilon)$ and, because there is no duality gap, there are measures $\mu_1, \mu_2 \in \mathbb{M}_P$ such that $D_T\mathbb{E}_{\mu_1}[\lambda_T(S_T)] < X^\dagger_0 + \epsilon$ and $D_T\mathbb{E}_{\mu_2}[\lambda_T(S_T)] > X^*_0 - \epsilon$. A convex combination μ of μ_1 and μ_2 then satisfies $D_T\mathbb{E}_\mu[\lambda_T(S_T)] = P_\lambda$ and one constructs a market model, for example by using Skorokhod embedding as explained in Section 2 above. If existence holds in D_{LB} then it was shown in the proof of Proposition 3.1 that the minimizing measure $\mu^\dagger$ satisfies

$$X^\dagger_0 = D_T \int_{\mathbf{K}} \lambda_T(F_T x)\mu^\dagger(\mathrm{d}x),$$

so that if $P_\lambda = X^\dagger_0$ then $\mu^\dagger$ is a martingale measure that consistently prices the convex payoff λ_T and the given set of put options. The same argument applies on the upper bound side.

Next, suppose that condition (3.17) holds and $P_\lambda = X^\dagger_0$ but no minimizing measure μ exists in D_{LB}. Let $\mathcal{M}$ be a model and $\hat\mu$ the distribution of S_T under $\mathcal{M}$. We can, at zero initial cost, buy $\lambda_T(S_T)$ and sell the portfolio $X^\dagger_T$ and this strategy realizes an arbitrage under $\mathcal{M}$ if $\hat\mu(\{\lambda_T(S_T) > X^\dagger_T\}) > 0$. Suppose now that $\hat\mu(\{\lambda_T(S_T) > X^\dagger_T\}) = 0$ and consider two cases. First, if λ_T is not affine on some $[z, \infty)$ then, by Proposition 3.3, $\bar n = \infty$ and $\lambda_T(S_T) > X^\dagger_T$ for $S_T \geq K_n$ so that in particular $\hat\mu([K_n, \infty)) = 0$. A strategy of going short a call option with strike K_n (which has strictly positive price because $\bar n = \infty$) gives an arbitrage because $S_T < K_n$ a.s. in $\mathcal{M}$. Second, suppose that $\lambda''_T(x) \equiv 0$ for $x \geq z$ but $\lambda''_T(x) > 0$ for $x < z$. If $\hat\mu([K_n, \infty)) = 0$ then we construct the arbitrage as previously so suppose this does not hold. Recall the atomic measure $\mu^\dagger$ defined in Case 2 in the proof of Proposition 3.3 and that $\iota_1 = 1$ as we do not have existence of a minimizer for the dual problem. $\lambda(x)$ in (3.1) is strictly convex on $[0, \tilde z)$ and linear on $[\tilde z, \infty)$ with $\tilde z = z/F_T$. It is not hard to see that $\mu^\dagger$ has to have an atom in some $x^*_n \in (k_{n-1}, k_n)$ and that either $\tilde z = x^*_n$ or else $\tilde z \geq k_n$. Otherwise we could modify $\mu^\dagger$ to obtain a minimizer for the dual problem. Strict convexity of λ_T on $[0, z)$ implies that there exist at most n points $s_1, \dots, s_m$ such that $s_i \in (K_{i-1}, K_i)$ and $\lambda_T(S_T)$ strictly dominates $X^\dagger_T$ for other values of $S_T \leq z$ and hence $\hat\mu([0, z)) = \hat\mu(\{s_1, \dots, s_n\})$. It follows that support of $\mu^\dagger$ is a subset of $\{s_1/F_T, \dots, s_n/F_T = x^*_n\}$. Let $\psi > 0$ and consider a portfolio Y with

$$Y_T = Y_T(S_T) = \sum_{i=1}^{n} \gamma_i(K_i - S_T)^+ + \psi, \quad \text{such that} \quad Y_T(s_j) = 0, \quad j = 1, \dots n.$$

This uniquely specifies $\gamma_i \in \mathbb{R}$. The payoff of Y is simply a zigzag line with kinks in K_i, zero in each s_i and equal to ψ for $S_T \geq K_n$. It follows that $\hat\mu\{Y_T \geq 0\} = 1$ and $\hat\mu\{Y_T > 0\} > 0$ as $\hat\mu([K_n, \infty)) > 0$. However $\mu^\dagger$ prices all put options correctly so that the initial price of Y is

$$\begin{aligned}(3.18)\quad Y_0 &= \sum_{i=1}^{n}\gamma_i P_i + D_T\psi = D_T\sum_{i=1}^{n}\gamma_i\int_0^{k_i}(K_i - F_T m)\mu^{\dagger}(\mathrm{d}m) + D_T\psi\\ &= D_T\int_0^{k_n}\left(\sum_{i=1}^{n}\gamma_i(K_i - F_T m)^+ + \psi\right)\mu^{\dagger}(\mathrm{d}m) = D_T\int_0^{k_n} Y_T(F_T m)\mu^{\dagger}(\mathrm{d}m) = 0,\end{aligned}$$

by construction of Y because $\mu^{\dagger}(\{s_1/F_T,\dots,s_n/F_T\}) = 1$ as remarked above, and where we used $\iota_1 = 1$. It follows that Y is an arbitrage strategy in $\mathcal{M}$.

Now suppose that condition (3.17) holds and $P_\lambda = X_0^* < \infty$ and there is no maximizing measure in the dual problem D_{UB}. Then, by Proposition 3.4, $\bar{n} = \infty$ and $\lambda_T(S_T) < X_T^*$ for $S_T > K_n$. Straightforward arguments as in the first case above show that there is a weak arbitrage.

Finally, if $P_\lambda < X_0^{\dagger}$ a model independent arbitrage is given by buying the European option with payoff $\lambda_T(S_T)$ and selling the subheding portfolio. This initial cost is negative while the payoff, because $X^{\dagger}$ subhedges $\lambda_T(S_T)$, is nonnegative. If $P_\lambda > X_0^*$ we go short in the option and long in the superhedge.

Now suppose (3.17) does not hold, so that $P_2/K_2 = P_1/K_1$ (this is the only case other than (3.17) consistent with absence of arbitrage among the put options). Consider portfolios with exercise values

$$\begin{aligned}H_1(S_T) &= [K_2 - S_T]^+ - \frac{K_2}{K_1}[K_1 - S_T]^+\\ H_2(S_T) &= \lambda_T(S_T) - \lambda_T(K_2) - \lambda_0'(K_2)(S_T - K_2)\\ &\quad - \frac{1}{P_1}(P_\lambda - \lambda(K_2) - \lambda_0'(F_T - K_2))\,[K_1 - S_T]^+,\end{aligned}$$

where λ_0' denotes the left derivative. The setup cost for each of these is zero, and $H_1(s) > 0$ for $s \in (0, K_2)$ while $H_2(s) \to \infty$ as $s \to 0$. There is a number $\theta \geq 0$ such that $H(s) = \theta H_1(s) + H_2(s) > 0$ for $s \in (0, K_2)$. Weak arbitrage is realized by a portfolio whose exercise value depends on a given model $\mathcal{M}$ and is specified via

$$X_T(S_T) = \begin{cases}-[K_2 - S_T]^+ & \text{if} \quad \mathbb{P}[S_T \in [0, K_2)] = 0\\ H(S_T) & \text{if} \quad \mathbb{P}[S_T \in [0, K_2)] > 0.\end{cases}$$

This completes the proof. □

4. WEIGHTED VARIANCE SWAPS

We come now to the second part of the paper where we consider weighted variance swaps. The main idea, as indicated in the Introduction, is to show that a weighted variance swap contract is equivalent to a European option with a convex payoff and hence their prices have to be equal. The equivalence here means that the difference of the two derivatives may be replicated through trading in a model-independent way. In order to formalize this we need to define (continuous) trading in absence of a model, i.e., in absence of a fixed probability space. This poses technical difficulties as we need to define pathwise stochastic integrals.

One possibility is to define stochastic integrals as limits of discrete sums. The resulting object may depend on the sequence of partitions used to define the limit. This approach was used in Bick and Willinger (1994) who interpreted the difference resulting from different sequences of partitions as broker's method of implementing continuous-time trading order. They were then interested in what happens if they apply the pricing-through-replication arguments on the set of paths with a fixed (σ^2) realized quadratic variation and wanted to recover Black-Scholes pricing and hedging. However for our purposes the ideas of Bick and Willinger (1994) are not suitable. We are interested in a much wider set of paths and then the replication of a weighted variance swap combining trading and a position in a European option would depend on the "broker" (i.e., sequence of partitions used). Instead, as in Lyons (1995), we propose an approach inspired by the work of Föllmer (1981). We restrict the attention to paths which admit quadratic variation or pathwise local time. For such paths we can develop pathwise stochastic calculus including Itô and Tanaka formulae. As this subject is self-contained and of independent interest we isolate it in Appendix B. Insightful discussions of this topic are found in Bick and Willinger (1994) and Lyons (1995).

To the standing assumptions (i)–(iii) of Section 2 we add another one:

(iv) $(S_t : t \leq T) \in \mathcal{L}^+$ – the set of *strictly* positive, continuous functions on $[0, T]$ which admit a finite, non-zero, quadratic variation and a pathwise local time, as formally defined in Definitions B.1, B.3 and Proposition B.4 of Appendix B.

Thus, our idea for the framework, as opposed to fixing a specific model $\mathcal{M}$, is to assume we are given a set of possible paths for the price process: $(S_t : t \leq T) \in \mathcal{P}$. This could be, for example, the space of continuous nonnegative functions, the space of functions with finite non-zero quadratic variation, or the space of continuous functions with a constant fixed realized volatility. The choice of $\mathcal{P}$ is supposed to reflect our beliefs about characteristics of price dynamics as well as modeling assumptions we are willing to take. Our choice above, $\mathcal{P} = \mathcal{L}^+$, is primarily dictated by the necessity to develop a pathwise stochastic calculus. It would be interesting to understand if an appropriate notion of no-arbitrage implies (iv). A recent paper of Vovk (2011), based on a game-theoretic approach to probability, suggests one may exclude paths with infinite quadratic variation through a no-arbitrage-like restriction, an interesting avenue for further investigation.

We introduce now a continuous time analogue of the weighted realized variance (1.1). Namely, we consider a market in which, in addition to finite family of put options as above, a w-weighted variance swap is traded. It is specified by its payoff at maturity T:

$$RV_T^w - P_T^{\mathrm{RV}(w)} := \int_0^T w(S_t/F_t)\mathrm{d}\langle \log S\rangle_t - P_T^{\mathrm{RV(w)}}, \tag{4.1}$$

where $P_T^{\mathrm{RV}(w)}$ is the swap rate, and has null entry cost at time 0. The above simplifies (1.1) in two ways. First, similarly to the classical works on variance swaps going back to Neuberger (1994), we consider a continuously and not discretely sampled variance swap which is easier to analyze with tools of stochastic calculus. Secondly, the weighting in (1.1) is a function of the asset price $h(S_{t_i})$ and in (4.1) it is a function of the ratio of the actual and the forward prices $w(S_t/F_t)$. This departure from the market contract definition is unfortunate but apparently necessary to apply our techniques. In practice, if $\hat{w}(S_t)$ is the function appearing in the contract definition we would apply our results with $w(x) = \hat{w}(S_0 x)$, so that $w(S_t/F_t) = \hat{w}((S_0/F_t)S_t)$. Because maturity times are short and, at present, interest rates are low, we have $S_0/F_t \approx 1$. See below and Section 5 for further remarks.

Our assumption (iv) and Proposition B.6 imply that ($\log S_t$, $t \leq T$) $\in \mathcal{L}$. Theorem B.5 implies that (4.1) is well defined as long as $w \in \mathrm{L}^2_{\mathrm{loc}}$, we can integrate with respect to S_t or M_t and obtain an Itô formula. This leads to the following representation.

LEMMA 4.1. *Let $w : \mathbb{R}_+ \to [0, \infty)$ be a locally square integrable function and consider a convex C^1 function λ_w with $\lambda''_w(a) = \frac{w(a)}{a^2}$. The extended Itô formula (B.1) then holds and reads*

$$\lambda_w(M_T) = \lambda_w(1) + \int_0^T \lambda'_w(M_u)\mathrm{d}M_u + \frac{1}{2}\int_{[0,T]} w(M_u)\mathrm{d}\langle \ln M\rangle_u. \tag{4.2}$$

The function λ_w is specified up to an addition of an affine component which does not affect pricing or hedging problems for a European option with payoff λ_w, see Remark 3.7 above. In what follows we assume that w and λ_w are fixed. Three motivating choices of w, as discussed in the Introduction, and the corresponding functions λ_w, are:

(1) Realized variance swap: $w \equiv 1$ and $\lambda_w(x) = -\ln(x)$. In this case there is of course no distinction between w and the contract function $\hat{w}$.

(2) Corridor variance swap: $w(x) = \mathbf{1}_{(0,a)}(x)$ or $w(x) = \mathbf{1}_{(a,\infty)}(x)$, where $0 < a < \infty$ and

$$\lambda_w(x) = \left(-\ln\left(\frac{x}{a}\right) + \frac{x}{a} - 1\right) w(x).$$

Here we would take $a = b/S_0$ if the contract corridor is $(0, b)$ or (b, ∞)

(3) Gamma swap: $w(x) = S_0 x$ and $\lambda_w(x) = S_0(x\ln(x) - x)$.

Clearly, (4.2) suggests that we should consider portfolios which trade dynamically and this will allow us to link w-weighted realized variance RV^w_T with a European option with a convex payoff λ_w. Note however that it is sufficient to allow only for relatively simple dynamic trading where the holdings in the asset only depend on asset's current price. More precisely, we extend the definition of portfolio X from static portfolios as in (2.1)-(2.2) to a class of dynamic portfolios. We still have a static position in traded options. These are options with given market prices at time zero and include n put options but could also include another European option, a weighted variance swap or other options. At time t we also hold $\Gamma_t\phi(M_t)$ assets S_t and ψ_t/D_t in cash. The portfolio is self-financing on $(0, T]$ so that

$$\psi_t := \phi(M_0)S_0 + \psi(0, S_0) + \int_0^t \phi(M_u)\mathrm{d}M_u - \Gamma_t\phi(M_t)S_tD_t, \quad t \in (0, T], \tag{4.3}$$

and where ϕ is implicitly assumed continuous and with a locally square integrable weak derivative so that the integral above is well defined, cf. Theorem B.5. We further assume that there exist: a linear combination of options traded at time zero with total payoff $Z = Z(S_t: t \leq T)$, a convex function G and constants $\tilde{\phi}$, $\tilde{\psi}$ such that

$$\Gamma_t\phi(M_t)S_t + \psi_t/D_t \geq Z - G(M_t)/D_t + \tilde{\phi}\Gamma_tS_t + \tilde{\psi}/D_t, \quad \forall t \leq T. \tag{4.4}$$

Such a portfolio X is called *admissible*. Observe that, in absence of a model, the usual *integrability* of Z is replaced by *having finite price at time zero*. In the classical setting, the admissibility of a trading strategy may depend on the model. Here admissibility of a strategy X may depend on which options are assumed to trade in the market. The presence of the term $G(M_t)$ on the right-hand side will become clear from the proof of

Theorem 4.3 below. It allows us to enlarge the space of admissible portfolios for which Lemma 4.2 below holds.

The two notions of arbitrage introduced in Section 2 are consequently extended by allowing not only static portfolios but possibly dynamic admissible portfolios as above. All the previous results remain valid with the extended notions of arbitrage. Indeed, if given prices admit no *dynamic* weak arbitrage then in particular they admit no *static* weak arbitrage. And for the reverse, we have the following general result.

LEMMA 4.2. *Suppose that we are given prices for a finite family of co-maturing options.*[3] *If a market model $\mathcal{M}$ exists for these options then any admissible strategy X satisfies $\mathbb{E}[D_T X_T] \le X_0$. In particular, the prices do not admit a weak arbitrage.*

Proof. Let $\mathcal{M}$ be a market model and X be an admissible strategy. We have $X_T = Z^1 + Y_T$, where Z^1 is a linear combination of payoffs of traded options and $Y_t = \Gamma_t \phi(M_t) S_t + \psi_t / D_t$ satisfies (4.4). Using (4.3) it follows that

$$D_t Y_t = \phi(1) S_0 + \psi(0, S_0) + \int_0^t \phi(M_u) \mathrm{d} M_u \ge D_t Z - G(M_t) + \tilde{\phi}^+ S_0 M_t + \tilde{\psi}.$$

We may assume that $G \ge 0$, it suffices to replace G by G^+. $\mathcal{M}$ is a market model and in particular Z is an integrable random variable. Because the traditional stochastic integral and our pathwise stochastic integral coincide a.s. in $\mathcal{M}$, see Theorem B.7, we conclude that $D_t Y_t$ is a local martingale and so is $N_t := D_t Y_t - \tilde{\phi}^+ S_0 M_t$. We will argue that this implies $\mathbb{E} N_t \le N_0$. Let ρ_n be the localizing sequence for N so that $\mathbb{E} N_{t \wedge \rho_n} = N_0$. In what follows all the limits are taken as $n \to \infty$. Fatou's lemma shows that $\mathbb{E} N_t^+ \le \liminf \mathbb{E} N_{t \wedge \rho_n}^+$ and $\mathbb{E} N_t^- \le \liminf \mathbb{E} N_{t \wedge \rho_n}^-$. By Jensen's inequality the process $G(M_t)$ is a submartingale, in particular the expectation is increasing and $\mathbb{E} G(M_t) \le \mathbb{E} G(M_T)$ which is finite because $\mathcal{M}$ is a market model. Using the Fatou lemma we have

$$\lim \mathbb{E} G(M_{t \wedge \rho_n}) \le \mathbb{E} G(M_t) = \mathbb{E} \liminf G(M_{t \wedge \rho_n}) \le \liminf \mathbb{E} G(M_{t \wedge \rho_n}),$$

showing that $\lim \mathbb{E} G(M_{t \wedge \rho_n}) = \mathbb{E} G(M_t)$. Observe that N_t is bounded below by $\tilde{Z} - G(M_t)$, where $\tilde{Z}$ is an integrable random variable and G is convex. Using Fatou's lemma again we can write

$$\begin{aligned} \mathbb{E} G(M_t)^+ + \mathbb{E} \tilde{Z}^- - \liminf \mathbb{E} N_{t \wedge \rho_n}^- &= \liminf \mathbb{E}[G(M_{t \wedge \rho_n})^+ + \tilde{Z}^- - N_{t \wedge \rho_n}^-] \\ &\ge \mathbb{E}[G(M_t)^+ + \tilde{Z}^- - N_t^-] \end{aligned}$$

which combined with the above gives $\mathbb{E} N_t^- = \liminf \mathbb{E} N_{t \wedge \rho_n}^-$ and in consequence $\mathbb{E} N_t \le \liminf \mathbb{E} N_{t \wedge \rho_n} = N_0$, as required. This shows that $\mathbb{E}[D_T Y_T] \le Y_0$. Because in a market model expectations of discounted payoffs of the traded options coincide with their initial prices it follows that $\mathbb{E}[D_T X_T] \le X_0 \le 0$. In particular if $X_T \ge 0$ and $X_0 \le 0$ then $X_T = 0$ a.s. and the prices do not admit a weak arbitrage. □

Having extended the notions of (admissible) trading strategy and arbitrage, we can now state the main theorem concerning robust pricing of weighted variance swaps. It is essentially a consequence of the hedging relation (4.2) and the results of Section 3.

[3] These could include European as well as exotic options.

THEOREM 4.3. *Suppose in the market which satisfies assumptions (i)–(iv) the following are traded at time zero: n put options with prices P_i, a w–weighted variance swap with payoff (4.1) and a European option with payoff $\lambda_T(S_T) = F_T\lambda_w(M_T)$ and price P_{λ_w}. Assuming the put prices do not admit a weak arbitrage, the following are equivalent*

1. *The "option" prices (European options and weighted variance swap) do not admit a weak arbitrage.*
2. *$P_1, \dots, P_n, P_{\lambda_w}$ do not admit a weak arbitrage and*

$$P_T^{\mathrm{RV}(w)} = \frac{2P_{\lambda_w}}{D_T F_T} - 2\lambda_w(1). \tag{4.5}$$

3. *A market model for all $n+2$ options exists.*

REMARK 4.4. It is true that under (4.5) a market model for $P_1, \dots, P_n, P_T^{\mathrm{RV}(w)}$ exists if and only if a market model for $P_1, \dots, P_n, P_{\lambda_w}$ exists. By Theorem 3.6, this is yet equivalent to $P_1, \dots, P_n, P_{\lambda_w}$ being consistent with absence of arbitrage. However it is not clear if this is equivalent to $P_1, \dots, P_n, P_T^{\mathrm{RV}(w)}$ being consistent with absence of arbitrage. This is because the portfolio of the variance swap and dynamic trading necessary to synthesize $-\lambda_T(S_T)$ payoff may not be admissible when $\lambda_T(S_T)$ is not a traded option.

REMARK 4.5. The formulation of Theorem 4.3 involves no-arbitrage prices but these are enforced via robust hedging strategies detailed in the proof. They involve the European option with payoff $\lambda_T(S_T)$ which in practice may not be traded and should be super-/sub-replicated using Propositions 3.1 and 3.4.

REMARK 4.6. It would be interesting to combine our study with the results of Hobson and Klimmek (2011) (HK) already alluded to in the Introduction. We note however that this may not be straightforward because the European option constituting the static part of the hedge in HK need not be convex and the variance kernels we consider are not necessarily monotone (in the terminology of HK), for example the Gamma swap $y(\log(y/x))^2$. Finally we note that the bounds in HK are attained by models where quadratic variation is generated entirely by a single large jump which is a radical departure from the assumption of continuous paths. Whether it is possible to obtain sharper bounds which only work for "reasonable" discontinuous paths is an interesting problem. We leave these challenges to future research.

Proof. We first show that 2$\Longrightarrow$3. Suppose that $P_1, \dots, P_n, P_{\lambda_w}$ do not admit a weak arbitrage and let $\mathcal{M}$ be a market model which prices correctly the n puts and the additional European option with payoff $\lambda_T(S_T)$. Note that, from the proof of Theorem 3.6, $\mathcal{M}$ exists and may be taken to satisfy (i)–(iv). Consider the Itô formula (4.2) evaluated at $\tau_n \wedge T = \inf\{t \geq 0 : M_t \notin (1/n, n)\} \wedge T$ instead of T. The continuous function λ_w' is bounded on $(1/n, n)$, the stochastic integral is a true martingale and taking expectations we obtain

$$\mathbb{E}[\lambda_w(M_{\tau_n \wedge T})] = \lambda_w(1) + \frac{1}{2}\mathbb{E}\left[\int_0^{\tau_n \wedge T} w(M_u)\mathrm{d}\langle M\rangle_u\right].$$

Subject to adding an affine function to λ_w we may assume that $\lambda_w \geq 0$. Jensen's inequality shows that $\lambda_w(M_{\tau_n \wedge T})$, $n \geq 2$, is a submartingale. Together with the Fatou lemma this shows that the left-hand side converges to $\mathbb{E}[\lambda_w(M_T)]$ as $n \to \infty$. Applying the monotone convergence theorem to the right-hand side we obtain

$$\mathbb{E}[\lambda_w(M_T)] = \lambda_w(1) + \frac{1}{2}\mathbb{E}\big[RV_T^w\big],$$

where either both quantities are finite or infinite. Because $\mathcal{M}$ is a market model for puts and $\lambda_T(S_T)$, we have $\mathbb{E}[\lambda_w(M_T)] = P_{\lambda_w}/(D_T F_T)$. Combining (4.5) with the above it follows that $\mathbb{E}[RV_T^w - P_T^{\mathrm{RV}(w)}] = 0$ and hence $\mathcal{M}$ is a market model for puts, $\lambda_T(S_T)$ and the w–wighted variance swap.

Lemma 4.2 implies that 3$\Longrightarrow$1. We note also that, if we have a market model $\mathcal{M}$ for all the puts, $\lambda_w(S_T)$ option and the w–wighted variance swap then by the above (4.5) holds. Then Lemma 4.2 also implies that 3$\Longrightarrow$2.

It remains to argue that 1$\Longrightarrow$2, i.e., that if $P_1, \ldots, P_n, P_{\lambda_w}$ are consistent with absence of arbitrage but (4.5) fails then there is a weak arbitrage. Consider a portfolio X with no put options, $\psi(0, S_0) = 2D_T(\lambda_w(1) - S_0\lambda_w'(1)) + D_T P_T^{\mathrm{RV}(w)}$ and $\phi(m) = 2D_T\lambda_w'(m)$. The setup cost of X is $X_0 = D_T(P_T^{\mathrm{RV}(w)} + 2\lambda_w(1))$. By assumption, λ_w is C^1 with $\lambda_w''(x) = w(x)/x^2 \in \mathrm{L}^2_{\mathrm{loc}}$ so that can apply Theorem B.5. Then, using (4.3) and (4.2) we obtain

$$\begin{aligned} X_t &= \Gamma_t\phi(M_t)S_t + \psi_t/D_t = \phi(S_0)S_0/D_t + \psi(0, S_0)/D_t + \frac{1}{D_t}\int_0^t \phi(M_u)\mathrm{d}M_u \\ &= \frac{D_T}{D_t}\left(2\lambda_w(1) + 2\int_0^t \lambda_w'(M_u)\mathrm{d}M_u + P_T^{\mathrm{RV}(w)}\right) \\ &= \frac{2D_T}{D_t}\lambda_w(M_t) - \frac{D_T}{D_t}\int_{[0,t]} w(S_u)\mathrm{d}\langle \ln M\rangle_u + \frac{D_T}{D_t}P_T^{\mathrm{RV}(w)}. \end{aligned}$$

Observing that $1 \geq D_T/D_t \geq D_T$ and $\int_{[0,t]} w(S_u)\mathrm{d}\langle \ln M\rangle_u$ is increasing in t it follows that both X and $-X$ are admissible. Suppose first that

$$P_T^{\mathrm{RV}(w)} < \frac{2P_{\lambda_w}}{D_T F_T} - 2\lambda_w(1). \tag{4.6}$$

Consider the following portfolio Y: short $2/F_T$ options with payoff $\lambda_T(S_T)$, long portfolio X and long a w–weighted variance swap. Y is admissible, the initial cost is

$$Y_0 = -2P_{\lambda_w}/F_T + X_0 = -2P_{\lambda_w}/F_T + D_T\big(P_T^{\mathrm{RV}(w)} + 2\lambda_w(1)\big) < 0,$$

while $Y_T = 0$ and hence we have a model independent arbitrage. If a reverse inequality holds in (4.6) then the arbitrage is attained by $-Y$. □

5. COMPUTATION AND COMPARISON WITH MARKET DATA

5.1. Solving the Lower Bound Dual Problem

When one of the sufficient conditions given in Proposition 3.3 is satisfied, for existence in the lower bound dual problem, then this problem can be solved by a dynamic programming algorithm. We briefly outline this here, referring the reader to Raval (2010) for complete details. For simplicity, we restrict attention to the practically relevant case $\underline{n} = 0$, $\bar{n} = \infty$. Of course, once the dual problem is solved, the maximal sub-hedging portfolio is immediately determined.

The measure $\mu^\dagger$ to be determined satisfies

$$\int \lambda(x)\mu^\dagger(\mathrm{d}x) = \inf_{\mu \in \mathbb{M}_P} \left\{ \int_{\mathbb{R}^+} \lambda(x)\mu(\mathrm{d}x) \right\},$$

and we recall from Lemma 3.2 that we can restrict our search to measures of the form $\mu(\mathrm{d}x) = \sum_{i=1}^{n+1} w_i \delta_{\chi_i}(\mathrm{d}x)$ where $\chi_i \in [k_{i-1}, k_i)$, $w_i \geq 0$, $\sum_i w_i = 1$. We denote $\zeta_0 = 0$ and, for $i \geq 1$, $\zeta_i = \sum_1^i w_j$, the cumulative weight on the interval $[0, k_i)$. For consistency of the put prices $r_1, \ldots, r_n$ with absence of arbitrage, Proposition 2.1 dictates that

$$\zeta_i \in A_i = \left[\frac{r_i - r_{i-1}}{k_i - k_{i-1}}, \frac{r_{i+1} - r_i}{k_{i+1} - k_i} \right] \quad \text{for } 1 \leq i < n$$

and

$$\zeta_n \in A_n = \left[\frac{r_n - r_{n-1}}{k_n - k_{n-1}}, 1 \right].$$

Given $\zeta = (\zeta_1, \ldots, \zeta_n)$ (the final weight is of course $w_{n+1} = 1 - \zeta_n$), the positions χ_i are determined by pricing the put options. We find that when $\zeta_{i-1} < \zeta_i$

$$\chi_i = \chi_i(\zeta_{i-1}, \zeta_i) = k_i + \frac{\zeta_{i-1}(k_i - k_{i-1}) - (r_i - r_{i-1})}{\zeta_i - \zeta_{i-1}} \quad \text{for} \quad i = 1, \ldots, n,$$

$$\chi_{n+1} = \chi_{n+1}(\zeta_n) = k_n + \frac{1 + r_n - k_n}{1 - \zeta_n}.$$

The measure corresponding to policy ζ is thus

$$\sum_{i=1}^n (\zeta_i - \zeta_{i-1}) \delta_{\chi_i(\zeta_{i-1}, \zeta_i)} + (1 - \zeta_n) \delta_{\chi_{n+1}(\zeta_n)}. \tag{5.1}$$

It follows that the minimization problem $\inf_{\mu \in \mathbb{M}_P} \int_{\mathbf{K}} \lambda(x)\mu(\mathrm{d}x)$ has the same value as

$$v_0 = \inf_{\zeta_1 \in A_1} \ldots \inf_{\zeta_n \in A_n} \left\{ \sum_{i=1}^n (\zeta_i - \zeta_{i-1}) \lambda(\chi_i(\zeta_{i-1}, \zeta_i)) + (1 - \zeta_n) \lambda(\chi_{n+1}(\zeta_n)) \right\}. \tag{5.2}$$

We can solve this by backwards recursion as follows. Define

(5.3)

$$V_n(\zeta_n) = (1 - \zeta_n)\chi_{n+1}(\zeta_n)$$
$$V_j(\zeta_j) = \inf_{\zeta_{j+1} \in A_{j+1},\, \zeta_{j+1} \geq \zeta_j} \{(\zeta_{j+1} - \zeta_j)\lambda(\chi_i(\zeta_j, \zeta_{j+1})) + V_{j+1}(\zeta_{j+1})\}, \quad j = n - 1, \ldots, 0.$$

Then $V_0(0) = v_0$. For a practical implementation one has only to discretize the sets A_j, and then (5.3) reduces to a discrete-time, discrete-state dynamic program in which the minimization at each step is just a search over a finite number of points.

5.2. Market Data

The vanilla variance swap is actively traded in the over-the-counter (OTC) markets. We have collected variance swap and European option data on the S&P 500 index from the recent past. Using put option prices, the lower arbitrage-bound for the variance swap

TABLE 5.1
Historical Variance Swap (VS) Quotes for the S&P 500 Index and the Lower Bound (LB) for It, Implied by the Bid Prices of Liquid European Put Options with the Same Maturity.

Term	Quote date	VS quote	LB	# puts	Libor
2M	21/04/2008	21.24	20.10	50	2.79
2M	21/07/2008	22.98	22.51	50	2.79
2M	20/10/2008	48.78	46.58	93	4.06
2M	20/01/2009	52.88	47.68	82	1.21
3M	31/03/2008	25.87	23.59	42	2.78
3M	20/06/2008	22.99	21.21	46	2.76
3M	19/09/2008	26.78	25.68	67	3.12
3M	19/12/2008	45.93	45.38	112	1.82
3M	20/12/2008	45.93	65.81	137	1.82
6M	24/03/2008	25.81	25.34	33	2.68
6M	20/06/2008	23.38	23.20	38	3.10

The units for the variance price and LB are volatility percentage points, $100 \times \sqrt{P_T^{\mathrm{VS}}}$, and M stands for months. The european option price data is courtesy of UBS investment bank, and the variance-swap data was provided by peter carr.

rate is computed in each case, and summarized in Table 5.1. One sees that the traded price of the variance swap frequently lies very close to the lower bound. Nonetheless, under our standing assumptions of frictionless markets, all but one of the prices were consistent with absence of arbitrage. The crash in October 2008 of the S&P 500, and indeed the financial markets in general, gave rise to significant increase in expected variance, which can be seen in the cross-section of data studied. One data point, the 3-month contract on December 20, 2008 lies below our lower bound and, at first sight, appears to represent an arbitrage opportunity. However, this data point should probably be discarded. First, practitioners tell us that this was a day of extreme disruption in the market, and indeed the final column of Table 5.1, giving the number of traded put prices used in the calculations, shows that December 19 and 20 were far from typical days. Second, the fact that the quotes for 19 and 20 December are exactly the same makes it almost certain that the figure for 20 December is a stale quote, not a genuine trade. It is a positive point of our method that we are able to pick up such periods of market dislocation, just from the raw price data.

A further point relates to our discussion in Section 4 about the weight function w and its relation to the contract weight $\hat{w}$. This is a moot point here, because $w = \hat{w} = 1$, but note from the final column in Table 5.1 that Libor rates are generally around 3% (albeit with some outliers), while the S&P 500 dividend yield increased from around 2.0% to around 3.1% over the course of 2008. Because the rate closely matches the dividend yield we have $F_t = S_0$ to a close approximation for t up to a few months.

APPENDIX A: THE KARLIN & ISII THEOREM

Let $a_1, \ldots, a_m, f$ be real-valued, continuous functions on $\mathbf{K} \subset \mathbb{R}^d$ and let $\mathbf{M}$ denote the collection of all finite Borel measures μ on $\mathbf{K}$ fulfilling the integrability conditions

$\int_{\mathbf{K}} |a_i(\mathbf{x})|\mu(\mathrm{d}\mathbf{x}) < \infty$ for $i = 1, \ldots, m$. For a fixed vector $\mathbf{b} = (b_1, \ldots, b_m)^T$, and letting $\mathbf{a}(\mathbf{x}) = (a_1(\mathbf{x}), \ldots, a_m(\mathbf{x}))^T$ for $\mathbf{x} \in \mathbf{K}$, consider the optimization problem

$$(P): \sup_{\mathbf{y}\in\mathbb{R}^m} \mathbf{y}^T\mathbf{b} \quad \text{s.t.} \quad \mathbf{y}^T\mathbf{a}(\mathbf{x}) \le f(\mathbf{x}) \ \forall \mathbf{x} \in \mathbf{K}.$$

Now, define

$$(D): \inf_{\mu\in\mathbf{M}} \int_{\mathbf{K}} f(\mathbf{x})\mu(\mathrm{d}\mathbf{x}) \quad \text{s.t.} \quad \int_{\mathbf{K}} \mathbf{a}(\mathbf{x})\mu(\mathrm{d}\mathbf{x}) = \mathbf{b},$$

where the constraint should be interpreted as $\int_{\mathbf{K}} a_i(\mathbf{x})\mu(\mathrm{d}\mathbf{x}) = b_i,\ i = 1, \ldots, m$. The values of the problems (P) and (D) will respectively be denoted by $V(P)$ and $V(D)$. Finally, $M_m \subset \mathbb{R}^m$ will denote the moment cone defined by

$$\text{(A.1)} \qquad M_m = \left\{ \tilde{\mathbf{b}} = (\tilde{b}_1, \ldots, \tilde{b}_m)^T \,|\, \tilde{b}_i = \int_{\mathbf{K}} a_i(\mathbf{x})\mu(\mathrm{d}\mathbf{x}),\ i = 1, \ldots, m,\ \mu \in \mathbf{M} \right\}.$$

THEOREM A.1. *Suppose*

1. $a_1, \ldots, a_m$ *are linearly independent over* $\mathbf{K}$,
2. $\mathbf{b}$ *is an interior point of* M_m, *and*
3. $V(D)$ *is finite.*

Then $V(P) = V(D)$ *and* (P) *has a solution.*

REMARKS.

(i) The beauty in the Karlin & Isii theorem is that the proof draws upon no more than a finite dimensional separating hyperplane theorem. Proofs can be found in Glashoff (1979) and Karlin and Studden (1966, chapter XII, section 2). We follow the latter below.
(ii) Condition 1 of the theorem ensures that the moment cone M_m of (A.1) is m-dimensional, i.e., not contained in some lower-dimensional hyperplane.

Proof of Theorem A.1. Define the enlarged moment cone

$$M = \left\{ \tilde{\mathbf{b}} = (\tilde{b}_1, \ldots, \tilde{b}_{m+1})^T : \tilde{\mathbf{b}} = \int_{\mathbf{K}} \begin{pmatrix} \mathbf{a}(\mathbf{x}) \\ f(\mathbf{x}) \end{pmatrix} \mu(\mathrm{d}\mathbf{x}),\ \mu \in \mathbf{M} \right\},$$

and let $\bar{M}$ denote its closure. Then $\bar{M}$ is a closed convex cone and, moreover, the vector $(b_1, \ldots, b_m,\ V(D))$ lies on its boundary. There exists a supporting hyperplane to $\bar{M}$ through this vector, specified by real constants $\{z_i\}_1^{m+1}$ not all zero, such that

$$\text{(A.2)} \qquad \sum_{i=1}^{m} z_i b_i + z_{m+1} V(D) = 0$$

and

$$\text{(A.3)} \qquad \sum_{i=1}^{m+1} z_i \tilde{b}_i \ge 0, \quad \forall \tilde{\mathbf{b}} \in \bar{M}.$$

In particular, on considering Dirac measures, one has

$$\sum_{i=1}^{m} z_i a_i(\mathbf{x}) + z_{m+1} f(\mathbf{x}) \geq 0, \quad \forall \mathbf{x} \in \mathbf{K}. \tag{A.4}$$

We now show that $z_{m+1} > 0$. Indeed, for any $\delta > 0$, the vector $(b_1, \ldots, b_m, V(D) - \delta)$ lies in the half-space complementary to (A.3). Therefore

$$\sum_{i=1}^{m} z_i b_i + z_{m+1}(V(D) - \delta) < 0, \quad \forall \delta > 0.$$

This clearly implies $z_{m+1} \geq 0$. However, $z_{m+1} = 0$ is not possible, because this would contradict the assumption that $\mathbf{b}$ lies in the interior M_m, which is m dimensional. Thus, it must be that $z_{m+1} > 0$. Then from (A.4), it follows that

$$f(\mathbf{x}) \geq \sum_{i=1}^{m} \hat{y}_i a_i(\mathbf{x}),$$

where $\hat{y}_i := -z_i / z_{m+1}$ for $i = 1, \ldots, m$. Then (A.2) becomes

$$V(D) = \sum_{i=1}^{m} \hat{y}_i b_i,$$

and this completes the proof. □

APPENDIX B: PATHWISE STOCHASTIC CALCULUS

This section describes a non-probabilistic approach to stochastic calculus, due to Föllmer (1981), that will enable us to define the continuous-time limit of the finite sums (1.1) defining the realized variance, without having to assume that the realized price function $t \mapsto S_t$ is a sample function of a semimartingale defined on some probability space.

For $T > 0$ let $\mathcal{T} = [0, T]$ with Borel sets $\mathcal{B}_{\mathcal{T}}$. A *partition* is a finite, ordered sequence of times $\pi = \{0 = t_0 < t_1 < \cdots < t_k = T\}$ with *mesh size* $m(\pi) = \max_{1 \leq j \leq k}(t_j - t_{j-1})$. We fix a nested sequence of partitions $\{\pi_n, n = 1, 2, \ldots\}$ such that $\lim_{n \to \infty} m(\pi_n) = 0$. All statements below relate to this specific sequence. An obvious choice would be the set of dyadic partitions $t_j^n = jT/2^n$, $j = 0, \ldots, 2^n$, $n = 1, 2, \ldots$.

DEFINITION B.1. *A continuous function $X : \mathcal{T} \to \mathbb{R}$ has the quadratic variation property if the sequence of measures on $(\mathcal{T}, \mathcal{B}_{\mathcal{T}})$*

$$\mu_n = \sum_{t_j \in \pi_n} (X(t_{j+1}) - X(t_j))^2 \delta_{t_j}$$

(where δ_t denotes the Dirac measure at t) converges weakly to a measure μ, possibly along some sub-sequence. The distribution function of μ is denoted $\langle X \rangle_t$, and we denote by $\mathcal{Q}$ the set of continuous functions having the quadratic variation property.

THEOREM B.2. (Föllmer 1981). *For $X \in \mathcal{Q}$ and $f \in C^2$, the limit*

$$\int_0^t f'(X_s)\mathrm{d}X_s := \lim_{n\to\infty} \sum_{t_j \in \pi_n} f'(X_{t_j})(X_{t_{j+1}} - X_{t_j})$$

is well defined and satisfies the pathwise Itô formula:

$$f(X_t) - f(X_0) = \int_0^t f'(X_s)\mathrm{d}X_s + \frac{1}{2}\int_0^t f''(X_s)\mathrm{d}\langle X\rangle_s. \tag{B.1}$$

Proof. (Sketch). The proof proceeds by writing the expansion

$$\begin{aligned} f(X_t) - f(X_0) = &\sum_j f'\big(X_{t_j}\big)\big(X_{t_{j+1}} - X_{t_j}\big) \\ &+ \frac{1}{2}\sum_j f''\big(X_{t_j}\big)\big(X_{t_{j+1}} - X_{t_j}\big)^2 + \sum_j R\big(X_{t_j}, X_{t_{j+1}}\big) \end{aligned} \tag{B.2}$$

where $R(a, b) \leq \phi(|b - a|)(b - a)^2$ with ϕ an increasing function, $\phi(c) \to 0$ as $c \to 0$. As $m(\pi) \to 0$, the third term on the right of (B.2) converges to 0 and the second term converges to the second term in (B.1). This shows that the first term converges, so that essentially the "pathwise stochastic integral" is defined by (B.1) for arbitrary C^1 functions f'. □

An important remark is that X_t, being continuous, achieves its minimum and maximum X_*, X^* in $[0, T]$, so only the values of f in the compact interval $[X_*, X^*]$ (depending on the path X) are relevant in (B.1).

For the applications in Section 4 we need an "Itô formula" valid when f'' is merely locally integrable, rather than continuous as it is in (B.1). In the theory of continuous semimartingales, such an extension proceeds via local time and the Tanaka formula—see theorem VI.1.2 of Rogers and Williams (2000). Here we present a pathwise version, following the diploma thesis of Wuermli (1980). For a C^2 function f we have $f(b) = f(a) + \int_a^b f'(y)dy$, and applying the same formula to f' we obtain

$$\begin{aligned} f(b) - f(a) &= \int_a^b \left(f'(a) + \int_a^y f''(u)\mathrm{d}u\right)\mathrm{d}y = f'(a)(b-a) + \int_a^b (b-u)f''(u)\mathrm{d}u \\ &= f'(a)(b-a) + \int_{-\infty}^{\infty} \mathbf{1}_{[a\wedge b, a\vee b]}(u)|b-u|f''(u)\mathrm{d}u. \end{aligned}$$

Hence for any partition $\pi = \{t_j\}$ of $[0, t]$ we have the identity

$$f(X_t) - f(X_0) = \sum_j f'(X_{t_j})(X_{t_{j+1}} - X_{t_j}) + \int_{-\infty}^{\infty} \sum_j \left(1_{[X_j^{\min}, X_j^{\max}]}(u)|X_{t_{j+1}} - u|\right) f''(\mathrm{d}u), \tag{B.3}$$

where $X_j^{\min} = \min\{X_{t_j}, X_{t_{j+1}}\}$ and $X_j^{\max} = \max\{X_{t_j}, X_{t_{j+1}}\}$. We define

$$L_t^{\pi}(u) = 2\sum_j 1_{[X_j^{\min}, X_j^{\max}]}(u)|X_{t_{j+1}} - u|, \tag{B.4}$$

and note that $L_t^{\pi}(u) = 0$ for $u \notin (X_*, X^*)$.

DEFINITION B.3. *Let $\mathcal{L}$ be the set of continuous paths on* $[0, T]$ *such that the discrete pathwise local time* $L_t^{\pi_n}(u)$ *converges weakly in* $\mathrm{L}^2(\mathrm{d}u)$ *to a limit* $L_t(\cdot)$, *for each* $t \in [0, T]$.

PROPOSITION B.4. $\mathcal{L} \subset \mathcal{Q}$. *For* $X \in \mathcal{L}$ *and* $t \in [0, T]$ *we have the occupation density formula*

$$\int_A L_t(u)\mathrm{d}u = \int_0^t \mathbf{1}_A(X_s)\mathrm{d}\langle X\rangle_s, \qquad A \in \mathcal{B}(\mathbb{R}). \tag{B.5}$$

Proof. Suppose $X \in \mathcal{L}$, and let π_n be a sequence of partitions of $[0, T]$ with $m(\pi_n) \to 0$. From (B.3) we have, for $f \in C^2$ and $t \in [0, T]$,

$$f(X_{\tilde{t}_n}) - f(X_0) - \sum_j f'\left(X_{t_j}\right)\left(X_{t_{j+1}} - X_{t_j}\right) = \frac{1}{2}\int_{-\infty}^{\infty} L_t^{\pi_n}(u) f''(u)\mathrm{d}u, \tag{B.6}$$

where $\tilde{t}_n$ is the nearest partition point in π_n to t. As $n \to \infty$, the first term on the left of (B.6) converges to $f(X_t)$ and, because $f'' \in \mathrm{L}^2([X_*, X^*], \mathrm{d}u)$, $X \in \mathcal{L}$ implies that the right-hand side converges to $(1/2)\int L_t(u)f''(u)du$. Hence the "integral" term on the left of (B.6) also converges to, say, I_t. If we take $f(x) = x^2$ then the left-hand side of (B.6) is equal to

$$\sum_{j:t_{j+1}\le\tilde{t}_n} X_{t_{j+1}}^2 - X_{t_j}^2 - 2X_{t_j}\left(X_{t_{j+1}} - X_{t_j}\right) = \sum_{j:t_{j+1}\le\tilde{t}_n} \left(X_{t_{j+1}} - X_{t_j}\right)^2,$$

showing that (a) $X \in \mathcal{Q}$ and (b) $I_t = \int_0^t f'(X_s)dX_s$, the Föllmer integral of Theorem B.2. It now follows from (B.6) that

$$\int_{-\infty}^{\infty} L_t(u) f''(u)\mathrm{d}u = \int_0^t f''(X_s)\mathrm{d}\langle X\rangle_s, \tag{B.7}$$

i.e., for any continuous function g we have

$$\int_{-\infty}^{\infty} L_t(u) g(u)\mathrm{d}u = \int_0^t g(X_s)\mathrm{d}\langle X\rangle_s.$$

Approximating the indicator function 1_A by continuous functions and using the monotone convergence theorem, we obtain (B.5). □

For the next result, let $\mathcal{W}_2$ be the set of functions f in $C^1(\mathbb{R})$ such that f' is weakly differentiable with derivative f'' in $\mathrm{L}^2_{\mathrm{loc}}(\mathbb{R})$.

THEOREM B.5. *If* $X \in \mathcal{L}$ *the Föllmer integral extends in such a way that the pathwise Ito formula (B.1) is valid for* $f \in \mathcal{W}_2$.

Proof. Let ϕ be a mollifier function, a nonnegative C^∞ function on $\mathbb{R}$ such that $\phi(x) = 0$ for $|x| > 1$ and $\int\phi(x)dx = 1$, define $\phi_n(x) = n\phi(nx)$ and let $f_n(x) = \int f(x - y)\phi_n(y)dy$. Then f_n, f_n' converge pointwise to f, f' and $f_n'' \to f''$ weakly in L^2. From Proposition B.4 we know that $X \in \mathcal{Q}$ and as $f_n \in C^2$ we have, using (B.1) and (B.7),

$$f_n(X_t) - f_n(X_0) = \int_0^t f_n'(X_s)\mathrm{d}X_s + \frac{1}{2}\int_{-\infty}^{\infty} L_t(u) f_n''(u)\mathrm{d}u.$$

As $n \to \infty$, the left-hand side converges to $f(X_t) - f(X_0)$ and the second term on the right converges to $\frac{1}{2}\int_{-\infty}^{\infty} L_t(u) f''(u)du$, and we can define

$$\int_0^t f_n'(X_s)\mathrm{d}X_s = \lim_{n\to\infty}\left\{ f_n(X_t) - f_n(X_0) - \frac{1}{2}\int_{-\infty}^{\infty} L_t(u) f_n''(u)\mathrm{d}u \right\}.$$

It now follows from Proposition B.4 that the Itô formula (B.1) holds for this extended integral. □

We need one further result.

PROPOSITION B.6. *Let $X \in \mathcal{L}$ and let $f : \mathbb{R} \to \mathbb{R}$ be a monotone C^2 function. Then $Y = f(X) \in \mathcal{L}$ and the pathwise local times are related by*

$$L_t^Y(u) = |f'(f^{-1}(u))|\, L_t^X(f^{-1}(u)). \tag{B.8}$$

Proof. We assume f is increasing—the argument is the same if it is decreasing—and denote $v = f^{-1}(u)$, so $u = f(v)$. For the partition π_n we have for fixed $t \in \mathcal{T}$, from (B.4),

$$\begin{aligned} L_t^{Y,\pi_n}(u) &= \sum_j \mathbf{1}_{\{f(X_{n,j}^{\min})\le f(v)\le f(X_{n,j}^{\max})\}} \left| f\big(X_{t_{j+1}^n}\big) - f(v)\right| \\ &= \sum_j \mathbf{1}_{\{X_{n,j}^{\min})\le v\le X_{n,j}^{\max})\}} \left| f'(v)\big(X_{t_{j+1}^n} - v)\big) + \frac{1}{2} f''(\xi)\big(X_{t_{j+1}^n} - v\big)^2 \right| \\ &= f'(v)\sum_j \mathbf{1}_{\{X_{n,j}^{\min})\le v\le X_{n,j}^{\max})\}} \left| X_{t_{j+1}^n} - v + \frac{f''}{2f'}\big(X_{t_{j+1}^n} - v\big)^2 \right| \end{aligned} \tag{B.9}$$

for ξ between v and $X_{t_{j+1}^n}$. Noting that $f''(\xi)/2f'(v)$ is bounded for $\xi, v \in [X_*, X^*]$ and that $\lim_{n\to\infty} \max_j |X(t_{j+1}^n) - X(t_j^n)| = 0$, we easily conclude that if $X \in \mathcal{L}$ then the expression at (B.9) converges in $\mathrm{L}^2(\mathrm{d}v)$ to $f'(v)L_t^X(v)$, so that $Y \in \mathcal{L}$ with local time given by (B.8). □

We note that in the above result it suffices to assume that f is defined on $[X_*, X^*]$. This allows us to apply the Proposition for $f = \log$ and and stock price trajectories in Section 4. Indeed, from (B.8) we obtain the elegant formula

$$L_t^{\log X}(u) = e^{-u} L_t^X(e^u).$$

Finally, we build the connection between the pathwise calculus and the classical Itô (stochastic) calculus.

THEOREM B.7. *Let $(X_t,\, t \in [0, T])$ be a continuous semimartingale on some complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then there is a set $N \in \mathcal{F}$ such that $\mathbb{P}N = 0$ and for $\omega \notin N$ the path $t \mapsto X(t, \omega)$ belongs to $\mathcal{L}$ (and hence to $\mathcal{Q}$), and $L_t(u)$ defined in Definition B.3 coincides with the semimartingale local time of X_t at u.*

Proof. First, $X_t \in \mathcal{Q}$ a.s. Indeed, theorem IV.1.3 of Revuz and Yor (1994) asserts that discrete approximations to the quadratic variation always converge in probability, so a sub-sequence converges almost surely, showing that $X_t \in \mathcal{Q}$ in accordance with Definition B.1. It is shown by Föllmer (1981) that the Itô integral and the pathwise integral defined by

(B.1) coincide almost surely. Every semimartingale S_t has an associated local time which satisfies the occupation density formula (B.5). It remains to show that with probability 1 the discrete approximations L_t^π defined by (B.4) converge in $\mathrm{L}^2(\mathrm{d}u)$. This is proved in Wuermli (1980), by detailed estimates which we cannot include here. □

REFERENCES

ACCIAIO, B., H. FÖLLMER, and I. PENNER (2011): Risk Assessment for Uncertain Cash Flows: Model Ambiguity, Discounting Ambiguity, and the Role of Bubbles, *Finance Stoch.* 16(4): 669–709.

BICK, A., and W. WILLINGER (1994): Dynamic Spanning without Probabilities, *Stochastic Proc. Appl.* 50(2), 349–374.

CONT, R. (2006): Model Uncertainty and Its Impact on the Pricing of Derivative Instruments, *Math. Finance* 16(3), 519–547.

COX, A. M. G., and J. OBŁÓJ (2011a): Robust Hedging of Double Touch Barrier Options, *SIAM J. Financ. Math.* 2, 141–182.

COX, A. M. G., and J. OBŁÓJ (2011b): Robust Hedging of Double No-Touch Barrier Options, *Finance Stoch.* 15(3), 573–605.

CROSBY, J., and M. DAVIS (2012): Variance Derivatives: Pricing and Convergence, Preprint, Imperial College London. Available at http://ssrn.com/abstract=2049278.

DAVIS, M., and D. HOBSON (2007): The Range of Traded Option Prices, *Math. Finance* 17, 1–14.

DELBAEN, F., and W. SCHACHERMAYER (1994): A General Version of the Fundamental Theorem of Asset Pricing, *Math. Ann.* 300(3), 463–520.

FÖLLMER, H. (1981): Calcul d'Itô Sans Probabilités, in *Séminaire de Probabilités XV (Univ. Strasbourg, 1979/1980)*, vol. 850 of Lecture Notes in Math., Berlin: Springer, pp. 143–150. Available at www.numdam.org. English translation in Sondermann 2006.

FÖLLMER, H., A. SCHIED, and S. WEBER (2009): Robust Preferences and Robust Portfolio Choice, in *Mathematical Modelling and Numerical Methods in Finance*, P. CIARLET, A. BENSOUSSAN and Q. ZHANG, eds., vol. 15 of Handbook of Numerical Analysis, Amsterdam: Elsevier, pp. 29–87.

GATHERAL, J. (2006): *The Volatility Surface: A Practitioner's Guide*. New York: Wiley.

GLASHOFF, K. (1979): Duality Theory of Semi Infinite Programming, in *Semi Infinite Programming (Proc. Workshop, Bad Honnef, 1978)*, vol. 15 of Lecture Notes in Control and Information Sci., Berlin: Springer, pp. 1–16.

HANSEN, L. P., and T. J. SARGENT (2010): Wanting Robustness in Macroeconomics, in *Handbook of Monetary Economics*, B. M. FRIEDMAN and M. WOODFORD, eds., vol. 3 of Handbook of Monetary Economics, Amsterdam: Elsevier, pp. 1097–1157.

HOBSON, D., and M. KLIMMEK (2012): Model Independent Hedging Strategies for Variance Swaps, *Finance Stoch.* 16(4), 611–649.

HOBSON, D. G. (1998): Robust Hedging of the Lookback Option, *Finance Stoch.* 2(4), 329–347.

ISII, K. (1960): The Extrema of Probability Determined by Generalized Moments. I. Bounded Random Variables, *Ann. Inst. Statist. Math.* 12, 119–134; errata, 280.

KARLIN, S., and W. STUDDEN (1966): Tchebycheff Systems: With Applications in Analysis and Statistics, *Pure Appl. Math.* vol. XV, New York: Wiley Interscience.

KELLER-RESSEL, M., and J. MUHLE-KARBE (2013): Asymptotic and Exact Pricing of Options on Variance, *Finance Stoch.* 17(1), 107–133.

LYONS, T. J. (1995): Uncertain Volatility and the Risk-Free Synthesis of Derivatives, *Appl. Math Fin.* 2(2), 117–133.

NEUBERGER, A. (1994): The Log Contract: A New Instrument to Hedge Volatility, *J. Portfolio Manag.* 20(2), 74–80.

OBŁÓJ, J. (2004): The Skorokhod Embedding Problem and Its Offspring, *Prob. Surv.* 1, 321–392.

RAVAL, V. (2010): *Arbitrage Bounds for Prices of Options on Realized Variance*, Ph.D. thesis, University of London (Imperial College).

REVUZ, D., and M. YOR (1994): *Continuous Martingales and Brownian Motion*, Berlin: Springer-Verlag.

ROGERS, L. C. G., and D. WILLIAMS (2000): *Diffusions, Markov Processes, and Martingales*, Vol. 2, Cambridge, UK: Cambridge University Press.

SONDERMANN, D. (2006): *Introduction to Stochastic Calculus for Finance: A New Didactic Approach*, vol. 579 of Lecture Notes in Economics and Mathematical Systemtems. Berlin: Springer-Verlag.

VOVK, V. (2012): Continuous-Time Trading and the Emergence of Probability, *Finance Stoch.* 16(4): 561–609.

WUERMLI, M. (1980): *Lokalzeiten für Martingale*, unpublished diploma thesis supervised by Professor H. Föllmer, Universität Bonn.

MATHEMATICAL FINANCE

An International Journal of Mathematics, Statistics and Financial Economics

Mathematical Finance, Vol. 24, No. 4 (October 2014), 855–881

CLOSED FORM PRICING FORMULAS FOR DISCRETELY SAMPLED GENERALIZED VARIANCE SWAPS

WENDONG ZHENG AND YUE KUEN KWOK

Hong Kong University of Science and Technology

Most of the existing pricing models of variance derivative products assume continuous sampling of the realized variance processes, though actual contractual specifications compute the realized variance based on sampling at discrete times. We present a general analytic approach for pricing discretely sampled generalized variance swaps under the stochastic volatility models with simultaneous jumps in the asset price and variance processes. The resulting pricing formula of the gamma swap is in closed form while those of the corridor variance swaps and conditional variance swaps take the form of one-dimensional Fourier integrals. We also verify through analytic calculations the convergence of the asymptotic limit of the pricing formulas of the discretely sampled generalized variance swaps under vanishing sampling interval to the analytic pricing formulas of the continuously sampled counterparts. The proposed methodology can be applied to any affine model and other higher moments swaps as well. We examine the exposure to convexity (volatility of variance) and skew (correlation between the equity returns and variance process) of these discretely sampled generalized variance swaps. We explore the impact on the fair strike prices of these exotic variance swaps with respect to different sets of parameter values, like varying sampling frequencies, jump intensity, and width of the monitoring corridor.

KEY WORDS: generalized variance swaps, stochastic volatility models, Fourier transform, discrete sampling.

1. INTRODUCTION

Volatility measures the standard deviation of the logarithm of returns of an underlying asset, thus it gives a measure of the risk of holding that asset. Volatility risk has drawn a wider attention in the financial markets in recent years, especially after the global financial crisis. Volatility trading becomes an important topic of risk management. In a bearish market environment, volatility typically stays at a high level, so holding a long position of volatility may be useful in hedging an equity portfolio. Indeed, volatility can be viewed as an asset class in its own right. Investors may use volatility derivatives to perform directional trading of volatility levels, say, trading the spread between the realized and implied volatility levels, or hedging an implicit volatility exposure. These volatility derivatives are investment tools for investors with specific views on the future market volatility or with particular risk exposures by allowing them to deal with these views or risks without taking a direct position in the underlying asset and/or delta-hedging their position (Brockhaus and Long 2000).

This work was supported by the Hong Kong Research Grants Council under Project 642110 of the General Research Funds.

Manuscript received October 2011; final revision received July 2012.

Address correspondence to Yue Kuen Kwok, Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China; e-mail: maykwok@ust.hk.

DOI: 10.1111/mafi.12016

Volatility products can be generally classified into two types. The historical-variance-based volatility derivatives include products whose payoff depends on the realized variance of the underlying asset. Another class of volatility products are the implied-volatility-based products, like the VIX futures traded in the Chicago Board Options Exchange (CBOE). The VIX stands for the CBOE Volatility Index, and it measures the 30-day expected future volatility of the S&P 500 index (Carr and Wu 2006). In recent years, the third generation of volatility products, which are known as the generalized variance swaps, including the corridor variance swaps, conditional variance swaps, and gamma swaps, have gained wider popularity as volatility trading instruments. These exotic variance swaps can offer investors a more finely tuned volatility exposure than the traditional variance swaps (see the review articles by Carr and Lewis 2004; Bouzoubaa and Osseiran 2010; Lee 2010). Their product specifications and potential uses in hedging or betting the various forms of volatility exposures will be presented in later sections. One of the objectives of this paper is to present a systematic and efficient analytic approach for pricing these discretely sampled generalized variance swaps under the stochastic volatility models with simultaneous jumps in the asset price and variance processes.

Assuming that the stock prices evolve without jumps, Neuberger (1994) shows how a continuously sampled variance swap can be theoretically equivalent to a dynamically adjusted constant dollar exposure to the stock, in combination with a static long position in a portfolio of options and a forward contract that replicate the payoff of a log contract. Carr and Madan (1998) propose various methods of trading the realized volatility, like taking a static position in options, delta-hedging an option position, etc. They demonstrate that the delta-hedged option approach exhibits a large amount of path dependency in the underlying in the final profit/loss. On the other hand, volatility derivatives are shown to provide the pure exposure to realized market volatility without an inherent price path dependency. Demeterfi et al. (1999) provide a nice review on the pricing behavior and theory of both variance and volatility swaps.

As a common approximation assumption in the pricing models of volatility derivatives in the literature, the discretely sampled realized variance in the actual contractual specification is approximated by a continuously sampled variance (as quantified by the quadratic variation of the log asset price process). For some volatility products, it occurs that the assumption of continuous sampling falls short of providing pricing results with sufficient accuracy when the actual discrete sampling becomes less frequent. Since it is not so straightforward to estimate the approximation errors in a unified framework, practitioners trading on contracts that are based on the realized variance with a low sampling frequency cannot properly assess the pricing errors caused by the continuous sampling assumption. A review on the replication errors for the discretely monitored variance swaps can be found in Carr and Lee (2009). Keller-Ressel and Muhle-Karbe (2010) discuss the rate of convergence of the approximation of the realized variance via the notion of quadratic variation and examine the errors of the approximation in pricing short-dated options with nonlinear payoffs.

There have been numerous papers that consider pricing variance product contracts on the discretely sampled realized variance. Little and Pant (2001) develop a finite difference approach for the valuation of the discretely sampled variance swaps in an extended Black–Scholes framework with a local volatility function. They adopt an effective numerical technique to capture the jumps in the realized variance across the sampling dates. Windcliff, Forsyth, and Vetzal (2006) improve the pricing algorithm for the discretely sampled volatility derivatives by allowing jumps in the asset price process. Using the Monte Carlo simulation method, Broadie and Jain (2008) investigate the effect of discrete sampling and asset price jumps on the fair strike prices of variance and volatility swaps under various stochastic volatility models. Carr and Lee (2009) consider the replication of discretely sampled variance products (including exotic path dependent payoff structures) using options, futures, and bonds with the same sampling frequency as that of the variance products. Itkin and Carr (2010) use a forward characteristic function approach to price discretely monitored variance and volatility swaps under various Lévy models with stochastic time change. Crosby and Davis (2011) consider the pricing of generalized variance swaps, such as self-quantoed variance swaps, gamma swaps, skewness swaps, and proportional variance swaps under the time-changed Lévy processes. They show that the prices of discretely monitored variance swaps and their generalizations all converge to the prices of continuously monitored counterparts as $O(1/N)$, where N is the number of monitoring instants. Sepp (2011) analyzes the impact of discrete sampling on the pricing of options on the realized variance under Heston's stochastic volatility model. He proposes a method of mixing the discrete variance in a log-normal model and the quadratic variance in a stochastic volatility model that approximates well the distribution of the discrete variation. Drimus and Farkas (2010) show that conditioning on the realization of the instantaneous variance process, the residual randomness arising from discrete sampling follows a normal distribution. They also provide a practical analysis of the greeks sensitivity of options on discretely sampled variance. Following the Little-Pant pricing formulation, Zhu and Lian (2011) manage to derive closed form pricing formulas for the vanilla variance swaps under Heston's stochastic volatility model for the underlying asset price process by solving a coupled system of partial differential equations. Lian (2010) extends the above analytic pricing approach to the underlying asset price process that allows stochastic volatility with simultaneous jumps in both the asset price and variance process (SVSJ model). The success of analytic tractability in the Zhu–Lian approach lies on the exponential affine structure of the SVSJ model, where the corresponding analytic formulas of the marginal characteristic functions can be derived. When the payoff structures of the variance swap contracts become more exotic, like those of the corridor variance swaps, conditional variance swaps, and gamma swaps, the knowledge of the marginal characteristic functions alone may not be enough in the derivation of the corresponding closed form pricing formulas. Instead, the joint moment generating function plays a vital role in deriving the pricing formulas for exotic variance swaps.

In this paper, we propose a general analytic approach for pricing various types of discretely sampled generalized variance swaps, thanks to the availability of the analytical expression of the joint moment generating function of the underlying processes. The analytic derivation of the associated moment generating function under the SVSJ model can be accomplished via the solution to a Riccati system of ordinary differential equations. Provided that the payoff function of a generalized variance swap can be transformed into an exponential function of the state variables, closed form or semianalytic (in terms of one-dimensional Fourier integrals) pricing formula of the derivative product can be derived. Duffie, Pan, and Singleton (2000) and Chacko and Das (2002) demonstrate the versatility of this analytic approach in pricing various types of fixed income derivatives. These papers explore various invariant properties of the solutions to the Riccati systems and manage to express the pricing formulas in terms of these solutions. Sepp (2007) applies similar techniques to price continuously sampled variance derivatives and conditional variance swaps via the derivation of the analytic representation of the Green function associated with the governing partial integral-differential equation under the SVSJ model.

This paper is organized as follows. In the next section, we present the formulation of the SVSJ model with a discussion on various possible extensions of the underlying joint dynamics of the asset returns and its variance. Thanks to the exponential affine structures of the SVSJ model, we manage to obtain an analytic representation of the corresponding joint moment generating function by solving a Riccati system of ordinary differential equations. In Section 3, we present the product specification and potential uses of various generalized variance swaps. We then show how to derive the closed form pricing formula of each of these discretely sampled generalized variance swaps under the SVSJ model. The continuously sampled gamma swap is known to provide a constant *share* gamma exposure. We illustrate how this gamma exposure property is modified under discrete sampling. The pricing of the conditional variance swap requires the computation of the expected occupation time of the asset price within a specified corridor. For the formulas of the fair strikes of the discretely monitored variance swaps and gamma swaps, we take the asymptotic limit by letting the sampling interval approach zero and illustrate that one can recover the same set of formulas of their continuously sampled counterparts (Sepp 2008). Also, we manage to obtain the pricing formulas of corridor and conditional variance swaps under continuous sampling. In Section 4, we report the numerical tests that examine the convergence of the fair strike prices with increasing sampling frequencies to the fair strike price of the continuously sampled counterparts. Our numerical tests show that an almost linear rate of convergence with respect to the sampling interval of the fair strike price under discrete sampling to that under continuously sampling is revealed for variance swaps. However, such convergence behavior may not always be observed for other exotic variance swaps. We also examine the exposure to convexity (volatility of variance), skew (correlation between equity returns and variance process), and jump intensities of these exotic swap products under the more realistic framework of discrete sampling of the realized variance. The fair strike prices in these generalized variance swaps are shown to be dependent on their contractual specifications. Summary and conclusive remarks are presented in the last section.

2. STOCHASTIC VOLATILITY MODELS WITH SIMULTANEOUS JUMPS AND JOINT MOMENT GENERATING FUNCTION

There have been numerous empirical studies on the dynamics of asset returns that illustrate evidence for both jumps in the price level and its volatility. A prominent continuous time model that has been widely adopted is the affine simultaneous jump model (Duffie et al. 2000) where the asset return and its variance follow the jump-diffusion process for which the drift, covariance, and jump intensities are assumed to have an affine dependence on the state vector. The analytic pricing under the affine simultaneous jump model can be performed by solving the corresponding Riccati system of ordinary differential equations (Chacko and Das 2002).

In this paper, we adopt the following stochastic volatility model with simultaneous jumps (SVSJ) to describe the joint dynamics of the stock price S_t and its instantaneous variance V_t. Under the risk neutral pricing measure Q, the joint dynamics of S_t and V_t assumes the form:

$$\begin{cases} \frac{\mathrm{d}S_t}{S_t} = (r - d - \lambda m)\,\mathrm{d}t + \sqrt{V_t}\,\mathrm{d}W_t^S + (e^{J^S} - 1)\,\mathrm{d}N_t, \\ \mathrm{d}V_t = \kappa(\theta - V_t)\,\mathrm{d}t + \varepsilon\sqrt{V_t}\,\mathrm{d}W_t^V + J^V\mathrm{d}N_t, \end{cases} \tag{2.1}$$

where W_t^S and W_t^V are a pair of correlated standard Brownian motions with $\mathrm{d}W_t^S\mathrm{d}W_t^V = \rho\,\mathrm{d}t$, and N_t is a Poisson process with constant intensity λ that is independent of the two Brownian motions. Here, ρ is the constant correlation coefficient. We let J^S and J^V denote the random jump size of the log price and variance, respectively, and these random jump sizes are assumed to be independent of W_t^S, W_t^V, and N_t. Also, we let r and d denote the riskless interest rate and the constant dividend yield, respectively, and $m = E_t^Q[e^{J^S} - 1]$. Throughout the paper, all the expectation calculations $E_t^Q[\,\cdot\,]$ are performed under the risk neutral pricing measure Q and conditional on filtration $\mathcal{F}_t$ at the current time t. In the sequel, we suppress the superscript and subscript in the expectation operator for notational convenience.

It is well known that jumps in the stock price provide a more realistic description of the short-term behavior of the stock price dynamics while jumps in the variance give the more accurate modeling of the volatility skew. Various empirical studies reveal that jumps in the price level and variance in general occur together, and they are strongly interdependent and have opposite sign. One may argue that the above SVSJ model with specific affine forms of the parameter functions may be somewhat restrictive. Some recent nonparametric studies of the high frequency movements in stock market volatility reveal that volatility may follow quite different forms of jump behavior (Todorov and Tauchen 2010). Various extended versions of the stochastic volatility models have also been proposed. For example, Kangro, Parna, and Sepp (2004) propose to generalize the intensity of the Poisson process to be a nonreverting stochastic process. Carr and Wu (2007) assume stochastic hazard rate for the Poisson process, where the stochastic hazard rate parameter is assumed to be the sum of the instantaneous variance and a latent risk factor that follows a diffusion process with mean reversion drift rate. Cont and Kokholm (2008) model directly the forward variance swap rates for a discrete tenor of maturities, somewhat analogous to the LIBOR market model in interest rates modeling. In some of these extended models, analytic tractability that is similar to the SVSJ model can be maintained though the analytic procedures tend to become more involved. In this paper, we illustrate the set of analytic procedures of deriving the pricing formulas of discretely sampled exotic variance products under the popular SVSJ model and relegate the research on analytic pricing methods under other types of stochastic volatility models to future works.

2.1. Joint Moment Generating Function

For convenience, we let $X_t = \ln S_t$. The joint moment generating function of the joint process X_t and V_t is defined to be

$$E[\exp(\phi X_T + b V_T + \gamma)],$$

where ϕ, b, and γ are constant parameters. Let $U(X_t, V_t, t)$ denote the nondiscounted time-t value of a contingent claim with the terminal payoff function: $U(X_T, V_T, T)$, where T is the maturity date. By adopting the temporal variable, $\tau = T - t$, it can be deduced from the Feynman–Kac theorem that $U(X, V, \tau)$ is governed by the following partial integral-differential equation (PIDE):

$$\begin{aligned}(2.2)\qquad \frac{\partial U}{\partial \tau} &= \left(r - d - m\lambda - \frac{V}{2}\right)\frac{\partial U}{\partial X} + \kappa(\theta - V)\frac{\partial U}{\partial V} + \frac{V}{2}\frac{\partial^2 U}{\partial X^2} + \frac{\varepsilon^2 V}{2}\frac{\partial^2 U}{\partial V^2}\\ &\quad + \rho\varepsilon V\frac{\partial^2 U}{\partial X \partial V} + \lambda E[U(X + J^S, V + J^V, \tau) - U(X, V, \tau)].\end{aligned}$$

The terminal payoff function of the contingent claim becomes the initial condition of the PIDE. Note that the joint moment generating function (MGF) can be regarded as the time-t forward value of the contingent claim with the terminal payoff: $\exp(\phi X_T + bV_T + \gamma)$, so the MGF also satisfies PIDE (2.2).

Using the analytic procedures similar to those in interest rate derivatives pricing under the stochastic volatility with simultaneous jumps model (Chacko and Das 2002), analytic solution to the joint MGF can be obtained by solving a Riccati system of ordinary differential equations. Due to the exotic payoff structure of the generalized variance swaps, analytic form of the joint MGF is required in deriving analytic pricing formulas for these discretely sampled variance products. As a remark, it suffices to use the marginal MGFs to price discretely sampled vanilla variance swaps (Zhu and Lian 2011) due to their simpler payoff structure. Once the joint MGF is known, the respective marginal MGF can be obtained easily by setting the irrelevant parameters in the joint MGF to be zero. For example, the marginal MGF with respect to the state variable V can be obtained by setting $\phi = \gamma = 0$.

Thanks to the affine structure in the SVSJ model, $U(X, V, \tau)$ admits an analytic solution of the following form (Duffie et al. 2000):

$$U(X, V, \tau) = \exp(\phi X + B(\Theta; \tau, \mathbf{q})V + \Gamma(\Theta; \tau, \mathbf{q}) + \Lambda(\Theta; \tau, \mathbf{q})), \tag{2.3}$$

where the parameter functions $B(\Theta; \tau, \mathbf{q})$, $\Gamma(\Theta; \tau, \mathbf{q})$, and $\Lambda(\Theta; \tau, \mathbf{q})$ satisfy the following Riccati system of ordinary differential equations:

$$\begin{cases} \dfrac{\partial B}{\partial \tau} = -\dfrac{1}{2}(\phi - \phi^2) - (\kappa - \rho\varepsilon\phi)B + \dfrac{\varepsilon^2}{2}B^2, \\ \dfrac{\partial \Gamma}{\partial \tau} = (r - d)\phi + \kappa\theta B, \\ \dfrac{\partial \Lambda}{\partial \tau} = \lambda\big(E[\exp(\phi J^S + BJ^V) - 1] - m\phi\big), \end{cases} \tag{2.4}$$

with the initial conditions: $B(0) = b, \Gamma(0) = \gamma$, and $\Lambda(0) = 0$. Here, $\mathbf{q} = (\phi\ b\ \gamma)^T$ and we use Θ to indicate the dependence of these parameter functions on the model parameters in the SVSJ model. One has to specify the distributions for J^S and J^V in order to obtain a complete solution to these parameter functions. For example, suppose we assume that $J^V \sim \exp(1/\eta)$ (exponential distribution with parameter rate $1/\eta$) and J^S follows:

$$J^S \mid J^V \sim \text{Normal}(\nu + \rho_J J^V, \delta^2),$$

that is the Gaussian distribution with mean $\nu + \rho_J J^V$ and variance δ^2, we obtain

$$m = E[e^{J^S} - 1] = \frac{e^{\nu+\delta^2/2}}{1 - \eta\rho_J} - 1, \tag{2.5}$$

provided that $\eta\rho_J < 1$. Under the above assumption on J^S and J^V, the parameter functions can be found to be

$$\begin{cases} B(\Theta;\tau,\mathbf{q}) = \dfrac{b(\xi_- e^{-\zeta\tau} + \xi_+) - (\phi - \phi^2)(1 - e^{-\zeta\tau})}{(\xi_+ + \varepsilon^2 b)e^{-\zeta\tau} + \xi_- - \varepsilon^2 b}, \\ \Gamma(\Theta;\tau,\mathbf{q}) = (r - d)\phi\tau + \gamma - \dfrac{\kappa\theta}{\varepsilon^2}\left[\xi_+\tau + 2\ln\dfrac{(\xi_+ + \varepsilon^2 b)e^{-\zeta\tau} + \xi_- - \varepsilon^2 b}{2\zeta}\right], \\ \Lambda(\Theta;\tau,\mathbf{q}) = -\lambda(m\phi + 1)\tau + \lambda e^{\phi\nu + \delta^2\phi^2/2}\left[\dfrac{k_2}{k_4}\tau - \dfrac{1}{\zeta}\left(\dfrac{k_1}{k_3} - \dfrac{k_2}{k_4}\right)\ln\dfrac{k_3 e^{-\zeta\tau} + k_4}{k_3 + k_4}\right], \end{cases} \tag{2.6}$$

with $\mathbf{q} = (\phi\ b\ \gamma)^T$ and

$$\begin{aligned} \zeta &= \sqrt{(\kappa - \rho\varepsilon\phi)^2 + \varepsilon^2(\phi - \phi^2)}, \\ \xi_\pm &= \zeta \mp (\kappa - \rho\varepsilon\phi), \\ k_1 &= \xi_+ + \varepsilon^2 b, \\ k_2 &= \xi_- - \varepsilon^2 b, \\ k_3 &= (1 - \phi\rho_J\eta)k_1 - \eta(\phi - \phi^2 + \xi_- b), \\ k_4 &= (1 - \phi\rho_J\eta)k_2 + \eta(\phi - \phi^2 - \xi_+ b). \end{aligned}$$

The derivation of the parameter functions in equation (2.6) is provided in Appendix A. Once the joint MGF is available, an effective analytic pricing approach can be constructed to derive the pricing formulas of the various types of discretely sampled generalized variance swaps. As an illustration, we demonstrate how the pricing formula of the vanilla variance swap can be readily derived. Let the sampling dates be denoted by $0 = t_0 < t_1 < \cdots < t_N = T$. On the maturity date T, the payoff of the vanilla variance swap is defined to be

$$\frac{A}{N}\sum_{k=1}^{N}\left(\ln\frac{S_{t_k}}{S_{t_{k-1}}}\right)^2 - K,$$

where K is the strike price of the variance swap and A is the annualized factor (say, we take $A = 252$ for daily sampling). We write $\Delta t_k = t_k - t_{k-1}$ and express the time interval $t_{k-1} - t_0$ simply as t_{k-1} since t_0 is taken to be zero. The pricing problem amounts to finding the fair strike price K such that the value of the vanilla variance swap at initiation is zero. The fair strike price K is then given by

$$K = E\left[\frac{A}{N}\sum_{k=1}^{N}\left(\ln\frac{S_{t_k}}{S_{t_{k-1}}}\right)^2\right].$$

We now show how to evaluate each term in the above summation. Using the known analytic expression of the marginal MGFs, we apply the tower rule in conditional expectation to obtain the expectation of a typical term as follows:

$$
\begin{aligned}
E\left[\left(\ln \frac{S_{t_k}}{S_{t_{k-1}}}\right)^2\right] &= E\left[\frac{\partial^2}{\partial \phi^2} e^{\phi(X_{t_k} - X_{t_{k-1}})}\right]\bigg|_{\phi=0} \\
&= \frac{\partial^2}{\partial \phi^2} E\big[E\big[e^{\phi X_{t_k}} \big| X_{t_{k-1}}, V_{t_{k-1}}\big] e^{-\phi X_{t_{k-1}}}\big]\big|_{\phi=0} \\
&= \frac{\partial^2}{\partial \phi^2} E\big[e^{B(\Theta;\Delta t_k,\mathbf{q}_1) V_{t_{k-1}} + \Gamma(\Theta;\Delta t_k,\mathbf{q}_1) + \Lambda(\Theta;\Delta t_k,\mathbf{q}_1)}\big]\big|_{\phi=0} \\
&= \frac{\partial^2}{\partial \phi^2} e^{B(\Theta;t_{k-1},\mathbf{q}_2) V_0 + \Gamma(\Theta;t_{k-1},\mathbf{q}_2) + \Lambda(\Theta;t_{k-1},\mathbf{q}_2)}\big|_{\phi=0},
\end{aligned}
$$

where $\mathbf{q}_1 = (\phi\ 0\ 0)^T$, and

$$
\mathbf{q}_2 = \begin{pmatrix} 0 \\ B(\Theta; \Delta t_k, \mathbf{q}_1) \\ \Gamma(\Theta; \Delta t_k, \mathbf{q}_1) + \Lambda(\Theta; \Delta t_k, \mathbf{q}_1) \end{pmatrix}.
$$

The fair strike price of the variance swap is then given by

$$
K_V(T, N) = \frac{A}{N} \sum_{k=1}^{N} \frac{\partial^2}{\partial \phi^2} e^{B(\Theta;t_{k-1},\mathbf{q}_2) V_0 + \Gamma(\Theta;t_{k-1},\mathbf{q}_2) + \Lambda(\Theta;t_{k-1},\mathbf{q}_2)}\bigg|_{\phi=0}. \tag{2.7}
$$

Note that the derivation procedure is less involved compared to the pricing approach used by Zhu and Lian (2011).

2.2. Asymptotic Limit of Vanishing Sampling Interval

It would be instructive to examine whether we can deduce the formula for the fair strike price of the continuously sampled variance swap by taking the asymptotic limit as $\Delta t \to 0$, where $\Delta t = \max_k \Delta t_k$, in formula (2.7). By expanding the parameter functions B, Γ, and Λ in powers of Δt_k and taking $\Delta t \to 0$ subsequently, we manage to obtain the following closed form formula for the fair strike of the continuously sampled variance swap:

$$
\begin{aligned}
K_V(T, \infty) = \frac{1}{T}\bigg\{ &\frac{1 - e^{-\kappa T}}{\kappa} V_0 - \frac{\lambda \eta}{\kappa^2}(1 - e^{-\kappa T} - \kappa T) \\
&+ \lambda\big[\delta^2 + \rho_J^2 \eta^2 + (\nu + \rho_J \eta)^2\big] T + \frac{\theta}{\kappa}(\kappa T - 1 + e^{-\kappa T})\bigg\},
\end{aligned} \tag{2.8}
$$

where we have used the convention $\frac{A}{N} = \frac{1}{T}$. The above formula is in agreement with a similar pricing formula in Sepp (2008). The proof of formula (2.8) is presented in Appendix B.

In the next section, we illustrate how to generalize the above pricing approach to find the pricing formulas for the various types of generalized variance swaps.

3. GENERALIZED VARIANCE SWAPS

Vanilla variance swaps are known to be the appropriate instruments to provide investors with pure volatility exposure. In recent years, various types of generalized variance swaps have been introduced in the financial markets to enhance volatility trading by providing asymmetric bets or hedges on volatility. Given a tenor structure $\{t_0, t_1, \ldots, t_N\}$ as before, the generalized realized variance over the period $[t_0, t_N]$ is defined to be

$$\frac{A}{N}\sum_{k=1}^{N} w_k \left(\ln \frac{S_{t_k}}{S_{t_{k-1}}}\right)^2,$$

where w_k is some discrete weight process chosen so as to target a specific form of volatility exposure. In this section, analytic pricing of discretely sampled gamma swaps, corridor variance swaps, and conditional variance swaps would be considered. For each type of these exotic variance products, we start with the description of their product nature and potential uses in volatility trading and hedging.

3.1. Gamma Swaps

In a gamma swap, the weight w_k is chosen to be $\frac{S_{t_k}}{S_{t_0}}$, $k = 1, 2, \ldots, N$. Accordingly, the terminal payoff of the gamma swap is defined by

$$\frac{A}{N}\sum_{k=1}^{N} \frac{S_{t_k}}{S_{t_0}} \left(\ln \frac{S_{t_k}}{S_{t_{k-1}}}\right)^2 - K.$$

The motivation of choosing the weight to be the underlying level is to provide the embedded damping of the large downside variance when the stock price falls close to zero. This feature serves to protect the swap seller from crash risk, thus provides an advantage over the noncapped variance swaps. Another motivation is related to variance dispersion trade, which refers to trading the difference between the realized index volatility and the market-cap weighted sum of the realized volatilities of its constituents. Gamma swaps provide better means to trade dispersion than vanilla variance swaps since the risk associated with changes in weights of the constituent stocks over the life of the variance swap is reduced (Jacquier and Slaoui 2010).

To find the analytic fair strike price of a discretely sampled gamma swap, we compute the expectation of a typical term in the floating leg of the gamma swap as follows:

$$\begin{aligned}
E\left[\frac{S_{t_k}}{S_{t_0}}\left(\ln\frac{S_{t_k}}{S_{t_{k-1}}}\right)^2\right] &= e^{-X_0} E\big[e^{X_{t_k}-X_{t_{k-1}}}(X_{t_k}-X_{t_{k-1}})^2 e^{X_{t_{k-1}}}\big] \\
&= e^{-X_0} E\left[\frac{\partial^2}{\partial\phi^2} e^{\phi(X_{t_k}-X_{t_{k-1}})+X_{t_{k-1}}}\right]\Bigg|_{\phi=1} \\
&= e^{-X_0}\frac{\partial^2}{\partial\phi^2} E\big[E\big[e^{\phi X_{t_k}}\,\big|\,X_{t_{k-1}},\, V_{t_{k-1}}\big]e^{(1-\phi)X_{t_{k-1}}}\big]\big|_{\phi=1} \\
&= e^{-X_0}\frac{\partial^2}{\partial\phi^2} E\big[e^{X_{t_{k-1}}+B(\Theta;\Delta t_k,\mathbf{q}_1)V_{t_{k-1}}+\Gamma(\Theta;\Delta t_k,\mathbf{q}_1)+\Lambda(\Theta;\Delta t_k,\mathbf{q}_1)}\big]\big|_{\phi=1} \\
&= \frac{\partial^2}{\partial\phi^2} e^{B(\Theta;t_{k-1},\mathbf{q}_2)V_0+\Gamma(\Theta;t_{k-1},\mathbf{q}_2)+\Lambda(\Theta;t_{k-1},\mathbf{q}_2)}\Bigg|_{\phi=1},
\end{aligned}$$

where $\mathbf{q}_1 = (\phi\ 0\ 0)^T$, and

$$\mathbf{q}_2 = \begin{pmatrix} 1 \\ B(\Theta; \Delta t_k, \mathbf{q}_1) \\ \Gamma(\Theta; \Delta t_k, \mathbf{q}_1) + \Lambda(\Theta; \Delta t_k, \mathbf{q}_1) \end{pmatrix}.$$

In the above derivation procedure, we have made use of the analytic form of the joint moment generating function of X_t and V_t. The fair strike price of the gamma swap is then given by

$$(3.1) \qquad K_\Gamma(T, N) = \frac{A}{N} \sum_{k=1}^{N} \frac{\partial^2}{\partial \phi^2} e^{B(\Theta; t_{k-1}, \mathbf{q}_2) V_0 + \Gamma(\Theta; t_{k-1}, \mathbf{q}_2) + \Lambda(\Theta; t_{k-1}, \mathbf{q}_2)} \Bigg|_{\phi=1},$$

which is seen to have no dependence on the initial price level S_{t_0}.

It is well known that the gamma exposure of vanilla variance swaps is insensitive to the underlying level of share price (Demeterifi et al. 1999), a property known as constant *cash* gamma exposure. Suppose an investor is interested in the gamma exposure in the number of portfolio units rather than the initial cash value of the portfolio, the gamma swaps provide constant *share* gamma exposure by choosing weights that are set equal to the underlying level at the sampling dates. The proof of this property of constant (almost constant) share gamma exposure for the continuously (discretely) sampled gamma swaps is shown below.

3.1.1. Constant Share Gamma Exposure. We would like to compute the share gamma exposure of an in-progress gamma swap at time t, where $t_{i-1} < t \leq t_i, i \geq 1$. The share gamma is defined to be

$$(3.2) \qquad \Gamma_S = \Gamma S_t := \frac{\partial^2 V_t}{\partial S_t^2} S_t,$$

where Γ is the usual greek gamma of a derivative product with value function V_t. Consider a hedging portfolio Π whose differential change of value is given by

$$\mathrm{d}\Pi = \Delta \mathrm{d}S_t + \frac{1}{2}\Gamma(\mathrm{d}S_t)^2.$$

Since $\mathrm{d}S_t$ is in proportion to the share price S_t, the "cash gamma" of the derivative ΓS_t^2 is in units of dollars. After normalizing the cash gamma by the share price, the resulting quantity ΓS_t is in units of shares and hence it is called the "share gamma." To compute the share gamma of a discretely sampled gamma swap, it is necessary to find the time-t value of the gamma swap as follows:

$$(3.3) \qquad V_t = e^{-r(t_N - t)} \Bigg\{ \frac{A}{N} \sum_{k=1}^{i-1} \frac{S_{t_k}}{S_{t_0}} \left(\ln \frac{S_{t_k}}{S_{t_{k-1}}} \right)^2 + \frac{A}{N} E_t \left[\frac{S_{t_i}}{S_{t_0}} \left(\ln \frac{S_{t_i}}{S_{t_{i-1}}} \right)^2 \right]$$
$$+ \frac{A}{N} E_t \left[\sum_{k=i+1}^{N} \frac{S_{t_k}}{S_{t_0}} \left(\ln \frac{S_{t_k}}{S_{t_{k-1}}} \right)^2 \right] - K_\Gamma(T, N) \Bigg\}$$

$$= e^{-r(t_N-t)} \left\{ \frac{A}{N} \sum_{k=1}^{i-1} \frac{S_{t_k}}{S_{t_0}} \left(\ln \frac{S_{t_k}}{S_{t_{k-1}}} \right)^2 + \frac{A}{N} \frac{S_{t_{i-1}}}{S_{t_0}} \right.$$
$$\times \frac{\partial^2}{\partial \phi^2} \left[\left(\frac{S_t}{S_{t_{i-1}}} \right)^{\phi} e^{B(\Theta;t_i-t,\mathbf{q}_1)V_t+\Gamma(\Theta;t_i-t,\mathbf{q}_1)+\Lambda(\Theta;t_i-t,\mathbf{q}_1)} \right] \Bigg|_{\phi=1} + \frac{A}{N} \frac{S_t}{S_{t_0}}$$
$$\left. \times \sum_{k=i+1}^{N} \frac{\partial^2}{\partial \phi^2} e^{B(\Theta;t_{k-1}-t,\mathbf{q}_2)V_t+\Gamma(\Theta;t_{k-1}-t,\mathbf{q}_2)+\Lambda(\Theta;t_{k-1}-t,\mathbf{q}_2)} \Bigg|_{\phi=1} - K_\Gamma(T, N) \right\},$$

where $\mathbf{q}_1 = (\phi\ 0\ 0)^T$, and

$$\mathbf{q}_2 = \begin{pmatrix} 1 \\ B(\Theta; \Delta t_k, \mathbf{q}_1) \\ \Gamma(\Theta; \Delta t_k, \mathbf{q}_1) + \Lambda(\Theta; \Delta t_k, \mathbf{q}_1) \end{pmatrix}.$$

By differentiating equation (3.3) with respect to S_t twice and multiplying by S_t, we obtain the share gamma as follows:

$$\Gamma_S = e^{-r(t_N-t)} \frac{A}{N} \frac{2}{S_{t_0}} \left[\left(\ln \frac{S_t}{S_{t_{i-1}}} + 1 \right) F(1) + F'(1) \right], \tag{3.4a}$$

where

$$F(\phi) = e^{B(\Theta;t_i-t,\mathbf{q}_1)V_t+\Gamma(\Theta;t_i-t,\mathbf{q}_1)+\Lambda(\Theta;t_i-t,\mathbf{q}_1)}.$$

The dependence of Γ_S on S_t appears in the term $\ln \frac{S_t}{S_{t_{i-1}}}$. Suppose we take the limit of continuous sampling, $S_t \to S_{t_{i-1}}$, we then obtain

$$\Gamma_S \to e^{-r(t_N-t)} \frac{A}{N} \frac{2}{S_{t_0}}. \tag{3.4b}$$

The above limit has no dependence on S_t, so it verifies the property of constant gamma exposure under continuous sampling. A similar result has been obtained by Jacquier and Slaoui (2010) using a replication argument.

3.1.2. Fair Strike Prices of Continuously Monitored Gamma Swaps. By following a similar procedure of taking the limit of vanishing sampling time interval (see Appendix B), we can obtain the following closed form formula for the fair strike price of a continuously monitored gamma swap

$$K_\Gamma(T, \infty) = \frac{1}{T} \left[\left(V_0 - \frac{\kappa\theta}{\kappa - \rho\varepsilon} - C_2 \right) \frac{e^{(r-d-\kappa+\rho\varepsilon)T} - 1}{r - d - \kappa + \rho\varepsilon} \right.$$
$$\left. + \left(\frac{\kappa\theta}{\kappa - \rho\varepsilon} + C_1 + C_2 \right) \frac{e^{(r-d)T} - 1}{r - d} \right], \tag{3.5}$$

where

$$C_1 = \frac{\lambda e^{\nu+\delta^2/2}}{1-\rho_J\eta}\left[\left(\nu+\delta^2+\frac{\rho_J\eta}{1-\rho_J\eta}\right)^2+\delta^2+\left(\frac{\rho_J\eta}{1-\rho_J\eta}\right)^2\right],$$

$$C_2 = \frac{\lambda\eta e^{\nu+\delta^2/2}}{(1-\rho_J\eta)^2(\kappa-\rho\varepsilon)}.$$

3.2. Corridor Variance Swaps

A corridor variance swap differs from the vanilla variance swap in that the underlying price must fall inside a specified corridor $(L, U]\,(L \geq 0, U < \infty)$, in order for its squared return to be included in the floating leg of the corridor variance swap. For a discretely sampled corridor variance swap with the tenor $0 = t_0 < t_1 < t_2 < \cdots < t_N = T$, suppose the corridor is monitored on the underlying price at the old time level t_{k-1} for the kth squared log return, the floating leg with the corridor $(L, U]$ is given by

$$\frac{A}{N}\sum_{k=1}^{N}\left(\ln\frac{S_{t_k}}{S_{t_{k-1}}}\right)^2 \mathbf{1}_{\{L<S_{t_{k-1}}\leq U\}}.$$

Here, $\mathbf{1}_{\{\,.\,\}}$ denotes the indicator function. Corridor variance swaps with a one-sided barrier are also widely traded in the financial markets, where the downside-variance swap and upside-variance swap can be obtained by taking $L = 0$ and $U \to \infty$, respectively. As further generalizations, one can choose to have the corridor monitored on the underlying price at the new time level t_k (Sepp 2007) or even at both time levels (Carr and Lewis 2004).

Corridor variance swaps allow the investors to take their views on the implied volatility skew. Suppose the implied volatility skew is expected to steepen, the investor may benefit from buying a downside-variance swap and selling an upside-variance swap if this view is realized. Also, investors seeking crash protection may buy the downside-variance swap since it can provide almost the same level of crash protection as the vanilla variance swap but at a lower premium.

It suffices to consider pricing downside-variance swaps alone since the payoffs of downside-variance swaps of varying values of the upper barrier are sufficient to span all different payoffs of various corridor variance swaps. We would like to find the fair strike price of a downside-variance swap with an upper barrier U whose payoff at maturity T is given by

$$\frac{A}{N}\sum_{k=1}^{N}\left(\ln\frac{S_{t_k}}{S_{t_{k-1}}}\right)^2 \mathbf{1}_{\{S_{t_{k-1}}\leq U\}} - K.$$

Let us consider the expectation calculation of a typical term:

(3.6)

$$\begin{aligned}
E\left[\left(\ln\frac{S_{t_k}}{S_{t_{k-1}}}\right)^2 \mathbf{1}_{\{S_{t_{k-1}}\leq U\}}\right] &= E\left[E\left[\frac{\partial^2}{\partial\phi^2}e^{\phi(X_{t_k}-X_{t_{k-1}})}\,\middle|\, X_{t_{k-1}},\, V_{t_{k-1}}\right]\mathbf{1}_{\{X_{t_{k-1}}\leq \ln U\}}\right]\Bigg|_{\phi=0}\\
&= E\left[\frac{\partial^2}{\partial\phi^2}e^{B(\Theta;\Delta t_k,\mathbf{q}_1)V_{t_{k-1}}+\Gamma(\Theta;\Delta t_k,\mathbf{q}_1)+\Lambda(\Theta;\Delta t_k,\mathbf{q}_1)}\mathbf{1}_{\{X_{t_{k-1}}\leq \ln U\}}\right]\Bigg|_{\phi=0}\\
&= \frac{\partial^2}{\partial\phi^2}E\left[e^{B(\Theta;\Delta t_k,\mathbf{q}_1)V_{t_{k-1}}+\Gamma(\Theta;\Delta t_k,\mathbf{q}_1)+\Lambda(\Theta;\Delta t_k,\mathbf{q}_1)}\mathbf{1}_{\{X_{t_{k-1}}\leq \ln U\}}\right]\Bigg|_{\phi=0},
\end{aligned}$$

where $\mathbf{q}_1 = (\phi\ 0\ 0)^T$. For $k = 1$, the above expectation is readily seen to be

$$\text{(3.7a)}\quad E\left[\left(\ln\frac{S_{t_1}}{S_{t_0}}\right)^2 \mathbf{1}_{\{S_{t_0}\leq U\}}\right] = \left.\frac{\partial^2}{\partial\phi^2} e^{B(\Theta;\Delta t_1,\mathbf{q}_1)V_0+\Gamma(\Theta;\Delta t_1,\mathbf{q}_1)+\Lambda(\Theta;\Delta t_1,\mathbf{q}_1)}\mathbf{1}_{\{X_0\leq \ln U\}}\right|_{\phi=0}.$$

For $k \geq 2$, the evaluation of expectation in formula (3.6) requires the representation of the indicator function $\mathbf{1}_{\{X_{t_{k-1}}\leq \ln U\}}$ in terms of an inverse Fourier transform integral. As a result, formula (3.6) can be expressed in terms of a Fourier integral as follows:

(3.7b)

$$E\left[\left(\ln\frac{S_{t_k}}{S_{t_{k-1}}}\right)^2 \mathbf{1}_{\{S_{t_{k-1}}\leq U\}}\right] = \frac{e^{w_i(X_0-u)}}{\pi}\int_0^\infty \mathrm{Re}\left(e^{-\mathrm{i}w_r(X_0-u)}\frac{F_k(w_r+\mathrm{i}w_i)}{\mathrm{i}w_r - w_i}\right)\mathrm{d}w_r, \qquad k\geq 2,$$

where $w = w_r + \mathrm{i}w_i$, $u = \ln U$, and

$$F_k(w) = \left.\frac{\partial^2}{\partial\phi^2} e^{B(\Theta;t_{k-1},\mathbf{q}_2)V_0+\Gamma(\Theta;t_{k-1},\mathbf{q}_2)+\Lambda(\Theta;t_{k-1},\mathbf{q}_2)}\right|_{\phi=0}, \qquad k\geq 2,$$

with

$$\mathbf{q}_2 = \begin{pmatrix} -\mathrm{i}w \\ B(\Theta;\Delta t_k,\mathbf{q}_1) \\ \Gamma(\Theta;\Delta t_k,\mathbf{q}_1)+\Lambda(\Theta;\Delta t_k,\mathbf{q}_1)\end{pmatrix}.$$

The above Fourier integral is regular provided that w_i is appropriately chosen to lie within $(-\infty, 0)$. The proof of equation (3.7b) is presented in Appendix C. The fair strike price of the downside-variance swap is then given by

$$\text{(3.8)}\quad K_D(T,N) = \frac{A}{N}\left[\left.\frac{\partial^2}{\partial\phi^2} e^{B(\Theta;\Delta t_1,\mathbf{q}_1)V_0+\Gamma(\Theta;\Delta t_1,\mathbf{q}_1)+\Lambda(\Theta;\Delta t_1,\mathbf{q}_1)}\mathbf{1}_{\{X_0\leq \ln U\}}\right|_{\phi=0} + \frac{e^{w_i(X_0-u)}}{\pi}\int_0^\infty \mathrm{Re}\left(e^{-\mathrm{i}w_r(X_0-u)}\frac{\sum_{k=2}^{N}F_k(w_r+\mathrm{i}w_i)}{\mathrm{i}w_r-w_i}\right)\mathrm{d}w_r\right].$$

The evaluation of the Fourier integral in equation (3.8) can be effected by adopting the fast Fourier transform (FFT) algorithm. Actually, by following a similar FFT calculation approach as in Carr and Madan (1999), one can produce the fair strike prices for all downside-variance swaps with varying values of the upper barrier using one single FFT calculation.

3.2.1. An Alternative Definition for the Corridor Variance. We have considered the discretely sampled downside-variance swaps with the breaching of the downside corridor $(0,\ U]$ being monitored on the stock price at the old time level. However, there is an alternative definition in the literature, where the floating leg payoff is defined to be

$$\frac{A}{N}\sum_{k=1}^{N}\left(\ln\frac{S_{t_k}}{S_{t_{k-1}}}\right)^2 \mathbf{1}_{\{S_{t_k}\leq U\}}.$$

In this new definition, the breaching of the downside corridor for the kth squared log return is monitored on the stock price at the new time level. For $k = 1, 2, \ldots, N$, by following a similar procedure as shown in equation (3.6), we manage to obtain

$$E\left[\left(\ln\frac{S_{t_k}}{S_{t_{k-1}}}\right)^2 \mathbf{1}_{\{S_{t_k}\leq U\}}\right] = \frac{e^{w_i(X_0-u)}}{\pi}\int_0^\infty \mathrm{Re}\left(e^{-\mathrm{i}w_r(X_0-u)}\frac{F_k(w_r+\mathrm{i}w_i)}{\mathrm{i}w_r-w_i}\right)\mathrm{d}w_r, \tag{3.9}$$

where $w = w_r + \mathrm{i}w_i$, w_i is chosen to lie in $(-\infty, 0)$ as before, $u = \ln U$, and

$$F_k(w) = \left.\frac{\partial^2}{\partial\phi^2}e^{B(\Theta;t_{k-1},\mathbf{q}_2)V_0+\Gamma(\Theta;t_{k-1},\mathbf{q}_2)+\Lambda(\Theta;t_{k-1},\mathbf{q}_2)}\right|_{\phi=0},$$

with $\mathbf{q}_1 = (\phi - \mathrm{i}w\ 0\ 0)^T$ and

$$\mathbf{q}_2 = \begin{pmatrix} -\mathrm{i}w \\ B(\Theta;\Delta t_k,\mathbf{q}_1) \\ \Gamma(\Theta;\Delta t_k,\mathbf{q}_1)+\Lambda(\Theta;\Delta t_k,\mathbf{q}_1)\end{pmatrix}.$$

In Section 4, we will investigate the impact of the alternative definition on the fair strike price of a downside-variance swap.

3.2.2. Fair Strike Prices for Continuously Monitored Downside-Variance Swaps. By taking the asymptotic limit of vanishing sampling time interval, the fair strike price of the continuously sampled downside-variance swaps is given by (see Appendix B)

$$K_D(T,\infty) = \frac{e^{w_i(X_0-u)}}{T\pi}\int_0^\infty\int_0^T \mathrm{Re}\left(e^{-\mathrm{i}w_r(X_0-u)}\frac{F(w_r+\mathrm{i}w_i,t)}{\mathrm{i}w_r-w_i}\right)\mathrm{d}t\,\mathrm{d}w_r, \tag{3.10}$$

where

$$\begin{aligned}F(w,t) = e^{B^0(-\mathrm{i}w,t)V_0+\Gamma^0(-\mathrm{i}w,t)+\Lambda^0(-\mathrm{i}w,t)}\big\{&B^1(-\mathrm{i}w,t)V_0+\Gamma^1(-\mathrm{i}w,t)+\Lambda^1(-\mathrm{i}w,t)\\ &+\lambda\big[(\nu+\rho_J\eta)^2+\delta^2+\rho_J^2\eta^2\big]\big\},\end{aligned}$$

and the coefficient functions $B^0(-\mathrm{i}w,t)$, $B^1(-\mathrm{i}w,t)$, $\Gamma^0(-\mathrm{i}w,t)$, $\Gamma^1(-\mathrm{i}w,t)$, $\Lambda^0(-\mathrm{i}w,t)$, and $\Lambda^1(-\mathrm{i}w,t)$ are defined in Appendix B [see equation (B.3)].

3.3. Conditional Variance Swaps

A conditional variance swap is similar to a corridor variance swap, though they differ in the following two aspects:

(i) The accumulated sum of squared returns is divided by the number of observations D that the underlying asset price stays within the corridor instead of the total number of sampling observations N.
(ii) The final payoff to the holder is scaled by the ratio D/N.

Let K be the strike price of a conditional downside-variance swap and K' be the strike price of its corridor variance swap counterpart. The holder's payoff of the conditional downside-variance swap with corridor's upper barrier U is given by

$$\text{(3.11)}\quad \frac{D}{N}\left[\frac{A}{D}\sum_{k=1}^{N}\left(\ln\frac{S_{t_k}}{S_{t_{k-1}}}\right)^2\mathbf{1}_{\{S_{t_{k-1}}\le U\}}-K\right]=\left[\frac{A}{N}\sum_{k=1}^{N}\left(\ln\frac{S_{t_k}}{S_{t_{k-1}}}\right)^2\mathbf{1}_{\{S_{t_{k-1}}\le U\}}-K'\right]+\left(K'-K\frac{D}{N}\right),$$

where $D=\sum_{k=1}^{N}\mathbf{1}_{\{S_{t_{k-1}}\le U\}}$.

The above formula reveals that the conditional variance swap can be decomposed into a corridor variance swap with the same upper barrier plus a range accrual note. Since we have shown how to find the fair strike price of a downside-variance swap, it suffices to compute the expected number of the sampling dates at which the underlying stock price stays below the upper barrier U. For a typical term in $E[\sum_{k=1}^{N}\mathbf{1}_{\{S_{t_{k-1}}\le U\}}]$, the expectation calculation involves

$$E[\mathbf{1}_{\{S_{k-1}\le U\}}]=E[\mathbf{1}_{\{X_{k-1}\le u\}}],\quad \text{where } u=\ln U.$$

Following similar calculations as in Section 3.1.1, we obtain

$$\text{(3.12)}\quad E[\mathbf{1}_{\{X_{k-1}\le u\}}]=\frac{e^{w_i(X_0-u)}}{\pi}\int_0^\infty \mathrm{Re}\left(e^{-\mathrm{i}w_r(X_0-u)}\frac{G_k(w_r+\mathrm{i}w_i)}{\mathrm{i}w_r-w_i}\right)\mathrm{d}w_r,\quad k\ge 2,$$

where $w_i\in(-\infty,0)$ and

$$G_k(w)=e^{B(\Theta;t_{k-1},\mathbf{q}_1)V_0+\Gamma(\Theta;t_{k-1},\mathbf{q}_1)+\Lambda(\Theta;t_{k-1},\mathbf{q}_1)},$$

with $\mathbf{q}_1=(-\mathrm{i}w\ 0\ 0)^T$. Similarly, the numerical evaluation of the above Fourier integral can be done via the FFT algorithm. Finally, the fair strike price of the conditional downside-variance swap is given by

$$\text{(3.13)}\quad K_C(T,N)=K_D(T,N)\left[\frac{\mathbf{1}_{\{X_0\le u\}}}{N}+\frac{e^{w_i(X_0-u)}}{\pi N}\int_0^\infty \mathrm{Re}\left(e^{-\mathrm{i}w_r(X_0-u)}\frac{\sum_{k=2}^{N}G_k(w_r+\mathrm{i}w_i)}{\mathrm{i}w_r-w_i}\right)\mathrm{d}w_r\right]^{-1}.$$

The payoff of a conditional variance swap counts only the sampling dates at which the realized variance does accumulate (conditional on the underlying price lying within the corridor). Compared to the corridor variance swaps, the conditional variance swaps are structured specifically for investors who would like to be exposed *only* to volatility risk within a prespecified corridor. In a corridor variance swap, the actual amount of the occupation time that the underlying price falls within the corridor over the whole life of the swap has a significant effect on the profit and loss to its holder. However, the conditional variance swap is immunized from this risk since only the realized variance within the corridor matters. Indeed, the decomposition formula (3.11) shows that the holder of a conditional variance swap receives compensation from the range accrual note when the occupation time attains a lower value leading to a lower payoff in the corridor variance swap counterpart.

3.3.1. Fair Strike Prices for Continuously Monitored Conditional Variance Swaps. The fair strike price of the continuously sampled conditional downside-variance swap is related to that of the downside-variance swap presented in the previous section. The analytical representation formula is given by

(3.14)
$$K_C(T,\infty) = K_D(T,\infty)\left[\frac{e^{w_i(X_0-u)}}{\pi T}\int_0^\infty\int_0^T \mathrm{Re}\left(e^{-\mathrm{i}w_r(X_0-u)}\frac{G(w_r+\mathrm{i}w_i,t)}{\mathrm{i}w_r-w_i}\right)\mathrm{d}t\,\mathrm{d}w_r\right]^{-1},$$

where

$$G(w,t) = e^{B(\Theta;t,\mathbf{q}_1)V_0+\Gamma(\Theta;t,\mathbf{q}_1)+\Lambda(\Theta;t,\mathbf{q}_1)},$$

with $\mathbf{q}_1 = (-\mathrm{i}w\ 0\ 0)^T$.

4. NUMERICAL TESTS

In this section, we report the numerical calculations that were performed for testing accuracy of the analytic pricing formulas obtained for the various types of discretely sampled generalized variance swaps. We examine the impact of the sampling frequency of the realized variance on the fair strike prices of the discretely sampled variance swaps, gamma swaps, corridor variance swaps, and conditional variance swaps, and the convergence of the fair strike prices to those of their continuously monitored counterparts. In particular, we used different sets of model parameters in the numerical calculations in order to have a more comprehensive view of the behavior of convergence to the continuous limits. To show the sensitivity of the fair strike prices of these discretely sampled generalized variance swaps to different model parameters, we provide various plots of the fair strike prices against three key model parameters, namely, correlation coefficient, volatility of variance, and jump intensity. Finally, we investigate the impact of the corridor convention (whether breaching of the corridor is monitored on the stock price at the old time level or new time level) on the pricing of the weekly sampled downside-variance swaps with varying values of the maturity date. In our numerical examples, we adopted the set of parameter values shown in Table 4.1 that are calibrated to S&P 500 option prices on November 2, 1993 (Duffie et al. 2000) as our basic set of parameter values. In addition, we take $r = 3.19\%$, $d = 0$, $S_0 = 1$, and assume the number of trading days in 1 year to be 252. Unless otherwise stated, we consider 1-year swap contracts so that $A = N$ ($T = 1$) and take $U = 1$ as the upper barrier of the downside corridor in the corridor and conditional variance swaps.

TABLE 4.1
The Basic Set of Parameter Values of the SVSJ Model

κ	3.46	ν	−0.086
θ	$(0.0894)^2$	η	0.05
ε	0.14	λ	0.47
ρ	−0.82	ρ_J	−0.38
$\sqrt{V_0}$	0.087	δ	0.0001

4.1. Impact of Sampling Frequency

First, we present numerical results that explore the convergence behavior of the fair strike prices of different types of discretely monitored generalized variance swaps under varying sampling frequencies. In Table 4.2, we report the fair strike prices of the vanilla variance swaps, gamma swaps, downside-variance swaps, and conditional downside-variance swaps with varying sampling frequencies and different values of the correlation coefficient ρ. The choice of $\rho = -1$ infers the scenario where the market is highly leveraged as the asset return has perfect negative correlation with its variance while the less negative correlation coefficient $\rho = -0.3$ represents the market condition with mild leverage effect. The choice of $\rho = -0.82$ is taken from Table 4.1 (based on calibration from actual data of option prices). The values of the fair strike prices are all presented in variance points, which is the expected realized variance multiplied by 100^2. The fair strike prices for these discretely sampled generalized variance swaps are numerically calculated using the closed form pricing formulas derived in previous sections [see equations (2.7), (3.1), (3.8), and (3.13)]. The fair strike prices of the continuously sampled generalized variance swaps, corresponding to $N = \infty$, are computed using the formulas (2.8), (3.5), (3.10), and (3.14) that are deduced by our asymptotic analysis. It is seen from Table 4.2 that the fair strike prices of all variance swap products (with the exception of the continuously sampled variance swaps) have dependence on ρ. The variance swaps and gamma swaps are seen to be less sensitive to ρ for any sampling frequency specification when compared to the downside-variance and conditional downside-variance swaps. As $N \to \infty$ (vanishing sampling interval), all the fair strike prices of the discretely sampled generalized variance swaps converge to those of their continuously sampled counterparts. The good agreement between these numerical values of fair strike prices provide a check for accuracy of all these analytic pricing formulas. As an interesting observation of the

TABLE 4.2
Comparison of the Numerical Values of the Fair Strike Prices of Variance Swaps, Gamma Swaps, Downside-Variance Swaps, and Conditional Downside-Variance Swaps with Varying Sampling Frequencies and Different Values of the Correlation Coefficient ρ. Here, N is the Number of Return Samples within 1 year

Sampling frequency		$N = 4$ quarterly	$N = 12$ monthly	$N = 26$ biweekly	$N = 52$ weekly	$N = 252$ daily	$N = \infty$ continuous
Variance	$\rho = -1$	187.0839	183.4365	182.2551	181.7172	181.2759	181.1590
swaps	$\rho = -0.82$	186.7823	183.3154	182.1961	181.6870	181.2695	181.1590
	$\rho = -0.3$	185.9113	182.9654	182.0257	181.5998	181.2512	181.1590
Gamma	$\rho = -1$	170.1311	169.2752	169.2176	169.2203	169.2350	169.2407
swaps	$\rho = -0.82$	171.0131	169.9908	169.8749	169.8504	169.8426	169.8423
	$\rho = -0.3$	173.6134	172.0962	171.8081	171.7036	171.6293	171.6113
Downside	$\rho = -1$	111.5139	102.5147	101.3211	101.0009	100.8345	100.8043
variance	$\rho = -0.82$	110.5369	101.0294	99.6504	99.2447	99.0083	98.9599
swaps	$\rho = -0.3$	107.8140	96.8144	94.8855	94.2254	93.7809	93.6779
Conditional	$\rho = -1$	216.8810	250.5501	265.4668	272.9108	279.2977	281.0162
variance	$\rho = -0.82$	213.6660	244.5615	258.3023	265.1702	271.0668	272.6579
swaps	$\rho = -0.3$	204.5881	227.7824	238.2826	243.5650	248.1260	249.3580

pattern of convergence of the fair strike prices with respect to sampling frequency, the convergence can be from above or below. To present a better visual view on the convergence behavior of the four types of generalized variance swaps, we show various plots of the percentage difference in the fair strike price against varying values of sampling interval Δt (in units of year) for different sets of model parameters (see Figure 4.1). The percentage difference, as a measurement of the discretization effect, is defined to be $100(\frac{K(\Delta t)}{K(0)} - 1)$. Here, $K(\Delta t)$ is the fair strike price with sampling interval Δt and $K(0)$ is the corresponding continuous limit of $\Delta t \to 0^+$. Among the four curves of percentage difference in fair strike plotted for each product, the curve labeled "basic" is obtained based on the basic set of model parameters in Table 4.1. The curve labeled "$\rho = -1$" is computed by changing the parameter value of ρ to -1 while keeping other model parameters unchanged, and similar interpretation for the other two curves labeled "$\varepsilon = 0.25$" and "$\lambda = 0.6$." It is observed that the impact of sampling frequency on the fair strike prices for most generalized variance swaps is small for the chosen range of Δt (the lowest sampling frequency being weekly in the plots). In particular, the fair strike prices of the gamma swaps show the least sensitivity to sampling frequency which may

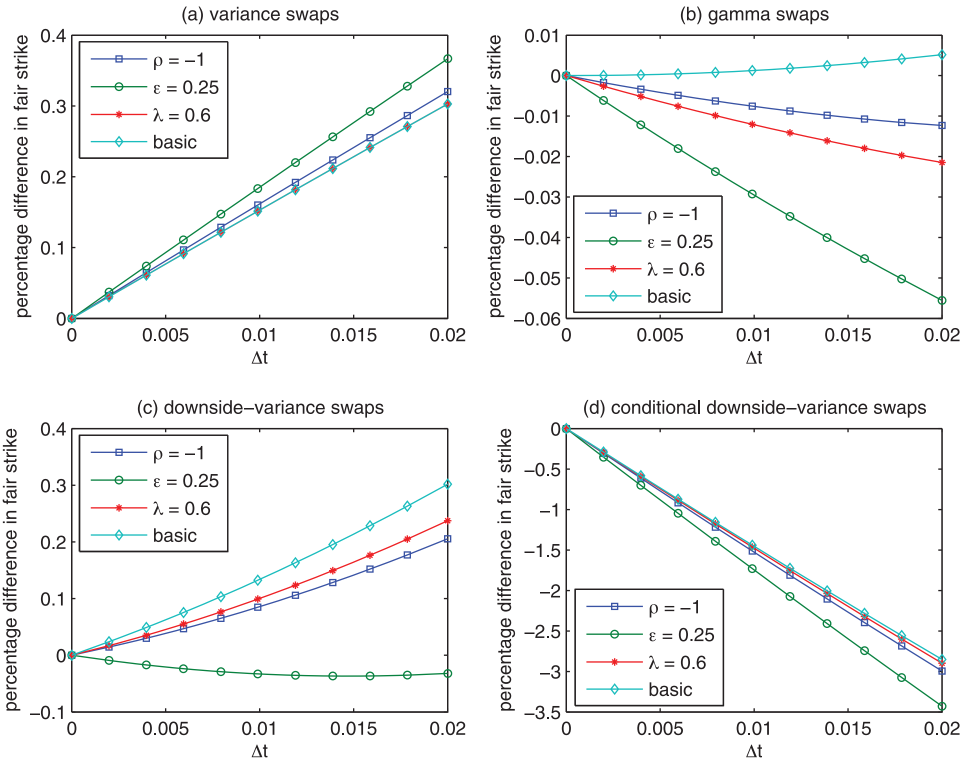

FIGURE 4.1. Plot of the percentage difference in the fair strike prices of various discretely sampled generalized variance swaps against sampling time interval Δt (in units of year): (a) variance swaps, (b) gamma swaps, (c) downside-variance swaps, and (d) conditional downside-variance swaps. The convergence of the fair strike prices to the continuous limit can be from above or below with vanishing sampling interval. For most cases, the convergence trend is "almost linear" [with a few exceptions, like the curves in (b) and (c)].

be possibly related to the constant share gamma property. An exception to insensitivity to sampling frequency is shown by the conditional downside-variance swap. This is not surprising since its fair strike price is equal to that of a downside-variance swap rescaled by $N/E[D]$, where $E[D]$ is the expected number of observations that the asset price falls within the corridor. The sensitivity to sampling frequency is then somehow enlarged by the rescaling factor.

The plots in Figure 4.2(a) and (d) show that the fair strike prices of the vanilla variance swaps and conditional downside-variance swaps exhibit an almost linear rate of convergence. However, careful observation of the plots in Figure 4.2(b) and (c) of the fair strike prices of the gamma swaps and downside-variance swaps reveal some convex curvature in the curves. This convexity behavior is not in good agreement with the results reported in the literature. The numerical tests performed by Broadie and Jain (2008) reveal a linear rate of convergence of vanilla variance swaps under various stochastic volatility models of the asset price process. Under the time-changed Lévy processes, Crosby and Davis (2011) manage to establish mathematically (under certain assumptions) the linear

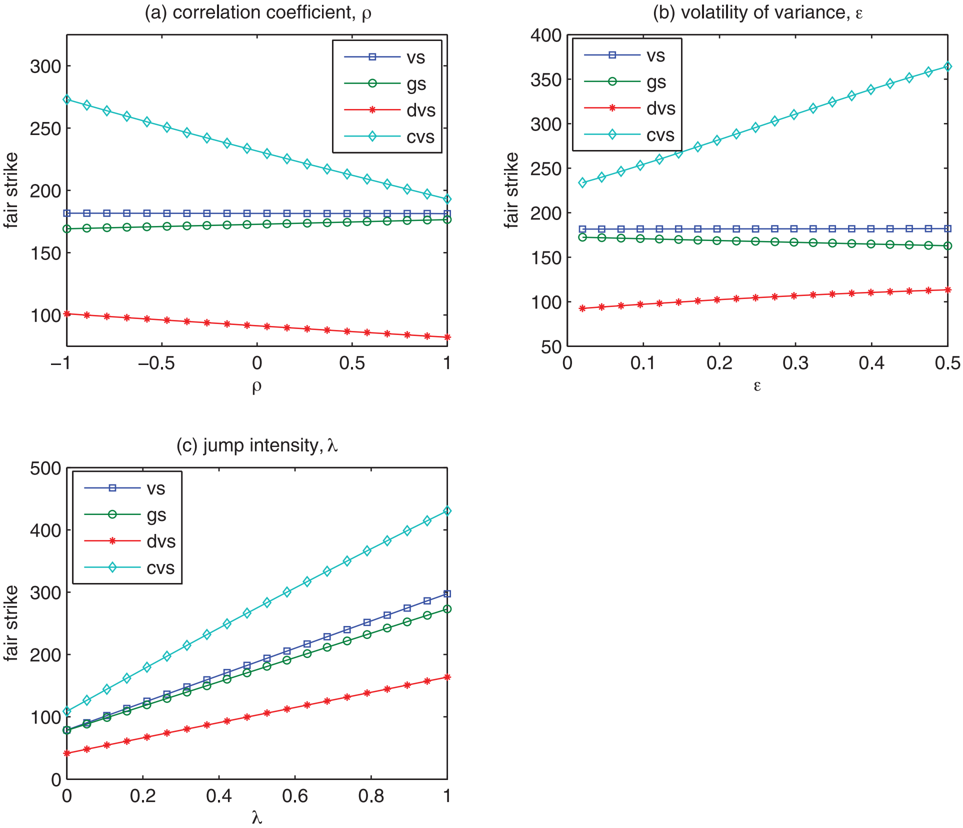

FIGURE 4.2. Plot of the fair strike of various discretely sampled generalized variance swaps against (a) correlation coefficient, ρ; (b) volatility of variance, ε; (c) jump intensity, λ. The variance swap products are variance swap (labeled as "vs"), gamma swap (labeled as "gs"), downside-variance swap (labeled as "dvs"), and conditional downside-variance swap (labeled as "cvs"). The fair strike prices of the conditional downside-variance swaps exhibit the highest sensitivity to model parameter values.

rate of convergence of various generalized variance swaps. However, due to the limitation of their approach, they have neither been able to perform a similar analysis for the generalized variance swaps with corridor restriction on the realized variance, nor make the same conclusion under the general asset price dynamics, say the SVSJ model. Our numerical results provide exceptions that the property of the linear convergence may not be always valid for discretely sampled generalized variance swaps.

4.2. Sensitivity of Fair Strike Prices to Key Model Parameters

Next, we examine the sensitivity of the fair strike price of various weekly sampled generalized variance swaps on the following key model parameters: (i) correlation coefficient ρ, (ii) volatility of variance ε, and (iii) jump intensity λ. The comparison of the fair strike prices of various generalized variance swaps with varying values of the above model parameters are shown in Figure 4.2. These plots reveal the different degrees of impact on the fair strike prices of different types of generalized variance swaps with respect to these three model parameters.

Figure 4.2(a) shows that the fair strike price of the conditional downside-variance swap has the highest sensitivity to ρ, followed by that of the downside-variance swap. Moreover, the fair strike prices of the downside-variance swap and conditional downside-variance swap tend to increase as ρ becomes more negative. To explain this phenomenon by an intuitive argument, we observe that when the leverage effect becomes stronger it is more likely that the spike of volatility is accompanied by a plunge in the asset price. The fair strike prices of the variance swap and gamma swap exhibit less sensitivity to ρ. For the variance swap, insensitivity to ρ is not surprising since its continuously sampled counterpart has no dependence on ρ. In the event of volatility running high, the gamma swap assigns lower weights to the sampled values of higher realized variance due to the decline in the underlying asset price in view of the negative correlation. This explains the slight drop in the fair strike price of the gamma swap when ρ becomes more negative.

Figure 4.2(b) shows that the fair strike price of the conditional downside-variance swap is most sensitive to ε, followed by that of the downside-variance swap. Normally, we expect that a higher ε would lead to a higher level of accumulation of the realized variance as in the case of the downside-variance swap or conditional downside-variance swap. Again, the vanilla variance swap exhibits insensitivity to ε as the fair strike price of its continuously sampled counterpart does not depend on ε. The dependence of the fair strike price on ε for the gamma swap is reversed, which may be attributed to the use of a negative value of correlation in the numerical calculations. Given a negative correlation, it is more likely for the asset price to sink when the variance process is more volatile.

Various earlier papers (e.g., Broadie and Jain 2008) report the strong dependence of the fair strike prices of variance swaps on jumps. The jumps in the underlying price and variance under the SVSJ model can be characterized by a set of jump parameters, namely, $\lambda, \nu, \eta, \rho_J$, and δ. The jump intensity λ is considered to be the most crucial parameter. In our calculations, we take ν and ρ_J to be negative. As a result, each jump would most likely lead to a decline in the underlying asset price. Actually, a larger value of λ leads to a higher chance of crash in the underlying asset price. Figure 4.2(c) shows how the fair strike prices of various swap products change with varying values of λ. All these four types of swap products are seen to show high sensitivity to λ. The conditional downside-variance swap is most sensitive to λ, followed by the variance swap, gamma swap, and then downside-variance swap. The least sensitivity to λ of the downside-variance swap may be attributed to an offsetting effect since a spike of the underlying asset price may result from a potential positive jump.

4.3. Impact of Upper Barrier of Corridor and Breaching Convention on Downside-Variance Swaps

Finally, we investigate the impact on the fair strike prices of the discretely sampled downside-variance swaps with varying values of the corridor's upper barrier and different maturities.

In Figure 4.3(a) and (b), we show the plot of the fair strike price of the weekly sampled downside-variance swap with varying values of the corridor's upper barrier U and different maturities. The plots shown in Figure 4.3(a) and (b) reveal the significant impact of the choice of upper barrier U and corridor breaching convention on the fair strike prices of the downside-variance swaps, in particular when U is chosen close to the current stock price S_0 (inferred from the steep slopes there). The difference in the fair strike prices of the two different types of downside-variance swaps, corresponding to the corridor's upper barrier U being monitored on the stock price at the old time level ("convention 1") or new time level ("convention 2"), can be quite substantial when the upper barrier is close to the current stock price S_0 (set equal to 1). Moreover, the fair strike prices of the swap contracts with the corridor monitored at the new time level are consistently larger than those of the swap contracts having the corridor monitored at the old time level. The difference in the fair strike prices vanishes when U is set extremely low or high. For downside-variance swaps with shorter maturity of half a year, the difference in the fair strike prices based on the two different breaching conventions can be more profound.

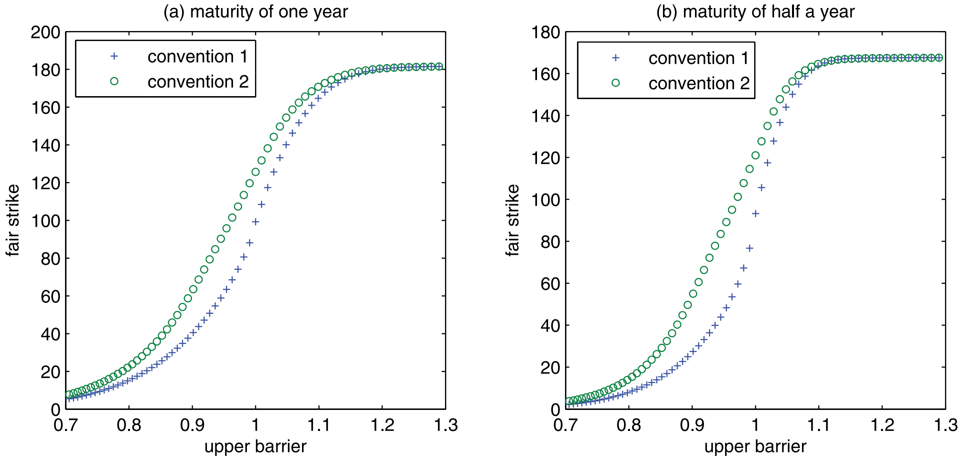

FIGURE 4.3. Comparison of the fair strike prices of the weekly sampled downside-variance swaps with varying values of the corridor's upper barrier when breaching of the corridor is monitored on the stock price at the old time level ("convention 1") or new time level ("convention 2"). The difference in the fair strike prices is shown to be more profound for downside-variance swaps with shorter maturities (comparing downside-variance swap contracts with maturities of one year and half a year).

5. CONCLUSION

In this paper, we demonstrate an analytic approach of deriving closed form pricing formulas of various types of discretely sampled generalized variance swaps under the dynamics of stochastic volatility with simultaneous jumps in the underlying asset price and its variance. The success of the analytic approach relies on the availability of the analytic expression of the joint moment generating function of the SVSJ model. We manage to derive analytic pricing formulas for the gamma swaps, corridor variance swaps, and conditional variance swaps. The last two types of generalized variance swaps with corridor constraints have their terminal payoffs dependent on the stochastic occupation time during which the underlying asset price lies within a specified corridor. The semianalytic pricing formulas for the corridor and conditional variance swaps are expressed in terms of Fourier integrals. The numerical evaluation of these Fourier integrals can be performed effectively, thanks to the fast Fourier transform algorithm. We also demonstrate through analytic asymptotic analysis that the pricing formulas of the discretely monitored generalized variance swaps converge to those of their continuously monitored counterparts. Though this paper is focused on pricing discretely monitored generalized variance swaps under the SVSJ framework, the analytic procedure can be applied to any affine model of the underlying asset price and payoff structures of higher moments swaps.

We performed numerical evaluation of these pricing formulas to examine the impact of sampling frequency on the fair strike prices of the gamma swaps, corridor variance swaps, and conditional variance swaps. We present various plots that demonstrate the convergence of the fair strike prices of the discretely monitored generalized variance swaps to those of their continuously monitored counterparts, and illustrate the sensitivity of the fair strike price to different choices of model parameter values in the SVSJ model. The fair strike prices of these generalized variance swaps are seen to be more sensitive to the contractual terms in the swap contracts and the choices of model parameter values. For example, there may exist significant difference in the fair strike prices of the downside-variance swaps with respect to whether the stock price of the old time level or new time level is used when the corridor feature is monitored. Our studies show that the impact of sampling frequency on the fair strike price can be quite insignificant for some types of generalized variance swaps, like the gamma swaps. Interestingly, the convergence of the fair strike prices with vanishing sampling interval to those of the continuously monitored counterparts can be from above or below and the linear convergence property may not be always valid for discretely sampled generalized variance swaps.

APPENDIX A: DERIVATION OF EQUATION (2.6)

A.1. Solution for $B(\tau) := B(\Theta; \tau, \mathbf{q})$

Note that $B(\tau)$ is governed by a nonlinear differential equation. To find the solution, we introduce the following transformation, where

$$B(\tau) = -\frac{2}{\varepsilon^2}\frac{E'(\tau)}{E(\tau)},$$

which leads to the following linear differential equation:

$$E''(\tau) + (\kappa - \rho\varepsilon\phi)E'(\tau) - \frac{1}{4}\varepsilon^2(\phi - \phi^2)E(\tau) = 0.$$

The initial condition $B(0) = b$ gives the derived initial condition for $E(\tau)$, where $E'(0) = -\frac{\varepsilon^2 b}{2} E(0)$. Solving the equation for $E(\tau)$, we obtain

$$E(\tau) = E(0)\left[\frac{(\xi_+ + \varepsilon^2 b)}{2\zeta} e^{-\frac{1}{2}\xi_-\tau} + \frac{(\xi_- - \varepsilon^2 b)}{2\zeta} e^{\frac{1}{2}\xi_+\tau}\right],$$

which then gives

$$B(\tau) = \frac{b(\xi_- e^{-\zeta\tau} + \xi_+) - (\phi - \phi^2)(1 - e^{-\zeta\tau})}{(\xi_+ + \varepsilon^2 b)e^{-\zeta\tau} + \xi_- - \varepsilon^2 b}.$$

A.2. Solutions for $\Gamma(\tau)$ and $\Lambda(\tau)$

Given the solution $B(\tau)$, it is relatively easy to obtain $\Gamma(\tau)$ by direct integration as shown below:

$$\begin{aligned}\Gamma(\tau) &= \gamma + (r - d)\phi\tau + \kappa\theta \int_0^\tau B(u)\,\mathrm{d}u \\ &= \gamma + (r - d)\phi\tau - \frac{2\kappa\theta}{\varepsilon^2} \ln \frac{E(\tau)}{E(0)} \\ &= (r - d)\phi\tau + \gamma - \frac{\kappa\theta}{\varepsilon^2}\left[\xi_+\tau + 2\ln \frac{(\xi_+ + \varepsilon^2 b)e^{-\zeta\tau} + \xi_- - \varepsilon^2 b}{2\zeta}\right].\end{aligned}$$

The evaluation of $\Lambda(\tau)$ requires an expectation calculation followed by integration, where

$$\Lambda(\tau) = -\lambda(m\phi + 1)\tau + \lambda \int_0^\tau E[\exp(\phi J^S + B(u)J^V)]\,\mathrm{d}u.$$

We employ the iterated expectation as follows:

$$\begin{aligned}E[\exp(\phi J^S + B(u)J^V)] &= E[E[\exp(\phi J^S + B(u)J^V)]|J^V] \\ &= E\left[\frac{e^{B(u)J^V}}{\sqrt{2\pi}\delta}\int_{-\infty}^{\infty} \exp\left(\phi x - \frac{(x - \nu - \rho_J J^V)^2}{2\delta^2}\right)\mathrm{d}x\right] \\ &= \exp\left(\phi\nu + \frac{\delta^2\phi^2}{2}\right) E[\exp([\rho_J\phi + B(u)]J^V)] \\ &= \exp\left(\phi\nu + \frac{\delta^2\phi^2}{2}\right)\int_0^\infty \frac{1}{\eta}\exp\left([\rho_J\phi + B(u)]y - \frac{y}{\eta}\right)\mathrm{d}y \\ &= \exp\left(\phi\nu + \frac{\delta^2\phi^2}{2}\right)\frac{1}{1 - [\rho_J\phi + B(u)]\eta},\end{aligned}$$

where the final integration step requires the technical condition: $\mathrm{Re}(\rho_J\phi + B(u))\eta < 1$. Since η is generally small, this requirement is usually fulfilled. Finally, we perform the integration as follows:

$$\int_0^\tau \frac{1}{1 - [\rho_J\phi + B(u)]\eta}\,\mathrm{d}u = \int_0^\tau \frac{k_1 e^{-\zeta u} + k_2}{k_3 e^{-\zeta u} + k_4}\,\mathrm{d}u = \frac{k_2}{k_4}\tau - \frac{1}{\zeta}\left(\frac{k_1}{k_3} - \frac{k_2}{k_4}\right)\ln\frac{k_3 e^{-\zeta\tau} + k_4}{k_3 + k_4},$$

where

$$\begin{aligned} k_1 &= \xi_+ + \varepsilon^2 b, \quad k_2 = \xi_- - \varepsilon^2 b, \\ k_3 &= (1 - \phi\rho_J\eta)k_1 - \eta(\phi - \phi^2 + \xi_- b), \\ k_4 &= (1 - \phi\rho_J\eta)k_2 + \eta(\phi - \phi^2 - \xi_+ b). \end{aligned}$$

In actual implementation, it may be necessary to consider the following two degenerate cases. Suppose $k_3 = 0$, then the integral becomes

$$\int_0^\tau \frac{1}{1 - [\rho_J\phi + B(u)]\eta}\,\mathrm{d}u = \frac{k_2}{k_4}\tau - \frac{k_1}{k_4}\frac{e^{-\zeta\tau} - 1}{\zeta},$$

and when $k_4 = 0$, the integral takes the form:

$$\int_0^\tau \frac{1}{1 - [\rho_J\phi + B(u)]\eta}\,\mathrm{d}u = \frac{k_1}{k_3}\tau + \frac{k_2}{k_3}\frac{e^{\zeta\tau} - 1}{\zeta}.$$

APPENDIX B: PROOF OF FORMULAS (2.8), (3.5), AND (3.10)

For notational convenience, we write $B_{\Delta t_k}$ as $B(\Theta; \Delta t_k, \mathbf{q}_1)$, and similar interpretation for other parameter functions $\Gamma(\Theta; \Delta t_k, \mathbf{q}_1)$ and $\Lambda(\Theta; \Delta t_k, \mathbf{q}_1)$. When $\mathbf{q}_1 = (\alpha\ 0\ 0)^T$, we expand $B_{\Delta t_k}$, $\Gamma_{\Delta t_k}$, and $\Lambda_{\Delta t_k}$ in powers of Δt_k, where

$$\begin{aligned} B_{\Delta t_k} &= \frac{1}{2}(\alpha^2 - \alpha)\Delta t_k + O(\Delta t_k^2), \\ \Gamma_{\Delta t_k} &= (r - d)\alpha\Delta t_k + O(\Delta t_k^2), \\ \Lambda_{\Delta t_k} &= -\lambda(m\alpha + 1)\Delta t_k + \frac{\lambda e^{\alpha\nu + \delta^2\alpha^2/2}}{1 - \alpha\rho_J\eta}\Delta t_k + O(\Delta t_k^2). \end{aligned}$$

Also, we write $B_{t_{k-1}}$ as $B(\Theta; t_{k-1}, \mathbf{q}_2)$, and similar notational interpretation for $\Gamma_{t_{k-1}}$ and $\Lambda_{t_{k-1}}$. Now, we expand $B_{t_{k-1}}$, $\Gamma_{t_{k-1}}$, and $\Lambda_{t_{k-1}}$ in powers of $B_{\Delta t_k}$, $\Gamma_{\Delta t_k}$, and $\Lambda_{\Delta t_k}$, and keep only the linear terms. This gives

(B.1)
$$\begin{cases} B_{t_{k-1}} = \dfrac{(\phi^2 - \phi)(1 - e^{-\zeta t_{k-1}})}{\xi_+ e^{-\zeta t_{k-1}} + \xi_-} + \left\{ \dfrac{\xi_- e^{-\zeta t_{k-1}} + \xi_+}{\xi_+ e^{-\zeta t_{k-1}} + \xi_-} + (\phi^2 - \phi)\left[\dfrac{(1 - e^{-\zeta t_{k-1}})\varepsilon}{\xi_+ e^{-\zeta t_{k-1}} + \xi_-}\right]^2 \right\} B_{\Delta t_k} \\ \Gamma_{t_{k-1}} = (r - d)\phi t_{k-1} - \dfrac{\kappa\theta}{\varepsilon^2}\left(\xi_+ t_{k-1} + 2\ln\dfrac{\xi_+ e^{-\zeta t_{k-1}} + \xi_-}{2\zeta}\right) \\ \qquad + \dfrac{2\kappa\theta(1 - e^{-\zeta t_{k-1}})}{\xi_+ e^{-\zeta t_{k-1}} + \xi_-} B_{\Delta t_k} + \Gamma_{\Delta t_k} + \Lambda_{\Delta t_k} \\ \Lambda_{t_{k-1}} = \lambda\varepsilon^2 \exp\left(\phi\nu + \dfrac{\delta^2\phi^2}{2}\right)\left[\dfrac{t_{k-1}}{J_2} - \dfrac{2\eta}{J_1 J_2}\ln\dfrac{\xi_+ J_1 e^{-\zeta t_{k-1}} + \xi_- J_2}{2\varepsilon^2\zeta(1 - \phi\rho_J\eta)}\right] - \lambda(m\phi + 1)t_{k-1} \\ \qquad - \dfrac{2\eta\lambda\varepsilon^2}{J_1 J_2}\exp\left(\phi\nu + \dfrac{\delta^2\phi^2}{2}\right)\left[\dfrac{\varepsilon^2(J_1 e^{-\zeta t_{k-1}} - J_2)}{\xi_+ J_1 e^{-\zeta t_{k-1}} + \xi_- J_2} + \dfrac{\eta}{1 - \phi\rho_J\eta}\right] B_{\Delta t_k}, \end{cases}$$

where ζ, ξ_+, and ξ_- are defined in equation (2.6) and

$$J_1 = (1 - \phi\rho_J\eta)\varepsilon^2 - \eta\xi_-, \;\; J_2 = (1 - \phi\rho_J\eta)\varepsilon^2 + \eta\xi_+.$$

In the above expansion procedure, we manage to maintain first-order accuracy with respect to Δt_k. The second-order derivative of $\exp(B_{t_{k-1}}V_0 + \Gamma_{t_{k-1}} + \Lambda_{t_{k-1}})$ [see equation (2.7)] with respect to α can be expressed as

(B.2)
$$\frac{\partial^2}{\partial\alpha^2}e^{B_{t_{k-1}}V_0+\Gamma_{t_{k-1}}+\Lambda_{t_{k-1}}} = e^{B_{t_{k-1}}V_0+\Gamma_{t_{k-1}}+\Lambda_{t_{k-1}}}\left(V_0\frac{\partial^2 B_{t_{k-1}}}{\partial\alpha^2} + \frac{\partial^2\Gamma_{t_{k-1}}}{\partial\alpha^2} + \frac{\partial^2\Lambda_{t_{k-1}}}{\partial\alpha^2}\right) + O\big(\Delta t_k^2\big).$$

For the variance swap, we set $\phi = 0$ in $B_{t_{k-1}}$, $\Gamma_{t_{k-1}}$, and $\Lambda_{t_{k-1}}$ in equation (B.2). By substituting all the relations between $B_{\Delta t_k}$, $\Gamma_{\Delta t_k}$, $\Lambda_{\Delta t_k}$, $B_{t_{k-1}}$, $\Gamma_{t_{k-1}}$, and $\Lambda_{t_{k-1}}$, and setting $\alpha = 0$, we obtain

$$\left.\frac{\partial^2}{\partial\alpha^2}e^{B_{t_{k-1}}V_0+\Gamma_{t_{k-1}}+\Lambda_{t_{k-1}}}\right|_{\alpha=0} = \Big\{e^{-\kappa t_{k-1}}V_0 + \frac{\lambda\eta}{\kappa}(1-e^{-\kappa t_{k-1}}) + \lambda\big[\delta^2 + \rho_J^2\eta^2 + (\nu+\rho_J\eta)^2\big]$$
$$+\,\theta(1-e^{-\kappa t_{k-1}})\Big\}\,\Delta t_k + O\big(\Delta t_k^2\big).$$

The fair strike price of the variance swap with N sampling dates is then given by

$$K_V(T, N) = \frac{1}{T}\sum_{k=1}^{N}\Big\{e^{-\kappa t_{k-1}}V_0 + \frac{\lambda\eta}{\kappa}(1-e^{-\kappa t_{k-1}}) + \lambda\big[\delta^2 + \rho_J^2\eta^2 + (\nu+\rho_J\eta)^2\big]$$
$$+\,\theta(1-e^{-\kappa t_{k-1}})\Big\}\,\Delta t_k + O(\Delta t^2),$$

where $\Delta t = \max_{1\le k\le N}\Delta t_k$. The principal term in the above expression is a Riemann left sum which converges to formula (2.8) by taking the limit $\Delta t \to 0$.

For the gamma swap, we set $\phi = 1$ in equation (B.2). By repeating similar calculations as above, we can obtain the pricing formula (3.5).

For notational convenience, we express the relations in equation (B.1) in terms of the coefficient functions $B^0(\phi, t_{k-1})$, $B^1(\phi, t_{k-1})$, $\Gamma^0(\phi, t_{k-1})$, $\Gamma^1(\phi, t_{k-1})$, $\Lambda^0(\phi, t_{k-1})$, and $\Lambda^1(\phi, t_{k-1})$ as follows:

$$\text{(B.3)}\qquad \begin{cases} B_{t_{k-1}} = B^0(\phi, t_{k-1}) + B^1(\phi, t_{k-1})B_{\Delta t_k}, \\ \Gamma_{t_{k-1}} = \Gamma^0(\phi, t_{k-1}) + \Gamma^1(\phi, t_{k-1})B_{\Delta t_k} + \Gamma_{\Delta t_k} + \Lambda_{\Delta t_k}, \\ \Lambda_{t_{k-1}} = \Lambda^0(\phi, t_{k-1}) + \Lambda^1(\phi, t_{k-1})B_{\Delta t_k}. \end{cases}$$

To derive the fair strike price of the continuously sampled downside-variance swap, we set $\phi = -\mathrm{i}w$ in equation (B.2). Again, by repeating a similar asymptotic analysis as above, we obtain the fair strike price of the continuously sampled conditional downside-variance swap in formula (3.10).

APPENDIX C: PROOF OF EQUATION (3.7b)

We consider the generalized Fourier transform of the indicator function $\mathbf{1}_{\{X_{t_{k-1}}\le u\}}$ visualized as a function of u, where

$$\int_{-\infty}^{\infty}\mathbf{1}_{\{X_{t_{k-1}}\le u\}}e^{-\mathrm{i}uw}\,\mathrm{d}u = \int_{X_{t_{k-1}}}^{\infty}e^{-\mathrm{i}uw}\,\mathrm{d}u = \frac{e^{-\mathrm{i}X_{t_{k-1}}w}}{\mathrm{i}w}.$$

Here, the Fourier transform variable w is taken to be complex and we write $w = w_r + \mathrm{i}w_i$. Provided that w_i is appropriately chosen to lie within $(-\infty, 0)$, the above generalized Fourier transform exists. By taking the corresponding generalized inverse Fourier transform, we obtain

$$\mathbf{1}_{\{X_{t_{k-1}} \leq u\}} = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{iuw} \frac{e^{-\mathrm{i}X_{t_{k-1}} w}}{\mathrm{i}w} \mathrm{d}w_r.$$

This analytic representation of the indicator function expressed in terms of a generalized Fourier integral is then substituted into equation (3.6). By interchanging the order of performing integration with the two operations of differentiation and expectation, we manage to obtain

$$\begin{aligned} &E\left[\left(\ln \frac{S_{t_k}}{S_{t_{k-1}}}\right)^2 \mathbf{1}_{\{S_{t_{k-1}} \leq U\}}\right] \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\partial^2}{\partial \phi^2} E\big[e^{-\mathrm{i}w\, X_{t_{k-1}} + B(\Theta;\Delta t_k,\mathbf{q}_1)\, V_{t_{k-1}} + \Gamma(\Theta;\Delta t_k,\mathbf{q}_1) + \Lambda(\Theta;\Delta t_k,\mathbf{q}_1)}\big]\big|_{\phi=0} \frac{e^{\mathrm{i}uw}}{\mathrm{i}w} \mathrm{d}w_r \\ &= \frac{e^{w_i(X_0-u)}}{\pi} \int_0^{\infty} \mathrm{Re}\left(e^{-\mathrm{i}w_r(X_0-u)} \frac{F_k(w_r + \mathrm{i}w_i)}{\mathrm{i}w_r - w_i}\right) \mathrm{d}w_r, \quad k \geq 2, \end{aligned}$$

as shown in equation (3.7b).

REFERENCES

BOUZOUBAA, M., and A. OSSEIRAN (2010): *Exotic Options and Hybrids: A Guide to Structuring, Pricing and Trading*, Chichester, UK: Wiley, pp. 250–254.

BROADIE, M., and A. JAIN (2008): The Effect of Jumps and Discrete Sampling on Volatility and Variance Swaps, *Int. J. Theor. Appl. Finance* 11(8), 761–797.

BROCKHAUS, O., and D. LONG (2000): Volatility Swaps Made Simple, *Risk* 13(1), 92–95.

CARR, P., and R. LEE (2009): Volatility Derivatives, *Ann. Rev. Financ. Econ.* 1, 319–339.

CARR, P., and K. LEWIS (2004): Corridor Variance Swaps, *Risk* 17(2), 67–72.

CARR, P., and D. MADAN (1998): Towards a Theory of Volatility Trading, in *Volatility*, R. Jarrow, ed., London: Risk Books, pp. 417–427.

CARR, P., and D. MADAN (1999): Option Valuation Using the Fast Fourier Transform, *J. Comput. Finan.* 2(4), 61–73.

CARR, P., and L. WU (2006): A Tale of Two Indices, *J. Derivatives* 13, 13–29.

CARR, P., and L. WU (2007): Theory and Evidence on Dynamic Interactions between Sovereign Credit Default Swaps and Currency Options, *J. Banking Finance* 31, 2383–2403.

CHACKO, G., and S. DAS (2002): Pricing Interest Rate Derivatives: A General Approach, *Rev. Financ. Stud.* 15(1), 195–241.

CONT, R., and T. KOKHOLM (2008): A Consistent Pricing Model Index Options and Volatility Derivatives. Working Paper, Finance Research Group, Aarhus University.

CROSBY, J., and M. H. DAVIS (2011): Variance Derivatives: Pricing and Convergence. Working Paper, Imperial College, London.

DEMETERFI, K., E. DERMAN, M. KAMAL, and J. ZOU (1999): A Guide to Volatility and Variance Swaps, *J. Derivatives* 6(4), 9–32.

DRIMUS, G. G., and W. FARKAS (2010): Options on Discretely Sampled Variance: Discretization Effect and Greeks. Working Paper, University of Copenhagen and University of Zurich.

DUFFIE, D., J. PAN, and K. SINGLETON (2000): Transform Analysis and Option Pricing for Affine Jump-Diffusions, *Econometrica* 68, 1343–1376.

ITKIN, A., and P. CARR (2010): Pricing Swaps and Options on Quadratic Variation under Stochastic Time Change Models—Discrete Observations Case, *Rev. Derivatives Res.* 13, 141–176.

JACQUIER, A., and S. SLAOUI (2010): Variance Dispersion and Correlation Swaps. Working Paper, Imperial College London.

KANGRO, R., K. PARNA, and A. SEPP (2004): Pricing European-Style Options under Jump Diffusion Processes with Stochastic Volatility: Applications of Fourier Transform. Working Paper, Tartu University.

KELLER-RESSEL, M., and J. MUHLE-KARBE (2010): Asymptotic and Exact Pricing of Options on Variance. Working Paper, ETH, Zürich.

LEE, R. (2010): Gamma Swap and Corridor Variance Swap, *Encyclopedia Quant. Finance*, Chichester, UK: Wiley.

LIAN, G. (2010): Pricing Volatility Derivatives with Stochastic Volatility. Ph.D. thesis, University of Wollongong.

LITTLE, T., and V. PANT (2000): A Finite-Difference Method for the Valuation of Variance Swaps, *J. Comput. Finance* 5(1), 81–101.

NEUBERGER, A. (1994): The Log Contract, *J. Portfolio Manag.* 20(2), 74–80.

SEPP, A. (2007): Variance Swaps under No Conditions, *Risk Mag.* 20(1), 82–87.

SEPP, A. (2008): Pricing Options on Realized Variance in the Heston Model with Jumps in Returns and Volatility, *J. Comput. Finance* 11(4), 33–70.

SEPP, A. (2011): Pricing Options on Realized Variance in the Heston Model with Jumps in Returns and Volatility II: An Approximation of the Discrete Variance, *J. Comp. Finance.* 16(2), 3–32.

TODOROV, V., and G. TAUCHEN (2010): Volatility Jumps. Working Paper, Northwestern University.

WINDCLIFF, H., P. A. FORSYTH, and K. R. VETZAL (2006): Pricing Methods and Hedging Strategies for Volatility Derivatives, *J. Banking Finance* 30, 409–431.

ZHU, S., and G. LIAN (2011): A Closed-Form Exact Solution for Pricing Variance Swaps with Stochastic Volatility, *Math. Finance* 21(2), 233–256.