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

Mathematical Finance, Vol. 25, No. 4 (October 2015), 673–701

NO-ARBITRAGE PRICING FOR DIVIDEND-PAYING SECURITIES IN DISCRETE-TIME MARKETS WITH TRANSACTION COSTS

TOMASZ R. BIELECKI, IGOR CIALENCO, AND RODRIGO RODRIGUEZ

Illinois Institute of Technology

We prove a version of First Fundamental Theorem of Asset Pricing under transaction costs for discrete-time markets with dividend-paying securities. Specifically, we show that the no-arbitrage condition under the efficient friction assumption is equivalent to the existence of a risk-neutral measure. We derive dual representations for the superhedging ask and subhedging bid price processes of a contingent claim contract. Our results are illustrated with a vanilla credit default swap contract.

KEY WORDS: arbitrage, fundamental theorem of asset pricing, transaction costs, consistent pricing system, liquidity, dividends, credit default swaps, interest rate swaps.

1. INTRODUCTION

One of the central themes in mathematical finance is no-arbitrage pricing and its applications. At the foundation of no-arbitrage pricing is the First Fundamental Theorem of Asset Pricing (FFTAP). We prove a version of the FFTAP under transaction costs for discrete-time markets with dividend-paying securities.[1]

The FFTAP has been proved in varying levels of generality for frictionless markets. In a discrete-time setting for a finite state space, the theorem was first proved in Harrison and Pliska (1981). Almost a decade later, Dalang, Morton, and Willinger (1990) proved the FFTAP for the more technically challenging setting in which the state space is general. Their approach requires the use of advanced, measurable selection arguments, which motivated several authors to provide alternative proofs using more accessible techniques (see Schachermayer 1992; Kabanov and Kramkov 1994; Rogers 1994; Jacod and Shiryaev 1998; and Kabanov and Stricker 2001b). Using advanced concepts from functional and stochastic analysis, the FFTAP was first proved in a general continuous-time setup in the celebrated paper by Delbaen and Schachermayer (1994). A comprehensive review of the literature pertaining to no-arbitrage pricing theory in frictionless markets can be found in Delbaen and Schachermayer (2006).

The first rigorous study of the FFTAP for markets with transaction costs in a discrete-time setting was carried out by Kabanov and Stricker (2001a). Under the assumption that the state space is finite, it was proved that *no-arbitrage condition* (**NA**) is equivalent to the existence of a consistent pricing system (CPS, using the terminology introduced

Tomasz R. Bielecki and Igor Cialenco acknowledge support from the NSF grant DMS-0908099. The authors thank the anonymous referees and the associate editor for their helpful comments and suggestion.
Manuscript received June 2012; final revision received December 2012.

Address correspondence to Igor Cialenco, Department of Applied Mathematics, Illinois Institute of Technology, 10 West 32nd Street, Bldg E1 Room 208, Chicago, IL 60616, USA; e-mail: igor@math.iit.edu.

[1] In this study, a transaction cost is defined as the cost incurred in trading in a market in which securities' quoted prices have a bid-ask spread. We do not consider other costs such as broker's fees and taxes.

DOI: 10.1111/mafi.12038

in Schachermayer 2004). However, their results did not extend to the case of a general state space. As in the frictionless case, the transition from a finite state space to a general state space is nontrivial due to measure-theoretic and topological related difficulties. These difficulties were overcome in Kabanov, Rásonyi, and Stricker (2002), where a version of the FFTAP was proven under the *efficient friction assumption* (**EF**). It was shown that the *strict no-arbitrage condition*, a condition which is stronger than **NA**, is equivalent to the existence of a strictly CPS. Therein, it was asked whether **EF** can be discarded. Schachermayer (2004) answered this question negatively by showing that neither **NA** nor the strict no-arbitrage condition alone is sufficiently strong to yield the existence of a CPS. More importantly, Schachermayer (2004) proved a new version of the FFTAP that does not require **EF**. Specifically, he proved that the *robust no-arbitrage condition*, which is stronger than the strict no-arbitrage condition, is equivalent to the existence of a strictly consistent pricing system. Subsequent studies that treat the robust no-arbitrage condition are Bouchard (2006), De Vallière, Kabanov, and Stricker (2007), Jacka, Berkaoui, and Warren (2008). Recently, Pennanen (2011a,b,c,d) studied no-arbitrage pricing in a general context in which markets can have constraints and transaction costs may depend nonlinearly on traded amounts. Therein, the problem of superhedging a claims process (e.g., swaps) is also investigated. An excellent survey of the literature pertaining to no-arbitrage pricing in markets with transaction costs can be found in Kabanov and Safarian (2009). Let us mention that versions of the FFTAP for markets with transaction costs in a continuous-time setting have also been studied in the literature. This literature considers stronger conditions than **NA** (see, for instance, Jouini and Kallal 1995; Cherny 2007; Guasoni, Rásonyi, and Schachermayer 2010; Guasoni, Lépinette, and Rásonyi 2011; Denis and Kabanov 2012).

The fundamental difference between no-arbitrage pricing theory for dividend-paying securities and no-arbitrage pricing theory for nondividend paying securities is that transaction costs associated with trading dividend-paying securities may not be proportional to the number of units of securities purchased or sold. Transaction costs associated with dividend-paying securities may *accrue* over time by merely holding the security—for a nondividend paying security transaction costs are only charged whenever the security is bought or sold. Our consideration of transaction costs on dividends distinguishes this study from other studies.

The contribution of this paper is summarized as follows:

- We define and study the value process and the self-financing condition under transaction costs for discrete-time markets with dividend-paying securities (Section 2).
- We define and investigate **NA** and **EF** in our context (Section 3).
- We prove a key closedness property of the set of claims that can be superhedged at zero cost (Section 3.1).
- Using classic separation arguments, we prove a version of the FFTAP that is relevant to our setup. Specifically, we prove that **NA** under **EF** is satisfied if and only if there exists a risk-neutral measure for the price processes of the traded securities (Section 3.2).
- We introduce an appropriate notion of CPSs in our setup, and we study the relationship between them and **NA** under **EF** (Section 4). We demonstrate that, if there are no transaction costs on the dividends paid by securities, **NA** under **EF** is equivalent to the existence of a CPS (Section 4.1).

- We derive a dual representation for the superhedging ask and subhedging bid price processes for a contingent claim contract (Section 5).

2. THE VALUE PROCESS AND THE SELF-FINANCING CONDITION

Let T be a fixed time horizon, and let $\mathcal{T} := \{0, 1, \ldots, T\}$ and $\mathcal{T}^* := \{1, 2, \ldots, T\}$. Next, let $(\Omega, \mathcal{F}_T, \mathbb{F} = (\mathcal{F}_t)_{t\in\mathcal{T}}, \mathbb{P})$ be the underlying filtered probability space.

On this probability space, we consider a market consisting of a savings account B and of N traded securities satisfying the following properties:

- The savings account can be purchased and sold according to the price process $B := ((\prod_{s=0}^{t}(1 + r_s)))_{t=0}^{T}$, where $(r_t)_{t=0}^{T}$ is a nonnegative process specifying the risk-free rate.
- The N securities can be purchased according to the ex-dividend price process $P^{\text{ask}} := ((P_t^{\text{ask},1}, \ldots, P_t^{\text{ask},N}))_{t=0}^{T}$, and pay (cumulative) dividends specified by the process $A^{\text{ask}} := ((A_t^{\text{ask},1}, \ldots, A_t^{\text{ask},N}))_{t=1}^{T}$. The quantity ΔA_t^{ask} is the dividends per unit of securities held.
- The N securities can be sold according to the ex-dividend price process $P^{\text{bid}} := ((P_t^{\text{bid},1}, \ldots, P_t^{\text{bid},N}))_{t=0}^{T}$, and pay (cumulative) dividend specified by the process $A^{\text{bid}} := ((A_t^{\text{bid},1}, \ldots, A_t^{\text{bid},N}))_{t=1}^{T}$. The quantity ΔA_t^{bid} is the dividends per unit of securities held short.

We assume that the processes introduced above adapted. In what follows, we shall denote by Δ the backward difference operator: $\Delta X_t := X_t - X_{t-1}$, and we take the convention that $A_0^{\text{ask}} = A_0^{\text{bid}} = 0$.

REMARK 2.1. For any $t = 1, 2, \ldots, T$ and $j = 1, 2, \ldots, N$, the random variable $\Delta A_t^{\text{ask},j}$ is interpreted as amount of dividend associated with holding one unit of security j from time $t-1$ to time t, and the random variable $\Delta A_t^{\text{bid},j}$ is interpreted as amount of dividend associated with selling one unit of security j from time $t-1$ to time t.

We now illustrate the processes introduced above in the context of a vanilla Credit Default Swap (CDS) contract.

EXAMPLE 2.2. A CDS contract is a contract between two parties, a *protection buyer* and a *protection seller*, in which the protection buyer pays periodic fees to the protection seller in exchange for some payment made by the protection seller to the protection buyer if a prespecified credit event of a reference entity occurs. Let τ be the nonnegative random variable specifying the time of the credit event of the reference entity. Suppose the CDS contract admits the following specifications: initiation date $t = 0$, expiration date $t = T$, and nominal value \$1. For simplicity, we assume that the loss-given-default is a nonnegative scalar δ and is paid at default. Typically, CDS contracts are traded on over-the-counter markets in which dealers quote CDS spreads to investors. Suppose that the CDS spread quoted by the dealer to sell a CDS contract with above specifications is κ^{bid} (to be received every unit of time), and the CDS spread quoted by the dealer to buy a CDS contract with above specifications is κ^{ask} (to be paid every unit of time). We remark that the CDS spreads κ^{ask} and κ^{bid} are specified in the CDS contract, so the CDS contract to sell protection is technically a different contract than the CDS contract to buy protection.

The cumulative dividend processes A^{ask} and A^{bid} associated with buying and selling the CDS with specifications above, respectively, are defined as

$$A_t^{\text{ask}} := 1_{\{\tau \le t\}}\delta - \kappa^{\text{ask}} \sum_{u=1}^{t} 1_{\{u<\tau\}}, \quad A_t^{\text{bid}} := 1_{\{\tau \le t\}}\delta - \kappa^{\text{bid}} \sum_{u=1}^{t} 1_{\{u<\tau\}}$$

for $t \in \mathcal{T}^*$. In this case, the ex-dividend ask and bid price processes P^{bid} and P^{ask} specify the mark-to-market values of the CDS for the protection seller and protection buyer, respectively, from the perspective of the protection buyer. The CDS spreads κ^{ask} and κ^{bid} are set so that $P_0^{\text{bid}} = P_0^{\text{ask}} = 0$. Also, we have that $P_T^{\text{ask}} = P_T^{\text{bid}} = 0$ since they are ex-dividend prices.

Next, we illustrate the processes above with a vanilla Interest Rate Swap (IRS) contract.

EXAMPLE 2.3. An IRS contract is a contract between two parties, in which one party agrees to periodically pay a fixed rate (the swap rate) to the other party, in exchange for a floating rate (usually the Libor rate). We suppose that the floating rate from $i-1$ to i, denoted by L_i, is exchanged for the swap rate every unit of time. Also, we assume that the IRS admits the following specifications: initiation date $t = 0$, expiration date $t = T$, and nominal value \$1. IRS contracts are traded on over-the-counter markets in which dealers quote swap rates to investors. For the contract specified above, we denote by s^{ask} the swap rate quoted by the dealer for a Payer IRS (pays the swap rate and receives the floating rate), and denote by s^{bid} the swap rate quoted by the dealer for a Receiver IRS (pays the floating rate and receives the swap rate). We remark that the spreads s^{ask} and s^{bid} are specified in the IRS contract.

The cumulative dividend processes A^{bid} and A^{ask} associated with the Payer and Receiver swap with specifications above, respectively, are defined as

$$A_t^{\text{ask}} := \sum_{i=1}^{t}(L_i - s^{\text{ask}}), \quad A_t^{\text{bid}} := \sum_{i=1}^{t}(s^{\text{bid}} - L_i)$$

for $t \in \mathcal{T}^*$. The ex-dividend ask and bid price processes P^{bid} and P^{ask} specify the mark-to-market values of the IRS for the Payer IRS and Receiver IRS, respectively. The values of swap spreads s^{ask} and s^{bid} are set so that $P_0^{\text{bid}} = P_0^{\text{ask}} = 0$ are null at initiation date, and also note that $P_T^{\text{ask}} = P_T^{\text{bid}} = 0$ since they are ex-dividend prices.

From now on, we make the following standing assumption.

BID–ASK ASSUMPTION. $P^{\text{ask}} \ge P^{\text{bid}}$ and $\Delta A^{\text{ask}} \le \Delta A^{\text{bid}}$.

For convenience we define $\mathcal{J} := \{0, 1, \ldots, N\}$ and $\mathcal{J}^* := \{1, 2, \ldots, N\}$. Unless stated otherwise, all inequalities and equalities between processes and random variables are understood $\mathbb{P}$-a.s. and coordinate-wise.

2.1. The Value Process and Trading Strategies

A *trading strategy* is a predictable process $\phi := (\phi_t^0, \phi_t^1, \ldots, \phi_t^N)_{t=1}^T$, where ϕ_t^j is interpreted as the number of units of security j held from time $t-1$ to time t. Processes $\phi^1, \ldots, \phi^N$ correspond to the holdings in the N securities, and process ϕ^0 corresponds to the holdings in the savings account B. We take the convention $\phi_0 = (0, \ldots, 0)$.

DEFINITION 2.4. The *value process* $(V_t(\phi))_{t=0}^T$ associated with a trading strategy ϕ is defined as

$$V_t(\phi) = \begin{cases} \phi_1^0 + \sum_{j=1}^N \phi_1^j \big(1_{\{\phi_1^j \geq 0\}} P_0^{\text{ask},j} + 1_{\{\phi_1^j < 0\}} P_0^{\text{bid},j}\big), & \text{if } t = 0, \\ \phi_t^0 B_t + \sum_{j=1}^N \phi_t^j \big(1_{\{\phi_t^j \geq 0\}} P_t^{\text{bid},j} + 1_{\{\phi_t^j < 0\}} P_t^{\text{ask},j}\big) \\ \quad + \sum_{j=1}^N \phi_t^j \big(1_{\{\phi_t^j < 0\}} \Delta A_t^{\text{bid},j} + 1_{\{\phi_t^j \geq 0\}} \Delta A_t^{\text{ask},j}\big), & \text{if } 1 \leq t \leq T. \end{cases}$$

For $t = 0$, $V_0(\phi)$ is interpreted as the cost of the portfolio ϕ, and for $t \in \{1, \ldots, T\}$, $V_t(\phi)$ is interpreted as the liquidation value of the portfolio ϕ before any time t transactions, including any dividends acquired from time $t-1$ to time t.

REMARK 2.5. Due to the presence of transaction costs, the value process V may not be linear in its argument, i.e., $V_t(\phi) + V_t(\psi) \neq V_t(\phi + \psi)$, and $V_t(\alpha\phi) \neq \alpha V_t(\phi)$ for $\alpha \in \mathbb{R}$, and some trading strategies ϕ, ψ, some time $t \in \mathcal{T}$. This is the major difference from the frictionless setting.

Next, we introduce the self-financing condition, which is appropriate in the context of this paper.

DEFINITION 2.6. A trading strategy ϕ is self-financing if

$$\begin{aligned} B_t \Delta\phi_{t+1}^0 + \sum_{j=1}^N \Delta\phi_{t+1}^j \big(1_{\{\Delta\phi_{t+1}^j \geq 0\}} P_t^{\text{ask},j} + 1_{\{\Delta\phi_{t+1}^j < 0\}} P_t^{\text{bid},j}\big) \\ = \sum_{j=1}^N \phi_t^j \big(1_{\{\phi_t^j \geq 0\}} \Delta A_t^{\text{ask},j} + 1_{\{\phi_t^j < 0\}} \Delta A_t^{\text{bid},j}\big) \end{aligned} \tag{2.1}$$

for $t = 1, 2, \ldots, T-1$.

The self-financing condition imposes the restriction that no money can flow in or out of the portfolio. We note that if $P := P^{\text{ask}} = P^{\text{bid}}$ and $\Delta A := \Delta A^{\text{ask}} = \Delta A^{\text{bid}}$, then the self-financing condition in the frictionless case is recovered.

REMARK 2.7. Note that the self-financing condition not only takes into account transaction costs due to purchases and sales of securities (left-hand side of (2.1)), but also transaction costs accrued through the dividends (right-hand side of (2.1)).

The following two propositions give useful characterizations of the self-financing condition in terms of the value process. The proofs are straightforward, and are left for the reader.

PROPOSITION 2.8. *A trading strategy ϕ is self-financing if and only if the value process $V(\phi)$ satisfies*

$$
\begin{aligned}
V_t(\phi) = V_0(\phi) &+ \sum_{u=1}^{t} \phi_u^0 \Delta B_u + \sum_{j=1}^{N} \phi_t^j \big(1_{\{\phi_t^j \geq 0\}} P_t^{\mathrm{bid},j} + 1_{\{\phi_t^j < 0\}} P_t^{\mathrm{ask},j}\big) \\
&- \sum_{j=1}^{N} \sum_{u=1}^{t} \Delta \phi_u^j \big(1_{\{\Delta\phi_u^j \geq 0\}} P_{u-1}^{\mathrm{ask},j} + 1_{\{\Delta\phi_u^j < 0\}} P_{u-1}^{\mathrm{bid},j}\big) \\
&+ \sum_{j=1}^{N} \sum_{u=1}^{t} \phi_u^j \big(1_{\{\phi_u^j \geq 0\}} \Delta A_u^{\mathrm{ask},j} + 1_{\{\phi_u^j < 0\}} \Delta A_u^{\mathrm{bid},j}\big)
\end{aligned} \tag{2.2}
$$

for all $t \in \mathcal{T}^*$.

The next proposition extends the previous result in terms of our *numéraire* B. For convenience, we let $V^*(\phi) := B^{-1} V(\phi)$ for all trading strategies ϕ.

PROPOSITION 2.9. *A trading strategy* ϕ *is self-financing if and only if the discounted value process* $V^*(\phi)$ *satisfies*

$$
\begin{aligned}
V_t^*(\phi) = V_0(\phi) &+ \sum_{j=1}^{N} \phi_t^j B_t^{-1} \big(1_{\{\phi_t^j \geq 0\}} P_t^{\mathrm{bid},j} + 1_{\{\phi_t^j < 0\}} P_t^{\mathrm{ask},j}\big) \\
&- \sum_{j=1}^{N} \sum_{u=1}^{t} \Delta \phi_u^j B_{u-1}^{-1} \big(1_{\{\Delta\phi_u^j \geq 0\}} P_{u-1}^{\mathrm{ask},j} + 1_{\{\Delta\phi_u^j < 0\}} P_{u-1}^{\mathrm{bid},j}\big) \\
&+ \sum_{j=1}^{N} \sum_{u=1}^{t} \phi_u^j B_u^{-1} \big(1_{\{\phi_u^j \geq 0\}} \Delta A_u^{\mathrm{ask},j} + 1_{\{\phi_u^j < 0\}} \Delta A_u^{\mathrm{bid},j}\big)
\end{aligned} \tag{2.3}
$$

for all $t \in \mathcal{T}^*$.

REMARK 2.10. If $P = P^{\mathrm{ask}} = P^{\mathrm{bid}}$ and $\Delta A = \Delta A^{\mathrm{ask}} = \Delta A^{\mathrm{bid}}$, then we recover the classic result: a trading strategy ϕ is self-financing if and only if the value process satisfies

$$
V_t^*(\phi) = V_0(\phi) + \sum_{j=1}^{N} \sum_{u=1}^{t} \phi_u^j \Delta \left(B_u^{-1} P_u^j + \sum_{w=1}^{u} B_w^{-1} \Delta A_w^j \right)
$$

for all $t \in \mathcal{T}^*$.

For convenience, we define $P^{\mathrm{ask},*} := B^{-1} P^{\mathrm{ask}}$ and $P^{\mathrm{bid},*} := B^{-1} P^{\mathrm{bid}}$, and with a slight abuse of notation $A^{\mathrm{ask},*} := B^{-1} \Delta A^{\mathrm{ask}}$ and $A^{\mathrm{bid},*} := B^{-1} \Delta A^{\mathrm{bid}}$.

In frictionless markets, the set of all self-financing trading strategies is a linear space because securities' prices are not influenced by the direction of trading. This is no longer the case if the direction of trading matters: the strategy $\phi + \psi$ may not be self-financing even if ϕ and ψ are self-financing. Intuitively this is true because transaction costs can be avoided whenever $\phi_t^j \psi_t^j < 0$ by combining orders. However, the strategy $(\theta^0, \phi^1 + \psi^1, \phi^2 + \psi^2, \ldots, \phi^N + \psi^N)$ can enjoy the self-financing property if the units in the savings account θ^0 are properly adjusted. The next proposition states that such θ^0 exists, is unique, and satisfies $\phi^0 + \psi^0 \leq \theta^0$. The proof is straightforward, and is left to the reader.

PROPOSITION 2.11. *Let* ψ *and* ϕ *be any two self-financing trading strategies with* $V_0(\psi) = V_0(\phi) = 0$. *Then there exists a unique predictable process* θ^0 *such that the trading strategy*

θ defined as $\theta := (\theta^0, \phi^1 + \psi^1, \ldots, \phi^N + \psi^N)$ is self-financing with $V_0(\theta) = 0$. Moreover, $\phi^0 + \psi^0 \leq \theta^0$.

The next result is the natural extension of Proposition 2.11 to value processes. It is intuitively true since some transaction costs may be avoided by combining orders. The proof is straightforward and therefore it is omitted.

THEOREM 2.12. *Let ϕ and ψ be any two self-financing trading strategies such that $V_0(\phi) = V_0(\psi) = 0$. There exists a unique predictable process θ^0 such that the trading strategy defined as $\theta := (\theta^0, \phi^1 + \psi^1, \ldots, \phi^N + \psi^N)$ is self-financing with $V_0(\theta) = 0$, and $V_T(\theta)$ satisfies*

$$V_T(\phi) + V_T(\psi) \leq V_T(\theta).$$

The following technical lemma, which is easy to verify, will be used in the next section.

LEMMA 2.13. *The following hold:*

- *Let Y^a and Y^b be any random variables, and suppose X^m is a sequence of $\mathbb{R}$-valued random variables converging a.s. to X. Then, $1_{\{X^m \geq 0\}} X^m Y^b + 1_{\{X^m < 0\}} X^m Y^a$ converges a.s. to $1_{\{X \geq 0\}} X Y^b + 1_{\{X < 0\}} X Y^a$.*
- *If a sequence of trading strategies ϕ^m converges a.s. to ϕ, then $V(\phi^m)$ converges a.s. to $V(\phi)$.*

2.2. The Set of Claims That Can Be Superhedged at Zero Cost

For all $t \in \mathcal{T}$, denote by $L^0(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R}^{(N+1)})$ the space of all ($\mathbb{P}$-equivalence classes of) $\mathbb{R}^{(N+1)}$-valued, $\mathcal{F}_t$-measurable random variables. We equip $L^0(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$ with the topology of convergence in measure $\mathbb{P}$. Also, let $\mathcal{S}$ be the set of all self-financing trading strategies. For the sake of conciseness, we will refer to sets that are closed with respect to convergence in measure $\mathbb{P}$ simply as $\mathbb{P}$-closed.

We define the sets

$$\mathcal{K} := \left\{ V_T^*(\phi) : \phi \in \mathcal{S},\ V_0(\phi) = 0 \right\},$$

$$L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R}) := \{X \in L^0(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R}) : X \geq 0\},$$

$$\mathcal{K} - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R}) := \{Y - X : Y \in \mathcal{K} \text{ and } X \in L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})\}.$$

The set $\mathcal{K}$ is the *set of attainable claims at zero cost*. On the other hand, $\mathcal{K} - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$ is *the set of claims that can be superhedged at zero cost*: for any $X \in \mathcal{K} - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$, there exists an attainable claim at zero cost $K \in \mathcal{K}$ so that $X \leq K$.

The following lemma asserts that the set of claims that can be superhedged at zero cost is a convex cone. The proof follows from Theorem 2.12.

LEMMA 2.14. *The set $\mathcal{K} - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$ is a convex cone.*

REMARK 2.15. The set $\mathcal{K}$ is not necessarily a convex cone. To see this, let us suppose that $T = 1$, $\mathcal{J} = \{0, 1\}$, and $r = 0$. Consider the trading strategies $\phi = \{\phi^0, 1\}$ and $\psi = \{\psi^0, -1\}$, where ϕ^0 and ψ^0 are chosen so that $V_0(\phi) = V_0(\psi) = 0$. By definition, $V_1(\phi), V_1(\psi) \in \mathcal{K}$. However,

$$V_1^*(\phi) + V_1^*(\psi) = P_1^{\text{bid}} - P_1^{\text{ask}} + A_1^{\text{ask}} - A_1^{\text{bid}} + P_0^{\text{bid}} - P_0^{\text{ask}}$$

is generally not in the set $\mathcal{K}$.

3. THE NO-ARBITRAGE CONDITION

We begin by introducing the definition of the no-arbitrage condition.

DEFINITION 3.1. *The no-arbitrage condition* **(NA)** is satisfied if for each $\phi \in \mathcal{S}$ such that $V_0(\phi) = 0$ and $V_T(\phi) \geq 0$, we have $V_T(\phi) = 0$.

In the present context, **NA** has the usual interpretation that "it is impossible to make something out of nothing." The next lemma provides us equivalent conditions to **NA** in terms of the set of attainable claims at zero cost, and also in terms of the set of claims that can be superhedged at zero cost. They are straightforward to verify.

LEMMA 3.2. *The following conditions are equivalent:*

(i) ***NA*** *is satisfied.*
(ii) $\big(\mathcal{K} - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})\big) \cap L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R}) = \{0\}$.
(iii) $\mathcal{K} \cap L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R}) = \{0\}$.

We proceed by defining The Efficient Friction Assumption in our context.

THE EFFICIENT FRICTION ASSUMPTION (EF):

$$\{\phi \in \mathcal{S} : \ V_0(\phi) = V_T(\phi) = 0\} = \{0\}. \tag{3.1}$$

Note that if (3.1) is satisfied, then for each $\phi \in \mathcal{S}$, we have $V_0(\phi) = V_T(\phi) = 0$ if and only if $\phi = 0$. The efficient friction assumption, which was introduced by Kabanov et al. (2002), states that the only portfolio that can be liquidated into the zero portfolio that is available at zero price is the zero portfolio. In the present context, **EF** has the same interpretation: the only zero-cost, self-financing strategy that can be liquidated into the zero portfolio is the zero portfolio.

We will denote by **NAEF** *the no-arbitrage condition under the efficient friction assumption*.

In what follows, we denote by $\mathcal{P}$ the set of all $\mathbb{R}^N$-valued, $\mathbb{F}$-predictable processes. Also, we define the mapping

$$\begin{aligned} F(\phi) := {} & \sum_{j=1}^{N} \phi_T^j \big(1_{\{\phi_T^j \geq 0\}} P_T^{\text{bid},j,*} + 1_{\{\phi_T^j < 0\}} P_T^{\text{ask},j,*}\big) \\ & - \sum_{j=1}^{N} \sum_{u=1}^{T} \Delta\phi_u^j \big(1_{\{\Delta\phi_u^j \geq 0\}} P_{u-1}^{\text{ask},j,*} + 1_{\{\Delta\phi_u^j < 0\}} P_{u-1}^{\text{bid},j,*}\big) \\ & + \sum_{j=1}^{N} \sum_{u=1}^{T} \phi_u^j \big(1_{\{\phi_u^j \geq 0\}} A_u^{\text{ask},j,*} + 1_{\{\phi_u^j < 0\}} A_u^{\text{bid},j,*}\big) \end{aligned} \tag{3.2}$$

for all $\mathbb{R}^N$-valued stochastic processes

$$(\phi_s)_{s=1}^T \in L^0(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R}^N) \times \cdots \times L^0(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R}^N),$$

and let $\mathbb{K} := \{F(\phi) : \phi \in \mathcal{P}\}$. In view of Proposition 2.9, we note that $V_T^*(\phi) = V_0(\phi) + F(\phi)$ for all self-financing trading strategies ϕ, and that $\mathcal{K} = \mathbb{K}$.

REMARK 3.3.

(i) Note that F is defined on the set of all $\mathbb{R}^N$-valued stochastic processes. On the contrary, the value process is defined on the set of trading strategies, which are $\mathbb{R}^{N+1}$-valued predictable processes.
(ii) The set $\mathbb{K}$ has the same financial interpretation as the set $\mathcal{K}$. We introduce the set $\mathbb{K}$ because it is more convenient to work with from the mathematical point of view.
(iii) $F(\alpha\phi) = \alpha F(\phi)$ for any nonnegative random variable α.

The next result, which is easy to prove, provides an equivalent condition for **EF** to hold.

LEMMA 3.4. *The efficient friction assumption **(EF)** is satisfied if and only if* $\{\psi \in \mathcal{P} : F(\psi) = 0\} = \{0\}$.

3.1. Closedness Property of the Set of Claims That Can Be Superhedged at Zero Cost

In this section, we prove that the set of claims that can be superhedged at zero cost, $\mathcal{K} - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$, is $\mathbb{P}$-closed whenever **NAEF** is satisfied. This property plays a central role in the proof of the FFTAP (Theorem 3.13).

We will denote by $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^N$.

Let us first recall lemma A.2 from Schachermayer (2004), which is closely related to lemma 1 in Kabanov et al. (2002).

LEMMA 3.5. *For a sequence of random variables $X^m \in L^0(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^N)$ there is a strictly increasing sequence of positive, integer-valued, $\mathcal{F}$-measurable random variables τ^m such that X^{τ^m} converges a.s. in the one-point-compactification $\mathbb{R}^N \cup \{\infty\}$ to some random variable $X \in L^0(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^N \cup \{\infty\})$. Moreover, we may find the subsequence such that $\|X\| = \limsup_m \|X^m\|$, where $\|\infty\| = \infty$.*

The next result extends the previous lemma to processes. For the proof, see lemma 5.2 in Pennanen (2011c).

LEMMA 3.6. *Let $\mathcal{F}^i$ be a σ-algebra, and $Y_i^m \in L^0(\Omega, \mathcal{F}^i, \mathbb{P}; \mathbb{R}^N)$ for $i = 1, \ldots, M$. Suppose that $\mathcal{F}^i \subseteq \mathcal{F}^j$ for all $i \le j$, and that Y_i^m satisfies $\limsup_m \|Y_i^m\| < \infty$ for $i = 1, \ldots, M$. Then there is a strictly increasing sequence of positive, integer-valued, $\mathcal{F}^M$-measurable random variables τ^m such that, for $i = 1, \ldots, M$, the sequence $Y_i^{\tau^m}$ converges a.s. to some $Y_i \in L^0(\Omega, \mathcal{F}^i, \mathbb{P}; \mathbb{R}^N)$.*

We proceed with a technical lemma.

LEMMA 3.7. *Let $\mathcal{F}^i$ be a σ-algebra, and $Y_i^m \in L^0(\Omega, \mathcal{F}^i, \mathbb{P}; \mathbb{R}^N)$ for $i = 1, \ldots, M$. Suppose that $\mathcal{F}^i \subseteq \mathcal{F}^j$ for all $i \le j$, and that there exists $k \in \{1, \ldots, M\}$ and $\Omega' \subseteq \Omega$ with $\mathbb{P}(\Omega') > 0$ such that $\limsup_m \|Y_k^m(\omega)\| = \infty$ for a.e. $\omega \in \Omega'$, and $\limsup_m \|Y_i^m(\omega)\| < \infty$ for $i = 1, \ldots, k-1$ and for a.e. $\omega \in \Omega$. Then there exists a strictly increasing sequence of positive, integer-valued, $\mathcal{F}^k$-measurable random variables τ^m such that $\lim_m \|Y_k^{\tau^m}(\omega)\| = \infty$,*

for a.e. $\omega \in \Omega'$, *and*[2]

$$X_i^m(\omega) := 1_{\Omega'}(\omega)\frac{Y_i^{\tau^m(\omega)}(\omega)}{\left\| Y_k^{\tau^m(\omega)}(\omega)\right\|}, \quad \omega \in \Omega,\ i = 1, \dots, M,$$

satisfies $\lim_m X_i^m(\omega) = 0$, *for* $i = 1, \dots, k-1$ *and for a.e.* $\omega \in \Omega$.

Proof. Since $\limsup_m \| Y_k^m(\omega)\| = \infty$ for a.e. $\omega \in \Omega'$, we may apply Lemma 3.5 to the sequence Y_k^m to find a strictly increasing sequence of positive, integer-valued, $\mathcal{F}^k$-measurable random variables τ^m so that $\| Y_k^{\tau^m(\omega)}(\omega)\|$ diverges for a.e. $\omega \in \Omega'$.

Because $\limsup_m \| Y_i^m\| < \infty$ for $i = 1, \dots, k-1$, we have $\limsup_m \| Y_i^{\tau^m}\| < \infty$ for $i = 1, \dots, k-1$. Now since $\| Y_k^{\tau^m(\omega)}(\omega)\|$ diverges for a.e. $\omega \in \Omega'$,

$$\lim_{m\to\infty} \left\| X_i^m(\omega)\right\| = 1_{\Omega'}(\omega)\frac{\left\| Y_i^{\tau^m(\omega)}(\omega)\right\|}{\left\| Y_k^{\tau^m(\omega)}(\omega)\right\|} = 0, \quad \text{a.e. } \omega \in \Omega,\ i = 1, \dots, k-1.$$

Thus, $\| X_i^m\|$ converges a.s. to 0 for $i = 1, \dots, k-1$, which implies that X_i^m converges a.s. to 0 for $i = 1, \dots, k-1$. Hence, the claim holds. □

We are now ready to prove the crucial result in this paper.

THEOREM 3.8. *If the no-arbitrage condition under the efficient friction assumption* ***(NAEF)*** *is satisfied, then the set* $\mathcal{K} - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$ *is* $\mathbb{P}$*-closed.*

Proof. Since $\mathcal{K} = \mathbb{K}$, we may equivalently prove that $\mathbb{K} - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$ is $\mathbb{P}$-closed. Suppose that $X^m \in \mathbb{K} - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$ converges in probability to X. Then there exists a subsequence X^{k_m} of X^k so that X^{k_m} converges a.s. to X. With a slight abuse of notation, we will denote by X^m the sequence X^{k_m} in what follows. By the definition of $\mathbb{K} - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$, there exists $Z^m \in L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$ and $\phi^m \in \mathcal{P}$ so that

$$X^m = F(\phi^m) - Z^m. \tag{3.3}$$

We proceed the proof in two steps. In the first step, we show by contradiction that $\limsup_m \|\phi_s^m\| < \infty$ for all $s \in \mathcal{T}^*$.

Step 1a: Let us assume that $\limsup_m \|\phi_s^m\| < \infty$ for all $s \in \mathcal{T}^*$ does not hold. Then

$$\mathcal{I}^0 := \left\{ s \in \mathcal{T}^* : \exists\ \Omega' \subseteq \Omega \text{ such that } \mathbb{P}(\Omega') > 0,\ \limsup_{m\to\infty} \left\|\phi_s^m(\omega)\right\| = \infty \text{ for a.e. } \omega \in \Omega' \right\}$$

is nonempty. Let $t_0 := \min \mathcal{I}^0$, and define the $\mathcal{F}_{t_0-1}$-measurable set

$$E^0 := \left\{ \omega \in \Omega : \limsup_{m\to\infty} \left\|\phi_{t_0}^m(\omega)\right\| = \infty \right\}.$$

Note that $\mathbb{P}(E^0) > 0$ by assumption. We now apply Lemma 3.7 to ϕ^m: there exists a strictly increasing sequence of positive, integer-valued, $\mathcal{F}_{t_0-1}$-measurable random variables τ_0^m such that

$$\lim_{m\to\infty} \left\|\phi_{t_0}^{\tau_0^m(\omega)}(\omega)\right\| = \infty, \quad \text{a.e. } \omega \in E^0, \tag{3.4}$$

[2]We take $X_i^m(\omega) = 0$ whenever $\| Y_k^{\tau^m(\omega)}(\omega)\| = 0$. We will take the convention $x/0 = 0$ throughout this section.

and

$$\psi_s^{m,(0)} := 1_{E^0} \frac{\phi_s^{\tau_0^m}}{\left\|\phi_{t_0}^{\tau_0^m}\right\|}, \quad s \in \mathcal{T}^* \tag{3.5}$$

satisfies $\lim_m \psi_s^{m,(0)}(\omega) = 0$, for $s = 1, \ldots, t_0 - 1$, for a.e. $\omega \in \Omega$.

We proceed as follows.

Recursively for $i = 1, \ldots, T$

If $\limsup_m \|\psi_s^{m,(i-1)}\| < \infty$ for all $s \in \{t_{i-1} + 1, \ldots, T\}$, then define $k := i$ and $\varphi^m := \psi^{m,(k-1)}$, and proceed to Step 1b.

Else, define

$$t_i := \min \Big\{ s \in \{t_{i-1} + 1, \ldots, T\} : \exists\, \Omega' \subseteq E^{i-1} \text{ s.t. } \mathbb{P}(\Omega') > 0, \\ \limsup_{m \to \infty} \left\| \psi_s^{m,(i-1)}(\omega) \right\| = \infty \text{ for a.e. } \omega \in \Omega' \Big\},$$

and

$$E^i := \left\{ \omega \in E^{i-1} : \limsup_{m \to \infty} \left\| \psi_{t_i}^{m,(i-1)}(\omega) \right\| = \infty \right\}.$$

Next, apply Lemma 3.7 to $\psi^{m,(i)}$: there exists a strictly increasing sequence of positive, integer-valued, $\mathcal{F}_{t_i - 1}$-measurable random variables τ_i^m such that

$$\left\{ \tau_i^1(\omega), \tau_i^2(\omega), \ldots \right\} \subseteq \cdots \subseteq \left\{ \tau_0^1(\omega), \tau_0^2(\omega), \ldots \right\}, \quad \text{a.e. } \omega \in \Omega, \tag{3.6}$$

the sequence $\psi_{t_i}^{\tau_i^m,(i-1)}$ satisfies

$$\lim_{m \to \infty} \left\| \psi_{t_i}^{\tau_i^m(\omega),(i-1)}(\omega) \right\| = \infty, \quad \text{a.e. } \omega \in E^i, \tag{3.7}$$

and the sequence $\psi^{m,(i)}$ defined as

$$\psi_s^{m,(i)} := 1_{E^i} \frac{\psi_s^{\tau_i^m,(i-1)}}{\left\|\psi_{t_i}^{\tau_i^m,(i-1)}\right\|}, \quad s \in \mathcal{T}^* \tag{3.8}$$

satisfies $\lim_m \psi_s^{m,(i)}(\omega) = 0$ for $s = 1, \ldots, t_i - 1$, for a.e. $\omega \in \Omega$.

Repeat: $i \to i + 1$.

Given this construction, we define

$$\beta_i^m(\omega) := \tau_i \circ \tau_{i+1} \circ \cdots \circ \tau_k^m(\omega), \qquad i \in \{0, \ldots, k\},\ \omega \in \Omega,$$

$$U^m(\omega) := \left\| \phi_{t_0}^{\beta_0^m(\omega)}(\omega) \right\| \prod_{i=1}^{k} \left\| \psi_{t_i}^{\beta_i^m(\omega),(i-1)}(\omega) \right\|, \quad \omega \in \Omega.$$

We make the following observations on this construction:

(i) The construction always produces a sequence φ^m such that $\limsup_m \|\varphi_s^m\| < \infty$ for all $s \in \mathcal{T}^*$. Indeed, if $t_i = T$ for some $i = 1, \ldots, T$, then $\lim_m \psi_s^{m,(i)}(\omega) = 0$ for

$s = 1, \dots, T-1$, for a.e. $\omega \in \Omega$, and $\lim_m \|\psi_T^{m,(i)}(\omega)\| = 1_{E^i}(\omega)$, for a.e. $\omega \in \Omega$. The sequence $\psi^{m,(i)}$ clearly satisfies $\limsup_m \|\psi_s^{m,(i)}\| < \infty$ for all $s \in \mathcal{T}^*$.

(ii) We have that $\varphi_s^m \in L^0(\Omega, \mathcal{F}_{t_k-1}, \mathbb{P}, \mathbb{R}^N)$ for $s = 1, \dots, t_k - 1$, and $\varphi_s^m \in L^0(\Omega, \mathcal{F}_{s-1}, \mathbb{P}, \mathbb{R}^N)$ for $s = t_k, \dots, T$. Hence, the sequence φ^m *is not* a sequence of predictable processes. However, the limit of any a.s. convergent subsequence of φ^m *is* predictable because φ_s^m converges a.s. to 0 for $s = 1, \dots, t_k - 1$.

(iii) $E^k \subseteq \cdots \subseteq E^0$, and $\mathbb{P}(E^k) > 0$.

(iv) Any a.s. convergent subsequence of φ^m converges a.s. to a nonzero process since $\|\varphi_{t_k}^m\|$ converges a.s. to 1_{E^k}, which is nonzero a.s. since $\mathbb{P}(E^k) > 0$.

(v) From (3.5) and (3.8), we have $\varphi_s^m = 1_E \phi_s^{\beta_0^m} / U^m$ for all $s \in \mathcal{T}^*$, where $E := \bigcap_{i=1}^k E^i$. Because $E^k \subseteq \cdots \subseteq E^0$,

$$\varphi_s^m = 1_{E^k} \frac{\phi_s^{\beta_0^m}}{U^m}, \quad s \in \mathcal{T}^*. \tag{3.9}$$

(vi) $U^m(\omega)$ diverges for a.e. $\omega \in E^k$ since (3.4), (3.6), and (3.7) hold.

Step 1b: By the previous step, $\limsup_m \|\varphi_s^m\| < \infty$ for all $s \in \mathcal{T}^*$. We apply Lemma 3.6 to φ^m to find a strictly increasing sequence of positive, integer-valued, $\mathcal{F}_{T-1}$-measurable random variables ρ^m so that φ^{ρ^m} converges a.s. to some process φ such that $\varphi_s \in L^0(\Omega, \mathcal{F}_{t_k-1}, \mathbb{P}; \mathbb{R}^N)$ for $s = 1, \dots, t_k - 1$, and $\varphi_s \in L^0(\Omega, \mathcal{F}_{s-1}, \mathbb{P}; \mathbb{R}^N)$ for $s = t_k, \dots, T$.[3] By observation (ii) in Step 1a, we have that φ is predictable.

Step 1c: We proceed by showing that **NAEF** implies $\mathbb{P}(E^0) = 0$. Toward this, we first show that the process φ constructed in Step 1b satisfies $F(\varphi) \in \mathbb{K}$. For the sake of notation, we define $\eta^m := \beta_0^{\rho^m}$. From (3.9), we have $\varphi^{\rho^m} = 1_{E^k} \phi^{\eta^m} / U^{\rho^m}$. Since 1_{E^k} and U^{ρ^m} are nonnegative, $\mathbb{R}$-valued random variables,

$$1_{E^k} \frac{F(\phi^{\eta^m})}{U^{\rho^m}} = F\left(1_{E^k} \frac{\phi^{\eta^m}}{U^{\rho^m}}\right) = F(\varphi^{\rho^m}). \tag{3.10}$$

Because φ^{ρ^m} converges a.s. to φ, we may apply Lemma 2.13 to see that $F(\varphi^{\rho^m})$ converges a.s. to $F(\varphi)$. Since φ is predictable, we have from the definition of $\mathbb{K}$ that $F(\varphi) \in \mathbb{K}$.

We proceed by showing that $F(\varphi) \in L_+^0(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$. Let us begin by defining $\widetilde{X}^m := X^{\eta^m} / U^{\rho^m}$ and $\widetilde{Z}^m := Z^{\eta^m} / U^{\rho^m}$. From (3.3),

$$F(\phi^{\eta^m}) = X^{\eta^m} + Z^{\eta^m}. \tag{3.11}$$

By multiplying both sides of (3.11) by $1_{E^k} / U^{\rho^m}$, we see from (3.10) that

$$F(\varphi^{\rho^m}) = 1_{E^k}(\widetilde{X}^m + \widetilde{Z}^m). \tag{3.12}$$

The sequence X^m converges a.s. by assumption, so the sequence X^{η^m} also converges a.s. Recall that the sequence $U^m(\omega)$ diverges for a.e. $\omega \in E^k$, so $U^{\rho^m}(\omega)$ diverges for a.e. $\omega \in E^k$ since $\{\rho^1(\omega), \rho^2(\omega), \dots, \} \subseteq \mathbb{N}$ for a.e. $\omega \in \Omega$.[4] Hence, $1_{E^k} \widetilde{X}^m$ converges a.s. to 0. Since $F(\varphi^{\rho^m})$ and $1_{E^k} \widetilde{X}^m$ converge a.s., the sequence $1_{E^k} \widetilde{Z}^m$ also converges a.s. to some $Z \in L_+^0(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$. Thus, $F(\varphi^{\rho^m})$ converges a.s. to Z, which implies $F(\varphi) \in L_+^0(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$.

[3] See observation (ii) in Step 1a.

[4] See observation (vi) in Step 1a.

Since $F(\varphi) \in \mathbb{K}$, we immediately see that $F(\varphi) \in \mathbb{K} \cap L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$. It is assumed that **NA** is satisfied, so by Lemma 3.2 we deduce that $F(\varphi) = 0$. We are supposing that **EF** holds, so according to Lemma 3.4 we have $\varphi = 0$. This cannot happen given our assumption that $\mathbb{P}(E^k) > 0$ because $\|\varphi_{t_k}\| = 1_{E^k}$.[5] Therefore, we must have that $\mathbb{P}(E^k) = 0$. This contradicts the construction in Step 1a, so $\mathbb{P}(E^0) = 0$.

Step 2: By the conclusion in Step 1, we obtain that $\limsup_m \|\phi^m_s\| < \infty$ for $s \in \mathcal{T}^*$. By applying Lemma 3.6 to ϕ^m, we may find a strictly increasing sequence of positive, integer-valued, $\mathcal{F}_{T-1}$-measurable random variables σ^m such that ϕ^{σ^m} converges a.s. to some predictable process ϕ.

By Lemma 2.13, the sequence $F(\phi^{\sigma^m})$ converges a.s. to $F(\phi)$. Since $\phi \in \mathcal{P}$, we have $F(\phi) \in \mathbb{K}$. Because X^m converges a.s. to X, the sequence X^{σ^m} also converges a.s. to X. From (3.3), it is true that $X^{\sigma^m} = F(\phi^{\sigma^m}) - Z^{\sigma^m}$. Since X^{σ^m} and $F(\phi^{\sigma^m})$ converges a.s., the sequence Z^{σ^m} also converges a.s. Thus, $F(\phi^{\sigma^m}) - X^{\sigma^m}$ converges a.s. to some nonnegative random variable $Z := F(\phi) - X$, which gives us that $X = F(\phi) - Z$. We conclude that $X \in \mathbb{K} - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$. □

3.2. The FFTAP

In this section, we formulate and prove a version of the FFTAP. We define the following set for convenience:

$$\mathcal{Z} := \{\mathbb{Q} : \ \mathbb{Q} \sim \mathbb{P}, \ P^{\text{ask},*}, P^{\text{bid},*}, A^{\text{ask},*}, A^{\text{bid},*} \text{ are } \mathbb{Q}\text{-integrable}\}.$$

We now define a risk-neutral measure in our context.

DEFINITION 3.9. A probability measure $\mathbb{Q}$ is a *risk-neutral measure* if $\mathbb{Q} \in \mathcal{Z}$, and if $\mathbb{E}_{\mathbb{Q}}[V^*_T(\phi)] \leq 0$ for all $\phi \in \mathcal{S}$ such that ϕ^j is bounded a.s., for $j \in \mathcal{J}^*$, and $V_0(\phi) = 0$.

We note that $V^*_T(\phi)$ is $\mathbb{Q}$-integrable under any $\mathbb{Q} \in \mathcal{Z}$, and any $\phi \in \mathcal{S}$ such that ϕ^j is bounded a.s., for $j \in \mathcal{J}^*$, and $V_0(\phi) = 0$.

REMARK 3.10. For frictionless markets ($P := P^{\text{ask}} = P^{\text{bid}}$, $A := A^{\text{ask}} = A^{\text{bid}}$), a risk-neutral measure is classically defined to be an equivalent probability measure such that the discounted cumulative price process $(P_t + \sum_{u=1}^t \Delta A_u)_{u=0}^T$ is a martingale under $\mathbb{Q}$. The present definition of a risk-neutral coincides with this classic definition of a risk-neutral measure if the market is frictionless. Indeed, if there are no frictions the value process satisfies $V^*_T(-\phi) = -V^*_T(\phi)$ for all trading strategies. Also, by Proposition 2.9, we have that $\mathbb{E}_{\mathbb{Q}}[V^*_T(\phi)] = 0$ for all $\phi \in \mathcal{S}$ such that ϕ^j is bounded a.s. and $V_0(\phi) = 0$ for $j \in \mathcal{J}^*$ if and only if

$$\sum_{j=1}^N \sum_{u=1}^T \mathbb{E}_{\mathbb{Q}} \left[\phi^j_u \mathbb{E}_{\mathbb{Q}} \left[\Delta \left(B^{-1}_u P^j_u + \sum_{w=1}^u B^{-1}_w \Delta A^j_w \right) \middle| \mathcal{F}_{u-1} \right] \right] = 0$$

for all $\phi \in \mathcal{S}$ such that ϕ^j is bounded a.s., for $j \in \mathcal{J}^*$.

The next lemma provides a mathematically convenient condition that is equivalent to **NA**. The proof is straightforward and is skipped.

[5] See observation (iv) in Step 1a.

LEMMA 3.11. *The no-arbitrage condition* ***(NA)*** *is satisfied if and only if for each* $\phi \in \mathcal{S}$ *such that* ϕ^j *is bounded a.s. for* $j \in \mathcal{J}^*$, $V_0(\phi) = 0$, *and* $V_T(\phi) \geq 0$, *we have* $V_T(\phi) = 0$.

Next, we recall the well-known Kreps–Yan Theorem. It was first proved by Yan (1980), and then obtained independently by Kreps (1981) in the context of financial mathematics. For a proof of the version presented in this paper, see Schachermayer (1992). Theorem 3.8 and the Kreps–Yan Theorem will essentially imply the FFTAP (Theorem 3.13).

THEOREM 3.12 (Kreps–Yan). *Let* $\mathcal{C}$ *be a closed convex cone in* $L^1(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R})$ *containing* $L^1_-(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R})$ *such that* $\mathcal{C} \cap L^1_+(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}) = \{0\}$. *Then there exists a functional* $f \in L^\infty(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R})$ *such that, for each* $h \in L^1_+(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R})$ *with* $h \neq 0$, *we have that* $\mathbb{E}_{\mathbb{P}}[fh] > 0$ *and* $\mathbb{E}_{\mathbb{P}}[fg] \leq 0$ *for any* $g \in \mathcal{C}$.

We are now ready to prove the following version of the FFTAP. To prove the FFTAP, we will employ the usual separation argument (cf. Schachermayer 1992).

THEOREM 3.13 (FFTAP). *The following conditions are equivalent:*

(i) The no-arbitrage condition under the efficient friction assumption ***(NAEF)*** *is satisfied.*
(ii) There exists a risk-neutral measure.
(iii) There exists a risk-neutral measure $\mathbb{Q}$ *so that* $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P} \in L^\infty(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$.

Proof. In order to prove these equivalences, we show that $(ii) \Rightarrow (i)$, $(i) \Rightarrow (iii)$, and $(iii) \Rightarrow (ii)$. The implications $(iii) \Rightarrow (ii)$ and $(ii) \Rightarrow (i)$ are immediate, so we only show $(i) \Rightarrow (iii)$.

Let us first define the weight function

$$w := 1 + \sum_{u=0}^{T} \left\| P_u^{\mathrm{ask},*} \right\| + \sum_{u=0}^{T} \left\| P_u^{\mathrm{bid},*} \right\| + \sum_{u=1}^{T} \left\| A_u^{\mathrm{ask},*} \right\| + \sum_{u=1}^{T} \left\| A_u^{\mathrm{bid},*} \right\|, \tag{3.13}$$

and let $\widetilde{\mathbb{P}}$ be the measure on $\mathcal{F}_T$ with Radon–Nikodým derivative $\mathrm{d}\widetilde{\mathbb{P}}/\mathrm{d}\mathbb{P} = \tilde{c}/w$, where $\tilde{c}$ is an appropriate normalizing constant. The processes $P^{\mathrm{ask},*}$, $P^{\mathrm{bid},*}$, $A^{\mathrm{ask},*}$, $A^{\mathrm{bid},*}$ are $\widetilde{\mathbb{P}}$-integrable, so $\widetilde{\mathbb{P}} \in \mathcal{Z}$. We consider the convex cone $\mathcal{C} := \left(\mathcal{K} - L^0_+(\Omega, \mathcal{F}_T, \widetilde{\mathbb{P}}; \mathbb{R})\right) \cap L^1(\Omega, \mathcal{F}_T, \widetilde{\mathbb{P}}; \mathbb{R})$, which is closed in $L^1(\Omega, \mathcal{F}_T, \widetilde{\mathbb{P}}; \mathbb{R})$ according to Theorem 3.8. By Theorem 3.12 there exists a strictly positive functional $f \in L^\infty(\Omega, \mathcal{F}_T, \widetilde{\mathbb{P}}; \mathbb{R})$ such that $\mathbb{E}_{\widetilde{\mathbb{P}}}[Kf] \leq 0$ for all $K \in \mathcal{K} \cap L^1(\Omega, \mathcal{F}_T, \widetilde{\mathbb{P}}; \mathbb{R})$.[6] This implies that $\mathbb{E}_{\widetilde{\mathbb{P}}}[V_T^*(\phi) f] \leq 0$ for all $\phi \in \mathcal{S}$ such that ϕ^j is bounded a.s., for $j \in \mathcal{J}^*$, $V_0(\phi) = 0$, and $V_T^*(\phi) \in L^1(\Omega, \mathcal{F}_T, \widetilde{\mathbb{P}}; \mathbb{R})$. Now let $\mathbb{Q}$ be the measure on $\mathcal{F}_T$ with Radon–Nikodým derivative $\mathrm{d}\mathbb{Q}/\mathrm{d}\widetilde{\mathbb{P}} := cf$, where c is an appropriate normalizing constant. We conclude that note that $\mathbb{Q}$ is a risk-neutral measure satisfying $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P} \in L^\infty(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$. □

REMARK 3.14.

(i) Note that **EF** is not needed to prove the implication $(ii) \Rightarrow (i)$.
(ii) In practice, it is typically required for a market model to satisfy **NA**. According to Theorem 3.13, it is enough to check that there exists a risk-neutral measure. However, this is not straightforward because it has to be verified whether there exists a probability measure $\mathbb{Q} \in \mathcal{Z}$ so that $\mathbb{E}_{\mathbb{Q}}[V_T^*(\phi)] \leq 0$ *for all* $\phi \in \mathcal{S}$ so that ϕ^j is bounded a.s., for $j \in \mathcal{J}^*$, and $V_0(\phi) = 0$. We will show in the following section that CPSs help solve this issue (see Proposition 4.3 and Theorem 4.7).

[6] For each $h \in L^1_+(\Omega, \mathcal{F}_T, \widetilde{\mathbb{P}}; \mathbb{R})$ with $h \neq 0$, we have $\mathbb{E}_{\widetilde{\mathbb{P}}}[fh] > 0$.

4. CONSISTENT PRICING SYSTEMS

CPSs are instrumental in the theory of arbitrage in markets with transaction costs—they provide a bridge between martingale theory in the theory of arbitrage in frictionless markets and more general concepts in the theory of arbitrage in markets with transaction costs. Essentially, CPSs are interpreted as corresponding auxiliary frictionless markets. In this section, we explore the relationship between CPSs and **NA**.

We begin by defining a CPS in our context.

DEFINITION 4.1. A *consistent pricing system* corresponding to the market $(B, P^{\text{ask}}, P^{\text{bid}}, A^{\text{ask}}, A^{\text{bid}})$ is a quadruplet $\{\mathbb{Q}, P, A, M\}$ consisting of

(i) a probability measure $\mathbb{Q} \in \mathcal{Z}$;
(ii) an adapted process P satisfying $P^{\text{bid},*} \leq P \leq P^{\text{ask},*}$;
(iii) an adapted process A satisfying $A^{\text{ask},*} \leq A \leq A^{\text{bid},*}$;
(iv) a martingale M under $\mathbb{Q}$ satisfying $M_t = P_t + \sum_{u=1}^{t} A_u$ for all $t \in \mathcal{T}$.

REMARK 4.2. Since our market is fixed throughout the paper, we shall simply refer to $\{\mathbb{Q}, P, A, M\}$ as a CPS, rather than a CPS corresponding to the market $(B, P^{\text{ask}}, P^{\text{bid}}, A^{\text{ask}}, A^{\text{bid}})$. Note that one can drop M from the definition of a CPS since it is uniquely expressed in terms of A and P. However, we will keep it to simplify notation, and because M has the following key interpretation: for a CPS $\{\mathbb{Q}, P, A, M\}$, the process P is interpreted as the corresponding auxiliary frictionless ex-dividend price process, and the process A has the interpretation of the corresponding auxiliary frictionless cumulative dividend process, whereas M is viewed as the corresponding auxiliary frictionless cumulative price process.

The next result establishes a relationship between **NA** and CPSs in our context.

PROPOSITION 4.3. *If there exists a CPS, then the no-arbitrage condition* ***(NA)*** *is satisfied.*

Proof. Suppose there exists a CPS, call it $\{\mathbb{Q}, P, A, M\}$, and suppose $\phi \in \mathcal{S}$ is a trading strategy such that ϕ^j is bounded a.s., for $j \in \mathcal{J}^*$, and $V_0(\phi) = 0$. In view of Proposition 2.9, and because $P^{\text{bid}} \leq P \leq P^{\text{ask}}$ and $A^{\text{ask}} \leq A \leq A^{\text{bid}}$, we deduce that

$$V_T^*(\phi) \leq \sum_{j=1}^{N} \left(\phi_T^j P_T^j + \sum_{u=1}^{T} \left(-\Delta\phi_u^j P_{u-1}^j + \phi_u^j A_u^j \right) \right).$$

Since $M = P + \sum_{u=1}^{\cdot} A_u$ is a martingale under $\mathbb{Q}$, and because ϕ^j is bounded a.s., for $j \in \mathcal{J}^*$, we have

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}}\left[V_T^*(\phi) \right] &\leq \sum_{j=1}^{N} \mathbb{E}_{\mathbb{Q}} \left[\phi_T^j P_T^j + \sum_{u=1}^{T} \left(-\Delta\phi_u^j P_{u-1}^j + \phi_u^j A_u^j \right) \right] \\ &= \sum_{j=1}^{N} \sum_{u=1}^{T} \mathbb{E}_{\mathbb{Q}}\left[\Delta\phi_u^j \mathbb{E}_{\mathbb{Q}}\left[M_T^j - M_{u-1}^j \middle| \mathcal{F}_{u-1} \right] \right] = 0. \end{aligned}$$

Therefore $\mathbb{Q}$ is a risk-neutral measure. According to Theorem 3.13, **NA** holds. □

At this point, a natural question to ask is whether there exists a CPS whenever **NA** is satisfied. In general, this is still an open question. However, for the special case in which

there are no transaction costs in the dividends paid by the securities, $A^{\text{ask}} = A^{\text{bid}}$, will show in Theorem 4.7 that there exists a CPS if and only if **NAEF** is satisfied.

Proposition 4.3 is important from the modeling point of view because it provides a sufficient condition for a model to satisfy **NA**. In the next example, we construct a model for which there exists a CPS.

EXAMPLE 4.4. Let us consider the CDS specified in Example 2.2. Recall that the cumulative dividend processes A^{ask} and A^{bid} corresponding to the CDS are defined as

$$A_t^{\text{ask}} := 1_{\{\tau \le t\}}\delta - \kappa^{\text{ask}} \sum_{u=1}^{t} 1_{\{u<\tau\}}, \quad A_t^{\text{bid}} := 1_{\{\tau \le t\}}\delta - \kappa^{\text{bid}} \sum_{u=1}^{t} 1_{\{u<\tau\}}$$

for all $t \in \mathcal{T}^*$. Let us fix any probability measure $\mathbb{Q}$ equivalent to $\mathbb{P}$. We postulate that the ex-dividend prices P^{ask} and P^{bid} satisfy

$$P_t^{\text{ask},*} = \mathbb{E}_{\mathbb{Q}}\left[\sum_{u=t+1}^{T} A_u^{\text{bid},*}\middle|\mathcal{F}_t\right], \quad P_t^{\text{bid},*} = \mathbb{E}_{\mathbb{Q}}\left[\sum_{u=t+1}^{T} A_u^{\text{ask},*}\middle|\mathcal{F}_t\right],$$

for all $t \in \mathcal{T}^*$. By substituting $A^{\text{ask},*}$ and $A^{\text{bid},*}$ into the equations for $P^{\text{ask},*}$ and $P^{\text{bid},*}$ above, we see that

$$P_t^{\text{ask},*} = \mathbb{E}_{\mathbb{Q}}\left[1_{\{t<\tau\le T\}} B_\tau^{-1}\delta - \kappa^{\text{bid}} \sum_{u=t+1}^{T} B_u^{-1} 1_{\{u<\tau\}}\middle|\mathcal{F}_t\right],$$

$$P_t^{\text{bid},*} = \mathbb{E}_{\mathbb{Q}}\left[1_{\{t<\tau\le T\}} B_\tau^{-1}\delta - \kappa^{\text{ask}} \sum_{u=t+1}^{T} B_u^{-1} 1_{\{u<\tau\}}\middle|\mathcal{F}_t\right].$$

For a fixed $\kappa \in [\kappa^{\text{bid}}, \kappa^{\text{ask}}]$, we define

$$A_t := B_t^{-1}(1_{\{\tau=t\}}\delta - \kappa 1_{\{t<\tau\}}), \qquad t \in \mathcal{T}^*,$$

$$P_t := \mathbb{E}_{\mathbb{Q}}\left[\sum_{u=t+1}^{T} A_u\middle|\mathcal{F}_t\right] = \mathbb{E}_{\mathbb{Q}}\left[1_{\{t<\tau\le T\}} B_\tau^{-1}\delta - \kappa \sum_{u=t+1}^{T} B_u^{-1} 1_{\{u<\tau\}}\middle|\mathcal{F}_t\right], \quad t \in \mathcal{T},$$

$$M_t := P_t + \sum_{u=1}^{t} A_u, \qquad t \in \mathcal{T}.$$

The quadruplet $\{\mathbb{Q}, P, A, M\}$ is a CPS. To see this, first observe that A and P are $\mathbb{Q}$-integrable since A is bounded $\mathbb{Q}$-a.s. Thus, $\mathbb{Q} \in \mathcal{Z}$. Next, M satisfies

$$M_t = \mathbb{E}_{\mathbb{Q}}\left[1_{\{\tau\le T\}} B_\tau^{-1}\delta - \kappa \sum_{u=1}^{T} B_u^{-1} 1_{\{u<\tau\}}\middle|\mathcal{F}_t\right], \quad t \in \mathcal{T},$$

so M is a Doob martingale under $\mathbb{Q}$. Also, since $\kappa \in [\kappa^{\text{bid}}, \kappa^{\text{ask}}]$, we have $A^{\text{ask},*} \le A \le A^{\text{bid},*}$ and $P^{\text{bid},*} \le P \le P^{\text{ask},*}$. Thus, $\{\mathbb{Q}, P, A, M\}$ is a CPS. According to Proposition 4.3, we may additionally conclude that the financial market model $\{B, P^{\text{ask}}, P^{\text{bid}}, A^{\text{ask}}, A^{\text{bid}}\}$ satisfies **NA**.

4.1. CPSs under the Assumption $A^{\text{ask}} = A^{\text{bid}}$

In this section, we investigate the relationship between risk-neutral measures and CPSs under the assumption $A^{\text{ask}} = A^{\text{bid}}$. We begin by proving two preliminary lemmas that hold in general (without the assumption $A^{\text{ask}} = A^{\text{bid}}$).

LEMMA 4.5. *If $\mathbb{Q}$ is a risk-neutral measure, then*

$$P^{\text{bid,j},*}_{\sigma_1} \leq \mathbb{E}_{\mathbb{Q}}\left[P^{\text{ask,j},*}_{\sigma_2} + \sum_{u=\sigma_1+1}^{\sigma_2} A^{\text{bid,j},*}_u \middle| \mathcal{F}_{\sigma_1} \right],$$

$$P^{\text{ask,j},*}_{\sigma_1} \geq \mathbb{E}_{\mathbb{Q}}\left[P^{\text{bid,j},*}_{\sigma_2} + \sum_{u=\sigma_1+1}^{\sigma_2} A^{\text{ask,j},*}_u \middle| \mathcal{F}_{\sigma_1} \right],$$

for all $j \in \mathcal{J}^*$ and stopping times $0 \leq \sigma_1 < \sigma_2 \leq T$.

Proof. Suppose $\mathbb{Q}$ is a risk-neutral measure. For stopping times $0 \leq \sigma_1 < \sigma_2 \leq T$ and random variables $\xi_{\sigma_1} \in L^\infty(\Omega, \mathcal{F}_{\sigma_1}, \mathbb{P};\ \mathbb{R}^N)$, we define the trading strategy

$$\theta(\sigma_1, \sigma_2, \xi_{\sigma_1}) := \left(\theta^0_t(\sigma_1, \sigma_2, \xi_{\sigma_1}), 1_{\{\sigma_1+1 \leq t \leq \sigma_2\}}\xi^1_{\sigma_1}, \ldots, 1_{\{\sigma_1+1 \leq t \leq \sigma_2\}}\xi^N_{\sigma_1}\right)_{t=1}^T,$$

where $\theta^0(\sigma_1, \sigma_2, \xi_{\sigma_1})$ is chosen such that $\theta(\sigma_1, \sigma_2, \xi_{\sigma_1})$ is self-financing and $V_0(\theta(\sigma_1, \sigma_2, \xi_{\sigma_1})) = 0$. Due to Proposition 2.9, the value process associated with θ satisfies

$$\begin{aligned} V^*_T(\theta(\sigma_1, \sigma_2, \xi_{\sigma_1})) = &\sum_{j=1}^N 1_{\{\xi^j_{\sigma_1} \geq 0\}}\xi^j_{\sigma_1}\left(P^{\text{bid,j},*}_{\sigma_2} + \sum_{u=\sigma_1+1}^{\sigma_2} A^{\text{ask,j},*}_u - P^{\text{ask,j},*}_{\sigma_1} \right) \\ &+ \sum_{j=1}^N 1_{\{\xi^j_{\sigma_1} < 0\}}\xi^j_{\sigma_1}\left(P^{\text{ask,j},*}_{\sigma_2} + \sum_{u=\sigma_1+1}^{\sigma_2} A^{\text{bid,j},*}_u - P^{\text{bid,j},*}_{\sigma_1} \right). \end{aligned}$$

Since $\mathbb{Q}$ is a risk-neutral measure, we have $\mathbb{E}_{\mathbb{Q}}[V^*_T(\theta(\sigma_1, \sigma_2, \xi_{\sigma_1}))] \leq 0$ for all stopping times $0 \leq \sigma_1 < \sigma_2 \leq T$ and $\xi_{\sigma_1} \in L^\infty(\Omega, \mathcal{F}_{\sigma_1}, \mathbb{P};\ \mathbb{R}^N)$. Hence, we are able to obtain

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}}\Bigg[&\sum_{j=1}^N 1_{\{\xi^j_{\sigma_1} \geq 0\}}\xi^j_{\sigma_1}\mathbb{E}_{\mathbb{Q}}\left[P^{\text{bid,j},*}_{\sigma_2} + \sum_{u=\sigma_1+1}^{\sigma_2} A^{\text{ask,j},*}_u - P^{\text{ask,j},*}_{\sigma_1} \middle| \mathcal{F}_{\sigma_1} \right] \\ &+ \sum_{j=1}^N 1_{\{\xi^j_{\sigma_1} < 0\}}\xi^j_{\sigma_1}\mathbb{E}_{\mathbb{Q}}\left[P^{\text{ask,j},*}_{\sigma_2} + \sum_{u=\sigma_1+1}^{\sigma_2} A^{\text{bid,j},*}_u - P^{\text{bid,j},*}_{\sigma_1} \middle| \mathcal{F}_{\sigma_1} \right] \Bigg] \leq 0. \end{aligned}$$

for all stopping times $0 \leq \sigma_1 < \sigma_2 \leq T$ and random variables $\xi_{\sigma_1} \in L^\infty(\Omega, \mathcal{F}_{\sigma_1}, \mathbb{P};\ \mathbb{R}^N)$. This implies that the claim is satisfied. □

The next result is motivated by theorem 4.5 in Cherny (2007). We will denote by $\mathcal{T}_t$ the set of stopping times in $\{t, t+1, \ldots, T\}$, for all $t \in \mathcal{T}$.

LEMMA 4.6. *Suppose $\mathbb{Q}$ is a risk-neutral measure, and let*

$$X_s^{b,j} := \operatorname*{ess\,sup}_{\sigma\in\mathcal{T}_s} \mathbb{E}_{\mathbb{Q}}\left[P_\sigma^{\mathrm{bid,j},*} + \sum_{u=1}^{\sigma} A_u^{\mathrm{ask,j},*}\middle|\mathcal{F}_s\right],$$

$$X_s^{a,j} := \operatorname*{ess\,inf}_{\sigma\in\mathcal{T}_s} \mathbb{E}_{\mathbb{Q}}\left[P_\sigma^{\mathrm{ask,j},*} + \sum_{u=1}^{\sigma} A_u^{\mathrm{bid,j},*}\middle|\mathcal{F}_s\right]$$

for all $j \in \mathcal{J}^*$ and $s \in \mathcal{T}$. Then , X^b is a supermartingale and X^a is a submartingale, both under $\mathbb{Q}$, and satisfy $X^b \le X^a$.

Proof. Let us fix $j \in \mathcal{J}^*$. The processes $X^{b,j}$ and $X^{a,j}$ are Snell envelopes, so $X^{a,j}$ is a supermartingale and $X^{b,j}$ is a submartingale, both under $\mathbb{Q}$ (see, for instance, Föllmer and Schied 2004).

We now show that $X^{b,j} \le X^{a,j}$. For stopping times $\tau_1, \tau_2 \in \mathcal{T}_t$, let us define the process

$$X_t^j := \mathbb{E}_{\mathbb{Q}}\left[P_{\tau_1}^{\mathrm{bid,j},*} + \sum_{u=1}^{\tau_1} A_u^{\mathrm{ask,j},*}\middle|\mathcal{F}_t\right] - \mathbb{E}_{\mathbb{Q}}\left[P_{\tau_2}^{\mathrm{ask,j},*} + \sum_{u=1}^{\tau_2} A_u^{\mathrm{bid,j},*}\middle|\mathcal{F}_t\right], \quad t \in \mathcal{T}.$$

We see that

$$\begin{aligned} X_t^j &= \mathbb{E}_{\mathbb{Q}}\left[\mathbb{E}_{\mathbb{Q}}\left[P_{\tau_1}^{\mathrm{bid,j},*} + \sum_{u=1}^{\tau_1} A_u^{\mathrm{ask,j},*} - P_{\tau_2}^{\mathrm{ask,j},*} - \sum_{u=1}^{\tau_2} A_u^{\mathrm{bid,j},*}\middle|\mathcal{F}_{\tau_1\wedge\tau_2}\right]\middle|\mathcal{F}_t\right] \\ &= \mathbb{E}_{\mathbb{Q}}\left[1_{\{\tau_1\le\tau_2\}}\left(P_{\tau_1}^{\mathrm{bid,j},*} + \sum_{u=1}^{\tau_1} A_u^{\mathrm{ask,j},*} - \mathbb{E}_{\mathbb{Q}}\left[P_{\tau_2}^{\mathrm{ask,j},*} + \sum_{u=1}^{\tau_2} A_u^{\mathrm{bid,j},*}\middle|\mathcal{F}_{\tau_1}\right]\right)\middle|\mathcal{F}_t\right] \\ &\quad + \mathbb{E}_{\mathbb{Q}}\left[1_{\{\tau_1>\tau_2\}}\left(\mathbb{E}_{\mathbb{Q}}\left[P_{\tau_1}^{\mathrm{bid,j},*} + \sum_{u=1}^{\tau_1} A_u^{\mathrm{ask,j},*}\middle|\mathcal{F}_{\tau_2}\right] - P_{\tau_2}^{\mathrm{ask,j},*} - \sum_{u=1}^{\tau_2} A_u^{\mathrm{bid,j},*}\right)\middle|\mathcal{F}_t\right]. \end{aligned}$$

After rearranging terms and since $A^{\mathrm{ask},*} \le A^{\mathrm{bid},*}$, we obtain that

$$\begin{aligned} X_t^j &\le \mathbb{E}_{\mathbb{Q}}\left[1_{\{\tau_1\le\tau_2\}}\left(P_{\tau_1}^{\mathrm{bid,j},*} - \mathbb{E}_{\mathbb{Q}}\left[P_{\tau_2}^{\mathrm{ask,j},*} + \sum_{u=\tau_1+1}^{\tau_2} A_u^{\mathrm{bid,j},*}\middle|\mathcal{F}_{\tau_1}\right]\right)\middle|\mathcal{F}_t\right] \\ &\quad + \mathbb{E}_{\mathbb{Q}}\left[1_{\{\tau_1>\tau_2\}}\left(\mathbb{E}_{\mathbb{Q}}\left[P_{\tau_1}^{\mathrm{bid,j},*} + \sum_{u=\tau_2+1}^{\tau_1} A_u^{\mathrm{ask,j},*}\middle|\mathcal{F}_{\tau_2}\right] - P_{\tau_2}^{\mathrm{ask,j},*}\right)\middle|\mathcal{F}_t\right]. \end{aligned} \tag{4.1}$$

Since $\mathbb{Q}$ is a risk-neutral measure, we see from Lemma 4.5 and (4.1) that $X_t^j \le 0$. The stopping times τ_1 and τ_2 are arbitrary in the definition of X^j, so we conclude that $X^{b,j} \le X^{a,j}$. □

The next theorem gives sufficient and necessary conditions for there to exist a CPS (cf. Kabanov et al. 2002; Schachermayer 2004).

THEOREM 4.7. *Under the assumption that $A^{\mathrm{ask}} = A^{\mathrm{bid}}$, there exists a CPS if and only if the no-arbitrage condition under the efficient condition **(NAEF)** is satisfied.*

Proof. Let us denote by A the process A^{ask}. Necessity is shown in Proposition 4.3, so we only prove sufficiency. Suppose that **NAEF** is satisfied. According to Theorem 3.13,

there exists a risk-neutral measure $\mathbb{Q}$. By Lemma 4.5,

$$P^{\mathrm{bid,j},*}_{\sigma_1} \le \mathbb{E}_{\mathbb{Q}}\left[P^{\mathrm{ask,j},*}_{\sigma_2} + \sum_{u=\sigma_1+1}^{\sigma_2} A^{j,*}_u \middle| \mathcal{F}_{\sigma_1} \right], \quad P^{\mathrm{ask,j},*}_{\sigma_1} \ge \mathbb{E}_{\mathbb{Q}}\left[P^{\mathrm{bid,j},*}_{\sigma_2} + \sum_{u=\sigma_1+1}^{\sigma_2} A^{j,*}_u \middle| \mathcal{F}_{\sigma_1} \right]$$

for all $j \in \mathcal{J}^*$ and stopping times $0 \le \sigma_1 < \sigma_2 \le T$. Now, let us define the processes

$$\begin{aligned} Y^{b,j}_t &:= \operatorname*{ess\,sup}_{\sigma \in \mathcal{T}_t} \mathbb{E}_{\mathbb{Q}}\left[P^{\mathrm{bid,j},*}_{\sigma} + \sum_{u=t+1}^{\sigma} A^{j,*}_u \middle| \mathcal{F}_t \right], \\ Y^{a,j}_t &:= \operatorname*{ess\,inf}_{\sigma \in \mathcal{T}_t} \mathbb{E}_{\mathbb{Q}}\left[P^{\mathrm{ask,j},*}_{\sigma} + \sum_{u=t+1}^{\sigma} A^{j,*}_u \middle| \mathcal{F}_t \right], \\ X^{b,j}_t &:= Y^{b,j}_t + \sum_{u=1}^{t} A^{j,*}_t, \quad X^{a,j}_t := Y^{a,j}_t + \sum_{u=1}^{t} A^{j,*}_t \end{aligned} \tag{4.2}$$

for all $t \in \mathcal{T}$ and $j \in \mathcal{J}^*$. From Lemma 4.6, we know that under $\mathbb{Q}$ the process X^a is a submartingale and the process X^b is a supermartingale, and that they satisfy $X^b \le X^a$.

For $t = 0, 1, \ldots, T-1$ and $j \in \mathcal{J}^*$, recursively define

$$\begin{aligned} M^j_0 &:= Y^{a,j}_0, \quad P^j_0 := Y^{a,j}_0, \quad P^j_{t+1} := \lambda^j_t Y^{a,j}_{t+1} + \left(1 - \lambda^j_t\right) Y^{b,j}_{t+1}, \\ M^j_{t+1} &:= P^j_{t+1} + \sum_{u=1}^{t+1} A^j_u, \end{aligned} \tag{4.3}$$

where λ^j_t satisfies

$$\lambda^j_t = \begin{cases} \dfrac{M^j_t - \mathbb{E}_{\mathbb{Q}}\left[X^{b,j}_{t+1} \middle| \mathcal{F}_t\right]}{\mathbb{E}_{\mathbb{Q}}\left[X^{a,j}_{t+1} - X^{b,j}_{t+1} \middle| \mathcal{F}_t\right]}, & \text{if } \mathbb{E}_{\mathbb{Q}}\left[X^{a,j}_{t+1} \middle| \mathcal{F}_t\right] \ne \mathbb{E}_{\mathbb{Q}}\left[X^{b,j}_{t+1} \middle| \mathcal{F}_t\right], \\ \dfrac{1}{2}, & \text{otherwise.} \end{cases} \tag{4.4}$$

Let us fix $j \in \mathcal{J}^*$ for the rest of the proof.

Step 1: In this step, we show that the processes above are well-defined and adapted. First, note that P_0 and M_0 are well-defined, and that, by (4.4),

$$\lambda^j_0 = \frac{M^j_0 - \mathbb{E}_{\mathbb{Q}}\left[X^{b,j}_1 \middle| \mathcal{F}_0\right]}{\mathbb{E}_{\mathbb{Q}}\left[X^{a,j}_1 - X^{b,j}_1 \middle| \mathcal{F}_0\right]}, \quad \text{or } \lambda^j_0 = \frac{1}{2}.$$

Thus, λ^j_0 is well-defined and $\mathcal{F}_0$-measurable. Next, we compute P^j_1 and M^j_1, and consequently we compute λ^j_1; all of them being $\mathcal{F}_1$-measurable. Inductively, we see that P^j_t, M^j_t, and λ^j_t, for $t = 2, \ldots, T$ are well-defined and $\mathcal{F}_t$-measurable.

Step 2: We inductively show that $\lambda^j_t \in [0, 1]$ for $t = 0, 1, \ldots, T-1$. We first show that $\lambda^j_0 \in [0, 1]$. If $\mathbb{E}_{\mathbb{Q}}[X^{a,j}_1 - X^{b,j}_1 | \mathcal{F}_0] = 0$, then $\lambda^j_0 \in [0, 1]$ automatically, so suppose that $\mathbb{E}_{\mathbb{Q}}[X^{a,j}_1 - X^{b,j}_1 | \mathcal{F}_0] > 0$. Now, by the definition of M^j, we have that $M^j_0 = X^{a,j}_0$, so (4.4)

gives that

$$\lambda_0^j = \frac{X_0^{a,j} - \mathbb{E}_{\mathbb{Q}}\big[X_1^{b,j}\big|\mathcal{F}_0\big]}{\mathbb{E}_{\mathbb{Q}}\big[X_1^{a,j} - X_1^{b,j}\big|\mathcal{F}_0\big]}\,. \tag{4.5}$$

The process $X^{a,j}$ is a submartingale under $\mathbb{Q}$, so it immediately follows that $\lambda_0^j \le 1$. On the other hand, since $X^{b,j}$ is a supermartingale under $\mathbb{Q}$,

$$\lambda_0^j \ge \frac{X_0^{a,j} - X_0^{b,j}}{\mathbb{E}_{\mathbb{Q}}\big[X_1^{a,j} - X_1^{b,j}\big|\mathcal{F}_0\big]}\,.$$

Because $X_0^{a,j} \ge X_0^{b,j}$, we deduce that $\lambda_0^j \ge 0$.

Suppose that $\lambda_t^j \in [0,1]$ for $t = 0, 1, \ldots, T-2$. We now prove that $\lambda_{T-1}^j \in [0,1]$. If $\mathbb{E}_{\mathbb{Q}}[X_T^{a,j} - X_T^{b,j}|\mathcal{F}_{T-1}] = 0$, then $\lambda_{T-1}^j = 1/2$, so assume that $\mathbb{E}_{\mathbb{Q}}[X_T^{a,j} - X_T^{b,j}|\mathcal{F}_{T-1}] > 0$. According to (4.4) and the definition of M^j, we have that

$$\lambda_{T-1}^j = \frac{\lambda_{T-2}^j X_{T-1}^{a,j} + \big(1-\lambda_{T-2}^j\big) X_{T-1}^{b,j} - \mathbb{E}_{\mathbb{Q}}\big[X_T^{b,j}\big|\mathcal{F}_{T-1}\big]}{\mathbb{E}_{\mathbb{Q}}\big[X_T^{a,j} - X_T^{b,j}\big|\mathcal{F}_{T-1}\big]}\,. \tag{4.6}$$

Since $\lambda_{T-2}^j \le 1$, and because $X^{b,j}$ is a supermartingale under $\mathbb{Q}$, we get that

$$\lambda_{T-1}^j \ge \frac{\lambda_{T-2}^j\big(X_{T-1}^{a,j} - X_{T-1}^{b,j}\big)}{\mathbb{E}_{\mathbb{Q}}\big[X_T^{a,j} - X_T^{b,j}\big|\mathcal{F}_{T-1}\big]}\,.$$

Because $X^{a,j} \ge X^{b,j}$, we arrive at $\lambda_{T-1}^j \ge 0$. Now, since $X_{T-1}^{a,j} \ge X_{T-1}^{b,j}$ and $\lambda_{T-2}^j \le 1$, we see from (4.6) that

$$\lambda_{T-1}^j \le \frac{X_{T-1}^{a,j} - \mathbb{E}_{\mathbb{Q}}\big[X_T^{b,j}\big|\mathcal{F}_{T-1}\big]}{\mathbb{E}_{\mathbb{Q}}\big[X_T^{a,j} - X_T^{b,j}\big|\mathcal{F}_{T-1}\big]}\,.$$

The process $X^{a,j}$ is a submartingale under $\mathbb{Q}$, so it follows that $\lambda_{T-1}^j \le 1$. We conclude that $\lambda_t^j \in [0,1]$ for $t = 0, 1, \ldots, T-1$.

Step 3: Next, we show that M is a martingale under $\mathbb{Q}$. First, we note that by (4.2) and (4.3) we have

$$M_{t+1}^j = \lambda_t^j X_{t+1}^{a,j} + \big(1-\lambda_t^j\big) X_{t+1}^{b,j}. \tag{4.7}$$

From here, the $\mathbb{Q}$-integrability of M^j follows from $\mathbb{Q}$-integrability of $X^{a,j}$, $X^{b,j}$ and boundedness of λ^j. From (4.4) and (4.7), we get that $\mathbb{E}_{\mathbb{Q}}[M_{t+1}^j|\mathcal{F}_t] = M_t^j$, for $t = 0, 1, \ldots, T-1$. Hence, M^j is a martingale under $\mathbb{Q}$.

Step 4: We continue by showing that P^j satisfies $P^{\mathrm{bid},*,\mathrm{j}} \le P^j \le P^{\mathrm{ask},*,\mathrm{j}}$. Let us first show that $P_0^{\mathrm{bid},\mathrm{j}} \le P_0^j \le P_0^{\mathrm{ask},\mathrm{j}}$. By definition of P_0^j, we have that $P_0^j = Y_0^{a,j}$, and by (4.2) we see that $Y_0^{a,j} = X_0^{a,j}$. Therefore, the claim holds since $P_0^{\mathrm{bid},\mathrm{j}} \le X_0^{a,j} \le P_0^{\mathrm{ask},\mathrm{j}}$.

We proceed by proving that $P_t^{\mathrm{bid},\mathrm{j}} \le P_t^j \le P_t^{\mathrm{ask},\mathrm{j}}$ for all $t \in \{1, \ldots, T\}$. Toward this, let $t \in \{1, \ldots, T\}$. By the definition of P_t^j, we have $P_t^j = \lambda_{t-1}^j Y_t^{a,j} + (1-\lambda_{t-1}^j) Y_t^{b,j}$. From (4.2), it is true that $X_t^{a,j} \ge X_t^{b,j}$ if and only if $Y_t^{a,j} \ge Y_t^{b,j}$. Also, since $t \in \mathcal{T}_t$, we see from (4.2) that $Y_t^{b,j} \ge P_t^{\mathrm{bid},\mathrm{j},*}$ and $Y_t^{a,j} \le P_t^{\mathrm{ask},\mathrm{j}}$. According to Step 1, $\lambda_{t-1}^j \in [0,1]$. So, putting

everything together, we obtain

$$P_t^{\text{bid,j}} \leq Y_t^{b,j} \leq P_t^j \leq Y_t^{a,j} \leq P_t^{\text{ask,j}}.$$

We conclude that $\{\mathbb{Q}, P, A, M\}$ is a CPS. □

The general case $A^{\text{ask}} \leq A^{\text{bid}}$ is harder to handle because transaction costs on dividends accrue over time. This makes it hard to construct a process A from A^{ask} and A^{bid}, satisfying $A^{\text{ask},*} \leq A \leq A^{\text{bid},*}$, so that $P + \sum_{u=1}^{\cdot} A_u$ is a martingale under some $\mathbb{Q} \in \mathcal{Z}$.

5. SUPERHEDGING AND SUBHEDGING THEOREM

In this section, we define the superhedging ask and subhedging bid prices for a dividend-paying contingent claim, and then we provide an important representation theorem for these prices. The representation theorem is important because it provides an alternative way of computing the superhedging ask prices and the superhedging bid prices. Also, it is an application of the Fundamental Theorem of Asset Pricing: the theorem relates how the no-arbitrage condition (and hence the existence of risk-neutral measures) is related to the pricing of contingent claims.

For results related to this topic, both for discrete-time and continuous-time markets with transaction costs, we refer to, among others, Soner, Shreve, and Cvitanić (1995); Levental and Skorohod (1997); Cvitanic, Pham, and Touzi (1999); Touzi (1999); Bouchard and Touzi (2000); Kabanov et al. (2002); Schachermayer (2004); Campi and Schachermayer (2006); Cherny (2007); Pennanen (2011a,b,c,d). Our contribution to this literature is that we consider dividend-paying securities such as swap contracts as hedging securities.

A *contingent claim* D is any a.s. bounded, $\mathbb{R}$-valued, $\mathbb{F}$-adapted process. Here, D is interpreted as the spot cash flow process (not the cumulative cash flow process). We remark that the boundedness assumption on contingent claims is satisfied for fixed income securities.

Let us now define the set of self-financing trading strategies initiated at time $t \in \{0, 1, \ldots, T-1\}$ with bounded components ($j = 1, \ldots, N$) as

$$\mathcal{S}(t) := \{\phi \in \mathcal{S} : \phi^j \text{ is bounded a.s. for } j \in \mathcal{J}^*,\ \phi_s = 0 \text{ for all } s \leq t\},$$

and the set of attainable claims at zero cost initiated at time $t \in \{0, 1, \ldots, T-1\}$ as

$$\mathcal{K}(t) := \{V_T^*(\phi) : \phi \in \mathcal{S}(t) \text{ such that } V_0(\phi) = 0\}.$$

REMARK 5.1.

(i) $\mathcal{S}(t)$ and $\mathcal{K}(t)$ are closed with respect to multiplication by random variables in $L_+^\infty(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$.[7]

(ii) $\mathcal{S} \supset \mathcal{S}(0) \supset \mathcal{S}(1) \supset \cdots \supset \mathcal{S}(T-1)$ and $\mathcal{K} \supset \mathcal{K}(0) \supset \mathcal{K}(1) \supset \cdots \supset \mathcal{K}(T-1)$. Moreover, if $\mathbb{Q}$ is a risk-neutral measure, then $\mathbb{E}_{\mathbb{Q}}[K] \leq 0$ for all $K \in \mathcal{K}(t)$, for $t = 0, 1, \ldots, T-1$.

We proceed by defining the main objects of this section.

[7] $L_+^\infty(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R}) := \{X \in L^\infty(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R}) : X \geq 0\}$.

DEFINITION 5.2. The discounted superhedging ask and subhedging bid prices of a contingent claim D at time $t \in \{0, \dots, T-1\}$ are defined as $\pi_t^{\text{ask}}(D) := \operatorname{ess\,inf} \mathcal{W}^a(t, D)$ and $\pi_t^{\text{bid}}(D) := \operatorname{ess\,sup} \mathcal{W}^b(t, D)$, where

$$\mathcal{W}^a(t, D) := \left\{ W \in L^0(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R}) : -W + \sum_{u=t+1}^{T} D_u^* \in \mathcal{K}(t) - L_+^0(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R}) \right\},$$

$$\mathcal{W}^b(t, D) := \left\{ W \in L^0(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R}) : W - \sum_{u=t+1}^{T} D_u^* \in \mathcal{K}(t) - L_+^0(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R}) \right\}.$$

Note that $\pi_t^{\text{ask}}(D) = -\pi_t^{\text{bid}}(-D)$ and

$$\mathcal{W}^a(t, D) = \left\{ W \in L^0(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R}) : \exists\, K \in \mathcal{K}(t) \text{ such that } \sum_{u=t+1}^{T} D_u^* \le K + W \right\},$$

$$\mathcal{W}^b(t, D) = \left\{ W \in L^0(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R}) : \exists\, K \in \mathcal{K}(t) \text{ such that } -\sum_{u=t+1}^{T} D_u^* \le K - W \right\}.$$

REMARK 5.3.

(i) For each $t \in \{0, 1, \dots, T-1\}$, the prices $\pi_t^{\text{ask}}(D)$ and $\pi_t^{\text{bid}}(D)$ have the following interpretations: The price $\pi_t^{\text{ask}}(D)$ is interpreted as the least discounted cash amount W at time t so that the gain $-W + \sum_{u=t+1}^{T} D_u^*$ can be superhedged at zero cost. On the other hand, the random variable $\pi_t^{\text{bid}}(D)$ is interpreted as the greatest discounted cash amount W at time t so that the gain $W - \sum_{u=t+1}^{T} D_u^*$ can be superhedged at zero cost.

(ii) In view of (i) above, it is unreasonable for the discounted ex-dividend ask price at time $t \in \{0, 1, \dots, T-1\}$ of a contingent claim D to be a.s. greater than $\pi_t^{\text{ask}}(D)$, and for the ex-dividend bid price at time $t \in \{0, 1, \dots, T-1\}$ of a contingent claim D to be a.s. less than $\pi_t^{\text{bid}}(D)$.

(iii) Direction of trade matters: a market participant can buy a contingent claim D at price $\pi_t^{\text{ask}}(D)$ and sell D at price $\pi_t^{\text{bid}}(D)$. This is in contrast to frictionless markets, where a contingent claim can be bought and sold at the same price.

(iv) The prices $\pi_t^{\text{ask}}(D)$ and $\pi_t^{\text{bid}}(D)$ satisfy $\pi_t^{\text{ask}}(D) < \infty$ and $\pi_t^{\text{bid}}(D) > -\infty$. Indeed, since $0 \in \mathcal{K}(t)$, $1 \in L_{++}^0(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$, and $\sum_{u=t+1}^{T} D_u^*$ is a.s. bounded, say by M, we have that $-M + \sum_{u=t+1}^{T} D_u^* \in L_-^0(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$. Thus, $\pi_t^{\text{ask}}(D) \le M$. Similarly, $\pi_t^{\text{bid}}(D) \ge -M$.

Next, we define *the sets of extended attainable claims initiated at time* $t \in \{0, 1, \dots, T-1\}$ *associated with cash amount* $W \in L^0(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$:

$$\mathcal{K}^a(t, W) := \mathcal{K}(t) + \left\{ \xi \left(-W + \sum_{u=t+1}^{T} D_u^* \right) : \xi \in L_+^\infty(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R}) \right\},$$

$$\mathcal{K}^b(t, W) := \mathcal{K}(t) + \left\{ \xi \left(W - \sum_{u=t+1}^{T} D_u^* \right) : \xi \in L_+^\infty(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R}) \right\}.$$

REMARK 5.4.

(i) The sets $\mathcal{K}^a(t, W)$ and $\mathcal{K}^b(t, W)$ are closed with respect to multiplication by random variables in the set $L^\infty_+(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$, and in view of Lemma 2.14 they are convex cones. Also, $\mathcal{K}(t) \subset \mathcal{K}^a(t, W) \cap \mathcal{K}^b(t, W)$ since $0 \in L^\infty_+(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$.

(ii) In view of Proposition 2.9,

$$\left\{ \xi \left(-\pi_t^{\text{ask}}(D) + \sum_{u=t+1}^{T} D_u^* \right) : \ \xi \in L^\infty_+(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R}) \right\} \tag{5.1}$$

is the set of all discounted terminal claims associated with zero-cost, self-financing, buy-and-hold trading strategies in the contingent claim D with discounted ex-dividend ask price $\pi_t^{\text{ask}}(D)$. On the other hand, the convex cone

$$\left\{ \xi \left(\pi_t^{\text{bid}}(D) - \sum_{u=t+1}^{T} D_u^* \right) : \ \xi \in L^\infty_+(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R}) \right\} \tag{5.2}$$

is the set of all discounted terminal claims associated with zero-cost, self-financing, sell-and-hold trading strategies in the contingent claim D with discounted ex-dividend bid price $\pi_t^{\text{bid}}(D)$.

We will now introduce definitions related to the sets of extended attainable claims. For each $t \in \{0, 1, \ldots, T-1\}$ and $X \in L^0(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$, a probability measure $\mathbb{Q}$ is *risk-neutral for* $\mathcal{K}^a(t, X)$ ($\mathcal{K}^b(t, X)$) if $\mathbb{Q} \in \mathcal{Z}$ and X is $\mathbb{Q}$-integrable, and if $\mathbb{E}_{\mathbb{Q}}[K] \leq 0$ for all $K \in \mathcal{K}^a(t, X)$ ($K \in \mathcal{K}^b(t, X)$). We denote by $\mathcal{R}^a(t, X)$ ($\mathcal{R}^b(t, X)$) the set of all risk-neutral measures $\mathbb{Q}$ for $\mathcal{K}^a(t, X)$ ($\mathcal{K}^b(t, X)$) so that $d\mathbb{Q}/d\mathbb{P} \in L^\infty(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$. We say that **NA** holds for $\mathcal{K}^a(t, X)$ if $\mathcal{K}^a(t, X) \cap L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R}) = \{0\}$, and likewise we say that **NA** holds for $\mathcal{K}^b(t, X)$ if $\mathcal{K}^b(t, X) \cap L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R}) = \{0\}$.

We will say that $\mathcal{K}^a(t, X)$ satisfies **EF** if

$$\left\{ (\phi, \xi) \in \mathcal{S}(t) \times L^\infty_+(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R}) : \ V_0(\phi) = 0, \ V_T^*(\phi) + \xi \left(-X + \sum_{u=t+1}^{T} D_u^* \right) = 0 \right\}$$
$$= \{(0, 0)\},$$

and say that $\mathcal{K}^b(t, X)$ satisfies **EF** if

$$\left\{ (\phi, \xi) \in \mathcal{S}(t) \times L^\infty_+(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R}) : \ V_0(\phi) = 0, \ V_T^*(\phi) + \xi \left(X - \sum_{u=t+1}^{T} D_u^* \right) = 0 \right\}$$
$$= \{(0, 0)\}.$$

REMARK 5.5. According to Lemma A.1, for any $t \in \{0, 1, \ldots, T-1\}$ and $X \in L^0(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{P})$, **NAEF** holds for $\mathcal{K}^a(t, X)$ ($\mathcal{K}^b(t, X)$) if and only if $\mathcal{R}^a(t, X) \neq \emptyset$ ($\mathcal{R}^b(t, X) \neq \emptyset$).

For each $t \in \{0, 1, \ldots, T-1\}$, we denote by $\mathcal{R}(t)$ the set of all risk-neutral measures for $\mathcal{K}(t)$ so that $d\mathbb{Q}/d\mathbb{P} \in L^\infty(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$. Specifically, we define $\mathcal{R}(t)$ as

$$\mathcal{R}(t) := \{\mathbb{Q} \in \mathcal{Z} : \ \mathbb{E}_{\mathbb{Q}}[K] \leq 0 \ \text{ for all } K \in \mathcal{K}(t)\}.$$

We note that $\mathcal{R}^a(t, X) \cup \mathcal{R}^b(t, X) \subseteq \mathcal{R}(t)$ for any $X \in L^0(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$ since $\mathcal{K}(t) \subseteq \mathcal{K}^a(t, X) \cap \mathcal{K}^b(t, X)$. Also, by the definition of a risk-neutral measure, it immediately follows that any risk-neutral measure $\mathbb{Q}$ (as in Definition 3.9) satisfies $\mathbb{Q} \in \mathcal{R}(t)$ for any $t \in \{0, 1, \ldots, T-1\}$.

The next technical lemma is needed to derive the dual representations of the superhedging ask and subhedging bid prices. The proof is straightforward and therefore skipped.

LEMMA 5.6.

(i) For each $t \in \{0, 1, \ldots, T-1\}$, if $\mathcal{R}(t) \neq \emptyset$ and $\mathbb{Q} \in \mathcal{R}(t)$, then we have that $\mathbb{E}_{\mathbb{Q}}[K|\mathcal{F}_t] \leq 0$ $\mathbb{Q}$-a.s. for all $K \in \mathcal{K}(t)$.

(ii) For each $t \in \{0, 1, \ldots, T-1\}$ and $X \in L^0(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$, if $\mathcal{R}^a(t, X) \neq \emptyset$ and $\mathbb{Q} \in \mathcal{R}^a(t, X)$, then we have that $\mathbb{E}_{\mathbb{Q}}[K^a|\mathcal{F}_t] \leq 0$ $\mathbb{Q}$-a.s. for all $K^a \in \mathcal{K}^a(t, X)$.

(iii) For each $t \in \{0, 1, \ldots, T-1\}$ and $X \in L^0(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$, if $\mathcal{R}^b(t, X) \neq \emptyset$ and $\mathbb{Q} \in \mathcal{R}^b(t, X)$, then we have that $\mathbb{E}_{\mathbb{Q}}[K^b|\mathcal{F}_t] \leq 0$ $\mathbb{Q}$-a.s. for all $K^b \in \mathcal{K}^b(t, X)$.

We are ready to prove the main result of this section: the dual representations of the superhedging ask price and subhedging bid price.

THEOREM 5.7. *Suppose that the no-arbitrage condition under the efficient friction assumption **(NAEF)** is satisfied. Let $t \in \{0, 1, \ldots, T-1\}$ and D be a contingent claim. Then the following hold:*

(i) The essential infimum of $\mathcal{W}^a(t, D)$ and the essential supremum of $\mathcal{W}^b(t, D)$ are attained.

*(ii) Suppose that for each $t \in \{0, 1, \ldots, T-1\}$ and $X \in L^0(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$, the efficient friction assumption **(EF)** holds for $\mathcal{K}^a(t, X)$ and $\mathcal{K}^b(t, X)$. Then the discounted superhedging ask and subhedging bid prices for contingent claim D at time t satisfy*

$$\pi_t^{\text{ask}}(D) = \operatorname*{ess\,sup}_{\mathbb{Q} \in \mathcal{R}(t)} \mathbb{E}_{\mathbb{Q}}\left[\sum_{u=t+1}^{T} D_u^* \middle| \mathcal{F}_t\right], \tag{5.3}$$

$$\pi_t^{\text{bid}}(D) = \operatorname*{ess\,inf}_{\mathbb{Q} \in \mathcal{R}(t)} \mathbb{E}_{\mathbb{Q}}\left[\sum_{u=t+1}^{T} D_u^* \middle| \mathcal{F}_t\right]. \tag{5.4}$$

Proof. Since $\pi_t^{\text{ask}}(D) = -\pi_t^{\text{bid}}(-D)$ holds for all $t \in \{0, \ldots, T-1\}$ and contingent claims D, it suffices to show that the essential infimum of $\mathcal{W}^a(t, D)$ is attained and (5.3) holds. Let us fix $t \in \{0, 1, \ldots, T-1\}$ throughout the proof.

We first prove (i). Let W^m be a sequence decreasing a.s. to $\pi_t^{\text{ask}}(D)$, and fix $K^m \in \mathcal{K}(t)$ and $Z^m \in L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$ so that $-W^m + \sum_{u=t+1}^{T} D_u^* = K^m - Z^m$. Since a.s. converges implies convergence in probability, we see that the sequence $K^m - Z^m$ converges in probability to some Y. Due to Theorem 3.8, we have that $\mathcal{K}(t) - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$ is $\mathbb{P}$-closed. Therefore, $Y \in \mathcal{K}(t) - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$. This proves that $-\pi_t^{\text{ask}}(D) + \sum_{u=t+1}^{T} D_u^* \in \mathcal{K}(t) - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$.

Next, we show that (ii) holds. We begin by showing that

$$\pi_t^{\text{ask}}(D) \geq \operatorname*{ess\,sup}_{\mathbb{Q} \in \mathcal{R}(t)} \mathbb{E}_{\mathbb{Q}}\left[\sum_{u=t+1}^{T} D_u^* \middle| \mathcal{F}_t\right]. \tag{5.5}$$

By (i), we have that $\pi_t^{\text{ask}}(D) \in \mathcal{W}^a(t, D)$, so there exists $K^* \in \mathcal{K}(t)$ so that

$$K^* + \pi_t^{\text{ask}}(D) - \sum_{u=t+1}^{T} D_u^* \geq 0. \tag{5.6}$$

We are assuming that **NAEF** is satisfied, so according to Theorem 3.13 there exists a risk-neutral measure $\mathbb{Q}^*$. Because any risk-neutral measure $\mathbb{Q}$ satisfies $\mathbb{Q} \in \mathcal{R}(t)$, we obtain that $\mathbb{Q}^* \in \mathcal{R}(t)$. By taking the conditional expectation with respect to $\mathcal{F}_t$ under $\mathbb{Q}^*$ of both sides of the last inequality we deduce that

$$\pi_t^{\text{ask}}(D) + \mathbb{E}_{\mathbb{Q}^*}[K^*|\mathcal{F}_t] \geq \mathbb{E}_{\mathbb{Q}^*}\left[\sum_{u=t+1}^{T} D_u^* \middle| \mathcal{F}_t\right].$$

According to part (i) of Lemma 5.6, we have that $\mathbb{E}_{\mathbb{Q}^*}[K^*|\mathcal{F}_t] \leq 0$. As a result, $\pi_t^{\text{ask}}(D) \geq \mathbb{E}_{\mathbb{Q}^*}\big[\sum_{u=t+1}^{T} D_u^* \big| \mathcal{F}_t\big]$. Taking the essential supremum of both sides of the last inequality over $\mathcal{R}(t)$ proves that (5.5) holds.

Next, we show that

$$\pi_t^{\text{ask}}(D) \leq \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{R}(t)} \mathbb{E}_{\mathbb{Q}}\left[\sum_{u=t+1}^{T} D_u^* \middle| \mathcal{F}_t\right]. \tag{5.7}$$

By (i), we have that $\pi_t^{\text{ask}}(D) > -\infty$, so we may take $X \in L^0(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$ so that $\pi_t^{\text{ask}}(D) > X$. We now prove by contradiction that **NA** holds for $\mathcal{K}^b(t, X)$. Toward this aim, we assume that there exist $K \in \mathcal{K}(t)$, $\xi \in L_+^\infty(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$, and $\Omega^0 \subseteq \Omega$ with $\mathbb{P}(\Omega^0) > 0$ so that

$$K + \xi\left(X - \sum_{u=t+1}^{T} D_u^*\right) \geq 0 \quad \text{a.s.}, \quad K + \xi\left(X - \sum_{u=t+1}^{T} D_u^*\right) > 0 \quad \text{a.s. on } \Omega^0. \tag{5.8}$$

Since **NA** is satisfied for underlying market $\mathcal{K}$, we have from (5.8) that there exists $\Omega^1 \subseteq \Omega^0$ with $\mathbb{P}(\Omega^1) > 0$ such that $\Omega^1 \in \mathcal{F}_t$ and $\xi > 0$ a.s. on Ω^1. Otherwise, our assumption that **NA** holds is contradicted. Of course, if $\Omega^1 \subseteq \Omega^0$ is any set such that $\Omega^1 \in \mathcal{F}_t$, $\mathbb{P}(\Omega^1) > 0$, and $\xi = 0$ a.s. on Ω^1, then $1_{\Omega^1} K \in \mathcal{K}(t) \in \mathcal{K}$ satisfies $1_{\Omega^1} K \geq 0$ a.s., and $1_{\Omega^1} K > 0$ a.s. on Ω^1, which violates that **NA** is satisfied.

Moreover, we observe that $X - \sum_{u=t+1}^{T} D_u^* \geq 0$ a.s. on Ω^1. If there exists $\Omega^2 \subseteq \Omega^1$ with $\mathbb{P}(\Omega^2) > 0$ such that $\Omega^2 \in \mathcal{F}_t$ and $X - \sum_{u=t+1}^{T} D_u^* < 0$ a.s. on Ω^2, then from (5.8) we see that $K \geq$ a.s., and $K > 0$ a.s. on Ω^2, which contradicts that **NA** holds for $\mathcal{K}$.

Now, let us define

$$\widetilde{X} := 1_{\Omega^1} X + 1_{(\Omega^1)^c} \pi_t^{\text{ask}}(D), \quad \widetilde{K} := 1_{\Omega^1} \frac{K}{\sup_{\omega\in\Omega^1}\{\xi(\omega)\}} + 1_{(\Omega^1)^c} K^*.$$

From (5.6), we immediately have that

$$\widetilde{K} + \widetilde{X} - \sum_{u=t+1}^{T} D_u^* = K^* + \pi_t^{\text{ask}}(D) - \sum_{u=t+1}^{T} D_u^* \geq 0 \quad \text{a.s. on } (\Omega^1)^c.$$

On the other hand, from (5.8) and since $X - \sum_{u=t+1}^{T} D_u^* \geq 0$ a.s. on Ω^1, we see that

$$\widetilde{K} + \widetilde{X} - \sum_{u=t+1}^{T} D_u^* = \frac{K}{\sup_{\omega\in\Omega^1}\{\xi(\omega)\}} + X - \sum_{u=t+1}^{T} D_u^* \geq 0 \quad \text{a.s. on } \Omega^1.$$

Consequently, $\widetilde{K} + \widetilde{X} - \sum_{u=t+1}^{T} D_u^* \geq 0$ a.s. on Ω. Now, since $0 \leq 1/\sup_{\omega\in\Omega^1}\{\xi(\omega)\} < \infty$, and because $\mathcal{K}(t)$ is a convex cone that is closed with respect to multiplication by random variables in $L_+^\infty(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$, we have that $\widetilde{K} \in \mathcal{K}(t)$. Therefore, $\widetilde{X} \in \mathcal{W}^a(t, D)$. However, since $\widetilde{X}$ satisfies $\widetilde{X} \leq \pi_t^{\text{ask}}(D)$ and $\mathbb{P}(\widetilde{X} < \pi_t^{\text{ask}}(D)) > 0$, we have that $\widetilde{X} \in \mathcal{W}^a(t, D)$ contradicts $\pi_t^{\text{ask}}(D) = \operatorname{ess\,inf} \mathcal{W}^a(t, D)$. Thus, **NA** holds for $\mathcal{K}^b(t, X)$.

By assumption, **EF** holds for $\mathcal{K}^b(t, X)$, so **NAEF** is satisfied for $\mathcal{K}^b(t, X)$. According to Lemma A.1 there exists $\hat{\mathbb{Q}} \in \mathcal{R}^b(t, X)$. In view of the claim (iii) in Lemma 5.6, we see that

$$\zeta X + \mathbb{E}_{\hat{\mathbb{Q}}}[K|\mathcal{F}_t] \leq \zeta\, \mathbb{E}_{\hat{\mathbb{Q}}}\left[\sum_{u=t+1}^{T} D_u^* \middle| \mathcal{F}_t\right], \quad K \in \mathcal{K}(t),\ \zeta \in L_+^\infty(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R}).$$

Since $0 \in \mathcal{K}(t)$ and $1 \in L_+^\infty(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$, we obtain that $X \leq \mathbb{E}_{\hat{\mathbb{Q}}}[\sum_{u=t+1}^{T} D_u^*|\mathcal{F}_t]$. Now, because $\mathcal{R}^b(t, X) \subseteq \mathcal{R}(t)$, we have that $\hat{\mathbb{Q}} \in \mathcal{R}(t)$. Hence,

$$X \leq \mathbb{E}_{\hat{\mathbb{Q}}}\left[\sum_{u=t+1}^{T} D_u^* \middle| \mathcal{F}_t\right] \leq \sup_{\mathbb{Q}\in\mathcal{R}(t)} \mathbb{E}_{\mathbb{Q}}\left[\sum_{u=t+1}^{T} D_u^* \middle| \mathcal{F}_t\right]. \tag{5.9}$$

The random variable $X < \pi_t^{\text{ask}}(D)$ is arbitrary, so for any scalar $\epsilon > 0$ we may take $X := \pi_t^{\text{ask}}(D) - \epsilon$. From (5.9), we see that

$$\pi_t^{\text{ask}}(D) \leq \sup_{\mathbb{Q}\in\mathcal{R}(t)} : \mathbb{E}_{\mathbb{Q}}\left[\sum_{u=t+1}^{T} D_u^* \middle| \mathcal{F}_t\right] + \epsilon, \quad \epsilon > 0.$$

Letting ϵ approach zero shows that (5.7) holds. This completes the proof of (i). □

REMARK 5.8. An open question that remains is whether $\mathcal{R}(t)$ can be replaced by $\mathcal{R}(0)$ in the representations in Theorem 5.7. In the arguments presented in this paper, it is more convenient to work with $\mathcal{R}(t)$ than $\mathcal{R}(0)$ because $\mathcal{K}(t)$ is closed under multiplication by random variables in $L_+^\infty(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$. In contrast, the set $\mathcal{K}(0)$ is only closed under multiplication by random variables in $L_+^\infty(\Omega, \mathcal{F}_0, \mathbb{P}; \mathbb{R})$.

6. CONCLUSIONS

In this paper, no-arbitrage pricing theory is extended to dividend-paying securities in discrete-time markets with transaction costs. A version of the Fundamental Theorem of Asset Pricing is proved under the efficient friction assumption, and the representations for the superhedging ask prices and the subhedging bid prices are given. As usual, the proof of the Fundamental Theorem of Asset Pricing relies on showing that the set of all claims that can be superhedged at zero cost is closed (under convergence in probability). In the special case when there are no transaction costs on the dividends paid by the security, the no-arbitrage condition under the efficient friction assumption is proved to be equivalent to the existence of a CPS. The general case, in which there are transaction costs on the dividends, is open.

The study conducted in this paper has been motivated by questions of valuation of credit default swaps and interest rate swaps subject to bid–ask transaction costs. The extensions to other types costs, such as funding costs, is important and should be studied in the future.

APPENDIX

LEMMA A.1. *For each $t \in \{0, 1, \dots, T-1\}$ and $W \in L^0(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$, if the no-arbitrage condition under the efficient friction assumption is satisfied for $\mathcal{K}^a(t, W)$ ($\mathcal{K}^b(t, W)$), then $\mathcal{R}^a(t, W) \neq \emptyset$ ($\mathcal{R}^b(t, W) \neq \emptyset$).*

Proof. Let us first fix $t \in \{0, 1, \dots, T-1\}$ and $W \in L^0(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$. We only prove the lemma for $\mathcal{K}^a(t, W)$, because the proof for $\mathcal{K}^b(t, W)$ is similar. Instead of working with $\mathcal{K}^a(t, X)$, we will work with the more mathematically convenient set

$$\mathbb{K}^a(t, W) := \big\{G(\phi, \xi, t, W) : \phi \in \mathcal{P}(t), \xi \in L^\infty_+(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})\big\},$$

where $\mathcal{P}(t)$ is the set

$$\mathcal{P}(t) := \{\phi \in \mathcal{P} : \phi^j \text{ is a.s. bounded for } j \in \mathcal{J}^*, \phi_s = 1_{\{t+1 \leq s\}}\phi_s \text{ for all } s \in \mathcal{T}^*\},$$

and

(A.1)

$$\begin{aligned} G(\phi, \xi, t, W) := {} & \sum_{j=1}^N \phi_T^j \big(1_{\{\phi_T^j \geq 0\}} P_T^{\text{bid,j,*}} + 1_{\{\phi_T^j < 0\}} P_T^{\text{ask,j,*}}\big) \\ & - \sum_{j=1}^N \sum_{u=t+1}^T \Delta\phi_u^j \big(1_{\{\Delta\phi_u^j \geq 0\}} P_{u-1}^{\text{ask,j,*}} + 1_{\{\Delta\phi_u^j < 0\}} P_{u-1}^{\text{bid,j,*}}\big) \\ & + \sum_{j=1}^N \sum_{u=t+1}^T \phi_u^j \big(1_{\{\phi_u^j \geq 0\}} A_u^{\text{ask,j,*}} + 1_{\{\phi_u^j < 0\}} A_u^{\text{bid,j,*}}\big) + \xi \left(-W + \sum_{u=t+1}^T D_u^*\right) \end{aligned}$$

for all $\mathbb{R}^N$-valued stochastic processes

$$(\phi_s)_{s=1}^T \in L^0(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R}^N) \times \cdots \times L^0(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R}^N),$$

and random variables $\xi \in L^\infty_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$.

Since $\mathbb{K}^a(t, W) = \mathcal{K}^a(t, W)$, we may equivalently prove that $\mathbb{K}^a(t, W) - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$ is $\mathbb{P}$-closed whenever **NAEF** is satisfied for $\mathcal{K}^a(t, W)$. Let $X^m \in \mathbb{K}^a(t, W) - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$ be a sequence converging in probability to some X. We may find a subsequence X^{k_m} that converges a.s. to X. With an abuse of notation we denote this subsequence by X^m. By the definition of $\mathbb{K}^a(t, W)$, we may find $\phi^m \in \mathcal{P}(t)$, $\xi^m \in L^0_+(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$, and $Z^m \in L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$ so that $X^m = G(\phi^m, \xi^m, t, W) - Z^m$. Using the same arguments as in Step 1 in the proof of Theorem 3.8, we prove that $\limsup_m \|\phi_s^m\| < \infty$ for all $t \in \mathcal{T}^*$ and $\limsup_m \xi^m < \infty$. Then, we apply Lemma 3.6 to show that we may find a strictly increasing set of positive, integer-valued, $\mathcal{F}_{T-1}$ measurable random variables σ^m such that ϕ^{σ^m} converges a.s. to some bounded a.s.

predictable process ϕ, and ξ^{σ^m} converges a.s. to some $\xi \in L^\infty_+(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbb{R})$. This gives us that $G(\phi^{\sigma^m}, \xi^{\sigma^m}, t, W) - X^{\sigma^m}$ converges a.s. to some nonnegative random variable. Therefore, $\mathbb{K}^a(t, W) - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$ is $\mathbb{P}$-closed.

We now argue that there exists a risk-neutral measure for $\mathcal{K}^a(t, W)$. Toward this, we define the convex cone $\mathcal{C}^a := (\mathcal{K}^a(t, W) - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})) \cap L^1(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$. Due to the closedness property of $\mathbb{K}^a(t, W) - L^0_+(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$, we have that the set $\mathcal{C}^a$ is closed in $L^1(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$. As in the proof of Theorem 3.13, we may construct a measure $\mathbb{Q} \in \mathcal{Z}$ such that W is $\mathbb{Q}$-integrable, and $\mathbb{E}_{\mathbb{Q}}[K^a] \leq 0$ for all $K^a \in \mathcal{K}^a(t, X)$. This completes the proof. □

REFERENCES

BOUCHARD, B. (2006): No-Arbitrage in Discrete-Time Markets with Proportional Transaction Costs and General Information Structure, *Finance Stoch.* 10, 276–297.

BOUCHARD, B., and N. TOUZI (2000): Explicit Solution to the Multivariate Super-Replication Problem under Transaction Costs, *Ann. Appl. Probab.* 10(3), 685–708.

CAMPI, L., and W. SCHACHERMAYER (2006): A Super-Replication Theorem in Kabanov's Model of Transaction Costs, *Finance Stoch.* 10, 579–596.

CHERNY, A. (2007): General Arbitrage Pricing Model. II. Transaction Costs, in *Séminaire de Probabilités XL*, Vol. 1899 of Lecture Notes in Mathematics, Berlin: Springer, pp. 447–461.

CVITANIC, J., H. PHAM, and N. TOUZI (1999): A Closed-Form Solution to the Problem of Super-Replication under Transaction Costs, *Finance Stoch.* 3, 35–54.

DALANG, R. C., A. MORTON, and W. WILLINGER (1990): Equivalent Martingale Measures and No-Arbitrage in Stochastic Securities Market Models, *Stoch. Stoch. Rep.* 29(2), 185–201.

DELBAEN, F., and W. SCHACHERMAYER (1994): A General Version of the Fundamental Theorem of Asset Pricing, *Math. Ann.* 300, 463–520.

DELBAEN, F., and W. SCHACHERMAYER (2006): *The Mathematics of Arbitrage*, Berlin: Springer-Verlag.

DENIS, E., and Y. M. KABANOV (2012): Consistent Price Systems and Arbitrage Opportunities of the Second Kind in Models with Transaction Costs, *Finance Stoch.* 16, 135–154.

DE VALLIRE, D., Y. M. KABANOV, and C. STRICKER (2007): No-Arbitrage Criteria for Financial Markets with Transaction Costs and Incomplete Information, *Finance Stoch.* 11, 237–251.

FÖLLMER, H., and A. SCHIED (2004): *Stochastic Finance: An Introduction in Discrete Time*, Berlin: Walter de Gruyter.

GUASONI, P., E. LÉPINETTE, and M. RÁSONYI (2012): The Fundamental Theorem of Asset Pricing under Transaction Costs, *Finance Stoch.* 16(4), 741–777.

GUASONI, P., M. RÁSONYI, and W. SCHACHERMAYER (2010): The Fundamental Theorem of Asset Pricing for Continuous Processes under Small Transaction Costs, *Ann. Finance* 6, 157–191.

HARRISON, J. M., and S. R. PLISKA (1981): Martingales and Stochastic Integrals in the Theory of Continuous Trading, *Stoch. Process. Appl.* 11(3), 215–260.

JACKA, S., A. BERKAOUI, and J. WARREN (2008): No Arbitrage and Closure Results for Trading Cones with Transaction Costs, *Finance Stoch.* 12, 583–600.

JACOD, J., and A. N. SHIRYAEV (1998): Local Martingales and the Fundamental Asset Pricing Theorems in the Discrete-Time Case, *Finance Stoch.* 2(3), 259–273.

JOUINI, E., and H. KALLAL (1995): Martingales and Arbitrage in Securities Markets with Transaction Costs, *J. Econ. Theory*, 66(1), 178–197.

KABANOV, Y. M., and D. O. KRAMKOV (1994): No-Arbitrage and Equivalent Martingale Measures: An Elementary Proof of the Harrison–Pliska Theorem, *Theory Probab. Appl.* 39(3), 523–527.

KABANOV, Y. M., M. RÁSONYI, and C. STRICKER (2002): No-Arbitrage Criteria for Financial Markets with Efficient Friction, *Finance Stoch.* 6(3), 371–382.

KABANOV, Y. M., and M. SAFARIAN (2009): *Markets with Transaction Costs: Mathematical Theory*, Berlin: Springer-Verlag.

KABANOV, Y. M., and C. STRICKER (2001a): The Harrison–Pliska Arbitrage Pricing Theorem under Transaction Costs, *J. Math. Econ.* 35(2), 185–196.

KABANOV, Y. M., and C. STRICKER (2001b): A Teachers' Note on No-Arbitrage Criteria, in *Séminaire de Probabilités XXXV*, Vol. 1755 of Lecture Notes in Mathematics, Berlin: Springer, pp. 149–152.

KREPS, D. M. (1981): Arbitrage and Equilibrium in Economies with Infinitely Many Commodities, *J. Math. Econ.* 8(1), 15–35.

LEVENTAL, S., and A. V. SKOROHOD (1997): On the Possibility of Hedging Options in the Presence of Transaction Costs, *Ann. Appl. Probab.* 7(2), 410–443.

PENNANEN, T. (2011a): Arbitrage and Deflators in Illiquid Markets, *Finance Stoch.* 15, 57–83.

PENNANEN, T. (2011b): Convex Duality in Stochastic Optimization and Mathematical Finance, *Math. Oper. Res.* 36, 340–362.

PENNANEN, T. (2011c): Dual Representation of Superhedging Costs in Illiquid Markets, *Math. Financ. Econ.* 5, 233–248.

PENNANEN, T. (2011d): Superhedging in Illiquid Markets, *Math. Finance*, 21, 519–540.

ROGERS, L. C. G. (1994): Equivalent Martingale Measures and No-Arbitrage, *Stoch. Stoch. Rep.* 51(1–2), 41–49.

SCHACHERMAYER, W. (1992): A Hilbert Space Proof of the Fundamental Theorem of Asset Pricing in Finite Discrete Time, *Insurance: Math. Econ.* 11(4), 249–257.

SCHACHERMAYER, W. (2004): The Fundamental Theorem of Asset Pricing under Proportional Transaction Costs in Finite Discrete Time, *Math. Finance* 14(1), 19–48.

SONER, H. M., S. E. SHREVE, and J. CVITANIĆ (1995): There Is No Nontrivial Hedging Portfolio for Option Pricing with Transaction Costs, *Ann. Appl. Probab.* 5(2), 327–355.

TOUZI, N. (1999): Super-Replication under Proportional Transaction Costs: From Discrete to Continuous-Time Models, *Math. Methods Oper. Res.* 50(2), 297–320.

YAN, J. A. (1980): Caractérisation d'une Classe d'Ensembles Convexes de L^1 ou H^1, in *Seminar on Probability, XIV (Paris, 1978/1979) (French)*, Volume 784 of Lecture Notes in Mathematics, Berlin: Springer, pp. 220–222.

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

Mathematical Finance, Vol. 25, No. 4 (October 2015), 702–723

OPTION PRICING AND HEDGING WITH SMALL TRANSACTION COSTS

JAN KALLSEN

Christian-Albrechts-Universität zu Kiel

JOHANNES MUHLE-KARBE

ETH Zürich and Swiss Finance Institute

An investor with constant absolute risk aversion trades a risky asset with general Itô-dynamics, in the presence of small proportional transaction costs. In this setting, we formally derive a leading-order optimal trading policy and the associated welfare, expressed in terms of the local dynamics of the frictionless optimizer. By applying these results in the presence of a random endowment, we obtain asymptotic formulas for utility indifference prices and hedging strategies in the presence of small transaction costs.

KEY WORDS: transaction costs, indifference pricing and hedging, exponential utility, asymptotics.

1. INTRODUCTION

The pricing and hedging of derivative securities is a central theme of mathematical finance. In complete markets, the risk incurred by selling any claim can be offset completely by dynamic trading in the underlying. Then, there is only one price compatible with the absence of arbitrage, namely, the initial value of the replicating portfolio. This line of reasoning is torn to pieces by the presence of even the small bid-ask spreads present in the most liquid financial markets. Transaction costs make (super-)replication prohibitively expensive (Soner et al. 1995), thereby calling for approaches that explicitly balance the gains and costs of trading.

An economically appealing choice is the utility-indifference approach put forward by Hodges and Neuberger (1989) as well as Davis, Panas, and Zariphopoulou (1993).[1] For an investor with given preference structure, the idea is to determine a "fair" price by matching the maximal expected utilities that can be attained with and without the claim. Both Hodges and Neuberger (1989) and Davis et al. (1993) focus on investors with constant absolute risk aversion for tractability. Nevertheless, the numerical computation of the solution turns out to be quite challenging, involving multidimensional nonlinear

Johannes Muhle-Karbe is partially supported by the National Centre of Competence in Research Financial Valuation and Risk Management (NCCR FINRISK), Project D1 (Mathematical Methods in Financial Risk Management), of the Swiss National Science Foundation (SNF). The authors are grateful to Aleš Černý, Christoph Czichowsky, Paolo Guasoni, Marcel Nutz, and Mete Soner for fruitful discussions. They also thank Ren Liu for his careful reading of the manuscript, and two anonymous referees for their pertinent remarks. Part of this work was completed while the second author was visiting Columbia University. He thanks Ioannis Karatzas and the university for their hospitality.

Manuscript received September 2012; final revision received December 2012.

Address correspondence to Johannes Muhle-Karbe, ETH Zürich-Mathematics, Rämistrasse 101 Zürich CH-8092, Switzerland, e-mail: Johannes.muhle-karbe@math.ethz.ch.

[1] Cf. Leland (1985) for an alternative approach, where trading only takes place at exogenous discrete times.

DOI: 10.1111/mafi.12035

free boundary problems already for plain vanilla call options written on a single risky asset with constant investment opportunities.

In reality, transaction costs are *small*, having further declined substantially since stock market decimalization in 2001.[2] Therefore, asymptotic expansions for small spreads have been proposed to "reveal the salient features of the problem while remaining a good approximation to the full but more complicated model" (Whalley and Wilmott 1997). For small costs, a formal asymptotic analysis of the model of Davis et al. (1993) has been carried out by Whalley and Wilmott (1997).[3] A different limiting regime, where absolute risk version becomes large as the spread tends to zero, is studied by Barles and Soner (1998). In both cases, the computation of indifference prices boils down to the solution of certain inhomogeneous Black–Scholes equations, whereas the corresponding hedging strategies are determined explicitly at the leading order.

For small costs, the present study provides formal asymptotics for essentially general continuous asset price dynamics and arbitrary contingent claims. As in the extant literature, we also focus on investors with constant absolute risk aversion, for which the cash additivity of the corresponding exponential utility functions allows to handle the option position by a change of measure.[4] Both with and without an option position, the leading-order optimal trading strategies consist of keeping the number of risky shares in a time and state-dependent no-trade region around their frictionless counterparts. The width of the latter is determined by the following trade-off: large fluctuations of the frictionless optimizer call for a wide buffer in order to reduce trading costs. Conversely, wildly fluctuating asset prices cause the investor's positions to deviate substantially from the frictionless target, thereby necessitating closer tracking. Accordingly, the ratio of local fluctuations—measured in terms of the local quadratic variation both for the frictionless optimizer and the risky asset—is the crucial statistic determining the optimal trading boundaries in the presence of small transaction costs. The corresponding welfare loss—and, in turn, the indifference price adjustments—turn out to be given by the squared width of the no-trade region, suitably averaged with respect to both time and states.

The pricing implications of small transaction costs depend on the interplay between the frictionless pure investment and hedging strategies. If no trading takes place in the absence of an option position, then the costs incurred by hedging the claims necessitate a higher premium. This changes, however, if trades prescribed by the hedge partially offset the rebalancing of the investor's pure investment position. Then, maybe surprisingly, a smaller compensation may, in fact, be sufficient for the risk incurred by selling the claims if transaction costs are taken into consideration.

The remainder of the paper is organized as follows. The main results—explicit formulas for the leading-order optimal policy and welfare—are presented in Section 2. Subsequently, we discuss how they can be adapted to deal with utility-indifference pricing and hedging. The derivations of all results are collected in Appendix A. They are based on formal perturbation arguments, applied to the martingale optimality conditions for a frictionless "shadow price," which admits the same leading-order optimal strategy

[2]The CRSP database reports drops by almost one order of magnitude from 2001 to 2010.

[3]See Bichuch (2011) for a rigorous proof.

[4]Similar but more involved asymptotics for the optimal policy and the associated welfare can also be obtained for general utilities and with intermediate consumption, see Soner and Touzi (2012) as well as the forthcoming companion paper of the present study (Kallsen and Muhle-Karbe 2013). However, it is then no longer possible to deal with option positions by a change of measure as in Section 3.

and utility as the original market with transaction costs. A rigorous verification theorem is a major challenge for future research.

2. OPTIMAL INVESTMENT

Consider a market with two assets, a riskless one with price normalized to one and a risky one trading with proportional transaction costs. This means that one has to pay a higher ask price $(1+\varepsilon)S_t$ when purchasing the risky asset but only receives a lower bid price $(1-\varepsilon)S_t$ when selling it. Here, $\varepsilon > 0$ is the relative bid-ask spread and the mid price S_t follows a general, not necessarily Markovian, Itô process:

$$dS_t = b_t^S dt + \sqrt{c_t^S} dW_t, \tag{2.1}$$

for a standard Brownian motion W_t. In this setting, an investor with exponential utility function $U(x) = -e^{-px}$, i.e., with constant absolute risk aversion $p > 0$, trades to maximize the certainty equivalent $-\frac{1}{p}\log E[e^{-pX_T^\phi}]$ over all terminal wealths X_T^ϕ at time T corresponding to self-financing trading strategies ϕ_t.[5] The optimal number of shares in the absence of frictions ($\varepsilon = 0$) is denoted by φ_t; we assume it to be a sufficiently regular Itô process with local quadratic variation $d\langle\varphi\rangle_t/dt$, which is satisfied in most applications.

The dynamics (2.1) are formulated in *discounted* terms. If the safe asset earns a constant interest rate $r > 0$, one can reduce to this case by discounting, replacing risk aversion p by $e^{rT}p$.

2.1. Optimal Policy

For small transaction costs ε, an approximately optimal[6] strategy φ_t^ε is to engage in the minimal amount of trading necessary to keep the number of risky shares within the following buy and sell boundaries around the frictionless optimizer φ_t:

$$\Delta\varphi_t^\pm = \pm\left(\frac{3}{2p}\frac{d\langle\varphi\rangle_t}{d\langle S\rangle_t}\varepsilon S_t\right)^{1/3}. \tag{2.2}$$

The random and time varying no-trade region $[\varphi_t + \Delta\varphi_t^-, \varphi_t + \Delta\varphi_t^+]$ is symmetric around the frictionless optimizer φ_t, and its half-width is given by the cubic root of three parts:

(i) the constant $3/2p$ that only depends on the investor's risk aversion but not on the underlying probabilistic model;
(ii) the observable (absolute) half-width εS_t of the bid-ask spread;
(iii) the fluctuations of the frictionless optimizer, measured in terms of its local quadratic variation $d\langle\varphi\rangle_t$, normalized by the market's local fluctuations $d\langle S\rangle_t$. Tracking more wildly fluctuating strategies requires a wider buffer to reduce trading costs. Conversely, large fluctuations in the asset prices cause large fluctuations of the investor's risky position, thereby necessitating closer tracking to reduce losses due to displacement from the frictionless target position.

[5]The costs of setting up and liquidating the portfolio are only incurred once and are therefore of order $O(\varepsilon)$. Hence, they do not impact the investor's welfare at the leading order $\varepsilon^{2/3}$, and we disregard them throughout.

[6]That is, this strategy matches the optimal certainty equivalent, at the leading order for small costs. See Appendix A.2 for more details.

Unless the planning horizon is postponed to infinity (Constantinides 1986; Davis and Norman 1990; Shreve and Soner 1994), transaction costs introduce horizon effects even with a constant investment opportunity set (Liu and Loewenstein 2002). Nevertheless, the *local* dynamics of the frictionless optimizer φ_t alone always act as a sufficient statistic for the asymptotically optimal trading boundaries (2.2)—the investor does not hedge against the presence of a small constant friction. These optimal trading boundaries are "myopic" in the sense that they are of the same form as for the *local* utility maximizers considered by Martin (2011). By definition, these also behave myopically in the absence of frictions, unlike the exponential investors considered here, whose optimal policies generally include an intertemporal hedging term reflecting future investment opportunities. Somewhat surprisingly, the generally different frictionless strategies enter the leading-order trading boundaries in the same way through their local fluctuations.

More general preferences and intermediate consumption are studied by Soner and Touzi (2012) and in a forthcoming companion paper of the present study (Kallsen and Muhle-Karbe, 2013).

2.2. Welfare

The utility associated with the above policy can also be quantified, thereby allowing to assess the welfare impact of transaction costs. To this end, let CE and CE^ε denote the certainty equivalents without and with transaction costs ε, respectively, i.e., the cash amounts that yield the same utility as trading optimally in the market. Then, for small transaction costs ε:

$$\mathrm{CE}^\varepsilon \sim \mathrm{CE} - \frac{p}{2} E_Q \left[\int_0^T \left(\Delta \varphi_t^+ \right)^2 d\langle S \rangle_t \right], \tag{2.3}$$

and this leading-order optimal performance is achieved by the policy from Section 2.1. As the trading boundaries $\Delta\varphi_t^+$ are proportional to $\varepsilon^{1/3}$ by (2.2), the leading-order loss due to transaction costs is therefore of order $O(\varepsilon^{2/3})$ as in the Black–Scholes model (Shreve and Soner 1994). It is given by an average of the squared half-width of the optimal no-trade region. The latter has to be computed with respect to a clock that runs at the speed $d\langle S \rangle_t = c_t^S dt$ of the market's local variance, i.e., losses due to transaction costs accrue more rapidly in times of frequent price moves. Moreover, this average has to be determined under the marginal pricing measure Q associated with the frictionless utility maximization problem, i.e., the equivalent martingale measure that minimizes the entropy with respect to the physical probability. The leading-order effect of transaction costs can therefore be interpreted as the price of a path-dependent contingent claim, computed under the investor's marginal pricing measure.

As first pointed out by Rogers (2004) (also compare Goodman and Ostrov 2010), the utility loss due to transaction costs is composed of two parts. On the one hand, there is the displacement loss due to following the strategy φ_t^ε instead of the frictionless maximizer φ_t. In addition, there is the loss due to the costs directly incurred by trading. Rogers observed that for small transaction costs, the leading orders of these two losses coincide for the optimal policy. For investors with constant absolute risk aversion, we complement this by the insight that two-thirds of the leading-order welfare loss are incurred due to trading costs, whereas the remaining one-third is caused by displacement. Surprisingly, this holds irrespective of the model for the risky asset and the investor's risk aversion.

3. INDIFFERENCE PRICING AND HEDGING

Due to the cash additivity of the exponential utility function, the above results can be adapted to optimal investment in the presence of a random endowment, thereby leading to asymptotic formulas for utility-based prices and hedging strategies.

Indeed, suppose that at time $t = 0$, the investor sells a claim H maturing at time T, for a premium $\pi(H)$. Then, her investment problem becomes

$$\sup_\phi E\big[-e^{-p(X_T^\phi+\pi(H)-H)}\big] = e^{-p\pi(H)}E[e^{pH}]\sup_\phi E_{P^H}\big[-e^{-pX_T^\phi}\big],$$

where P^H is the equivalent probability with density $dP^H/dP = e^{pH}/E[e^{pH}]$. Hence, we are back in the above setting of pure investment; only the dynamics of the risky asset S_t change when passing from the physical probability P to P^H, and the frictionless optimizer changes accordingly. Then, with small transaction costs ε, the optimal policy in the presence of the random endowment $-H$ corresponds to the minimal amount of trading to keep the number of risky assets within the following buy and sell boundaries around the frictionless optimizer φ_t^H:[7]

$$\Delta\varphi_t^{H,\pm} = \left(\frac{3}{2p}\frac{d\langle\varphi^H\rangle_t}{d\langle S\rangle_t}\varepsilon S_t\right)^{1/3}. \tag{3.1}$$

Hence, at the leading order, the optimal investment strategy in the presence of a random endowment again prescribes the minimal amount of trading to remain within a buffer around its frictionless counterpart, whose width can be calculated from the local variations of the latter. Again, by appealing to the results for the pure investment problem, the corresponding certainty equivalent is found to be

$$\mathrm{CE}^{\varepsilon,H} \sim \mathrm{CE}^H - \frac{p}{2}E_{Q^H}\left[\int_0^T \left(\Delta\varphi_t^{H,+}\right)^2 d\langle S\rangle_t\right], \tag{3.2}$$

where CE^H and Q^H denote the frictionless certainty equivalent and minimal entropy martingale measure in the presence of the claim H, respectively. With this result at hand, the corresponding *utility indifference price* $\pi^\varepsilon(H)$ of Hodges and Neuberger (1989) can be computed by matching (3.2) with the investor's certainty equivalent (2.3) in the absence of the claim. At the leading order, it turns out that the frictionless indifference price $\pi^0(H)$—which makes the frictionless certainty equivalents CE^H, CE with and without the claim coincide—has to be corrected by the difference between the effects of transaction costs with and without the claim:

$$\pi^\varepsilon(H) \sim \pi^0(H) + \frac{p}{2}\left(E_{Q^H}\left[\int_0^T \left(\Delta\varphi_t^{H,+}\right)^2 d\langle S\rangle_t\right] - E_Q\left[\int_0^T \left(\Delta\varphi_t^{+}\right)^2 d\langle S\rangle_t\right]\right). \tag{3.3}$$

As a consequence, the indifference price can be either higher or lower than its frictionless counterpart, depending on whether higher or lower trading costs are incurred due to the presence of the claim.

[7]Note that the square bracket processes of continuous processes are invariant under equivalent measure changes, i.e., it does not matter whether they are computed under P or P^H.

3.1. Complete Markets

Even in the absence of frictions, utility-based prices and hedging strategies are typically hard to compute unless the market is *complete*, and we first focus on this special case in the sequel. Then, there is a *unique* equivalent martingale measure Q, and frictionless indifference prices coincide with expectations under the latter:

$$\pi^0(H) = E_Q[H].$$

Moreover, any claim H can then be hedged perfectly by a replicating strategy Δ_t^H:[8]

$$H = E_Q[H] + \int_0^T \Delta_t^H dS_t.$$

Consequently, the random endowment can simply be removed from the optimal investment problem:

$$\sup_\phi E\big[-e^{-p(x+\int_0^T \phi_t dS_t + \pi(H) - H)}\big] = e^{-p(\pi(H)-E_Q[H])} \sup_\phi E\big[-e^{-p(x+\int_0^T (\phi_t - \Delta_t^H) dS_t)}\big].$$

As a result, the optimal strategy for investors with constant absolute risk aversion is to hedge away the random endowment and, in addition to that, invest as in the pure investment problem:

$$\varphi_t^H = \varphi_t + \Delta_t^H. \tag{3.4}$$

The (monetary) boundaries of the corresponding leading-order optimal no-trade region for small transaction costs ε can, in turn, be determined by inserting (3.4) into (3.1):

$$\Delta\varphi_t^{\pm,H} S_t = \pm \left(\frac{3}{2p} \frac{d\langle \varphi + \Delta^H \rangle_t}{d\langle S \rangle_t} S_t^4 \right)^{1/3} \varepsilon^{1/3}. \tag{3.5}$$

To shed more light on this first-order correction due to the presence of small transaction costs, write[9]

$$d\varphi_t = \Gamma_t^\varphi \, dS_t + a_t^\varphi dt, \quad d\Delta_t^H = \Gamma_t^H dS_t + a_t^H dt. \tag{3.6}$$

The *gammas* Γ_t^φ and Γ_t^H describe the sensitivities (of the diffusive parts) of the strategies φ_t and Δ_t^H with respect to price moves.[10] With this notation,

$$\frac{d\langle \varphi + \Delta^H \rangle_t}{d\langle S \rangle_t} S_t^4 = \big(\Gamma_t^\varphi S_t^2 + \Gamma_t^H S_t^2\big)^2. \tag{3.7}$$

Hence, the width of the no-trade region (3.5) is determined by the *cash-gamma* of the investor's portfolio, that is, the sensitivity of the frictionless optimal risky position to changes in the risky asset. If shocks to the risky asset cause the frictionless position to move a lot, then the investor should keep a wider buffer around it to save transaction

[8] As this notation indicates, this is just the usual *delta hedge* in a Markovian setting, i.e., the derivative of the option price with respect to the underlying.

[9] For $d\varphi_t = b_t^\varphi dt + \sqrt{c_t^\varphi} dW_t$, this is obtained by setting $\Gamma_t^\varphi = \sqrt{c_t^\varphi / c_t^S}$ and $a_t^\varphi = b_t^\varphi - b_t^S \sqrt{c_t^\varphi / c_t^S}$; the argument for Δ^H is analogous.

[10] In a Markovian setting, Itô's formula shows that these processes indeed coincide with the usual notion of an option's "gamma," i.e., the second derivative of the option price with respect to the underlying.

costs. For the Black–Scholes model, where the frictionless optimal risky position in the pure investment problem is constant, (3.5) reduces to the formula derived by Whalley and Wilmott (1997, Section 4).

Inserting (3.5) into (3.3), the corresponding utility indifference price for small transaction costs is found to be:

$$\pi^{\varepsilon}(H) \sim E_Q[H] + \left(\frac{9p}{32}\right)^{1/3} \varepsilon^{2/3} E_Q\left[\int_0^T \left[\left(\frac{d\langle \varphi + \Delta^H\rangle_t}{d\langle S\rangle_t}\right)^{2/3} - \left(\frac{d\langle \varphi\rangle_t}{d\langle S\rangle_t}\right)^{2/3}\right] S_t^{2/3} d\langle S\rangle_t\right]$$

$$= E_Q[H] + \left(\frac{9p}{32}\right)^{1/3} \varepsilon^{2/3} E_Q\left[\int_0^T \left[\left|\Gamma_t^{\varphi} S_t^2 + \Gamma_t^H S_t^2\right|^{4/3} - \left|\Gamma_t^{\varphi} S_t^2\right|^{4/3}\right] \frac{d\langle S\rangle_t}{S_t^2}\right]. \tag{3.8}$$

The correction compared to the frictionless model is therefore—up to a constant—determined by the Q-expected time average of the difference between (suitable powers of) the future cash-gamma of the investor's optimal position with and without the option, scaled by the infinitesimal variance of the relative returns. For the Black–Scholes model, one readily verifies that (3.8) can be rewritten in terms of the inhomogeneous Black–Scholes equation of Whalley and Wilmott (1997, Section 3.3).

Representation (3.8) implies that the investor should charge a higher price than in the frictionless case if delta-hedging the option increases the sensitivity of her position with respect to price changes of the risky asset. Conversely, a lower premium is sufficient if delta hedging the claim reduces this sensitivity. The impact of trade size and risk aversion depends on the relative importance of investment and hedging. Since the underlying frictionless market is complete, the perfect hedge Δ_t^H and, in turn, its gamma Γ_t^H are independent of the investor's risk aversion, but scale linearly with the number n of claims sold. Conversely, the optimal investment strategy φ_t and its gamma Γ_t^{φ} are independent of the option position sold, but scale linearly with the inverse of risk aversion. Consequently, the comparative statics of utility-based prices and hedges with transaction costs—which depend on both quantities—are ambiguous in general. However, they can be analyzed in more detail if either the option position or the pure investment dominates.

Marginal Investment. First, we focus on the case where the investor's primary focus lies on the pricing and risk management of her option position. This regime applies if the cash-gamma Γ_t^{φ} of the pure investment under consideration is negligible compared to its counterpart Γ_t^{nH} for the option position nH. In particular, this occurs if the risky asset is assumed to be a martingale with vanishing risk premium, as in Hodges and Neuberger (1989), so that no investment is optimal without the option position, or in the asymptotic regime of Barles and Soner (1998), where the option position increases as the spread becomes small.

If the contribution of the pure investment strategy is negligible, formula (3.8) for the indifference price per claim reads as

$$\frac{\pi^{\varepsilon}(nH)}{n} \sim E_Q[H] + \left(\frac{9pn\varepsilon^2}{32}\right)^{1/3} E_Q\left[\int_0^T \left|\Gamma_t^H S_t^2\right|^{4/3} \frac{d\langle S\rangle_t}{S_t^2}\right]. \tag{3.9}$$

For a marginal pure investment, small transaction costs therefore always lead to a positive price correction compared to the frictionless case. The interpretation is that the transaction costs incurred by carrying out the approximate hedge necessitate a higher premium. The size of this effect depends on the relative magnitudes of trade size, risk aversion, and

transaction costs. Trade size n and risk aversion p both enter the leading-order correction through their cubic roots, and therefore in an interchangeable manner.[11]

If the contribution of the pure investment strategy is negligible, the optimal trading strategy in the presence of the claims can be directly interpreted as a utility-based hedge. In view of (3.5), the latter corresponds to keeping the number of risky shares within a no-trade region around the frictionless delta-hedge $\Delta_t^{nH} = n\Delta_t^H$; the maximal monetary deviations allowed are

$$\Delta\varphi^{nH,\pm}S_t = \pm\frac{n^{2/3}\varepsilon^{1/3}}{p^{1/3}}\left(\frac{3}{2}\left(\Gamma_t^H S_t^2\right)^2\right)^{1/3}. \tag{3.10}$$

Higher risk aversion induces closer tracking of the frictionless target here, leading to more trading and, in turn, higher prices (cf. formula (3.9)).

Semistatic Delta-Gamma Hedging. "Delta-gamma hedging" is often advocated in order to reduce the impact of transaction costs (cf., e.g., Björk 2003, p. 129). The above results allow to relate this idea to the semistatic hedging of a claim H, by dynamic trading in the underlying risky asset and a static position n' in some other claim H' setup at time zero. In the frictionless case, the choice of n' does not matter, since any such position can be offset by delta-hedging with the underlying. With transaction costs, this no longer remains true. Suppose the risky asset is a martingale ($P = Q$) and H and H' are traded at their frictionless prices $E_Q[H]$, $E_Q[H']$ with some transaction costs of order $O(\varepsilon)$. Then, (3.9) applied to the claim $H - n'H'$ shows that the leading-order frictional certainty equivalent of selling one unit of H—and hedging it with a static position of n' units of H' and optimal dynamic trading in the underlying—is given by

$$-\left(\frac{9p}{32}\right)^{1/3}\varepsilon^{2/3}E_Q\left[\int_0^T\left|\Gamma_t^H S_t^2 - n'\Gamma_t^{H'}S_t^2\right|^{4/3}\frac{d\langle S\rangle_t}{S_t^2}\right].$$

Maximizing this certainty equivalent to n' therefore amounts to minimizing a suitable average of the future (cash) gamma of the total option position $H - n'H'$.

If other options are only used once for hedging, the total position should thus not be made gamma-neutral at any one point in time. The exception is when both option gammas are approximately constant over the horizon under consideration; then it is optimal to make the total option position gamma-neutral. This situation occurs if the static option position is only held briefly, and before maturity of either claim. Hence, it seems reasonable to conjecture that one should trade to remain close to a delta-gamma neutral position if hedging dynamically with both the underlying and an option.[12]

Marginal Option Position. Now, let us turn to the converse situation of a *marginal option position*, i.e., the sale of a *small* number n of claims H. For incomplete markets without frictions, the limiting price for $n \to 0$, called *marginal-utility-based price*, is

[11] Barles and Soner (1998) consider the case where the product $pn\varepsilon^2$ converges to a finite limit. In the Black–Scholes model, they characterize the limiting price per claim as the solution to an inhomogeneous Black–Scholes equation. However, this limit does not coincide with the right-hand side of (3.9), whose derivation assumes only transaction costs to be small while risk aversion is fixed.

[12] Compare Goodman and Ostrov (2011) for related results in a pure investment problem with a stock and an option.

a linear pricing rule, namely, the expectation under the frictionless minimal entropy martingale measure, independent of both trade size and risk aversion. The leading-order correction for small n is linear both in the trade size n and in the investor's risk aversion p (Becherer 2005; Mania and Schweizer 2006; Kallsen and Rheinländer 2011). Hence, both quantities are interchangeable also in this setting: doubling risk aversion has the same effect as doubling the trade size.

Let us now derive corresponding results for the incompleteness caused by imposing small transaction costs in an otherwise complete market. Then, Taylor expanding (3.8) for small n yields:[13]

$$\frac{\pi^{\varepsilon}(nH)}{n} \sim E_Q[H] + \left(\frac{9p}{32}\right)^{1/3} \varepsilon^{2/3} E_Q\left[\int_0^T \frac{4}{3}\left|\Gamma_t^{\varphi} S_t^2\right|^{4/3} \frac{\Gamma_t^H}{\Gamma_t^{\varphi}} \frac{d\langle S\rangle_t}{S_t^2}\right] + n\left(\frac{9p}{32}\right)^{1/3} \varepsilon^{2/3} E_Q\left[\int_0^T \frac{2}{9}\left|\Gamma_t^{\varphi} S_t^2\right|^{4/3} \left(\frac{\Gamma_t^H}{\Gamma_t^{\varphi}}\right)^2 \frac{d\langle S\rangle_t}{S_t^2}\right].$$

Recall that the optimal frictionless strategy φ_t and its gamma Γ_t^{φ} are independent of the trade size n but scale linearly with the inverse of risk aversion p. Hence, the frictionless scalings are robust with respect to small transaction costs: The limiting price per claim for $n \to 0$ is also linear in the claim, and independent of trade size and risk aversion. Moreover, these quantities both enter linearly, and therefore in an interchangeable manner, in the leading-order correction term for small trade sizes.

For a small option position, the sign of the prize correction compared to the complete frictionless market depends on the interplay between the pure investment strategy and the hedge for the claim. The pure investment strategy is typically negatively correlated with price shocks, $\Gamma_t^{\varphi} < 0$ (e.g., in the Black–Scholes model). Then, the difference between the frictionless price and the limiting price with transaction costs is determined by the sign of the option's gamma. If the latter is positive as for European call or put options, then the marginal utility-based price taking into account transaction costs is smaller than its frictionless counterpart. This is because utility-based investment strategies are typically of contrarian type, i.e., decreasing when the risky price rises, whereas the delta-hedge of a European call or put is increasing with the value of its underlying. Consequently, hedging the claim allows the investor to save transaction costs so that she is willing to sell the claim for a smaller premium. This rationale, however, is only applicable if the fluctuations of the hedging position are small enough to be absorbed by the investor's other investments. In particular, the price adjustment is always positive for a marginal pure investment.

Let us also consider how the investor's trading strategy changes in the presence of a small number of claims. Taylor expanding (3.5) for small n shows that it is optimal to refrain from trading as long as the risky position remains within a bandwidth of

$$\Delta\varphi_t^{nH,\pm} S_t \sim \pm\Delta\varphi_t^{\pm} S_t \left(1 + \frac{2}{3} n \frac{\Gamma_t^H}{\Gamma_t^{\varphi}}\right)$$

around the frictionless optimal position $(\varphi_t + n\Delta_t^H)S_t$. The interpretation for the ratio of gammas is the same as for the corresponding utility-based prices above: if the trades prescribed by the delta hedge partially offset moves of the pure investment strategy, then

[13] Here and in the sequel, we report the leading-order terms for small transaction costs ε *and* small trade size n.

the resulting reduced sensitivity to price shocks allows the investor to use a smaller no-trade region than in the absence of the claims. As for risk aversion, note that whereas the investor's pure investment and the corresponding trading boundaries $\Delta\varphi_t^{\pm}$ in the absence of the claim scale with her risk tolerance $1/p$, the adjustment due to the presence of the claims does not. For small option positions, it is linear in trade size but independent of risk aversion as is true for the frictionless hedge.

3.2. Incomplete Markets

In incomplete markets, simple formulas for indifference prices and hedging strategies can typically only be obtained in the limit for a small number of claims, even in the absence of frictions. If the trade size n is small, Mania and Schweizer (2006), Becherer (2005), and Kallsen and Rheinländer (2011) show that the optimal strategy φ_t for the pure investment problem should be complemented be $n\xi_t$, where ξ_t is the mean-variance optimal hedge for the claim, determined under the marginal pricing measure Q, i.e.,

$$\xi_t = \frac{d\langle V, S\rangle_t}{d\langle S\rangle_t},$$

where V_t denotes the Q-martingale generated by the payoff H. As a consequence, the leading-order adjustment of the portfolio due to the presence of the claim is linear in trade size and independent of risk aversion, as in the complete case discussed above. The corresponding indifference price per claim converges to the expectation under the marginal pricing measure, which is again independent of trade size and risk aversion. The leading-order adjustment for larger trade sizes is given by the $pn/2$-fold of the minimal Q-expected squared hedging error, i.e.,

$$\frac{\pi^0(nH)}{n} \sim E_Q[H] + \frac{pn}{2} E_Q\left[\left(H - E_Q[H] - \int_0^T \xi_t\, dS_t\right)^2\right]. \tag{3.11}$$

As a result, it is linear both in trade size and risk aversion. In this setting of a small option position held in a potentially incomplete frictionless market, we now discuss the implications of small transaction costs.

Negligible Risk Premium. Let us first consider the case where the risky asset is a martingale under the physical probability. Then, no trading is optimal for the pure investment problem, $\varphi_t = 0$, and the minimal entropy martingale measure coincides with the physical probability, $Q = P$. As a consequence, the monetary trading boundaries (3.1) around the frictionless strategy $n\xi_t$ are given by

$$\Delta\varphi_t^{nH,\pm} S_t \sim \pm \frac{n^{2/3}\varepsilon^{1/3}}{p^{1/3}} \left(\frac{3}{2}\frac{d\langle\xi\rangle_t}{d\langle S\rangle_t} S_t^4\right)^{1/3}.$$

In view of (3.7), this is the same formula as in the complete case (3.10), with the perfect hedge replaced by the mean-variance optimal one. In particular, the scalings in trade size and risk aversion are robust to incompleteness in the frictionless market, as long as the option position is small.

To determine the corresponding leading-order price correction, insert the above trading boundaries into (3.3) and note that $Q = P$, $\Delta\varphi_t^{+} = 0$, as well as $dQ^{nH}/dP = 1 + O(n)$. As a consequence,

$$\frac{\pi^{\varepsilon}(nH)}{n} \sim E[H] + \frac{pn}{2} E\left[\left(H - E[H] - \int_0^T \xi_t dS_t\right)^2\right] + \left(\frac{9pn\varepsilon^2}{32}\right)^{1/3} E\left[\int_0^T \left(\frac{d\langle \xi\rangle_t}{d\langle S\rangle_t} S_t^4\right)^{2/3} \frac{d\langle S\rangle_t}{S_t^2}\right].$$

The second term is the correction due to transaction costs, which once more parallels the complete case (3.9), with the mean-variance optimal hedge again replacing the replicating strategy. The first term is the correction (3.11) due to the incompleteness of the frictionless market, which is proportional to the minimal squared hedging error and hence vanishes in the complete case. At the leading order, the two price corrections therefore separate; their relative sizes are determined by the magnitude of risk aversion p times trade size n, compared to the spread ε.

Nontrivial Risk Premium. If the pure investment strategy φ_t is not negligible, the trading boundaries (3.1) around the frictionless strategy $\varphi_t + n\xi_t$ are given by

$$\Delta\varphi_t^{nH,\pm} \sim \Delta\varphi_t^{\pm}\left(1 + \frac{2n}{3}\frac{d\langle \varphi, \xi\rangle_t}{d\langle \varphi\rangle_t}\right).$$

In the small claim limit, the interpretations from the complete case are therefore robust as well: If shocks to the frictionless investment and hedging strategies are negatively correlated, one should keep a smaller buffer with the claim, and conversely for the case of a positive correlation.

Concerning the pricing implications of small transaction costs added to incomplete frictionless markets, the situation is somewhat more involved. The reason is that the presence of the claim changes the impact of the transaction costs in two different ways. On the one hand, it affects the trading strategy that is used: the investor passes from staying within $\Delta\varphi_t^{\pm}$ around the pure investment strategy φ_t, to keeping within $\Delta\varphi_t^{nH,\pm}$ around $\varphi_t^{nH} = \varphi_t + n\xi_t + O(n^2)$. On the other hand, even after hedging the claim, the latter still induces some unspanned risk in incomplete markets, and therefore affects the investor's marginal evaluation rule. That is, the marginal pricing measure changes from Q, with density proportional to the marginal utility $U'(\int_0^T \varphi_t dS_t)$ associated with the pure investment strategy, to Q^{nH}, with density proportional to the marginal utility augmented by the n claims, $U'(\int_0^T \varphi_t^H dS_t - nH)$ (compare Owen 2002, Theorem 1.1). By formula (3.3) and Taylor expansion for small n, the leading-order price impact of small transaction costs is then found to be given by

$$\frac{p}{2}\left(E_Q\left[\int_0^T \frac{4n}{3}\left(\Delta\varphi_t^{+}\right)^2 \frac{d\langle \varphi, \xi\rangle_t}{d\langle \varphi\rangle_t} d\langle S\rangle_t\right] + E\left[\left(\frac{dQ^{nH}}{dP} - \frac{dQ}{dP}\right)\int_0^T \left(\Delta\varphi_t^{+}\right)^2 d\langle S\rangle_t\right]\right).$$

The first term is due to changing the trading strategy; it is already visible for complete frictionless markets. The second term reflects the change of the marginal pricing measure due to the presence of the claim, which does not take place in complete frictionless markets with a unique equivalent martingale measure. As for the hedging strategy above, the sign of the first term depends on the correlation of shocks to investment and hedging strategies. To examine the sign of the second term, notice that $\varphi_t^H = \varphi_t + n\xi_t + O(n^2)$,

Taylor expansion, and the Q-martingale property of the wealth process $\int_0^T \xi_t dS_t$ yield

$$\begin{aligned}\frac{dQ^{nH}}{dP} &= \frac{e^{-p(\int_0^T(\varphi_t+n\xi_t)dS_t+O(n^2)-nH)}}{E\left[e^{-p(\int_0^T(\varphi_t+n\xi_t)dS_t+O(n^2)-nH)}\right]} \\ &= \frac{dQ}{dP}\left(1+np\left(H-E_Q[H]-\int_0^T \xi_t dS_t\right)\right)+O(n^2).\end{aligned}$$

As a result, the second term in the price impact of small transaction costs is given by the covariance between the shortfall of the frictionless utility-based hedge and the cumulated transaction costs effect, measured by the average squared width of the no-trade region:

$$\begin{aligned}&E\left[\left(\frac{dQ^{nH}}{dP}-\frac{dQ}{dP}\right)\int_0^T \left(\Delta\varphi_t^+\right)^2 d\langle S\rangle_t\right] \\ &\quad\sim npE_Q\left[\left(H-E_Q[H]-\int_0^T \xi_t dS_t\right)\int_0^T \left(\Delta\varphi_t^+\right)^2 d\langle S\rangle_t\right].\end{aligned}$$

Hence, incompleteness of the frictionless market increases the impact of transaction costs, if these tend to accrue more rapidly when the imperfect utility-based hedge also does badly, i.e., when the different sources of incompleteness tend to cluster. In contrast, the premium for the option is decreased if the two risks are negatively correlated and thereby diversify the investor's portfolio. The same correlation adjustment also occurs in frictionless markets, when passing from the marginal pricing measure for the pure investment problem to its counterpart in the presence of a small option position. In this sense, the marginal pricing implications of transaction costs are therefore the same as for selling a path-dependent option with payoff $\int_0^T (\Delta\varphi_t^+)^2 d\langle S\rangle_t$. It is important to emphasize, however, that this is not the case at all for the corresponding hedge.

APPENDIX A: DERIVATION OF THE MAIN RESULTS

In the following, the main results are derived by applying formal perturbation arguments to the martingale optimality conditions for a frictionless *shadow price*. The latter is a "least favorable" frictionless market extension in the sense that it evolves in the bid-ask spread, thereby leading to potentially more favorable trading prices, but admits an optimal policy that only entails the purchase resp. sale of risky shares when the shadow price coincides with the ask resp. bid price.

The observation that such a shadow price should always exist can be traced back to Jouini and Kallal (1995) as well as Cvitanić and Karatzas (1996) (also cf. Loewenstein 2002). Starting with Kallsen and Muhle-Karbe (2010), this concept has recently also been used for the computation and verification of optimal policies in simple settings. Since shadow prices are not known a priori, they have to be determined simultaneously with the optimal policy. Here, we show how to do so for general continuous asset prices, *approximately* for small costs. In contrast to most previous asymptotic results, we do not first solve the problem for arbitrary costs $\varepsilon > 0$, and then expand the solution around $\varepsilon = 0$. Instead, we directly tackle the much simpler approximate problem for $\varepsilon \sim 0$, in the same spirit as in the approach of Soner and Touzi (2012).

Throughout, mathematical formalism is treated liberally. For example, we do not state and verify technical conditions warranting the uniform integrability of local

martingales, interchange of integration and differentiation, and the uniformity of estimates. In particular, the Landau symbols $O(\cdot)$ and $o(\cdot)$ refer to pointwise estimates, with the implicit assumption of enough regularity in time and states to eventually turn these into an estimate of the expected utility generated by the approximately optimal policy. Rigorous proofs have been worked out in the present setting for the Black–Scholes model (Bichuch 2011; Guasoni and Muhle-Karbe 2015), and by Soner and Touzi (2012) for an infinite-horizon consumption problem in a Markovian setup.

A.1. Notation

Throughout, we write $\phi \bullet S$ for the stochastic integral $\int_0^{\cdot} \phi_t dS_t$. The identity process is denoted by $I_t = t$ and for any Itô process X, we write b^X and σ^X for its local drift and diffusion coefficients, respectively, in the sense that $dX_t = b_t^X dt + \sigma_t^X dW_t$ for a standard Brownian motion W. Finally, for Itô processes X and Y, we denote by $c_t^{X,Y} = d\langle X, Y\rangle_t/dt$ their local quadratic covariation; if $X = Y$, we abbreviate to $d\langle X\rangle_t/dt = c_t^{X,X} = c_t^X$.

A.2. Martingale Optimality Conditions

In this section, we formally derive conditions ensuring that a family $(\varphi^\varepsilon)_{\varepsilon>0}$ of frictional strategies is *approximately optimal* as the spread ε becomes small. For the convenience of the reader, we first briefly recapitulate their *exact* counterparts in the frictionless case.

Frictionless Optimality Conditions. In the absence of transaction costs ($\varepsilon = 0$), the following duality result is well known (cf., e.g., Delbaen et al. 2002): The wealth process $x + \varphi \bullet S$ corresponding to a trading strategy φ is optimal, if (and essentially only if) there exists a process Z satisfying the following optimality conditions:

(i) Z is a martingale.
(ii) ZS is a martingale.
(iii) $Z_T = U'(x + \varphi \bullet S_T)$.

The first two conditions imply that Z is—up to normalization—the density of an equivalent martingale measure Q for S. The third identifies it as the solution to a dual minimization problem, linked to the primal maximizer by the usual first-order condition.

Let us briefly recall why conditions (i)–(iii) imply the optimality of φ. To this end, let ψ be any competing strategy. Then, the concavity of the utility function U and condition (iii) imply

$$\begin{aligned} E[U(x + \psi \bullet S_T)] &\leq E[U(x + \varphi \bullet S_T)] + E[U'(x + \varphi \bullet S_T)(\psi - \varphi) \bullet S_T] \\ &= E[U(x + \varphi \bullet S_T)] + E[Z_0] E_Q[(\psi - \varphi) \bullet S_T]. \end{aligned}$$

Since S and, in turn, the wealth process $(\psi - \varphi) \bullet S$ is a Q-martingale by conditions (i) and (ii), the second expectation vanishes and the optimality of φ follows.

Approximate Optimality Conditions with Transaction Costs. Now, let us derive *approximate* versions of conditions (i)–(iii) in the presence of small transaction costs ε, ensuring the *approximate optimality* of a family $(\varphi^\varepsilon)_{\varepsilon>0}$ of strategies, at the leading order $O(\varepsilon^{2/3})$ as

ε becomes small. The exact optimal strategies converge to their frictionless counterpart. Hence, it suffices to consider families $(\psi^\varepsilon)_{\varepsilon>0}$ of strategies converging to the frictionless optimizer φ, i.e., $\psi^\varepsilon = \varphi + o(1)$.

Let $(\varphi^\varepsilon)_{\varepsilon>0}$ be a candidate family of strategies whose optimality we want to verify. As above, for any family of competitors $(\psi^\varepsilon)_{\varepsilon>0}$, the concavity of U implies:

$$E\big[U\big(X_T^{\psi^\varepsilon}\big)\big] \le E\big[U\big(X_T^{\varphi^\varepsilon}\big)\big] + E\big[U'\big(X_T^{\varphi^\varepsilon}\big)\big(X_T^{\psi^\varepsilon} - X_T^{\varphi^\varepsilon}\big)\big], \tag{A.1}$$

where $X_T^{\psi^\varepsilon}$, $X_T^{\varphi^\varepsilon}$ denote the payoffs generated by trading the strategies with transaction costs. Now, suppose we can find *shadow prices* S^ε evolving within the bid-ask spreads $(1 \pm \varepsilon)S$, matching the trading prices $(1 \pm \varepsilon)S$ in the original market with transaction costs whenever the respective strategies φ^ε trade. Then, the frictional wealth process associated with φ^ε evidently coincides with its frictionless counterpart for S^ε, i.e., $X_T^{\varphi^\varepsilon} = x + \varphi^\varepsilon \bullet S_T^\varepsilon$. For any other strategy, trading in terms of S^ε rather than with the original bid-ask spread can only increase wealth, since trades are carried out at potentially more favorable prices: $X_T^{\psi^\varepsilon} \le x + \psi^\varepsilon \bullet S_T^\varepsilon$. Together with (A.1), this implies:

$$E\big[U\big(X_T^{\psi^\varepsilon}\big)\big] \le E\big[U\big(X_T^{\varphi^\varepsilon}\big)\big] + E\big[U'\big(X_T^{\varphi^\varepsilon}\big)\big(\psi^\varepsilon - \varphi^\varepsilon\big) \bullet S_T^\varepsilon\big]. \tag{A.2}$$

Now, suppose we can find a process Z^ε satisfying the following approximate versions of the frictionless optimality conditions (i)–(iii) above:
(i$^\varepsilon$) Z^ε is approximately a martingale, in that its drift rate b^{Z^ε} is of order $O(\varepsilon^{2/3})$.
(ii$^\varepsilon$) $Z^\varepsilon S^\varepsilon$ is approximately a martingale, in that its drift rate $b^{Z^\varepsilon S^\varepsilon}$ is of order $O(\varepsilon^{2/3})$.
(iii$^\varepsilon$) $Z_T^\varepsilon = U'(x + \varphi^\varepsilon \bullet S_T^\varepsilon) + O(\varepsilon^{2/3})$.

Then, since $\psi^\varepsilon - \varphi^\varepsilon = o(1)$, condition (iii$^\varepsilon$) implies that (A.2) can be rewritten as

$$E\big[U\big(X_T^{\psi^\varepsilon}\big)\big] \le E\big[U\big(X_T^{\varphi^\varepsilon}\big)\big] + E\big[Z_T^\varepsilon\big((\psi^\varepsilon - \varphi^\varepsilon) \bullet S_T^\varepsilon\big)\big] + o(\varepsilon^{2/3}).$$

Applying integration by parts twice yields

$$\begin{aligned} Z^\varepsilon((\psi^\varepsilon - \varphi^\varepsilon) \bullet S^\varepsilon) &= Z^\varepsilon(\psi^\varepsilon - \varphi^\varepsilon) \bullet S^\varepsilon + ((\psi^\varepsilon - \varphi^\varepsilon) \bullet S^\varepsilon) \bullet Z^\varepsilon + (\psi^\varepsilon - \varphi^\varepsilon) \bullet \langle Z^\varepsilon, S^\varepsilon \rangle \\ &= ((\psi^\varepsilon - \varphi^\varepsilon) \bullet S^\varepsilon - (\psi^\varepsilon - \varphi^\varepsilon)S^\varepsilon) \bullet Z^\varepsilon + (\psi^\varepsilon - \varphi^\varepsilon) \bullet (Z^\varepsilon S^\varepsilon), \end{aligned}$$

and, in turn,

$$\begin{aligned} E\big[U\big(X_T^{\psi^\varepsilon}\big)\big] \le{}& E\big[U\big(X_T^{\varphi^\varepsilon}\big)\big] \\ &+ E\big[((\psi^\varepsilon - \varphi^\varepsilon) \bullet S^\varepsilon - (\psi^\varepsilon - \varphi^\varepsilon)S^\varepsilon) \bullet Z_T^\varepsilon + (\psi^\varepsilon - \varphi^\varepsilon) \bullet (Z^\varepsilon S^\varepsilon)_T\big] + o(\varepsilon^{2/3}). \end{aligned}$$

The second expectation is given by the integrated expected drift rate of its argument

$$((\psi^\varepsilon - \varphi^\varepsilon) \bullet S^\varepsilon - (\psi^\varepsilon - \varphi^\varepsilon)S^\varepsilon)\, b^{Z^\varepsilon} + (\psi^\varepsilon - \varphi^\varepsilon) b^{Z^\varepsilon S^\varepsilon},$$

which is of order $o(\varepsilon^{2/3})$, by conditions (i$^\varepsilon$) and (ii$^\varepsilon$) above and because $\psi^\varepsilon - \varphi^\varepsilon = o(1)$. Hence,

$$E\big[U\big(X_T^{\psi^\varepsilon}\big)\big] \le E\big[U\big(X_T^{\varphi^\varepsilon}\big)\big] + o(\varepsilon^{2/3}),$$

and the expected utilities of the candidate family $(\varphi^\varepsilon)_{\varepsilon>0}$ therefore dominate those of the competitors $(\psi^\varepsilon)_{\varepsilon>0}$ at the leading order $O(\varepsilon^{2/3})$. In summary, the strategies $(\varphi^\varepsilon)_{\varepsilon>0}$ are indeed approximately optimal if we can find a shadow price S^ε and an approximate martingale density Z^ε satisfying the approximate optimality conditions (i$^\varepsilon$)–(iii$^\varepsilon$).

A.3. Derivation of a Candidate Policy

We now look for strategies φ^ε, shadow prices S^ε, and approximate martingale densities Z^ε satisfying the approximate optimality conditions (i^ε) – (iii^ε). Write

$$\varphi^\varepsilon = \varphi + \Delta\varphi, \quad S^\varepsilon = S + \Delta S.$$

Motivated by previous asymptotic results (Whalley and Wilmott 1997; Bichuch 2011; Guasoni and Muhle-Karbe 2015; Soner and Touzi 2012), we assume that the deviations of the optimal strategy with transaction costs from the frictionless optimizer are asymptotically proportional to the cubic root of the spread:

$$\Delta\varphi = O(\varepsilon^{1/3}). \tag{A.3}$$

Since the shadow price S^ε has to lie in the bid-ask spread $(1 \pm \varepsilon)S$, we must have

$$\Delta S = O(\varepsilon). \tag{A.4}$$

In addition, we assume that ΔS is an Itô process with drift and diffusion coefficients satisfying

$$b^{\Delta S} = O(\varepsilon^{1/3}), \quad \sigma^{\Delta S} = O(\varepsilon^{2/3}). \tag{A.5}$$

All of these assumptions will turn out to be consistent with the results of our calculations below, see Section 1.4. Now, notice that

$$\varphi^\varepsilon \bullet S^\varepsilon = \varphi \bullet S + \Delta\varphi \bullet S + O(\varepsilon^{2/3}),$$

because (A.3) and (A.5) give $\Delta\varphi \bullet \Delta S = O(\varepsilon^{2/3})$, and integration by parts in conjunction with (A.4) and (A.5) shows $\varphi \bullet \Delta S = O(\varepsilon^{2/3})$. Therefore,

$$U'\big(x + \varphi^\varepsilon \bullet S_T^\varepsilon\big) = pe^{-p(x+\varphi^\varepsilon \bullet S_T^\varepsilon)} = pe^{-p(x+\varphi \bullet S_T)}(1 - p\Delta\varphi \bullet S_T) + O(\varepsilon^{2/3}).$$

The factor $pe^{-p(x+\varphi \bullet S_T)}$ coincides with the terminal value of the frictionless martingale density Z (cf. the frictionless optimality condition (iii) above). The process

$$Z^\varepsilon = Z(1 - p\Delta\varphi \bullet S)$$

therefore is a martingale (because Z is the density of a martingale measure for S), satisfying condition (i^ε), for which (iii^ε) holds as well. It remains to determine ΔS and $\Delta\varphi$ for which (ii^ε) holds, too. Integration by parts yields

$$Z^\varepsilon S^\varepsilon - Z_0^\varepsilon S_0^\varepsilon = S^\varepsilon \bullet Z^\varepsilon + Z^\varepsilon \bullet S^\varepsilon + \langle Z^\varepsilon, S^\varepsilon \rangle.$$

Since the martingale Z^ε has zero drift, it follows that the drift rate of $Z^\varepsilon S^\varepsilon$ is given by

$$b^{Z^\varepsilon S^\varepsilon} = Z^\varepsilon(b^S + b^{\Delta S}) + c^{Z^\varepsilon, S+\Delta S}. \tag{A.6}$$

As $b^{\Delta S} = O(\varepsilon^{1/3})$ by assumption, and $Z^\varepsilon = Z(1 - p\Delta\varphi \bullet S)$, it follows that

$$Z^\varepsilon(b^S + b^{\Delta S}) = Z\big(b^S - p(\Delta\varphi \bullet S)b^S + b^{\Delta S}\big) + O(\varepsilon^{2/3}). \tag{A.7}$$

Moreover, writing the frictionless martingale density as a stochastic exponential $Z = \mathcal{E}(N) = 1 + Z \bullet N$, it follows from (A.5) and integration by parts that

$$\begin{aligned}\langle Z^\varepsilon, S + \Delta S\rangle &= \langle Z(1 - p\Delta\varphi \bullet S), S\rangle + O(\varepsilon^{2/3}) \\ &= Z \bullet (\langle N, S\rangle - p(\Delta\varphi \bullet S) \bullet \langle N, S\rangle - p\Delta\varphi \bullet \langle S, S\rangle) + O(\varepsilon^{2/3}),\end{aligned}$$

so that

$$c^{Z^\varepsilon, S+\Delta S} = Z(c^{N,S} - p(\Delta\varphi \bullet S)c^{N,S} - p\Delta\varphi c^S) + O(\varepsilon^{2/3}). \tag{A.8}$$

Then, inserting (A.7) and (A.8) into (A.6) and using that $b^S + c^{N,S} = 0$ by Girsanov's theorem because $ZS = \mathcal{E}(N)S$ is a martingale by the frictionless optimality condition (ii) gives

$$\begin{aligned}b^{Z^\varepsilon S^\varepsilon} &= Z(b^S + c^{N,S} - p(\Delta\varphi \bullet S)(b^S + c^{N,S}) + b^{\Delta S} - p\Delta\varphi c^S) + O(\varepsilon^{2/3}) \\ &= Z(b^{\Delta S} - p\Delta\varphi c^S) + O(\varepsilon^{2/3}).\end{aligned}$$

To make the drift of $Z^\varepsilon S^\varepsilon$ vanish—up to terms of order $O(\varepsilon^{2/3})$—in accordance with (ii$^\varepsilon$), it is therefore necessary that

$$b^{\Delta S} = p\Delta\varphi c^S + O(\varepsilon^{2/3}). \tag{A.9}$$

This drift condition naturally leads to an ansatz of the form $\Delta S = f(\Delta\varphi)$. Then, since the shadow price $S^\varepsilon = S + \Delta S$ has to move from the ask price $(1+\varepsilon)S$ to the bid price $(1-\varepsilon)S$ as $\Delta\varphi$ varies between some buy boundary $\Delta\varphi^-$ and some sell boundary $\Delta\varphi^+$, the function f has to satisfy the boundary conditions

$$f(\Delta\varphi^-) = \varepsilon S, \quad f(\Delta\varphi^+) = -\varepsilon S. \tag{A.10}$$

Moreover, even though the process $\Delta\varphi$ is reflected to remain between the trading boundaries, these singular terms should vanish in the dynamics of ΔS so that the shadow price $S^\varepsilon = S + \Delta S$ does not allow for arbitrage. By Itô's formula, this implies that the derivative of f should vanish at the boundaries:

$$f'(\Delta\varphi^-) = 0, \quad f'(\Delta\varphi^+) = 0. \tag{A.11}$$

The simplest family of functions capable of matching these boundary conditions is given by the symmetric cubic polynomials

$$f(x) = \alpha x^3 - \gamma x.$$

With this ansatz, (A.11) gives

$$\Delta\varphi^\pm = \pm\sqrt{\frac{\gamma}{3\alpha}},$$

and (A.10) implies

$$\gamma = \left(\frac{1}{2}\varepsilon S\right)^{2/3} 3\alpha^{1/3}.$$

Moreover, Itô's formula applied to $f(x) = \alpha x^3 - \gamma x$ yields

$$\Delta S - \Delta S_0 = (3\alpha\Delta\varphi^2 - \gamma) \bullet \Delta\varphi + 3\alpha\Delta\varphi \bullet \langle \Delta\varphi, \Delta\varphi \rangle.$$

Now, notice that the optimal trading strategy $\varphi^\varepsilon = \varphi + \Delta\varphi$ with transaction costs is necessarily of finite variation. Assuming it is also continuous then implies $\langle \Delta\varphi \rangle = \langle \varphi \rangle$. Moreover, since φ^ε is constant except at the trading boundaries (where $\Delta\varphi = \Delta\varphi^\pm$ and, in turn, $3\alpha\Delta\varphi^2 - \gamma = 0$), we also have

$$(3\alpha\Delta\varphi^2 - \gamma) \bullet \Delta\varphi = -(3\alpha\Delta\varphi^2 - \gamma) \bullet \varphi.$$

Thus,

$$\Delta S - \Delta S_0 = -(3\alpha\Delta\varphi^2 - \gamma) \bullet \varphi + 3\alpha\Delta\varphi \bullet \langle \varphi \rangle,$$

and the drift coefficient of ΔS is given by

$$b^{\Delta S} = 3\alpha\Delta\varphi c^\varphi + O(\varepsilon^{2/3}).$$

Comparing this to the leading-order term in (A.9), we obtain

$$\alpha = \frac{p}{3}\frac{c^S}{c^\varphi},$$

and, in turn,

$$\gamma = \left(\frac{3p^{1/2}}{2}\sqrt{\frac{c^S}{c^\varphi}}S\right)^{2/3} \varepsilon^{2/3}$$

as well as

$$\Delta\varphi^\pm = \pm\sqrt{\frac{\gamma}{3\alpha}} = \pm\left(\frac{3}{2p}\frac{c^\varphi}{c^S}\varepsilon S\right)^{1/3}.$$

At the first order, this determines the optimal strategy φ^ε with transaction costs as the minimal amount of trading necessary to remain in the randomly changing interval $[\varphi + \Delta\varphi^-, \varphi + \Delta\varphi^+]$ around the frictionless optimizer φ.

A.4. Approximate Optimality

The above considerations assumed that the coefficients α and γ are constant, but then lead to stochastic processes α_t and γ_t, which seems contradictory at first glance. However, we can verify a fortiori that this choice does indeed satisfy (i$^\varepsilon$)–(iii$^\varepsilon$). To see this set, for α, γ as above,

$$\Delta S_t = \alpha_t \Delta\varphi_t^3 - \gamma_t \Delta\varphi_t,$$

and let the strategy $\varphi^\varepsilon = \varphi + \Delta\varphi$ correspond to the minimal amount of trading necessary to remain within the boundaries $\Delta\varphi^\pm$ around the frictionless optimizer φ. Then, by definition, the process $S^\varepsilon := S + \Delta S$ takes values in $[(1-\varepsilon)S, (1+\varepsilon)S]$ and coincides with the bid resp. ask price whenever φ^ε reaches the selling boundary $\varphi + \Delta\varphi^+$ resp. the buying boundary $\varphi - \Delta\varphi^+$ as required for a shadow price. Concerning the dynamics of

ΔS, notice that integration by parts (now taking into account the stochasticity of α and γ) and Itô's formula give

$$\begin{aligned}\Delta S - \Delta S_0 &= \alpha \bullet (\Delta\varphi^3) + \Delta\varphi^3 \bullet \alpha + \langle \alpha, \Delta\varphi^3 \rangle - \gamma \bullet \Delta\varphi - \Delta\varphi \bullet \gamma - \langle \gamma, \Delta\varphi \rangle \\ &= (3\alpha\Delta\varphi^2 - \gamma) \bullet \Delta\varphi + (3\alpha\Delta\varphi) \bullet \langle \Delta\varphi \rangle \\ &\quad + \Delta\varphi^3 \bullet \alpha - \Delta\varphi \bullet \gamma + (3\Delta\varphi^2) \bullet \langle \alpha, \Delta\varphi \rangle - \langle \gamma, \Delta\varphi \rangle \\ &= -(3\alpha\Delta\varphi^2 - \gamma) \bullet \varphi + (3\alpha\Delta\varphi) \bullet \langle \varphi \rangle \\ &\quad + \Delta\varphi^3 \bullet \alpha - \Delta\varphi \bullet \gamma - (3\Delta\varphi^2) \bullet \langle \alpha, \varphi \rangle + \langle \gamma, \varphi \rangle. \end{aligned} \tag{A.12}$$

Here, we have used for the last equality that $\varphi^\varepsilon = \varphi + \Delta\varphi$ only moves on the set $\Delta\varphi = \Delta\varphi^\pm$ where $3\alpha\Delta\varphi^2 - \gamma = 0$, and that $\Delta\varphi = -\varphi + \varphi^\varepsilon$ only differs from $-\varphi$ by a finite variation term. If the risky asset S, the frictionless optimizer φ, as well as their local quadratic variation processes c^S, c^φ (and, in turn, the processes α and γ) follow sufficiently regular Itô processes, this representation shows that this property is passed on to ΔS. Moreover, since $\Delta\varphi = O(\varepsilon^{1/3})$, $\gamma = O(\varepsilon^{2/3})$, and $\alpha = O(1)$ (and the same asymptotics are valid for the drift and diffusion coefficients of α and γ), its diffusion coefficient is indeed of order $O(\varepsilon^{2/3})$ and its drift rate is of order $O(\varepsilon^{1/3})$. More specifically, the latter is given by $b^{\Delta S} = 3\alpha\Delta\varphi c^\varphi + O(\varepsilon^{2/3})$; hence, by definition of α, the drift condition (A.9) and, in turn, the approximate optimality condition (ii$^\varepsilon$) are indeed satisfied for the shadow price S^ε and the strategy φ^ε. Consequently, the latter is approximately optimal for small spreads.

A.5. Computation of the Leading-Order Utility Loss

Let us now compute—at the leading order $O(\varepsilon^{2/3})$—the expected utility that can be obtained by applying the strategy φ^ε. Since the latter is approximately optimal, this will then also determine the leading-order impact of transaction costs on the certainty equivalent of trading optimally in the market.

To do this, the analysis of the previous section needs to be refined. Including a second term in the Taylor expansion of the utility function, and taking into account $\varphi^\varepsilon \bullet S^\varepsilon = \varphi \bullet S + \Delta\varphi \bullet S + \varphi^\varepsilon \bullet \Delta S$, where $\varphi^\varepsilon \bullet \Delta S$ is of order $O(\varepsilon^{2/3})$,[14] gives

$$\begin{aligned}E\big[U\big(x + \varphi^\varepsilon \bullet S_T^\varepsilon\big)\big] &= E[U(x + \varphi \bullet S_T)] + E[U'(x + \varphi \bullet S_T)(\Delta\varphi \bullet S_T + \varphi^\varepsilon \bullet \Delta S_T)] \\ &\quad + \frac{1}{2} E[U''(x + \varphi \bullet S_T)(\Delta\varphi \bullet S_T)^2] + O(\varepsilon).\end{aligned}$$

For the exponential utility function $U(x) = -e^{-px}$, the marginal utility $U'(x + \varphi \bullet S_T)$ needs to be normalized by $E[U'(x + \varphi \bullet S_T)] = -pE[U(x + \varphi \bullet S_T)]$ to obtain the density of an equivalent martingale measure Q for S. Since, moreover, the absolute risk-aversion $-U''/U' = p$ is constant, it follows that

$$\begin{aligned}E\big[U\big(x + \varphi^\varepsilon \bullet S_T^\varepsilon\big)\big] &= E[U(x + \varphi \bullet S_T)] \\ &\quad \times \left(1 - pE_Q[\varphi^\varepsilon \bullet \Delta S_T] + \frac{p^2}{2} E_Q[(\Delta\varphi \bullet S_T)^2]\right) + O(\varepsilon),\end{aligned}$$

[14] This follows using integration by parts to write $\varphi^\varepsilon \bullet \Delta S = \Delta\varphi \bullet \Delta S + \varphi\Delta S - \varphi_0 \Delta S_0 - \Delta S \bullet \varphi - \langle \varphi, \Delta S \rangle$, and recalling that the drift and diffusion coefficients of ΔS are of order $O(\varepsilon^{1/3})$ and $O(\varepsilon^{2/3})$, respectively, whereas ΔS and $\Delta\varphi$ are of order $O(\varepsilon)$ resp. $O(\varepsilon^{1/3})$.

where we have used that the expectation of the Q-martingale $\Delta\varphi \bullet S$ vanishes. The second correction term $\frac{p^2}{2} E_Q[(\Delta\varphi \bullet S_T)^2]$ represents the leading-order relative utility loss due to displacement, incurred by trading φ^ε instead of the frictionless optimizer φ at the frictionless price S. The first correction term $-pE_Q[\varphi^\varepsilon \bullet \Delta S_T]$ measures the utility loss incurred directly due to transaction costs, when trades are carried out at the shadow price S^ε rather than at the mid price S.

Let us first focus on the displacement loss $\frac{p^2}{2} E_Q[(\Delta\varphi \bullet S_T)^2]$. Integration by parts gives

$$(\Delta\varphi \bullet S_T)^2 = 2(\Delta\varphi \bullet S)\Delta\varphi \bullet S_T + \Delta\varphi^2 \bullet \langle S\rangle_T.$$

As the first term is a Q-martingale, it follows that the leading-order displacement loss is given by

$$\frac{p^2}{2} E_Q[(\Delta\varphi \bullet S_T)^2] = \frac{p^2}{2} E_Q[\Delta\varphi^2 \bullet \langle S\rangle_T].$$

Now, consider the direct transaction cost loss $-pE_Q[\varphi^\varepsilon \bullet \Delta S_T]$. Integration by parts and $\Delta S = O(\varepsilon)$ yield

$$\varphi \bullet \Delta S = -\langle \varphi, \Delta S\rangle + O(\varepsilon).$$

First taking into account the dynamics of ΔS (cf. (A.12)), and then inserting the definitions of α and γ and the trading boundaries $\Delta\varphi^+$, as well as $c^S \bullet I = \langle S\rangle$, it follows that

$$\begin{aligned} -pE_Q[\varphi \bullet \Delta S_T] &= -pE_Q[(3\alpha\Delta\varphi^2 - \gamma)c^\varphi \bullet I_T] + O(\varepsilon) \\ &= -p^2 E_Q[(\Delta\varphi^2 - (\Delta\varphi^+)^2) \bullet \langle S\rangle_T] + O(\varepsilon). \end{aligned}$$

The remaining term $-pE_Q[\Delta\varphi \bullet \Delta S_T]$ can again be computed by integrating the drift rate of the argument of the expectation (here, $b^{\Delta S, Q}$ denotes the drift of ΔS under the measure Q):

$$\begin{aligned} -pE_Q[\Delta\varphi \bullet \Delta S_T] &= -pE_Q[\Delta\varphi b^{\Delta S, Q} \bullet I_T] \\ &= -pE_Q[\Delta\varphi(b^{\Delta S} + c^{N, \Delta S}) \bullet I_T] = -pE_Q[\Delta\varphi b^{\Delta S} \bullet I_T] + O(\varepsilon). \end{aligned}$$

Here, the second equality follows from Girsanov's theorem, and the third one holds since $\Delta\varphi = O(\varepsilon^{1/3})$ and the diffusion coefficient of ΔS is of order $O(\varepsilon^{2/3})$ by (A.12). Combined with the drift condition (A.9) and $c^S \bullet I = \langle S\rangle$, this yields

$$-pE_Q[\Delta\varphi \bullet \Delta S_T] = -p^2 E_Q[\Delta\varphi^2 \bullet \langle S\rangle_T] + O(\varepsilon).$$

As a consequence, the total relative utility loss directly caused by transaction costs is given by

$$-pE_Q[\varphi^\varepsilon \bullet \Delta S_T] = p^2 E_Q[((\Delta\varphi^+)^2 - 2\Delta\varphi^2) \bullet \langle S\rangle_T] + O(\varepsilon).$$

To further simplify the formulas for both parts of the utility loss, replace—at the leading order $O(\varepsilon^{2/3})$—the terms $\Delta\varphi^2$ by their expectation $\frac{1}{3}(\Delta\varphi^+)^2$ under the uniform distribution on $[\Delta\varphi^-, \Delta\varphi^+]$ (compare Rogers 2004; Goodman and Ostrov 2010), which is justified below.

Then, the displacement loss is determined as $\frac{p^2}{6} E_Q[(\Delta\varphi^+)^2 \bullet \langle S, S\rangle_T] + o(\varepsilon^{2/3})$, and the transaction cost loss is found to be given by twice that value. Hence, the total utility loss due to transaction costs is given by

$$E\big[U\big(x+\varphi^\varepsilon \bullet S_T^\varepsilon\big)\big] = E[U(x+\varphi \bullet S_T)]\left(1+\frac{p^2}{2}E_Q\left[(\Delta\varphi^+)^2 \bullet \langle S\rangle_T\right]\right) + o(\varepsilon^{2/3}),$$

and the claimed formula for the certainty equivalent follows by taking logarithms and Taylor expansion.

To complete the argument, it remains to verify that we can indeed pass to the uniform distribution for $\Delta\varphi$ at the leading order. To this end, define $D = \Delta\varphi\sigma^S$, which is an Itô process reflected to stay between the boundaries $D^\pm = \Delta\varphi^\pm\sigma^S$. In the interior of $[D^-, D^+]$, the strategy φ^ε is constant so that $\Delta\varphi = -\varphi$. Hence, the drift rate b^D and the diffusion coefficient σ^D of D are both of order $O(1)$. Now, fix a mesh $0 = t_0^\varepsilon < \cdots < t_{N^\varepsilon}^\varepsilon = T$ with mesh size of order $O(\varepsilon^{1/3})$, and write

$$\int_0^T \Delta\varphi_u^2 d\langle S\rangle_u = \int_0^T D_u^2 du = \sum_{i=1}^{N^\varepsilon} \int_{t_{i-1}^\varepsilon}^{t_i^\varepsilon} D_u^2 du. \tag{A.13}$$

Rescale D by dividing by $\varepsilon^{1/3}$ and integrating over $v = u/\varepsilon^{2/3}$ instead of u, obtaining

$$\int_{t_{i-1}^\varepsilon}^{t_i^\varepsilon} D_u^2 du = \varepsilon^{4/3} \int_{t_{i-1}^\varepsilon/\varepsilon^{2/3}}^{t_i^\varepsilon/\varepsilon^{2/3}} \left(\frac{D_{\varepsilon^{2/3}v}}{\varepsilon^{1/3}}\right)^2 dv. \tag{A.14}$$

The drift and diffusion coefficients of the rescaled integrand $(\varepsilon^{-1/3} D_{\varepsilon^{2/3}v})_{v\geq 0}$ are given by $b^D_{\varepsilon^{2/3}v}\varepsilon^{1/3} = O(\varepsilon^{1/3})$ and $\sigma^D_{\varepsilon^{2/3}v} = O(1)$, respectively. Hence, at the leading order, it equals driftless Brownian motion with constant volatility $\sigma^D_{t_i^\varepsilon}$ on $t_{i-1}^\varepsilon/\varepsilon^{2/3} \leq v \leq t_i^\varepsilon/\varepsilon^{2/3}$. Reflected Brownian motion on the interval $[D^-_{\varepsilon^{2/3}v}/\varepsilon^{1/3}, D^+_{\varepsilon^{2/3}v}/\varepsilon^{1/3}]$ (whose boundaries are constant and equal to $D^\pm_{t_{i-1}^\varepsilon}$, at the leading order) has a uniform stationary distribution with second moment $(D^+_{t_{i-1}^\varepsilon})^2/3\varepsilon^{2/3}$. Consequently, the ergodic theorem (Borodin and Salminen, 2002, II.35 and II.36) implies

$$\int_{t_{i-1}^\varepsilon/\varepsilon^{2/3}}^{t_i^\varepsilon/\varepsilon^{2/3}} \left(\frac{D_{\varepsilon^{2/3}v}}{\varepsilon^{1/3}}\right)^2 dv = \frac{t_i^\varepsilon - t_{i-1}^\varepsilon}{\varepsilon^{2/3}}\left(\frac{\big(D^+_{t_{i-1}^\varepsilon}\big)^2}{3\varepsilon^{2/3}} + o(1)\right).$$

Combining this with (A.13) and (A.14) and letting the mesh size go to zero, we obtain the assertion:

$$\int_0^T \Delta\varphi_u^2 d\langle S, S\rangle_u = \int_0^T D_u^2 du = \frac{1}{3}\int_0^T \big(D_u^+\big)^2 du + o(\varepsilon^{2/3}).$$

REFERENCES

BARLES, G., and H. M. SONER (1998): Option Pricing with Transaction Costs and a Nonlinear Black-Scholes Equation, *Finance Stoch.* 2(4), 369–397.

BECHERER, D. (2005): Bounded Solutions to Backward SDE's with Jumps for Utility Optimization and Indifference Hedging, *Ann. Appl. Probab.* 15(3), 2113–2143.

BICHUCH, M. (2011): Pricing a Contingent Claim Liability Using Asymptotic Analysis for Optimal Investment in Finite Time with Transaction costs, Preprint, arXiv:1112.3012.

BJÖRK, T. (2003): *Arbitrage Theory in Continuous Time*, 2nd ed., Oxford:Oxford University Press.

BORODIN, A. N., and P. SALMINEN (2002): *Handbook of Brownian Motion—Facts and Formulae*, 2nd ed., Basel: Birkhäuser Verlag.

CONSTANTINIDES, G. (1986): Capital Market Equilibrium with Transaction Costs, *J. Polit. Economy* 94(4), 842–862.

CVITANIĆ, J., and I. KARATZAS (1996): Hedging and Portfolio Optimization Under Transaction Costs: A Martingale Approach, *Math. Finance* 6(2), 133–165.

DAVIS, M. H. A., and A. R. NORMAN (1990): Portfolio Selection with Transaction Costs, *Math. Oper. Res.* 15(4), 676–713.

DAVIS, M. H. A., V. G. PANAS, and T. ZARIPHOPOULOU (1993): European Option Pricing with Transaction Costs, *SIAM J. Control Optim.* 31(2), 470–493.

DELBAEN, F., P. GRANDITS, T. RHEINLÄNDER, D. SAMPERI, M. SCHWEIZER, and C. STRICKER (2002): Exponential Hedging and Entropic Penalties, *Math. Finance* 12(2), 99–123.

GOODMAN, J., and D. N. OSTROV (2010): Balancing Small Transaction Costs with Loss of Optimal Allocation in Dynamic Stock Trading Strategies, *SIAM J. Appl. Math.* 70(6), 1977–1998.

GOODMAN, J., and D. OSTROV (2011): An Option to Reduce Transaction Costs, *SIAM J. Financ. Math.* 2(1), 512–537.

GUASONI, P., and J. MUHLE-KARBE (2015): Long Horizons, High Risk Aversion, and Endogeneous Spreads, *Math. Finance* 25(4), 724–753.

HODGES, S., and A. NEUBERGER (1989): Optimal Replication of Contingent Claims Under Transaction Costs, *Rev. Futures Markets* 8, 222–239.

JOUINI, E., and H. KALLAL (1995): Martingales and Arbitrage in Securities Markets with Transaction Costs, *J. Econ. Theory* 66(1), 178–197.

KALLSEN, J., and J. MUHLE-KARBE (2010): On Using Shadow Prices in Portfolio Optimization with Transaction Costs, *Ann. Appl. Probab.* 20(4), 1341–1358.

KALLSEN, J., and J. MUHLE-KARBE (2013): The General Structure of Optimal Investment and Consumption with Small Transaction Costs, Preprint, arXiv:1303.3148.

KALLSEN, J., and T. RHEINLÄNDER (2011): Asymptotic Utility-Based Pricing and Hedging for Exponential Utility, *Statist. Decisions* 28(1), 17–36.

LELAND, H. E. (1985): Option Pricing and Replication with Transaction Costs, *J. Finance* 40(5), 1283–1301.

LIU, H., and M. LOEWENSTEIN (2002): Optimal Portfolio Selection with Transaction Costs and Finite Horizons, *Rev. Financ. Stud.* 15(3), 805–835.

LOEWENSTEIN, M. (2000): On Optimal Portfolio Trading Strategies for an Investor Facing Transactions Costs in a Continuous Trading Market, *J. Math. Econ.* 33(2), 209–228.

MANIA, M., and M. SCHWEIZER (2006): Dynamic Exponential Utility Indifference Valuation, *Ann. Appl. Probab.* 16(4), 2027–2054.

MARTIN, R. (2011): Optimal Multifactor Trading Under Proportional Transaction Costs, Preprint, arXiv:1204.6488.

OWEN, M. P. (2002): Optimal Replication of Contingent Claims Under Transaction Costs, *Ann. Appl. Probab.* 12(2), 691–709.

ROGERS, L. C. G. (2004): Why Is the Effect of Proportional Transaction Costs $O(\delta^{2/3})$?, in *Mathematics of Finance*, G. Yin and Q. Zhang, eds. Providence, RI: Amer. Math. Soc., pp. 303–308.

SHREVE, S. E., and H. M. SONER (1994): Optimal Investment and Consumption with Transaction Costs, *Ann. Appl. Probab.* 4(3), 609–692.

SONER, H. M., S. E. SHREVE, and J. CVITANIĆ (1995): There Is No Nontrivial Hedging Portfolio for Option Pricing with Transaction Costs, *Ann. Appl. Probab.* 5(2), 327–355.

SONER, H. M., and N. TOUZI (2012): Homogenization and Asymptotics for Small Transaction Costs, Preprint. Available at: http://www.cmap.polytechnique.fr/∼touzi/Soner-Touzi-march13.pdf. Accessed April 23, 2013.

WHALLEY, A. E., and P. WILMOTT (1997): An Asymptotic Analysis of an Optimal Hedging Model for Option Pricing with Transaction Costs, *Math. Finance* 7(3), 307–324.

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

Mathematical Finance, Vol. 25, No. 4 (October 2015), 724–753

LONG HORIZONS, HIGH RISK AVERSION, AND ENDOGENOUS SPREADS

PAOLO GUASONI

Boston University and Dublin City University

JOHANNES MUHLE-KARBE

ETH Zürich

For an investor with constant absolute risk aversion and a long horizon, who trades in a market with constant investment opportunities and small proportional transaction costs, we obtain explicitly the optimal investment policy, its implied welfare, liquidity premium, and trading volume. We identify these quantities as the limits of their isoelastic counterparts for high levels of risk aversion. The results are robust with respect to finite horizons, and extend to multiple uncorrelated risky assets. In this setting, we study a Stackelberg equilibrium, led by a risk-neutral, monopolistic market maker who sets the spread as to maximize profits. The resulting endogenous spread depends on investment opportunities only, and is of the order of a few percentage points for realistic parameter values.

KEY WORDS: transaction costs, long-run portfolio choice, exponential utility, trading volume

1. INTRODUCTION

Despite their singular behavior, investors with constant absolute risk aversion are familiar figures in financial economics, thanks to their tractable character. Such investors, defined by exponential utility functions, are indeed peculiar for both portfolios and prices. With constant investment opportunities, they insist on keeping in risky assets a fixed amount of money, regardless of their wealth. If asked to price a claim, their answer depends neither on wealth, nor on risk aversion. Consuming over time, they do not disdain negative consumption, especially at later dates. Yet, their actions are often easier to grasp than the proper but impenetrable behavior of isoelastic investors.[1] Hence, exponential utility remains a central tool to glean insights from complex models, such as the ones with frictions.

This paper examines the implications of exponential utility for long-run portfolio choice with small transaction costs. For an exponential investor, with constant investment

We thank Ales Černý, Stefan Gerhold, Marcel Nutz, and Walter Schachermayer for helpful discussions. We also thank two anonymous referees and Ren Liu for their careful reading of the paper. Part of this work was completed while the second author was visiting Columbia University. He thanks Ioannis Karatzas and the university for their hospitality. Guasoni is supported by the ERC (278295), NSF (DMS-1109047), SFI (07/MI/008, 07/SK/M1189, 08/SRC/FMC1389), and FP7 (RG-248896). Muhle-Karbe is supported by the National Centre of Competence in Research "Financial Valuation and Risk Management" (NCCR FINRISK), Project D1 (Mathematical Methods in Financial Risk Management), of the Swiss National Science Foundation (SNF).

Manuscript received October 2011; final revision received March 2013.

Address correspondence to Paolo Guasoni, Department of Mathematics and Statistics, Boston University, 111 Cummington Mall, Boston, MA 02215, USA; e-mail: guasoni@bu.edu.

[1] An isoelastic investor is one with constant *relative* risk aversion, i.e., with either power or logarithmic utility.

DOI: 10.1111/mafi.12046

opportunities and a long planning horizon, we find the optimal trading policy, its welfare, the liquidity premium, and trading volume. We then allow a risk-neutral, monopolistic market maker to set the spread as to maximize profits, obtaining an endogenous spread that depends on investment opportunities alone.

Our analysis leads to new economic implications, and sheds new light on existing results. We find that investing optimally with a long horizon is equivalent to receiving, over the same period, a fixed *equivalent annuity*, found explicitly, which does not depend on the horizon. This fact is well known in the frictionless case, but fails for transaction costs with a finite horizon, due to the spurious effects of portfolio setup and liquidation. As in the isoelastic case, transaction costs entail a small reduction in the equivalent annuity.

The equivalent annuity, trading boundaries, and absolute turnover are inversely proportional to risk aversion, as in the frictionless case, while relative turnover and the liquidity premium are independent of risk aversion. We identify all these quantities as the limits of their isoelastic counterparts, as relative risk aversion becomes large. This result suggests that exponential utility is a useful tool to study isoelastic investors with high risk aversion.

Our results are robust to finite horizons and to several assets. For a finite horizon, we derive bounds on the investor's certainty equivalent, the time-average of which converges to the equivalent annuity. In a market with several uncorrelated assets, one-dimensional trading policies remain optimal, leading to the same liquidity premia and trading volumes, while equivalent annuities add in the cross-section.

Finally, we endogenize the spread by allowing a risk-neutral, monopolistic market maker to fix it to maximize expected profits. Unlike the investor, whose policy and welfare depend on the bid and ask prices alone, the market maker's profits depend on the book price, which lies within the bid-ask spread, and represents the price at which the market maker's inventory is valued. The resulting endogenous spread is independent of risk aversion, and hence depends only on investment opportunities, and on the book price. When the latter is chosen close to the ask price, realistic values of investment opportunities lead to spreads of a few percentage points.

Our results also have novel mathematical implications. We obtain new finite-horizon bounds, which measure the monetary value of investment opportunities for an exponential investor, and are expressed in terms of the risk-neutral probability. For isoelastic investors, such bounds involve the myopic probability, under which a hypothetical logarithmic investor adopts the same policy as the original investor under the physical probability. Thus, for exponential investors, the risk-neutral probability plays a similar role as the myopic probability for isoelastic investors. This analogy is central to obtain a new kind of verification theorem, which stems from the finite-horizon bounds.

Our paper touches on three main strands of literature: asymptotics, shadow prices, and exponential utility. Our contribution to asymptotics is to prove the first rigorous expansions for small transaction costs and exponential utility, complementing the heuristics of Whalley and Wilmott (1997), Mokkhavesa and Atkinson (2002), and Goodman and Ostrov (2010), as well as results for the isoelastic case, cf. Shreve and Soner (1994), Rogers (2004), Janeček and Shreve (2004), Gerhold, Muhle-Karbe, and Schachermayer (2012, 2013), and Bichuch (2012). Our solution is based on shadow prices, as in the recent papers Kallsen and Muhle-Karbe (2010), Gerhold et al. (2012, 2013), Gerhold et al. (2013), Herczegh and Prokaj (2011), and Choi, Sirbu, and Žitković (2012).

More broadly, our results are relevant for the literature on exponential utility with transaction costs, both in the context of portfolio choice (Mokkhavesa and Atkinson 2002; Liu 2004; Goodman and Ostrov 2010) and of option pricing (Davis, Panas, and

Zariphopoulou 1993; Whalley and Wilmott 1997; Barles and Soner 1998). In contrast to these papers, we remove consumption and random endowment from our model, focusing instead on long-horizon asymptotics for tractability. This approach is common for isoelastic utilities (Taksar, Klass, and Assaf 1988; Dumas and Luciano 1991; Gerhold et al. 2013), but, unaccountably, it seems unexplored in the exponential class. Our analysis of endogenous spreads is closest to the work of Luciano (2012) on equilibrium between isoelastic investors and dealers, with the difference that we pair an exponential investor with a risk-neutral dealer.

Finally, this paper on constant *absolute* risk aversion complements the analysis of constant *relative* risk aversion in Gerhold et al. (2013) (henceforth GGMKS). To facilitate comparison, the main results in both papers are stated in the same format and with the same notation. Yet, each paper addresses a different set of implications, which is specific to the preference class considered.

The rest of the paper is organized as follows. Section 2 presents the model and our main results. Their main implications are discussed in Section 3. The main results are first derived informally in Section 4, then proved in Section 5.

2. MODEL AND MAIN RESULT

2.1. Market

The market has a safe asset $S^0_t = 1$ and a risky asset, trading at ask (buying) price S. The bid (selling) price is $S(1-\varepsilon)$, hence $\varepsilon \in (0,1)$ is the relative bid-ask spread. Denoting by W a standard Brownian motion, the ask price S follows

$$dS_t/S_t = \mu dt + \sigma dW_t,$$

where $\mu > 0$ is the expected excess return and $\sigma > 0$ is the volatility. The *mean–variance ratio* $\bar{\mu} = \mu/\sigma^2$ turns out to be a key parameter in the solution.

A self-financing *trading strategy* is an $\mathbb{R}^2$-valued predictable process (φ^0, φ) of finite variation: $(\varphi^0_{0-}, \varphi_{0-}) = (\xi^0, \xi) \in \mathbb{R}^2$ represents the initial positions (in units) in the safe and risky asset, and φ^0_t and φ_t denote the positions held at time $t \geq 0$. Writing $\varphi_t = \varphi^\uparrow_t - \varphi^\downarrow_t$ as the difference between the cumulative number of shares bought ($\varphi^\uparrow_t$) and sold ($\varphi^\downarrow_t$) by time t, the *self-financing condition* dictates that the cash balance φ^0 changes only due to trading activity in the number of shares φ:

$$d\varphi^0_t = -S_t d\varphi^\uparrow_t + (1-\varepsilon) S_t d\varphi^\downarrow_t.$$

Trading strategies are further restricted from unlimited borrowing by the following admissibility condition, which rules out doubling strategies: [2]

DEFINITION 2.1. A self-financing strategy (φ^0, φ) is admissible if the risky position φS is a.s. uniformly bounded.

The *liquidation value* of the wealth associated to an admissible strategy is denoted by

$$\Xi^\varphi_T = \varphi^0_T + \varphi^+_T(1-\varepsilon)S_T - \varphi^-_T S_T.$$

[2]The definition of admissibility given here is sufficient to guarantee an interior solution in our setting, and has a clear interpretation. In more abstract models, admissibility is typically defined in terms of pricing measures, cf. Delbaen et al. (2002), Kabanov and Stricker (2002), and Schachermayer (2003).

2.2. Preferences

An investor with constant absolute risk aversion $\alpha > 0$, which corresponds to the exponential utility function $U(x) = -e^{-\alpha x}$, maximizes the certainty equivalent $U^{-1}(E[U(\Xi)])$, which for exponential utility reduces to:

$$-\frac{1}{\alpha}\log E\left[e^{-\alpha \Xi_T^{\varphi}}\right].$$

In a market with constant investment opportunities, the certainty equivalent increases linearly with the horizon. Hence, we focus on the certainty equivalent per unit of time, which is the same as a fixed annuity.

DEFINITION 2.2. A strategy (φ^0, φ) is long-run optimal if it maximizes the equivalent annuity

$$EA_\alpha^{\varphi} = \liminf_{T\to\infty} -\frac{1}{\alpha T}\log\ E\left[e^{-\alpha \Xi_T^{\varphi}}\right]. \tag{2.1}$$

$EA_\alpha = \max_\varphi EA_\alpha^{\varphi}$ denotes the maximal equivalent annuity.

2.3. Main Result

The next theorem contains our main results. Recall that $\bar{\mu} = \mu/\sigma^2$ is the mean–variance ratio.

THEOREM 2.3. *An investor with constant absolute risk aversion $\alpha > 0$ trades to maximize the equivalent annuity. Then, for a small bid-ask spread $\varepsilon > 0$:*

(i) (Equivalent Annuity) For the investor, trading the risky asset with transaction costs is equivalent to leaving all wealth in the safe asset, while receiving the equivalent annuity (the gap $\bar{\lambda}$ is defined in (iv) below):

$$EA_\alpha = \sigma^2\bar{\beta} = \frac{\sigma^2}{2\alpha}(\bar{\mu}^2 - \bar{\lambda}^2).$$

(ii) (Liquidity Premium) Trading the risky asset with transaction costs is equivalent to trading a hypothetical asset, at no transaction costs, with the same volatility σ, but with a lower expected excess return $\sigma^2\sqrt{\bar{\mu}^2 - \bar{\lambda}^2}$. Thus, the liquidity premium is

$$Li\,Pr = \sigma^2\left(\bar{\mu} - \sqrt{\bar{\mu}^2 - \bar{\lambda}^2}\right).$$

(iii) (Trading Policy) It is optimal to keep the value of the risky position within the buy and sell boundaries

$$\eta_{\alpha-} = \frac{\bar{\mu} - \bar{\lambda}}{\alpha}, \qquad \eta_{\alpha+} = \frac{\bar{\mu} + \bar{\lambda}}{\alpha},$$

where $\eta_{\alpha-}$ and $\eta_{\alpha+}$ are evaluated at ask and bid prices, respectively.

(iv) (Gap) The gap $\bar{\lambda}$ is the unique value for which the solution of the initial value problem

$$w'(y) - w(y)^2 + (2\bar{\mu} - 1)\,w(y) - (\bar{\mu} - \bar{\lambda})(\bar{\mu} + \bar{\lambda}) = 0$$
$$w(0) = \bar{\mu} - \bar{\lambda},$$

also satisfies the terminal value condition

$$w\left(\log\left(\frac{u(\bar\lambda)}{l(\bar\lambda)}\right)\right) = \bar\mu + \bar\lambda, \qquad \textit{where} \qquad u(\bar\lambda) = \frac{\bar\mu + \bar\lambda}{(1-\varepsilon)\alpha} \quad \textit{and} \quad l(\bar\lambda) = \frac{\bar\mu - \bar\lambda}{\alpha}.$$

In view of the explicit formula for $w(y, \bar\lambda)$ in Lemma 5.1 below, this is a scalar equation for $\bar\lambda$.

(v) (Trading Volume) Let $\mu \neq \sigma^2/2$.[3] *Then, relative turnover, defined as shares traded $d||\varphi||_t$ divided by shares held $|\varphi_t|$, has long-term average*

$$ShTu = \lim_{T\to\infty} \frac{1}{T}\int_0^T \frac{d\|\varphi\|_t}{|\varphi_t|} = \frac{\sigma^2}{2}\left(\frac{1-2\bar\mu}{(u(\bar\lambda)/l(\bar\lambda))^{1-2\bar\mu}-1} + \frac{2\bar\mu - 1}{(u(\bar\lambda)/l(\bar\lambda))^{2\bar\mu-1}-1}\right).$$

Absolute turnover, defined as value of wealth traded, has long-term average:[4]

$$\begin{aligned} WeTu_\alpha &= \lim_{T\to\infty}\frac{1}{T}\left(\int_0^T (1-\varepsilon)S_t d\varphi_t^\downarrow + \int_0^T S_t d\varphi_t^\uparrow\right) \\ &= \frac{\sigma^2}{2}\left(\frac{\eta_{\alpha+}(1-2\bar\mu)}{(u(\bar\lambda)/l(\bar\lambda))^{1-2\bar\mu}-1} + \frac{\eta_{\alpha-}(2\bar\mu-1)}{(u(\bar\lambda)/l(\bar\lambda))^{2\bar\mu-1}-1}\right). \end{aligned}$$

(vi) (Asymptotics) The following expansions in terms of the bid-ask spread ε hold:[5]

$$\begin{aligned} \bar\lambda &= \left(\frac{3}{4}\bar\mu^2\right)^{1/3}\varepsilon^{1/3} + O(\varepsilon), \\ EA_\alpha &= \frac{\sigma^2}{2\alpha}\left(\bar\mu^2 - \left(\frac{3}{4}\bar\mu^2\right)^{2/3}\varepsilon^{2/3} + O(\varepsilon^{4/3})\right), \\ LiPr &= \frac{\sigma^2}{2\bar\mu}\left(\frac{3}{3}\bar\mu^2\right)^{2/3}\varepsilon^{2/3} + O(\varepsilon^{4/3}), \\ \eta_{\alpha\pm} &= \frac{1}{\alpha}\left(\bar\mu \pm \left(\frac{3}{4}\bar\mu^2\right)^{1/3}\varepsilon^{1/3} + O(\varepsilon)\right), \\ ShTu &= \frac{\sigma^2}{2}\bar\mu\left(\frac{3}{4}\bar\mu^2\right)^{-1/3}\varepsilon^{-1/3} + O(\varepsilon^{1/3}), \\ WeTu_\alpha &= \frac{2\sigma^2}{3\alpha}\left(\frac{3}{4}\bar\mu^2\right)^{2/3}\varepsilon^{-1/3} + O(\varepsilon^{1/3}). \end{aligned}$$

The proof of Theorem 2.3 exploits the construction of a shadow price, i.e., a fictitious risky asset evolving within the bid-ask spread, which is equivalent to the transaction cost market in terms of both welfare *and* the optimal policy. This is the approach used for power utility by Gerhold et al. (2013).

[3]The special case $\mu = \sigma^2/2$ leads to analogous results, see GGMKS, lemma D.2.

[4]The number of shares is written as the difference $\varphi_t = \varphi_t^\uparrow - \varphi_t^\downarrow$ of cumulative shares bought and sold, and wealth is evaluated at trading prices, i.e., at the bid price $(1-\varepsilon)S_t$ when selling, and at the ask price S_t when buying.

[5]Algorithmic calculations can deliver terms of arbitrarily high order.

THEOREM 2.4 (Shadow price). *The policy in Theorem 2.3 (iii) and the equivalent annuity in Theorem 2.3 (i) are also optimal for a frictionless risky asset with shadow price $\tilde{S}$, which always lies within the bid-ask spread, and coincides with the trading price at times of trading for the optimal policy. The shadow price follows*

$$d\tilde{S}_t/\tilde{S}_t = \tilde{\mu}(\Upsilon_t)dt + \tilde{\sigma}(\Upsilon_t)dW_t,$$

for deterministic functions $\tilde{\mu}(\cdot)$ and $\tilde{\sigma}(\cdot)$ given explicitly in Lemma 5.6. The state variable $\Upsilon_t = \log(\varphi_t S_t/l(\bar{\lambda}))$ represents the centered logarithm of the risky position, which is a Brownian motion with drift reflected to remain in the interval $[0, \log(u(\bar{\lambda})/l(\bar{\lambda}))]$, i.e.,

$$d\Upsilon_t = (\mu - \sigma^2/2)dt + \sigma dW_t + dL_t - dU_t.$$

Here, L_t and U_t are nondecreasing processes, corresponding to the relative purchases and sales, respectively (cf. Lemma 5.9).[6] In the interior of the no-trade region, i.e., when the risky position lies in $(l(\bar{\lambda}), u(\bar{\lambda}))$, the numbers of units of the safe and risky asset are constant, and the state variable Υ_t follows Brownian motion with drift. When Υ_t reaches the boundary of the no-trade region, buying or selling takes place so as to keep it within $[l(\lambda), u(\lambda)]$.

3. IMPLICATIONS AND APPLICATIONS

Theorem 2.3 presents both analogies with and departures from the isoelastic case in GGMKS. One analogy is that the trading policy depends on the market only through the mean–variance ratio $\bar{\mu}$. All other quantities also only depend on $\bar{\mu}$ in *business* time, i.e., when measured with respect to the clock $\tau = \sigma^2 t$ running at the speed of the market's variance. Measured in usual calendar time t, they scale linearly with the market's variance σ^2.

In a departure from power utility, the gap $\bar{\lambda}$ is independent of the investor's risk aversion α here. Hence, the liquidity premium and relative turnover are also common to all investors with constant absolute risk aversion. The equivalent annuity and trading boundaries, however, are inversely proportional to risk aversion, as in the frictionless case, and so is absolute turnover.

Further, in this model a strategy of full investment in the risky asset is never optimal, regardless of the risk aversion α, and the only solution without trading obtains with a null risk premium $\bar{\mu} = 0$. The absence of full risky investment is a consequence of constant trading boundaries in terms of monetary amounts rather than fractions of wealth.

In spite of these differences, the model- and preference-free relationships for power utilities (GGMKS, section 3.5) carry over to the exponential case. For example, the universal relation

$$Li\,Pr \approx \frac{3}{4}\varepsilon\, Sh\,Tu,$$

between the liquidity premium $Li\,Pr$, the spread ε, and the relative turnover $Sh\,Tu$ remains valid for exponential utilities. Thus, this relation not only holds both across market and preference parameters, but is also the same for both families.

[6]The increasing processes L_t and U_t are explicitly identified by the double Skorokhod map in a finite interval, see Kruk et al. (2007).

3.1. Trading Policy and Average Risky Position

In analogy with the isoelastic case, the optimal policy is to keep the risky position between two boundaries. Unlike the isoelastic case, but as in the frictionless setting, trading boundaries are measured not as fractions of wealth, but as monetary amounts. The novelty is that these boundaries are symmetric around the frictionless optimum. In apparent contrast, Liu (2004) finds numerically that average risky holdings (over the investment period) increase with transaction costs, suggesting that more transaction costs reduce an investor's *effective* risk aversion. In fact, our results are consistent with his findings, once we observe that, if the risk premium is large enough, the risky position is on average closer to the sell than to the buy boundary, even as the boundaries are equidistant from the frictionless solution at the order $\varepsilon^{2/3}$. To see this, it suffices to calculate the average risky position, using the stationary distribution of the reflected geometric Brownian motion $Y_t = \varphi_t S_t = l e^{\Upsilon_t}$:

$$\begin{aligned}\lim_{T\to\infty} \frac{1}{T}\int_0^T Y_t dt &= \int_0^{\log u/l} l e^y \frac{2\bar{\mu}-1}{(u(\bar{\lambda})/l(\bar{\lambda}))^{2\bar{\mu}-1}-1} e^{(2\bar{\mu}-1)y} dy \\ &= \frac{\bar{\mu}-\frac{1}{2}}{\bar{\mu}} \frac{(u(\bar{\lambda})/l(\bar{\lambda}))^{2\bar{\mu}}-1}{(u(\bar{\lambda})/l(\bar{\lambda}))^{2\bar{\mu}-1}-1}.\end{aligned}$$

To obtain an asymptotic expansion for small transaction costs, recall that $l(\bar{\lambda}) = (\bar{\mu} - \bar{\lambda})/\alpha$ and $u(\bar{\lambda}) = (\bar{\mu} + \bar{\lambda})/(1-\varepsilon)\alpha$. Then, the expansion for $\bar{\lambda}$ yields:

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T Y_t dt = \frac{\bar{\mu}}{\alpha}\left(1 + \frac{\bar{\mu}-1}{6^{1/3}\bar{\mu}^{2/3}}\varepsilon^{2/3} + O(\varepsilon)\right). \tag{3.1}$$

As a result, for $\bar{\mu} > 1$ the average risky position tends to be higher than the frictionless value $\bar{\mu}/\alpha$, and vice versa for $\bar{\mu} < 1$. This effect is entirely due to the skewness of the stationary distribution toward the upper boundary, because the boundaries $(\bar{\mu} \pm \lambda)/\alpha$ are symmetric around $\bar{\mu}/\alpha$ at the order $\varepsilon^{2/3}$. Thus, the estimator $\hat{\alpha}$ obtained by comparing the average risky holding in (3.1) to the frictionless formula $\bar{\mu}/\hat{\alpha}$ is given by:

$$\hat{\alpha} = \alpha\left(1 - \frac{\bar{\mu}-1}{6^{1/3}\bar{\mu}^{2/3}}\varepsilon^{2/3} + O(\varepsilon)\right). \tag{3.2}$$

This estimator underestimates true risk aversion for $\bar{\mu} > 1$, with larger transaction costs leading to a larger bias, and explains the observation of Liu (2004) that transaction costs seem to reduce effective risk aversion.

3.2. Convergence and High Risk-Aversion Asymptotics

The formulae in Theorem 2.3 for exponential utilities are closely related to the limits of the corresponding isoelastic quantities in GGMKS. The next result makes this relation precise, and justifies the dual interpretation of the aforementioned formulas as high risk-aversion asymptotics for isoelastic investors, in the same spirit as in Černý (2009) and Nutz (2012).

THEOREM 3.1 (High risk-aversion asymptotics). *An investor with constant relative risk aversion $\gamma > 0$ trades to achieve the maximal equivalent safe rate*

$ESR_\gamma = \max_\varphi ESR_\gamma(\varphi)$, *where*

$$ESR_\gamma(\varphi) = \lim_{T\to\infty} \frac{1}{T} \log E\left[(\Xi_T^\varphi)^{1-\gamma}\right]^{\frac{1}{1-\gamma}}.$$

Denote by $\bar{\lambda}_\gamma$,[7] $Li\,Pr_\gamma$, $\pi_{\gamma\pm}$, $ShTu_\gamma$, *and* $WeTu_\gamma$ *the corresponding gap, liquidity premium, trading boundaries (as wealth fractions), share turnover, and (relative) wealth turnover (see GGMKS for details). Then, as* $\gamma \uparrow \infty$, *the following properties hold for a small spread* $\varepsilon > 0$*:*

(i) The equivalent safe rate times relative risk aversion converges to the equivalent annuity times absolute risk aversion, i.e.,

$$\lim_{\gamma\uparrow\infty} \gamma\, ESR_\gamma = \alpha\, EA_\alpha = \frac{\sigma^2}{2}(\bar{\mu}^2 - \bar{\lambda}^2).$$

(ii) The gap $\bar{\lambda}_\gamma$ *converges to the gap* $\bar{\lambda}$.

(iii) The liquidity premium $Li\,Pr_\gamma$ *converges to* $Li\,Pr$.

(iv) The trading boundaries $\pi_{\gamma\pm}$ *as wealth fractions, times* γ, *converge to the trading boundaries* $\eta_{\alpha\pm}$, *as position values, times* α, *i.e.,*

$$\lim_{\gamma\uparrow\infty} \gamma\pi_{\gamma\pm} = \alpha\eta_{\alpha\pm}.$$

(v) Share turnover $ShTu_\gamma$ *converges to relative turnover* $ShTu$.

(vi) Relative wealth turnover, times γ, *converges to absolute turnover times* α.

This result clarifies some properties of the isoelastic quantities. For example GGMKS, section 3.4, share turnover converges to a finite limit as relative risk aversion increases. Theorem 2.3 identifies this limit as the relative turnover of exponential utility. By contrast, relative wealth turnover declines to zero as risk aversion increases but, once rescaled by γ, it converges to absolute turnover for exponential utility.

3.3. Trading Volume and Endogenous Spreads

An attraction of exponential utility is that, because the value of the investor's risky position is bounded, then also total rebalancing costs are bounded. These costs are, in turn, related to the profits of a market maker, who earns the costs paid by the investor. Thus, exponential utility is well-suited to develop a valuation model of a market maker, in which both trading and spreads are endogenous.

If the market maker acts as a monopolist, and fixes the spread to maximize profits, the model yields an endogenous optimal spread, which depends on investment opportunities only. The monopolist trade-off is clear: a larger spread increases the profit of each transaction, but reduces demand for trading.

Bid and ask prices alone are not sufficient to determine profits. What is missing is the "book" price $\bar{S}_t$, at which the market maker values his inventory.[8] Such a price

[7] This is the gap in business time, replacing μ and σ^2 with $\bar{\mu}$ and 1, respectively, in GGMKS, theorem 2.2.

[8] The book price also admits the interpretation of production cost of the asset for the monopolist market maker.

must lie within the bid-ask spread: whichever policy the investor chooses, the average execution price will be within the bid-ask spread. Thus, we denote the book value by $\bar{S}_t = S_t(1 - \varepsilon\delta)$, with $\delta = 0$ and $\delta = 1$ leading to the ask and bid prices, respectively. With this notation, the average profits are:

$$Profit_T = \int_0^T (S_t - \bar{S}_t) d\varphi_t^{\uparrow} + \int_0^T (\bar{S}_t - (1-\varepsilon)S_t) d\varphi_t^{\downarrow}. \tag{3.3}$$

In other words, while the investor is only sensitive to the final bid and ask prices, the market maker's profits depend separately on sales or purchases, depending on the choice of the book price.

Plugging $\bar{S}_t = S_t(1 - \varepsilon\delta)$ in the above expression, and passing to the limit as $T \uparrow \infty$, average profits become equal to:

$$Profit = \lim_{T\to\infty} \frac{1}{T} Profit_T = \varepsilon \left(\delta \, WeTu_{\alpha-} + \frac{1-\delta}{1-\varepsilon} \, WeTu_{\alpha+} \right), \tag{3.4}$$

where $WeTu_-$ and $WeTu_+$ denote the expressions for expected purchases and sales, which add to absolute turnover in Theorem 2.3:

$$WeTu_{\alpha-} = \frac{\sigma^2}{2} \frac{\eta_{\alpha-}(2\bar{\mu} - 1)}{(u(\bar{\lambda})/l(\bar{\lambda}))^{2\bar{\mu}-1} - 1}, \qquad WeTu_{\alpha+} = \frac{\sigma^2}{2} \frac{\eta_{\alpha+}(1 - \bar{\mu})}{(u(\bar{\lambda})/l(\bar{\lambda}))^{1-2\bar{\mu}} - 1}.$$

Figure 3.1 shows the market maker's expected profit, as a function of the spread ε, for $\delta = 1, 0.5, 0$. When profits are concentrated on purchases ($\delta = 1$), the market maker optimally sets the spread in the range of 3–4%. This value is high compared to the spreads currently observed in US and European equities, in which market makers are no longer monopolists. However, such a figure is typical of the spreads observed on small capitalization stocks (Amihud and Mendelson 1986), which are traded by fewer dealers.

By contrast, the model leads to unreasonably high spreads, well above 20%, if the profits are split equally between sales and purchases, or concentrated on sales ($\delta = 0.5$ or 0). This result, if counterintuitive initially, is implied by the asymmetry between expected sales ($\delta = 0$) and expected purchases ($\delta = 1$). Because the exponential investor wants to keep the risky position approximately fixed, and because the risky asset grows on average, the investor steadily realizes past gains over time, as the risky position reaches the selling boundary. By contrast, purchases occur after large drops in the asset price, which are much less frequent.

This observation in fact hints at a weakness of the model, the absence of dividends, which in practice allow the investor to avoid selling shares, cashing dividends instead. Thus, using a value of δ close to 1 is a convenient assumption, which remedies in part the absence of dividends in the model.

3.4. Finite-Horizon Bounds: Myopic Probability as Risk Neutral

A novel aspect of our analysis is the derivation of finite-horizon bounds (Guasoni and Robertson 2012) in the context of exponential utility. These bounds offer estimates on the performance of the candidate's long-run optimal policy on any intermediate horizon T. They are both a mathematical device to prove the verification theorem, and a diagnostic tool to determine at which horizons the long-run optimal policy is effective enough.

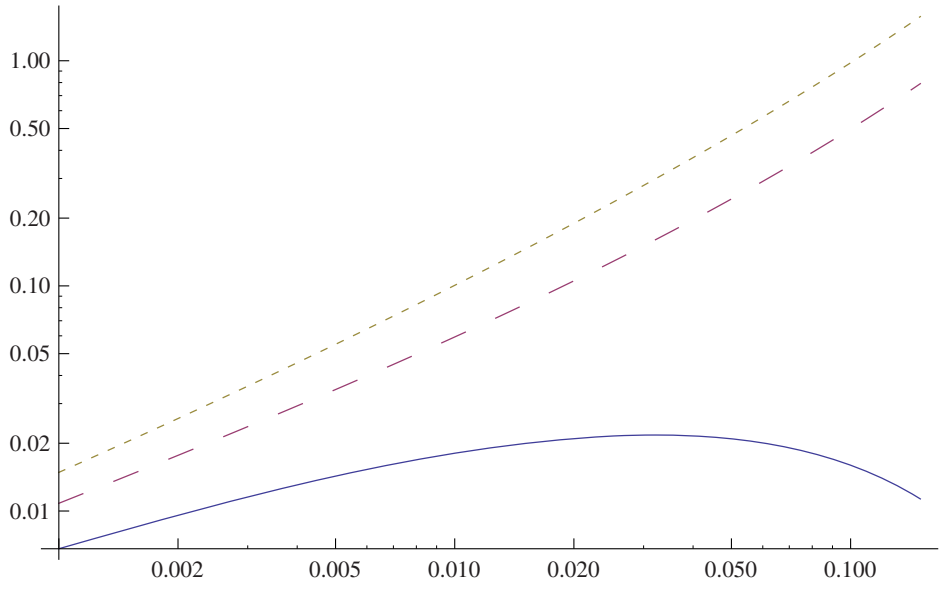

FIGURE 3.1. Expected value of future fees (vertical axis, in dollars) against the spread ε (horizontal axis). The plot compares the case of fees earned 100% on purchases (solid line), 50% on purchases and sales (long dashing), and 100% on sales (short dashing). Parameters are $\mu = 8\%, \sigma = 16\%$, and $\alpha = (\mu/\sigma^2)/100$, corresponding to a frictionless position of $\eta = \mu/\alpha\sigma^2 = 100$ dollars, and both axes are in logarithmic scale.

For exponential utility, finite-horizon bounds admit an especially appealing form—in monetary units. As for isoelastic utilities, the respective asymptotics show that—at the first order—our stationary long-run policy is also optimal for any fixed finite horizon $T > 0$, because the corresponding utility matches the finite-horizon value function at the first nontrivial order.

THEOREM 3.2. *(Finite-horizon bounds) For any horizon $T > 0$, the payoff $\tilde{X}_T^{\phi}$ of a generic admissible strategy ϕ in the frictionless shadow market $\tilde{S}$ satisfies*

$$-\frac{1}{\alpha}\log E\left[e^{-\alpha \tilde{X}_T^{\phi}}\right] \leq \tilde{X}_0^{\phi} + \sigma^2\bar{\beta}T + \frac{1}{\alpha}\tilde{E}\left[\tilde{q}(\Upsilon_0) - \tilde{q}(\Upsilon_T)\right] = \tilde{X}_0^{\phi} + \sigma^2\bar{\beta}T + O\left(\frac{\varepsilon}{\alpha}\right).$$

For the shadow payoff corresponding to the long-run optimal strategy φ from Theorem 2.3,

$$-\frac{1}{\alpha}\log E\left[e^{-\alpha \tilde{X}_T^{\varphi}}\right] = \tilde{X}_0^{\varphi} + \sigma^2\bar{\beta}T - \frac{1}{\alpha}\log \tilde{E}[e^{\tilde{q}(\Upsilon_T)-\tilde{q}(\Upsilon_0)}] = \tilde{X}_0^{\varphi} + \sigma^2\bar{\beta}T + O\left(\frac{\varepsilon}{\alpha}\right).$$

Here $\tilde{E}[\cdot]$ denotes the expectation with respect to the unique risk-neutral probability for $\tilde{S}$, Υ the centered logarithm of the risky position in Theorem 2.4, and $\tilde{q}$ the deterministic function defined in Lemma 5.7 below.

A new mathematical insight of this theorem is that, with exponential utility, the risk neutral probability replaces the myopic probability in the finite-horizon bounds. By definition, under the myopic probability the logarithmic policy coincides with the optimal

policy under the real probability. Thus, strictly speaking, a risk-neutral probability is never myopic, because a logarithmic investor (in fact, any investor) takes a zero position in a risky asset with null return.

Yet, the finite-horizon bounds in the theorem above involve expectations under the risk-neutral probability, just as the similar bounds in Guasoni and Robertson (2012) involve expectations under the myopic probability. The intuition is that, as relative risk aversion increases, the risky weight in the optimal isoelastic policy decreases to zero, and so does the drift under the myopic probability. Even as the weight decreases to zero, the monetary position can converge to a finite amount, as it happens for exponential utility.

Duality theory sheds further light on the bounds. For any payoff X and any stochastic discount factor (or martingale density) $M = dQ/dP$, Jensen's inequality implies that:

$$-\frac{1}{\alpha}\log E\left[e^{-\alpha X}\right] = -\frac{1}{\alpha}\log E_Q\left[e^{-\alpha X-\log M}\right] \leq E_Q[X] + \frac{1}{\alpha}E[M\log M]. \tag{3.5}$$

For a fixed payoff X, this inequality still holds passing to the infimum over M, thereby indicating that equality may only hold when M has minimal entropy. This abstract inequality also shows that the minimal entropy is interpreted as a monetary certainty equivalent, which represents the opportunity value of trading in the market over a given horizon. In the statement in Theorem 3.2, this opportunity value decomposes into the integral term, which increases linearly with the horizon, and leads to the equivalent annuity, and to the transitory term with $\tilde{q}$, which oscillates with the relative position of the portfolio within the no-trade region.

3.5. Multiple Risky Assets

Consider a market with risky assets $S^1, \ldots, S^d$ following

$$dS_t^i/S_t^i = \mu_i dt + \sigma_i dW_t^i,$$

for excess returns $\mu_i > 0$, volatilities $\sigma_i > 0$, and *independent* standard Brownian motions W^i.

With exponential utility, the optimal policy in a market with several such independent risky assets entails no-trade regions for each asset, which coincide with the no-trade regions obtained for each risky asset alone, that is, in a market with a single risky asset (Liu 2004). In our setting, the following verification theorem applies, which allows to avoid the technical conditions in Liu (2004). The implication is that the equivalent annuity for multiple independent risky assets is the sum of the equivalent annuities for each asset.

THEOREM 3.3. *For $i = 1, \ldots, d$, let $\tilde{S}^i$ be the shadow price from Lemma 5.6, with corresponding optimal strategy $(\varphi^{0,i}, \varphi^i)$ from Lemma 5.8 in the market with safe asset S^0 and risky asset S^i. Then, $\tilde{S} = (\tilde{S}^1, \ldots, \tilde{S}^d)$ is a shadow price in the market with safe asset S^0 and risky assets $S^1, \ldots, S^d$, with optimal strategy $(\sum_{i=1}^d \varphi^{0,i}, \varphi^1, \ldots, \varphi^d)$ and corresponding equivalent annuity*

$$\mathrm{EA}_\alpha = \sum_{i=1}^d \mathrm{EA}_\alpha^i = \sum_{i=1}^d \frac{\sigma_i^2}{2\alpha}(\bar{\mu}_i^2 - \bar{\lambda}_i^2).$$

Here, $\bar{\mu}_i = \mu_i/\sigma_i^2$ and each gap $\bar{\lambda}_i$ is defined as in item (iv) of Theorem 2.3. Moreover, like the equivalent annuity, relative and absolute turnover also add across independent assets.

These decompositions are unique for exponential utility, and fail for utilities in the isoelastic class. For example, Akian, Menaldi, and Sulem (1996) show that, in a market with two identical and independent assets, the no-trade region for each asset is wider than the no-trade region for a market with that asset alone. As a result, the equivalent safe rate for the two-asset market is greater than the sum of the equivalent safe rates.

3.6. Safe Rate

Throughout the paper, we assume a zero safe rate. This choice is made in part to ease notation, and the results can be adapted to the case of a constant safe rate r, with an important caveat. For an exponential investor with a long horizon, an arbitrarily small safe rate r is preferable to *any* risky investment opportunity $\bar{\mu}$. The reason is that the optimal policy of this particularly risk-averse investor is to keep only a bounded amount of money in the risky asset, whence her wealth on average grows linearly over time. By contrast, a positive safe rate allows wealth to grow exponentially—without risk. Therefore, full investment in the safe asset is eventually preferred by the exponential investor in the long run.

Nevertheless, the finite-horizon bounds in Theorem 3.2 remain valid even for a positive safe rate. Indeed, setting $\hat{\alpha} = e^{rT}\alpha$ in Theorem 3.2 for *discounted* payoffs, it follows that the *undiscounted* payoff $\tilde{X}_T^{\phi}$ of any admissible strategy ϕ in the frictionless shadow market $\tilde{S}$ satisfies

$$-\frac{1}{\alpha}\log E\big[e^{-\alpha \tilde{X}_T^{\phi}}\big] \leq e^{rT}\tilde{X}_0^{\phi} + \sigma^2\bar{\beta}T + \frac{1}{\alpha}\tilde{E}[\tilde{q}(\Upsilon_0) - \tilde{q}(\Upsilon_T)] = e^{rT}\tilde{X}_0^{\phi} + \sigma^2\bar{\beta}T + O\Big(\frac{\varepsilon}{\alpha T}\Big).$$

Likewise, the shadow payoff of the long-run optimal strategy φ in Theorem 2.3 satisfies

$$-\frac{1}{\alpha}\log E\big[e^{-\alpha \tilde{X}_T^{\varphi}}\big] = e^{rT}\tilde{X}_0^{\varphi} + \sigma^2\bar{\beta}T - \frac{1}{\alpha}\log \tilde{E}[e^{\tilde{q}(\Upsilon_T)-\tilde{q}(\Upsilon_0)}] = e^{rT}\tilde{X}_0^{\varphi} + \sigma^2\bar{\beta}T + O\Big(\frac{\varepsilon}{\alpha T}\Big).$$

Hence, our long-run optimal policy still matches the finite-horizon value function up to terms of order $O(\varepsilon/T)$. However, as the horizon becomes large, the contribution of the risky investment grows only linearly with the horizon T. Hence it becomes negligible, as the certainty equivalent grows exponentially.

4. HEURISTICS

In this section, we first use informal arguments from stochastic control to determine a candidate for the optimal policy. Then, we derive a candidate shadow price process, which is key for the subsequent verification.

4.1. Optimal Policy

For a trading strategy (φ_t^0, φ_t), write the number of shares $\varphi = \varphi^{\uparrow} - \varphi^{\downarrow}$ as the difference of the cumulative numbers of shares purchased and sold, and denote by

$$X_t = \varphi_t^0, \quad Y_t = \varphi_t S_t,$$

the values of the safe and risky positions in terms of the ask price S_t. Then, the self-financing condition, and the dynamics of S imply

$$\begin{aligned} dX_t &= -S_t d\varphi_t^{\uparrow} + (1-\varepsilon) S_t d\varphi_t^{\downarrow}, \\ dY_t &= \mu Y_t dt + \sigma Y_t dW_t + S_t d\varphi_t^{\uparrow} - S_t d\varphi_t^{\downarrow}. \end{aligned}$$

Consider the problem of maximizing the expected exponential utility $U(x) = -e^{-\alpha x}$ from terminal wealth at time T. Denote by $V(t, x, y)$ its value function, which depends on time as well as the safe and risky positions. Itô's formula yields:

$$\begin{aligned} dV(t, X_t, Y_t) &= V_t dt + V_x dX_t + V_y dY_t + \frac{1}{2} V_{yy} d\langle Y, Y\rangle_t \\ &= \left(V_t + \mu Y_t V_y + \frac{\sigma^2}{2} Y_t^2 V_{yy} \right) dt \\ &\quad + S_t (V_y - V_x) d\varphi_t^{\uparrow} + S_t ((1-\varepsilon) V_x - V_y) d\varphi_t^{\downarrow} + \sigma Y_t V_y dW_t, \end{aligned}$$

where the arguments of the functions are omitted for brevity. Because $V(t, X_t, Y_t)$ must be a supermartingale for any choice of the cumulative purchases and sales $\varphi^{\uparrow}, \varphi^{\downarrow}$ (which are increasing processes), it follows that $V_y - V_x \le 0$ and $(1-\varepsilon) V_x - V_y \le 0$, that is,

$$1 \le \frac{V_x}{V_y} \le \frac{1}{1-\varepsilon}.$$

In the interior of this region, the drift of $V(t, X_t, Y_t)$ cannot be positive, and must become zero for the optimal policy,

$$V_t + \mu Y_t V_y + \frac{\sigma^2}{2} Y_t^2 V_{yy} = 0 \qquad \text{if} \qquad 1 < \frac{V_x}{V_y} < \frac{1}{1-\varepsilon}.$$

To simplify further, we use the usual scaling for exponential utility (cf., e.g., Davis et al. 1993). Moreover, in the long run, the value function should grow exponentially with the horizon at a constant rate. This leads to the following ansatz for the value function:

$$V(t, X_t, Y_t) = -e^{-\alpha X_t} e^{\alpha\sigma^2 \bar{\beta} t} v(Y_t), \tag{4.1}$$

which reduces the Hamilton–Jacobi–Bellman (HJB) equation to

$$\frac{1}{2} y^2 v''(y) + \bar{\mu} y v'(y) + \alpha \bar{\beta} v(y) = 0 \qquad \text{if} \qquad 1 < \frac{-\alpha v(y)}{v'(y)} < \frac{1}{1-\varepsilon}.$$

Conjecturing that the set $\{y : 1 < \frac{-\alpha v(y)}{v'(y)} < \frac{1}{1-\varepsilon}\}$ coincides with some interval $l < y < u$ to be determined, the following free boundary problem arises:

$$\frac{1}{2} y^2 v''(y) + \bar{\mu} y v'(y) + \alpha \bar{\beta} v(y) = 0 \qquad \text{if } l < y < u, \tag{4.2}$$

$$v'(l) + \alpha v(l) = 0, \tag{4.3}$$

$$(1/(1-\varepsilon)) v'(u) + \alpha v(u) = 0. \tag{4.4}$$

These conditions are not enough to identify the solution, because they can be matched for any choice of the trading boundaries l, u. The optimal boundaries are the ones that also satisfy the smooth pasting conditions (cf. Dumas 1991), formally obtained by differentiating (4.3) and (4.4) with respect to l and u, respectively:

$$v''(l) + \alpha v'(l) = 0, \tag{4.5}$$

$$(1/(1-\varepsilon))v''(u) + \alpha v'(u) = 0. \tag{4.6}$$

In addition to the reduced value function v, this system requires to solve for $\bar{\beta}$ (and hence the equivalent annuity $\sigma^2\bar{\beta}$) as well as the trading boundaries l and u. Substituting (4.5) and (4.3) into (4.2) yields

$$\frac{1}{2}(-\alpha)^2 l^2 v + \bar{\mu}(-\alpha) l v + \alpha\bar{\beta} v = 0.$$

Setting $\eta_{\alpha-} = l$, and factoring out $-\alpha v$, it follows that

$$-\frac{\alpha}{2}\eta_{\alpha-}^2 + \bar{\mu}\eta_{\alpha-} - \bar{\beta} = 0.$$

Note that $\eta_{\alpha-}$ is the risky position when it is time to buy, and hence the risky position is valued at the ask price. The same argument for u shows that the other solution of the quadratic equation is $\eta_{\alpha+} = u(1-\varepsilon)$, i.e., the risky position when it is time to sell, and hence the risky position is valued at the bid price. Thus, the optimal policy is to buy when the "ask" position falls below $\eta_{\alpha-}$, sell when the "bid" position rises above $\eta_{\alpha+}$, and do nothing in between. Since $\eta_{\alpha-}$ and $\eta_{\alpha+}$ solve the same quadratic equation, they are related to $\bar{\beta}$ via

$$\eta_{\alpha\pm} = \frac{\bar{\mu}}{\alpha} \pm \frac{\sqrt{\bar{\mu}^2 - 2\bar{\beta}\alpha}}{\alpha}.$$

It is convenient to set $\bar{\beta} = (\bar{\mu}^2 - \bar{\lambda}^2)/2\alpha$, because $\bar{\beta} = \bar{\mu}^2/2\alpha$ without transaction costs. With this notation, the buy and sell boundaries are just

$$\eta_{\alpha\pm} = \frac{\bar{\mu} \pm \bar{\lambda}}{\alpha}.$$

Now that $l(\bar{\lambda})$, $u(\bar{\lambda})$ are identified by $\eta_{\alpha\pm}$ in terms of $\bar{\lambda}$, it remains to find $\bar{\lambda}$. After deriving $l(\bar{\lambda})$ and $u(\bar{\lambda})$, the boundaries in the problem (4.2)–(4.4) are no longer free, but fixed. With the substitution

$$v(y) = e^{-\int_0^{\log(y/l(\bar{\lambda}))} w(z)dz}, \quad \text{i.e.,} \quad w(y) = -\frac{l(\bar{\lambda})e^y v'(l(\bar{\lambda})e^y)}{v(l(\bar{\lambda})e^y)},$$

the boundary problem (4.2)–(4.4) simplifies to a Riccati ordinary differential equation (ODE):

$$w'(y) - w(y)^2 + (2\bar{\mu} - 1)\, w(y) - \left(\bar{\mu} - \bar{\lambda}\right)\left(\bar{\mu} + \bar{\lambda}\right) = 0,$$

$$y \in [0, \log u(\bar{\lambda})/l(\bar{\lambda})], \tag{4.7}$$

$$w(0) = \bar{\mu} - \bar{\lambda}, \tag{4.8}$$

$$w(\log(u(\bar{\lambda})/l(\bar{\lambda}))) = \bar{\mu} + \bar{\lambda}, \tag{4.9}$$

where

$$\frac{u(\bar{\lambda})}{l(\bar{\lambda})} = \frac{1}{(1-\varepsilon)}\frac{\eta_{\alpha+}}{\eta_{\alpha-}} = \frac{1}{(1-\varepsilon)}\frac{\bar{\mu}+\bar{\lambda}}{\bar{\mu}-\bar{\lambda}}.$$

For each $\bar{\lambda}$, the initial value problem (4.7) and (4.8) has a solution $w(\cdot)$, which we now denote by $w(\bar{\lambda}, \cdot)$ with a slight abuse of notation. Thus, the correct value of $\bar{\lambda}$ is identified by the second boundary condition (4.9):

$$w(\bar{\lambda}, \log(u(\bar{\lambda})/l(\bar{\lambda}))) = \bar{\mu} + \bar{\lambda}. \tag{4.10}$$

4.2. Shadow Market

The key to making the above arguments rigorous is to find a frictionless shadow price $\tilde{S}$, which yields the same optimal policy as the one derived in the previous section. This step requires another heuristic argument.

As for logarithmic utility (Gerhold et al. 2012, 2013) and power utility (GGMKS), the idea is that $\tilde{S}/S$, the ratio between the shadow and the ask price, should only depend on the state variable. Hence, we look for a shadow price of the form

$$\tilde{S}_t = \frac{S_t}{e^{\Upsilon_t}} g(e^{\Upsilon_t}),$$

where $e^{\Upsilon_t} = Y_t/l$ is the risky position at the ask price S, and centered at the buying boundary $l = \eta_{\alpha-} = (\bar{\mu} - \bar{\lambda})/\alpha$. The number of units φ remains constant inside the no-trade region, so that in its interior the dynamics of $\Upsilon = \log(\varphi/l) + \log(S)$ coincides with that of $\log(S)$. Moreover, since Υ must remain in $[0, \log(u/l)]$ by definition, Υ is reflected at the boundaries. Hence,

$$d\Upsilon_t = \left(\mu - \frac{\sigma^2}{2}\right)dt + \sigma\, dW_t + dL_t - dU_t,$$

for nondecreasing local time processes L, U that only increase on $\{\Upsilon_t = 0\}$ (resp. $\{\Upsilon_t = \log(u/l)\}$). The function $g : [1, u/l] \to [1, (1-\varepsilon)u/l]$ is a C^2-function satisfying the smooth pasting conditions (cf. Gerhold et al. 2013):

$$g(1) = 1, \quad g(u/l) = (1-\varepsilon)u/l, \quad g'(1) = 1, \quad g'(u/l) = 1 - \varepsilon. \tag{4.11}$$

The first two conditions ensure that $\tilde{S}$ equals the ask price S (resp. the bid price $S(1-\varepsilon)$) when Υ sits at the buying boundary 0 (resp. at the selling boundary $\log(u/l)$), while the latter two conditions ensure that the diffusion coefficient of $\tilde{S}_t/S_t$ vanishes both at the bid $1-\varepsilon$ and at the ask 1, and hence that these bounds are not breached. The boundary conditions for g', and Itô's formula imply that $\tilde{S}$ is an Itô process with dynamics

$$d\tilde{S}_t/\tilde{S}_t = \tilde{\mu}(\Upsilon_t)dt + \tilde{\sigma}(\Upsilon_t)dW_t,$$

where

$$\tilde{\mu}(y) = \frac{\mu g'(e^y)e^y + \frac{\sigma^2}{2} g''(e^y)e^{2y}}{g(e^y)}, \quad \text{and} \quad \tilde{\sigma}(y) = \frac{\sigma g'(e^y)e^y}{g(e^y)}.$$

To identify the function g, first derive the HJB equation for a generic g. Then, compare this equation to the one obtained in the previous section for the market with transaction costs. Because the value function of the two problems must be the same, matching the two HJB equations identifies the function g.

The wealth process corresponding to a policy[9] $\tilde{\eta}$ in terms of the shadow price $\tilde{S}$ is

$$d\tilde{X}_t = \tilde{\eta}_t \tilde{\mu}(\Upsilon_t)dt + \tilde{\eta}_t \tilde{\sigma}(\Upsilon_t) dW_t.$$

With the standard ansatz $\tilde{V}(t, \tilde{X}_t, \Upsilon_t)$ for the value function, Itô's formula yields

$$\begin{aligned} d\tilde{V}(t, \tilde{X}_t, \Upsilon_t) = &\left(\tilde{V}_t + \tilde{\mu}\tilde{\eta}_t \tilde{V}_x + \frac{\tilde{\sigma}^2}{2}\tilde{\eta}_t^2 \tilde{V}_{xx} + \left(\mu - \frac{\sigma^2}{2}\right)\tilde{V}_y + \frac{\sigma^2}{2}\tilde{V}_{yy} + \sigma\tilde{\sigma}\tilde{\eta}_t \tilde{V}_{xy} \right) dt \\ &+ \tilde{V}_y(dL_t - dU_t) + (\tilde{\sigma}\tilde{\eta}_t \tilde{V}_x + \sigma \tilde{V}_y) dW_t, \end{aligned}$$

where the arguments of the functions are omitted for brevity. Since $\tilde{V}$ must be a supermartingale for any strategy, and a martingale for the optimal strategy, the HJB equation reads as:

$$\sup_{\tilde{\eta}} \left(\tilde{V}_t + \tilde{\mu}\tilde{\eta}\tilde{V}_x + \frac{\tilde{\sigma}^2}{2}\tilde{\eta}^2 \tilde{V}_{xx} + \left(\mu - \frac{\sigma^2}{2}\right)\tilde{V}_y + \frac{\sigma^2}{2}\tilde{V}_{yy} + \sigma\tilde{\sigma}\tilde{\eta}\tilde{V}_{xy} \right) = 0,$$

with the Neumann boundary conditions

$$\tilde{V}_y(0) = \tilde{V}_y(\log(u/l)) = 0.$$

The homogeneity of the value function, (i.e., $\tilde{V}(t, x, y) = -e^{-\alpha x}\tilde{v}(t, y)$) leads to the first-order condition:

$$\tilde{\eta}_t = \frac{1}{\alpha}\left(\frac{\tilde{\mu}}{\tilde{\sigma}^2} + \frac{\sigma}{\tilde{\sigma}} \frac{\tilde{v}_y}{\tilde{v}} \right).$$

Plugging this equality back into the HJB equation yields the nonlinear equation

$$\tilde{v}_t + \left(\mu - \frac{\sigma^2}{2}\right)\tilde{v}_y + \frac{\sigma^2}{2}\tilde{v}_{yy} - \frac{1}{2}\left(\frac{\tilde{\mu}}{\tilde{\sigma}} + \sigma \frac{\tilde{v}_y}{\tilde{v}} \right)^2 \tilde{v} = 0.$$

Now, the equivalent annuity of the optimal policy must be the same for the shadow market as for the transaction cost market in the previous section. Thus, in view of (4.1), set

$$\tilde{v}(t, y) = e^{\alpha\sigma^2\bar{\beta}t} e^{-\int_0^y \tilde{w}(z)dz},$$

[9]Note that this is the monetary *amount* rather than the fraction of wealth in the risky asset.

which implies that $\tilde{v}_t = \alpha\bar{\beta}\sigma^2\tilde{v}$ and $\tilde{v}_y/\tilde{v} = -\tilde{w}$. Then, the HJB equation reduces to the inhomogeneous Riccati ODE,

$$\tilde{w}' - \tilde{w}^2 + (2\bar{\mu} - 1)\,\tilde{w} - (\bar{\mu}^2 - \bar{\lambda}^2) + \left(\frac{\tilde{\mu}}{\sigma\tilde{\sigma}} - \tilde{w}\right)^2 = 0, \tag{4.12}$$

with the boundary conditions

$$\tilde{w}(0) = \tilde{w}(\log(u/l)) = 0.$$

For $\tilde{S}$ to be a shadow price, its value function

$$\tilde{V}_t = -e^{\alpha\sigma^2\bar{\beta}t - \alpha\tilde{X}_t - \int_0^y \tilde{w}(z)dz},$$

must coincide with the value function

$$V_t = -e^{\alpha\sigma^2\bar{\beta}t - \alpha X_t - \int_0^y w(z)dz},$$

for the transaction cost problem derived above. By definition, the safe position X and the wealth $\tilde{X}$ in terms of $\tilde{S} = Se^{-\Upsilon}g(e^{\Upsilon})$ are related via

$$\tilde{X}_t - X_t = \varphi_t^0 + \varphi_t\tilde{S}_t - \varphi_t^0 = g(e^{\Upsilon_t})l.$$

Now, the condition $\tilde{V} = V$ implies that

$$0 = \alpha g(e^y)l + \int_0^y (\tilde{w}(z) - w(z))dz,$$

which in turn means that

$$\tilde{w}(y) = w(y) - \alpha g'(e^y)e^y l.$$

Plugging this relation into the ODE (4.12) for $\tilde{w}$, using the ODE (4.7) for w, and simplifying gives

$$\left(-w(y) + \frac{\tilde{\mu}(y)}{\sigma\tilde{\sigma}(y)}\right)^2 = 0.$$

Inserting the definitions of $\tilde{\mu}(y)$ and $\tilde{\sigma}(y)$, this relation is tantamount to the following ODE for g:

$$\frac{g''(e^y)e^y}{g'(e^y)} + 2\bar{\mu} - 2w(y) = 0. \tag{4.13}$$

Now, the substitution[10]

$$k(y) = \frac{1}{g'(e^y)e^y}, \quad \text{i.e.,} \quad g(e^y) = 1 + \int_0^y \frac{1}{k(z)}dz,$$

reduces this ODE to the inhomogeneous *linear* equation

$$k'(y) = k(y)\,(2\bar{\mu} - 1 - 2w(y))\,. \tag{4.14}$$

[10]For the second representation, we already use the boundary condition $g(1) = 1$.

The smooth pasting condition for g implies $k(0) = 1/g'(1) = 1$. The solution to (4.14) then follows from the variation of constants formula. Plugging in the explicit formula (5.1) for w, and integrating, leads to (with $a = a(\bar{\lambda})$ and $b = b(\bar{\lambda})$ as in (5.2)) the solution[11]

$$k(y) = \left(1 + \frac{b^2}{a^2}\right) \cos^2\left[\tan^{-1}\left(\frac{b}{a}\right) + ay\right].$$

Now the chain of substitutions is reversed starting from k, which is known explicitly up to the constant $\bar{\lambda}$. First, set $\tilde{w}(y) = w(y) - \alpha l/k(y)$; then $\tilde{w}(0) = 0$ by the initial conditions for w and k. To establish the other boundary condition $\tilde{w}(\log(u/l)) = 0$, it suffices to check that $k(\log(u/l)) = (\bar{\mu} - \bar{\lambda})/(\bar{\mu} + \bar{\lambda})$. To see this, insert the boundary condition for w',

$$\bar{\mu} + \bar{\lambda} = w'(\log(u/l)) = \frac{a^2}{\cos^2\left[\tan^{-1}\left(\frac{b}{a}\right) + a\log\left(\frac{u}{l}\right)\right]}, \tag{4.15}$$

into the explicit formula for $k(y)$.[12] Now, observe that the function

$$g(e^y) = 1 + \int_0^y \frac{1}{k(z)} dz = 1 + \frac{a}{a^2 + b^2}\left(\tan\left[\tan^{-1}\left(\frac{b}{a}\right) + ay\right] - \frac{b}{a}\right),$$

evidently satisfies $g(1) = 1$. Moreover, $g(u/l) = (1 - \varepsilon)u/l$, which follows by inserting the terminal condition for w,

$$\bar{\mu} + \bar{\lambda} = w(\log(u/l)) = a\tan\left[\tan^{-1}\left(\frac{b}{a}\right) + a\log\left(\frac{u}{l}\right)\right] + \left(\bar{\mu} - \frac{1}{2}\right),$$

into the explicit expression for g. Finally, these boundary conditions for g and those for k imply that $g'(1) = 1$ and $g'(u/l) = 1 - \varepsilon$, i.e., g satisfies the smooth pasting conditions (4.11) and, by construction, also the ODE (4.13).

5. PROOFS

In the previous section, we first used informal control arguments to find a candidate optimal policy and its corresponding value function. Then, we used this guess to derive a candidate shadow price, matching a generic shadow value function with the one of the transaction cost problem.

In this section, we prove a verification theorem for the optimal policy in the frictionless market corresponding to the candidate shadow price process, and show that the optimal shadow strategy only entails purchasing (selling) when the shadow price coincides with the ask (bid) price. Thus, the policy is also feasible and optimal in the market with transaction costs.

Key to this goal are the new finite-horizon bounds for exponential utility in Theorem 3.2.

[11] Here we only show the case $\bar{\mu} > 1/4$. The other case leads to another explicit formula, whence similar calculations follow.

[12] The first equalities in (4.15) follows from the ODE for w, whereas the second equality is obtained from the explicit formula (5.1).

5.1. Explicit Formulas and Their Properties

The first step to construct the shadow price is to determine, for a given small $\bar{\lambda} > 0$, an explicit expression for the solution w of the ODE (4.7), complemented by the initial condition (4.8).

LEMMA 5.1. *For sufficiently small $\bar{\lambda} > 0$, the function*

$$w(\bar{\lambda}, y) = \begin{cases} a(\bar{\lambda}) \coth[\coth^{-1}(b(\bar{\lambda})/a(\bar{\lambda})) - a(\bar{\lambda})y] + \left(\bar{\mu} - \frac{1}{2}\right), & \text{if } \bar{\mu} \le 1/4, \\ a(\bar{\lambda}) \tan[\tan^{-1}(b(\bar{\lambda})/a(\bar{\lambda})) + a(\bar{\lambda})y] + \left(\bar{\mu} - \frac{1}{2}\right), & \text{if } \bar{\mu} > 1/4, \end{cases} \tag{5.1}$$

with

$$a(\bar{\lambda}) = \sqrt{\left|\bar{\mu}^2 - \bar{\lambda}^2 - \left(\frac{1}{2} - \bar{\mu}\right)^2\right|} \quad \text{and} \quad b(\lambda) = \frac{1}{2} + \bar{\lambda}, \tag{5.2}$$

is a local solution of

$$w'(y) - w^2(y) + (2\bar{\mu} - 1)\, w(y) - (\bar{\mu}^2 - \bar{\lambda}^2) = 0, \quad w(0) = \bar{\mu} - \bar{\lambda}. \tag{5.3}$$

Moreover, $y \mapsto w(\bar{\lambda}, y)$ is increasing in both cases.

Proof. The first part of the assertion is verified by taking derivatives. The second follows by inspection of the explicit formulas. □

Next, establish that the crucial constant $\bar{\lambda}$, which determines both the no-trade region and the equivalent annuity, is well defined.

LEMMA 5.2. *Let $w(\bar{\lambda}, \cdot)$ be defined as in Lemma 5.1, and set*

$$l(\bar{\lambda}) = \frac{\bar{\mu} - \bar{\lambda}}{\alpha}, \quad u(\bar{\lambda}) = \frac{1}{(1-\varepsilon)} \frac{\bar{\mu} + \bar{\lambda}}{\alpha}.$$

Then, for sufficiently small $\varepsilon > 0$, there exists a unique solution $\bar{\lambda}$ of

$$w\left(\bar{\lambda}, \log\left(\frac{u(\bar{\lambda})}{l(\bar{\lambda})}\right)\right) - (\bar{\mu} + \bar{\lambda}) = 0. \tag{5.4}$$

As $\varepsilon \downarrow 0$, it has the asymptotics

$$\bar{\lambda} = \left(\frac{3}{4}\bar{\mu}^2\right)^{1/3} \varepsilon^{1/3} + O(\varepsilon).$$

Proof. The explicit expression for w in Lemma 5.1 implies that $w(\bar{\lambda}, x)$ in Lemma 5.1 is analytic in both variables at (0, 0). By the initial condition in (5.3), its power series has the form

$$w(\bar{\lambda}, y) = (\bar{\mu} - \bar{\lambda}) + \sum_{i=1}^{\infty} \sum_{j=0}^{\infty} W_{ij} y^i \bar{\lambda}^j,$$

where expressions for the coefficients W_{ij} are computed by expanding the explicit expression for w. Hence, the left-hand side of the boundary condition (5.4) is an analytic function of ε and $\bar{\lambda}$. Its power series expansion shows that the coefficients of $\varepsilon^0\bar{\lambda}^j$ vanish for $j = 0, 1, 2$, so that the condition (5.4) reduces to

$$\bar{\lambda}^3 \sum_{i\geq 0} A_i \bar{\lambda}^i = \varepsilon \sum_{i,j\geq 0} B_{ij}\varepsilon^i \bar{\lambda}^j, \tag{5.5}$$

with (computable) coefficients A_i and B_{ij}. This equation has to be solved for $\bar{\lambda}$. Since

$$A_0 = -\frac{4}{3\bar{\mu}} \qquad \text{and} \qquad B_{00} = \bar{\mu}$$

are nonzero, divide the equation (5.5) by $\sum_{i\geq 0} A_i\bar{\lambda}^i$, and take the third root, obtaining that, for some C_{ij},

$$\bar{\lambda} = \varepsilon^{1/3} \sum_{i,j\geq 0} C_{ij}\varepsilon^i \bar{\lambda}^j = \varepsilon^{1/3} \sum_{i,j\geq 0} C_{ij}(\varepsilon^{1/3})^{3i}\bar{\lambda}^j\,.$$

The right-hand side is an analytic function of $\bar{\lambda}$ and $\varepsilon^{1/3}$, so that the implicit function theorem (Gunning and Rossi 2009, theorem I.B.4) yields a unique solution $\bar{\lambda}$ (for ε sufficiently small), which is an analytic function of $\varepsilon^{1/3}$. Its power series coefficients can be computed at any order. □

Henceforth, consider a small relative bid-ask spread $\varepsilon > 0$, and let $\bar{\lambda}$ denote the constant in Lemma 5.2. Moreover, set $w(y) := w(\bar{\lambda}, y)$, $a := a(\bar{\lambda})$, $b := b(\bar{\lambda})$, and $u := u(\bar{\lambda})$, $l := l(\bar{\lambda})$. By inspection, it follows that

LEMMA 5.3. *In both cases of Lemma 5.1,*

$$w'(0) = \bar{\mu} - \bar{\lambda}, \quad w'\left(\log\left(\frac{u}{l}\right)\right) = \bar{\mu} + \bar{\lambda}.$$

The next lemma states the properties of the function k.

LEMMA 5.4. *Define*

$$k(y) = \begin{cases} \left(\dfrac{b^2}{a^2} - 1\right)\sinh^2\left[\coth^{-1}\left(\dfrac{b}{a}\right) - ay\right], & \text{if } \bar{\mu} \leq 1/4, \\ \left(\dfrac{b^2}{a^2} + 1\right)\cos^2\left[\tan^{-1}\left(\dfrac{b}{a}\right) + ay\right], & \text{if } \bar{\mu} > 1/4. \end{cases}$$

Then k satisfies the linear ODE

$$k'(y) = k(y)\,(2\bar{\mu} - 1 - 2w(y)), \quad 0 \leq y \leq \log\left(\frac{u}{l}\right),$$

with boundary conditions

$$k(0) = 1, \quad k\left(\log\left(\frac{u}{l}\right)\right) = \frac{\bar{\mu} - \bar{\lambda}}{\bar{\mu} + \bar{\lambda}}.$$

Moreover, k is strictly decreasing and, in particular, strictly positive on $[0, \log(u/l)]$.

Proof. That k satisfies the ODE follows by insertion. The identities $\cos^2[\tan^{-1}(x)] = 1/(1+x^2)$ and $\sinh^2[\coth^{-1}(x)] = 1/(x^2-1)$ yield the boundary condition at zero, whereas the boundary condition at $\log(u/l)$ follows by inserting $w'(\log(u/l)) = \bar{\mu} + \bar{\lambda}$. Finally, the ODE and a comparison argument yield that k is strictly decreasing. □

LEMMA 5.5. *For* $0 \le y \le \log(u/l)$, *define*

$$g(e^y) := 1 + \int_0^y \frac{1}{k(z)} dz.$$

Then

$$g(e^y) = \begin{cases} 1 + \dfrac{a}{b^2 - a^2}\left(\sinh\left[\sinh^{-1}\left(\dfrac{b}{a}\right) - ay\right] - \dfrac{b}{a}\right), & \text{if } \bar{\mu} \le 1/4, \\ 1 + \dfrac{a}{b^2 + a^2}\left(\tan\left[\tan^{-1}\left(\dfrac{b}{a}\right) + ay\right] - \dfrac{b}{a}\right), & \text{if } \bar{\mu} > 1/4, \end{cases}$$

and g *satisfies the boundary and smooth pasting conditions*

$$g(1) = 1, \quad g(u/l) = (1-\varepsilon)u/l, \quad g'(1) = 1, \quad g'(u/l) = 1 - \varepsilon.$$

Moreover, $g' > 0$ *so that* g *maps* $[1, u/l]$ *onto* $[1, (1-\varepsilon)u/l]$. *Finally,* g *solves the ODE*

$$\frac{g''(e^y)e^y}{g'(e^y)} + 2\bar{\mu} - 2w(y) = 0. \tag{5.6}$$

Proof. The explicit representation follows by elementary integration. Evidently, $g(1) = 1$. Moreover, $g(u/l) = (1-\varepsilon)u/l$ follows by inserting $\bar{\mu} + \bar{\lambda} = w(\log(u/l))$. Next, since $g'(e^y) = 1/e^y k(y)$, the boundary conditions for g and k imply the smooth pasting conditions $g'(1) = 1$ and $g'(u/l) = 1 - \varepsilon$. Furthermore, $k > 0$ and $g'(e^y) = 1/e^y k(y)$ show that $g' > 0$. Finally, computing the derivatives verifies that g indeed satisfies the ODE (5.6). □

5.2. The Shadow Price and Verification

The construction of the shadow price proceeds in analogy to logarithmic utilities (Gerhold et al. 2012, 2013) and power utilities (GGMKS). For $y \in [0, \log(u/l)]$, let Υ be a Brownian motion with drift, reflected at 0 and $\log(u/l)$, that is, the continuous, adapted process with values in $[0, \log(u/l)]$ such that

$$d\Upsilon_t = (\mu - \sigma^2/2)dt + \sigma\, dW_t + dL_t - dU_t, \quad \Upsilon_0 = y_0, \tag{5.7}$$

for nondecreasing adapted local time processes L and U increasing only on the sets $\{\Upsilon_t = 0\}$ and $\{\Upsilon_t = \log(u/l)\}$, respectively.

LEMMA 5.6. *Define*

$$y_0 = \begin{cases} 0, & \text{if } \xi S_0 \le l, \\ \log(u/l), & \text{if } \xi S_0 \ge u, \\ \log(\xi S_0/l), & \text{otherwise,} \end{cases} \tag{5.8}$$

and let Υ be defined as in (5.7), started at $\Upsilon_0 = y_0$. Then $\tilde{S} = Se^{-\Upsilon}g(e^{\Upsilon})$, with g as in Lemma 5.5, is a positive Itô process with dynamics

$$d\tilde{S}_t/\tilde{S}_t = \tilde{\mu}(\Upsilon_t)dt + \tilde{\sigma}(\Upsilon_t)dW_t, \quad \tilde{S}_0 = S_0 e^{-y_0}g(e^{y_0}),$$

for

$$\tilde{\mu}(y) = \frac{\mu g'(e^y)e^y + \frac{\sigma^2}{2}g''(e^y)e^{2y}}{g(e^y)}, \quad \tilde{\sigma}(y) = \frac{\sigma g'(e^y)e^y}{g(e^y)},$$

and $\tilde{S}$ takes values in the bid-ask spread $[(1-\varepsilon)S, S]$.

Note that the first (resp. second) case in (5.8) occurs if the initial position ξS_0 in the risky asset lies below the buying boundary l or above the selling boundary u. Then, there is a jump from the initial position $(\varphi^0_{0-}, \varphi_{0-}) = (\xi^0, \xi)$, which moves the position in the risky asset to the nearest boundary of the interval $[l, u]$. Since this initial trade involves the purchase (resp. sale) of shares, the initial value of $\tilde{S}$ is chosen to match the initial ask (resp. bid) price.

Proof of Lemma 5.6. The first part of the assertion follows from the smooth pasting conditions for g and Itô's formula. As for the second part, since $g''(1) \leq 0$, a comparison argument yields that the derivative $(g'(y)y - g(y))/y^2$ of $g(y)/y$ is nonpositive. Hence, $g(1)/1 = 1$ and $g(u/l)/(u/l) = 1-\varepsilon$ yield that $\tilde{S} = Sg(e^{\Upsilon})e^{-\Upsilon}$ is indeed $[(1-\varepsilon)S, S]$-valued. □

The long-run optimal portfolio in the frictionless "shadow market" with price process $\tilde{S}$ can be determined by calculating finite-horizon bounds, similarly as in Guasoni and Robertson (2012) for power utility. Note that for the exponential utilities considered here, the myopic probability coincides with the (unique) risk-neutral probability for $\tilde{S}$.

LEMMA 5.7. *$\tilde{w}(y) = w(y) - \alpha g'(e^y)e^y l$, with w and g as in Lemmas 5.1 and 5.5, solves the ODE*

$$\tilde{w}' - \tilde{w}^2 + (2\bar{\mu} - 1)\tilde{w} - (\bar{\mu}^2 - \bar{\lambda}^2) + \left(\frac{\tilde{\mu}}{\sigma\tilde{\sigma}} - \tilde{w}\right)^2 = 0, \tag{5.9}$$

with boundary conditions $\tilde{w}(0) = \tilde{w}(\log(u/l)) = 0$. Moreover, denoting by $\tilde{q}(y) = \int_0^y \tilde{w}(z)dz$, the shadow payoff $\tilde{X}_T$ corresponding to the policy $\tilde{\eta} = \frac{1}{\alpha}\left(\frac{\tilde{\mu}}{\tilde{\sigma}^2} - \frac{\sigma}{\tilde{\sigma}}\tilde{w}\right)$ (in terms of $\tilde{S}$) and the shadow discount factor $M_T = \mathcal{E}(-\int_0^{\cdot}\frac{\tilde{\mu}}{\tilde{\sigma}}dW_t)_T$ satisfies the following bounds:

$$E\left[e^{-\alpha\tilde{X}_T}\right] = e^{-\alpha\tilde{X}_0}e^{-\alpha\sigma^2\bar{\beta}T}\tilde{E}\left[e^{\tilde{q}(\Upsilon_T)-\tilde{q}(\Upsilon_0)}\right] = e^{-\alpha\tilde{X}_0}e^{-\alpha\sigma^2\bar{\beta}T} + O(\varepsilon), \tag{5.10}$$

$$e^{-\alpha\tilde{X}_0 - \tilde{E}[\log M_T]} = e^{-\alpha\tilde{X}_0}e^{-\alpha\sigma^2\bar{\beta}T}e^{\tilde{E}[\tilde{q}(\Upsilon_T)-\tilde{q}(\Upsilon_0)]} = e^{-\alpha\tilde{X}_0}e^{-\alpha\sigma^2\bar{\beta}T} + O(\varepsilon). \tag{5.11}$$

Here, $\bar{\beta} = (\bar{\mu}^2 - \bar{\lambda}^2)/2\alpha$ and $\tilde{E}[\cdot]$ denotes the expectation with respect to the risk-neutral probability $\tilde{Q}$ for $\tilde{S}$ with density process M.

Proof. That $\tilde{w}$ solves the ODE (5.9) is easily verified by taking derivatives, while the boundary conditions immediately follow from their counterparts for w and g.

Next, note that $\tilde{\mu}, \tilde{\sigma}, \tilde{\eta}, \tilde{w}$ are functions of Υ_t, but their argument is omitted throughout to ease notation. To prove the first bound (5.10), note that the shadow wealth process $\tilde{X}$

satisfies:

$$
\begin{aligned}
e^{-\alpha \tilde{X}_T} &= e^{-\alpha \tilde{X}_0} \exp\left(-\int_0^T \alpha \tilde{\eta} \tilde{\mu} dt - \int_0^T \alpha \tilde{\eta} \tilde{\sigma} \, dW_t \right) \\
&= e^{-\alpha \tilde{X}_0} \exp\left(\int_0^T \left(-\frac{\tilde{\mu}^2}{\tilde{\sigma}^2} + \frac{\tilde{\mu}\sigma}{\tilde{\sigma}} \tilde{w} \right) dt + \int_0^T \left(-\frac{\tilde{\mu}}{\tilde{\sigma}} + \sigma \tilde{w} \right) dW_t \right), \qquad (5.12)
\end{aligned}
$$

where the second equality follows by substituting $\tilde{\eta} = \frac{1}{\alpha}(\frac{\tilde{\mu}}{\tilde{\sigma}^2} - \frac{\sigma}{\tilde{\sigma}} \tilde{w})$. Now, Itô's formula and the boundary conditions $\tilde{w}(0) = \tilde{w}(\log(u/l)) = 0$ imply

$$
\tilde{q}(\Upsilon_T) - \tilde{q}(\Upsilon_0) = \int_0^T \left(\left(\mu - \frac{\sigma^2}{2} \right) \tilde{w} + \frac{\sigma^2}{2} \tilde{w}' \right) dt + \int_0^T \sigma \tilde{w} \, dW_t.
$$

Plugging in the ODE for $\tilde{w}$, it follows that

$$
\tilde{q}(\Upsilon_T) - \tilde{q}(\Upsilon_0) = \int_0^T \left(-\frac{\tilde{\mu}^2}{2\tilde{\sigma}^2} + \frac{\tilde{\mu}\sigma}{\tilde{\sigma}} \tilde{w} + \alpha \sigma^2 \bar{\beta} \right) dt + \int_0^T \sigma \tilde{w} \, dW_t. \qquad (5.13)
$$

Using this identity to replace $\int_0^T \sigma \tilde{w} \, dW_t$ in (5.12) and taking expectations then yields

$$
E\left[e^{-\alpha \tilde{X}_T} \right] = e^{-\alpha \tilde{X}_0} e^{-\alpha \sigma^2 \bar{\beta} T} E\left[M_T e^{\tilde{q}(\Upsilon_T) - \tilde{q}(\Upsilon_0)} \right].
$$

The first bound now follows by noting that, since $\tilde{\mu}(\cdot)/\tilde{\sigma}(\cdot)$ is bounded on the support $[0, \log(u/l)]$ of its argument, the nonnegative local martingale M is in fact a true martingale, such that $\tilde{Q}$ is well defined.

As for the second bound, first note that by definition of M and Girsanov's theorem,

$$
e^{-\alpha \tilde{X}_0 - \tilde{E}[\log M_T]} = \exp\left(-\alpha \tilde{X}_0 + \tilde{E}\left[-\int_0^T \frac{\tilde{\mu}^2}{2\tilde{\sigma}^2} dt + \int_0^T \frac{\tilde{\mu}}{\tilde{\sigma}} d\tilde{W}_t \right] \right), \qquad (5.14)
$$

where $\tilde{W}_t = W_t + \int_0^t \frac{\tilde{\mu}}{\tilde{\sigma}} ds$ denotes a $\tilde{Q}$-Brownian motion. Again by using that the process $\tilde{\mu}/\tilde{\sigma}$ is bounded, it follows that the stochastic integral in (5.14) is a $\tilde{Q}$-martingale with vanishing expectation. Using (5.13), also rewritten in terms of $\tilde{W}$, to replace the Lebesgue integral in (5.14) then shows

$$
e^{-\alpha \tilde{X}_0 - \tilde{E}[\log M_T]} = e^{-\alpha \tilde{X}_0 - \alpha \sigma^2 \bar{\beta} T} \exp\left(\tilde{E}\left[\tilde{q}(\Upsilon_T) - \tilde{q}(\Upsilon_0) + \int_0^T \left(\frac{\tilde{\mu}}{\tilde{\sigma}} - \sigma \tilde{w} \right) d\tilde{W}_t \right] \right).
$$

Since $\tilde{w}(\cdot)$ and $\frac{\tilde{\mu}(\cdot)}{\tilde{\sigma}(\cdot)}$ are bounded on $[0, \log(u/l)]$, the $d\tilde{W}_t$-term in this expression is a $\tilde{Q}$-martingale, which yields the second bound (5.11).

The asymptotics follow by expanding the function $\tilde{q}$ as in the proof of GGMKS, theorem 3.1 □

With the finite-horizon bounds at hand, we can now establish that the policy $\tilde{\eta}$ is indeed long-run optimal in the frictionless market with price $\tilde{S}$.

LEMMA 5.8. *The policy*

$$
\tilde{\eta}(\Upsilon_t) = \frac{1}{\alpha} \left(\frac{\tilde{\mu}(\Upsilon_t)}{\tilde{\sigma}^2(\Upsilon_t)} - \frac{\sigma}{\tilde{\sigma}(\Upsilon_t)} \tilde{w}(\Upsilon_t) \right) = g(e^{\Upsilon_t}) l \qquad (5.15)
$$

is long-run optimal with equivalent annuity $\sigma^2\bar{\beta}$ in the frictionless market with price process $\tilde{S}$. The corresponding wealth process (in terms of $\tilde{S}$), and the numbers of safe and risky units satisfy

$$\tilde{X} = (\xi^0 + \xi\tilde{S}_0) + \int_0^{\cdot} \tilde{\eta}(\Upsilon_t)\tilde{\mu}(\Upsilon_t)dt + \int_0^{\cdot} \tilde{\eta}(\Upsilon_t)\tilde{\sigma}(\Upsilon_t)dW_t,$$
$$\varphi^0_{0-} = \xi^0, \quad \varphi^0_t = \tilde{X}_t - \tilde{\eta}(\Upsilon_t) \quad \textit{for } t \geq 0,$$
$$\varphi_{0-} = \xi, \quad \varphi_t = \tilde{\eta}(\Upsilon_t)/\tilde{S}_t \quad \textit{for } t \geq 0.$$

Proof. The formulae for the trading strategy and the wealth process associated to $\tilde{\eta}$ are immediate consequences of the respective definitions. The second representation for $\tilde{\eta}$ follows by inserting the definitions of $\tilde{\mu}$, $\tilde{\sigma}$ from Lemma 5.6, the ODE (5.6) for g, and $w(y) - \tilde{w}(y) = \alpha g'(e^y)e^y l$.

Next, note that φ is admissible for $\tilde{S}$, because (5.15) shows that the corresponding risky position $\tilde{\eta}$ is bounded. Now, standard duality arguments for exponential utility imply that the shadow payoff $\tilde{X}^\phi$ corresponding to *any* admissible strategy ϕ satisfies the inequality

$$E\big[e^{-\alpha\tilde{X}^\phi_T}\big] \geq e^{-\alpha\tilde{X}^\phi_0 - \tilde{E}[\log M_T]}. \tag{5.16}$$

Indeed, since $\tilde{\sigma}(\Upsilon)$ is uniformly bounded and the same holds for $\phi\tilde{S}$ by admissibility of ϕ and $(1-\varepsilon)S \leq \tilde{S} \leq S$, the local $\tilde{Q}$-martingale $\tilde{X}^\phi = \tilde{X}^\phi_0 + \int_0^{\cdot} \phi_t d\tilde{S}_t$ is in fact a true $\tilde{Q}$-martingale. Now, $d\tilde{Q}|_{\mathcal{F}_T}/dP|_{\mathcal{F}_T} = M_T$ and Jensen's inequality yield

$$E\big[e^{-\alpha\tilde{X}^\phi_T}\big] = \tilde{E}\big[e^{-\alpha\tilde{X}^\phi_T - \log M_T}\big] \geq e^{-\alpha\tilde{E}[\tilde{X}^\phi_T] - \tilde{E}[\log M_T]},$$

such that (5.16) follows from the $\tilde{Q}$-martingale property of the shadow wealth process $\tilde{X}^\phi$.

Inequality (5.16) in turn yields the following upper bound, valid for any admissible strategy ϕ in the frictionless market with price process $\tilde{S}$:

$$\liminf_{T\to\infty}\left(\frac{1}{-\alpha T}\log E\big[e^{-\alpha\tilde{X}^\phi_T}\big]\right) \leq \liminf_{T\to\infty}\frac{1}{T}\left(\tilde{X}^\phi_0 + \frac{1}{\alpha}\tilde{E}[\log M_T]\right).$$

The function $\tilde{q}$ in Lemma 5.7 is bounded on the compact support of its argument Υ. Hence, the bound (5.11) in Lemma 5.7 implies that the right-hand side equals $\sigma^2\bar{\beta}$.

Likewise, the bound (5.10) in the same lemma implies that the shadow payoff $\tilde{X}^\varphi$ (corresponding to the number of units φ, defined in terms of the policy $\tilde{\eta}$) satisfies, using again that $\tilde{q}$ is bounded,

$$\liminf_{T\to\infty}\frac{1}{\alpha T}\log E\big[e^{-\alpha\tilde{X}^\varphi_T}\big] = \liminf_{T\to\infty}\left(\sigma^2\bar{\beta} + \frac{\tilde{X}^\varphi_0}{T} - \frac{1}{\alpha T}\log\tilde{E}\big[e^{\tilde{q}(\Upsilon_T)-\tilde{q}(\Upsilon_0)}\big]\right) = \sigma^2\bar{\beta},$$

which shows that the policy $\tilde{\eta}$ attains this upper bound, and concludes the proof. □

The next lemma establishes that $\tilde{S}$ is a shadow price.

LEMMA 5.9. *The number of shares $\varphi = \tilde{\eta}/\tilde{S}$ in the portfolio $\tilde{\eta}$ in Lemma 5.8 has the dynamics*

$$d\varphi_t/\varphi_t = dL_t - dU_t.$$

Thus, φ increases only when $\Upsilon_t = 0$, that is, when $\tilde{S}$ equals the ask price, and decreases only when $\Upsilon_t = \log(u/l)$, that is, when $\tilde{S}$ equals the bid price.

Proof. Itô's formula applied to (5.15) yields

$$\frac{d\tilde{\eta}(\Upsilon_t)}{\tilde{\eta}(\Upsilon_t)} = \frac{\mu g'(e^{\Upsilon_t})e^{\Upsilon_t} + \frac{\sigma^2}{2} g''(e^{\Upsilon_t})e^{2\Upsilon_t}}{g(e^{\Upsilon_t})} dt + \frac{\sigma g'(e^{\Upsilon_t})e^{\Upsilon_t}}{g(e^{\Upsilon_t})} dW_t + \frac{g'(e^{\Upsilon_t})e^{\Upsilon_t}}{g(e^{\Upsilon_t})} d(L_t - U_t).$$

Integrating $\varphi = \tilde{\eta}/\tilde{S}$ by parts, inserting the dynamics of $\tilde{\eta}$ and $\tilde{S}$, and simplifying, it follows that

$$\frac{d\varphi_t}{\varphi_t} = \frac{g'(e^{\Upsilon_t})e^{\Upsilon_t}}{g(e^{\Upsilon_t})} d(L_t - U_t).$$

Since L and U only increase on the sets $\{\Upsilon = 0\}$ and $\{\Upsilon = \log(u/l)\}$, respectively, the assertion now follows from the boundary conditions for g and g'. □

The equivalent annuity for any frictionless price within the bid-ask spread must be greater or equal than in the original market with bid-ask process $((1-\varepsilon)S, S)$, because the investor trades at more favorable prices. For a *shadow price*, there is an optimal strategy that only entails buying (resp. selling) stocks when $\tilde{S}$ coincides with the ask (resp. bid) price. Hence, this strategy yields the same payoff when executed at bid-ask prices, and is also optimal in the original model with transaction costs. The corresponding equivalent annuity must also be the same, since the difference due to the liquidation costs vanishes as the horizon grows in (2.1):

PROPOSITION 5.10. *Let $\tilde{S}$ be the shadow price for $((1-\varepsilon)S, S)$ from Lemma 5.6, and (φ^0, φ) the corresponding long-run optimal strategy from Lemma 5.8. Then (φ^0, φ) is long-run optimal for the bid-ask process $((1-\varepsilon)S, S)$ as well, with the same equivalent annuity $\sigma^2\bar{\beta}$.*

Proof. As φ only increases (resp. decreases) when $\tilde{S} = S$ (resp. $\tilde{S} = (1-\varepsilon)S$), the strategy (φ^0, φ) is self-financing for the bid-ask process $((1-\varepsilon)S, S)$ as well. Moreover, (5.15) and $(1-\varepsilon)S \le \tilde{S} \le S$ show that it is also admissible for $((1-\varepsilon)S, S)$ in the sense of Definition 2.1.

Now, since $S \ge \tilde{S} \ge S(1-\varepsilon)$ and the number φ_t of shares is always positive and bounded from above by $u/S > 0$,

$$\varphi_t^0 + \varphi_t \tilde{S}_t \ge \varphi_t^0 + \varphi_t^+(1-\varepsilon)S_t - \varphi_t^- S_t \ge \varphi_t^0 + \varphi_t \tilde{S}_t - \varepsilon u.$$

These upper and lower bounds yield:

$$\begin{aligned} &\liminf_{T\to\infty} \frac{1}{-\alpha T} \log E\big[e^{-\alpha(\varphi_T^0 + \varphi_T^+(1-\varepsilon)S_T - \varphi_T^- S_T)}\big] \\ &= \liminf_{T\to\infty} \frac{1}{-\alpha T} \log E\big[e^{-\alpha(\varphi_T^0 + \varphi_T \tilde{S}_T)}\big], \end{aligned} \tag{5.17}$$

that is, (φ^0, φ) has the same growth rate, either with $\tilde{S}$ or with $[(1-\varepsilon)S, S]$.

Now let (ψ^0, ψ) be any admissible strategy for the bid-ask spread $[(1-\varepsilon)S, S]$, and define the corresponding cash position in the shadow market as $\tilde{\psi}^0 = \psi_{0-}^0 - \int_0^{\cdot} \tilde{S}_t d\psi_t$. Then $(\tilde{\psi}^0, \psi)$ is a self-financing trading strategy for the shadow price $\tilde{S}$, and $\tilde{\psi}^0 \ge \psi^0$

because $(1-\varepsilon)S \le \tilde{S} \le S$ implies that $\int_0^\cdot \tilde{S}_t d\psi_t \le \int_0^\cdot S_t d\psi_t^{\uparrow} - (1-\varepsilon)\int_0^\cdot S_t d\psi_t^{\downarrow}$. Together with $\tilde{S} \in [(1-\varepsilon)S, S]$, the long-run optimality of (φ^0, φ) for $\tilde{S}$, and (5.17), it follows that:

$$\begin{aligned}
&\liminf_{T\to\infty}\left(\frac{1}{-\alpha T}\log E\left[e^{-\alpha(\psi_T^0+\psi_T^+(1-\varepsilon)S_T-\psi_T^- S_T)}\right]\right)\\
&\le \liminf_{T\to\infty}\left(\frac{1}{-\alpha T}\log E\left[e^{-\alpha(\tilde{\psi}_T^0+\psi_T\tilde{S}_T)}\right]\right)\\
&\le \liminf_{T\to\infty}\left(\frac{1}{-\alpha T}\log E\left[e^{-\alpha(\varphi_T^0+\varphi_T\tilde{S}_T)}\right]\right)\\
&= \liminf_{T\to\infty}\left(\frac{1}{-\alpha T}\log E\left[e^{-\alpha(\varphi_T^0+\varphi_T^+(1-\varepsilon)S_T-\varphi_T^- S_T)}\right]\right).
\end{aligned}$$

Hence (φ^0, φ) is also long-run optimal for the bid-ask process $((1-\varepsilon)S, S)$. □

Putting everything together, we can now complete the proofs of our main results:

Proofs of Theorem 2.3 (*i*)–(*iv*), (*vi*), *Theorem 2.4, Theorem 3.2.* By Lemma 5.8, the strategy (φ^0, φ) is optimal in the frictionless market with price process $\tilde{S}$. Since the latter is a shadow price process by Lemma 5.9, Proposition 5.10 yields that the same strategy is also optimal with the same equivalent annuity $\sigma^2\bar{\beta} = \frac{\sigma^2}{2\alpha}(\bar{\mu}^2 - \bar{\lambda}^2)$ in the original market with transaction costs. This proves Theorem 2.4 and also Item (*i*) of Theorem 2.3, since the definition of $\bar{\lambda}$ in Lemma 5.1 matches (*iv*) by Lemma 5.2. Item (*ii*) of Theorem 2.3 follows immediately by comparing the growth rate to its frictionless value. Next, since $\tilde{S}$ is a shadow price, the buy (resp. sell) boundaries for (φ^0, φ) are quoted in terms of the ask (resp. bid) price. Item (*iii*) then follows from the representation in Lemma 5.8, combined with the boundary conditions for g and the definitions of u, l in Lemma 5.2. The corresponding asymptotic expansions in (*vi*) are an immediate consequence of the fractional power series for $\bar{\lambda}$ (cf. Lemma 5.2) and Taylor expansion. Finally, Theorem 3.2 has been established in Lemma 5.7 and the proof of Lemma 5.8. □

Next, we prove Theorem 3.3, which generalizes the finite-horizon bounds to a market with several uncorrelated assets.

Proof of Theorem 3.3. Let $M^i = \mathcal{E}(-\int_0^\cdot \frac{\tilde{\mu}_i(\Upsilon_t^i)}{\tilde{\sigma}_i(\Upsilon_t^i)} dW_t^i)$ be the stochastic discount factor in the market $(S^0, \tilde{S}^i)$, where the coefficients $\tilde{\mu}_i, \tilde{\sigma}_i$ and the reflected Brownian motions Υ^i are defined as in Lemma 5.6. Then, since the risky assets S^i are independent, the same holds for the processes Υ^i, M^i, $\tilde{S}^i$. For the shadow wealth process $\tilde{X}^\varphi = \tilde{X}_0^\varphi + \sum_{i=1}^d \int_0^\cdot \varphi_t^i d\tilde{S}_t^i$, the first univariate finite-horizon bound in Lemma 5.7, therefore, yields

$$-\frac{1}{\alpha}\log E\big[e^{-\alpha\tilde{X}_T^\varphi}\big] = \tilde{X}_0^\varphi + \sum_{i=1}^d \sigma_i^2\bar{\beta}_i T - \frac{1}{\alpha}\sum_{i=1}^d \log \tilde{E}_i\big[e^{\tilde{q}_i(\Upsilon_T^i)-\tilde{q}_i(\Upsilon_0^i)}\big],$$

where $\tilde{E}_i[\cdot]$ denotes expectation with respect to the measure with density process M_i, and the constants $\bar{\beta}_i$ and the functions $\tilde{q}_i$ are defined as in Lemma 5.7 for $i = 1, \dots, d$. Since

the mappings $\tilde{q}_i$ are bounded on the compact supports of the Υ^i, it follows that

$$\liminf_{T\to\infty} -\frac{1}{\alpha T}\log E\big[e^{-\alpha X_T^{\varphi}}\big] = \sum_{i=1}^{d} \sigma_i^2 \bar{\beta}_i. \tag{5.18}$$

Next, note that since each $\tilde{S}^i$ only depends on one of the independent Brownian motions, Girsanov's theorem implies that $M = \mathcal{E}(-\sum_{i=1}^{d} \int_0^{\cdot} \frac{\tilde{\mu}_i(\Upsilon_t^i)}{\tilde{\sigma}_i(\Upsilon_t^i)} dW_t^i)$ is a stochastic discount factor for the market $(S^0, \tilde{S}^1, \ldots, \tilde{S}^d)$. Hence, it follows verbatim as in the proof of Lemma 5.8 that the shadow wealth process $\tilde{X}^{\phi}$ associated to any admissible strategy ϕ satisfies

$$E\big[e^{-\alpha \tilde{X}_T^{\phi}}\big] \geq e^{-\alpha \tilde{X}_0^{\phi} - \tilde{E}[\log M_T]},$$

where $\tilde{E}[\cdot]$ denotes the expectation with respect to the measure with density process M. Since each of the M^i only depends on one of the independent Brownian motions, Yor's formula implies that $M_T = \prod_{i=1}^{d} M_T^i$ and hence, by independence of the M^i,

$$E\big[e^{-\alpha \tilde{X}_T^{\phi}}\big] \geq e^{-\alpha \tilde{X}_0^{\phi}} \prod_{i=1}^{d} e^{-\tilde{E}_i[\log M_T^i]},$$

where $\tilde{E}_i[\cdot]$ denotes the expectation with respect to the measure with density process M^i. Combined with the second univariate finite-horizon bound from Lemma 5.7, it follows that

$$E\big[e^{-\alpha \tilde{X}_T^{\phi}}\big] \geq e^{-\alpha \tilde{X}_0^{\phi}} \prod_{i=1}^{d} e^{-\alpha\sigma_i^2\bar{\beta}_i T + \tilde{E}_i[\tilde{q}_i(\Upsilon_T^i) - \tilde{q}_i(\Upsilon_0^i)]}.$$

In view of the boundedness of the $\tilde{q}_i$, this inequality yields

$$\liminf_{T\to\infty} -\frac{1}{\alpha T}\log E\big[e^{-\alpha \tilde{X}_T^{\phi}}\big] \leq \sum_{i=1}^{d} \sigma_i^2 \bar{\beta}_i.$$

Together with (5.18), it follows that the strategy φ is optimal in the frictionless market with risky asset $\tilde{S}$. Since, by definition of $\tilde{S}^i$ and Lemma 5.9, the strategy φ only purchases (resp. sells) shares of the risky asset i when $\tilde{S}^i = S^i$ (resp. $\tilde{S}^i = (1-\varepsilon)S^i$), the process $\tilde{S}$ is a shadow price. It then follows as in the proof of Proposition 5.10 that the same strategy is also optimal with the same equivalent annuity in the original market with transaction costs, completing the proof. □

5.3. Trading Volume

As above, let $\varphi_t = \varphi_t^{\uparrow} - \varphi_t^{\downarrow}$ denote the number of risky units at time t, written as the difference of the cumulated numbers of shares bought (resp. sold) until t. *Relative turnover*, defined as the measure $\tilde{S}_t d\|\varphi\|_t / \tilde{S}_t \varphi_t = d\|\varphi\|_t / \varphi_t = d\varphi_t^{\uparrow}/|\varphi_t| + d\varphi_t^{\downarrow}/|\varphi_t|$, is a scale-invariant indicator of trading volume, compare Lo and Wang (2000). The

long-term average relative turnover is defined as

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \frac{d\|\varphi\|_t}{|\varphi_t|}.$$

Similarly, *absolute turnover* $(1-\varepsilon)S_t d\varphi_t^{\downarrow} + S_t d\varphi_t^{\uparrow}$ is defined as the amount of wealth traded, evaluated in terms of the bid price $(1-\varepsilon)S_t$ when selling resp. in terms of the ask price S_t when buying. As above, the *long-term average absolute turnover* is then defined as

$$\lim_{T\to\infty} \frac{1}{T} \left(\int_0^T (1-\varepsilon) S_t d\varphi_t^{\downarrow} + \int_0^T S_t d\varphi_t^{\uparrow} \right).$$

These quantities can be expressed in terms of the long-run averages of the local times L and U:

PROPOSITION 5.11. *The long-term average relative turnover is*

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \frac{d\|\varphi\|_t}{|\varphi_t|} = \lim_{T\to\infty} \frac{U_T}{T} + \lim_{T\to\infty} \frac{L_T}{T}.$$

The long-term average absolute turnover is

$$\lim_{T\to\infty} \frac{1}{T} \left(\int_0^T (1-\varepsilon) S_t d\varphi_t^{\downarrow} + \int_0^T S_t d\varphi_t^{\uparrow} \right) = \frac{\bar{\mu}+\bar{\lambda}}{\alpha} \lim_{T\to\infty} \frac{U_T}{T} + \frac{\bar{\mu}-\bar{\lambda}}{\alpha} \lim_{T\to\infty} \frac{L_T}{T}.$$

Proof. The formula for the relative turnover follows from the representation for $d\varphi/\varphi$ in Lemma 5.9. The formulae for the absolute turnover follow analogously by noting that $S_t\varphi_t = (\bar{\mu}-\bar{\lambda})/\alpha$ on the set $\{\Upsilon_t = 0\}$ where L increases, and likewise $(1-\varepsilon)S_t\varphi_t = (\bar{\mu}+\bar{\lambda})/\alpha$ on the set $\{\Upsilon_t = \log(u/l)\}$ where U increases. □

Using the long-term limits of the local times L and U determined in GGMKS, lemma D.2, it follows that the long-run averages of the local times admit explicit formulae in terms of the gap $\bar{\lambda}$. These in turn yield the asymptotic expansions for $\varepsilon \downarrow 0$ stated in Theorem 2.3 via Taylor expansion.

5.4. Connection to Constant Relative Risk Aversion

Finally, we prove Theorem 3.1, which states that all relevant quantities, i.e., liquidity premium, optimal policy, equivalent annuity, and trading volume, for an investor with constant *absolute* risk aversion arise in the limit for increasing constant *relative* risk aversion.

To this end it suffices to show that the gap of GGMKS, lemma B.2 for relative risk aversion γ converges to its counterpart in our Lemma 5.2 for constant absolute risk aversion α, as $\gamma \uparrow \infty$. Note that this convergence holds for any level of absolute risk aversion, since our gap is independent of the latter.

THEOREM 5.12. *As the relative risk aversion γ in GGMKS, lemma B.2 tends to infinity, their gap $\bar{\lambda}_\gamma$ converges to our counterpart $\bar{\lambda}$ in Lemma 5.2, which is the gap for all levels α of absolute risk aversion.*

Proof. For small ε, the gap[13] $\bar{\lambda}_\gamma$ of GGMKS is given by the unique root of the function

$$f_\gamma(\bar{\lambda}) = \gamma w_\gamma(\bar{\lambda}, \log[u_\gamma(\bar{\lambda})/l_\gamma(\bar{\lambda})]) - (\bar{\mu} + \bar{\lambda}),$$

where the function w_γ is given explicitly in GGMKS,lemma B.1. Now, note that as the relative risk aversion γ becomes large, Case 2 of GGMKS, lemma B.1 applies if $\bar{\mu} > 1/4$ and Case 3 applies if $\bar{\mu} \leq 1/4$. By inspection of the explicit formulas in GGMKS, lemma B.1 resp. our Lemma 5.1, it follows that, as $\gamma \uparrow \infty$, the function $\gamma w_\gamma(\cdot)$ converges uniformly on compacts to w from our Lemma 5.1. Since the same holds for the functions $u_\gamma(\cdot), l_\gamma(\cdot)$ from GGMKS, lemma B.2 and $u(\cdot), l(\cdot)$ from our Lemma 5.2, this in turn yields that $f_\gamma(\cdot)$ converges uniformly on compacts to

$$f(\bar{\lambda}) = w(\bar{\lambda}, \log[u(\bar{\lambda})/l(\bar{\lambda})]) - (\bar{\mu} + \bar{\lambda}).$$

Since our gap $\bar{\lambda}$ is the unique root of this function, it suffices to show that the zeros of f_γ also converge as $\gamma \uparrow \infty$. But this follows, because a calculation shows that, for small ε, the derivative $\frac{\partial}{\partial \bar{\lambda}} f$ is bounded away from zero in a neighborhood of the root of f, completing the proof. □

The convergence of all other—suitably rescaled—quantities follows immediately from the explicit formulas in GGMKS and in this paper.

REFERENCES

AKIAN, M., J. L. MENALDI, and A. SULEM (1996): On an Investment-Consumption Model with Transaction Costs, *SIAM J. Control Optim.* 34(1), 329–364.

AMIHUD, Y., and H. MENDELSON (1986): Asset Pricing and the Bid-Ask Spread, *J. Finan. Econ.* 17(2), 223–249.

BARLES, G., and H. M. SONER (1998): Option Pricing with Transaction Costs and a Nonlinear Black–Scholes Equation, *Finance Stoch.* 2(4), 369–397.

BICHUCH, M. (2012): Asymptotic Analysis for Optimal Investment in Finite Time with Transaction Costs, *SIAM J. Financial Math.* 3(1), 433–458.

ČERNÝ, A. (2009): *Mathematical Techniques in Finance*, 2nd ed., Princeton, NJ: Princeton University Press.

CHOI, J., M. SIRBU, and G. ŽITKOVIĆ (2012): Shadow Prices and Well-Posedness in the Problem of Optimal Investment and Consumption with Transaction Costs. Preprint arXiv:1204.0305.

DAVIS, M. H. A., V. G. PANAS, and T. ZARIPHOPOULOU (1993): European Option Pricing with Transaction Costs, *SIAM J. Control Optim.* 31(2), 470–493.

DELBAEN, F., P. GRANDITS, T. RHEINLÄNDER, D. SAMPERI, M. SCHWEIZER, and C. STRICKER (2002): Exponential Hedging and Entropic Penalties, *Math. Finance* 12(2), 99–123.

DUMAS, B. (1991): Super Contact and Related Optimality Conditions, *J. Econom. Dynam. Control* 15(4), 675–685.

DUMAS, B., and E. LUCIANO (1991): An Exact Solution to a Dynamic Portfolio Choice Problem under Transactions Costs, *J. Finance* 46(2), 577–595.

[13]In GGMKS, appendix B, this is the gap in business time, obtained by their formulae with μ and σ replaced by $\bar{\mu}$ and 1, respectively.

GERHOLD, S., P. GUASONI, J. MUHLE-KARBE, and W. SCHACHERMAYER (2013): Transaction Costs, Trading Volume, and the Liquidity Premium. *Finance Stoch*.

GERHOLD, S., J. MUHLE-KARBE, and W. SCHACHERMAYER (2012): Asymptotics and Duality for the Davis and Norman Problem, *Stoch. Int. J. Probab. Stoch. Proc*. 84(5–6), 625–641.

GERHOLD, S., J. MUHLE-KARBE, and W. SCHACHERMAYER (2013): The Dual Optimizer for the Growth-Optimal Portfolio under Transaction Costs, *Finance Stoch*.

GOODMAN, J., and D. N. OSTROV (2010): Balancing Small Transaction Costs with Loss of Optimal Allocation in Dynamic Stock Trading Strategies, *SIAM J. Appl. Math.* 70(6), 1977–1998.

GUASONI, P., and S. ROBERTSON (2012): Portfolios and Risk Premia for the Long Run, *Ann. Appl. Probab.* 22(1), 239–284.

GUNNING, R. C., and H. ROSSI (2009): *Analytic Functions of Several Complex Variables*, Providence, RI: AMS Chelsea Publishing.

HERCZEGH, A., and V. PROKAJ (2011): Shadow Price in the Power Utility Case. Preprint arXiv:1112.4385.

JANEČEK, K., and S. E. SHREVE (2004): Asymptotic Analysis for Optimal Investment and Consumption with Transaction Cost, *Finance Stoch.* 8(2), 181–206.

KABANOV, Y. M., and C. STRICKER (2002): On the Optimal Portfolio for the Exponential Utility Maximization: Remarks to the Six-Author Paper, *Math. Finance* 12(2), 125–134.

KALLSEN, J., and J. MUHLE-KARBE (2010): On Using Shadow Prices in Portfolio Optimization with Transaction Costs, *Ann. Appl. Probab.* 20(4), 1341–1358.

KRUK, L., J. LEHOCZKY, K. RAMANAN, and S. SHREVE (2007): An Explicit Formula for the Skorokhod Map on [0, a], *Ann. Probab.* 35(5), 1740–1768.

LIU, H. (2004): Optimal Consumption and Investment with Transaction Costs and Multiple Risky Assets, *J. Finance* 59(1), 289–338.

LO, A., and J. WANG (2000): Trading Volume: Definitions, Data Analysis, and Implications of Portfolio Theory, *Rev. Finan. Stud.* 13(2), 257–300.

LUCIANO, E. (2012): Equilibrium Price of Immediacy and Infrequent Trade. Carlo Alberto Notebook 221 (November 221). Available at SSRN 2021331.

MOKKHAVESA, S., and C. ATKINSON (2002): Perturbation Solution of Optimal Portfolio Theory with Transaction Costs for Any Utility Function, *IMA J. Manag. Math.* 13(2), 131–151.

NUTZ, M. (2012): Risk Aversion Asymptotics for Power Utility Maximization, *Probab. Theory Relat. Fields* 152(3–4), 703–749.

ROGERS, L. C. G. (2004): Why Is the Effect of Proportional Transaction Costs $O(\delta^{2/3})$?, in *Mathematics of Finance*, G. Yin and Q. Zhang, eds., Vol. 351 of Contemporary Mathematics, Providence, RI: American Mathematical Society, pp. 303–308.

SCHACHERMAYER, W. (2003): A Super-Martingale Property of the Optimal Portfolio Process, *Finance Stoch.* 7(4), 433–456.

SHREVE, S. E., and H. M. SONER (1994): Optimal Investment and Consumption with Transaction Costs, *Ann. Appl. Probab.* 4(3), 609–692.

TAKSAR, M., M. J. KLASS, and D. ASSAF (1988): A Diffusion Model for Optimal Portfolio Selection in the Presence of Brokerage Fees, *Math. Oper. Res.* 13(2), 277–294.

WHALLEY, A. E., and P. WILMOTT (1997): An Asymptotic Analysis of an Optimal Hedging Model for Option Pricing with Transaction Costs, *Math. Finance* 7(3), 307–324.

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

Mathematical Finance, Vol. 25, No. 4 (October 2015), 754–788

OPTIMAL SELLING RULES FOR MONETARY INVARIANT CRITERIA: TRACKING THE MAXIMUM OF A PORTFOLIO WITH NEGATIVE DRIFT

ROMUALD ELIE

Université Paris-Dauphine and CREST

GILLES-EDOUARD ESPINOSA

Department of Mathematics, ETH Zurich

Considering a positive portfolio diffusion X with negative drift, we investigate optimal stopping problems of the form

$$\inf_{\theta} \mathbb{E}\left[f\left(\frac{X_\theta}{\sup_{s\in[0,\tau]} X_s} \right) \right],$$

where f is a nonincreasing function, τ is the next random time where the portfolio X crosses zero and θ is any stopping time smaller than τ. Hereby, our motivation is the obtention of an optimal selling strategy minimizing the relative distance between the liquidation value of the portfolio and its highest possible value before it reaches zero. This paper unifies optimal selling rules observed for the quadratic absolute distance criteria in this stationary framework with bang–bang type ones observed for monetary invariant criteria but in finite horizon. More precisely, we provide a verification result for the general stopping problem of interest and derive the exact solution for two classical criteria f of the literature. For the power utility criterion $f : y \mapsto -y^\lambda$ with $\lambda > 0$, instantaneous selling is always optimal, which is consistent with previous observations for the Black-Scholes model in finite observation. On the contrary, for a relative quadratic error criterion, $f : y \mapsto (1-y)^2$, selling is optimal as soon as the process X crosses a specified function φ of its running maximum X^*. These results reinforce the idea that optimal stopping problems of similar type lead easily to selling rules of very different nature. Nevertheless, our numerical experiments suggest that the practical optimal selling rule for the relative quadratic error criterion is in fact very close to immediate selling.

KEY WORDS: optimal stopping, optimal prediction, running maximum, free boundary partial differential equation, verification.

1. INTRODUCTION

One of the main objectives of a hedge fund manager, an asset manager, or a proprietary trader is to take positions on the financial markets in a way that allows him to generate

The authors wish to thank two anonymous referees as well as Mete Soner and Nizar Touzi for fruitful discussions. Research partly supported by the European Research Council under the grant 228053-FiRM as well as the Chair *Finance and sustainable development*.

Manuscript received March 2011; final revision received December 2012.

Address correspondence to Romuald Elie, CEREMADE, CNRS, UMR 7534, Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, 75116 Paris, France; e-mail: elie@ceremade.dauphine.fr.

DOI: 10.1111/mafi.12036

profits, with a proper risk exposure. In order to achieve this goal, he will try to find some particular behavior of an asset or a combination of assets. Two classical examples are mean-reverting assets (or combinations of assets) and trend following assets. Besides finding a combination of assets that meets the required behavior, the fund manager requires to determine the right time to invest. To make it simple, he wants to sell high and buy low. Different techniques to detect the good times to invest are used by practitioners, such as statistical analysis, technical analysis, or financial intuition. Another idea would be to fit a mathematical model on financial data and try to solve an optimization problem. Indeed, what could be a best way of selling high than selling as close as possible to the maximum of the associated process? However, when you think it twice, the random time of interest is not a stopping time, and therefore this may seem to be a hopeless ambition.

Nevertheless, this kind of optimal prediction problem has remarkably already been addressed in the literature. Graversen, Peskir, and Shiryaev were the first authors who tackled successfully this challenging problem. Considering a standard Brownian motion W on the time interval [0, 1], they solved in Graversen, Peskir, and Shiryaev (2001) the optimal stopping problem $\inf_\theta \mathbb{E}[|W_\theta - W_1^*|^2]$, where W_1^* denotes the maximum of the Brownian motion W at time 1 and θ is any stopping time smaller than 1. They found out that selling was optimal as soon as the drawdown of the portfolio, that is, the gap between its current maximum and its value, went above the function $t \mapsto c^*\sqrt{1-t}$, for a specified constant c^*. Urusov (2005) then observed that this strategy also provides a good approximation of the last time τ^* where the portfolio reaches its maximum, since it solves the problem $\inf_\theta \mathbb{E}[|\theta - \tau^*|]$. For a portfolio driven by a drifted Brownian motion, this property is no longer satisfied, and Du Toit and Peskir (2007, 2008) characterized the solution of both problems in this more general framework. Once again, stopping is optimal as soon as the drawdown of the portfolio enters a time-to-horizon dependent region.

Considering instead a portfolio consisting of one stock $(S_t)_{0\leq t\leq 1}$ with Black-Scholes dynamics, several authors (Shiryaev, Xu, and Zhou 2008, Du Toit and Peskir 2009, or Dai et al. 2010) tried to minimize the relative distance between the stopped stock S_θ and its ultimate maximum S_1^*. In particular they characterized the optimal selling rule associated to the natural stopping problem $\sup_\theta \mathbb{E}[S_\theta/S_1^*]$. As pointed out in Du Toit and Peskir (2009), the formulation in terms of ratio between the stopped process and its maximum has the effect of stripping away the monetary value of the stock, focusing only on the underlying randomness. Using either probabilistic or deterministic methods, the common interpretation of the solution derived in these papers is that one should "sell bad stocks and keep good ones." Indeed, introducing the "goodness index" α of the stock as the ratio between its excess return and its square volatility, the optimal strategy appears to be of "bang–bang" type: one should immediately sell the stock if $\alpha \leq 1/2$ and keep it until maturity otherwise. Focusing also on the problem $\inf_\theta \mathbb{E}[S_1^*/S_\theta]$, Du Toit and Peskir (2009) observed that one should sell immediately if $\alpha < 0$, keep until time 1 if $\alpha > 1$, and stop as soon as the ratio S^*/S hits a specified deterministic function of time in the intermediate case. It is worth noticing that these two optimal prediction problems of similar type offer therefore different optimal selling strategies. It is also interesting to point out that both problems with a relative criterion exhibit an optimal selling rule which is very different from the ones obtained for an absolute criterion in Du Toit and Peskir (2007, 2008), Graversen et al. (2001), and Urusov (2005). We also refer to Hobson (2007).

Of course, the only consideration of stocks with Black-Scholes dynamics is unrealistic and limitative. A recent paper of Espinosa and Touzi (2012) allows for the consideration of more general diffusion dynamics and, as a by-product, requires to focus on a stationary version of this problem. As pointed out by the authors, considering the first hitting

of zero does not only have the advantage of reducing the problem to a stationary one, but in fact corresponds to a classical mean-reversion strategy of fund managers. They want to detect the maximum of a signal on each excursion above its mean or symmetrically detect its minimum on each excursion below its mean. Such a signal can be, for example, generated by the difference between two (normalized) stocks of the same sector. Considering a diffusion process X with general dynamics and negative drift starting at $X_0 > 0$, the authors in Espinosa and Touzi (2012) study the infinite time horizon problem: $\inf_\theta \mathbb{E}[|X^*_\tau - X_\theta|^2]$ where τ is the first time where X hits zero and θ is any stopping time smaller than τ. They solve explicitly this stationary problem as a free boundary one and verify that the fund manager should sell the portfolio whenever its running maximum X^* and its drawdown $X^* - X$ are both large enough. The underlying question, which is very important for a fund manager, is whether or not the optimal selling rule is robust with respect to the criterion of interest. Indeed, if two criteria seem very close and are relevant, then they should bring close allocation strategies (here selling/buying rules), because there is no a priori reason to choose one instead of the other.

Therefore, the motivation for this paper is twofold. First, we want to determine optimal selling rules for monetary invariant criteria in a realistic stationary framework, and then we intend to discuss the robustness of the solution with respect to the criterion. The consideration of the ratio between the stopped portfolio and its upcoming maximum allows to capture the scale of the prices themselves. We thus consider the following problem:

$$\inf_\theta \mathbb{E}\left[f\left(\frac{X_\theta}{X^*_\tau}\right)\right],$$

where τ is the first time where X hits 0, θ is any stopping time smaller than τ, and f is a penalization function. X^*_τ represents the best price the fund manager could have reached, while X_θ is what he is able to obtain in practice, so that the ratio X_θ / X^*_τ may interpret as the efficiency of the manager. The penalization function f captures how he will be penalized by his management, by his clients, or by his self-esteem when missing his target. It is therefore directly related to his investment profile. More precisely, we focus on the two following problems:

$$V_1 = \sup_\theta \mathbb{E}\left[\left(\frac{X_\theta}{X^*_\tau}\right)^\lambda\right], \quad \text{for } \lambda > 0, \quad \text{and} \quad V_2 = \inf_\theta \mathbb{E}\left[\left(\frac{X^*_\tau - X_\theta}{X^*_\tau}\right)^2\right].$$

The first problem consists in maximizing the power utility of the possible relative value of the portfolio. The case $\lambda = 1$ corresponds to the criterion considered in Dai et al. (2010), Du Toit and Peskir (2009), and Shiryaev et al. (2008) in the finite horizon Black-Scholes framework. Solving the second problem consists simply in minimizing the classical relative quadratic distance between the stopped portfolio and its maximum possible value. The power utility function with $\lambda > 1$ corresponds to an investment profile where the fund manager absolutely wants to be as close as possible to the maximum. In case he missed it, he does not really make a big difference between being away or far away. On the contrary, for $\lambda < 1$ and for the quadratic error, the fund manager absolutely wants to avoid poor efficiency, whereas as soon as he is doing well, it is not very important how good he is. As we will see hereafter, those two problems lead to optimal selling rules of very different nature.

For the first problem V_1, we prove that the optimal stopping strategy consists in liquidating the portfolio immediately. For $\lambda \leq 1$, immediate selling is still optimal even if

the portfolio value starts at a point below its current maximum. This is no longer the case when $\lambda > 1$, where the fund manager should wait until the value of the portfolio gets close enough to its running maximum. These conclusions extend and are in accordance with the conclusions of Dai et al. (2010), Du Toit and Peskir (2009), and Shiryaev et al. (2008) obtained in the finite horizon Black-Scholes framework for $\lambda = 1$. Conversely for V_2, when minimizing the relative quadratic distance between the portfolio and its maximum, the optimal selling time is the first time where the process X goes below a specified function φ of its running maximum X^*. Similarly to Peskir (1998) or Espinosa and Touzi (2012), this function φ (or more precisely its inverse) is characterized as the "biggest" solution of an ordinary differential equation (ODE) and can easily be approximated numerically. Even though for V_2, the solution seems to be of the same nature as the one derived in Espinosa and Touzi (2012), they in fact differ a lot in practice. Indeed, here $\varphi(0) = 0$ whereas φ is bounded from below by a positive constant in Espinosa and Touzi (2012). Hence, for the absolute criterion considered in Espinosa and Touzi (2012), this implies that one should hold the portfolio no matter how small its value gets, as long as the running maximum of the portfolio has not reached a given minimal threshold. It is worth noticing that this is not the case here.

As already observed by Du Toit and Peskir (2009), our results confirm that optimal prediction problems of similar nature can lead to very different types of optimal solution. Predicting the maximum of a portfolio is really intricate and the corresponding optimal selling rule strongly depends on the criterion choice of the fund manager. However, we shall temper a bit this conclusion in our framework, since numerical experiments provided hereafter show that the function φ is close to the identity function. Hence, even if immediate stopping is not optimal, a fund manager will not wait long until the drawdown of the portfolio $X^* - X$ exceeds $X^* - \varphi(X^*)$. Thus, immediate selling is for both criteria a reasonable strategy.

The paper is organized as follows. The next section provides the set-up of the problem and derives preliminary properties. Section 3 is dedicated to the obtention of a general verification theorem allowing to treat the first and the second optimal stopping problems at once. Sections 4 and 5 tackle successively the power utility type criterion and the quadratic distance one. In both cases, the value function solution is presented and discussed at the beginning of the section, numerical results are provided, and the technical proofs are postponed to the end of it.

2. OPTIMAL LIQUIDATION OF A PORTFOLIO

2.1. The Optimal Stopping Problem of Interest

Let W be a scalar Brownian motion on the complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and denote by $\mathbb{F} = \{\mathcal{F}_t, t \geq 0\}$ the corresponding augmented natural filtration. Let X be a diffusion process given by the following dynamics:

$$dX_t = -\mu(X_t)dt + \sigma(X_t)dW_t, \quad t \geq 0, \tag{2.1}$$

together with an initial data $x := X_0 > 0$, where μ and σ are Lipschitz continuous functions. We will assume that the portfolio process X is directed toward the origin in the sense that:

$$\mu(x) \geq 0\,, \quad \text{for } x \geq 0. \tag{2.2}$$

We denote by $\tau := \inf\{t \geq 0,\ X_t = 0\}$ the first time where the process X hits the origin, $\mathcal{T}$ the set of $\mathbb{F}$-stopping times θ such that $\theta \leq \tau$ a.s and $X^*_\cdot := \sup_{s \leq \cdot} X_s$ the running maximum of the portfolio X. As motivated in the introduction, we consider the following optimization problem:

$$V_0 := \inf_{\theta \in \mathcal{T}} \mathbb{E}\ f\left(\frac{X_\theta}{X^*_\tau}\right), \tag{2.3}$$

where f is a nonincreasing continuous function on $[0, 1]$, C^1 on $(0, 1]$, and such that, there exist two constants $A > 0$ and $\eta > 0$ satisfying

$$|f'(x)| \leq Ax^{\eta-2}, \quad 0 < x \leq 1. \tag{2.4}$$

Since f is nonincreasing, solving this optimal stopping problem consists in minimizing the relative distance, quantified according to the criterion f, between the liquidation price of the portfolio and its highest possible value before hitting 0. We intend to focus in Sections 4 and 5 on two particular classical cases of criteria f: $y \mapsto -y^\lambda$, $\lambda > 0$ and $y \mapsto (1 - y)^2$. Before doing so, we first exhibit preliminary properties and provide a verification argument for the optimal stopping problem (2.3) in its general form.

In order to use dynamic programming techniques, we introduce as usual the process Z defined by $Z_t := z \vee X^*_t$ for a given $z > 0$, and define the corresponding value function associated to the optimization problem (2.3):

$$V(x, z) := \inf_{\theta \in \mathcal{T}} \mathbb{E}_{x,z}\ f\left(\frac{X_\theta}{Z_\tau}\right), \quad (x, z) \in \mathbf{\Delta}, \tag{2.5}$$

where $\mathbf{\Delta}$ is defined by

$$\mathbf{\Delta} := \{(x, z), \quad 0 \leq x \leq z \quad \text{and} \quad z > 0\}, \tag{2.6}$$

and corresponds to the domain where (X, Z) lies.

REMARK 2.1. Notice that the definition of $\mathbf{\Delta}$ differs from the one in Espinosa and Touzi (2012). Observe also that contrary to Espinosa and Touzi (2012), the problem is not invariant by translation. More precisely, if one considers the criterion $\inf_\theta f(\frac{b+X_\theta}{b+Z_\tau})$, the problem might not be well defined for $b < 0$ since we can have $b + Z_\tau = 0$. For $b > 0$ (or $z > -b$ if $b < 0$) however, the problem makes sense and could be studied in a similar fashion, but is not a particular case of what we do here.

Defining the reward function g from immediate stopping

$$g(x, z) := \mathbb{E}_{x,z} f\left(\frac{x}{Z_\tau}\right), \quad (x, z) \in \mathbf{\Delta}, \tag{2.7}$$

observe from the dynamics of (X, Z) that

$$\mathbb{E}_{x,z}\ f\left(\frac{X_\theta}{Z_\tau}\right) = \mathbb{E}_{x,z}\ \mathbb{E}_{X_\theta, Z_\theta}\ f\left(\frac{X_\theta}{Z_\tau}\right) = \mathbb{E}_{x,z}\ g(X_\theta, Z_\theta), \quad (x, z) \in \mathbf{\Delta}, \quad \theta \in \mathcal{T}.$$

We may thus rewrite this problem in the standard form of an optimal stopping problem

$$V(x, z) = \inf_{\theta \in \mathcal{T}} \mathbb{E}_{x,z}\ g(X_\theta, Z_\theta), \quad (x, z) \in \mathbf{\Delta}. \tag{2.8}$$

2.2. Assumptions and First Properties

Let introduce the so-called scale function S defined, for $x \geq 0$, by

$$S(x) := \int_0^x e^{\int_0^u \alpha(r)dr}\, du, \quad \text{with} \quad \alpha := \frac{2\mu}{\sigma^2}. \tag{2.9}$$

REMARK 2.2. Since the portfolio process X is directed toward 0, the function α is non-negative. Therefore, the scale function S is increasing, convex and dominates the Identity function.

By construction, S satisfies $S_{xx} = \alpha S_x$ and is related to the law of Z_τ via the estimate

$$\mathbb{P}_{x,z}[Z_\tau \leq u] = \mathbb{P}_x\big[X^*_\tau \leq u\big]\mathbf{1}_{z\leq u} = \left(1 - \frac{S(x)}{S(u)}\right)\mathbf{1}_{z\leq u}, \quad (x,z) \in \mathbf{\Delta}, \quad u > 0.$$

Using the scale function S, the reward function g rewrites as

$$g(x,z) = f\left(\frac{x}{z}\right)\left(1 - \frac{S(x)}{S(z)}\right) + S(x)\int_z^\infty f\left(\frac{x}{u}\right)\frac{S'(u)}{S(u)^2}\, du, \quad (x,z) \in \mathbf{\Delta}, \tag{2.10}$$

which is well defined since f is continuous on [0, 1] so that we have

$$\int_z^\infty \left|f\left(\frac{x}{u}\right)\right| \frac{S'(u)}{S(u)^2}\, du \leq \|f\|_\infty \int_z^\infty \frac{S'(u)}{S(u)^2}\, du = \frac{\|f\|_\infty}{S(z)}, \quad (x,z) \in \mathbf{\Delta}.$$

Via an integration by parts, we deduce

$$g(x,z) = f\left(\frac{x}{z}\right) - xS(x)\int_z^\infty f'\left(\frac{x}{u}\right)\frac{du}{u^2 S(u)}, \quad (x,z) \in \mathbf{\Delta}, \quad x > 0. \tag{2.11}$$

Observe that the previous integral is well defined since, combining Remark 2.2 with estimate (2.4), we compute

$$\left|\frac{f'\left(\frac{x}{u}\right)}{u^2 S(u)}\right| \leq A\left(\frac{x}{u}\right)^{\eta-2}\frac{1}{u^3} = A\,\frac{x^{\eta-2}}{u^{1+\eta}}, \quad 0 < x \leq u.$$

If f is C^1 on [0, 1], since $g(0,z) = f(0)$, (2.11) also holds true for $x = 0$ and $z > 0$.

In this paper, we aim at considering a general framework including the classical types of mean reverting portfolio processes. In particular, we intend to treat the following diffusion dynamics:

- Brownian motion with negative drift: α constant, positive and $S(x) = \frac{e^{\alpha x}-1}{\alpha}$;
- Displaced Cox-Ingersol-Ross process: $dX_t = -\mu X_t dt + \sigma\sqrt{X_t + b}\, dW_t$ for some positive constants μ, σ, b so we get $\alpha(x) = \frac{\alpha x}{x+b}$ and $S'(x) = e^{\alpha x}(\frac{b}{x+b})^{\alpha b}$ with $\alpha > 0$;
- Ornstein-Uhlenbeck process: $dX_t = -\mu X_t dt + \sigma dW_t$ for some positive constants μ, σ so we get $\alpha(x) = \alpha x$ and $S'(x) = e^{\alpha x^2/2}$.

For this purpose, we impose on α similar but less restrictive conditions as in Espinosa and Touzi (2012) and work under the following standing assumption:

$$\alpha = \frac{2\mu}{\sigma^2} : (0,\infty) \to \mathbb{R} \text{ is a } C^2 \text{ positive nondecreasing concave function.} \tag{2.12}$$

REMARK 2.3. As in Espinosa and Touzi (2012), we observe for later use that the restriction (2.12) implies in particular that the function $2S' - \alpha S - 2$ is a non-negative increasing function.

REMARK 2.4. One can reasonably wonder if the solution of the present problem can be deduced from Espinosa and Touzi (2012) using the following change of variable: $Y := \ln(1 + X)$, since Y would also be a process directed toward 0 if X is. We claim that this is not the case. Indeed, we can observe that $f(\frac{1+X_\theta}{1+Z_\tau}) = f \circ \exp(\ln(1 + X_\theta) - \ln(1 + Z_\tau))$, and define $\ell : x \mapsto f \circ \exp(-x)$. First, as briefly explained in Remark 2.1, the problem considered here where $b = 0$ cannot be deduced from the one with $b = 1$. Moreover, for the functions f that we intend to study, $f : x \mapsto -x^\lambda$ or $f : x \mapsto \frac{1}{2}(1 - x)^2$, the convexity of ℓ required in Espinosa and Touzi (2012) is not satisfied. Finally, if X is, for example, an Ornstein-Uhlenbeck process, one can compute that the function α_Y associated to Y is of the form $\alpha_Y : y \mapsto 2\alpha(e^{2y} - e^y) + 1$, which is convex on $\mathbb{R}_+$, and therefore does not satisfy the assumptions of Espinosa and Touzi (2012).

3. A PDE VERIFICATION ARGUMENT

This section is devoted to the obtention of a partial differential equation (PDE) characterization for the solution of the optimal stopping problem of interest (2.5). We first derive the corresponding Hamilton-Jacobi-Bellman equation and then provide a verification theorem.

3.1. The Corresponding Dynamic Programming Equation

The linear second order Dynkin operator associated to the diffusion (2.1) is simply given by

$$\mathcal{L} \; : \; v \mapsto v_{xx} - \alpha(x)v_x, \quad \text{with} \quad \alpha(x) = \frac{2\mu(x)}{\sigma^2(x)}, \quad \text{for } x \geq 0.$$

By construction, observe that the scale function S satisfies in particular $\mathcal{L}S = 0$. Since the value function of interest V rewrites as the solution of a classical optimal stopping problem (2.8), we expect V to be the solution of the associated dynamic programming equation. Namely, V should be a solution of the Hamilton-Jacobi-Bellman equation:

$$\min(\mathcal{L}v; g - v) = 0; \quad v(0, z) = f(0), \quad v_z(z, z) = 0, \quad (x, z) \in \mathbf{\Delta}. \tag{3.1}$$

The first term indicates that V is dominated by the immediate reward function g and that the dynamics of v in the continuation domain are given by the Dynkin operator of the diffusion X. The second relation manifests that only immediate stopping is possible whenever the diffusion X has reached 0. Finally, the last one is the classical Von Neumann condition encountered whenever the diffusion process hits its maximum.

As in any optimal stopping problem, the domain of definition $\mathbf{\Delta}$ of the value function subdivides into two subsets: the stopping region $\mathcal{S}$ where immediate liquidation of the portfolio is optimal and the continuation region where the optimal strategy consists in waiting until the portfolio process enters the stopping region. The optimal stopping time is the first time where the process arrives in the stopping region, and, in order to obtain a stopping time in $\mathcal{T}$, we expect the region $\mathcal{S}$ to be a closed subset of $\mathbf{\Delta}$. Of course, the stopping region is characterized by the relation $v = g$ since g is the reward function from

immediate stopping. Depending on the position of $(x, z) \in \mathbf{\Delta}$ with respect to the region $\mathcal{S}$, we expect the dynamics of (3.1) to rewrite

$$\begin{aligned} \text{On the stopping region:} \quad & v(x,z) = g(x,z), \quad \mathcal{L}g(x,z) \geq 0\,; \\ \text{On the continuation region:} \quad & v(x,z) < g(x,z), \quad \mathcal{L}v(x,z) = 0\,; \\ \text{Everywhere:} \quad & v_z(x,z)\mathbf{1}_{\{x=z\}} = 0\,. \end{aligned}$$

Observe that the last Neumann condition is expected on all the domain Δ, and thus surprisingly in the stopping region where the portfolio X does not diffuse. This feature is particular to our framework, in which the reward function g given in (2.10) satisfies by construction $g_z(z, z) = 0$ for $z > 0$.

In the next sections of the paper, we exhibit different shapes of stopping and continuation regions depending on the choice of the objective function f. We observe that, although the objective functions may appear rather similar, the optimal selling strategies can be very different.

3.2. The Verification Theorem

As detailed earlier, we expect the value function V given by (2.5) to be solution of the Hamilton-Jacobi-Bellman equation (3.1). The solution of this problem is intimately related to the form of the associated stopping region $\mathcal{S}$. Afterwards, we shall not prove that V is indeed a (weak) solution of this PDE but instead try to guess a regular solution to the PDE and verify that it satisfies the assumptions of the following verification theorem.

We first introduce the definition of piecewise $\mathcal{C}^{2,1}$ functions.

DEFINITION 3.1. We say that a function v defined on $\mathbf{\Delta}$ is piecewise $\mathcal{C}^{2,1}$ on a subset $E \subset \mathbf{\Delta}$ if for any subset K which is compact in $\mathbf{\Delta}$, there exist $K_1, \ldots, K_n$ compacts in $\mathbf{\Delta}$ such that $K \cap E \subset \cup_{i=1}^n K_i$ and v can be extended as a $\mathcal{C}^{2,1}$ function on K_i for each $i = 1, \ldots, n$.

THEOREM 3.2. *Let v be a bounded from below function $\mathcal{C}^{1,0}$ on $\mathbf{\Delta}$ and piecewise $\mathcal{C}^{2,1}$ on $\mathbf{\Delta}\backslash\{(0, z), z > 0\}$ in the sense of Definition 3.1.*

(i) If v satisfies $\mathcal{L}v \geq 0$, $v \leq g$ as well as $v_z(z, z) \geq 0$ for $z > 0$, then $v \leq V$.

(ii) More precisely, if $v_z(z, z) = 0$ for $z > 0$ and there exists a closed in Δ set $\mathcal{S} \subset \mathbf{\Delta}$ containing the axis $\{(0, z)\,, z > 0\}$ such that

$$\begin{aligned} & v = g \quad \text{on } \mathcal{S}\,, \qquad \mathcal{L}v \geq 0 \quad \text{on } \mathcal{S}\backslash\{(0,z), z > 0\}, \\ & v \leq g \quad \text{and } \mathcal{L}v = 0 \quad \text{on } \mathbf{\Delta}\backslash\mathcal{S}, \end{aligned} \tag{3.2}$$

then $v = V$ and $\theta^* := \inf\{t \geq 0, (X_t, Z_t) \in \mathcal{S}\}$ is an optimal stopping time.

(iii) If in addition $v < g$ on $\mathbf{\Delta}\backslash\mathcal{S}$, then θ^* is the "smallest" optimal stopping time, in the sense that $\theta^* \leq \nu$ a.s. for any optimal stopping time ν.

Proof. We prove each assertion separately.

(i) Fix $(X_0, Z_0) := (x, z) \in \mathbf{\Delta}$. Let $\theta \in \mathcal{T}$ and define $\theta_n = n \wedge \theta \wedge \inf\{t \geq 0; Z_t \geq n \text{ or } Z_t \leq \frac{1}{n}\}$ for $n \in \mathbb{N}$. Since (X, Z) takes value in a compact subset of $\mathbf{\Delta}$, a direct

application of Itô's formula gives

$$v(x,z) = v(X_{\theta_n}, Z_{\theta_n}) - \int_0^{\theta_n} \mathcal{L}v(X_t, Z_t)\frac{\sigma(X_t)^2}{2}dt - \int_0^{\theta_n} v_x(X_t, Z_t)\sigma(X_t)dW_t - \int_0^{\theta_n} v_z(X_t, Z_t)dZ_t\,.$$

Combining estimates $\mathcal{L}v \geq 0$ and $v_z(X_t, Z_t)dZ_t = v_z(Z_t, Z_t)dZ_t \geq 0$ with the fact that (X, Z) lies in a compact subset of $\mathbf{\Delta}$, we deduce

$$v(x,z) \leq \mathbb{E}_{x,z}v(X_{\theta_n}, Z_{\theta_n}). \tag{3.3}$$

Since $v \leq g$, this leads directly to

$$v(x,z) \;\leq\; \mathbb{E}_{x,z}g(X_{\theta_n}, Z_{\theta_n}) = \mathbb{E}_{x,z}\mathbb{E}_{X_{\theta_n}, Z_{\theta_n}} f\left(\frac{X_{\theta_n}}{Z_\tau}\right) = \mathbb{E}_{x,z} f\left(\frac{X_{\theta_n}}{Z_\tau}\right).$$

Clearly as $n \to \infty$, $\theta_n \to \theta$ almost surely. Since $0 \leq X_{\theta_n}/Z_\tau \leq 1$ and f is continuous, Lebesgue's dominated convergence theorem gives: $\mathbb{E}_{x,z} f(X_{\theta_n}/Z_\tau) \to_{n\to\infty} \mathbb{E}_{x,z} f(X_\theta/Z_\tau)$, leading to

$$v(x,z) \leq V(x,z)\,, \quad (x,z) \in \mathbf{\Delta}.$$

(ii) Observe that this framework is more restrictive than the previous one, so that $v \leq V$ on $\mathbf{\Delta}$. For $(x,z) \in \mathcal{S}$, we have $v(x,z) = g(x,z) \geq V(x,z)$ by definition of g. We now fix $(x,z) \in \mathbf{\Delta}\backslash\mathcal{S}$ and prove that $v(x,z) \geq V(x,z)$.

Let $\theta^* := \inf\{t \geq 0; (X_t, Z_t) \in \mathcal{S}\}$. Observe that $\theta^* \in \mathcal{T}$ since $\mathcal{S}$ is closed in Δ and contains the axis $\{(0,z),\ z \geq 0\}$. The regularity of v implies $\mathcal{L}v(X_t, Z_t) = 0$ for any $t \in [0, \theta^*)$. As before, we define $\theta_n^* := n \wedge \theta^* \wedge \inf\{t \geq 0;\, Z_t \geq n \text{ or } Z_t \leq \frac{1}{n}\}$, which is a stopping time. A very similar computation leads directly to

$$v(x,z) = \mathbb{E}_{x,z}v\big(X_{\theta_n^*}, Z_{\theta_n^*}\big).$$

Since v is bounded from below and $v \leq g \leq \|f\|_\infty$, v is bounded. Therefore the sequence $(v(X_{\theta_n^*}, Z_{\theta_n^*}))_n$ is uniformly integrable and we deduce that $v(x,z) = \mathbb{E}_{x,z}v(X_{\theta^*}, Z_{\theta^*})$. Since $(X_{\theta^*}, Z_{\theta^*}) \in \mathcal{S}$ and $v = g$ on $\mathcal{S}$, we get

$$v(x,z) = \mathbb{E}_{x,z}g(X_{\theta^*}, Z_{\theta^*}) = \mathbb{E}_{x,z} f\left(\frac{X_{\theta^*}}{Z_\tau}\right) \;\geq\; V(x,z).$$

Thus $v = V$ on $\mathbf{\Delta}$ and θ^* is an optimal stopping time.

(iii) For a given $(x,z) \in \mathbf{\Delta}$, we argue by contradiction and suppose the existence of a stopping time $\nu \in \mathcal{T}$ satisfying $\mathbb{P}(\nu < \theta^*) > 0$ and $V(x,z) = \mathbb{E}_{x,z} f(X_\nu/Z_\tau)$.

By assumption, we have $V(X_\nu, Z_\nu) < g(X_\nu, Z_\nu)$ on $\{\tau < \theta^*\}$, which combined with estimate $V \leq g$ implies

$$V(x,z) = \mathbb{E}_{x,z} f\left(\frac{X_\nu}{Z_\tau}\right) = \mathbb{E}_{x,z}g(X_\nu, Z_\nu) \;>\; \mathbb{E}_{x,z}V(X_\tau, Z_\tau) \;\geq\; V(x,z),$$

where the last inequality follows from the definition of V. This leads to a contradiction, which guarantees the minimality of θ^*. □

REMARK 3.3. From the definition of g, one easily checks that $g_z(z, z) = 0$ for any $z > 0$. Therefore, in the PDE dynamics (3.2), the Neumann boundary condition $v_z(z, z) = 0$ is only necessary for $(z, z) \in \mathbf{\Delta} \setminus \mathcal{S}$, since it is automatically satisfied otherwise.

REMARK 3.4. Whenever g is a function which is $\mathcal{C}^{1,0}$ with respect to (x, z) on $\mathbf{\Delta}$, $\mathcal{C}^{2,1}$ with respect to (x, z) on $\mathbf{\Delta} \setminus \{(0, z), z > 0\}$ and $\mathcal{L}g \geq 0$ on $\mathbf{\Delta} \setminus \{(0, z), z > 0\}$, then $v = g$ and $\mathcal{S} = \mathbf{\Delta}$ satisfy the assumptions of Theorem 3.2 (ii). In that case, immediate stopping is always optimal. We prove in Proposition 3.5 that the reverse is true. Notice also that expression (2.10) implies that an immediate sufficient condition for g to be in $\mathcal{C}^{1,0}(\mathbf{\Delta}) \cap \mathcal{C}^{2,1}(\mathbf{\Delta} \setminus \{(0, z), z > 0\})$ is that f is $\mathcal{C}^2$ on $(0, 1]$.

PROPOSITION 3.5. *Assume that g is $\mathcal{C}^{1,0}$ on $\mathbf{\Delta}$, $\mathcal{C}^{2,1}$ on $\mathbf{\Delta} \setminus \{(0, z), z > 0\}$, and that there exists $(x_0, z_0) \in \mathbf{\Delta} \setminus \{(0, z), z > 0\}$ such that $\mathcal{L}g(x_0, z_0) < 0$. Then, immediate stopping at (x_0, z_0) is not optimal (or equivalently $V(x_0, z_0) < g(x_0, z_0)$).*

Proof. Since $\mathcal{L}g$ is continuous at (x_0, z_0), there exists a neighborhood U_0 of (x_0, z_0) in $\mathbf{\Delta}$ such that $\mathcal{L}g(x, z) < 0$ for any $(x, z) \in U_0$. Without loss of generality, we can assume that U_0 is compact in $\mathbf{\Delta}$. Let $(X_0, Z_0) = (x_0, z_0)$. Since $x_0 > 0$, there exists $\theta_0 \in \mathcal{T}$ such that $\mathbb{E}_{x_0, z_0}\theta_0 > 0$ and let define $\theta_1 := 1 \wedge \theta_0 \wedge \inf\{t \geq 0; (X_t, Z_t) \notin U_0\} \in \mathcal{T}$. Since $\{\theta_0 > 0\} = \{\theta_1 > 0\}$, we also have $\mathbb{E}_{x_0, z_0}\theta_1 > 0$. Using Itô's formula, we compute:

$$g(x_0, z_0) = g\big(X_{\theta_1}, Z_{\theta_1}\big) - \int_0^{\theta_1} \mathcal{L}g(X_u, Z_u)\frac{\sigma(X_u)^2}{2}du$$
$$- \int_0^{\theta_1} g_x(X_u, Z_u)\sigma(X_u)dW_u - \int_0^{\theta_1} g_z(X_u, Z_u)\, dZ_u.$$

From Remark 3.3, $g_z(z, z) = 0$ for $z > 0$ so that the last term of the previous expression disappears. Since U_0 is compact and $\mathbb{E}_{x_0, z_0}\theta_1 > 0$, taking conditional expectations, we deduce that $g(x_0, z_0) > \mathbb{E}_{x_0, z_0}\, g(X_{\theta_1}, Z_{\theta_1}) \geq V(x_0, z_0)$. □

In the next sections, we investigate two particular cases of objective functions, for which we exhibit functions v and stopping regions $\mathcal{S}$, which satisfy the assumptions of Theorem 3.2 and are in general nontrivial.

4. THE POWER UTILITY CRITERION

Let first examine the case where the function f is given by $f : x \mapsto -\frac{x^\lambda}{\lambda}$, for $\lambda > 0$. In other words, we are computing the following value function

$$V^\lambda(x, z) := -\frac{1}{\lambda}\sup_{\theta \in \mathcal{T}} \mathbb{E}_{x,z}\left(\frac{X_\theta}{Z_\tau}\right)^\lambda, \quad (x, z) \in \mathbf{\Delta}, \quad \lambda > 0. \tag{4.1}$$

Consider an investor, whose relative preferences are given by a power utility function and suppose that he detains at time 0 a given portfolio X directed toward 0. The optimal stopping time at which he should liquidate his portfolio is the solution of the previous prediction problem. With a given finite time horizon T, Du Toit and Peskir (2009) as well as Shiryaev et al. (2008) investigate the case where X is a Geometric Brownian motion. They conclude that the optimal strategy consists in waiting until time T if the portfolio has promising returns (i.e. $1 < 2\mu/\sigma^2 = x\alpha(x)$, $x > 0$, with our notations), and sell immediately otherwise. In our stationary framework, waiting until the wealth reaches

0 is obviously a nonoptimal strategy. For a linear utility function ($\lambda = 1$), we prove in Theorem 4.1 below that immediate stopping is also optimal. Depending on the value of λ, the latter may no longer be the case for the nonlinear problem (4.1). Nevertheless, we observe that immediate selling is still optimal for the practical value function of interest $V^\lambda(x, x)$, for $x > 0$.

4.1. The Particular Case Where $\lambda \leq 1$

For $\lambda \leq 1$, we prove hereafter that immediate stopping is always optimal. For $\lambda = 1$, these conclusions are therefore in accordance with those of Du Toit and Peskir (2009), Shiryaev et al. (2008) obtained for the case of an exponential Brownian motion on a fixed time horizon.

A direct application of estimate (2.11) proves that the reward function g^λ associated to problem (4.1) is given by

$$g^\lambda(x, z) = -\frac{x^\lambda}{\lambda z^\lambda} + x^\lambda S(x) \int_z^\infty \frac{du}{S(u)u^{1+\lambda}}, \quad (x, z) \in \mathbf{\Delta}, \quad \lambda > 0.$$

The next theorem indicates that the framework of Remark 3.4 holds for $\lambda \leq 1$, so that g^λ coincides with the value function on $\mathbf{\Delta}$.

THEOREM 4.1. *For $\lambda \leq 1$, immediate stopping is optimal for problem (4.1), so that*

$$V^\lambda(x, z) = g^\lambda(x, z), \quad (x, z) \in \mathbf{\Delta}, \quad 0 < \lambda \leq 1.$$

Proof. For any $\lambda > 0$ and $(x, z) \in \mathbf{\Delta}$ with $x > 0$, we compute

$$g^\lambda_x(x, z) = -\frac{x^{\lambda-1}}{z^\lambda} + \{\lambda x^{\lambda-1} S(x) + x^\lambda S'(x)\} \int_z^\infty \frac{du}{S(u)u^{1+\lambda}}.$$

Differentiating one more time and using the relation $\mathcal{L}S = 0$, we get,

$$g^\lambda_{xx}(x, z) = (1-\lambda)\frac{x^{\lambda-2}}{z^\lambda} + \{\lambda(\lambda-1)x^{\lambda-2}S(x) + (2\lambda x^{\lambda-1} + x^\lambda \alpha(x))S'(x)\} \int_z^\infty \frac{du}{S(u)u^{1+\lambda}},$$

for any $\lambda \neq 1$ and $0 < x \leq z$. Combining the previous estimates, we deduce that

$$\mathcal{L}g^\lambda(x, z) = x^{\lambda-2}[x\alpha(x) + 1 - \lambda]\left(\frac{1}{z^\lambda} - \int_z^\infty \frac{S(x)}{S(u)}\frac{\lambda du}{u^{1+\lambda}}\right) + 2x^{\lambda-1}S'(x)\int_z^\infty \frac{\lambda du}{S(u)u^{1+\lambda}},$$

(4.2)

for any $\lambda \neq 1$ and $0 < x \leq z$. In the case, where $\lambda = 1$, we get similarly

$$\mathcal{L}g^1(x, z) = \alpha(x)\left(\frac{1}{z} - \int_z^\infty \frac{S(x)du}{u^2 S(u)}\right) + 2S'(x)\int_z^\infty \frac{du}{u^2 S(u)}, \quad (x, z) \in \Delta. \tag{4.3}$$

Furthermore, since S is increasing, we have

$$\int_z^\infty \frac{S(x)}{S(u)}\frac{\lambda du}{u^{1+\lambda}} \leq \int_z^\infty \frac{S(z)}{S(u)}\frac{\lambda du}{u^{1+\lambda}} \leq \int_z^\infty \frac{\lambda du}{u^{1+\lambda}} = \frac{1}{z^\lambda}, \quad (x, z) \in \mathbf{\Delta}, \quad \lambda > 0.$$

Plugging this estimate in (4.2) and (4.3), we see that $\mathcal{L}g^\lambda \geq 0$ on $\mathbf{\Delta}\setminus\{(0, z), z > 0\}$ for any $\lambda \leq 1$. As detailed in Remark 3.4, since g is C^0 on $\mathbf{\Delta}$ and $C^{2,1}$ on $\mathbf{\Delta}\setminus\{(0, z), z > 0\}$,

we deduce that $V^\lambda = g^\lambda$ on $\mathbf{\Delta}$ and consequently immediate stopping is optimal for any $\lambda \leq 1$. □

4.2. Construction of the Solution When $\lambda > 1$

We now turn to the more interesting and intricate case where $\lambda > 1$. Then, the function $\mathcal{L}g^\lambda$ is still given by expression (4.2) and we observe that:

$$\mathcal{L}g^\lambda(x, z) \sim_{x\sim 0} (1-\lambda)\frac{x^{\lambda-2}}{z^\lambda} < 0,$$

for any $z > 0$ and $\lambda > 1$. Therefore, $\mathcal{L}g^\lambda$ is not non-negative on $\mathbf{\Delta}$ and Proposition 3.5 ensures that the associated continuation region is nonempty. Since immediate stopping shall not be optimal close to the axis $\{(0, z);\ z > 0\}$, we expect to have a stopping region of the form $\mathcal{S}^\lambda := \{(x, z) \in \mathbf{\Delta};\ x \geq \varphi^\lambda(z)\} \cup \{(0, z); z > 0\}$. Hence, our objective is to find functions v^λ and φ^λ satisfying

$$\mathcal{L}v^\lambda(x, z) = 0 \quad \text{for } 0 < x < \varphi^\lambda(z) \quad \text{and} \quad (x, z) \in \mathbf{\Delta}, \tag{4.4}$$

$$v^\lambda(x, z) = g^\lambda(x, z) \quad \text{and} \quad \mathcal{L}g^\lambda(x, z) \geq 0 \quad \text{for } x \geq \varphi^\lambda(z) \quad \text{and } (x, z) \in \mathbf{\Delta}, \tag{4.5}$$

$$v^\lambda(0, z) = 0 \quad \text{for } z > 0, \tag{4.6}$$

$$v^\lambda_z(z, z) = 0 \quad \text{for } z > 0. \tag{4.7}$$

Since we look for regular solutions, we complement the above system by the continuity and the smooth fit conditions

$$v^\lambda(\varphi^\lambda(z), z) = g^\lambda(\varphi^\lambda(z), z) \quad \text{and} \quad v^\lambda_x(\varphi^\lambda(z), z) = g^\lambda_x(\varphi^\lambda(z), z), \quad \text{for } z > 0. \tag{4.8}$$

The stopping region $\mathcal{S}^\lambda$ will then be defined as

$$\mathcal{S}^\lambda := \{(x, z) \in \mathbf{\Delta}; x \geq \varphi^\lambda(z)\} \cup \{(0, z); z > 0\}. \tag{4.9}$$

Since the optimization problem of practical interest corresponds to the value of V^λ on the diagonal $\{(x, x);\ x > 0\}$, our main concern here is to find out if $\varphi^\lambda(z) \leq z$ for any $z \geq 0$, hence indicating if immediate stopping is always optimal on the diagonal. Surprisingly, we verify hereafter that $\varphi^\lambda(0) = 0$ and $\varphi^\lambda(z) < z$, for $z > 0$, so that immediate stopping is the optimal strategy for the practical problem of interest.

Due to the dynamics of (4.4) and since $\mathcal{L}S = 0$, the function v^λ must be of the form

$$v^\lambda(x, z) = A(z) + B(z)S(x), \quad (x, z) \in \mathbf{\Delta}\backslash\mathcal{S}.$$

Combined with the continuity and smooth fit conditions (4.8), this leads to

$$v(x, z) = g^\lambda(\varphi^\lambda(z), z) + \frac{g^\lambda_x(\varphi^\lambda(z), z)}{S' \circ \varphi^\lambda(z)}[S(x) - S \circ \varphi^\lambda(z)], \quad (x, z) \in \mathbf{\Delta}\backslash\mathcal{S}.$$

The free boundary φ^λ is then determined by the Dirichlet condition (4.6) and must satisfy:

$$g^\lambda(\varphi^\lambda(z), z)S' \circ \varphi^\lambda(z) = g^\lambda_x(\varphi^\lambda(z), z)S \circ \varphi^\lambda(z), \quad (x, z) \in \mathbf{\Delta}\backslash\mathcal{S}.$$

The next lemma introduces a free boundary function φ^λ satisfying this required condition. It also provides useful properties of this free boundary function and its technical proof is postponed to Section 4.4.

LEMMA 4.2. *For any $\lambda > 1$, the function φ^λ given by*

$$\varphi^\lambda : z \in (0, \infty) \mapsto \arg\min_{x \in [0,z]} \frac{g(x,z)}{S(x)},$$

is a well defined increasing C^1 function, satisfying:

(i) $0 \leq \varphi^\lambda(z) < z$, *for any* $z > 0$;
(ii) φ^λ *maps* $(0, \infty)$ *onto* $(0, y^\lambda)$, *where* y^λ *is the unique non-null zero of* $y \mapsto yS'(y) - \lambda S(y)$.

REMARK 4.3. As pointed out by a referee, the definition of the optimal frontier φ^λ interprets as the one associated to a one-dimensional classical stopping problem with fixed running maximum $z > 0$. Indeed, we observe ex post that the exhibited optimal policy does not allow the running maximum of the portfolio to increase. Hence solving this stopping problem with a fixed $z > 0$ is essentially the same.

REMARK 4.4. Observe that, for any fixed $z > 0$ and $\lambda > 1$, $g^\lambda(x, z)/S(x)$ converges to 0 as x goes to 0, since S dominates the Identity function as pointed out in Remark 2.2. Therefore, the function $g^\lambda(., z)/S(.)$ is well defined on $[0, z]$ for any $z > 0$.

Before providing the value function solution and verifying that it satisfies the requirements of Theorem 3.2, we still need to check that the stopping region $\mathcal{S}^\lambda$ associated to φ^λ is indeed a good candidate, that is, the second part of (4.5) holds. This is the purpose of the next lemma, which proof is also postponed to Section 4.4.

LEMMA 4.5. *For any $\lambda > 1$, the function $\mathcal{L}g^\lambda$ is non-negative on* $\{(x, z) \in \mathbf{\Delta},\ x \geq \varphi^\lambda(z)\}$.

Given the free boundary φ^λ defined above and the corresponding stopping region $\mathcal{S}^\lambda$, we are now in position to provide the optimal strategy and value function solutions of the problem (4.1).

THEOREM 4.6. *For any $\lambda > 1$, the value function V^λ solution of problem (4.1) is given by*

$$V^\lambda(x, z) = g^\lambda(\varphi^\lambda(z), z)\frac{S(x)}{S \circ \varphi^\lambda(z)}\, \mathbf{1}_{\{x < \varphi^\lambda(z)\}} + g^\lambda(x, z)\mathbf{1}_{\{x \geq \varphi^\lambda(z)\}}, \quad (x, z) \in \mathbf{\Delta}. \tag{4.10}$$

The smallest optimal stopping time associated to (4.1) is given by

$$\theta^\lambda := \inf\{t \geq 0,\ X_t \geq \varphi^\lambda(Z_t)\} \wedge \tau, \quad \lambda > 1.$$

Proof. Let denote by v^λ the candidate value function defined by the right-hand side of (4.10). We shall prove that v^λ coincides with the value function (4.1) by checking that it satisfies all the requirements of Theorem 3.2.

It is immediate that v^λ is bounded from below by 0 because $g^\lambda \geq 0$. Since g^λ is C^1 on $\mathbf{\Delta}$ and $C^{2,1}$ with respect to (x, z) on $\mathbf{\Delta}\backslash\{(0, z), z > 0\}$, and φ^λ is C^1 by Lemma 4.2, v^λ is $C^{2,1}$ with respect to (x, z) on both $\mathbf{\Delta}\backslash\mathcal{S}$ and $\mathcal{S}\backslash\{(0, z), z > 0\}$, so that it is piecewise $C^{2,1}$ on $\mathbf{\Delta}\backslash\{(0, z), z > 0\}$. By construction, v^λ is continuous on $\mathbf{\Delta}$ and we recall from the

definition of φ^λ that

$$g^\lambda(\varphi^\lambda(z), z)\frac{S' \circ \varphi^\lambda(z)}{S \circ \varphi^\lambda(z)} = g^\lambda_x(\varphi^\lambda(z), z), \quad z > 0.$$

Therefore, v^λ is C^1 on $\mathbf{\Delta}$.

The closed in Δ stopping region associated to the value function v^λ is naturally given by (4.9). By definition, $v^\lambda = g^\lambda$ on $\mathcal{S}^\lambda$ and we deduce from Lemma 4.5 that $\mathcal{L}g^\lambda \geq 0$ on the set $\mathcal{S}^\lambda \setminus \{(0, z), z > 0\}$. By construction, we have $\mathcal{L}v^\lambda = 0$ on $\mathbf{\Delta}\setminus\mathcal{S}^\lambda$. For any $z > 0$, since $g^\lambda(., z)/S$ achieves its minimum at a unique point $\varphi^\lambda(z)$, we get

$$g^\lambda(\varphi^\lambda(z), z)\frac{S(x)}{S \circ \varphi^\lambda(z)} < g^\lambda(x, z), \quad 0 \leq x < \varphi^\lambda(z),$$

and we deduce that $v^\lambda < g^\lambda$ on $\mathbf{\Delta}\setminus\mathcal{S}^\lambda$. Finally, since $v^\lambda_z(z, z) = g^\lambda_z(z, z) = 0$ for any $z > 0$, all the requirements of (ii)–(iii) in Theorem 3.2 are in force, and the proof is complete. □

4.3. Properties of the Optimal Liquidation Strategy

We first observe that the two previous cases where λ is above or below 1 seem to be of different natures. However, we prove hereafter that this is not the case and provide via simple arguments the continuity of V^λ with respect to the parameter λ.

PROPOSITION 4.7. *The mapping* $\lambda \mapsto V^\lambda$ *is continuous on* $(0, \infty)$.

Proof. We fix λ_1 and λ_2 in $(0, \infty)$ such that $\lambda_1 \leq \lambda_2$. First notice that since $X_\cdot \leq Z_\tau$ on $[0, \tau]$, we necessarily have

(4.11)

$$-\lambda_2 V^{\lambda_2}(x, z) = \sup_{\theta\in\mathcal{T}} \mathbb{E}_{x,z}\left(\frac{X_\theta}{Z_\tau}\right)^{\lambda_2} \leq \sup_{\theta\in\mathcal{T}} \mathbb{E}_{x,z}\left(\frac{X_\theta}{Z_\tau}\right)^{\lambda_1} = -\lambda_1 V^{\lambda_1}(x, z), \quad (x, z) \in \mathbf{\Delta}.$$

Now, using Jensen's inequality, we observe that

$$\left[\mathbb{E}_{x,z}\left(\frac{X_\theta}{Z_\tau}\right)^{\lambda_1}\right]^{\frac{\lambda_2}{\lambda_1}} \leq \mathbb{E}_{x,z}\left(\frac{X_\theta}{Z_\tau}\right)^{\lambda_2} \leq -\lambda_2 V^{\lambda_2}(x, z), \quad \theta \in \mathcal{T}, \quad (x, z) \in \mathbf{\Delta}.$$

Bringing this expression to the power λ_1/λ_2 and taking the supremum over θ, we deduce from (4.11) that

$$[-\lambda_1 V^{\lambda_1}(x, z)]^{\frac{\lambda_2}{\lambda_1}} \leq -\lambda_2 V^{\lambda_2}(x, z) \leq -\lambda_1 V^{\lambda_1}(x, z), \quad (x, z) \in \mathbf{\Delta}.$$

Therefore $\lambda_2 V^{\lambda_2} \to \lambda_1 V^{\lambda_1}$ whenever $\lambda_2 \to \lambda_1$ and we deduce the continuity of V^λ with respect to λ. □

For $\lambda > 1$, Theorem 4.6 indicates that the stopping region $\mathcal{S}^\lambda$ associated to problem (4.1) is given by (4.9). Since $\varphi^\lambda(z) < z$, we see that the stopping region $\mathcal{S}^\lambda$ includes in particular the axis $\{(x, x), x > 0\}$. Therefore, if an investor detains a portfolio directed toward zero and hopes to get close to its upcoming maximum before it reaches zero according to the criterion (4.1), he should liquidate the portfolio immediately. If ever the running maximum z of the portfolio exceeds its current value, he should wait until

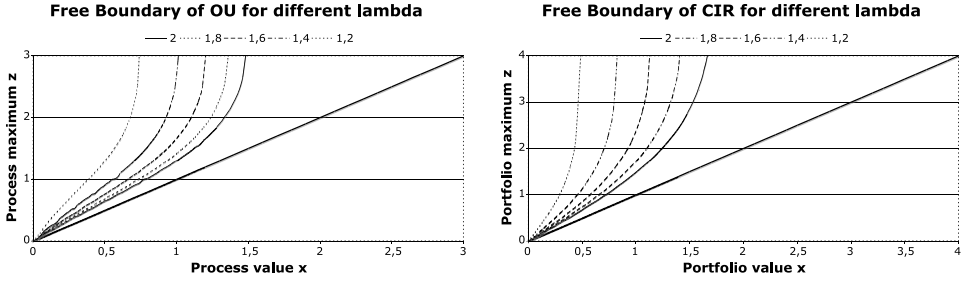

FIGURE 4.1. Optimal frontier for an OU ($\alpha = 1$) and a CIR ($\alpha = 1$, $b = 1$) with different parameter λ

the value of the portfolio hits zero or increases enough and reaches $\varphi^\lambda(z)$. In any case, the optimal strategy does not allow the running maximum of the portfolio to increase. Theorem 4.1 indicates that immediate selling is optimal for $\lambda \in (0, 1]$ and these results are in accordance with those of Shiryaev et al. (2008) for an exponential Brownian motion on a finite fixed horizon, since waiting until maturity is irrelevant in our framework. Nevertheless, changing the criterion of interest may lead to value functions where immediate stopping is not optimal on the axis $\{(x, x)\,, x > 0\}$. This is exactly the purpose of Section 5.

Figure 4.1 represents the frontier between the stopping and the continuation regions for different values of λ larger than 1 and associated to an Orstein-Uhlenbeck with parameter $\alpha = 1$ and a CIR-Feller process with parameters $\alpha = 1$ and $b = 1$. We first observe that the shape of the free boundary φ^λ is rather similar in both cases, and we observe indeed this feature for a large range of parameter set. Furthermore, the mapping $\lambda \mapsto \varphi^\lambda$ seems to be continuous, property which is easily verified from the definition of φ^λ. Second, we notice that the free boundary φ^λ is decreasing with respect to λ. Indeed, arguing as in Part 2. of the proof of Lemma 4.2, one can easily check that the function $x \in \mathbb{R}^+ \mapsto xS'(x)/S(x)$ is decreasing starting from 1. Hence, by definition of y^λ, the valuation domain $[0, y^\lambda]$ of φ^λ shrinks monotonically to $\{0\}$ as λ decreases to 1, hence leading to the absence of continuation region for the problem V^1.

REMARK 4.8. Considering, for example, an Ornstein-Uhlenbeck portfolio X, one verifies easily from their definitions that the free boundary φ^λ and the value function v^λ are continuous with respect to the parameter $\alpha \in \mathbb{R}$, characterizing the dynamics of the portfolio X. Hence, the continuation and stopping regions are not too sensitive to eventual estimation errors of this parameter of interest.

4.4. Proofs of Lemma 4.2 and Lemma 4.5

This section provides successively the proofs of Lemma 4.2 and Lemma 4.5.

Proof of Lemma 4.2. Fix $\lambda > 1$. Let introduce the functions

$$m : x \mapsto \frac{xS'(x)}{\lambda S(x)^2} - \frac{1}{S(x)} \quad \text{and} \quad \ell : z \mapsto -\int_z^\infty \frac{\lambda z^\lambda du}{S(u)u^{1+\lambda}},$$

so that the derivative of the function of interest rewrites

$$\frac{\partial}{\partial x}\left[\frac{g(x,z)}{S(x)}\right] = \frac{x^{\lambda-1}}{z^{\lambda}}\left\{m(x)-\ell(z)\right\}, \quad (x,z)\in\mathbf{\Delta}. \tag{4.12}$$

0. *A useful estimate*

We will use several times the following expansion as $x\to\infty$:

$$\alpha(x)S(x) \sim_{\infty} S'(x). \tag{4.13}$$

Indeed, recalling that $\mathcal{L}S=0$ and integrating by parts, we compute:

$$\begin{aligned} S(x) &= S(1)+\int_1^x \frac{\alpha(u)S'(u)}{\alpha(u)}du \\ &= S(1)+\frac{S'(x)}{\alpha(x)}-\frac{S'(1)}{\alpha(1)}-\int_1^x (1/\alpha)'(u)S'(u)du \sim_{\infty} \frac{S'(x)}{\alpha(x)}, \end{aligned}$$

since $S(x)\to\infty$ as $x\to\infty$ and (2.12) implies that $(1/\alpha)'(x)\to 0$ as $x\to\infty$.

1. *Definition of* φ^{λ}

In order to justify that φ^{λ} is well defined, we study separately the functions m and ℓ. We observe first that the function ℓ is negative, increasing and, according to (4.13), satisfies

$$\ell(z) \sim_{\infty} -\int_z^{\infty}\frac{\lambda z^{\lambda}\alpha(u)du}{u^{1+\lambda}S'(u)} \sim_{\infty} -\frac{\lambda}{zS'(z)} \to_{\infty} 0, \tag{4.14}$$

where the second equivalence comes from the following computation:

$$-\int_z^{\infty}\frac{\lambda z^{\lambda}\alpha(u)du}{u^{1+\lambda}S'(u)} = \left[\frac{\lambda z^{\lambda}}{u^{1+\lambda}S'(u)}\right]_z^{\infty}+\int_z^{\infty}\frac{\lambda(1+\lambda)z^{\lambda}du}{u^{2+\lambda}S'(u)}, \quad z>0.$$

We now turn to the study of m and compute, for any $x>0$,

$$\begin{aligned} m'(x) &= \frac{\{\lambda+1+x\alpha(x)\}}{\lambda S(x)^3}S'(x)M(x), \quad \text{with} \\ M: x &\mapsto S(x)-\frac{2x}{\lambda+1+x\alpha(x)}S'(x). \end{aligned}$$

Differentiating one more time, we obtain

$$M'(x) = \frac{x^2S'(x)}{(\lambda+1+x\alpha(x))^2}\left[\frac{\lambda^2-1}{x^2}-\alpha(x)^2+2\alpha'(x)\right], \quad x>0.$$

Since $\lambda>1$ while α is non-negative, increasing and concave, the term in between brackets is decreasing. Furthermore $M'(0)=\lambda-1>0$ and, since $x\alpha'(x)\le\alpha(x)$ for $x>0$, we get

$$M'(x) \le \frac{S'(x)}{(\lambda+1+x\alpha(x))^2}[\lambda^2-1-x^2\alpha(x)^2+2x\alpha(x)] \to_{x\to\infty} -\infty.$$

Thus M is first increasing and then decreasing. Furthermore, estimate (4.13) implies that

$$M(x) \sim_{\infty} \left[1-\frac{2x\alpha(x)}{\lambda+1+x\alpha(x)}\right]S(x) = \frac{\lambda+1-x\alpha(x)}{\lambda+1+x\alpha(x)}S(x) \sim_{\infty} -S(x).$$

Since $M(0) = 0$, we deduce that m is first increasing and then decreasing. Then we have as $x \to 0$, $m(x) \sim \frac{1-\lambda}{\lambda x} \to -\infty$, and, using (4.13),

$$m(x) \sim_{x\to\infty} \frac{x\alpha(x) - \lambda}{\lambda S(x)} > 0\,, \quad \text{for sufficiently large } x. \tag{4.15}$$

Since the function ℓ is negative, we deduce that, for any $z > 0$, there is a unique point in $(0, \infty)$, denoted $\varphi^\lambda(z)$, such that $m \circ \varphi^\lambda(z) = \ell(z)$, and is the unique minimum of $x \mapsto g(x, z)/S(x)$ on $[0, \infty)$. This point is also the unique solution of

$$g_x(x, z)S(x) - g(x, z)S'(x) = 0 \tag{4.16}$$

for any fixed z. The implicit functions theorem implies that φ^λ is C^1 on $(0, \infty)$. We prove hereafter that $\varphi^\lambda(z) < z$, for any $z > 0$, so that φ^λ corresponds to the definition given in the statement of the lemma.

2. $\varphi^\lambda(z) < z$ *, for any* $z > 0$

For any $z > 0$, since $x \mapsto m(x) - \ell(z)$ is first negative and then positive, on $(0, \infty)$, the property $\varphi^\lambda(z) < z$ will be a direct consequence of the estimate $m(z) - \ell(z) > 0$, that we prove now. First observe that the derivative of $h : z \mapsto [m(z) - \ell(z)]z^{-\lambda}$ is given by

$$h'(z) = \frac{1 + z\alpha(z)}{z^\lambda S(z)^3} S'(z)n(z), \quad z > 0, \quad \text{with}$$
$$n : z \mapsto S(z) - \frac{2z}{1 + z\alpha(z)} S'(z).$$

Hence h' has the same sign as n and, differentiating one more time, we compute

$$\begin{aligned} n'(z) &= S'(z)\left[1 - \frac{2 + 2z\alpha(z)}{1 + z\alpha(z)} + \frac{2z(\alpha(z) + z\alpha'(z))}{|1 + z\alpha(z)|^2}\right] \\ &= -\frac{1 + z^2\alpha(z)^2 - 2z^2\alpha'(z)}{|1 + z\alpha(z)|^2} S'(z), \end{aligned}$$

for any $z > 0$. Since α is concave and non-negative, we have $z\alpha'(z) \le \alpha(z)$ for $z > 0$, and, plugging this estimate in the previous expression, we obtain

$$n'(z) \le -\frac{|1 - z\alpha(z)|^2}{|1 + z\alpha(z)|^2} S'(z) \le 0, \quad z > 0.$$

Hence, n is nonincreasing starting from $n(0) = 0$, and therefore h is also nonincreasing on $(0, \infty)$. Furthermore, we know from (4.14) and (4.15) that $h(z) = [m(z) - \ell(z)]z^{-\lambda} > 0$ for sufficiently large z, so that we have $m(z) - \ell(z) > 0$, for any $z > 0$.

3. φ^λ *is increasing and valued in* $[0, y_\lambda]$

Recall that $m \circ \varphi^\lambda = \ell$ and ℓ is increasing and negative. Since m is also increasing when it is negative, we deduce that φ^λ is increasing. Since after crossing zero, the function m remains positive, $\varphi(z)$ must be smaller than the point where m crosses zero, for any $z > 0$. By definition of m, this point y^λ is implicitly defined by $y^\lambda S'(y^\lambda) = \lambda S(y^\lambda)$. Therefore $\varphi^\lambda(.) \le y^\lambda$ and, since $\ell(z) \to_{z\to\infty} 0$, we even have $\varphi^\lambda(z) \to_{z\to\infty} y^\lambda$. □

Proof of Lemma 4.5. We fix $\lambda > 1$ and recall from estimate (4.2) in the proof of Theorem 4.1 that $\mathcal{L}g^\lambda$ is given by

$$\begin{aligned}\mathcal{L}g^\lambda(x,z) &= x^{\lambda-2}[x\alpha(x)+1-\lambda] \\ &\quad\times\left(\frac{1}{z^\lambda}-\int_z^\infty \frac{S(x)}{S(u)}\frac{\lambda du}{u^{1+\lambda}}\right)+2x^{\lambda-1}S'(x)\int_z^\infty \frac{\lambda du}{S(u)u^{1+\lambda}},\end{aligned} \tag{4.17}$$

for any $0 < x \le z$. Since S is increasing, we first observe that $\mathcal{L}g^\lambda(x,.) \ge 0$ for any $x > 0$ such that $x\alpha(x)+1-\lambda \ge 0$. Denoting by x_λ the unique point of $\mathbb{R}^+$ defined implicitly by

$$x_\lambda\alpha(x_\lambda) = \lambda - 1,$$

we deduce that $\mathcal{L}g^\lambda(x,.) \ge 0$ for any $x \ge x_\lambda$.

It remains to treat the case where $x < x_\lambda$ and we compute

$$\frac{\partial}{\partial z}\mathcal{L}g^\lambda(x,z) = \lambda\frac{x^{\lambda-2}}{z^{1+\lambda}S(z)}\big\{[\lambda-1-x\alpha(x)](S(z)-S(x))-2xS'(x)\big\}, \quad 0 < x \le z.$$

For any fixed $x \in (0, x_\lambda)$, the previous expression in between brackets is increasing with respect to z, negative for $z = x$ and positive for z large enough. Hence, for any $x \in (0, x_\lambda)$, $\mathcal{L}g(x,.)$ is first decreasing, then increasing and $\mathcal{L}g(x,z)$ goes to 0 as z goes to infinity. Denoting by γ^λ the inverse of φ^λ, we deduce that

$$\mathcal{L}g(x,z) \ge 0, \quad \text{for any} \quad z \le \gamma^\lambda(x)\,, \quad \text{if and only if} \quad \mathcal{L}g(x,\gamma^\lambda(x)) \ge 0,$$

for any fixed $x \in (0, x^\lambda)$. Since φ^λ and hence γ^λ are increasing, it therefore only remains to verify that $\mathcal{L}g(.,\gamma(.)) \ge 0$ on $(0, x_\lambda)$.

We recall from the proof of Lemma 4.2 that γ^λ is defined implicitly by

$$\int_{\gamma^\lambda(x)}^\infty \frac{\lambda[\gamma^\lambda(x)]^\lambda du}{S(u)u^{1+\lambda}} = \frac{1}{S(x)} - \frac{xS'(x)}{\lambda S(x)^2}, \quad 0 < x < x_\lambda.$$

For a given $x \in (0, x_\lambda)$, plugging this estimate into (4.17), we deduce

$$\mathcal{L}g(x,\gamma^\lambda(x)) = (x\alpha(x)+1-\lambda)\frac{x^{\lambda-1}S'(x)}{\lambda[\gamma^\lambda(x)]^\lambda S(x)} + \left(\frac{1}{S(x)} - \frac{xS'(x)}{\lambda S^2(x)}\right)\frac{2x^{\lambda-1}S'(x)}{[\gamma^\lambda(x)]^\lambda},$$

which after simplifications leads to

$$\begin{aligned}\mathcal{L}g(x,\gamma^\lambda(x)) &= \frac{(x\alpha(x)+1+\lambda)x^{\lambda-1}S'(x)}{\lambda S(x)^2[\gamma^\lambda(x)]^\lambda}h(x), \quad \text{with} \\ h : x &\mapsto S(x) - \frac{2x}{x\alpha(x)+1+\lambda}S'(x)\,.\end{aligned}$$

In order to get the sign of $\mathcal{L}g(.,\gamma^\lambda(.))$, we look for the sign of h and compute

$$\begin{aligned}h'(x) &= S'(x)\left[1 - \frac{2+2x\alpha(x)}{1+\lambda+x\alpha(x)} + \frac{2x(\alpha(x)+x\alpha'(x))}{|1+\lambda+x\alpha(x)|^2}\right] \\ &= \frac{S'(x)}{|1+\lambda+x\alpha(x)|^2}[\lambda^2 - 1 - x^2\alpha(x)^2 + 2x^2\alpha'(x)] \\ &\ge \frac{S'(x)}{|1+\lambda+x\alpha(x)|^2}[\lambda^2 - (1-x\alpha(x))^2], \quad x < x_\lambda.\end{aligned}$$

Since $x\alpha'(x) \leq \alpha(x)$ for $x > 0$, due to the concavity of α. By definition of x_λ, we deduce that h is nondecreasing on $(0, x_\lambda)$. However, $h(0) = 0$ and therefore $\mathcal{L}g(., \gamma(.)) \geq 0$ on $(0, x_\lambda)$, which concludes the proof. □

5. MINIMIZATION OF THE RELATIVE QUADRATIC ERROR

Let us now consider the case where $f : x \mapsto \frac{1}{2}(1-x)^2$. Therefore, we are computing the following value function

$$V(x,z) := \frac{1}{2} \inf_{\theta \in \mathcal{T}} \mathbb{E}_{x,z} \left(1 - \frac{X_\theta}{Z_\tau}\right)^2, \quad (x,z) \in \mathbf{\Delta}. \tag{5.1}$$

With such a criterion, the investor tries to minimize the expected value of the squared relative error between the value of the stopped process and the maximal value of the process up to τ. In other words he wants to minimize the expectation of $[(Z_\tau - X_\theta)/Z_\tau]^2$, whereas in Section 4, $\lambda = 2$ would correspond to the minimization of $1 - (X_\theta/Z_\tau)^2$, which is not as natural. In contrast with the previous optimal stopping problem (4.1), we prove that stopping immediately even for $x = z$ is not optimal in general. We exhibit the optimal liquidation strategy, which is numerically not that different from immediate selling.

5.1. Construction of the Solution

From (2.11), we compute the corresponding reward function:

$$g(x,z) = \frac{1}{2}\left(1 - \frac{x}{z}\right)^2 + xS(x) \int_z^\infty \left(1 - \frac{x}{u}\right) \frac{du}{u^2 S(u)}, \quad (x,z) \in \mathbf{\Delta}\,.$$

In view of Proposition 3.5, we would require $\mathcal{L}g(x,z) \geq 0$ in order for some $(x,z) \in \mathbf{\Delta}$ to be in the stopping region. Let us first compute

$$g_x(x,z) = -\frac{1}{z}\left(1 - \frac{x}{z}\right) + [S(x) + xS'(x)] \int_z^\infty \frac{du}{u^2 S(u)} - [2xS(x) + x^2 S'(x)] \int_z^\infty \frac{du}{u^3 S(u)}\,,$$

$$\begin{aligned} g_{xx}(x,z) = {} & \frac{1}{z^2} + [2 + x\alpha(x)]S'(x) \int_z^\infty \frac{du}{u^2 S(u)} \\ & - [2S(x) + 4xS'(x) + x^2\alpha(x)S'(x)] \int_z^\infty \frac{du}{u^3 S(u)}, \end{aligned}$$

for any $(x,z) \in \Delta$. Combining these estimates, we deduce

$$\begin{aligned} \mathcal{L}g(x,z) = {} & \frac{1}{z^2}[1 + \alpha(x)(z-x)] + [2S'(x) - \alpha(x)S(x)] \int_z^\infty \frac{du}{u^2 S(u)} \\ & - [S(x) + 2xS'(x) - x\alpha(x)S(x)] \int_z^\infty \frac{2du}{u^3 S(u)}, \qquad (x,z) \in \mathbf{\Delta}. \end{aligned} \tag{5.2}$$

In view of Theorem 3.2 (i), if $\mathcal{L}g \geq 0$ on $\mathbf{\Delta}$, then immediate stopping is optimal, $v = g$ and the problem is trivial. However, the next result gives sufficient conditions such that it is not the case. Consider the following condition:

$$\alpha(0)^2 - 2\alpha'(0) < 0. \tag{5.3}$$

REMARK 5.1. Notice that (5.3) will be satisfied for an Ornstein-Uhlenbeck process as well as a CIR-Feller process with positive "mean," for which we respectively have $\alpha(x) = \alpha x$ and $\alpha(x) = \alpha\frac{x}{x+b}$, respectively, α and b being positive constants. More generally, as soon as $\alpha(0) = 0$, (5.3) is satisfied. However, for a drifted Brownian motion or a degenerated CIR-Feller process with "mean" equal to 0, (5.3) does not hold true.

PROPOSITION 5.2. *Assume that (5.3) is satisfied. Then, $\mathcal{L}g(z, z) < 0$ for z small enough so that immediate stopping is not optimal for the problem of interest $V(z, z)$.*

Proof. Using the asymptotic expansions from Proposition 5.9 in Section 5.4, we compute for z close to 0:

$$\begin{aligned}
\mathcal{L}g(z,z) &= [2S'(z) - \alpha(z)S(z)] \int_z^\infty \frac{du}{u^2 S(u)} \\
&\quad - [S(z) + 2zS'(z) - z\alpha(z)S(z)] \int_z^\infty \frac{2du}{u^3 S(u)} + \frac{1}{z^2} \\
&= \left(2 + \alpha(0)z + O(z^2)\right)\left(\frac{1}{2z^2} - \frac{\alpha(0)}{2z} - \frac{\alpha(0)^2 - 2\alpha'(0)}{12}\ln z + o(\ln z)\right) \\
&\quad - \left(3z + \frac{3}{2}\alpha(0)z^2 + O(z^3)\right)\left(\frac{2}{3z^3} - \frac{\alpha(0)}{2z^2} + O\left(\frac{1}{z}\right)\right) + \frac{1}{z^2} \\
&= -\frac{\alpha(0)^2 - 2\alpha'(0)}{6}\ln z + o(\ln z).
\end{aligned}$$

Since $\ln z \to -\infty$ when $z \to 0$, we see that if (5.3) holds, then $\mathcal{L}g(z, z) < 0$ for z in a neighborhood of 0, so that we have the result by continuity of $\mathcal{L}g$.

In particular, we deduce from Proposition 3.5 that immediate selling is not optimal for the problem $V(z, z)$ with z small enough. □

Hence stopping immediately is not optimal in general and the optimal strategy shall be very different from the one in the power utility case. Since we do not have $\mathcal{L}g \geq 0$ on the entire space $\mathbf{\Delta}$ but we can exercise only in that region, we first need to study the set

$$\Gamma^+ := \{(x, z) \in \mathbf{\Delta}, \mathcal{L}g(x, z) \geq 0\}\,, \tag{5.4}$$

and we define similarly:

$$\Gamma^- := \{(x, z) \in \mathbf{\Delta}, \mathcal{L}g(x, z) \leq 0\}\,. \tag{5.5}$$

In fact, observe that (5.2) rewrites as:

$$\begin{aligned}
\mathcal{L}g(x,z) &= \alpha(x)\frac{z-x}{z^2} + [2S'(x) - \alpha(x)S(x)] \int_z^\infty \left(1 - \frac{2x}{u}\right)\frac{du}{u^2 S(u)} \\
&\quad + \left(\frac{1}{z^2} - S(x)\int_z^\infty \frac{2du}{u^3 S(u)}\right), \qquad (x, z) \in \mathbf{\Delta}.
\end{aligned} \tag{5.6}$$

By Remark 2.3, we have $2S' - \alpha S - 2 \geq 0$ and therefore each of the three terms above are positive if $z \geq 2x$, and so

$$\mathcal{L}g(x, z) > 0 \quad \text{for} \quad z \geq 2x \quad \text{and} \quad (x, z) \in \mathbf{\Delta}, \tag{5.7}$$

which implies that $\{(x, z) \in \mathbf{\Delta}, z \geq 2x\} \subset \text{Int}(\Gamma^+)$.

Moreover we have the following result, which proof is given in Section 5.4 later.

LEMMA 5.3. For any $x > 0$, there exists $\delta_x \in (x, 2x)$ such that $\mathcal{L}g(x, .)$ is increasing on $[x, \delta_x)$ and decreasing on $(\delta_x, 2x]$.

In view of (5.7), we can define the following function on $\mathbb{R}_+ \setminus \{0\}$:

$$\Gamma(x) := \inf\{z \geq x, \mathcal{L}g(x, z) \geq 0\}. \tag{5.8}$$

Lemma 5.3 and (5.7) imply that, if $z > \Gamma(x)$, then $\mathcal{L}g(x, z) > 0$, while if $z \in (x, \Gamma(x))$, then $\mathcal{L}g(x, z) < 0$. We also deduce that $\Gamma(x) > x$ implies $\mathcal{L}g(x, \Gamma(x)) = 0$. Notice that Γ is continuous, and, from (5.7), we also know that $\Gamma(x) < 2x$.

The next result provides the main properties of Γ: it is increasing and equal to the Identity function for sufficiently large x. Again the proof is postponed to Section 5.4.

PROPOSITION 5.4. We have the two following properties:

(i) Γ is increasing on $(0, +\infty)$;
(ii) Denoting $\Gamma^\infty := \sup\{x \geq 0; \Gamma(x) > x\}$, we get $\Gamma^\infty < \infty$.

Notice that $\Gamma^+ \neq \mathbf{\Delta}$ implies directly $\Gamma^\infty > 0$.

Now that we have a better understanding of the set Γ^+, we expect to have a stopping region of the form $\{(x, z) \in \Delta\,;\ z \geq \gamma(x)\}$, and our objective is then to find functions v and γ, satisfying the following free-boundary problem:

$$\mathcal{L}v(x, z) = 0 \quad \text{for } x \leq z < \gamma(x) \quad \text{and} \quad (x, z) \in \mathbf{\Delta}, \tag{5.9}$$

$$v(x, z) = g(x, z) \quad \text{and} \quad \mathcal{L}g(x, z) \geq 0 \quad \text{for } z \geq \gamma(x) \quad \text{and} \quad (x, z) \in \mathbf{\Delta}, \tag{5.10}$$

$$v(0, z) = \frac{1}{2} \quad \text{for } z > 0, \tag{5.11}$$

$$v_z(z, z) = 0 \quad \text{for } z > 0. \tag{5.12}$$

In order to allow for the application of Itô's formula, the verification step requires a value function which is $C^{1,0}$ and piecewise $C^{2,1}$ with respect to (x, z). Therefore, as in Section 4, we complement the above system by the continuity and the smooth fit conditions

$$v(x, \gamma(x)) = g(x, \gamma(x)) \quad \text{and} \quad v_x(x, \gamma(x)) = g_x(x, \gamma(x)), \quad \text{for } x > 0. \tag{5.13}$$

The stopping region $\mathcal{S}$ will then be defined as:

$$\mathcal{S} := \{(x, z) \in \mathbf{\Delta}; z \geq \gamma(x)\} \cup \{(0, z); z > 0\}. \tag{5.14}$$

First by (5.9), on the continuation region, v is of the form:

$$v(x, z) = A(z) + B(z)S(x), \quad (x, z) \in \mathbf{\Delta} \setminus \mathcal{S}.$$

Then, on the interval where γ is one-to-one, the continuity and smooth fit conditions (5.13) imply that

$$v(x, z) = g(\gamma^{-1}(z), z) + \frac{g_x(\gamma^{-1}(z), z)}{S' \circ \gamma^{-1}(z)}[S(x) - S \circ \gamma^{-1}(z)], \quad (x, z) \in \mathbf{\Delta} \setminus \mathcal{S}.$$

Finally, the Neumann condition (5.12), implies that we expect the boundary γ to satisfy the following ODE:

$$\gamma'(x) = \frac{\gamma(x)^2 \mathcal{L}g(x, \gamma(x))}{\left(\frac{2x}{\gamma(x)} - 1\right)\left(1 - \frac{S(x)}{S \circ \gamma(x)}\right)}, \quad \text{for } x < \Gamma^\infty. \tag{5.15}$$

As in Espinosa and Touzi (2012), there is no a priori initial condition for this ODE. In the sequel, we take this ODE (with no initial condition) as a starting point to construct the boundary γ. Notice that this ODE has infinitely many solutions, as the Cauchy-Lipschitz condition is locally satisfied whenever (5.15) is complemented with the condition $\gamma(x_0) = z_0$ for any $0 < x_0 < z_0$ and $z_0 \neq 2x_0$. We will follow the ideas of Espinosa and Touzi (2012), however in our case, (5.15) is not well defined for $\gamma(x) = 2x$, so that our framework requires to be more cautious. Notice also that we encounter here a similar feature as in Peskir (1998). The following result selects an appropriate solution of (5.15), and its proof is given in Section 5.5.

PROPOSITION 5.5. *Let Int*(Γ^-) *be nonempty. Then, there exists an increasing continuous function* γ *defined on* $\mathbb{R}_+$ *with graph* $\{(x, \gamma(x)) : x > 0\} \subset \mathbf{\Delta}$*, such that:*

(i) *On the set* $\{x > 0 : \gamma(x) > x\}$, γ *is a* C^1 *solution of the ODE* (5.47),
(ii) $\{(x, \gamma(x)) : x > 0\} \subset \Gamma^+$, *and* $\{(x, \gamma(x)) : x > 0 \text{and} \gamma(x) > x\} \subset$ *Int*(Γ^+),
(iii) $\gamma(x) = x$ *for all* $x \geq \Gamma^\infty$.

Since γ is increasing, we can define:

$$\varphi := \gamma^{-1}. \tag{5.16}$$

Now that we have constructed the free-boundary φ, we are able to state the following result.

THEOREM 5.6. *Let Int*(Γ^-) *be nonempty,* γ *be given by Proposition 5.5 and* φ *be defined by (5.16). Then the value function V solution of problem (5.1) is given, for* $(x, z) \in \mathbf{\Delta}$*, by:*

$$V(x, z) := \begin{cases} g(x, z), & \text{if } x \leq \varphi(z) \\ g(\varphi(z), z) + g_x(\varphi(z), z)\dfrac{S(x) - S \circ \varphi(z)}{S' \circ \varphi(z)}, & \text{if } x > \varphi(z) \end{cases}. \tag{5.17}$$

Moreover, the smallest optimal stopping time associated to (5.1) is given by $\theta^* := \inf\{t \geq 0, \quad X_t \leq \varphi(Z_t)\}$.

Proof. Let v be defined by (5.17) and recall that $\mathcal{S}$ is defined by (5.14). The result follows from verifying that all the assumptions of Theorem 3.2 (ii) and (iii) are satisfied.

1. *Regularity of* v.

We know from Proposition 5.5 that γ and therefore φ are continuous and hence v is continuous on $\mathbf{\Delta}$ by construction. Furthermore, by Proposition 5.5 (i) and (ii) together with the dynamics of the ODE (5.15), γ is a C^1 function with positive derivative on the set $\{x > 0; \gamma(x) > x\}$. Therefore φ is C^1 as well on $\{z > 0; \varphi(z) < z\}$ so that it is immediate that v is C^0 and piecewise $C^{2,1}$ with respect to (x, z). Furthermore, since $\Gamma^\infty < \infty$ by

Proposition 5.4, $\mathbf{\Delta}\backslash\mathcal{S}$ is bounded. Since v is continuous and $g \geq 0$, v is bounded from below.

2. *Dynamics of* v.

By definition, we have $\mathcal{L}v = 0$ on $\mathbf{\Delta}\backslash\mathcal{S}$. By Proposition 5.5 (ii), $\mathcal{L}g(x, \gamma(x)) \geq 0$ for $x > 0$, and we deduce from Lemma 5.3 and (5.7) that $\mathcal{L}g(x, z) \geq 0$ for any $(x, z) \in \mathbf{\Delta}$ such that $z \geq \gamma(x)$. Hence, (5.7) ensures that $\mathcal{L}g \geq 0$ on $\mathcal{S}$.

It remains to prove that $v_z(z, z) = 0$ for $z > 0$. We fix $z > 0$. If $\varphi(z) \geq z$, since $g_z(z, z) = 0$, we have $v_z(z, z) = 0$ as well. Suppose now that $\varphi(z) < z$. Then, by Proposition 5.5 (i), γ satisfies (5.15) in a neighborhood of $\varphi(z)$, and by Proposition 5.5 (ii), $\mathcal{L}g(\varphi(z), z) > 0$, which implies $\gamma' \circ \varphi(z) > 0$, so that:

$$\varphi'(z)\mathcal{L}g(\varphi(z), z) = \frac{1}{z^2}\left(\frac{2\varphi(z)}{z} - 1\right)\left(1 - \frac{S \circ \varphi(z)}{S(z)}\right).$$

We then compute from the definitions of v and g that

$$\begin{aligned} v_z(z, z) &= g_z(\varphi(z), z) + g_{xz}\frac{S(z) - S \circ \varphi(z)}{S' \circ \varphi(z)} + \varphi'(z)\mathcal{L}g(\varphi(z), z)\frac{S(z) - S \circ \varphi(z)}{S' \circ \varphi(z)} \\ &= \left[\frac{1}{z^2}\left(1 - \frac{2\varphi(z)}{z}\right)\left(1 - \frac{S \circ \varphi(z)}{S(z)}\right) + \varphi'(z)\mathcal{L}g(\varphi(z), z)\right]\frac{S(z) - S \circ \varphi(z)}{S' \circ \varphi(z)} = 0. \end{aligned}$$

3. *Comparing* v *and* g.

Finally, the fact that $v \leq g$ on $\mathbf{\Delta}$ and $v < g$ on $\mathbf{\Delta}\backslash\mathcal{S}$ follows from similar arguments as in the proof of proposition 6.2 in Espinosa and Touzi (2012) but the demonstration is simpler in our context since $\Gamma^\infty < \infty$. For the sake of completeness, we detail this proof. For $(x, z) \in \mathbf{\Delta}$ such that $x > \varphi(z)$, we compute

$$v(x, z) - g(x, z) = g(\varphi(z), z) + g_x(\varphi(z), z)\frac{S(x) - S \circ \varphi(z)}{S' \circ \varphi(z)} - g(x, z),$$

and, differentiating twice with respect to x and using (5.13), we verify that

$$v_x(x, z) - g_x(x, z) = -S'(x)\int_{\varphi(z)}^{x} \frac{\mathcal{L}g(u, z)}{S'(u)}du. \tag{5.18}$$

Therefore, from Lemma 5.3 and Proposition 5.4 (i), for any fixed z, the function $x \mapsto (v - g)(x, z)$ is either decreasing on $[\varphi(z), z]$, or decreasing on $[\varphi(z), \delta)$ and then increasing on $(\delta, z]$ for a given $\delta \in (\varphi(z), z)$. For any $z > 0$, since $v(\varphi(z), z) = g(\varphi(z), z)$, we only need to prove that $n(z) := v(z, z) - g(z, z) < 0$ if $\varphi(z) < z$.

Since $v_z(z, z) = g_z(z, z) = 0$ for $z > 0$, we compute:

$$n'(z) = v_x(z, z) - g_x(z, z) = -S'(z)\int_{\varphi(z)}^{z} \frac{\mathcal{L}g(u, z)}{S'(u)}du, \quad z > 0.$$

We assume the existence of a fixed $z < \Gamma^\infty$ such that $n(z) \geq 0$ and $\varphi(z) < z$ and work toward a contradiction. We first observe that necessarily $n'(z) > 0$. If not, $\int_{\varphi(z)}^{z} \frac{\mathcal{L}g(u,z)}{S'(u)}du \geq 0$ implies that $\int_{\varphi(z)}^{x} \frac{\mathcal{L}g(u,z)}{S'(u)}du > 0$ for any $x \in (\varphi(z), z)$, and (5.18) combined with $v(\varphi(z), z) = g(\varphi(z), z)$ leads to $n(z) < 0$ which is impossible. Since n is continuous, this implies that n is increasing on any connected subset of $\{z' \geq z, \varphi(z') < z'\}$. Defining $a := \inf\{z' > z; \varphi(z') = z'\} \leq \Gamma^\infty < \infty$, we get $n(a) = v(a, a) - g(a, a) > 0$, which contradicts the definition of v.

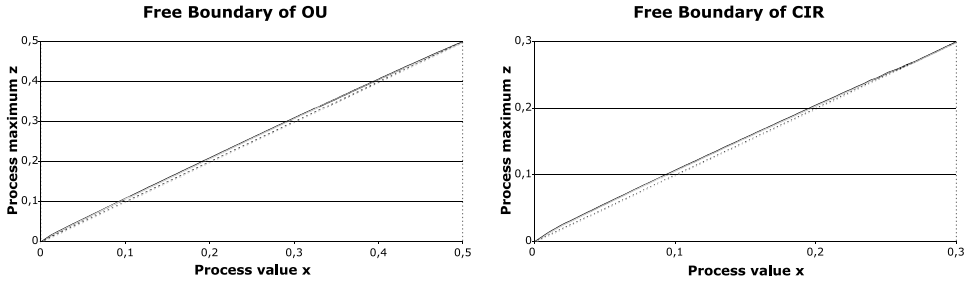

FIGURE 5.1. Optimal frontier for an OU ($\alpha = 1$) and a CIR ($\alpha = 0.1$, $b = 0.1$).

Therefore, $n(z) < 0$ for any $z > 0$ such that $\varphi(z) < z$ and we deduce that $v \leq g$ on $\mathbf{\Delta}$ and $v < g$ on $\mathbf{\Delta} \setminus \mathcal{S}$. □

5.2. Properties of the Optimal Liquidation Strategy

Theorem 5.6 and Proposition 5.2 indicate that, at least for processes satisfying (5.3), such as the Ornstein-Uhlenbeck process or the CIR-Feller process, the diagonal $\{(x, x); x > 0\}$ is not included in the stopping region $\mathcal{S}$. In other words, it is not always optimal to stop immediately, even when starting from points such that $x = z$. Therefore, the form of the solution and the nature of the optimal strategy to apply in order to be as close as possible to the maximum using this criterion is very different from the ones obtained in Section 4 or in Shiryaev et al. (2008).

The Ornstein-Uhlenbeck process as well as the CIR-Feller process are two examples for which the coefficient α satisfies Conditions (2.12) and $\text{Int}(\Gamma^-) \neq \emptyset$. Indeed we have $\alpha(x) = \alpha x$ and $\alpha(x) = \alpha \frac{x}{x+b}$, respectively, where α and b are two positive constants. Therefore, Condition (5.3) is satisfied, ensuring that $\text{Int}(\Gamma^-) \neq \emptyset$ by Proposition 5.2. Hence, Theorem 5.6 can be applied. Figure 5.1 represents the boundary φ for those two processes, with $\alpha = 1$ for the OU process and $(\alpha, b) = (0.1, 0.1)$ for the CIR process. We observe that the continuation region is in fact pretty small since the free boundary is very close to the diagonal axis. Therefore, even if immediate stopping is not optimal, an investor should not wait long until the process (X, X^*) enters the stopping region.

REMARK 5.7. Similarly to proposition 7.3 of Espinosa and Touzi (2012), an homogeneity result can be derived for the OU process, so that the free boundary for any $\alpha > 0$ can be deduced by a change of scale from the one for $\alpha = 1$.

The Brownian motion with negative drift is another example for which α satisfies Condition (2.12). However, since $\alpha(x) = \alpha > 0$ is constant, Condition (5.3) does not hold. Although we did not verify it, numerical computations suggest that $\mathcal{L}g \geq 0$ on $\mathbf{\Delta}$.

Finally, we can also consider the case of a Brownian motion. In this case, $\alpha(x) = 0$, so that α does not satisfy Condition (2.12). However, for any $(x, z) \in \mathbf{\Delta}$, we can compute from (5.2) that $\mathcal{L}g(x, z) = 2\frac{z-x}{z^3} \geq 0$ on $\mathbf{\Delta}$. Since the proofs of Theorem 3.2 and Remark 3.4 do not require Condition (2.12), we deduce that immediate stopping is always optimal.

REMARK 5.8. Let α be associated to an Ornstein-Uhlenbeck or a CIR process and hence be parameterized by a possibly bi-dimensional parameter set a. Since the parameter

set a may be badly estimated, let consider a sequence of parameter set (a_n) converging to a and denote by (α_n) the corresponding sequence of functions. Then, S_n, g_n, and all their derivatives converge, respectively, to S, g, and their derivatives in the sense of the uniform norm on the compact sets. Moreover, Γ_n converges to Γ in the same sense so that for n sufficiently large, $\mathrm{Int}(\Gamma_n^-) \neq \emptyset$. ODE (5.15) also depends continuously on a_n, so that $z_n^*(x_0)$ defined by (5.31) converges to $z^*(x_0)$, and γ_n given by Proposition 5.5 converges pointwise to γ. Since γ_n is increasing for any n and γ is continuous, Dini's theorem implies that the convergence is uniform on any compact set of $\mathbb{R}_+\backslash\{0\}$. Let us prove that φ_n converges to φ in the same sense. Let $y > 0$ be fixed, we define $x_n := \varphi_n(y)$ and $x := \varphi(y)$. We shall prove that $x_n \to x$. Indeed, since $\varphi_n(y) \in [\frac{y}{2}, y]$ for any n, $\{x_n; n \in \mathbb{N}\}$ is relatively compact in $\mathbb{R}_+\backslash\{0\}$. Now let x' be the limit of a subsequence of (x_n). For notational reasons, let us write $x_n \to x'$, forgetting that it is a subsequence. Since (γ_n) converges uniformly on compact sets of $\mathbb{R}_+\backslash\{0\}$, $\gamma_n(x_n) \to \gamma(x')$. Recalling that $\gamma_n(x_n) = y$ for any n, we get $\gamma(x') = y$ and therefore $x' = x$. In consequence, $x_n \to x$, or in other words, (φ_n) converges pointwise to φ on $\mathbb{R}_+\backslash\{0\}$. Noticing that $\varphi_n(0) = 0$ for any n and $\varphi(0) = 0$ and using again Dini's theorem, we see that (φ_n) converges to φ uniformly on the compact sets of $\mathbb{R}_+$. This finally implies that (V_n) converges pointwise to V. As a consequence, if one makes a small mistake estimating the parameters of the model, the induced mistake on the free boundary as well as the mistake on the value function will be small as well.

5.3. Generalization

As in Section 4, we may also consider, for any $\lambda > 0$, the following extension of the previous problem:

$$V_\lambda(x,z) := \frac{1}{\lambda} \inf_{\theta \in \mathcal{T}} \mathbb{E}_{x,z}\left(1 - \frac{X_\theta}{Z_\tau}\right)^\lambda, \quad (x,z) \in \mathbf{\Delta}. \tag{5.19}$$

In that case, (2.10) rewrites

$$g_\lambda(x,z) = \frac{1}{\lambda}\left(1 - \frac{x}{z}\right)^\lambda + xS(x)\int_z^\infty \left(1 - \frac{x}{u}\right)^{\lambda-1} \frac{du}{u^2 S(u)} du, \quad (x,z) \in \mathbf{\Delta}, \quad \lambda > 0. \tag{5.20}$$

If $\lambda = 2$, $\mathcal{L}g_2$ is given by (5.2). If $\lambda = 1$, the stopping problem has already been solved in Section 4 and $\mathcal{L}g_1$ is given by (4.3). For any $\lambda > 0$ such that $\lambda \notin \{1, 2\}$, we compute:

$$\begin{aligned}
\mathcal{L}g_\lambda(x,z) = (\lambda - 1)\Bigg\{ & \frac{1}{z^2}\left(1 - \frac{x}{z}\right)^{\lambda-2} \\
& - \int_z^\infty \frac{S(x)}{S(u)}\left[\frac{2}{u^3}\left(1 - \frac{x}{u}\right)^{\lambda-2} - \frac{(\lambda-2)x}{u^4}\left(1 - \frac{x}{u}\right)^{\lambda-3}\right] du \Bigg\} \\
& + [2S'(x) - \alpha(x)S(x)] \int_z^\infty \left(1 - \frac{x}{u}\right)^{\lambda-2} (u - \lambda x) \frac{du}{u^3 S(u)} + \alpha(x)\frac{(z-x)^{\lambda-1}}{z^\lambda},
\end{aligned} \tag{5.21}$$

for $0 \leq x < z$. In this case, the sign of $\mathcal{L}g_\lambda$ is hardly identifiable analytically, and we shall restrict our analysis to simple remarks and guesses on the solution of the problem (5.19).

Noticing that $\int_z^\infty (1-\frac{x}{u})^{\lambda-2}\frac{2}{u^3} - (\lambda-2)(1-\frac{x}{u})^{\lambda-3}\frac{x}{u^4}du = \frac{1}{z^2}(1-\frac{x}{z})^{\lambda-2}$ for $0 < x < z$, we deduce from (4.3), (5.7), and (5.21) that

$$\mathcal{L}g^\lambda(x,z) \geq 0, \quad \text{for } z \geq \lambda x > 0 \quad \text{and } 1 \leq \lambda \leq 2. \tag{5.22}$$

Therefore, for $1 \leq \lambda \leq 2$, we expect to obtain as for $\lambda = 2$ a free boundary γ_λ in between the axis $\{(x,x)\,;\,x>0\}$ and $\{(x,\lambda x)\,;\,x>0\}$. We verify easily as in Proposition (4.7) that $\lambda \mapsto V_\lambda$ is continuous and, as expected, we observe a disappearance of the free boundary γ_λ for $\lambda = 1$.

On the other hand, in the case where $\alpha(0) = 0$, for $\lambda < 1$, we observe that $\mathcal{L}g(x,z) < 0$ for x small enough and z large enough. Indeed, recalling from (4.13) that $S'(z) \sim \alpha(z)S(z)$ when $z \to \infty$, an integration by parts leads to $\int_z^\infty \frac{du}{u^2 S(u)} \sim_{z\to\infty} \frac{1}{z^2 S'(z)}$. Assuming moreover that $\alpha(0) = 0$ and plugging this estimate in (5.21), we get

$$\mathcal{L}g(0,z) \sim_{z\to\infty} \frac{\lambda-1}{z^2} \;<\; 0, \quad \text{for any } \lambda < 1.$$

In view of Proposition 3.5, this implies that the stopping region cannot have the same form as the one in the quadratic case $\lambda = 2$. It even suggests that the nature of the stopping region could be similar to the one of Section 4.2.

5.4. Proofs of Lemma 5.3 and Proposition 5.4

This section is dedicated to the proofs of Lemma 5.3 and Proposition 5.4, but we first state the asymptotic expansions used in Proposition 5.2.

PROPOSITION 5.9. *As $z \to 0$, we have the following expansions:*

$$S'(z) = 1 + \alpha(0)z + (\alpha'(0) + \alpha(0)^2)\frac{z^2}{2} + o(z^2);$$

$$S(z) = z + \alpha(0)\frac{z^2}{2} + (\alpha'(0) + \alpha(0)^2)\frac{z^3}{6} + o(z^3);$$

$$\alpha(z)S(z) = z\alpha(0) + \frac{z^2}{2}\big(\alpha(0)^2 + 2\alpha'(0)\big) + \frac{z^3}{6}(\alpha(0)^3 + 4\alpha(0)\alpha'(0) + 3\alpha''(0)) + o(z^3);$$

$$\int_z^\infty \frac{du}{u^2 S(u)} = \frac{1}{2z^2} - \frac{\alpha(0)}{2z} - \frac{\alpha(0)^2 - 2\alpha'(0)}{12}\ln(z) + o\big(\ln(z)\big);$$

$$\int_z^\infty \frac{2du}{u^3 S(u)} = \frac{2}{3z^3} - \frac{\alpha(0)}{2z^2} + \frac{\alpha(0)^2 - 2\alpha'(0)}{6z} + o\left(\frac{1}{z}\right).$$

Proof. As $z \to 0$, we directly compute the expansion:

$$S'(z) = S'(0) + zS''(0) + \frac{z^2}{2}S^{(3)}(0) + o(z^2) = 1 + \alpha(0)z + (\alpha'(0) + \alpha(0)^2)\frac{z^2}{2} + o(z^2).$$

The exact same reasoning also leads to

$$S(z) = z + \frac{\alpha(0)}{2}z^2 + \frac{\alpha'(0) + \alpha(0)^2}{6}z^3 + o(z^3);$$

$$\alpha(z)S(z) = z\alpha(0) + \frac{z^2}{2}(\alpha(0)^2 + 2\alpha'(0)) + \frac{z^3}{6}(\alpha(0)^3 + 4\alpha(0)\alpha'(0) + 3\alpha''(0)) + o(z^3).$$

Using one of the previous estimates, we get

$$\begin{aligned}\int_z^\infty \frac{du}{u^2 S(u)} &= \int_z^\infty \frac{du}{u^3\left(1+\frac{\alpha(0)}{2}u+\frac{\alpha'(0)+\alpha(0)^2}{6}u^2+o(u^2)\right)}\\ &= \int_z^\infty \left(1-\frac{\alpha(0)}{2}u-\frac{\alpha'(0)+\alpha(0)^2}{6}u^2+\left(\frac{\alpha(0)u}{2}\right)^2+o(u^2)\right)\frac{du}{u^3}\\ &= \int_z^\infty \left(\frac{1}{u^3}-\frac{\alpha(0)}{2u^2}+\frac{\alpha(0)^2-2\alpha'(0)}{12u}+o\left(\frac{1}{u}\right)\right)du\\ &= \frac{1}{2z^2}-\frac{\alpha(0)}{2z}-\frac{\alpha(0)^2-2\alpha'(0)}{12}\ln(z)+o\big(\ln(z)\big),\end{aligned}$$

which is justified since all the nonzero terms go to infinity when $z \to 0$. Similarly, we compute

$$\begin{aligned}\int_z^\infty \frac{2du}{u^3 S(u)} &= \int_z^\infty \left(\frac{2}{u^4}-\frac{\alpha(0)}{u^3}+\frac{\alpha(0)^2-2\alpha'(0)}{6u^2}+o\left(\frac{1}{u^2}\right)\right)du\\ &= \frac{2}{3z^3}-\frac{\alpha(0)}{2z^2}+\frac{\alpha(0)^2-2\alpha'(0)}{6z}+o\left(\frac{1}{z}\right).\end{aligned}$$

□

Proof of Lemma 5.3. Differentiating (5.2) with respect to z, we compute

$$\begin{aligned}\frac{\partial}{\partial z}\mathcal{L}g(x,z) &= -\frac{2S'(x)-\alpha(x)S(x)}{z^2S(z)}+\frac{[2-2x\alpha(x)]S(x)+4xS'(x)}{z^3S(z)}-\frac{2-2x\alpha(x)}{z^3}-\frac{\alpha(x)}{z^2}\\ &= [(2x-z)\alpha(x)-2]\frac{S(z)-S(x)}{z^3S(z)}+(2x-z)\frac{2S'(x)}{z^3S(z)},\quad (x,z)\in\mathbf{\Delta}.\end{aligned}$$

Let us introduce x_α as the unique solution of:

$$x_\alpha\alpha(x_\alpha)=2. \tag{5.23}$$

If $x \le x_\alpha$, then $z \mapsto (2x-z)\alpha(x)-2$ is negative on $[x, 2x)$, whereas if $x > x_\alpha$, then there exists $z_x \in (x, 2x)$ such that $z \mapsto (2x-z)\alpha(x)-2$ will be positive on (x, z_x), zero at z_x and negative on $(z_x, 2x)$.

Let x be fixed and let us introduce

$$F : z \mapsto S(z)-S(x)+\frac{2S'(x)(2x-z)}{(2x-z)\alpha(x)-2},$$

which is well defined and continuous on $[x, 2x]$ if $x < x_\alpha$, on $[x, 2x)$ if $x = x_\alpha$ and on $[x, 2x]\backslash\{z_x\}$ if $x > x_\alpha$. Furthermore, F is increasing, since we compute on the domain of definition of F:

$$F'(z)=S'(z)+\frac{4S'(x)}{((2x-z)\alpha(x)-2)^2}>0.$$

We consider first the case where $x \le x_\alpha$. Then F and $\frac{\partial}{\partial z}\mathcal{L}g(x,.)$ have opposite signs on $[x, 2x)$. Since F is increasing, $F(x) < 0$ while $F(2x) = S(2x)-S(x) > 0$, $\mathcal{L}g(x,.)$ is increasing on $[x, \delta_x)$ and decreasing on $(\delta_x, 2x]$, for a certain $\delta_x \in (x, 2x)$.

We now turn to the case where $x > x_\alpha$. Then F and $\frac{\partial}{\partial z}\mathcal{L}g(x, .)$ have the same sign on $[x, z_x)$ and opposite signs on $(z_x, 2x]$. Since F is increasing, $F(x) > 0$, $F(z_x^+) = -\infty$ and $F(2x) > 0$, we see that again $\mathcal{L}g(x, .)$ is increasing on $[x, \delta_x)$ and decreasing on $(\delta_x, 2x]$, for a certain $\delta_x \in (z_x, 2x) \subset (x, 2x)$. □

Proof of Proposition 5.4. We prove the two assertions separately.

(i) Γ *is increasing on* $(0, +\infty)$.

We fix $x > 0$ such that $\Gamma(x) > x$. Then $\mathcal{L}g(., \Gamma(.)) = 0$ in a neighborhood of x, and using the implicit functions theorem, Γ is C^1 in a neighborhood of x and we have:

$$\Gamma'(x)\frac{\partial}{\partial z}\mathcal{L}g(x, \Gamma(x)) + \frac{\partial}{\partial x}\mathcal{L}g(x, \Gamma(x)) = 0. \tag{5.24}$$

We will prove that $\Gamma'(x) > 0$. Denoting $m := 2S' - \alpha S$ which is increasing and positive, we get combining $\mathcal{L}g(x, \Gamma(x)) = 0$ and (5.6):

$$m(x)\int_{\Gamma(x)}^{\infty} \frac{u - 2x}{u^3 S(u)}du = S(x)\int_{\Gamma(x)}^{\infty} \frac{2du}{u^3 S(u)} - \frac{1 + \alpha(x)(\Gamma(x) - x)}{\Gamma(x)^2} \leq -\frac{\alpha(x)(\Gamma(x) - x)}{\Gamma(x)^2},$$

since S is increasing. Differentiating (5.6) with respect to x, we also compute

$$\begin{aligned}\frac{\partial}{\partial x}\mathcal{L}g(x, z) = {} & \frac{\alpha'(x)(z - x) - \alpha(x)}{z^2} + m'(x)\int_z^{\infty} \frac{u - 2x}{u^3 S(u)}du \\ & - [m(x) + S'(x)]\int_z^{\infty} \frac{2du}{u^3 S(u)},\end{aligned}$$

for $z \geq x$. Denoting $A := (\alpha m' - \alpha' m)(\Gamma - Id) + \alpha m$, the two previous estimates lead to

$$\frac{\partial}{\partial x}\mathcal{L}g(x, \Gamma(x)) \leq -\frac{A(x)}{m(x)\Gamma(x)^2} - [m(x) + S'(x)]\int_{\Gamma(x)}^{\infty} \frac{2du}{u^3 S(u)}. \tag{5.25}$$

Introducing $B := \alpha m' - \alpha' m$ and observing that $x \leq \Gamma(x) \leq 2x$, we obtain

$$A(x) \geq \alpha(x)m(x)\mathbf{1}_{\{B(x)\geq 0\}} + (xB(x) + \alpha(x)m(x))\mathbf{1}_{\{B(x)<0\}}. \tag{5.26}$$

Introducing finally $C : x \mapsto xB(x) + \alpha(x)m(x)$, we compute $C(0) = 2\alpha(0) \geq 0$ and

$$C'(x) = 2\alpha(x)m'(x) + x(\alpha(x)m''(x) - \alpha''(x)m(x)) \geq 0,$$

because $m'' \geq 0$ and $\alpha'' \leq 0$. Therefore C is non-negative, and, according to (5.26), A is also non-negative. As a consequence, combining $m > 0$ and (5.25), we deduce that $\frac{\partial}{\partial x}\mathcal{L}g(x, \Gamma(x)) < 0$. Using Lemma 5.3, we have $\frac{\partial}{\partial z}\mathcal{L}g(x, \Gamma(x)) > 0$, and (5.24) implies that $\Gamma'(x) > 0$.

Therefore, Γ is increasing on the set $\{x > 0, \Gamma(x) > x\}$. However, it is also increasing on the interior of the set $\{x > 0, \Gamma(x) = x\}$. Since Γ is continuous, it is increasing on $(0, +\infty)$.

(ii) We have $\Gamma^\infty := \sup\{x \geq 0; \Gamma(x) > x\} < \infty$.

The arguments used here are very close to the ones in the proof of proposition 4.3 in Espinosa and Touzi (2012). However, our conclusions cannot be deduced from theirs since the involved computations are different and we need to detail this proof.

From the definition of the scale function (2.9), we compute:

$$S(x) = S(1) + \frac{S'(x)}{\alpha(x)} - \frac{S'(1)}{\alpha(1)} - \int_1^x \left(\frac{1}{\alpha}\right)'(u) S'(u) du, \quad x > 0.$$

We then distinguish two cases depending on the explosion of the last term in the previous expression.

Case 1: $\int_1^\infty (1/\alpha)'(u) S'(u) du > -\infty$.

Then $S(x) = \frac{S'(x)}{\alpha(x)} + O(1)$ for x large enough. Recalling that $\mathcal{L}S = 0$, we compute

$$\begin{aligned}\int_x^\infty \frac{du}{u^2 S(u)} &= \int_x^\infty \frac{du}{u^2\left(\frac{S'(u)}{\alpha(u)} + O(1)\right)} = \int_x^\infty \frac{\alpha(u)}{u^2 S'(u)} \frac{du}{1 + O\left(\frac{\alpha(u)}{S'(u)}\right)} \\ &= \int_x^\infty \frac{\alpha(u) du}{u^2 S'(u)} + O\left(\int_x^\infty \frac{\alpha^2(u)}{u^2 [S'(u)]^2}\right),\end{aligned}$$

for x large enough. Integrating by parts, we observe that

$$\int_x^\infty \frac{\alpha(u) du}{u^2 S'(u)} = \frac{1}{x^2 S'(x)} - 2\int_x^\infty \frac{du}{u^3 S'(u)}, \quad x > 1.$$

We now prove that $\frac{x\alpha^2(x)}{S'(x)} \to 0$ when $x \to \infty$.

Indeed, since $\alpha(1) > 0$ by (2.12), and since α is nondecreasing, we get $S'(x) \geq e^{(x-1)\alpha(1)}$, for any $x \geq 1$. On the other hand, since α is concave, we also have $0 \leq \alpha(x) \leq x\alpha'(0)$, so that:

$$0 \leq \frac{x\alpha^2(x)}{S'(x)} \leq \frac{x^3[\alpha'(0)]^2}{e^{(x-1)\alpha(1)}} \to 0 \quad \text{when} \quad x \to \infty.$$

As a consequence, we get

$$\int_x^\infty \frac{du}{u^2 S(u)} = \frac{1}{x^2 S'(x)} - 2\int_x^\infty \frac{du}{u^3 S'(u)} + \circ\left(\int_x^\infty \frac{du}{u^3 S'(u)}\right).$$

Integrating by parts again, we finally compute

$$\int_x^\infty \frac{du}{u^2 S(u)} = \frac{1}{x^2 S'(x)} - \frac{2}{\alpha(x) x^3 S'(x)} + \circ\left(\frac{1}{\alpha(x) x^3 S'(x)}\right),$$

and similarly we get

$$\int_x^\infty \frac{du}{u^3 S(u)} = \frac{1}{x^3 S'(x)} - \frac{3}{\alpha(x) x^4 S'(x)} + \circ\left(\frac{1}{\alpha(x) x^4 S'(x)}\right).$$

Plugging these estimates in the expression of $\mathcal{L}g$ given by (5.2) leads to:

$$\begin{aligned}\mathcal{L}g(x,x) = \frac{1}{x^2} &+ [2S'(x) - \alpha(x)S(x)] \int_x^\infty \frac{du}{u^2 S(u)} \\ &- [S(x) + 2xS'(x) - x\alpha(x)S(x)] \int_x^\infty \frac{2du}{u^3 S(u)}\end{aligned}$$

$$= \frac{1}{x^2} + \big(S'(x) + O(\alpha(x))\big)\left(\frac{1}{x^2 S'(x)} - \frac{2}{\alpha(x)x^3 S'(x)} + \circ\left(\frac{1}{\alpha(x)x^3 S'(x)}\right)\right)$$
$$-2\left(xS'(x) + \frac{S'(x)}{\alpha(x)} + O(x\alpha(x))\right)$$
$$\times\left(\frac{1}{x^3 S'(x)} - \frac{3}{\alpha(x)x^4 S'(x)} + \circ\left(\frac{1}{\alpha(x)x^4 S'(x)}\right)\right).$$

Using the fact that $\frac{x\alpha^2(x)}{S'(x)} \to 0$ as $x \to \infty$, we get:

$$\mathcal{L}g(x, x) = \frac{2}{\alpha(x)x^3} + \circ\left(\frac{1}{\alpha(x)x^3}\right).$$

Hence $\mathcal{L}g(x, x) > 0$ and therefore $\Gamma(x) = x$ for x large enough, so that $\Gamma_\infty < \infty$.

Case 2: $\int_1^\infty (1/\alpha)'(u)S'(u)du = -\infty$.

For x large enough, we have

$$S(x) = \frac{S'(x)}{\alpha(x)}\left[1 - \left(\frac{1}{\alpha}\right)'(x) + \circ\left(\left(\frac{1}{\alpha}\right)'(x)\right)\right],$$

so that

$$\int_x^\infty \frac{du}{u^2 S(u)} = \int_x^\infty \frac{\alpha(u)}{u^2 S'(u)}\left[1 + \left(\frac{1}{\alpha}\right)'(u) + \circ\left(\left(\frac{1}{\alpha}\right)'(u)\right)\right]du$$
$$= \frac{1}{x^2 S'(x)} - \int_x^\infty \frac{2du}{u^3 S'(u)} - \int_x^\infty \frac{\alpha'(u)du}{u^2\alpha(u)S'(u)} + \circ\left(\int_x^\infty \frac{\alpha(u) + u\alpha'(u)}{u^3\alpha(u)S'(u)}\,du\right).$$

Noticing that $0 \le x\alpha'(x) \le \alpha(x)$ for $x > 0$, since α is concave, we have:

$$\circ\left(\int_x^\infty \frac{\alpha(u) + u\alpha'(u)}{u^3\alpha(u)S'(u)}\,du\right) = \circ\left(\int_x^\infty \frac{du}{u^3 S'(u)}\right).$$

Integrating by parts, we finally get

$$\int_x^\infty \frac{du}{u^2 S(u)} = \frac{1}{x^2 S'(x)} - \frac{2}{x^3\alpha(x)S'(x)} - \frac{\alpha'(x)}{x^2\alpha^2(x)S'(x)} + \circ\left(\frac{1}{x^3\alpha(x)S'(x)}\right),$$

where the third term in the previous expansion might be negligible or not (depending on α). Similarly, we compute:

$$\int_x^\infty \frac{du}{u^3 S(u)} = \frac{1}{x^3 S'(x)} - \frac{3}{x^4\alpha(x)S'(x)} - \frac{\alpha'(x)}{x^3\alpha^2(x)S'(x)} + \circ\left(\frac{1}{x^4\alpha(x)S'(x)}\right),$$

so that:

$$\mathcal{L}g(x, x) = \frac{1}{x^2} + [2S'(x) - \alpha(x)S(x)]\int_x^\infty \frac{du}{u^2 S(u)}$$
$$-[S(x) + 2xS'(x) - x\alpha(x)S(x)]\int_x^\infty \frac{2du}{u^3 S(u)}$$
$$= \frac{1}{x^2} - 2\left(xS'(x) - x\frac{\alpha'(x)}{\alpha^2(x)}S'(x) + \frac{S'(x)}{\alpha(x)}\right)$$

$$\times\left(\frac{1}{x^3S'(x)}-\frac{3}{x^4\alpha(x)S'(x)}-\frac{\alpha'(x)}{x^3\alpha^2(x)S'(x)}\right)+\left(S'(x)-\frac{\alpha'(x)}{\alpha^2(x)}S'(x)\right)$$
$$\times\left(\frac{1}{x^2S'(x)}-\frac{2}{x^3\alpha(x)S'(x)}-\frac{\alpha'(x)}{x^2\alpha^2(x)S'(x)}\right)+\circ\left(\frac{1}{x^3\alpha(x)}\right)$$
$$=\frac{2}{x^3\alpha(x)}+\frac{2\alpha'(x)}{x^2\alpha^2(x)}+\circ\left(\frac{1}{x^3\alpha(x)}\right),$$

where the second term might be or not negligible. In any case, we see that for sufficiently large x, $\mathcal{L}g(x,x)>0$, so that $\Gamma(x)=x$. Therefore, $\Gamma_\infty<\infty$ also holds in this case. □

5.5. Proof of Proposition 5.5

This section is dedicated to the proof of Proposition 5.5. As already explained, this proof uses the same ideas as the one developed in Espinosa and Touzi (2012). However, because of the specificity of (5.15), the properties of the flow are different and the analysis needs to be adapted to our framework. We will try to follow their notations and point out in the proofs the parts that are identical to their paper, but we choose to rewrite them for the sake of completeness.

First, for the convenience of the reader, we recall ODE (5.15) that γ needs to satisfy:

$$\gamma'(x)=\frac{\gamma(x)^2\mathcal{L}g(x,\gamma(x))}{\left(\frac{2x}{\gamma(x)}-1\right)\left(1-\frac{S(x)}{S\circ\gamma(x)}\right)},\quad x>0\,. \tag{5.27}$$

Let us first define

$$\mathbf{D}^-:=\{x>0:\mathcal{L}g(x,x)<0\}, \tag{5.28}$$

and, for all $x_0\in\mathbf{D}^-$, we introduce

$$d(x_0):=\sup\{x\le x_0:\mathcal{L}g(x,x)\ge 0\}\ \text{ and }\ u(x_0):=\inf\{x\ge x_0;\mathcal{L}g(x,x)\ge 0\}, \tag{5.29}$$

with the convention that $d(x_0)=0$ if $\{x\le x_0:\mathcal{L}g(x,x)\ge 0\}=\emptyset$. Observe that Proposition 5.4 ensures that $u(x_0)\le\Gamma^\infty<\infty$. Since $\mathcal{L}g$ is continuous and $x_0\in\mathbf{D}^-$ we must have $d(x_0)<x_0<u(x_0)<\infty$.

For any $x_0\in\mathbf{D}^-$ and $z_0>x_0$, we denote by $\gamma^{z_0}_{x_0}$ the maximal solution of the Cauchy problem (5.27) complemented by the additional condition $\gamma(x_0)=z_0$, and we denote by $I^{z_0}_{x_0}:=(\ell^{z_0}_{x_0},r^{z_0}_{x_0})$ the corresponding (open) interval of definition of $\gamma^{z_0}_{x_0}$. Since the right-hand side of ODE (5.27) is locally Lipschitz on either one of the sets $\{(x,\gamma),0<2x<\gamma\}$ or $\{(x,\gamma),x<\gamma<2x\}$ but is not defined on the set $\{(x,\gamma),2x=\gamma\}$, the maximal solution will be defined as long as $(x,\gamma(x))$ remains in one of those two sets. Since $\Gamma(x_0)<2x_0$, we restrict our attention to conditions $\gamma(x_0)=z_0$ satisfying $x_0<z_0<2x_0$.

The next lemma provides useful additional properties of the maximal solutions described above and their respective domains of definitions.

LEMMA 5.10. *Assume that α satisfies Conditions (2.12) and let $x_0\in\mathbf{D}^-$ be fixed.*

(i) For all $z_0\in(x_0,2x_0)$, $\ell^{z_0}_{x_0}\le d(x_0)$, we have $\lim_{x\to\ell^{z_0}_{x_0}}\gamma^{z_0}_{x_0}(x)=\ell^{z_0}_{x_0}$ and, if $\ell^{z_0}_{x_0}>0$, we get $\mathcal{L}g(\ell^{z_0}_{x_0},\ell^{z_0}_{x_0})\ge 0$;

(ii) for all $z_0\in(x_0,\Gamma(x_0))$, $\mathcal{L}g(x,\gamma^{z_0}_{x_0}(x))<0$ for any $x\in(x_0,r^{z}_{x_0})$;

(iii) *there exists* $a_0 \in (x_0, 2x_0)$ *such that for any* $z_0 \in [a_0, 2x_0)$, $\mathcal{L}g(x, \gamma^{z_0}_{x_0}(x)) > 0$ *for any* $x \in (x_0, r^{z_0}_{x_0})$.

Proof. We fix $x_0 \in \mathbf{D}^-$ and prove each property separately.

(i) Let us fix $z_0 \in (x_0, 2x_0)$. The right-hand side of (5.27) is locally Lipschitz as long as $0 < x < \gamma^{z_0}_{x_0}(x) < 2x$, so that this last estimate holds for any $x \in I^{z_0}_{x_0}$. We intend to prove that $\gamma^{z}_{x_0}$ hits the diagonal $\{(x, z); x = z\}$ at the left-hand side $\ell^{z_0}_{x_0}$ of $I^{z_0}_{x_0}$.

For this purpose, let us first prove that, for any $\zeta \in (0, x_0)$, the graph of $\gamma^{z_0}_{x_0}$ restricted to $[\zeta, x_0]$ cannot come too close to $\{(x, z); 2x = z\}$. Since $\Gamma(x) < 2x$ for $x > 0$ and Γ and $\mathcal{L}g$ are continuous, there exist $\varepsilon > 0$ and $\delta > 0 \in (0, \zeta)$ such that $\mathcal{L}g \geq \varepsilon$ on the compact set $\{(x, z); x \in [\zeta, x_0] \text{ and } z \in [2x - \delta, 2x]\}$. Observe that, for $x \in [\zeta, x_0]$ such that $\gamma^{z_0}_{x_0}(x) \in [\max(2x - \delta, \frac{4x}{2+\zeta^2\varepsilon}), 2x)$, we get from (5.27) that

$$\big(\gamma^{z_0}_{x_0}\big)'(x) \geq \frac{\gamma^{z_0}_{x_0}(x)^2 \mathcal{L}g\big(x, \gamma^{z_0}_{x_0}(x)\big)}{\dfrac{2x}{\gamma^{z_0}_{x_0}(x)} - 1} \;\geq\; \frac{2\gamma^{z_0}_{x_0}(x)^2\varepsilon}{\zeta^2\varepsilon} \;\geq\; 2,$$

where, for the last inequality, we used $\gamma(x) \geq x \geq \zeta$. Hence, the function $x \mapsto 2x - \gamma^{z_0}_{x_0}(x)$ is nonincreasing on the set $\{x \in [\zeta, x_0] \cap I^{z_0}_{x_0}; \gamma^{z_0}_{x_0}(x) \in [\max(2x - \delta, \frac{4x}{2+\zeta^2\varepsilon}), 2x)\}$. Therefore, by arbitrariness of $\zeta > 0$, the graph of $\gamma^{z_0}_{x_0}$ restricted to $(\ell^{z_0}_{x_0}, x_0]$ stays away from $\{(x, z); 2x = z\}$ and $\gamma^{z_0}_{x_0}$ necessarily hits the diagonal $\{(x, z); x = z\}$ at the left-hand side $\ell^{z_0}_{x_0}$ of the maximal interval $I^{z_0}_{x_0}$.

On the other hand, we observe from (5.27) that $\gamma^{z_0}_{x_0}$ is nonincreasing at the points x satisfying $(x, \gamma^{z_0}_{x_0}(x)) \in \Gamma^-$ and therefore $\ell^{z_0}_{x_0} \notin \mathbf{D}^-$ by minimality of $I^{z_0}_{x_0}$. Since $\big(d(x_0), u(x_0)\big) \subset \mathbf{D}^-$, we get $\ell^{z_0}_{x_0} \leq d(x_0)$ and $\mathcal{L}g(\ell^{z_0}_{x_0}, \ell^{z_0}_{x_0}) \geq 0$, or equivalently $\Gamma(\ell^{z_0}_{x_0}) = \ell^{z_0}_{x_0}$.

It still remains to prove properly that $\lim_{x \to \ell^{z_0}_{x_0}} \gamma^{z_0}_{x_0}(x) = \ell^{z_0}_{x_0}$. Assume first that $\ell^{z_0}_{x_0} > 0$. Notice from (5.27) that $\gamma^{z_0}_{x_0}$ is nondecreasing if $(x, \gamma^{z_0}_{x_0}(x)) \in \Gamma^+$, and $x \leq \gamma^{z_0}_{x_0}(x) \leq \Gamma(x)$ otherwise. Since Γ is also nondecreasing, $\tilde{\gamma}^{z_0}_{x_0} := \max(\gamma^{z_0}_{x_0}, \Gamma)$ is a nondecreasing function defined on $(\ell^{z_0}_{x_0}, x_0]$, which therefore admits a limit at $\ell^{z_0}_{x_0}$. Since $\ell^{z_0}_{x_0} > 0$, we have $\lim_{x \to \ell^{z_0}_{x_0}} \tilde{\gamma}^{z_0}_{x_0}(x) < 2\ell^{z_0}_{x_0}$ as observed earlier, so that, combining the maximality of $I^{z_0}_{x_0}$ with $\Gamma(\ell^{z_0}_{x_0}) = \ell^{z_0}_{x_0}$, we obtain $\lim_{x \to \ell^{z_0}_{x_0}} \tilde{\gamma}^{z_0}_{x_0}(x) = \ell^{z_0}_{x_0}$. Since $x \leq \gamma^{z_0}_{x_0}(x) \leq \tilde{\gamma}^{z_0}_{x_0}(x)$ for $x \in (\ell^{z_0}_{x_0}, x_0]$, we have $\lim_{x \to \ell^{z_0}_{x_0}} \gamma^{z_0}_{x_0}(x) = \ell^{z_0}_{x_0}$. Finally, if $\ell^{z_0}_{x_0} = 0$, since $x < \gamma^{z_0}_{x_0}(x) < 2x$, we also have the result.

(ii) Let us fix $z_0 \in \big(x_0, \Gamma(x_0)\big)$. As already observed, the dynamics of (5.27) imply that $\gamma^{z_0}_{x_0}$ is nonincreasing in the neighborhood of any point x such that $(x, \gamma^{z_0}_{x_0}(x)) \in \mathrm{Int}(\Gamma^-)$. On the other hand, Proposition 5.4 tells us that the function Γ is increasing on $[x_0, +\infty)$. Hence $x \mapsto (x, \gamma^{z_0}_{x_0}(x))$ remains in $\mathrm{Int}(\Gamma^-)$ on $\big[x_0, r^{z_0}_{x_0}\big)$.

We consider now the case where $z_0 = \Gamma(x_0)$. Since $\Gamma(x_0) > x_0$, the proof of Proposition 5.4 (i) tells us that Γ' is positive on a neighborhood of x_0. Since $z_0 = \Gamma(x_0)$, we deduce from (5.27) that $(\gamma^{z_0}_{x_0})'(x_0) = 0$, and the exact same reasoning as above implies that $x \mapsto \big(x, \gamma^{z_0}_{x_0}(x)\big) \in \mathrm{Int}(\Gamma^-)$ on $\big(x_0, r^{z_0}_{x_0}\big)$.

(iii) Recall that $\Gamma^\infty < \infty$. Therefore, as in (i), there exist $\varepsilon > 0$ and $\delta \in (0, 1)$ such that $\mathcal{L}g \geq \varepsilon$ on $\{(x, z); x \in [x_0, \Gamma^\infty] \text{ and } z \in [(2 - \delta)x, 2x]\}$. Let $b := \min(x_0^2\varepsilon, \delta)$. From (5.27), we see that if $x \in [x_0, \Gamma^\infty]$ and $\gamma^{z}_{x_0}(x) \in [(2 - b)x, 2x)$, then $(\gamma^{z}_{x_0})'(x) \geq \frac{2-b}{b} x_0^2\varepsilon \geq 2 - b$. We denote $a_0 := (2 - b)x_0$ and fix $z \in [a_0, 2x_0)$. We deduce from the previous reasoning that we must have

$$\gamma^{z}_{x_0}(x) \geq z + \int_{x_0}^{x} \big(\gamma^{z}_{x_0}\big)'(u)du \;\geq\; z + (2 - b)(x - x_0) \;>\; (2 - b)x \;>\; \Gamma(x),$$

for $x \in \big[x_0, \min\big(r^z_{x_0}, \Gamma^\infty\big)\big)$. If ever $r^z_{x_0} \le \Gamma^\infty$, we just obtained the announced result and, if ever $r^z_{x_0} > \Gamma^\infty$, we complete the proof noticing that the maximality of $I^z_{x_0}$ implies that $\gamma^z_{x_0}(x) > x = \Gamma(x)$ for $x \ge \Gamma^\infty$. □

We now construct the stopping boundary γ by selecting one of the previous maximal solutions. For a given $x_0 \in \mathbf{D}^-$, let

$$\mathbf{Z}(x_0) := \big\{z \in (x_0, 2x_0);\, \mathcal{L}g\big(x, \gamma^z_{x_0}(x)\big) < 0 \ \text{ for some } \ x \in [x_0, r^z_{x_0})\big\}, \tag{5.30}$$

$$z^*(x_0) := \sup \mathbf{Z}(x_0). \tag{5.31}$$

Moreover, whenever $z^*(x_0) < 2x_0$, we denote

$$\gamma^*_{x_0} := \gamma^{z^*(x_0)}_{x_0}, \quad \ell^*_{x_0} := \ell^{z^*(x_0)}_{x_0}, \quad r^*_{x_0} := r^{z^*(x_0)}_{x_0}, \ \text{ and } \ I^*_{x_0} := \big(\ell^*_{x_0}, r^*_{x_0}\big). \tag{5.32}$$

The next lemma provides useful properties on the function γ^* and its domain of definition. In particular, it discusses its dependance with respect to the starting point x_0.

LEMMA 5.11. *Assume that α satisfies Conditions (2.12) and let x_0 be arbitrary in $\mathbf{D}^-$. Then, the following holds.*

(i) $z^*(x_0) \in (\Gamma(x_0), 2x_0)$ *and* $\gamma^*_{x_0}$ *has a positive derivative on the interval* $I^*_{x_0}$.
(ii) $\big(d(x_0), u(x_0)\big) \subset I^*_{x_0}$ *and* $\lim_{x \to r^*_{x_0}} \gamma^*_{x_0}(x) = r^*_{x_0} \le \Gamma^\infty$ *with equality if* $u(x_0) = \Gamma^\infty$.
(iii) *For* $x_0, x_1 \in \mathbf{D}^-$, *we have either* $I^*_{x_0} \cap I^*_{x_1} = \emptyset$, *or* $I^*_{x_0} = I^*_{x_1}$ *and* $\gamma^*_{x_0} = \gamma^*_{x_1}$.

Proof. We fix $x_0 \in \mathbf{D}^-$ and prove each assertion separately. The proofs of points (i) and (iii) are very close to the proof of lemma 5.2 in Espinosa and Touzi (2012), but we rewrite and adapt them here.

(i) Lemma 5.10 (iii) ensures the existence of $a_0 < 2x_0$ such that $\mathcal{L}g\big(x, \gamma^z_{x_0}(x)\big) > 0$ for any $x \ge x_0$ and $z \ge a_0$. By definition of $z^*(x_0)$, we deduce that $z^*(x_0) \le a_0 < 2x_0$. Since $x_0 \in \mathbf{D}^-$, we obtain from Lemma 5.10 (ii) that $\Gamma(x_0) \in \mathbf{Z}(x_0)$ and deduce that $\Gamma(x_0) \le z^*(x_0)$.

In order to prove that $z^*(x_0) \in (\Gamma(x_0), 2x_0)$, we now assume that $z^*(x_0) = \Gamma(x_0)$ and work toward a contradiction. Since $\mathcal{L}g$ is continuous, $\mathbf{D}^-$ is an open set and there exists $\varepsilon > 0$ such that $(x_0, x_0 + 2\varepsilon) \subset \mathbf{D}^- \cap (x_0, r^*_{x_0})$ and $d(x) = d(x_0)$ for any $x \in (x_0, x_0 + \varepsilon)$. Let us denote $x_\varepsilon := x_0 + \varepsilon \in \mathbf{D}^-$ and $z_\varepsilon := \Gamma(x_\varepsilon) > \Gamma(x_0)$. By Lemma 5.10 (i), we have $\ell^{z_\varepsilon}_{x_\varepsilon} \le d(x_0) < x_0$, and it follows from Lemma 5.10 (ii) and the dynamics of (5.27) that $\gamma^{z_\varepsilon}_{x_\varepsilon}$ is decreasing on $(x_0, r^{z_\varepsilon}_{x_\varepsilon})$. Therefore, we compute

$$\gamma^{z_\varepsilon}_{x_\varepsilon}(x_0) \ > \ \gamma^{z_\varepsilon}_{x_\varepsilon}(x_\varepsilon) = \Gamma(x_\varepsilon) \ > \ \Gamma(x_0) = z^*. \tag{5.33}$$

On the other hand, since $\gamma^{\gamma^{z_\varepsilon}_{x_\varepsilon}(x_0)}_{x_0}(x_\varepsilon) = z^\varepsilon = \Gamma(x_\varepsilon)$, Lemma 5.10 (ii) ensures that $\gamma^{z^\varepsilon}_{x_\varepsilon}(x_0) \in \mathbf{Z}(x_0)$, leading to $z^* \ge \gamma^{z^\varepsilon}_{x_\varepsilon}(x_0) \in \mathbf{Z}(x_0)$, which contradicts (5.33).

The same line of argument implies also that $(x, \gamma^*(x)) \in \mathrm{Int}(\Gamma^+)$ for any $x \in I^*_{x_0}$. We deduce from the dynamics of (5.27) that γ^* has a positive derivative on $I^*_{x_0}$, and in particular $\lim_{x \to r^*_{x_0}} \gamma^*(x)$ exists.

(ii) For any $z \in \mathbf{Z}(x_0)$, since $\gamma^z_{x_0}$ is nonincreasing in Γ^-, we deduce that $\lim_{x \to r^z_{x_0}} \gamma^z_{x_0}(x) = r^z_{x_0} \le \Gamma^\infty$. Let us write $r_0 := \sup\{r^z_{x_0};\, z \in \mathbf{Z}(x_0)\} \le \Gamma^\infty$. Let us first prove that $r_0 \ge u(x_0)$. Assume on the contrary that $r_0 < u(x_0)$, so that $\Gamma(r_0) > r_0$. Let fix $z \in (r_0, \Gamma(r_0))$. By Lemma 5.10 (i), $\ell^z_{r_0} \le d(x_0)$, so that $x_0 \in I^z_{r_0}$ and, since Lemma 5.10 (ii) implies that

$\mathcal{L}g(x, \gamma^z_{r_0}(x)) < 0$ for $x > r_0$, we deduce that $\gamma^z_{r_0}(x_0) \in \mathbf{Z}(x_0)$. This contradicts the definition of r_0 since $z \in (r_0, \Gamma(r_0))$ implies that $r^{\gamma^z_{r_0}(x_0)}_{x_0} = r^z_{r_0} > r_0$. In conclusion, $r_0 \geq u(x_0)$. Besides, Lemma 5.10 (i) implies that $\ell^*_{x_0} \leq d(x_0)$, and we intend to prove that $r_0 = r^*_{x_0}$ in order to derive $(d(x_0), u(x_0)) \subset I^*_{x_0}$.

First, we derive the existence of a sequence $(z_n) \in \mathbf{Z}(x_0)$ such that $z_n \to z^*(x_0)$ and $r^{z_n}_{x_0} \to r_0$. Combining the one-to-one property of the flow with the property that $\lim_{x \to r^z_{x_0}} \gamma^z_{x_0}(x) = r^z_{x_0}$ for $z \in \mathbf{Z}(x_0)$, we deduce that $z \mapsto r^z_{x_0}$ is nondecreasing on $\mathbf{Z}(x_0)$. Hence, if $z^*(x_0) \notin \mathbf{Z}(x_0)$, any sequence (z_n) valued in $\mathbf{Z}(x_0)$ such that $z_n \to z^*(x_0)$ satisfies also $\sup\{r^{z_n}_{x_0}; n \in \mathbb{N}\} = \sup\{r^z_{x_0}; z \in \mathbf{Z}(x_0)\} = r_0$ and thus the required property. If ever $z^*(x_0) \in \mathbf{Z}(x_0)$, we simply pick the sequence $z_n := z^*(x_0)$, for any $n \in \mathbb{N}$.

We now prove that $r_0 = r^*_{x_0}$. Let $z \in (r_0, 2r_0)$ be arbitrary. Up to a subsequence, we have by construction $I^{z_n}_{x_0} \cap I^z_{r_0} \neq \emptyset$ for any $n \in \mathbb{N}$. We know that $\lim_{x \to r^{z_n}_{x_0}} \gamma^{z_n}_{x_0}(x) = r^{z_n}_{x_0}$ and $\gamma^z_{r_0}(r^{z_n}_{x_0}) > r^{z_n}_{x_0}$ since $r^z_{r_0} > r_0 \geq r^{z_n}_{x_0}$, for any $n \in \mathbb{N}$. Hence, the one-to-one property of the flow ensures that $\gamma^z_{r_0}(x) > \gamma^{z_n}_{x_0}(x)$ for any $x \in I^{z_n}_{x_0} \cap I^z_{r_0}$ and $n \in \mathbb{N}$. By Lemma 5.10 (i), $\lim_{x \to \ell^z_{r_0}} \gamma^z_{r_0}(x) = \ell^z_{r_0}$, so that $x_0 \in I^{z_n}_{x_0} \cap I^z_{r_0}$. Since $(z_n) \in \mathbf{Z}(x_0)$ converges to $z^*(x_0) = \sup \mathbf{Z}(x_0)$, we deduce that $\gamma^z_{r_0}(x_0) \geq \gamma^*_{x_0}(x_0) = z^*(x_0) \geq z^n = \gamma^{z_n}_{x_0}(x_0)$, for any $n \in \mathbb{N}$. Hence, the one-to-one property of the flow implies that

$$2x > \gamma^z_{r_0}(x) \geq \gamma^*_{x_0}(x) \geq \gamma^{z_n}_{x_0}(x), \quad x \in [x_0, r^{z_n}_{x_0} \wedge r^*_{x_0}), \quad n \in \mathbb{N}. \tag{5.34}$$

Therefore $r^*_{x_0} \geq r^{z_n}_{x_0}$ for $n \in \mathbb{N}$, and, passing to the limit, we get $r^*_{x_0} \geq r_0$. Besides, (5.34) implies that $\limsup_{x \to r_0} \gamma^*_{x_0}(x) \leq \gamma^z_{r_0}(r_0) = z$, and the arbitrariness of $z \in (r_0, 2r_0)$ leads to $\limsup_{x \to r_0} \gamma^*_{x_0}(x) \leq r_0$. Since $\gamma^*_{x_0}(x) \geq x$ for $x \in I^*_{x_0}$, we get $\lim_{x \to r_0} \gamma^*_{x_0}(x) = r_0$ and $r^*_{x_0} \leq r_0$. Hence, $r^*_{x_0} = r_0 \leq \Gamma^\infty$, and, if $u(x_0) = \Gamma^\infty$, $r_0 \geq u(x_0)$ implies that $r^*_{x_0} = \Gamma^\infty$.

(iii) Let x_1 in $\mathbf{D}^-$. Suppose that $x_0 < x_1$ and that there exists $x_2 \in I^*_{x_0} \cap I^*_{x_1}$. If ever $\gamma^*_{x_0}(x_2) = \gamma^*_{x_1}(x_2)$, the one-to-one property of the flow combined with the maximality of I^* imply that $I^*_{x_0} = I^*_{x_1}$ and $\gamma^*_{x_0} = \gamma^*_{x_1}$ and conclude the proof. It therefore only remains to prove that $\gamma^*_{x_0}(x_2) = \gamma^*_{x_1}(x_2)$.

We assume on the contrary that $\gamma^*_{x_0}(x_2) < \gamma^*_{x_1}(x_2)$, the case where $\gamma^*_{x_0}(x_2) > \gamma^*_{x_1}(x_2)$ being treated similarly. The one-to-one property of the flow implies that $\gamma^*_{x_0} < \gamma^*_{x_1}$ on all the interval $I^*_{x_0} \cap I^*_{x_1}$. Furthermore, Lemma 5.11 (i) and Lemma 5.10 (i) ensure that $\lim_{r^*_{x_1}} \gamma^*_{x_1} = r^*_{x_1}$ and $\lim_{\ell^*_{x_1}} \gamma^*_{x_1} = \ell^*_{x_1}$. Hence, we deduce from the maximality of $I^*_{x_1}$ that $I^*_{x_0} \subset I^*_{x_1}$. Combining the definition of $z^*(x_1)$ with the continuity of the flow with respect to initial data, we obtain the existence of $z \in \mathbf{Z}(x_1)$ such that $z < z^*(x_1)$ and $\gamma^*_{x_0}(x_2) < \gamma^z_{x_1}(x_2) < \gamma^*_{x_1}(x_2)$. Once again, the one-to-one property of the flow implies that $I^*_{x_0} \subset I^z_{x_1}$ and $\gamma^*_{x_0} < \gamma^z_{x_1} < \gamma^*_{x_1}$ on $I^*_{x_0}$. Since $z \in \mathbf{Z}(x_1)$, we deduce that $\gamma^z_{x_1}(x_0) \in \mathbf{Z}(x_0)$ while $\gamma^z_{x_1}(x_0) > z^*(x_0) = \gamma^*_{x_0}(x_0)$, which contradicts the definition of $z^*(x_0)$. □

Finally, we are in position to provide the proof of Proposition 5.5:

Proof of Proposition 5.5. This construction follows similar ideas as in the proof of proposition 5.1 in Espinosa and Touzi (2012), but turns out to be simpler since $\Gamma^\infty < \infty$.

Let

$$\mathcal{D} := \bigcup_{x_0 \in \mathbf{D}^-} I^*(x_0) \supset \mathbf{D}^-. \tag{5.35}$$

Lemma 5.11 (iii) ensures that , for any x_0 and x_1 in $\mathbf{D}^-$, we either have $I^*_{x_0} = I^*_{x_1}$ or $I^*_{x_0} \cap I^*_{x_1} = \emptyset$. Hence, there exists a subset $\mathbf{D}^-_0$ of $\mathbf{D}^-$ such that $\mathcal{D} = \bigcup_{x_0 \in \mathbf{D}^-_0} I^*(x_0)$ and, for any $x_0, x_1 \in \mathbf{D}^-_0$, $x_0 \neq x_1$ implies that $I^*_{x_0} \cap I^*_{x_1} = \emptyset$.

We now define the function γ on $\mathbb{R}_+\backslash\{0\}$ by:

$$\gamma(x) := \begin{cases} \gamma^*_{x_0}(x) & \text{if } x \in I^*_{x_0}, \text{ for } x_0 \in \mathbf{D}_0^- \\ x & \text{otherwise.} \end{cases} \tag{5.36}$$

According to Lemma 5.11, this definition does not depend on the choice of $\mathbf{D}_0^-$.

Lemmata 5.10 and 5.11 imply that γ is continuous at the endpoints $\ell^*_{x_0}$ and $r^*_{x_0}$, for any $x_0 \in \mathbf{D}_0^-$. Hence, setting $\gamma(0) := 0$, we obtain a continuous function γ on $\mathbb{R}_+$. For any $x_0 \in \mathbf{D}_0^-$, $\gamma^*_{x_0}$ is increasing on $I^*_{x_0}$ and the identity function is increasing as well, so that γ is increasing on $\mathbb{R}_+$. We now justify each assertion of the proposition separately. (i) is immediate from the definition of γ.

To prove (ii), we first notice that $\{x \geq 0 : \gamma(x) = x\} = \mathbb{R}_+\backslash\mathcal{D} \subset \mathbb{R}_+\backslash\mathbf{D}^-$, so that $\mathcal{L}g(x, x) \geq 0$ on the set $\{x > 0 : \gamma(x) = x\}$. On the set $\{x > 0 : \gamma(x) > x\}$, since γ has a positive derivative by Lemma 5.11 (ii) and satisfies (5.27), we have $\mathcal{L}g(x, \gamma(x)) > 0$. Finally, (iii) can be deduced from Lemma 5.11 (ii), since $r^*_{x_0} \leq \Gamma^\infty$ for any $x_0 \in \mathbf{D}_0^-$. □

REFERENCES

DAI, M., H. JIN, Y. ZHONG, and X. ZHOU (2010): Buy Low and Sell High, in *Contemporary Quantitative Finance: Essays in Honour of Eckhard Platen*, C. Chiarella and A. Novikov, eds. Berlin: Springer-Verlag 317–334.

DU TOIT, J., and G. PESKIR (2007): The Trap of Complacency in Predicting the Maximum, *Ann. Probab.* 35, 340–365.

DU TOIT, J., and G. PESKIR (2008): Predicting the Time of the Ultimate Maximum for Brownian Motion with Drift. *Proc. Math. Control Theory Finance* (Lisbon 2007), Berlin: Springer, pp. 95–112.

DU TOIT, J., and G. PESKIR (2009): Selling a Stock at the Ultimate Maximum, *Ann. Appl. Probab.* 19(3), 983–1014.

ESPINOSA, G. E., and N. TOUZI (2012), Detecting the Maximum of a Scalar Diffusion with Negative Drift, *Siam J. Control Optim.* 50(5), 2543–2572.

GRAVERSEN, S. E., G. PESKIR, and A.N. SHIRYAEV (2001): Stopping Brownian Motion without Anticipation as Close as Possible to Its Ultimate Maximum, *Theory Probab. Appl.* 45, 125–136.

HOBSON, D. (2007): Optimal Stopping of the Maximum Process: A Converse to the Results of Peskir, *Stochastics* 79(1), 85–102.

PESKIR, G. (1998): Optimal Stopping of the Maximum Process: The Maximality Principle. *Ann. Probab.* 26(4), 1614–1640.

SHIRYAEV, A. N. Z. XU, and X. ZHOU (2008): Thou Shalt Buy and Hold, *Quant. Finance* 8, 765–776.

URUSOV, M. A. (2005): On a Property of the Moment at Which Brownian Motion Attains Its Maximum and Some Optimal Stopping Problems, *Theory Probab. Appl.* 49, 169–176.

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

Mathematical Finance, Vol. 25, No. 4 (October 2015), 789–826

STATIC FUND SEPARATION OF LONG-TERM INVESTMENTS

PAOLO GUASONI

Boston University and Dublin City University

SCOTT ROBERTSON

Carnegie Mellon University

This paper proves a class of *static fund separation* theorems, valid for investors with a long horizon and constant relative risk aversion, and with stochastic investment opportunities. An optimal portfolio decomposes as a constant mix of a few preference-free funds, which are common to all investors. The weight in each fund is a constant that may depend on an investor's risk aversion, but not on the state variable, which changes over time. Vice versa, the composition of each fund may depend on the state, but not on the risk aversion, since a fund appears in the portfolios of different investors. We prove these results for two classes of models with a single state variable, and several assets with constant correlations with the state. In the *linear* class, the state is an Ornstein–Uhlenbeck process, risk premia are affine in the state, while volatilities and the interest rate are constant. In the *square root* class, the state follows a square root diffusion, expected returns and the interest rate are affine in the state, while volatilities are linear in the square root of the state.

KEY WORDS: portfolio choice, fund separation, long horizon.

1. INTRODUCTION

Fund separation means that investors need to trade only a few well-chosen funds, not all the securities in the market. Like primary colors, which mix to span the visible spectrum, these funds, combined in varying weights, span the optimal portfolios of all investors. The idea is as simple as it is important, and its implications range from equilibrium asset pricing to the theory of financial intermediation.

There are two main streams of results.[1] Two-fund separation holds if the risk-free asset and the market span all optimal portfolios, and in equilibrium leads to the Capital Asset Pricing Model. The main assumption of two-fund separation is that investment opportunities are either deterministic or unhedgeable. When they are constant, along with relative risk aversion, investors divide their wealth across funds in proportions that are constant over time, or *static*.

This study was partially supported by the ERC (278295), NSF (DMS-0807994, DMS-1109047), SFI (07/MI/008, 07/SK/M1189, 08/SRC/FMC1389), and FP7 (RG-248896). We thank Bernard Dumas and Hao Xing for useful comments.

Manuscript received June 2011; final revision received August 2012.

Address correspondence to Paolo Guasoni, Mathematics and Statistics, Boston University, 111 Cummington St, Boston, MA 02215, USA; e-mail: guasoni@bu.edu.

[1]Tobin (1958) first derives two-fund separation in a mean variance setting, while Cass and Stiglitz (1970) and Ross (1978) derive it under assumptions on preferences and return distributions, respectively. Chamberlain (1988) and Khanna and Kulldorff (1999) find separating conditions in diffusion models, while Schachermayer, Sîrbu, and Taflin (2009) characterize it in a semimartingale setting.

DOI: 10.1111/mafi.12017

A more complex fund separation obtains if investment opportunities are both stochastic and hedgeable (at least partially). If they depend on k state variables, as in Merton (1973), then $k + 2$ funds are needed to span optimal portfolios. The k funds, in addition to the safe asset and the myopic portfolio, mimic each state variable over time. Then, even investors with constant relative risk aversion (CRRA) need to dynamically change their weights in the $k + 2$ funds, according to complex and generally unknown trading strategies. This is *dynamic fund separation*.

This paper proves *static fund separation* theorems, valid for investors with a long horizon and constant relative risk aversion, and for investment opportunities depending on a single state variable. Unlike two-fund separation theorems, we allow for partially hedgeable investment opportunities, and obtain three or more spanning funds. Unlike the classical $k + 2$ separation, we obtain a static decomposition, whereby each investor holds the funds in constant proportions.

Our setting of a single state variable ($k = 1$) already implies dynamic fund separation with three funds: the safe asset, the myopic portfolio, and the hedging portfolio. But only the myopic weight is known in this decomposition, while the risk-free and hedging weights depend jointly on the residual horizon and on the state variable. This dependence severely limits the significance of dynamic separation, and calls for a more precise description of optimal portfolios.

With static fund separation, optimal portfolios are weighted sums of a fixed number of funds, which are common to all investors. The weight of each fund is constant: it may depend on the risk aversion and market parameters, but not on the state variable, which changes over time. Vice versa, the composition of each fund may depend on the state variable and on market parameters, but not on the risk aversion, since the same fund appears in the portfolios of different investors. Because static fund separation requires constant fund weights, in general it implies more funds than dynamic fund separation. In addition, the number of static funds depends on the model considered, and so do fund weights.

We prove two static separation results, corresponding to two mainstream classes of models. Their common features are a single state variable and several assets, with constant correlations with the state. In the *linear* class, the state is an Ornstein–Uhlenbeck process, risk premia are affine in the state, while volatilities and the interest rate are constant. In the *square root* class, the state follows Feller's (1951) square root diffusion, expected returns and the interest rate are affine in the state, while volatilities are linear in the square root of the state. The first class includes, in particular, the models considered in the literature on predictability of stock returns. Although the statistical significance of predictability remains controversial, its potential welfare gains are large, and this paper shows that they could be realized even by unsophisticated investors through a small number of mutual funds.[2]

While our static fund separation results are based on the joint assumption of CRRA preferences and long horizons, their implications have a broader scope, in both directions. First, Brandt (1999), Barberis (2000), and Wachter (2002) report that optimal portfolios for a 10-year horizon are already close to their long run limit. Second, turnpike theorems[3]

[2]For the debate on predictability, see Welch and Goyal (2008), Cochrane (2008), and Campbell and Thompson (2008). For models with predictable returns, see Kim and Omberg (1996), Xia (2001), Wachter (2002), and Bensoussan, Keppo, and Sethi (2009) among others.

[3]For turnpike theorems, see Leland (1972), Hakansson (1974), Huberman and Ross (1983), Cox and Huang (1992), Huang and Zariphopoulou (1999), Dybvig, Rogers, and Back (1999), and Guasoni et al. (2011).

suggest that, at long horizons, optimal CRRA portfolios are approximately optimal for a wide class of utility functions.

The rest of the paper is organized as follows: Section 2 describes the general framework, and outlines the main ideas behind our static fund separation results. These ideas are made precise in Sections 3 and 4, respectively, for the linear and square root models. In each model, static fund separation holds with four funds. We find explicit formulas for both the funds and their weights, and display their composition for typical values of risk aversion, using the market parameters estimated by Barberis (2000) and Pan (2002). Section 5 concludes, and all proofs are in the Appendixes.

2. PROBLEM AND HEURISTIC SOLUTION

This section describes the general setting of the paper, and shows the steps which unify the main results, leaving aside the technical details, which differ across models. Here arguments are presented at a heuristic level, while the next sections contain their precise versions for two classes of models.

2.1. Market and Preferences

The market has a safe asset and several risky assets. Investment opportunities are modeled by a single state variable Y, which drives the safe rate r, the excess returns μ, and the volatility matrix σ. In summary, the prices of the safe asset S^0 and risky assets $S^1, \ldots, S^n$ satisfy:

$$\frac{dS_t^0}{S_t^0} = r(Y_t)\,dt, \tag{2.1}$$

$$\frac{dS_t^i}{S_t^i} = r(Y_t)\,dt + dR_t^i \quad 1 \le i \le n, \tag{2.2}$$

where the cumulative excess returns $R = (R^1, \ldots, R^n)$ and the state variable Y follow the diffusion:

$$dR_t^i = \mu_i(Y_t)\,dt + \sum_{j=1}^{n} \sigma_{ij}(Y_t)\,dZ_t^j \qquad 1 \le i \le n, \tag{2.3}$$

$$dY_t = b(Y_t)\,dt + a(Y_t)\,dW_t, \tag{2.4}$$

$$\langle Z^i, W\rangle_t = \rho_i dt \qquad 1 \le i \le n, \tag{2.5}$$

$Z = (Z^1, \ldots, Z^n)$ and W are Brownian motions, and $\rho = (\rho_1, \ldots, \rho_n)$ denotes their vector of cross correlations. $\Sigma = \sigma\sigma' = d\langle R, R\rangle_t/dt$ defines the covariance matrix of returns, while $A = a^2 = d\langle Y, Y\rangle_t/dt$ is the variance rate of the state variable, and $\Upsilon = \sigma\rho a = d\langle R, Y\rangle_t/dt$ is the covariation rate between asset returns and the state variable (the prime sign denotes matrix transposition). Let $E = (\alpha, \beta)$ with $-\infty \le \alpha < \beta \le \infty$ be an open interval such that the state variable remains in E at all times. For example, $E = (-\infty, \infty)$ for the linear model of Section 3, and $E = (0, \infty)$ for the square root model of Section 4. The

market defined above is potentially incomplete, because asset returns do not perfectly span changes in the state variable.

An investor trades in the market according to a portfolio $\pi = (\pi_t)_{t\geq 0}$, which represents the proportions of wealth in each risky asset. Since the investor observes the state variable Y and the asset returns R, the portfolio π is adapted to the augmentation of the filtration generated by (R, Y), and is R-integrable. The corresponding wealth process $X^\pi = (X^\pi_t)_{t\geq 0}$ satisfies:

$$\frac{dX^\pi_t}{X^\pi_t} = r(Y_t)\,dt + \pi_t' dR_t. \tag{2.6}$$

Since a positive initial capital $X_0 \geq 0$ implies a positive wealth at all times ($X^\pi_t \geq 0$ a.s. for all $t \geq 0$), doubling strategies are excluded. Investors' preferences display CRRA, so that their marginal utilities are

$$U'(x) = x^{p-1} \quad p < 1. \tag{2.7}$$

For a fixed planning horizon $T > 0$ and a current time $0 \leq t < T$, the investor's goal is to maximize utility from terminal wealth, given the current wealth x and the current state y:

$$\max_{(\pi_u)_{u\geq t}} \frac{1}{p} E\big[\big(X^\pi_T\big)^p \mid Y_t = y,\, X_t = x\big]. \tag{2.8}$$

2.2. Stochastic Control

Start by defining the value function $V(t, X_t, Y_t)$ as

$$V(t, X_t, Y_t) = \sup_{(\pi_u)_{u\geq t}} \frac{1}{p} E\big[\big(X^\pi_T\big)^p \mid Y_t, X_t\big]. \tag{2.9}$$

Following usual control arguments, (2.9) leads to the Hamilton–Jacobi–Bellman (HJB) equation:

$$V_t + b V_y + \frac{a^2}{2} V_{yy} + r x V_x + \sup_\pi \left(\pi'(\mu V_x + \Upsilon V_{xy})x + \frac{x^2 V_{xx}}{2}\pi'\Sigma\pi\right) = 0 \tag{2.10}$$

with the terminal condition $V(T, x, y) = x^p/p$. Here subscripts denote partial derivatives, and a, b, μ, Σ, Υ are functions of y, although their dependence is omitted to simplify notation. Because V is concave in x, and recalling that $\sup_\pi(\pi' b + \frac{1}{2}\pi' A\pi) = -\frac{1}{2} b' A^{-1} b$ for A negative definite, the equation becomes

$$V_t + b V_y + \frac{a^2}{2} V_{yy} + r x V_x - (\mu V_x + \Upsilon V_{xy})' \frac{\Sigma^{-1}}{2V_{xx}} (\mu V_x + \Upsilon V_{xy}) = 0 \tag{2.11}$$

and the corresponding optimal portfolio is $\pi = -\Sigma^{-1}\mu \frac{V_x}{xV_{xx}} - \Sigma^{-1}\Upsilon \frac{V_{xy}}{xV_{xx}}$. Since power utility is homothetic, i.e., $U(cx) = c^p U(x)$, and payoffs can be scaled arbitrarily, the value function is also homothetic. Thus, writing $V(t, x, y) = u(t, y)x^p/p$, the HJB equation in terms of the reduced value function u becomes

$$u_t + (b - q\Upsilon'\Sigma^{-1}\mu)u_y + \frac{a^2}{2} u_{yy} + \left(pr - \frac{q}{2}\mu'\Sigma^{-1}\mu\right) u - q\frac{a^2}{2}\rho'\rho\frac{u_y^2}{u} = 0, \tag{2.12}$$

where $q = p/(p-1)$, and the terminal condition is now $u(T, y) = 1$. The optimal portfolio similarly reduces to

$$\pi(t, y) = \frac{1}{1-p}\left(\Sigma^{-1}\mu + \Sigma^{-1}\Upsilon\frac{u_y}{u}\right), \tag{2.13}$$

which is the traditional *dynamic* three-fund separation into the safe asset, the myopic portfolio $\Sigma^{-1}\mu$, and the intertemporal hedging component $\Sigma^{-1}\Upsilon$. The word *dynamic* refers to the dependence of the intertemporal weight $\frac{u_y}{u}$ on the value function, which entails a complex dynamic trading strategy, depending jointly on the horizon, the state variable, and the risk aversion.

Equation (2.12) simplifies further using the assumption that $\rho'\rho$ is constant. Then, the power substitution of Zariphopoulou (2001) makes this equation linear. Setting $u(t, y) = v(t, y)^{\delta}$, where $\delta = 1/(1 - q\rho'\rho)$, (2.12) turns into a linear parabolic equation for v:

(2.14)

$$\begin{aligned} v_t + \frac{1}{2}Av_{yy} + (b - q\Upsilon'\Sigma^{-1}\mu)v_y + \frac{1}{\delta}\left(pr - \frac{q}{2}\mu'\Sigma^{-1}\mu\right)v = 0 \qquad & (t, y) \in (0, T) \times E, \\ v(T, y) = 1 \qquad & y \in E, \end{aligned}$$

and the optimal portfolio accordingly reduces to

$$\pi(t, y) = \frac{1}{1-p}\left(\Sigma^{-1}\mu + \delta\Sigma^{-1}\Upsilon\frac{v_y}{v}\right). \tag{2.15}$$

By now, these steps have become standard, and underlie virtually all explicit solutions to portfolio choice problems.

2.3. Eigenvalue Representation

We exploit an eigenvalue expansion, in which the the principal eigenvalue and its eigenvector drive the long horizon limit. The first step in this direction is to rewrite (2.14) in self-adjoint form. Set

$$b^v = b - q\Upsilon'\Sigma^{-1}\mu \quad V = \frac{1}{\delta}\left(pr - \frac{q}{2}\mu'\Sigma^{-1}\mu\right) \tag{2.16}$$

and define the differential operator:

$$L^v = \frac{1}{2}A\frac{d^2}{dy^2} + b^v\frac{d}{dy}, \tag{2.17}$$

so that (2.14) becomes

$$\begin{aligned} v_t + L^v v + Vv = 0 \qquad & (t, y) \in (0, T) \times E, \\ v(T, y) = 1 \qquad & y \in E. \end{aligned} \tag{2.18}$$

To ease presentation, the symbols $\dot{f}$, $\ddot{f}$ replace f_y, f_{yy}, the partial derivatives with respect to the state y, for functions $f(y)$ of the state variable alone. Define m as the solution to the ordinary differential equation (ODE):

$$\frac{\dot{m}}{m} + \frac{\dot{A}}{A} = \frac{2b^{\nu}}{A}. \tag{2.19}$$

If m integrable, it can be normalized to unit mass, and represents the steady-state distribution of the state variable Y under the equivalent probability measure associated to L^{ν}. The self-adjoint version of (2.14) is then

$$\begin{aligned} v_t m + \frac{1}{2}(A v_y m)_y + V v m = 0 \qquad & (t, y) \in (0, T) \times E, \\ v(T, y) = 1 \qquad & y \in E. \end{aligned} \tag{2.20}$$

The classical strategy is to solve such an equation by separation of variables. Define the operator $-M$ with a certain domain $\mathcal{D}(-M) \subset L^2(E, m)$ as

$$M\phi = \frac{(a^2 \dot{\phi} m)^{\cdot}}{2m} + V\phi \qquad \phi \in \mathcal{D}(-M). \tag{2.21}$$

If $(-M, \mathcal{D}(-M))$ is Hilbert-Schmidt[4] then a natural guess for the solution to the boundary value problem is

$$v(t, y) = \sum_{n \geq 0} \alpha_n \phi_n(y) e^{-\lambda_n (T-t)} \tag{2.22}$$

provided that $1 \in L^2(E, m)$ (or that m is a probability density), because v must satisfy the terminal condition $v(T, y) = 1$. Then, the coefficients $(\alpha_n)_{n \geq 0}$ are $\alpha_n = \int_E \phi_n(x) m(x) dx$, and α_n does not depend on T.

The expansion in (2.22) has important consequences for portfolio choice. It implies that v has the representation:

$$v(t, y) = e^{-\lambda_0 (T-t)} \left(\alpha_0 \phi_0(y) + \sum_{n \geq 1} \alpha_n \phi_n(y) e^{-(\lambda_n - \lambda_0)(T-t)} \right). \tag{2.23}$$

Consider the terms within the parentheses in the above expression. Since each eigenfunction ϕ_n, $n \geq 1$ carries a weight $\alpha_n e^{-(\lambda_n - \lambda_0)(T-t)}$ that decreases in the horizon T, for long horizons both the value function u and the portfolio π are determined by the principal eigenfunction ϕ_0 alone. Indeed, (2.23) formally yields

$$\lim_{T \uparrow \infty} \frac{v_y(t, y)}{v(t, y)} = \lim_{T \uparrow \infty} \frac{\alpha_0 \dot{\phi}_0(y) + \sum_{n \geq 1} \alpha_n \dot{\phi}_n(y) e^{-(\lambda_n - \lambda_0)(T-t)}}{\alpha_0 \phi_0(y) + \sum_{n \geq 1} \alpha_n \phi_n(y) e^{-(\lambda_n - \lambda_0)(T-t)}} = \frac{\dot{\phi}_0(y)}{\phi_0(y)} \tag{2.24}$$

provided that $\alpha_0 \neq 0$. When $\phi_0 > 0$, $\alpha_0 = \int_E \phi_0(x) m(x) dx > 0$. Regarding the optimal portfolios, the above calculation implies, by (2.15)

$$\lim_{T \uparrow \infty} \pi(t, y) = \frac{1}{1-p} \Sigma^{-1} \mu + \frac{\delta}{1-p} \Sigma^{-1} \Upsilon \frac{\dot{\phi}_0(y)}{\phi_0(y)}. \tag{2.25}$$

Thus, as the horizon increases, the optimal portfolio converges to a time-homogeneous portfolio that depends only on the current state variable y.

[4]An operator is Hilbert–Schmidt if it is self-adjoint, with a discrete spectrum, bounded from below, tending to ∞, and the eigenfunctions $(\phi_n)_{n \geq 0}$ form an orthonormal basis for $L^2(E, m)$.

2.4. Static Fund Separation

Equation (2.25) paves the way to static fund separation for long-term investments. Optimal portfolios for long planning horizons decompose as combinations of a fixed number of funds, such that

(i) The composition of each fund may depend on the state variable y, but is independent of the risk aversion $1 - p$.
(ii) Each investor chooses fund weights that depend on risk aversion $1 - p$, but not on the state variable y.

The crucial point is that most models lead to a decomposition of the form:

$$\frac{\dot{\phi}_0(y)}{\phi_0(y)} = \sum_{i=1}^{m} w_i(p)\psi_i(y), \tag{2.26}$$

which implies that any optimal portfolio for a long horizon is the combination of $m + 2$ funds, independent of the preference parameter p: the safe asset, the myopic portfolio $\Sigma^{-1}\mu$, and the m hedging portfolios $\Sigma^{-1}\Upsilon\psi_i(y)$. Risk aversion determines optimal portfolios only through their weights in these funds, but does not affect the funds themselves. Thus, the funds are the same for all long-horizon investors, regardless of their risk aversion.

Static fund separation entails a clear division of labor between an investor, or her financial planner, on one side, and an intermediary, such as a mutual fund manager, on the other. The investor or her financial planner choose the weights in the funds, as they are in the best position to assess the investor's tolerance for risk. As in usual dynamic fund separation, investors do not need to trade the single securities in the funds' portfolios.

Furthermore—and this is the hallmark of *static* fund separation—investors do not trade in response to changes in the state variable y, but only rebalance as to keep fund proportions constant over time. Investors do not even need to *observe* the state variable. The manager, on the other hand, trades on behalf of several investors, with a broad range of risk attitudes. He does not need to know the risk tolerance of investors, because the composition of each fund is independent of preferences. On the contrary, the manager must observe the state variable, as this is the only trading signal affecting security weights within each fund.

The next sections carry out this program in detail for two classes of models, which nest several examples in the literature. For each class, the state variable follows one of the two basic stationary processes in finance: the Ornstein–Uhlenbeck and the Feller diffusions. In both cases, it is possible to turn the above heuristic arguments into precise results, but at the cost of some parametric restrictions, which guarantee the well-posedness of optimization problems. In this regard, we tend to favor simpler to sharper results, sometimes concentrating on the most relevant case of higher risk aversion than logarithmic utility ($p < 0$).

There are two main technical difficulties in turning the heuristic arguments into precise statements. First, although finite linear combinations of functions of the form $e^{-\lambda_n(T-t)}\phi_n(y)$ satisfy the HJB equation, infinite sums may not commute with derivatives, especially when the state space E is not compact, as in the models considered. Checking that the guess in (2.22) solves (2.15) involves some careful arguments, which exploit the specific properties of the eigenfunctions ϕ_n in each model. Second, the solution

to the HJB equation must correspond to the value function of the utility maximization problem—a verification theorem is needed.

3. LINEAR MODEL

This section studies a model with a state variable, which follows an Ornstein–Uhlenbeck process. The expected returns of the risky assets depend linearly on the state variable, while the interest rate and the volatilities are constant:

$$\begin{aligned} dR_t &= (\sigma\nu_0 + b\sigma\nu_1 Y_t)\,dt + \sigma\,dZ_t, \\ dY_t &= -bY_t dt + dW_t, \\ d\langle R, Y\rangle_t &= \rho dt, \\ r(Y_t) &= r_0, \end{aligned} \tag{3.1}$$

where $\sigma \in \mathbb{R}^{n\times n}$; $\nu_0, \nu_1, \rho \in \mathbb{R}^n$; $b, r_0 > 0$, $\nu_1'\nu_1 > 0$. We consider the parametric restriction:

ASSUMPTION 3.1.

$$1 + q\rho'\nu_1 > 0, \quad \textit{where } q = p/(p-1). \tag{3.2}$$

To understand how this restriction arises, observe that m in (2.19) takes the form:

$$m(y) = Ke^{-b(1+q\rho'\nu_1)y^2 - 2q\rho'\nu_0 y} \quad y \in \mathbb{R}, \tag{3.3}$$

where $K > 0$ is an arbitrary constant. Thus, unless (3.2) holds, there is no solution m with finite integral. The eigenvalue equation $-M\phi = \lambda\phi$ for M in (2.21) specifies to

$$\frac{1}{2}(\dot{\phi}m)^{\cdot} + \left(\frac{1}{\delta}\left(pr_0 - \frac{q}{2}\nu_0'\nu_0 - qb\nu_0'\nu_1 y - \frac{q}{2}b^2\nu_1'\nu_1 y^2\right) + \lambda\right)\phi m = 0. \tag{3.4}$$

The following lemma identifies the solutions of (3.4) with those of the differential equation of the harmonic oscillator, under the additional parameter restriction:

ASSUMPTION 3.2.

$$b^2\left((1+q\rho'\nu_1)^2 + \frac{q}{\delta}\nu_1'\nu_1\right) > 0. \tag{3.5}$$

REMARK 3.3. Note that (3.5) always holds if $p < 0$. For $0 < p < 1$, setting $\nu_1 = -\frac{1-\varepsilon}{q\rho'\rho}\rho$ for a small enough $\varepsilon > 0$ will cause (3.5) to fail even if Assumption 3.1 holds.

LEMMA 3.4. *Let Assumptions 3.1 and 3.2. hold. Let $\phi \in L^2(\mathbb{R}, m)$, and define ψ by the equality:*

$$\phi(y) = \sqrt{\alpha}m(y)^{-1/2}\psi(\alpha y + \beta), \tag{3.6}$$

where the constants $\alpha > 0$, β, $\eta = K_1\lambda + K_2$ are in (B.1). Then, ϕ solves (3.4) if and only if $\psi \in L^2(\mathbb{R})$ solves the ODE

$$-\ddot{\psi}(z) + z^2\psi(z) = \eta\psi(z). \tag{3.7}$$

For $n \geq 0$, consider the Hermite functions

TABLE 3.1
Static Funds in the Linear Model. $w_1(p)$ and $w_2(p)$ Are Given in (3.15)

Fund Name	Portfolio	Fund Weight
Myopic	$\Sigma^{-1}\mu(y)$	$w_v(p) = \frac{1}{1-p}$
Hedging constant	$\Sigma^{-1}\Upsilon$	$w_{hc}(p) = \frac{\delta}{1-p} w_1(p)$
Hedging linear	$\Sigma^{-1}\Upsilon y$	$w_{hl}(p) = \frac{\delta}{1-p} w_2(p)$

$$\psi_n(z) = \sqrt{\frac{1}{n!2^n\sqrt{\pi}}} e^{-\frac{1}{2}z^2} h_n(z), \tag{3.8}$$

where h_n is the nth Hermite polynomial defined by the recurrence relation

$$h_0(z) = 1, \quad h_{n+1}(z) = 2zh_n(z) - \dot{h}_n(z), n \geq 0.$$

It is well known (Miklavčič 1998, section 2.9) that

$$\begin{aligned} &\psi_n \text{ solves (3.7) with } \eta_n = 2n+1, n \geq 0, \\ &(\psi_n)_{n\in\mathbb{N}_0} \text{ is a complete orthonormal basis of } L^2(\mathbb{R}). \end{aligned} \tag{3.9}$$

Set

$$\varphi(z) = \left(\frac{1}{\alpha} m\left(\frac{z-\beta}{\alpha}\right)\right)^{1/2}. \tag{3.10}$$

Assumption 3.1 implies that $\varphi \in L^2(\mathbb{R})$, hence the series

$$\sum_{n=0}^{M} \alpha_n \psi_n, \quad \alpha_n = \int_{\mathbb{R}} \varphi(z)\psi_n(z)\,dz = \int_{\mathbb{R}} \phi_n(y)m(y)\,dy \tag{3.11}$$

converges to φ in $L^2(\mathbb{R})$ as $M \uparrow \infty$. By construction of φ and (3.11) the series $\sum_{n=0}^{M} \alpha_n\phi_n$ converges to 1 in $L^2(\mathbb{R}, m)$ as $M \uparrow \infty$ and hence (2.22) is a candidate solution to (2.18).

The series in (3.11) converges to φ in a very strong sense: since φ is in the space of functions of rapid decrease, $\sum_n \alpha_n\psi_n$ converges to φ in this space (Reed and Simon 1972, theorem V.13, page 143). In other words, for all nonnegative integers l, m

$$\lim_{M\uparrow\infty} \sup_{z\in\mathbb{R}} \left| z^l \frac{d^m}{dz^m} \left(\sum_{n=0}^{M} \alpha_n \psi_n(z) - \varphi(z) \right) \right| = 0. \tag{3.12}$$

This convergence entails that the series expansion for v solves the partial differential equation (PDE) in (2.18). In summary, the following result holds:

THEOREM 3.5. *Let Assumptions 3.1 and 3.2 hold. Define ψ_n, η_n as in (3.9), λ_n, ϕ_n as in Lemma 3.4, and α_n as in (3.11). Then*

(i) the function

$$v(t, y) = \sum_{n=0}^{\infty} e^{-\lambda_n (T-t)} \alpha_n \phi_n(y) \tag{3.13}$$

is a strictly positive $C^{1,2}((0, T) \times E)$ solution of the PDE in (2.18);

(ii) v satisfies the convergence property:

$$\lim_{T \uparrow \infty} \frac{v_y(t, y)}{v(t, y)} = \frac{\dot{\phi}_0(y)}{\phi_0(y)} \quad \text{for all } t > 0,\ y \in \mathbb{R}; \tag{3.14}$$

(iii) the value function of the utility maximization problem (2.9) is equal to $V(x, t, y) = \frac{x^p}{p} v(t, y)^\delta$;

(iv) the decomposition (2.26) (static fund separation) holds with $m = 2$, $\psi_1(y) = 1$ and $\psi_2(y) = y$, and with the corresponding weights in Table 3.1, where:

(3.15)

$$w_1(p) = q\rho' \nu_0 \left(1 - \frac{1}{\sqrt{1 + \dfrac{q\nu_1' \nu_1}{\delta(1 + q\rho'\nu_1)^2}}} \right) - \frac{q\nu_0' \nu_1}{\delta(1 + q\rho'\nu_1)\sqrt{1 + \dfrac{q\nu_1'\nu_1}{\delta(1 + q\rho'\nu_1)^2}}},$$

$$w_2(p) = b(1 + q\rho'\nu_1) \left(1 - \sqrt{1 + \frac{q\nu_1'\nu_1}{\delta(1 + q\rho'\nu_1)^2}} \right).$$

Wachter (2002) considers the following model with a single risky asset, in which the Sharpe ratio follows a mean reverting OU process:

$$\begin{aligned} dR_t &= \sigma X_t dt + \sigma dZ_t, \\ dX_t &= b\left(\bar{X} - X_t\right) dt + \sigma_X dW_t, \\ d\langle Z, W \rangle_t &= \rho dt, \\ r(X_t) &= r_0. \end{aligned} \tag{3.16}$$

TABLE 3.2
Parameter Values for the Linear Model, as in Barberis (2000) and Wachter (2002). Time is in Monthly Units

Parameter	Value
σ_X	0.0189
$\bar{X}$	0.0788
σ	0.0436
b	0.0226
ρ	−0.9350
r_0	0.0014
ν_0	0.0788
ν_1	0.8363

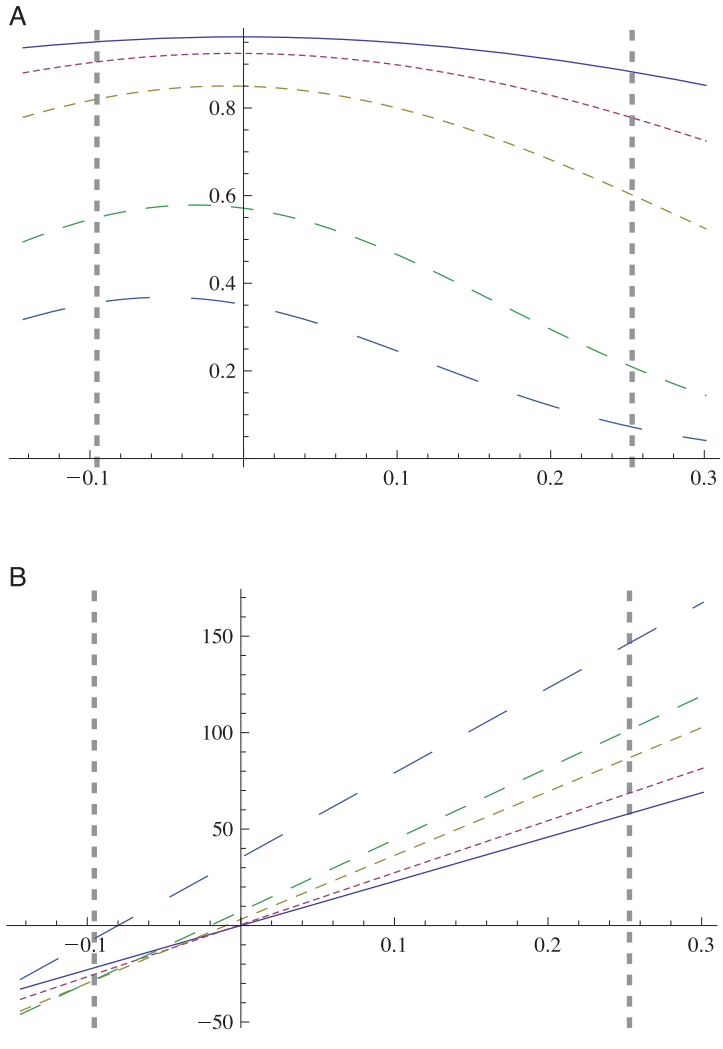

FIGURE 3.1. Reduced value function v^δ (upper panel, vertical axis) and risky portfolio weight (lower panel, vertical axis, in percent) against state variable (horizontal axis), as the planning horizon varies from 0 (solid), 1 year (tiny dashed), 5 years (short dashed), 10 years (medium dashed), and ∞ (long dashed). The state variable is the Sharpe ratio X from (3.16). Vertical lines are at the low (2.5%) and high (97.5%) percentiles of the state variable, under its stationary distribution. Risk aversion is 10 ($p = -9$), and market parameters are as in Table 3.2. All plots use 15 eigenfunctions from the series representation.

The transformation $Y = (X - \bar{X})/\sigma_X$ yields the model in (3.1) with $\nu_0 = \bar{X}$ and $\nu_1 = \sigma_X/b$.

For the parameter values in Table 3.2, Assumptions 3.1, 3.2 hold for risk aversion within the range [0.5, 10] used in the plots and tables below. The upper panel of Figure 3.1 shows the reduced value function v^δ as a function of the Sharpe ratio X as in (3.16), for a range of planning horizons, and for risk aversion of 10 ($p = -9$). The lower panel plots

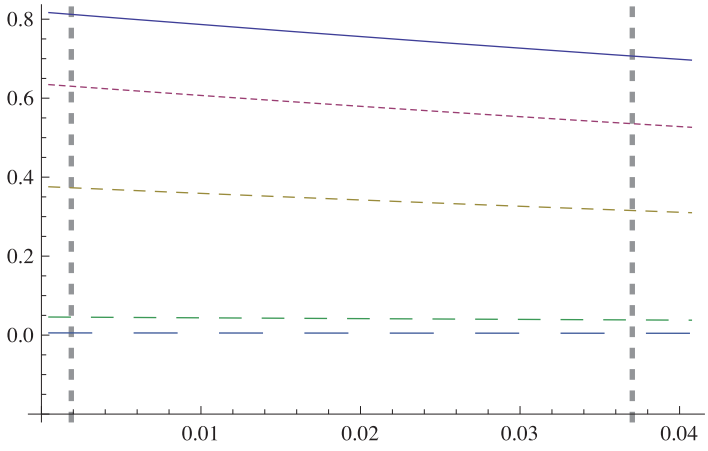

FIGURE 4.1. Reduced value function v^δ (vertical axis) against state variable (horizontal axis), as the planning horizon varies from 3 months (solid), 6 months (tiny dashed), 1 year (short dashed), 3 years (medium dashed), and 5 years (long dashed). Vertical lines are at the low (2.5%) and high (97.5%) percentiles of the state variable, under its stationary distribution. Risk aversion is 10 ($p = -9$), and market parameters are as in Table 4.2. All plots use 15 eigenfunctions from the series representation.

the optimal portfolios as a function of the Sharpe ratio X for a number of horizons, including the myopic ($T \downarrow 0$) and long run ($T \uparrow \infty$) limits, for risk aversion equal to 10. This plot directly compares to figure 3.1 in Wachter (2002): the two figures are obtained from the same set of parameters, with the difference that Wachter (2002) approximates the correlation $\rho = -0.935$ with the value -1 required by her assumption of market completeness. Thus, our results show the effects of this approximation.

Compared to Wachter (2002), Figure 4.1 shows substantially lower stock holdings, especially at longer horizons. This finding is consistent with economic intuition: in a complete market, intertemporal hedging is perfect, therefore the investor takes a larger hedging position. By contrast, in an incomplete setting, intertemporal hedging is imperfect, therefore there is a tradeoff between hedging more, and adding more idiosyncratic volatility. The net result is that an investor hedges less. Since in this model hedging is achieved with a positive stock position, hedging less entails lower stock holdings.

Table 3.3 gives the respective fund weights for various values of the risk aversion $1 - p$ using the parameter values in Table 3.2. Fund weights are calculated for the low (2.5%) and high (97.5%) quantiles of the state variable, which represents the Sharpe ratio. Comparing the second and the last columns in the table shows how intertemporal hedging changes the variability of stock holdings for a range of risk aversions and Sharpe ratios. For risk aversion greater than one, the difference between stock holdings in the high and low state is bigger in the last column: intertemporal hedging leads to a larger variation between stock holdings in good and bad times, relative to the oscillation implied by the myopic strategy in the second column. The effect is reversed for risk aversion less than one, since such investors have negative hedging demands.

Another pattern emerges from the table: for the typical risk aversion greater than one, the hedging component is large when the Sharpe ratio is high, while it is close to

TABLE 3.3
Fund Weights and Implied Risky Positions for the Optimal Long Run Portfolios, for Different Risk Aversion Levels (0.5, 1, 2, 3, 5, 10).

Risk Aversion $1-p$	Myopic Fund Weight $w_v(p)$	Hedging Constant Fund Weight $w_{hc}(p)$	Hedging Linear Fund Weight $w_{hl}(p)$	
Risky (low)	$w_v(p)\Sigma^{-1}\mu(y_{\min})$	$w_{hc}(p)\Sigma^{-1}\Upsilon$	$w_{hl}(p)\Sigma^{-1}\Upsilon\, y_{\min}$	Totals
Positions (high)	$w_v(p)\Sigma^{-1}\mu(y_{\max})$	$w_{hc}(p)\Sigma^{-1}\Upsilon$	$w_{hl}(p)\Sigma^{-1}\Upsilon\, y_{\max}$	
0.5	2	0.073	0.01	
	−4.4	−1.6	2	−3.9
	12	−1.6	−2	8.1
1	1	0	0	
	−2.2	0	0	−2.2
	5.8	0	0	5.8
2	0.5	−0.028	−0.0029	
	−1.1	0.6	−0.57	−1.1
	2.9	0.6	0.57	4.1
3	0.33	−0.033	−0.0031	
	−0.73	0.71	−0.61	−0.64
	1.9	0.71	0.61	3.3
5	0.2	−0.032	−0.0027	
	−0.44	0.68	−0.54	−0.29
	1.2	0.68	0.54	2.4
10	0.1	−0.024	−0.0019	
	−0.22	0.52	−0.37	−0.068
	0.58	0.52	0.37	1.5

In each subpanel, the first row contains the static fund weights, while the second and third rows report the weights in the risky asset, respectively, at the low (2.5%) and high (97.5%) quantiles of the state variable, under its stationary distribution. Market parameters are as in Table 3.2.

zero when the Sharpe ratio is low. In other words, the optimal portfolio is substantially different than myopic when current investment opportunities are good, but it is very close to myopic when they are poor.

In summary, the hedging component amplifies the response of investors to changes in the state variable, and (for typical levels of risk aversion) becomes more visible when investment opportunities are good.

4. SQUARE ROOT DIFFUSION

In this model (cf. Guasoni and Robertson 2012) a single state variable follows the square-root diffusion of Feller (1951), and simultaneously affects the interest rate, the volatilities

of risky assets, and the Sharpe ratios.

$$
\begin{aligned}
dR_t &= (\sigma \nu_0 + \sigma \nu_1 Y_t)\, dt + \sqrt{Y_t}\sigma\, dZ_t, \\
dY_t &= b(\theta - Y_t)\, dt + a\sqrt{Y_t} dW_t, \\
d\langle R, Y\rangle_t &= \rho dt, \\
r(Y_t) &= r_0 + r_1 Y_t.
\end{aligned} \tag{4.1}
$$

Here $\sigma \in \mathbb{R}^{n\times n}$; $\nu_0, \nu_1, \rho \in \mathbb{R}^n$ and b, θ, a, r_0, r_1 are all positive reals.

ASSUMPTION 4.1.

$$
b\theta - \frac{1}{2}a^2 > 0, \quad b\theta - qa\rho'\nu_0 > 0, \quad b + qa\rho'\nu_1 > 0. \tag{4.2}
$$

These parametric restrictions deserve some comment. $b\theta > \frac{1}{2}a^2$ ensures that Y remains strictly positive at all times, so that $E = (0, \infty)$ To understand (4.2), note that m in (2.19) takes the form:

$$
m(y) = K y^{\frac{2(b\theta - qa\rho'\nu_0)}{a^2} - 1} e^{-\frac{2(b+qa\rho'\nu_1)}{a^2} y} \quad y > 0, \tag{4.3}
$$

where $K > 0$ is an arbitrary constant. Thus, unless the second and third inequalities in (4.2) hold, there is no solution m with finite integral. The eigenvalue equation $-M\phi = \lambda\phi$ for M from (2.21) specifies to

$$
\frac{1}{2}\left(a^2 y \dot{\phi} m\right)^{\cdot} + \left(\frac{1}{\delta}\left(-\frac{1}{2}q\nu_0'\nu_0\frac{1}{y} + \left(pr_0 - q\nu_0'\nu_1\right) + \left(pr_1 - \frac{1}{2}q\nu_1'\nu_1\right) y\right) + \lambda\right)\phi m = 0. \tag{4.4}
$$

The following lemma identifies solutions of (4.4) with solutions of the generalized Laguerre differential equation, under the additional parameter restriction:

ASSUMPTION 4.2.

$$
p < 0. \tag{4.5}
$$

REMARK 4.3. For $0 < p < 1$ the results of this section still hold, but under very delicate parameter restrictions (similar to, but more involved than, those in (3.5) from Assumption 3.2). Since the proofs remain the same, only the case $p < 0$ is considered here for brevity.

LEMMA 4.4. *Let Assumptions 4.1 and 4.2 hold. Let $\phi \in L^2((0, \infty), m)$, and define ψ by the equality:*

$$
\phi(y)m(y)^{1/2} = \sqrt{\zeta}\,\psi(\zeta y)\hat{m}(\zeta y)^{1/2}, \tag{4.6}
$$

where

$$
\hat{m}(dz) = \frac{1}{\Gamma(\omega + 1)} z^{\omega} e^{-z} dz \tag{4.7}
$$

TABLE 4.1
Static Funds in the Square Root Model. $w_1(p)$ and $w_2(p)$ Are Given in (4.15)

Fund Name	Portfolio	Fund Weight
Myopic	$\Sigma^{-1}\mu(y)$	$w_v(p) = \frac{1}{1-p}$
Hedging constant	$\Sigma^{-1}\Upsilon$	$w_{hc}(p) = \frac{\delta}{1-p} w_1(p)$
Hedging harmonic	$\Sigma^{-1}\Upsilon \frac{1}{y}$	$w_{hh}(p) = \frac{\delta}{1-p} w_2(p)$

and the constants $\zeta > 0$, $\omega > 0$, $K_1 > 0$, K_2 and $\eta = K_1\lambda + K_2$ are defined in (C.1). Then ϕ solves (4.4) if and only if $\psi \in L^2((0,\infty), \hat{m})$ solves the ODE:

$$-z\ddot{\psi}(z) - (1 + \omega - z)\,\dot{\psi}(z) = \eta\psi(z). \tag{4.8}$$

For $n \geq 0$ consider the *generalized Laguerre polynomial*

$$\psi_n(z) = \sqrt{\frac{\Gamma(\omega+1)}{n!\Gamma(n+\omega+1)}} z^{-\omega} e^z \frac{d^n}{dz^n}\left(e^{-z} z^{n+\omega}\right). \tag{4.9}$$

It is well known (Magnus, Oberhettinger, and Soni 1966, chapter 5.5) that

$$\begin{aligned} &\psi_n \text{ solves (4.8) with } \eta_n = n; n \geq 0, \\ &(\psi_n)_{n\in\mathbb{N}_0} \text{ is an orthonormal set in } L^2((0,\infty), \hat{m}). \end{aligned} \tag{4.10}$$

Set

$$\varphi(z) = \left(\frac{1}{\zeta}\frac{m(z/\zeta)}{\hat{m}(z)}\right)^{1/2}. \tag{4.11}$$

By Assumption 4.1 $\varphi \in L^2((0,\infty), \hat{m})$. Because φ is smooth, it follows from Lebedev (1972, chapter 5, theorem 3) that the series:

$$\sum_{n=0}^{M} \alpha_n \psi_n; \qquad \alpha_n = \int_0^\infty \varphi(z)\psi_n(z)\hat{m}(z)\,dz = \int_0^\infty \phi_n(y) m(y)\,dy \tag{4.12}$$

converges to φ pointwise (i.e., for all $z > 0$) and in $L^2((0,\infty), \hat{m})$ as $M \uparrow \infty$. By the construction of φ and (4.12) the series $\sum_{n=0}^{M} \alpha_n\phi_n$ converges to 1 pointwise and in $L^2((0, \infty), m)$ as $M \uparrow \infty$, hence it is a candidate solution to the PDE (2.18). Unlike the linear model in Section 3, the convergence does not take place in the strong sense of (3.12). However, asymptotic growth estimates for the Laguerre polynomials ψ_n imply the following convergence result:

THEOREM 4.5. *Let Assumptions 4.1 and 4.2 hold. Define ψ_n as in (4.9), η_n as in (4.10), λ_n, ϕ_n as in Lemma 4.4, and α_n as in (4.12). Then*

(i) the function

$$v(t, y) = \sum_{n=0}^{\infty} e^{-\lambda_n(T-t)} \alpha_n \phi_n(y) \tag{4.13}$$

is a strictly positive $C^{1,2}((0, T) \times (0,\infty))$ solution of the PDE in (2.18);

TABLE 4.2
Parameter Values for the Square Root Model as in Pan (2002)

Parameter	Value
ν_0	0
ν_1	8.6
σ	1.00
b	7.1
θ	0.0137
a	0.32
ρ	−0.53
r_0	0.058
r_1	0

(ii) v satisfies the convergence property:

$$\lim_{T\uparrow\infty} \frac{v_y(t,y)}{v(t,y)} = \frac{\dot{\phi}_0(y)}{\phi_0(y)} \quad \textit{for all } t, y > 0; \tag{4.14}$$

(iii) the value function of the utility maximization problem (2.9) is equal to $V(x,t,y) = \frac{x^p}{p} v(t,y)^\delta$;

(iv) *the decomposition (2.26) (static fund separation) holds with* $m = 2$, $\psi_1(y) = 1$, *and* $\psi_2(y) = 1/y$, *with respective weights in Table 4.1, where:*

(4.15)

$$w_1(p) = \frac{1}{a^2}\left(b + qa\rho'\nu_1 - \sqrt{(b+qa\rho'\nu_1)^2 - \frac{2}{\delta}a^2\left(pr_1 - \frac{1}{2}q\nu_1'\nu_1\right)}\right),$$

$$w_2(p) = \frac{1}{a^2}\left(\sqrt{\left(b\theta - qa\rho'\nu_0 - \frac{1}{2}a^2\right)^2 + \frac{q}{\delta}a^2\nu_0'\nu_0} - \left(b\theta - qa\rho'\nu_0 - \frac{1}{2}a^2\right)\right).$$

Table 4.2 shows the parameter values estimated by Pan (2002, Table 1, Row "SV"). For these values, each of Assumptions 4.1, 4.2 holds for risk aversion within the range [1, 10] used in the plots and tables below. Figure 4.1 shows the reduced value function v^δ as a function of the state variable Y for various planning horizons and for risk aversion of 10 ($p = -9$). Figure 4.2 plots the optimal portfolios as a function of the state variable Y for a number of horizons, including the myopic ($T \downarrow 0$) and long run ($T \uparrow \infty$) limits when the risk aversion is 10. Figure 4.2 shows how the optimal portfolio shifts from the myopic to the long run limit, as the horizon increases. All plots use 15 eigenfunctions from the series representation.

Table 4.3 below gives the respective fund weights for various values of the risk aversion $1 - p$ using the parameter values in (4.2). Fund weights are given for high and low percentiles of the state variable.

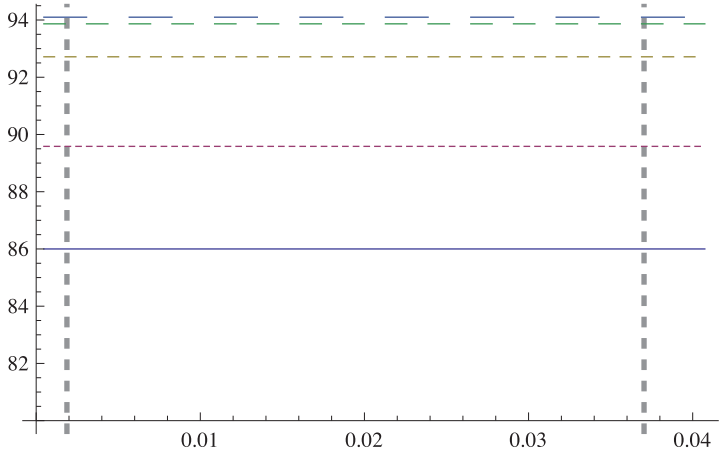

FIGURE 4.2. Risky portfolio weight (vertical axis), against state variable (horizontal axis). The planning horizon varies from $T \downarrow 0$ (solid), 1 month (tiny dashed), 3 months (short dashed), 6 months (medium dashed), and $T \uparrow \infty$ (long dashed). Portfolio weights are given in terms of percentages. Vertical lines are at the low (2.5%) and high (97.5%) percentiles of the state variable, under its stationary distribution. Risk aversion is 10 ($p = -9$), and market parameters are as in Table 4.2. All plots use 15 eigenfunctions from the series representation.

TABLE 4.3
Fund Weights and Implied Risky Positions for the Optimal Long Run Portfolios, for Different Risk Aversion Levels (1, 2, 3, 5, 10)

Risk Aversion $1-p$	Myopic Fund Weight $w_v(p)$	Hedging Constant Fund Weight $w_{hc}(p)$	
Risky Positions	$w_v(p)\Sigma^{-1}\mu$	$w_{hc}(p)\Sigma^{-1}\Upsilon$	Totals
1	1	0	
	8.6	0	8.6
2	0.5	−1.2	
	4.3	0.21	4.5
3	0.33	−1.1	
	2.9	0.19	3.1
5	0.2	−0.83	
	1.7	0.14	1.9
10	0.1	−0.48	
	0.86	0.081	0.94

In each subpanel, the first row contains the static fund weights, while the second row reports the weights in the risky asset. Market parameters are as in Table 4.2

5. CONCLUSION

If static fund separation holds, optimal portfolios for long-term CRRA investors are constant mixes of a few common funds. Then, hedging demands are explicit combinations of one or more hedging funds. The number of funds depends on the model, and may be larger than the number of state variables. Merton's (1973) dynamic fund separation implies that hedging funds are (locally) perfectly correlated with each other, but their covariance changes over time, as it depends on the state variable.

This paper establishes static fund separation for two common classes of models, in which the state variable is either an Ornstein–Uhlenbeck or a Feller diffusion. A central technique is the eigenvalue decomposition of value functions, made possible by the linear HJB equation, which follows from constant asset-state correlations, and a single state variable. Static fund separation in models with several state variables is an open area of research.

APPENDIX A: VERIFICATION

The following assumptions are on the coefficients μ, b, Σ, A, Υ given in (2.2)–(2.3). Note that these assumptions are satisfied by all the models considered within this paper. The first assumption requires the regularity and nondegeneracy of coefficients, while the second one guarantees that they identify the law of Y.

ASSUMPTION A.1. *There exists* $\alpha \in (0, 1]$ *such that* $r \in C(E, \mathbb{R})$, $b \in C^{1,\alpha}(E, \mathbb{R})$, Υ, $\mu \in C^{1,\alpha}(E, \mathbb{R}^n)$, $\Sigma \in C^{1,\alpha}(E, \mathbb{R}^{n\times n})$, *and* $A \in C^{2,\alpha}(E, \mathbb{R})$. $A > 0$ *and* Σ *is positive definite for all* $y \in E$, *uniformly on all compact subsets of* E.

$\Omega^d = C([0, \infty), \mathbb{R}^d)$ denotes the space of continuous paths from $[0, \infty)$ to $\mathbb{R}^d$, endowed with the Borel σ-algebra $\mathcal{F}$.

ASSUMPTION A.2. *For all* $y \in E$ *there exists a unique probability* P^y *on* $(\Omega^{n+1}, \mathcal{F})$ *such that* (R, Y) *satisfies (2.2)–(2.5) with initial condition* $(R_0, Y_0) = (0, y)$, *and* $P^y(Y_t \in E, \forall t \geq 0) = 1$. *The family* $(P^y)_{y\in E}$ *has the strong Markov property.*

The value function of the utility maximization problem is defined as the expected utility, conditional on the current state $y \in E$, the time $t \in [0, T]$, and the current wealth $x > 0$:

$$V(t, x, y) = \sup_{(\pi_u)_{u\geq t}} \frac{1}{p} E\big[(X_T^\pi)^p \mid Y_t = y, X_t = x\big].$$

The following lemma establishes conditions under which a solution to the HJB equation is indeed the value function.

LEMMA A.3. *Let Assumption A.1 hold. For* b^v, V *as in (2.16) and* L^v *as in (2.17) assume that*

(i) $v(t, y)$ *is a strictly positive* $C^{1,2}((0, T) \times E, \mathbb{R})$ *solution to*

$$\begin{aligned} v_t + L^v v + Vv = 0 \quad (t, y) \in (0, T) \times E, \\ v(T, y) = 1 \quad y \in E. \end{aligned} \tag{A.1}$$

(ii) both the original and the auxiliary models:

$$\text{(A.2)}\ (P)\begin{cases} dR_t = \mu dt + \sigma dZ_t \\ dY_t = bdt + adW_t \end{cases} \qquad (\hat{P}_T)\begin{cases} dR_t = \dfrac{1}{1-p}\left(\mu + \delta\Upsilon\dfrac{v_y}{v}\right)dt + \sigma d\hat{Z}_t \\ dY_t = \left(b^v + A\dfrac{v_y}{v}\right)dt + ad\hat{W}_t \end{cases}$$

satisfy Assumption A.2 on $[0, T]$ *under equivalent probabilities* P^y *and* $\hat{P}^y_T$, $y \in E$.

Then $V(t, x, y) = \frac{x^p}{p}(v(t, y))^\delta$. *The optimal trading strategy is given by*

$$\pi_t = \frac{1}{1-p}\Sigma^{-1}\left(\mu + \delta\Upsilon\frac{v_y}{v}\right) \tag{A.3}$$

evaluated at (t, Y_t).

Proof of Lemma A.3. For any function $\eta \in C((0, T) \times E, \mathbb{R})$ set

$$\begin{aligned} h(t, y) &= -\rho'\sigma^{-1}(y)\mu(y) + (1 - \rho'\rho)\, a(y)\eta(t, y), \\ g(t, y) &= -\bar{\rho}'\sigma^{-1}(y)\mu(y) - \bar{\rho}'\rho a(y)\eta(t, y), \end{aligned} \tag{A.4}$$

where $\bar{\rho}$ is the unique, symmetric, nonnegative definite square root of $1 - \rho\rho'$. Note that $1 - \rho\rho$ being nonnegative definite is equivalent to $1 - \rho'\rho \geq 0$. Using the decomposition $Z = \rho W + \bar{\rho} B$, where B is a Brownian Motion independent of W define the process:

$$Z^\eta_t = \mathcal{E}\left(\int_0^\cdot h(s, Y_s)'dW_s + \int_0^\cdot g(s, Y_s)'dB_s\right)_t \qquad 0 \leq t \leq T.$$

The continuity of η and Assumptions A.1 and A.2 ensure that Z^η is a strictly positive P^y local martingale on $[0, T]$ for any $T > 0$ and all $y \in E$. Furthermore, using stochastic integration by parts, it follows that for any trading strategy π the process:

$$e^{-\int_0^\cdot r\, ds} Z^\eta_\cdot X^\pi_\cdot$$

is a nonnegative P^y supermartingale. Set $M^\eta = e^{-\int_0^\cdot r\, ds} Z^\eta$. Using the well-known duality results relating terminal wealths from trading strategies and process which render trading strategies nonnegative supermartingales (see Kramkov and Schachermayer 1999), it suffices to show that under the assumption of well-posedness for $(\hat{P}^y_T)_{y \in E}$:

$$E\left[\frac{1}{p}\left(X^\pi_T\right)^p \mid X_t = x,\, Y_t = y\right] = \frac{x^p}{p}\left(E\left[\left(\frac{M^\eta_T}{M^\eta_t}\right)^q \,\middle|\, Y_t = y\right]\right)^{1-p}$$

for v solving (A.1), π as in (A.3) and $\eta = \delta\frac{v_y}{v}$. Plugging in for X^π_T, M^η_T / M^η_t, this is equivalent to showing

(A.5)

$$\begin{aligned} &E\left[\exp\left(p\int_t^T\left(r + \pi'\mu - \frac{1}{2}\pi'\Sigma\pi\right)ds + \int_t^T p\pi'\sigma\rho dW_s + \int_t^T p\pi'\sigma\bar{\rho}dB_s\right) \,\middle|\, Y_t = y\right] \\ &= \left(E\left[\exp\left(q\int_t^T\left(-r - \frac{1}{2}h'h - \frac{1}{2}g'g\right)ds + \int_t^T qh'dW_s + \int_t^T qg'dB_s\right) \,\middle|\, Y_t = y\right]\right)^{1-p}. \end{aligned}$$

Under Assumptions A.2 and A.1, the smoothness of v and the well-posedness under P^y and $\hat{P}^y_T$ imply that $\hat{Z}_t = d\hat{P}_T/dP\big|_{\mathcal{F}_t}$ is a martingale (see Cheridito, Filipović, and Yor 2005, theorem 2.4). $\hat{Z}$ takes the form:

(A.6)
$$\hat{Z}_t = \mathcal{E}\left(\int_0^{\cdot}\left(-q\Upsilon'\Sigma^{-1}\mu + A\frac{v_y}{v}\right)'a^{-1}dW_s - q\int_0^{\cdot}\left(\Sigma^{-1}\mu + \delta\Sigma^{-1}\Upsilon\frac{v_y}{v}\right)'\sigma\bar{\rho}dB_s\right)_t.$$

It is convenient to define $w(t, y) = \delta\log v(t, y)$. w solves the semilinear PDE:

(A.7)
$$w_t + \frac{1}{2}Aw_{yy} + \frac{1}{2\delta}Aw_y^2 + b^v w_y + \delta V = 0 \qquad (t, y) \in (0, T)\times E,$$
$$w(T, y) = 0 \qquad y \in E.$$

With this notation, $\pi = \frac{1}{1-p}\Sigma^{-1}\left(\mu + \Upsilon w_y\right)$ and $\eta = w_y$. It suffices to show that, for $Y_t = y$:

(A.8)
$$\exp\left(p\int_t^T\left(r + \pi'\mu - \frac{1}{2}\pi'\Sigma\pi\right)ds + \int_t^T p\pi'\sigma\rho dW_s + \int_t^T p\pi'\sigma\bar{\rho}dB_s\right) = \frac{\hat{Z}_T}{\hat{Z}_t}e^{w(t,y)},$$
$$\exp\left(-q\int_t^T\left(r + \frac{1}{2}h'h + \frac{1}{2}g'g\right)ds + \int_t^T qh'dW_s + \int_t^T qg'dB_s\right) = \frac{\hat{Z}_T}{\hat{Z}_t}e^{\frac{1}{1-p}w(t,y)}$$

from which (A.5) follows. Under P^y, using Itô's formula and (A.7) it follows that given $Y_t = y$:

$$\begin{aligned}-w(t, y) &= \int_t^T w_t ds + \int_t^T bw_y ds + \int_t^T aw_y dW_s + \frac{1}{2}\int_t^T Aw_{yy}ds\\ &= -\int_t^T\left(\frac{1}{2\delta}Aw_y^2 - qw_y\Upsilon'\Sigma^{-1}\mu + pr - \frac{q}{2}\mu'\Sigma^{-1}\mu\right)ds + \int_t^T w_y a dW_s\end{aligned}$$

plugging back in for b^v and V. At this point, verifying (A.8) amounts to plugging in for $\hat{Z}_T/\hat{Z}_t$, π, η (and hence h, g via (A.4)) and w and then matching up the ds, dW, and dB terms. Performing these steps readily yields the conclusion (recalling that $A/\delta = A - q\Upsilon'\Sigma^{-1}\Upsilon$). □

APPENDIX B: PROOFS OF SECTION 3

The following constants are used in Lemma 3.4 and Theorem 3.5:

$$\mathbf{A} = 2qb\left(\rho'\nu_0(1 + q\rho'\nu_1) + \frac{1}{\delta}\nu_0'\nu_1\right),$$
$$\mathbf{B} = b^2\left((1 + q\rho'\nu_1)^2 + \frac{q}{\delta}\nu_1'\nu_1\right),$$
$$\mathbf{C} = -q^2(\rho'\nu_0)^2 + b(1 + q\rho'\nu_1) - \frac{q}{\delta}\nu_0'\nu_0 + \frac{2pr_0}{\delta},$$

$$\alpha = \mathbf{B}^{1/4}, \quad \beta = \frac{\mathbf{A}}{2\mathbf{B}^{3/4}}, \tag{B.1}$$
$$K_1 = \frac{2}{\sqrt{\mathbf{B}}}, \quad K_2 = \frac{1}{\sqrt{\mathbf{B}}}\left(\frac{\mathbf{A}^2}{4\mathbf{B}} + \mathbf{C}\right).$$

Note that (3.5) in Assumption 3.2 is equivalent to $\mathbf{B} > 0$.

The proof of Lemma 3.4 and Proposition 3.5 require the following two lemmas:

LEMMA B.1. *Let Assumptions 3.1 and 3.2 hold. Let m be as in (3.3), α, β, K_1, and K_2 be as in (B.1), $\lambda \in \mathbb{R}$ and $\eta = K_1\lambda + K_2$. Let*

$$V(y) = \frac{1}{\delta}\left(pr_0 - \frac{q}{2}\nu_0'\nu_0 - qb\nu_0'\nu_1 y - \frac{q}{2}b^2\nu_1'\nu_1 y^2\right). \tag{B.2}$$

If $f, g \in C^{1,2}((0, T) \times \mathbb{R})$ are related by

$$f(t, y) = \sqrt{\alpha}m(y)^{-1/2}g(t, z) \Longleftrightarrow f(t, y) = \varphi(z)^{-1}g(t, z)$$

for $z = \alpha y + \beta$ then at $z = \alpha y + \beta$

$$\begin{aligned}\frac{1}{2}\left(\dot{f}(t, y)m(y)\right)^{\cdot} + V(y)f(t, y)m(y) + \lambda f(t, y)m(y) + f_t(t, y)m(y) \\ = -\frac{1}{2}\alpha^{5/2}m\left(\frac{z-\beta}{\alpha}\right)^{1/2}\left(-\ddot{g}(z) + z^2 g(z) - \eta g(z) - \frac{2}{\alpha^2}g_t(t, z)\right).\end{aligned} \tag{B.3}$$

Proof of Lemma B.1. That $f_t(t, y)m(y) = \sqrt{\alpha}m((z-\beta)/\alpha)^{1/2}g_t(t, z)$ is immediate. For the remaining terms, it suffices to think of f, g as functions of y (resp. z) alone. Define $\Xi(y)$ by

$$\Xi(y) = \frac{1}{2}\left(\dot{f}(y)m(y)\right)^{\cdot} + V(y)f(y)m(y) + \lambda f(y)m(y). \tag{B.4}$$

Let $h(y) = m^{1/2}(y)f(y)$. It follows that

$$\dot{f} = m^{-1/2}\left(\dot{h} - \frac{\dot{m}}{2m}h\right),$$
$$\ddot{f} = m^{-1/2}\left(\ddot{h} - 2\frac{\dot{m}}{2m}\dot{h} + \left(\left(\frac{\dot{m}}{2m}\right)^2 - \left(\frac{\dot{m}}{2m}\right)^{\cdot}\right)h\right).$$

Therefore, from (B.4) it follows that

$$\begin{aligned}\Xi &= \frac{1}{2}\left(\dot{f}m\right)^{\cdot} + (V + \lambda)fm \\ &= m^{1/2}\left(\frac{1}{2}\ddot{h} - \frac{\dot{m}}{2m}\dot{h} + \frac{1}{2}\left(\left(\frac{\dot{m}}{2m}\right)^2 - \left(\frac{\dot{m}}{2m}\right)^{\cdot}\right)h + \frac{\dot{m}}{2m}\dot{h} - \left(\frac{\dot{m}}{2m}\right)^2 h + (V + \lambda)h\right) \\ &= m^{1/2}\left(\frac{1}{2}\ddot{h} + \left(V + \lambda - \frac{1}{2}\left(\frac{\dot{m}}{2m}\right)^2 - \frac{1}{2}\left(\frac{\dot{m}}{2m}\right)^{\cdot}\right)h\right).\end{aligned}$$

Plugging in for V from (B.2) and in for $\frac{\dot{m}}{m}$ from (3.3) and multiplying by -2 yields

$$-2\Xi = m^{1/2}\left(-\ddot{h} + \left(\mathbf{B}y^2 + \mathbf{A}y - \mathbf{C} - 2\lambda\right)h\right) \tag{B.5}$$

for $\mathbf{A}, \mathbf{B}, \mathbf{C}$ as in (B.1). An affine transformation in the state variable normalizes the quadratic constant $\mathbf{B}$ and eliminates the linear constant $\mathbf{A}$. Let $z = \alpha y + \beta$. By construction, $g(z) = \frac{1}{\sqrt{\alpha}} h(y) = \frac{1}{\sqrt{\alpha}} h(\frac{z-\beta}{\alpha})$. Evaluating (B.5) at $y = \frac{z-\beta}{\alpha}$ yields

$$-2\Xi\left(\frac{z-\beta}{\alpha}\right) = \alpha^{5/2} m\left(\frac{z-\beta}{\alpha}\right)^{1/2} \left(-\ddot{g}(z) + \left(\frac{\mathbf{B}}{\alpha^4} z^2 + \left(\frac{\mathbf{A}}{\alpha^3} - \frac{2\beta\mathbf{B}}{\alpha^4}\right) z + \left(\frac{\beta^2\mathbf{B}}{\alpha^4} - \frac{\beta\mathbf{A}}{\alpha^3} - \frac{\mathbf{C}}{\alpha^2} - \frac{2\lambda}{\alpha^2}\right)\right) g(z)\right). \tag{B.6}$$

Using the representations for $\alpha, \beta, \mathbf{A}, \mathbf{B}, \mathbf{C}, K_1$, and K_2 from (B.1) and η from the statement of the lemma it follows that

$$-2\Xi\left(\frac{z-\beta}{\alpha}\right) = \alpha^{5/2} m\left(\frac{z-\beta}{\alpha}\right)^{1/2} \left(-\ddot{g}(z) + z^2 g(z) - \eta g(z)\right), \tag{B.7}$$

which is the desired result. □

LEMMA B.2. *Let $(h_n)_{n\geq 1}$ be a sequence of real numbers and let $(\gamma_n)_{n\geq 1}$ be a decreasing sequence of positive real numbers. Then, for all $m, M \in \mathbb{N}$, $m \leq M$*

$$\left|\sum_{n=m}^{M} \gamma_n h_n\right| \leq \gamma_m \max_{N=m,\dots M} \left|\sum_{n=m}^{N} h_n\right|.$$

Proof of Lemma B.2. For the γ_n of the statement

$$\begin{aligned}
\left|\sum_{n=m}^{M} \gamma_n h_n\right| &= \left|\gamma_M \sum_{n=m}^{M} h_n + (\gamma_{M-1} - \gamma_M) \sum_{n=m}^{M-1} h_n + \cdots + (\gamma_m - \gamma_{m+1}) h_m\right| \\
&\leq \gamma_M \left|\sum_{n=m}^{M} h_n\right| + (\gamma_{M-1} - \gamma_M) \left|\sum_{n=m}^{M-1} h_n\right| + \cdots + (\gamma_m - \gamma_{m+1}) |h_m| \\
&\leq \max_{N=m,\dots M} \left|\sum_{n=m}^{M} h_n\right| (\gamma_M + (\gamma_{M-1} - \gamma_M) + \cdots + (\gamma_m - \gamma_{m+1})) \\
&= \gamma_m \max_{N=m,\dots M} \left|\sum_{n=m}^{M} h_n\right|.
\end{aligned}$$

□

Proof of Lemma 3.4. By applying Lemma B.1 for $f(y) = \phi(y)$, $g(z) = \psi(z)$ (note there is no dependence upon t) it follows that ϕ satisfies (3.4) for some $\lambda \in \mathbb{R}$ if and only if ψ satisfies (3.7) for $\eta = K_1\lambda + K_2$. As for the respective L^2 norms, since

$$\int_{\mathbb{R}} \psi(z)^2 dz = \frac{1}{\alpha} \int_{\mathbb{R}} \phi\left(\frac{z-\beta}{\alpha}\right)^2 m\left(\frac{z-\beta}{\alpha}\right) dz = \int_{\mathbb{R}} \phi(y)^2 m(y)\, dy,$$

$\psi \in L^2(\mathbb{R})$ if and only if $\phi \in L^2(\mathbb{R}, m)$ and they both have the same norm. □

Proof of Theorem 3.5. It is first shown that v from (3.13) is a strictly positive $C^{1,2}((0, T) \times \mathbb{R})$ solution to the PDE in (2.18), or equivalently, to the PDE in (2.20),

specified to the model in (3.1). We have, using (3.6), (3.9), (3.10), and (B.1) that for any $y \in \mathbb{R}$, at $z = \alpha y + \beta$

$$v(t, y) = \sum_{n=0}^{\infty} e^{-\lambda_n (T-t)} \alpha_n \phi_n(y) = e^{K_2/K_1 (T-t)} \varphi(z)^{-1} \sum_{n=0}^{\infty} e^{-\eta_n/K_1 (T-t)} \alpha_n \psi_n(z). \tag{B.8}$$

Set

$$w(t, z) = \sum_{n=0}^{\infty} e^{-\eta_n/K_1 (T-t)} \alpha_n \psi_n(z) \tag{B.9}$$

so that (A.8) becomes

$$v(t, y) = e^{K_2/K_1 (T-t)} \varphi(z)^{-1} w(t, z). \tag{B.10}$$

By applying Lemma B.1 for

$$\lambda = 0, \quad f(t, y) = v(t, y), \quad g(t, z) = w(t, z) e^{K_2/K_1 (T-t)}$$

it follows that $v(t, y)$ solves (2.20) if and only if $w(t, z)$ solves (note $\alpha^2/2 = 1/K_1$)

$$\begin{aligned} K_1 w_t + w_{zz} - z^2 w = 0 \quad (t, z) \in (0, T) \times \mathbb{R}, \\ w(T, z) = \varphi(z) \quad z \in \mathbb{R}. \end{aligned} \tag{B.11}$$

Set

$$\tau_n(t) = e^{-\eta_n/K_1 (T-t)} = e^{-(2n+1)/K_1 (T-t)}. \tag{B.12}$$

In the light of (3.9), for each integer M the function

$$w_M(t, z) \equiv \sum_{n=0}^{M} \alpha_n \tau_n(t) \psi_n(z) \tag{B.13}$$

solves the differential expression in (B.11) with boundary condition $w_M(T, z) = \sum_{n=0}^{M} \alpha_n \psi_n(z)$. Thus, it suffices to prove that the function $w(t, z) = \lim_{M \uparrow \infty} w_M(t, z)$ is a well-defined $C^{1,2}((0, T) \times \mathbb{R})$ function, which is strictly positive and which solves the PDE in (A.11).

Since φ is a function of rapid decrease, taking $l = 0$, $m = 0$ in (3.12) shows that $w_M(T, z)$ converges uniformly to $\varphi(z)$ on $\mathbb{R}$. To show that $w(t, z) = \lim_{M \uparrow \infty} w_M(t, z)$ exists on $(0, T) \times \mathbb{R}$ and solves the full PDE in (B.11), it is enough to prove any integer l that $\frac{d^l}{dz^l} w_M(t, z)$ as well as $\partial_t w_M(t, z)$ are uniformly convergent as $M \uparrow \infty$ in $(0, T) \times \mathbb{R}$. Regarding the spatial derivatives for $l = 1$ it is clear that for each $t \leq T$, $\gamma_n = \tau_n(t)$ satisfies the hypothesis of Lemma B.2 with $\tau_n(t) \leq e^{-T/K_1}$ for all n, $0 \leq t \leq T$. Since $w_M(T, z)$ converges uniformly in $\mathbb{R}$ for any $\varepsilon > 0$ there is an M_ε such that $m, M > M_\varepsilon$ implies that

$$\sup_{z \in \mathbb{R}} \left| \sum_{n=m}^{M} \alpha_n \psi_n(z) \right| < \varepsilon. \tag{B.14}$$

Hence, by Lemma B.2 for any $0 \leq t \leq T$

$$\sup_{z \in \mathbb{R}} \left| \sum_{n=m}^{M} \alpha_n \tau_n(t) \psi_n(z) \right| \leq e^{-T/K_1} \varepsilon$$

and thus $w_M(t, y)$ converges uniformly on $(0, T) \times \mathbb{R}$. The proof for $\frac{d^l}{dz^l} w_M(t, z)$ for any l is analogous, using the convergence in the space of functions of rapid decrease at $t = T$. As for $\partial_t w_M$ note that

$$\partial_t w_M = \sum_{n=0}^{M} \frac{\eta_n}{K_1} \tau_n(t) \alpha_n \psi_n(z) = \frac{1}{K_1} \sum_{n=0}^{M} \tau_n(t) \alpha_n \left(-\ddot{\psi}_n(z) + z^2 \psi_n(z) \right)$$

since $\eta_n \psi_n = -\ddot{\psi}_n + z^2 \psi_n$. Thus, the uniform convergence of $\partial_t w_M$ follows from that of $\dot{w}_M$ and $z^2 \ddot{w}_M$. This proves that w satisfies the PDE in (B.11). Note that (B.11) can be written

$$\begin{aligned} w_t + L w_{zz} - \tfrac{z^2}{K_1} w = 0 \quad (t, z) \in (0, T) \times \mathbb{R}, \\ w(T, z) = \varphi(z) \quad z \in \mathbb{R} \end{aligned} \tag{B.15}$$

for the operator $L = \frac{1}{2} \frac{2}{K_1} d^2/dz^2$. Thus, using the fact that w solves (B.11) and that φ is a function of rapid decrease it clearly follows that $w(t, z)$ admits the stochastic representation:

$$w(t, z) = E_z^{\mathbb{Q}} \left[\varphi(Z_{T-t}) \exp\left(-\frac{1}{K_1} \int_0^{T-t} Z_s^2 ds \right) \right], \tag{B.16}$$

where $(\mathbb{Q}_z)_{z \in \mathbb{R}}$ is a solution to the martingale problem for L on $\mathbb{R}$. Such a solution clearly exists since $Z = \sqrt{2/K_1}\, W$, where W is a standard Brownian Motion. The strict positivity of $w(t, z)$ easily follows from (A.16) since $\varphi > 0$.

It is next shown that v satisfies the convergence relation in (3.14). By (B.10) it follows that at $z = \alpha y + \beta$:

$$\frac{v_y(t, y)}{v(t, y)} = \alpha \left(\frac{w_z(t, z)}{w(t, z)} - \frac{\dot{\varphi}(z)}{\varphi(z)} \right). \tag{B.17}$$

By (3.6) and (3.10) it follows that at $z = \alpha y + \beta$, $\phi_0(y) = \varphi(z)^{-1} \psi_0(z)$ and hence

$$\frac{v_y(t, y)}{v(t, y)} - \frac{\dot{\phi}_0(y)}{\phi_0(y)} = \zeta \left(\frac{w_z(t, z)}{w(t, z)} - \frac{\dot{\psi}_0(z)}{\psi_0(z)} \right)$$

and thus (3.14) will follow if

$$\lim_{T \uparrow \infty} \frac{w_z(t, z)}{w(t, z)} = \frac{\dot{\psi}_0(z)}{\psi_0(z)} \tag{B.18}$$

holds for all $(t, z) \in (0, \infty) \times \mathbb{R}$. Using the uniform convergence of $w_M, \dot{w}_M$ it follows that

$$\lim_{T\uparrow\infty}\frac{w_z(t,z)}{w(t,z)} = \lim_{T\uparrow\infty}\frac{\sum_{n=0}^{\infty}\tau_n(t)\alpha_n\dot{\psi}_n(z)}{\sum_{n=0}^{\infty}\tau_n(t)\alpha_n\psi_n(z)}$$

$$= \lim_{T\uparrow\infty}\frac{\alpha_0\dot{\psi}_0(z)+\sum_{n=1}^{\infty}(\tau_n(t)/\tau_0(t))\alpha_n\dot{\psi}_n(z)}{\alpha_0\psi_0(z)+\sum_{n=1}^{\infty}(\tau_n(t)/\tau_0(t))\alpha_n\psi_n(z)}.$$

Since $\tau_n(t)/\tau_0(t) = e^{-2n/K_1(T-t)}$ it suffices to show

$$\lim_{T\uparrow\infty}\lim_{M\uparrow\infty}\sum_{n=1}^{M}e^{-2n/K_1(T-t)}\alpha_n\dot{\psi}_n(z)=0 \quad \lim_{T\uparrow\infty}\lim_{M\uparrow\infty}\sum_{n=1}^{\infty}e^{-2n/K_1(T-t)}\alpha_n\psi_n(z)=0.$$

It will be shown for the partial sums of ψ_n, the proof for the partial sums of $\dot{\psi}_n$ being similar. Fix $t > 0$ and $z \in \mathbb{R}$. Let $\tilde{T}$ be such that $t < \tilde{T} < T$. Since $w_M(t, z)$ converges uniformly in z for $\tilde{T}$ for any $\varepsilon > 0$ there is an M_ε such that for $m, M > M_\varepsilon$

$$\left|\sum_{n=m}^{M}e^{-\eta_n/K_1(\tilde{T}-t)}\alpha_n\psi_n(z)\right| < \varepsilon.$$

By making ε smaller it can be assumed that, plugging in $\eta_n = 2n+1$,

$$\left|\sum_{n=m}^{M}e^{-2n/K_1(\tilde{T}-t)}\alpha_n\psi_n(z)\right| < \varepsilon.$$

It thus follows using Lemma B.2 with $\gamma_n = e^{-2n/K_1(T-\tilde{T})}$ and $h_n = e^{-2n/K_1(\tilde{T}-t)}\alpha_n\psi_n(z)$ that

$$\lim_{T\uparrow\infty}\lim_{M\uparrow\infty}\left|\sum_{n=1}^{M}e^{-2n/K_1(T-t)}\alpha_n\psi_n(z)\right| \le \lim_{T\uparrow\infty}\lim_{M\uparrow\infty}\left|\sum_{n=M_\varepsilon+1}^{M}e^{-2n/K_1(T-t)}\alpha_n\psi_n(z)\right|$$

$$= \lim_{T\uparrow\infty}\lim_{M\uparrow\infty}\left|\sum_{n=M_\varepsilon+1}^{M}e^{-2n/K_1(T-\tilde{T})}e^{-2n/K_1(\tilde{T}-t)}\alpha_n\psi_n(z)\right|$$

$$\le \lim_{T\uparrow\infty}\lim_{M\uparrow\infty}e^{-2(M_\varepsilon+1)/K_1(T-\tilde{T})}\max_{N=M_\varepsilon+1,\ldots,M}\left|\sum_{n=N}^{M}e^{-2n/K_1(\tilde{T}-t)}\alpha_n\psi_n(z)\right|$$

$$\le \lim_{T\uparrow\infty}\varepsilon e^{-2(M_\varepsilon+1)/K_1(T-\tilde{T})}=0.$$

It is now shown using Lemma A.3 that the value function $V(x, t, y)$ is equal to $\frac{x^p}{p}v(t, y)^\delta$. It has already been shown that v satisfies the PDE in (A.1) specified to (3.1). Thus, it only remains to prove that the model $\hat{P}_T$ satisfies Assumption A.2. Specified to the model in (3.1), $\hat{P}_T$ takes the form:

$$dR_t = \frac{1}{1-p}\left(\sigma\nu_0 + b\sigma\nu_1 Y_t + \delta\sigma\rho\frac{v_y}{v}(t, Y_t)\right)dt + \sigma dZ_t,$$
$$dY_t = \left(-q\rho'\nu_0 - b\left(1+q\rho'\nu_1\right)Y_t + \frac{v_y}{v}(t, Y_t)\right)dt + dW_t.$$

Notice that it is enough to prove there is a weak solution for the SDE involving Y. Indeed, if this is the case, there is a weak solution for Y, B where B is an n-dimensional Brownian Motion independent of Y. Then, setting $Z = \rho W + \bar{\rho} B$ and defining R accordingly will result in a weak solution for R, Y.

Clearly, for each $z \in \mathbb{R}$, there is a weak solution to the SDE:

$$dZ_t = \sqrt{\frac{2}{K_1}}dW_t \quad Z_0 = z.$$

Indeed, the measures $(\mathbb{Q}_z)_{z\in\mathbb{R}}$ from (B.16) provide such a solution. Using (A.16) it follows that

$$\frac{1}{w(0,z)}E_z^{\mathbb{Q}}\left[\varphi(Z_T)\exp\left(-\frac{1}{K_1}\int_0^T Z_t^2 dt\right)\right] = 1.$$

But, Itô's formula implies that under $\mathbb{Q}_z$

$$\frac{1}{w(0,z)}\varphi(Z_T)\exp\left(-\frac{1}{K_1}\int_0^T Z_t^2 dt\right) = \mathcal{E}\left(\int_0^{\cdot}\sqrt{\frac{2}{K_1}}\frac{w_z}{w}(t, Z_t)dW_t\right)_T$$

and hence by Girsanov's theorem there is a weak solution to the SDE

$$\begin{aligned} dZ_t &= \frac{2}{K_1}\frac{w_z}{w}(t, Z_t)\,dt + \sqrt{\frac{2}{K_1}}dW_t \\ &= \alpha^2\frac{w_z}{w}(t, Z_t)\,dt + \alpha dW_t \end{aligned}$$

using the definition of α. Setting $Y = \frac{1}{\alpha}(Z-\beta)$ and using (B.17), (3.10), and (3.3) gives

$$\begin{aligned} dY_t &= \alpha\frac{w_z}{w}(t, \alpha Y_t + \beta)\,dt + dW_t \\ &= \left(\frac{v_y}{v}(t, Y_t) + \alpha\frac{\dot{\varphi}}{\varphi}(Y_t)\right)dt + dW_t \\ &= \left(-q\rho'\nu_0 - b(1+q\rho'\nu_1)Y_t + \frac{v_y}{v}(t, Y_t)\right)dt + dW_t. \end{aligned}$$

Therefore, the model $\hat{P}_T$ is well posed.

Finally, the statements regarding (3.15) follow immediately from the fact that $\psi_0(z) = \pi^{-1/4}e^{-z^2/2}$ and that $\phi_0(y) = \sqrt{\alpha}m(y)^{-1/2}\psi_0(\alpha y + \beta)$.

It remains to show the static fund separation given in (3.15). By (3.6) and the fact that $\psi_0(z) = \pi^{-1/4}e^{-z^2/2}$ we have

$$\frac{\dot{\phi}_0(y)}{\phi_0(y)} = \left(q\rho'\nu_0 - \alpha\beta\right)\times 1 + \left(b(1+q\rho'\nu_1) - \alpha^2\beta\right)\times y$$

from which (3.15) follows by plugging in for the constants in (B.1). □

APPENDIX C: PROOFS OF SECTION 4

The following constants are used in the proof of Lemma 4.4 and Theorem 4.5:

$$
\begin{aligned}
&\alpha = b\theta - qa\rho'\nu_0 - \frac{1}{2}a^2, \quad \beta = b + qa\rho'\nu_1,\\
&\Lambda = \alpha^2 + \frac{qa^2}{\delta}\nu_0'\nu_0, \quad \Theta = \beta^2 - \frac{2a^2}{\delta}\left(pr_1 - \frac{1}{2}q\nu_1'\nu_1\right),\\
&\omega = \frac{2\sqrt{\Lambda}}{a^2}, \quad \zeta = \frac{2\sqrt{\Theta}}{a^2}, \quad K_1 = \frac{1}{\sqrt{\Theta}},\\
&K_2 = \frac{a^2}{2\sqrt{\Theta}}\left(\frac{\beta - \sqrt{\Theta}}{a^2} + \frac{2\left(\alpha\beta - \sqrt{\Lambda\Theta}\right)}{a^4} + \frac{2}{a^2\delta}\left(pr_0 - q\nu_0'\nu_1\right)\right).
\end{aligned}
\tag{C.1}
$$

The proof of Lemma 4.4 and Theorem 4.5 requires the following three lemmas.

LEMMA C.1. *Let Assumption 4.1 hold. Let m be as in (4.3), and $\hat{m}$ as in (4.7). Let ω, ζ, K_1, and K_2 be as in (C.1). Let $\lambda \in \mathbb{R}$ and $\eta = K_1\,\lambda + K_2$. Let*

$$
V(y) = \frac{1}{\delta}\left(pr_0 + pr_1 y - \frac{1}{2}q\nu_0'\nu_0\frac{1}{y} - q\nu_0'\nu_1 - \frac{1}{2}q\nu_1'\nu_1 y\right). \tag{C.2}
$$

If $f, g \in C^{1,2}((0,T)\times(0,\infty))$ are related by (φ is defined by (4.11))

$$
f(t,y)m(y)^{1/2} = \sqrt{\zeta}\,g(t,z)\hat{m}(z)^{1/2} \Longleftrightarrow f(t,y) = g(t,z)\varphi(z)^{-1}
$$

then, setting $z = \zeta y$,

(C.3)

$$
\begin{aligned}
&\frac{1}{2}\left(a^2 y\dot{f}(t,y)m(y)\right)^{\cdot} + V(y)f(t,y)m(y) + \lambda f(t,y)m(y) + f_t(t,y)m(y)\\
&\quad = \frac{1}{2}a^2\zeta^{3/2}\sqrt{\hat{m}(z)m(y)}\left(z\ddot{g}(t,z) + (1+\omega-z)\dot{g}(t,z) + \eta g(t,z) + \frac{1}{\sqrt{\Theta}}g_t(t,z)\right).
\end{aligned}
$$

Proof of Lemma C.1. It suffices to compare the terms in the left- and the right-hand sides. That $f_t(t,y)m(y) = \frac{1}{2}a^2\zeta^{3/2}(1/\sqrt{\Theta})\sqrt{m(y)\hat{m}(z)}g_t(t,z)$ follows from $\sqrt{\Theta} = (a^2/2)\zeta$. For the remaining terms, define $\Xi(y)$ by

$$
\Xi(y) = \frac{1}{2}\left(a^2 y\dot{f}(y)m(y)\right)^{\cdot} + V(y)f(y)m(y) + \lambda f(y)m(y). \tag{C.4}
$$

From (C.1) and (4.3) it follows that $\frac{\dot{m}}{2m} = \frac{\alpha}{a^2 y} - \frac{\beta}{a^2}$. Plugging this equality in yields (to ease notation, y is suppressed from the right-hand side of the following equations)

$$
\begin{aligned}
\Xi(y) &= m\left(\frac{1}{2}a^2\dot{f} + \frac{1}{2}a^2 y\ddot{f} + a^2 y\dot{f}\frac{\dot{m}}{2m} + (V+\lambda)f\right)\\
&= m\left(\frac{1}{2}a^2 y\ddot{f} + \left(\alpha + \frac{1}{2}a^2 - \beta y\right)\dot{f} + (V+\lambda)f\right).
\end{aligned}
\tag{C.5}
$$

Let $h(y) = \sqrt{m(y)/\hat{m}(\zeta y)}f(y) = (1/K)y^{-\mathbf{A}}e^{-\mathbf{B}y}f(y)$, where ζ is from (C.1), K is a normalizing constant and

$$\mathbf{A} = 1/a^2\left(\sqrt{\Lambda} - \alpha\right) \quad \mathbf{B} = 1/a^2\left(-\sqrt{\Theta} + \beta\right). \tag{C.6}$$

It follows that

$$\dot{f} = Ky^{\mathbf{A}}e^{\mathbf{B}y}\left(\frac{\mathbf{A}}{y}h + \mathbf{B}h + \dot{h}\right),$$
$$\ddot{f} = Ky^{\mathbf{A}}e^{\mathbf{B}y}\left(\frac{\mathbf{A}(\mathbf{A}-1) + 2\mathbf{A}\mathbf{B}y + \mathbf{B}^2y^2}{y^2}h + \frac{2\mathbf{A} + 2\mathbf{B}y}{y}\dot{h} + \ddot{h}\right).$$

Plugging these relations into (C.5), the expressions for V in (C.2), and collecting terms, yields

$$\begin{aligned}\Xi(y) &= \frac{1}{2}a^2mKy^{\mathbf{A}}e^{\mathbf{B}y}\left(y\ddot{h} + \left(\left(2\mathbf{A} + \frac{2\alpha}{a^2} + 1\right) + \left(2\mathbf{B} - \frac{2\beta}{a^2}\right)y\right)\dot{h}\right.\\ &\qquad\left. + \left(\frac{\hat{\mathbf{A}}}{y} + \hat{\mathbf{B}}y + \hat{\mathbf{C}} + \frac{2\lambda}{a^2}\right)h\right)\\ &= \frac{1}{2}a^2\sqrt{\hat{m}(\zeta y)m(y)}\left(y\ddot{h} + \left(\left(2\mathbf{A} + \frac{2\alpha}{a^2} + 1\right) + \left(2\mathbf{B} - \frac{2\beta}{a^2}\right)y\right)\dot{h}\right.\\ &\qquad\left. + \left(\frac{\hat{\mathbf{A}}}{y} + \hat{\mathbf{B}}y + \hat{\mathbf{C}} + \frac{2\lambda}{a^2}\right)h\right)\end{aligned} \tag{C.7}$$

for

$$\hat{\mathbf{A}} = \mathbf{A}^2 + \frac{2\alpha}{a^2}\mathbf{A} - \frac{q}{\delta a^2}\nu_0'\nu_0,$$
$$\hat{\mathbf{B}} = \mathbf{B}^2 - \frac{2\beta}{a^2}\mathbf{B} + \frac{2}{\delta a^2}\left(pr_1 - \frac{q}{2}\nu_1'\nu_1\right),$$
$$\hat{\mathbf{C}} = \mathbf{B} + 2\mathbf{A}\mathbf{B} + \frac{2\alpha\mathbf{B} - 2\beta\mathbf{A}}{a^2} + \frac{2}{a^2\delta}(pr_0 - q\nu_0'\nu_1).$$

By (C.6), it follows that $\hat{\mathbf{A}} = \hat{\mathbf{B}} = 0$ and

$$\hat{\mathbf{C}} = \frac{\beta - \sqrt{\Theta}}{a^2} + \frac{2(\alpha\beta - \sqrt{\Lambda\Theta})}{a^4} + \frac{2}{a^2\delta}(pr_0 - q\nu_0'\nu_1).$$

Plugging this equality into (C.7) and using the expressions for K_1, K_2 in (C.1) yields

$$\Xi(y) = \frac{1}{2}a^2\sqrt{\hat{m}(\zeta y)m(y)}\left(y\ddot{h} + \left(\left(1 + \frac{2\sqrt{\Lambda}}{a^2}\right) - \zeta y\right)\dot{h} + \zeta\left(K_1\lambda + K_2\right)h\right).$$

Setting $z = \zeta y$, $g(z) = (1/\sqrt{\zeta})h(y)$, and $\eta = K_1\lambda + K_2$ yields that, at $z = \zeta y$,

$$\Xi(y) = \frac{1}{2}a^2\zeta^{3/2}\sqrt{\hat{m}(z)m(y)}\left(z\ddot{g} + \left(1 + \frac{2\sqrt{\Lambda}}{a^2} - z\right)\dot{g} + \eta g\right),$$

which is the desired result. □

LEMMA C.2. *Let Assumptions 4.1 and 4.2 hold, and let the functions* $(\psi_n)_{n\geq 0}$ *be defined as in (4.10). If a sequence* $(\gamma_n)_{n\geq 0} \subset \mathbb{R}$ *satisfies, for some* $s > \frac{\omega}{2} + \frac{1}{4}$,

$$\sum_{n=0}^{\infty} n^s |\gamma_n| < \infty \tag{C.8}$$

then, the series

$$\sum_{n=0}^{\infty} \gamma_n \psi_n(z) \tag{C.9}$$

converges uniformly on compacts on $(0, \infty)$.

Proof of Lemma C.2. Hille (1926) shows that the series $\sum_{n=1}^{\infty} n^{-s}\psi_n(z)$ converges uniformly on compacts on $(0, \infty)$. For any positive integers $m < M$

$$\left|\sum_{n=m}^{M} \gamma_n \psi_n(z)\right| = \left|\sum_{n=m}^{M} n^s \gamma_n n^{-s} \psi(z)\right| \leq \max_{n=m,\dots M} \left|n^{-s}\psi_n(z)\right| \sum_{n=m}^{M} n^s \, |\gamma_n|$$

and thus the result follows by (C.8). □

LEMMA C.3. *Let* $\varphi \in L^2((0, \infty), \hat{m})$ *be strictly positive and continuously differentiable. Let the functions* $(\psi_n)_{n\geq 0}$ *be from (4.10). Let*

$$\beta_n = \int_0^{\infty} \varphi(y)\psi_n(y)\hat{m}(y)\, dy.$$

Let $(\mathbb{Q}_z)_{z>0}$ *denote the solution to the martingale problem on* $(0, \infty)$ *for*

$$L = \sqrt{\Theta} z \frac{d^2}{dz^2} + \sqrt{\Theta}\,(1 + \omega - z)\,\frac{d}{dz}. \tag{C.10}$$

Let Z *denote the coordinate mapping process. Let* $\tau(t) = e^{-\sqrt{\Theta}(T-t)}$. *Then, for* $t \in [0, T]$ and $z > 0$

$$\sum_{n=0}^{\infty} \tau(t)^n \beta_n \psi_n(z) = E_z^{\mathbb{Q}}\left[\varphi(Z_{T-t})\right]. \tag{C.11}$$

REMARK C.4. Note that under $\mathbb{Q}_z$, Z is a CIR process starting at z. Since $\sqrt{\Theta}(1 + \omega) > \sqrt{\Theta}$, Z does not hit 0, and thus there is a solution to the martingale problem for L on $(0, \infty)$.

Proof of Lemma C.3. Let $t < T$ and $\gamma_n = \tau(t)^n \beta_n$. Note that for any $s > 0$, by Parseval's equality, and the fact that $\sqrt{\Theta} > 0$

$$\sum_{n=0}^{\infty} n^s |\gamma_n| \leq 2\sum_{n=0}^{\infty} n^{2s}\tau(t)^{2n} + 2\sum_{n=0}^{\infty} \beta_n^2 < \infty. \tag{C.12}$$

Thus, Lemma C.2 implies that the series

$$\sum_{n=0}^{\infty} \tau(t)^n \beta_n \psi_n(z)$$

converges uniformly on compact subsets of $(0, \infty)$. At $t = T$ the pointwise convergence of

$$\sum_{n=0}^{\infty} \beta_n \psi_n(z)$$

follows from Lebedev (1972, chapter 5, theorem 3) since φ is continuously differentiable. Therefore, the series $\sum_{n=0}^{\infty} \tau(t)^n \beta_n \psi_n(z)$ defines a function on $(0, T) \times (0, \infty)$.

Three facts about the Feller diffusion and generalized Laguerre polynomials are necessary to prove (C.11). First, (Glasserman 2004, chapter 3.4), under $\mathbb{Q}_z$, for $0 \le t < T$, Z_{T-t} is distributed like KX, where

$$K = \frac{1}{2}(1 - e^{-\sqrt{\Theta}(T-t)}) = \frac{1}{2}(1 - \tau(t)) \tag{C.13}$$

and X is a noncentral Chi-square random variable with noncentrality parameter λ and degrees of freedom d given by

$$\lambda = \frac{2e^{-\sqrt{\Theta}(T-t)} z}{1 - e^{-\sqrt{\Theta}(T-t)}} = \frac{2\tau(t) z}{1 - \tau(t)}, \quad d = 2(1 + \omega). \tag{C.14}$$

Second, X has probability density equal to

$$\begin{aligned} f(x) &= \frac{1}{2} e^{-x/2 - \lambda/2} x^{d/4 - 1/2} \lambda^{-(d/4 - 1/2)} I_{d/2-1}(\sqrt{\lambda x}) \\ &= \frac{1}{2} e^{-x/2 - \lambda/2} x^{\omega/2} \lambda^{-\omega/2} I_{\omega}(\sqrt{\lambda x}), \end{aligned} \tag{C.15}$$

where $I_\omega(y)$ is the modified Bessel function of the first kind. Third, recall the following identity regarding generalized Laguerre polynomials: for $|\tau| < 1$ (Magnus et al. 1966, chapter 5.5, note the typo in this reference, in that yz/τ should read as $yz\tau$)

$$\sum_{n=0}^{\infty} \psi_n(y)\psi_n(z)\tau^n = \frac{\Gamma(1+\omega)}{1-\tau} (yz\tau)^{-\omega/2} \exp\left(-\frac{\tau(y+z)}{1-\tau}\right) I_{\omega}\left(\sqrt{\frac{4yz\tau}{(1-\tau)^2}}\right). \tag{C.16}$$

(C.15) readily gives that

$$\begin{aligned} E_z^{\mathbb{Q}}[\varphi(Z_{T-t})] &= \int_0^{\infty} \varphi(Kx) f(x) dx = \int_0^{\infty} \varphi(y) f(y/K) \frac{1}{K} dy \\ &= \int_0^{\infty} \varphi(y) \frac{1}{2K} e^{-y/(2K) - \lambda/2} y^{\omega/2} (\lambda K)^{-\omega/2} I_{\omega}\left(\sqrt{\frac{\lambda y}{K}}\right) dy. \end{aligned} \tag{C.17}$$

Thus, if the infinite sum and the integral commute, as proven later, from the definition of $\tau(t)$, β_n and (C.16), it follows that

$$\begin{aligned}\sum_{n=0}^{\infty}\tau(t)^n\beta_n\psi_n(z) &= \sum_{n=0}^{\infty}\tau(t)^n\psi_n(z)\int_0^{\infty}\varphi(y)\psi_n(y)\hat{m}(y)\,dy\\ &= \int_0^{\infty}\varphi(y)\hat{m}(y)\sum_{n=0}^{\infty}\psi_n(y)\psi_n(z)\tau(t)^n\\ &= \int_0^{\infty}\varphi(y)\hat{m}(y)\frac{\Gamma(1+\omega)}{1-\tau(t)}(yz\tau(t))^{-\omega/2}\exp\left(-\frac{\tau(t)(y+z)}{1-\tau(t)}\right)\\ &\quad\times I_{\omega}\left(\sqrt{\frac{4yz\tau(t)}{(1-\tau(t))^2}}\right)dy.\end{aligned}\tag{C.18}$$

Therefore, the result holds if the integrands in (C.17) and (C.18) are the same. Thus, it suffices to show that

$$\begin{aligned}&\varphi(y)\frac{1}{2K}e^{-y/(2K)-\lambda/2}y^{\omega/2}(\lambda K)^{-\omega/2}I_{\omega}\left(\sqrt{\frac{\lambda y}{K}}\right)\\ &= \varphi(y)\hat{m}(y)\frac{\Gamma(1+\omega)}{1-\tau(t)}(yz\tau(t))^{-\omega/2}\exp\left(-\frac{\tau(t)(y+z)}{1-\tau(t)}\right)I_{\omega}\left(\sqrt{\frac{4yz\tau(t)}{(1-\tau(t))^2}}\right)\\ &= \frac{\varphi(y)}{1-\tau(t)}e^{-y/(1-\tau(t))}y^{\omega/2}(z\tau(t))^{-\omega/2}\exp\left(-\frac{\tau(t)z}{1-\tau(t)}\right)I_{\omega}\left(\sqrt{\frac{4yz\tau(t)}{(1-\tau(t))^2}}\right),\end{aligned}\tag{C.19}$$

where the last line comes from plugging in for $\hat{m} = \frac{1}{\Gamma(1+\omega)}y^{\omega}e^{-y}$ and collecting terms. By (C.14) and (C.13), it follows that

$$\frac{\lambda y}{K} = \frac{4yz\tau(t)}{(1-\tau(t))^2},\quad 2K = 1-\tau(t),\quad \frac{\lambda}{2} = \frac{\tau(t)z}{1-\tau(t)},\quad \lambda K = \tau(t)z.$$

Plugging these equalities into the right-hand side of (C.19) gives the desired result

$$\varphi(y)\frac{1}{2K}e^{-y/(2K)}y^{\omega/2}(\lambda K)^{-\omega/2}\exp(-\lambda/2)\,I_{\omega}\left(\sqrt{\frac{\lambda y}{K}}\right).$$

It remains to prove that in (C.18) the infinite sum and the integral commute. To see this, for each integer M and a fixed $z > 0$, set

$$f_M(y) = \frac{1}{\varphi(y)}\sum_{n=0}^{M}\tau(t)^n\psi_n(z)\psi_n(y).$$

Thus, if the integral and infinite sum commute,

$$\lim_{M\uparrow\infty}\int_0^{\infty}f_M(y)\varphi^2(y)\hat{m}(y)\,dy = \int_0^{\infty}\left(\lim_{M\uparrow\infty}f_M(y)\right)\varphi^2(y)\hat{m}(y)\,dy.$$

Since $\varphi \in L^2((0,\infty),\hat{m})$, it suffices to show that

$$\sup_{M\in\mathbb{N}} \int_0^\infty f_M^2(y)\varphi^2(y)\hat{m}(y)\,dy < \infty.$$

Indeed, if this is the case, then $(f_M)_{M\in\mathbb{N}}$ is bounded in L^2, hence uniformly integrable, for the finite measure $\varphi^2\hat{m}$. To this end, the orthonormality of (ψ_n) in $L^2((0,\infty),\hat{m})$ implies that

$$\int_0^\infty f_M^2(y)\varphi^2(y)\hat{m}(y)\,dy = \sum_{n=0}^{M} \tau^{2n}\psi_n^2(z) \le \sum_{n=0}^{\infty} \tau^{2n}\psi_n^2(z)$$

and the right-hand side is finite, in view of the explicit formula in (C.16). □

Proof of Lemma 4.4. By applying Lemma C.1 for

$$f(y) = \phi(y), \quad g(z) = \psi(z),$$

it follows that ϕ satisfies (4.4) for some $\lambda \in \mathbb{R}$ if and only if ψ satisfies (4.8) for $\eta = K_1\lambda + K_2$. As for the respective L^2 norms, since

$$\phi(y)m(y)^{1/2} = \sqrt{\zeta}\,\psi(\zeta y)\hat{m}(\zeta y)^{1/2}$$

with m from (4.3) (normalized to be a probability measure) and $\hat{m}$ from (4.7) it follows that

$$\int_0^\infty \phi(y)^2 m(y)\,dy = \int_0^\infty \zeta\,\psi(\zeta y)^2\hat{m}(\zeta y)\,dy = \int_0^\infty \psi(z)^2\hat{m}(z)\,dz.$$

Thus, $\phi \in L^2((0,\infty), m)$ if and only if $\psi \in L^2((0,\infty),\hat{m})$ and they have the same norm. □

Proof of Lemma 4.5. We first show that v in (4.13) is a strictly positive $C^{1,2}((0,T)\times(0,\infty))$ solution to the PDE in (2.18) (equivalently, the PDE in (2.20)), specified to the model in (4.1). Equations (4.6), (4.7), (4.10), and (4.11) imply that for any $y > 0$, at $z = \zeta y$

$$v(t,y) = \sum_{n=0}^{\infty} e^{-\lambda_n(T-t)}\alpha_n\phi_n(y) = e^{K_2/K_1(T-t)}\varphi(z)^{-1}\sum_{n=0}^{\infty} e^{-n\sqrt{\Theta}(T-t)}\alpha_n\psi_n(z). \tag{C.20}$$

Set

$$w(t,z) = \sum_{n=0}^{\infty} e^{-n\sqrt{\Theta}(T-t)}\alpha_n\psi_n(z) \tag{C.21}$$

so that (C.20) becomes

$$v(t,y) = e^{K_2/K_1(T-t)}\varphi(z)^{-1}w(t,z). \tag{C.22}$$

Applying Lemma C.1 to

$$\lambda = 0, \quad f(t,y) = v(t,y), \quad g(t,z) = w(t,z)e^{K_2/K_1(T-t)},$$

it follows that $v(t, y)$ solves (2.20) if and only if $w(t, z)$ solves (note $K_1 = 1/\sqrt{\Theta}$)

$$w_t + \sqrt{\Theta} z w_{zz} + \sqrt{\Theta}(1 + \omega - z) w_z = 0 \quad (t, z) \in (0, T) \times (0, \infty), \tag{C.23}$$
$$w(T, z) = \varphi(z) \quad z \in (0, \infty).$$

For the operator L from (C.10), rewrite now (C.23) as

$$w_t + Lw = 0 \quad (t, z) \in (0, T) \times (0, \infty),$$
$$w(T, z) = \varphi(z) \quad z \in (0, \infty).$$

Since $\varphi \in L^2((0, \infty), \hat{m})$, Lemma C.3 implies that

$$w(t, z) = \sum_{n=0}^{\infty} e^{-n\sqrt{\Theta}(T-t)} \alpha_n \psi_n(z) = E_z^{\mathbb{Q}} [\varphi(Z_{T-t})]. \tag{C.24}$$

Proposition D.2 applies if, for any $T > 0$

$$\sup_{0 \le t \le T} E_z^{\mathbb{Q}} [\varphi(Z_{T-t})] < \infty.$$

To see this, note that, under Assumptions 4.1 and 4.2 for the operator L in (C.10), it follows from equations (4.3), (4.7), (4.11), and (C.1) that there is a constant K such that

$$L\varphi(z) \le K\varphi(z).$$

Therefore,

$$\partial_t (e^{K(T-t)} \varphi(z)) + L(e^{K(T-t)} \varphi(z)) \le 0$$

and hence using Itô's formula and Fatou's lemma it follows that

$$\sup_{0 \le t \le T} E_z^{\mathbb{Q}} [\varphi(Z_{T-t})] \le e^{KT} \varphi(z) < \infty.$$

Thus, Proposition D.2 implies that w satisfies (C.23). Because φ is strictly positive, so are v and w. Next, to see that v satisfies the convergence relation in (4.14), note that (C.22) implies that, at $z = \zeta y$,

$$\frac{v_y(t, y)}{v(t, y)} = \zeta \left(\frac{w_z(t, z)}{w(t, z)} - \frac{\dot{\varphi}(z)}{\varphi(z)} \right). \tag{C.25}$$

Now, (4.9) implies that $\psi_0(z) = 1$. Also, (4.6) and (4.11) imply that $\phi_0(y) = \varphi(z)^{-1}$ at $z = \zeta y$, and hence

$$\frac{v_y(t, y)}{v(t, y)} - \frac{\dot{\phi}_0(y)}{\phi_0(y)} = \zeta \frac{w_z(t, z)}{w(t, z)} = \zeta \left(\frac{w_z(t, z)}{w(t, z)} - \frac{\dot{\psi}_0(z)}{\psi_0(z)} \right)$$

and thus (4.14) follows if

$$\lim_{T \uparrow \infty} \frac{w_z(t, z)}{w(t, z)} = \frac{\dot{\psi}_0(z)}{\psi_0(z)} \tag{C.26}$$

holds for all $(t, z) \in (0, \infty) \times (0, \infty)$. Set w_M by

$$w_M(t, z) = \sum_{n=0}^{M} \tau(t)^n \alpha_n \psi_n(z).$$

(C.26) will follow if w_M and $\dot{w}_M$ locally (in z) converge uniformly as $M \uparrow \infty$. Repeating the argument at the beginning of the proof of Lemma C.3, it follows that, for $t < T$, $w_M(t, z)$ converges uniformly in z on compact subsets of $(0, \infty)$. To see the convergence of $\dot{w}_M(t, z)$, recall the recurrence relation for the Laguerre polynomials (Hille 1926):

$$z\dot{\psi}_n = (n+1)\psi_{n+1} - (n+1+\omega-z)\psi_n. \tag{C.27}$$

This relation yields, for each $z > 0$,

$$\begin{aligned}
\dot{w}_M(t, z) &= \frac{1}{z} \sum_{n=0}^{M} \tau(t)^n \alpha_n \left((n+1)\psi_{n+1}(z) - (n+1+\omega-z)\psi_n(z)\right) \\
&= \frac{1}{z} \sum_{n=0}^{M} \alpha_n \tau(t)^n (n+1)\psi_{n+1}(z) \\
&\quad - \frac{1}{z} \sum_{n=0}^{M} \left(\tau(t)^n n \alpha_n \psi_n(z) - \frac{1}{z}(1+\omega-z) w_m(t, z) \right).
\end{aligned}$$

It has already been shown that w_M converges uniformly in z on compact subsets of $(0, \infty)$. As for the other two sums

$$\sum_{n=0}^{M} \alpha_n \tau(t)^n (n+1)\psi_{n+1}(z) = \sum_{n=0}^{\infty} \gamma_n \psi_n(z) \quad \gamma_0 = 0, \gamma_n = n\alpha_{n-1}\tau(t)^{n-1}, n = 1, 2, \ldots,$$

$$\sum_{n=0}^{M} \tau(t)^n n \alpha_n \psi_n(z) = \sum_{n=0}^{\infty} \gamma_n \psi_n(z) \quad \gamma_n = n\alpha_n \tau(t)^n, n = 0, 1, \ldots.$$

Thus, Parseval's equality and $\sqrt{\Theta} > 0$ imply that (C.8) holds for any $s > 0$ uniformly for $t \in (0, T - \varepsilon)$ for any $\varepsilon > 0$. Thus, $\dot{w}_m$ converges uniformly for z on compact subsets of $(0, \infty)$, and for $t < T - \varepsilon$ for any $\varepsilon > 0$. Using the local uniform convergence of w_M and $\dot{w}_M$, it follows that

$$\begin{aligned}
\lim_{T \uparrow \infty} \frac{w_z(t, z)}{w(t, z)} &= \lim_{T \uparrow \infty} \frac{\displaystyle\sum_{n=0}^{\infty} \tau(t)^n \alpha_n \dot{\psi}_n(z)}{\displaystyle\sum_{n=0}^{\infty} \tau(t)^n \alpha_n \psi_n(z)} \\
&= \lim_{T \uparrow \infty} \frac{\alpha_0 \dot{\psi}_0(z) + \displaystyle\sum_{n=1}^{\infty} \tau(t)^n \alpha_n \dot{\psi}_n(z)}{\alpha_0 \psi_0(z) + \displaystyle\sum_{n=1}^{\infty} \tau(t)^n \alpha_n \psi_n(z)}
\end{aligned}$$

and thus it suffices to show that

$$\lim_{T\uparrow\infty}\lim_{M\uparrow\infty}\sum_{n=1}^{M}\tau(t)^n\alpha_n\dot{\psi}_n(z)=0 \quad \text{and} \quad \lim_{T\uparrow\infty}\lim_{M\uparrow\infty}\sum_{n=1}^{\infty}\tau(t)^n\alpha_n\psi_n(z)=0.$$

We prove the first limit, the second one being analogous. Fix $t, z > 0$, and let $\tilde{T}$ be such that $t < \tilde{T} < T$. By the local uniform convergence of $\dot{w}_M$ for $\tilde{T}$ instead of T, for any $\epsilon > 0$ there is a M_ε such that $m, M > M_\varepsilon$ imply that

$$\max_{N=m,\ldots,M}\left|\sum_{n=m}^{N}e^{-n\sqrt{\Theta}(\tilde{T}-t)}\alpha_n\dot{\psi}_n(z)\right|<\varepsilon.$$

Thus, Lemma B.2 with $\gamma_n = e^{-n\sqrt{\theta}(T-\tilde{T})}$ and $h_n = e^{-n\sqrt{\Theta}(\tilde{T}-t)}\alpha_n\dot{\psi}_n(z)$ implies that

$$\begin{aligned}
\lim_{T\uparrow\infty}\lim_{M\uparrow\infty}\left|\sum_{n=1}^{M}\tau(t)^n\alpha_n\dot{\psi}_n(z)\right| &\le \lim_{T\uparrow\infty}\lim_{M\uparrow\infty}\left|\sum_{n=M_\varepsilon+1}^{M}\tau(t)^n\alpha_n\dot{\psi}_n(z)\right| \\
&= \lim_{T\uparrow\infty}\lim_{M\uparrow\infty}\left|\sum_{n=M_\varepsilon+1}^{M}e^{-n\sqrt{\Theta}(T-\tilde{T})}e^{-n\sqrt{\Theta}(\tilde{T}-t)}\alpha_n\dot{\psi}_n(z)\right| \\
&\le \lim_{T\uparrow\infty}e^{-(M_\varepsilon+1)\sqrt{\Theta}(T-\tilde{T})}\max_{N=m,\ldots,M}\left|\sum_{n=m}^{N}e^{-n\sqrt{\Theta}(\tilde{T}-t)}\alpha_n\dot{\psi}_n(z)\right| \\
&\le \varepsilon\lim_{T\uparrow\infty}e^{-(M_\varepsilon+1)\sqrt{\Theta}(T-\tilde{T})} \\
&= 0.
\end{aligned}$$

Now we prove that the value function $V(x, t, y)$ equals $\frac{x^p}{p}v(t, y)^\delta$ using Lemma A.3. It has already been shown that v satisfies the PDE in (A.1) specified to (4.1). Thus, it only remains to prove that the model $\hat{P}_T$ satisfies Assumption A.2. Specified to the model in (4.1), $\hat{P}_T$ takes the form:

$$\begin{aligned}
dR_t &= \frac{1}{1-p}\left(\sigma\nu_0+\sigma\nu_1 Y_t+\delta Y_t\sigma\rho a\frac{v_y}{v}(t, Y_t)\right)dt+\sqrt{Y_t}\sigma\, dZ_t, \\
dY_t &= \left(\alpha+\frac{a^2}{2}-\beta Y_t+a^2 Y_t\frac{v_y}{v}(t, Y_t)\right)dt+a\sqrt{Y_t}dW_t.
\end{aligned}$$

Note that it suffices to prove there is a weak solution for the second SDE involving Y. If this is the case, there is a weak solution for Y, B, where B is an n-dimensional Brownian Motion independent of Y. Then, setting $Z = \rho W + \bar{\rho}B$ and defining R accordingly will result in a weak solution for R, Y.

Regarding Y, first note that there is a weak solution to the SDE:

$$dZ_t = \sqrt{\Theta}(1+\omega-Z_t)\,dt+\sqrt{2\sqrt{\Theta}}\sqrt{Z_t}dW_t.$$

Indeed, the operator L associated to this solution is given in (C.10) in Lemma C.3. Let $(\mathbb{Q}_z)_{z>0}$ denote the solution to the martingale problem for L on $(0, \infty)$. From Proposition D.2 and (C.24) it follows that $w(t, Z_t)/w(0, z)$, $0 \le t \le T$ is a $\mathbb{Q}_z$ martingale and

$$\frac{w(t, Z_t)}{w(0, z)} = \mathcal{E}\left(\int_0^{\cdot} \sqrt{2\sqrt{\Theta}}\sqrt{Z_s}\frac{w_z}{w}(t, Z_s)\, dW_t\right)_t.$$

Therefore, by Girsanov's theorem, there is a weak solution to the SDE:

$$dZ_t = \left(\sqrt{\Theta}(1+\omega) - \sqrt{\Theta}Z_t + 2\sqrt{\Theta}Z_t\frac{w_z}{w}(t, Z_t)\right) dt + \sqrt{2\sqrt{\Theta}}\sqrt{Z_t}\, dW_t. \tag{C.28}$$

Now, using (4.11) and (C.1) it follows that

$$dZ_t = \left(\zeta\left(\alpha + \frac{a^2}{2}\right) - \beta Z_t + a^2\zeta Z_t\left(\frac{w_z}{w}(t, Z_t) - \frac{\dot{\varphi}}{\varphi}(Z_t)\right)\right) dt + a\sqrt{\zeta}\sqrt{Z_t}\, dW_t.$$

Using (C.25) it follows that

$$dZ_t = \left(\zeta\left(\alpha + \frac{a^2}{2}\right) - \beta Z_t + a^2 Z_t\frac{v_y}{v}(t, Z_t/\zeta)\right) dt + a\sqrt{\zeta}\sqrt{Z_t}\, dW_t.$$

Setting $Y = Z/\zeta$ it follows that

$$dY_t = \left(\alpha + \frac{a^2}{2} - \beta Y_t + a^2 Y_t\frac{v_y}{v}(t, Y_t)\right) dt + a\sqrt{Y_t}\, dW_t.$$

Therefore, there is a weak solution for Y, and hence (R, Y), with $\hat{P}_T$ dynamics. It remains to show the static fund separation in (4.15). By (4.6), and the relation $\psi_0 = 1$, it follows that

$$\frac{\dot{\phi}_0(y)}{\phi_0(y)} = -\zeta\frac{\dot{\varphi}(\zeta y)}{\varphi(\zeta y)}.$$

Plugging in for $\dot{\varphi}/\varphi$ according to (4.11) and (C.1) gives

$$\frac{\dot{\phi}_0(y)}{\phi_0(y)} = -\zeta\frac{\dot{\varphi}(\zeta y)}{\varphi(\zeta y)} = \frac{\sqrt{\Lambda} - \alpha}{a^2} \times \frac{1}{y} + \frac{\beta - \sqrt{\Theta}}{a^2} \times 1,$$

and (C.1) again yields (4.15). □

APPENDIX D: EXTENSION OF THE FEYNMAN-KAČ FORMULA

Let $E = (\alpha, \beta)$, $-\infty \le \alpha < \beta \le \infty$ be an open interval in $\mathbb{R}$, and $C^{m,\alpha}(E, \mathbb{R}^d)$ the class of $\mathbb{R}^d$-valued continuous maps on E whose mth derivative is uniformly α-Hölder continuous on compact subsets of E. Set $C^m(E, \mathbb{R}^d) = C^{m,0}(E, \mathbb{R}^d)$. $\Omega^d = C([0, \infty), \mathbb{R}^d)$ denotes the space of continuous paths from $[0, \infty)$ to $\mathbb{R}^d$, endowed with the Borel σ-algebra $\mathcal{F}$. For notational simplicity $\Omega^1 = \Omega$.

The following extension of the Feynman–Kač formula, proved in Guasoni et al. (2011), is needed for the proofs of Theorem 3.5 and Theorem 4.5. Because Proposition D.2 is used in a number of contexts the A and b in Assumption D.1 and Proposition D.2 are not necessarily the same A and b as in (2.2)–(2.3).

ASSUMPTION D.1. *Let $\gamma \in (0, 1]$, $A \in C^{2,\gamma}(E)$; $A(z) > 0$, $z \in E$ and $b \in C^{1,\gamma}(E, \mathbb{R})$. Assume there exists a solution $(\mathbb{Q}_z)_{z\in E}$ to the martingale problem for L on E where L is given by*

$$L = \frac{1}{2} A(z) \frac{d^2}{dz^2} + b(z) \frac{d}{dz}.$$

Let Z denote the coordinate mapping process.

PROPOSITION D.2. *Let Assumption D.1 hold. Let $f \in C^2(E)$; $f(z) > 0$, $z \in E$* be *such that for any $T > 0$ and $z \in E$*

$$\sup_{0 \le t \le T} E_z^{\mathbb{Q}} [f(Z_{T-t})] < \infty. \tag{D.1}$$

Then the function $h(t, z)$: $(0, T) \times E \mapsto (0, \infty)$ defined by

$$h(t, z) = E_z^{\mathbb{Q}} [f(Z_{T-t})]$$

is in $C^{1,2}((0, T) \times E)$ and satisfies the following differential expression and terminal condition:

$$\partial_t h(t, z) + Lh(t, z) = 0 \quad (t, z) \in (0, T) \times E, \tag{D.2}$$

$$h(T, z) = f(z) \quad z \in E.$$

Furthermore, $h(t, Z_t)/h(0, z)$ is a $\mathbb{Q}_z$ martingale for all $z \in E$.

REFERENCES

BARBERIS, N. (2000): Investing for the Long Run When Returns Are Predictable, *J. Finance* 55(1), 225–264.

BENSOUSSAN, A., J. KEPPO, and S. SETHI (2009): Optimal Consumption and Portfolio Decisions with Partially Observed Real Prices, *Math. Finance* 19(2), 215–236.

BRANDT, M. (1999): Estimating Portfolio and Consumption Choice: A Conditional Euler Equations Approach, *J. Finance* 54(5), 1609–1645.

CAMPBELL, J., and S. THOMPSON (2008): Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? *Rev. Finan c*. 21(4), 1509–1531.

CASS, D., and J. E. STIGLITZ (1970): The Structure of Investor Preferences and Asset Returns, and Separability in Portfolio Allocation: A Contribution to the Pure Theory of Mutual Funds, *J. Econ. Theory* 2, 122–160.

CHAMBERLAIN, G. (1988): Asset Pricing in Multiperiod Securities Markets, *Econometrica* 56(6), 1283–1300. http://dx.doi.org/10.2307/1913098

CHERIDITO, P., D. FILIPOVIĆ, and M. YOR (2005): Equivalent and Absolutely Continuous Measure Changes for Jump-Diffusion Processes, *Ann. Appl. Probab.* 15(3), 1713–1732.

COCHRANE, J. (2008): The Dog that Did Not Bark: A Defense of Return Predictability, *Rev. Financ. Stud.* 21(4), 1533–1575.

COX, J. C., and C.-F. HUANG (1992): A Continuous-Time Portfolio Turnpike Theorem, *J. Econ. Dyn. Control* 16(3–4), 491–507.

DYBVIG, P. H., L. ROGERS, and K. BACK (1999): Portfolio Turnpikes, *Rev. Financ. Stud.* 12(1), 165–195.

FELLER, W. (1951): Two Singular Diffusion Problems, *Ann. Math.* 54(2), 173–182.

GLASSERMAN, P. (2004): *Monte Carlo Methods in Financial Engineering: Applications of Mathematics (New York)*, Stochastic Modelling and Applied Probability, Vol. 53, New York: Springer-Verlag.

GUASONI, P., C. KARDARAS, S. ROBERTSON, and H. XING (2011): Abstract, Classic, and Explicit Turnpikes, Preprint, arXiv:1101.0945.

GUASONI, P., and S. ROBERTSON (2012): Portfolios and Risk Premia for the Long Run, *Ann. Appl. Probab.* 22(1), 239–284.

HAKANSSON, N. H. (1974): Convergence to Isoelastic Utility and Policy in Multiperiod Portfolio Choice, *J. Financ. Econ.* 1, 201–224.

HILLE, E. (1926): On Laguerre's Series, Second Note, *Proc. Natl. Acad. Sci. U.S.A.* 12(4), 265–269.

HUANG, C.-F., and T. ZARIPHOPOULOU (1999): Turnpike Behavior of Long-Term Investments, *Finance Stoch.* 3(1), 15–34.

HUBERMAN, G., and S. ROSS (1983): Portfolio Turnpike Theorems, Risk Aversion, and Regularly Varying Utility Functions, *Econometrica* 51(5), 1345–1361.

KHANNA, A., and M. KULLDORFF (1999): A Generalization of the Mutual Fund Theorem, *Finance Stoch.* 3(2), 167–185. http://dx.doi.org/10.1007/s007800050056.

KIM, T., and E. OMBERG (1996): Dynamic Nonmyopic Portfolio Behavior, *Rev. Financ. Stud.* 9(1), 141–161.

KRAMKOV, D., and W. SCHACHERMAYER (1999): The Asymptotic Elasticity of Utility Functions and Optimal Investment in Incomplete Markets, *Ann. Appl. Probab.* 9(3), 904–950.

LEBEDEV, N. N. (1972): *Special Functions and Their Applications*, revised ed., Richard A. Silverman, ed., New York: Dover Publications, Inc. (translated from Russian, unabridged and corrected republication).

LELAND, H. (1972): On Turnpike Portfolios, in *Mathematical Models in Investment and Finance*, M. Harris and R Stulz, eds., Amsterdam: North-Holland.

MAGNUS, W., F. OBERHETTINGER, and R. P. SONI (1966): *Formulas and Theorems for the Special Functions of Mathematical Physics*, 3rd enlarged ed., Die Grundlehren der mathematischen Wissenschaften, Band 52, New York: Springer-Verlag.

MERTON, R. (1973): An Intertemporal Capital Asset Pricing Model, *Econometrica* 41(5), 867–887.

MIKLAVČIČ, M. (1998): *Applied Functional Analysis and Partial Differential Equations*, River Edge, NJ: World Scientific Publishing Co. Inc.

PAN, J. (2002): The Jump-Risk Premia Implicit in Options: Evidence from an Integrated Time-Series Study, *J. Financ. Econ.* 63(1), 3–50.

REED, M., and B. SIMON (1972): *Methods of Modern Mathematical Physics. I. Functional Analysis*, New York: Academic Press.

ROSS, S. A. (1978): Mutual Fund Separation in Financial Theory—The Separating Distributions, *J. Econom. Theory* 17(2), 254–286.

SCHACHERMAYER, W., M. SÎRBU, and E. TAFLIN (2009): In Which Financial Markets Do Mutual Fund Theorems Hold True? *Finance Stoch.* 13(1), 49–77. http://dx.doi.org/10.1007/s00780-008-0072-x.

TOBIN, J. (1958): Liquidity Preference as Behavior Towards Risk, *Rev. Econ. Stud.* 25(2), 65–86.

WACHTER, J. (2002): Portfolio and Consumption Decisions under Mean-Reverting Returns: An Exact Solution for Complete Markets, *J. Financ. Quant. Anal.* 37(1), 63–91.

WELCH, I., and A. GOYAL (2008): A Comprehensive Look at the Empirical Performance of Equity Premium Prediction, *Rev. Financ. Stud.* 21(4), 1455–1508.

XIA, Y. (2001): Learning about Predictability: The Effects of Parameter Uncertainty on Dynamic Asset Allocation, *J. Finance* 56(1), 205–246.

ZARIPHOPOULOU, T. (2001): A Solution Approach to Valuation with Unhedgeable Risks, *Finance Stoch.* 5(1), 61–82.

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

Mathematical Finance, Vol. 25, No. 4 (October 2015), 827–868

A GENERAL EQUILIBRIUM MODEL OF A MULTIFIRM MORAL-HAZARD ECONOMY WITH FINANCIAL MARKETS

JAEYOUNG SUNG

Ajou University

XUHU WAN

Hong Kong University of Science and Technology

We present a general equilibrium model of a moral-hazard economy with many firms and financial markets, where stocks and bonds are traded. Contrary to the principal-agent literature, we argue that optimal contracting in an infinite economy is not about a tradeoff between risk sharing and incentives, but it is all about incentives. Even when the economy is finite, optimal contracts do not depend on principals' risk aversion, but on market prices of risks. We also show that optimal contracting does not require relative performance evaluation, that the second best risk-free interest rate is lower than that of the first best, and that the second-best equity premium can be higher or lower than that of the first best. Moral hazard can contribute to the resolution of the risk-free rate puzzle. Its potential to explain the equity premium puzzle is examined.

KEY WORDS: moral hazard, general equilibrium, optimal contract, relative performance evaluation, equity premium, interest rate.

1. INTRODUCTION

We present an equilibrium model of product and financial markets for a moral-hazard economy with (infinitely/finitely) many firms which produce numeraire goods. In spite of the importance of moral hazard effects on multiasset economies as we witnessed from the recent subprime mortgage crisis, little is known about properties of equilibria beyond their existence. In particular, interactions of financial markets and agency problems are poorly understood. Our model enables us to examine equilibrium effects of moral hazard on such popular fundamental financial economic issues as optimal contracts, relative performance measure, real investment decisions, portfolio decisions, interest rates, and equity premia.

One may epitomize main results in the existing optimal contracting literature as follows: (1) optimal contracting is a tradeoff between risk sharing and incentives (see Stiglitz 1974; Shavell 1979); and (2) the performance measure for optimal contracting should be constructed based on relative performance evaluation (RPE) (see Lazear and

Sung was supported by WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-2009-000-20007-0). Wan was supported by Hong Kong government general research fund grant GRF 620909.

We would like to thank for valuable comments the editor (Jerome Detemple), an anonymous associate editor, an anonymous referee, and seminar participants at Ajou University, Shandong University, Fudan University, and March 2012 American Mathematical Society Meetings at Tampa, Florida.

Manuscript received November 2010; final revision received November 2012.

Address correspondence to Jaeyoung Sung, Ajou University-Financial Engineering, San 5, Woncheon-dong, Yeongtong-gu, Suwon 443-749, Republic of Korea; e-mail: jaeyoungsung@ajou.ac.kr.

DOI: 10.1111/mafi.12032

Rosen 1981; Holmstrom 1982; Green and Stokey 1983; Nalebuff and Stiglitz 1983; and recently Ou-Yang 2005). The RPE measure is to reward the agent for performance, but not for luck, and typically consists of signals for unobservable agent effort/decisions, being independent of systematic risks such as interest rate risks, exchange rate risks, and market-wide productivity shocks.

We argue that optimal contracting in an infinitely large economy with financial markets is not about a tradeoff between risk sharing and incentives, but all about incentives. In a finite economy with financial markets, the risk sharing role of each contract still matters, which is consistent with the literature on partial equilibrium agency models. However, unlike the literature, we find that the extent of risk sharing through optimal contracting in the finite economy directly depends on the market price of the unique risk of the firm, not on the principal's (shareholders') risk aversion. As the economy grows, the market price of each unique risk converges to zero, and the optimal contract becomes free from the role of risk sharing, and is only left with incentives issues.

Each agent's incentives are determined by his aggregate risk exposure resulting not only from the contract but from his financial portfolio position. Thus, the principal faces a fundamental challenge in optimal contracting. Given any contract, if the agent were allowed to privately trade stocks in financial markets including the stock of the firm under his own management, he could always undo incentives intended by the contract through financial transactions. Hence, without proper restrictions on the agent's financial transactions, the principal's desired optimum cannot be achieved in general. One way for her to attain her optimum is to forbid the agent to trade/short-sell the stock of her firm. (In this case, he may still be allowed to freely trade bonds and all other stocks.)

Under this trading restriction, we show that an optimal compensation contract should optimally depend on the unique risk of the firm which the agent manages. What is striking is that in an economy with financial markets, the optimal contract can still remain to be optimal even when nonperformance-related risks such as the systematic risk of the market and unique risks of all other firms are added into the contract, as long as the agent is compensated for those additional risks at market-determined risk-premia. In other words, the principal's optimum would not be affected by adding/removing those extra risks at market determined risk premia to/from an already optimal contract. A corollary to this result is that in the presence of financial markets, the RPE emphasized in the standard principal-agent literature becomes irrelevant.

Given the above-mentioned properties of optimal contracts, we investigate implications of moral hazard on asset prices and interest rates. We show that right after signing an optimal compensation contract, each principal optimally unloads her entire real-asset position through financial transactions, and holds the market portfolio, and each agent holds the market portfolio less the stock of the firm under his own management. In a finite economy, this difference in portfolio decisions by principals and agents competitively affects market prices of idiosyncratic risks, but not the market price of the systematic risk. In the limit of the finite economy, as is the case with the first-best economy, market prices of idiosyncratic risks in the second best approach zero, and eventually only the systematic risk matters.

The market prices of risks and interest rates influence the costs of capital for real investment and thus consumption decisions, which in turn affect the market prices of risks and interest rates, in equilibrium. Taking this mutual interaction into account, we show that the second best interest rate is less than that of the first best, and that the second best equity premium is higher (lower) than that of the first best if and only if the second-best equilibrium real investment level is lower (higher) than that of the first-best

equilibrium. In other words, in equilibrium, moral hazard decreases interest rates but it can increase or decrease the equity premium. These comparisons are qualitatively the same as those of the single-firm economy of Sung and Wan (2009) although each firm in this paper exerts no direct influence on the market price of market/systematic risk, whereas the single firm in Sung and Wan does. For intuitions about comparison results on interest rates and equity premium, interested readers are referred to that paper.

However, these comparisons of the first- and second-best risk-free rates and equity premia may not be directly related to the well-known Weil's (1989) risk-free and Mehra-Prescott's (1985) equity premium puzzles. The reason is that economic situations leading to the comparisons and to the puzzles are different. The puzzles are concerned with the level of the representative investor's risk aversion given observed risk-free rates, equity premia, and volatilities of the market portfolio, whereas our foregoing comparisons result from examining equilibrium risk-free rates, equity premia, and volatilities of the market portfolio, given risk-aversion levels of principals and agents.

Recast in terms of economic settings in the literature on the two puzzles, our above-mentioned equilibrium effects imply that moral hazard can help explain the low risk-free rate puzzle, but it cannot contribute to the resolution of the equity premium puzzle. Consider two economies: one is first best and the other second best. Suppose they happen to support the same risk-free rates. Recall that the interest rate in equilibrium is the marginal rate of substitution (MRS) between current and future certainty-equivalent consumption. If the degrees of risk aversion were the same, the second-best interest rate would be lower than that of the first best. In order to increase the second-best interest rate to the level of the first best, the MRS has to be increased, which can be achieved by lowering investors' (the principal's) risk aversion. Hence, our moral-hazard economy helps explain the low risk-free rate puzzle.

On the other hand, if the return volatilities of the market portfolios of the classical (first best) and moral-hazard economies were the same, investors in both economies would price the market risks in financial markets in exactly the same way, independently of the presence of moral hazard occurring in product markets. This inability of moral hazard to help explain the equity premium puzzle is in contrast with results in the literature. See Kahn (1990) and Kocherlakota and Pistaferri (2009) for moral hazard resulting in a high equity premium, and Kocherlakota (1998) for moral hazard causing a low equity premium. Thus, the moral-hazard effect on the equity premium may still remain unresolved. We leave this issue for future research.[1]

All the above results are obtained based on the linearity of optimal contracts for our continuous-time moral-hazard economy with product and financial markets that is presented in Appendix G. The product markets are modeled using the well-known Holmstrom and Milgrom's (1987) moral-hazard economy with all individuals exhibiting constant absolute risk aversion (CARA), and the financial markets are borrowed from the asset-pricing literature. Holmstrom and Milgrom show that the optimal contract is a linear function of the outcome. However, it is well known that any deviation from the Holmstrom–Milgrom setting may cause the failure of the linearity result. For example, if the contract is subject to limited liability which requires a lower bound for the agent

[1] It is well known that a classical economy with the representative investor with a utility function exhibiting habit persistence can result in a high-equilibrium equity premium. See Constantinides (1990). Thus, our low equilibrium risk-free rate under moral hazard seems to suggest a potential to explain the low risk-free rate and high equity premium simultaneously if the moral-hazard economy consists of individuals with habit persistent preferences.

compensation, then linear contracts can no longer be optimal. Moreover, it is not a priori clear whether the optimal contract still remains to be linear if the economy has financial markets as well as product markets. In Appendix G, we show that the Holmstrom–Milgrom linearity result still holds even in the presence of financial markets. Intuitively, we obtain this linearity again. The reason is that individuals with CARA preferences can undo undesired complexities in their overall wealth distribution generated from both product and financial markets through financial transactions affecting both the mean and variance of their financial wealth, to the extent that their resulting (self-financing) overall wealth process can be expressed in the form of arithmetic Brownian motion. Consequently, as seen in Sung (1995), Holmstrom and Milgrom's stationarity is still preserved, resulting in linear contracts. Our linearity result not only extends the Holmstrom–Milgrom linearity to financial markets, but enables us to solve moral hazard problems complicated with financial transactions in closed forms.

In terms of the model, our paper can be positioned in the literature on general equilibrium under moral hazard. Prescott and Townsend (1984) present a general equilibrium model of moral hazard where a central planner designs contracts for all agents, and argue that a constrained competitive equilibrium can exist and implement a (constrained) Pareto-efficient allocation. Extensions of Prescott-Townsend's seminal work include Bisin and Gottardi (1999) and Zame (2007). Citanna and Villanacci (2002) incorporate contracting problems into a general equilibrium framework for an economy with commodity markets but without financial markets, and show the existence of equilibria. Danthine and Donaldson (2007) model a principal-agent economy in general equilibrium from a growth theoretic perspective to examine the structure of first-best contracts. Magill and Quinzii (2005) argue that the second-best competitive equilibrium can be Pareto optimal only when principals are risk neutral. This paper is closest to Citanna and Villanacci, but unlike them, our model allows financial markets and enables us to examine properties of equilibria.

Our results are also related to the literature on RPE. In the literature, theoretical effects of financial markets on RPE are somewhat mixed. Maug (2000) argues that if the manager of the firm can trade on assets other than his own company's stock, then there are typically no or only small roles for RPE. Garvey and Milbourn (2003) claim that if there are hedging costs, the optimal level of RPE is determined by equalizing marginal hedging costs to the principal and agent.[2] Ou-Yang (2005) shows that if she is risk averse, the principal puts less emphasis on RPE than she would if she were risk neutral. Neither Maug nor Garvey and Milbourn explicitly consider financial markets, but Ou-Yang incorporates financial markets in his partial equilibrium model of asset pricing and moral hazard. On the other hand, Chiappori and Salanié (2002) in a survey article point out an empirical bewilderment known as "the RPE puzzle" in the literature, by stating that "firms do not seem to use relative performance evaluation of managers very much." In our general equilibrium model, our irrelevance result may help explain

[2]Garvey and Milbourn argue that the sensitivity of the contract becomes independent of systematic risks contained in the performance measure, provided that the manager can hedge systematic risks without affecting his expected wealth. If hedging (costly) affects individuals' expected wealth levels, the sensitivity decreases as the level of systematics risk in the performance measure increases. Ignoring portfolio choice problems, the authors also argue that the optimal level of RPE equalizes marginal hedging costs between shareholders and the manager. However, we argue that in the presence of financial markets, the sensitivity is not affected at all by the extent of RPE, and the RPE is irrelevant in optimal contracting, even when hedging (systematic risk) affects individuals' expected wealth levels.

this puzzle and supports Maug's results even without his assumption on the agent trading restriction.

This paper is organized as follows: The next section describes the economy with both product and capital/financial markets. In Section 3 we restate the expected utilities of principals and agents in terms of their certainty equivalent wealth levels, which will be used in later sections to simplify their problems using change of variables. Section 4 presents first-best solutions which will be used to compare with second-best solutions. Second-best solutions are provided in Section 5 where we show main results of the paper including the irrelevance of RPE and risk-sharing role of compensation contracts in the presence of financial markets. Then, in Section 6 we compare the first- and second-best solutions in terms of interest rates, equity premia, and real investment levels. Then, our equilibrium results are related to the well-known risk-free rate and equity premium puzzles. Finally, Section 7 briefly summarizes the results of the paper.

2. THE ECONOMY

Consider a risky economy with two dates 0 and 1 and with probability space $(\Omega, P^0, \mathcal{B})$, where $\Omega \equiv \mathcal{R}$, P^0 is a probability distribution function, and $\mathcal{B}$ is the augmented Borel sigma algebra.[3] There are N risky production opportunities in the economy to produce numeraire perishable consumption goods. The sources of uncertainty are from an $(N+1)$-vector of independent standard normal random variables, denoted by $\tilde{B}^0 := (\tilde{B}_1, \ldots, \tilde{B}_N, \tilde{B}_c)^\top$, where $\top$ indicates the transpose. Each production unit is initially owned by a principal (she) who hires an agent (he) for production. We assume that no principal initially owns and no agent works for more than one unit, and that no principal can be an agent, vice versa. In addition, our economy allows all individuals to trade in financial markets where stocks and bonds are traded, as will be seen shortly. Since it takes a lot of symbols for us to describe both product and financial markets complicated with agency problems, a table of notation used in this paper is provided in Appendix A, in case readers may want to refer to it when reading the paper.

We assume all principals and agents have time-additive CARA preferences. In particular, CARA coefficients of principal i and agent i are γ_P and γ_A, respectively. That is, there are N identical principals and N identical agents. Later it shall be seen that the aggregate risk tolerance τ of each firm (or the representative firm) turns out to be important in the determination of asset prices in the financial markets, where

$$\tau := \frac{1}{\gamma_A} + \frac{1}{\gamma_P}.$$

Individual ki's utility U_{ki} of consumption $c^{ki} = (c_0^{ki}, c_1^{ki}) \in \mathcal{R}^2$, for $k = P, A$ and $i = 1, \ldots, N$, is represented by the sum of two CARA utility functions as follows:

$$U_{ki}(c^{ki}) = -\exp\left(-\gamma_k c_0^{ki}\right) - \exp\left(-\gamma_k c_1^{ki}\right).$$

At time 0, each agent i is endowed with $M_A (\in \mathcal{R})$ units of perishable numeraire goods. Each principal i, endowed with $M_P (\in \mathcal{R})$, establishes a firm by making an irreversible investment of $I_i (\geq 0)$ in her production opportunity and hires agent i by signing a contract (C_i, K_i) where C_i is an $\mathcal{B}$-measurable compensation scheme, and $K_i (\subset \mathcal{R}^{N+1})$

[3]The model described in this section is a simplified version of the continuous-time model that is outlined in Appendix G.

denotes a restriction on the agent's financial transactions. More description about K^i will be provided later.

2.1. Product and Financial Markets

The accounting/liquidation value of firm i at date 1 before compensation to the agent, denoted by $\tilde{D}_i$ is given by

$$\tilde{D}_i = D_0 + f(I_i)[\sigma_u \tilde{B}_i + \sigma_c \tilde{B}_c] = D_0 + \Sigma^i \tilde{B}^0, \tag{2.1}$$

where $\sigma_u, \sigma_c > 0$, $D_0 \in \mathcal{R}$, and $\Sigma^i = (0, \ldots, \sigma_u f(I_i), 0, \ldots, 0, \sigma_c f(I_i))$. One may call $\sigma_c \tilde{B}_c$ the market/common risk and $\sigma_u \tilde{B}_i$ the unique risk of firm i. We assume that $\tilde{D}_i$ is public information at date 1, and $f(I) = I^\alpha$ with $\alpha < \frac{1}{2}$. Later it will be seen that the assumption $\alpha < \frac{1}{2}$ ensures the production of cashflows to exhibit decreasing returns to scale in real investment I. Note that the N production units in our economy are identical (in distribution) to each other. In addition, there is a pure financial asset/contract which pays at time 1

$$\tilde{Y} = A + \sigma_Y \tilde{B}_c, \tag{2.2}$$

where $A \in \mathcal{R}$ and $\sigma_Y > 0$. Since it is a financial asset, this asset will be priced to be in zero net supply in equilibrium.

Given compensation schedule C_i, agent i exerts effort $\mu_i (> 0)$ incurring a private cost of $\frac{\delta}{2}\mu_i^2$, where μ_i lies in a compact subset of $\mathcal{R}$. Here, $\delta (> 0)$ is a parameter for agent-effort efficiency: the higher the value of δ, the higher the cost the agent has to incur to increase μ_i. All agents' effort $\{\mu = (\mu_1, \ldots, \mu_N)\}$ jointly changes the probability measure of $\{\tilde{D}_i, i = 1, \ldots, N\}$ from P^0 to P^μ such that[4]

$$\frac{dP^\mu}{dP^0} = \exp\left\{-\frac{1}{2}\|\Sigma^{-1} l\|^2 + (\Sigma^{-1} l)^\top \tilde{B}^0\right\},$$

where $l(\mu) = [f(I_1)\mu_1, \ldots, f(I_N)\mu_N, 0]^\top$,

$$\Sigma = \begin{bmatrix} f(I_1)\sigma_u & 0 & \ldots & 0 & f(I_1)\sigma_c \\ 0 & f(I_2)\sigma_u & 0 & \ldots & f(I_2)\sigma_c \\ \ldots. & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & f(I_N)\sigma_u & f(I_N)\sigma_c \\ 0 & 0 & \ldots & \ldots & \sigma_Y \end{bmatrix},$$

[4]Let $\eta := (\tilde{D}_1 - D_0, \ldots, \tilde{D}_N - D_0, \tilde{Y} - A)^\top$. Then η is multivariate-normally distributed with its probability density function given by $(2\pi\sqrt{\det \Sigma\Sigma^\top})^{-1} \exp[-(1/2)(\eta^\top(\Sigma\Sigma^\top)^{-1}\eta]$. If the distribution is shifted to its new mean $l(\mu)$, (which is equivalent to moving the mean of B^0 by $\Sigma^{-1} l(\mu)$), then, the density function will change to $(2\pi\sqrt{\det \Sigma\Sigma^\top})^{-1} \exp[-(1/2)((\eta - l(\mu))^\top(\Sigma\Sigma^\top)^{-1}(\eta - l(\mu))]$. That is, after the change of its mean, η is distributed with its probability density $(2\pi\sqrt{\det \Sigma\Sigma^\top})^{-1} \exp[-(1/2)(\eta^\top(\Sigma\Sigma^\top)^{-1}\eta]\frac{dP^\mu}{dP^0}$.

and

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{f(I)\sigma_u} & 0 & \dots & 0 & -\frac{\sigma_c}{\sigma_u \sigma_Y} \\ 0 & \frac{1}{f(I)\sigma_u} & 0 & \dots & -\frac{\sigma_c}{\sigma_u \sigma_Y} \\ \dots. & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{f(I)\sigma_u} & -\frac{\sigma_c}{\sigma_u \sigma_Y} \\ 0 & 0 & 0 & \dots & \frac{1}{\sigma_Y} \end{bmatrix}.$$

Then under P_μ,

$$\tilde{D}_i = D_0 + f(I_i)\big[\mu_i + \sigma_u \tilde{B}_i^\mu + \sigma_c \tilde{B}_c\big], \tag{2.3}$$

and $\tilde{B}^\mu = \tilde{B}^0 - \Sigma^{-1} l$ is an $(N+1)$-vector of independent standard normal random variables. Note that given the structure of $\Sigma^{-1} l$, agent i affects the probability distribution of unique risks $\tilde{B}_i$, $i = 1, \dots, N$, but no agents can influence that of the market risk $\tilde{B}_c$. In particular, the distribution of $\tilde{B}_c$ under P^0 remains unaffected under P^μ.

We assume that P^μ is neither observable nor verifiable. (This kind of setup is called the weak formulation. See Schättler and Sung 1993.) Each agent i can observe his own effort level μ_i, but cannot observe other agents' effort levels μ^{-i}, where $\mu^{-i} := (\mu_1, \dots, \mu_{i-1}, \mu_{i+1}, \dots, \mu_N)$. We also assume all agents are Nash game players. Namely, given a contract (C_i, K_i), agent i chooses an effort level μ_i and a financial portfolio process to maximize his expected utility of compensation net of the cost of effort, as if all other agents' decisions had already been made and his decisions would not affect other agents' decisions.

Without loss of generality, we assume that contracts C_i, $i = 1, \dots, N$, are given in the following affine form in $\tilde{B}^0$:[5]

$$C_i = a_i + \sum_{j=1}^{N+1} \beta_{ij} \tilde{B}_j = a_i + (\beta^i)^\top \tilde{B}^0, \tag{2.4}$$

where $a_i \in \mathcal{R}$ is for fixed payment, $(\beta^i)^\top \tilde{B}^0$ is for performance based compensation subject to some noise, and $\beta^i := (\beta_{i1}, \dots \beta_{ic})^\top \in \mathcal{R}^{N+1}$ is an $(N+1)$-vector of contract sensitivities to $(N+1)$ risk sources $\tilde{B}^0$. By convention, we let $\beta_{i,N+1} = \beta_{ic}$, and $\tilde{B}_{N+1} = \tilde{B}_c$.

The date-1 cash flow (2.3) can be interpreted as an outcome that can be affected by both the agent's effort μ and the principal's real investment/project selection decision, I. We believe that this feature of our model is important when we compare asset prices between moral-hazard and classical economies, because moral hazard problems can affect not only the expected dollar-productivity but dollar-risk levels of real assets on which stock prices critically depend. Before describing full-fledged details of contracting

[5]One can show that optimal contracts are indeed linear in a continuous-time model which is analogous to the discrete-time model of this paper. Our continuous-time model in Appendix G incorporates financial markets into the well-known Holmstrom–Milgrom's economy. See Appendix G where the intuition of the linearity is provided.

problems between principals and agents (which shall be given in Section 2.2), we need to understand financial markets.

In addition to product markets described above, there are financial markets where $N+1$ risky assets as well as a bond are traded. The $N+1$ risky assets consist of N stocks and one risky financial asset. The stock market emerges right after principals and agents sign on their contracts, as each principal issues one share to the public based on her residual claim, $\tilde{D}_i - C_i$. The stock price of asset i at time t is denoted by S_t^i, for $i = 1, \ldots, N+1$ and $t = 0, 1$. We assume $S_0^i \geq I$, $i = 1, \ldots, N$, in equilibrium, because otherwise principal i's problem is not meaningful as she decides not to produce.

The market prices of stocks depend on investors' beliefs on agents' optimal effort levels. Let their beliefs be $\{\mu^*\}$ and the one-plus risk-free rate of return be R. We assume that in equilibrium, the beliefs are fulfilled, and prices of all assets traded in financial markets satisfy the following linear pricing rule:[6]

$$S_1^i = RS_0^i + \sum_j^{N+1} \sigma_{ij}^S \theta_j + \sum_j^{N+1} \sigma_{ij}^S \tilde{B}_j^{\mu^*}, \tag{2.5}$$

where θ_j is the market price of risk for the jth risk source, i.e., $\tilde{B}_j^{\mu^*}$; σ_{ij}^S is the dollar volatility of stock i caused by the jth risk source; $\theta_{N+1} = \theta_c$ and $\tilde{B}_{N+1}^{\mu^*} = \tilde{B}_c$ by convention. Note that this linear pricing rule is consistent with the Ross arbitrage pricing model (1976), or the Ross APT. In other words, no-arbitrage implies this linear pricing rule.[7] On the other hand, by the definition of stock, we have $S_1^i = \tilde{D}_i - C_i$.

Let Σ^S denote an $(N+1) \times (N+1)$ dollar-price volatility matrix of $N+1$ risky assets, i.e., the i-th row of Σ^S is $\sigma_i^S = (\sigma_{i1}^S, \ldots, \sigma_{ic}^S)$, $i = 1, \ldots, N+1$. Thus, $S_1^i = RS_0 + \sigma_i^S \theta + \sigma_i^S \tilde{B}^{\mu^*}$, where $\theta := (\theta_1, \ldots, \theta_c)^\top$. Or

$$S_1^i = RS_0^i + \sigma_i^S \tilde{B}^\theta, \tag{2.6}$$

where $\tilde{B}^\theta := \tilde{B}^{\mu^*} + \theta$. This equation is also consistent with the well-known no-arbitrage condition: Under P^θ that makes $\tilde{B}^\theta$ a standard Brownian motion, the expected rate of return on the stock is equal to the risk-free rate, and P^θ itself becomes the risk-neutral probability.

[6]Since we cannot guarantee $D_1 > C_1$ in this paper, the stock price is not necessarily positive. Negative stock prices are frequently found for tractability reasons in the asset-pricing as well as microstructure literature. See, for instance, Wang (1993) and Kyle (1985). If we require $D_1 > C_1$, then optimal C_1 will not be linear, and comparisons between the first and second-best economies will be too complex to generate meaningful economics.

[7]Recall that in the economy, there are $N+1$ risk sources, $(\tilde{B}_j^{\mu^*}, j = 1, \ldots, N+1)$. Let the rate of return on stock i be $\tilde{r}_i$ which can be written as follows: $\tilde{r}_i = r_f + \sum_{j=1}^N (\tilde{r}_{ij} - r_f)$, where r_f is the risk-free rate of return, and $(\tilde{r}_{ij} - r_f)$ is the realized risk premium due to the exposure of the stock to the jth risk $\tilde{B}_j^{\mu^*}$. Let $\tilde{r}_{ij} = r_{ij} + \sigma_{ij}^S \tilde{B}_j^{\mu^*}$ where $r_{ij} = E^{\mu^*}[\tilde{r}_{ij}]$. Then

$$\tilde{r}_i = r_f + \sum_{j=1}^N \frac{(r_{ij} - r_f)}{\sigma_{ij}^S} \sigma_{ij}^S + \sum_{j=1}^N \sigma_{ij}^S \tilde{B}_j^{\mu^*}.$$

By the Ross APT, no-arbitrage implies that $\frac{(r_{ij} - r_f)}{\sigma_{ij}^S} (= \theta_j)$ is equalized across all assets.

2.2. Contracting Problems

Right after capital investments and their initial consumptions, both principals and agents can trade in stocks and a bond in the financial markets. Let $\mathcal{W}_t^{Pi}$, $\mathcal{W}_t^{Ai}$, $t = 0, 1$, be wealth levels of principal i and agent i resulted from financial market transactions at time t. Initial wealth levels for financial market transactions after capital investment I_i and their initial consumptions (c_0^{Pi}, c_0^{Ai}) are as follows:

$$\mathcal{W}_0^{Pi} = M_P - I_i - c_0^{Pi}, \tag{2.7}$$

$$\mathcal{W}_0^{Ai} = M_A - c_0^{Ai}. \tag{2.8}$$

Let $\pi^{ki} = (\pi_{ki1}, \ldots, \pi_{ki,N+1})^\top$, $k = P, A$, be an $(N+1)$-vector of shares of the $N+1$ risky assets held by individual ki in addition to the individual's initial positions on risky assets. For $ki = Ai$, the agent starts with zero risky assets at time 0, and thus π^{Ai} represents his total position on the $N+1$ risky assets at time 1, whereas for $ki = Pi$, the principal initially holds one share of stock i, and thus her total position on the risky assets at time 1 is π^{Pi} plus one share of stock i, i.e., $(\pi_{Pi1}, \ldots, \pi_{Pi,i-1}, \pi_{Pii} + 1, \pi_{Pi,i+1}, \ldots, \pi_{Pi,N+1})^\top$.

Then, the principal's and agent's self-financed wealth at time 1 are

$$\mathcal{W}_1^{Pi} = \mathcal{W}_0^{Pi} R + (\pi^{Pi})^\top \Sigma^S \tilde{B}^\theta, \tag{2.9}$$

$$\mathcal{W}_1^{Ai} = \mathcal{W}_0^{Ai} R + (\pi^{Ai})^\top \Sigma^S \tilde{B}^\theta. \tag{2.10}$$

We allow principal i to impose restrictions on agent i's risky-asset transactions, by choosing sets K_i^k's, $k = 1, \ldots, N+1$.

For the remainder of the paper, we use F and S as superscripts or subscripts on various variables to denote "the first best" and "the second best," whenever clarity is desired. Given capital market variables (R^F, θ^F), principal i's first-best problem is stated as follows:

PROBLEM 1 (The First Best). *Given* $\{\mu^{-i}\}$, *each principal chooses* $(C_i, I_i, c_0^{Pi}, c_0^{Ai}, \mu^i, \pi^{Pi}, \pi^{Ai})$ *and trading restriction* K_i *to*

$$\begin{aligned}
\max \quad & E^\mu\left[-\exp\left\{-\gamma_P c_0^{Pi}\right\} - \exp\left\{-\gamma_P\left(\mathcal{W}_1^{Pi} + \tilde{D}_i - C_i\right)\right\}\right], \\
s.t. \quad & \mathcal{W}_1^{Pi} = \mathcal{W}_0^{Pi} R + (\pi^{Pi})^\top \Sigma^S \tilde{B}^\theta, \quad \mathcal{W}_0^{Pi} = M_P - c_0^{Pi} - I_i, \\
& \mathcal{W}_1^{Ai} = \mathcal{W}_0^{Ai} R + (\pi^{Ai})^\top \Sigma^S \tilde{B}^\theta, \quad \mathcal{W}_0^{Ai} = M_A - c_0^{Ai}, \\
& \pi^{Ai} \in \arg\max_{\tilde{\pi}^{Ai}} \quad E^\mu\left[-\exp\left\{-\gamma_A\left(\mathcal{W}_1^{Ai} + C_i - \frac{\delta}{2}\mu_i^2\right)\right\}\right], \\
& \qquad\qquad s.t. \quad \mathcal{W}_1^{Ai} = \mathcal{W}_0^{Ai} R + (\tilde{\pi}^{Ai})^\top \Sigma^S \tilde{B}^\theta, \quad \mathcal{W}_0^{Ai} = M_A - c_0^{Ai}, \\
& E^\mu\left[-\exp\left\{-\gamma_A c_0^{Ai}\right\} - \exp\left\{-\gamma_A\left(\mathcal{W}_1^{Ai} + C_i - \frac{\delta}{2}\mu_i^2\right)\right\}\right] \geq L.
\end{aligned}$$

The first and second constraints are for self-financing financial wealth levels of the principal and agent in the presence of financial markets, the third is the agent's portfolio choice problem, and the fourth is the participation constraint where the agent's

reservation utility level L is assumed to be exogenously given in the labor market.[8] In this paper, all results hold except Proposition 6.2 for which we assume that L is determined in such a way that the net present value (NPV) of the principal's real investment is equal to zero.

The principal's second best problem is as follows:

PROBLEM 2 (The Second Best). *Given* $\{\mu^{-i}\}$*, each principal chooses* $(C_i, I_i, c_0^{Pi}, c_0^{Ai}, \mu_i, \pi^{Pi}, \pi^{Ai})$ *and trading constraint* K_i *to*

$$
\begin{aligned}
\max \quad & E^{\mu}\left[-\exp\left\{-\gamma_P c_0^{Pi}\right\} - \exp\left\{-\gamma_P\left(\mathcal{W}_1^{Pi} + \tilde{D}_i - C_i\right)\right\}\right], \\
s.t. \quad & \mathcal{W}_1^{Pi} = \mathcal{W}_0^{Pi} R + (\pi^{Pi})^{\top} \Sigma^S \tilde{B}^{\theta}, \quad \mathcal{W}_0^{Pi} = M_P - c_0^{Pi} - I_i, \\
& \mathcal{W}_1^{Ai} = \mathcal{W}_0^{Ai} R + (\pi^{Ai})^{\top} \Sigma^S \tilde{B}^{\theta}, \quad \mathcal{W}_0^{Ai} = M_A - c_0^{Ai}, \\
& (\pi^{Ai}, \mu^i) \in \arg\max_{\tilde{\pi}^{Ai}, \tilde{\mu}_i} \quad E^{\mu}\left[-\exp\left\{-\gamma_A\left(\mathcal{W}_1^{Ai} + C_i - \frac{\delta}{2}\tilde{\mu}_i^2\right)\right\}\right], \\
& \qquad\qquad s.t. \quad \mathcal{W}_1^{Ai} = \mathcal{W}_0^{Ai} R + (\tilde{\pi}^{Ai})^{\top} \Sigma^S \tilde{B}^{\theta}, \quad \mathcal{W}_0^{Ai} = M_A - c_0^{Ai}, \\
& E^{\mu}\left[-\exp\left\{-\gamma_A c_0^{Ai}\right\} - \exp\left\{-\gamma_A\left(\mathcal{W}_1^{Ai} + C_i - \frac{\delta}{2}\mu_i^2\right)\right\}\right] \geq L.
\end{aligned}
$$

The main difference between the first- and second-best problems lies in the third constraint called the incentive compatibility condition which describes how the agent chooses not only his financial portfolio but his effort level. Both the principal and agent are allowed to observe all outcomes $\{D_i, i = 1, \ldots, N\}$ and Y. However, principal i cannot observe and verify (μ_i, μ^{-i}). The reservation utility level L is again exogenously determined in the labor market. In the second best world, if the agent buys out the firm and manages it by himself, the agency problem will disappear. We, however, assume that the agent prefers to stay as an agent rather than buy out the firm. This assumption can be justified if σ_u, the idiosyncratic risk parameter of the firm, is sufficiently high.[9]

[8]Note that the principal's wealth level at time 1 is $\mathcal{W}_1^{Pi} + \tilde{D}_i - C_i$. This does not however mean she is holding the initial real asset until the terminal date. She can always unload the initial real asset position using her financial account $\mathcal{W}_1^{Pi}$. One can equivalently write the principal's terminal utility as $-\exp\{-\gamma_P \mathcal{W}_1^{Pi}\}$ and $\mathcal{W}_0^{Pi} = M_P - c_0^{Pi} - I_i + E^{\theta}[R^{-1}(\tilde{D}_i - C_i)]$, without affecting results in this paper. In fact, in equilibrium, each principal liquidates her real-asset position right after the contract is signed and then holds the market portfolio and the risk-free asset. Consequently, in equilibrium, the diffuse ownership of each stock will arise. This point shall be made clear in Proposition 5.5.

[9]If the agent buys out the firm 100% equity-financed, then it eliminates the agency problem, but results in forcing himself to hold a seriously undiversified portfolio, as his entire asset-portfolio becomes highly exposed to the idiosyncratic risk of one real asset. If σ_u is sufficiently high, then it can be too costly for the agent to hold the undiversified portfolio, as his expected utility level with the 100% equity-financed purchase becomes lower than L. Consequently, it can happen that he prefers not to buy out the firm. If he leverage-buys out the firm, the agent will suffer from an even higher cost of nondiversification due to the well-known financial leverage effect. If he holds a partial ownership, then the same agency problem as in the paper will occur.

2.3. Equilibrium

Market clearing conditions in this paper can be summarized as follows:

$$\sum_{i=1}^{N} \tilde{D}_i = \sum_{i=1}^{N} \left(c_1^{Pi} + c_1^{Ai}\right) \tag{2.11}$$

$$\sum_{i=1}^{N} (\pi^{Pi} + \pi^{Ai}) = 0, \tag{2.12}$$

$$N(M_A + M_P) = \sum_{i=1}^{N} \left(c_0^{Ai} + c_0^{Pi} + I_i + \mathcal{W}_0^{Pi} + \mathcal{W}_0^{Ai}\right). \tag{2.13}$$

Condition (2.11) is to clear the market at time 1 by equating the aggregate realized production to the aggregate consumption; condition (2.12) to clear stock (risky asset) markets; and condition (2.13) to clear the bond market at time 0 by equating the aggregate source to its aggregate use of endowment.[10]

Since $c_1^{Pi} = \mathcal{W}_1^{Pi} + \tilde{D}_i - C_i$ and $c_1^{Ai} = \mathcal{W}_1^{Ai} + C_i$, condition (2.11) implies

$$0 = \sum_{i=1}^{N} \left(\mathcal{W}_1^{Pi} + \mathcal{W}_1^{Ai}\right). \tag{2.14}$$

On the other hand, by the self-financing budget constraints (2.9) and (2.10), and the stock market clearing condition (2.12), we also have

$$\sum_{i=1}^{N} \left(\mathcal{W}_0^{Pi} + \mathcal{W}_0^{Ai}\right) = 0. \tag{2.15}$$

Applying (2.15) to the initial bond market clearing condition, we have

$$N(M_A + M_P) = \sum_{i=1}^{N} \left(c_0^{Ai} + c_0^{Pi} + I_i\right).$$

By the symmetry of all production units in the economy, we have for all i,

$$M_A + M_P = c_0^{Ai} + c_0^{Pi} + I_i. \tag{2.16}$$

In this paper, we examine economies where both product and capital markets simultaneously equilibrate. The equilibria in the paper are defined as follows:

DEFINITION 2.1 (Second-best equilibrium). The tuple $(\theta, R; C_i, \pi^{Pi}, \pi^{Ai}, I_i, \mu_i, i = 1, \ldots, N)$ is called the second-best equilibrium if it satisfies the following conditions.

(i) Securities prices admit no arbitrage opportunities: in particular, securities prices satisfy (2.6) for all $i = 1, \ldots, N + 1$.

[10] A caution is in order. The condition (2.12) does not mean that the net supply of each stock is zero share. Recall that π only counts shares traded in financial markets, without counting initial positions held by principals. If all positions created both in financial and product markets are counted together, the net supply of each stock will turn out to be one share.

(ii) Given (θ_1, R) and $\{\mu^{-i}\}$, principal i chooses $(C_i, I_i, c_0^{Pi}, c_0^{Ai}, \mu_i, \pi^{Pi}, \pi^{Ai})$ to solve Problem *2,*

(iii) Markets clear:

a. The stock market clears, i.e., $\sum_{i=1}^{N}(\pi^{Ai} + \pi^{Pi}) = 0$.
b. The bond market clears, i.e., $M = c_0^{Ai} + c_0^{Pi} + I_i$.

The definition of the first-best equilibrium is the same as Definition 2.1, except that in the first best, principal i solves Problem 1.

3. EXPECTED UTILITY

Before proceeding to our investigation of the first- and second-best equilibria, we express expected utilities of principals and agents in terms of their certainty equivalent wealth levels, which shall be used repeatedly in later sections.

3.1. The Agent's Expected Utility

In the first best, agent i chooses (c_0^{Ai}, π^{Ai}) given μ_i, whereas in the second best, he chooses $(c_0^{Ai}, \pi^{Ai}, \mu_i)$, in order to

$$\max \; E^{\mu}\left[-\exp\left\{-\gamma_A c_0^{Ai}\right\} - \exp\left\{-\gamma_A\left(\mathcal{W}_1^{Ai} + C_i - \frac{\delta}{2}\mu_i^2\right)\right\}\right]$$
$$\text{s.t.} \quad \mathcal{W}_1^{Ai} = R\left(M_A - c_0^{Ai}\right) + (\pi^{Ai})^{\top}\Sigma^S\tilde{B}^{\theta}.$$

Let V_0^{Ai} be the certainty equivalent of the agent's expected utility over future cashflows. Then,

$$-\exp\left\{-\gamma_A V_0^{Ai}\right\} \equiv E^{\mu}\left[-\exp\left\{-\gamma_A\left(\mathcal{W}_1^{Ai} + a_i + (\beta^i)^{\top}\tilde{B}^0 - \frac{\delta}{2}\mu_i^2\right)\right\}\right],$$

and by the participation constraint, we have $V_0^{Ai} = -\frac{1}{\gamma_A}\ln(-L - e^{-\gamma_A c_0^{Ai}})$.

A straightforward application of the moment generating function property of normal random variables to the agent's expected utility reveals that the structure of both the first- and second-best optimal (linear) contracts are necessarily given as follows:

PROPOSITION 3.1. *Both the first- and second-best contracts in equilibrium can be represented in the following form:*

$$C_i = -\frac{1}{\gamma_A}\ln\left(-L - e^{-\gamma_A c_0^{Ai}}\right) - R\left(M_A - c_0^{Ai}\right) + (\beta^i)^{\top}B^0$$
$$+\left[\frac{\delta}{2}\mu_i^2 + \frac{\gamma_A}{2}\|\beta^i + (\Sigma^S)^{\top}\pi^{Ai}\|^2 - (\pi^{Ai})^{\top}\Sigma^S\theta - (\beta^i)^{\top}\Sigma^{-1}l\right]. \tag{3.1}$$

In the first best, this representation is constrained by the following condition from the agent's portfolio choice problem:

$$\pi^{Ai} \in \arg\max_{\tilde{\pi}^{Ai}\in K_i}\left\{-\frac{\gamma_A}{2}\|\beta^i + (\Sigma^S)^{\top}\pi^{Ai}\|^2 + (\pi^{Ai})^{\top}\Sigma^S\theta\right\}. \tag{3.2}$$

In the second best, the representation is constrained not only by condition (3.2), but by the following incentive compatibility condition:

$$\mu_i = \frac{1}{\delta\sigma_u}[\beta^i + (\Sigma^S)^\top \pi^{Ai}]_i, \tag{3.3}$$

where the subscript i denotes the i-th component of the vector.

COROLLARY 3.2. *Suppose that Σ^S is invertible, and that the agent is allowed to freely trade all financial assets including all stocks. Then, the agent's portfolio choice constraint (3.2) implies the following condition:*

$$\frac{\theta}{\gamma_A} = \beta^i + (\Sigma^S)^\top \pi^{Ai}. \tag{3.4}$$

Proposition 3.1 helps simplify Problems 1 and 2, and Corollary 3.2 will be useful when we later discuss an implementability issue. For both the first- and second-best optimal contracting, Proposition 3.1 implies that the principal may, without loss of generality, only consider the class of salary functions given in the form of (3.1), and that the third constraint of Problem 1 can be replaced with expression (3.2), and the same constraint of Problem 2 with two expressions (3.2) and (3.3).

3.2. The Principal's Expected Utility

Let V_0^{Pi} be the principal's certainty equivalent wealth at time 0:

$$-\exp\left(-\gamma_P V_0^{Pi}\right) = E^\mu\left[-\exp\left\{-\gamma_P\left(\mathcal{W}_1^{Pi} + \tilde{D}_i - C_i\right)\right\}\right].$$

Define

$$\xi^{Pi} = (\Sigma^S)^\top \pi^{Pi} - \beta^i,$$
$$\xi^{Ai} = (\Sigma^S)^\top \pi^{Ai} + \beta^i.$$

Then, ξ^{Pi} and ξ^{Ai} denote aggregate risk positions held by the principal and agent, respectively, resulting from their financial portfolios and the compensation contract.

PROPOSITION 3.3. *Given the contract representation in Proposition 3.1, V_0^{Pi} can be expressed as follows:*

$$\begin{aligned} V_0^{Pi} = {} & D_0 + \Sigma^i \Sigma^{-1} l + \frac{1}{\gamma_A}\ln\left(-L - e^{-\gamma_A c_0^{Ai}}\right) + R\left(\mathcal{W}_0^{Pi} + M_A - c_0^{Ai}\right) \\ & - \frac{\gamma_P}{2}\|(\xi^{Pi})^\top + \Sigma^i\|^2 - \frac{\delta}{2}\mu_i^2 - \frac{\gamma_A}{2}\|\xi^{Ai}\|^2 + (\xi^{Pi} + \xi^{Ai})^\top \theta. \end{aligned} \tag{3.5}$$

Note that in the principal's certainty equivalent wealth (3.5), terms involving π^{Pi}, π^{Ai} and β^i in Problems 1 and 2 and Proposition 3.1 are replaced with ξ^{Pi} and ξ^{Ai}. The reason is that in the presence of financial markets, both the principal and agent are concerned with aggregate risks resulting from their portfolio positions and the contract, and their aggregate risk positions are ξ^{Pi} and ξ^{Ai}.

4. THE FIRST BEST

Propositions 3.1 and 3.3 imply that principal i's first-best problem can be restated as follows:

PROBLEM 1′. *Choose* $(c_0^{Pi}, c_0^{Ai}, I_i, \xi^{Pi}, \xi^{Ai}, \mu^i, K^i)$ *to*

$$\begin{aligned}
\max \quad & -\exp\{-\gamma_P c_0^{Pi}\} - \exp\{-\gamma_P^i V_0^{Pi}\}, \\
s.t. \quad & V_0^{Pi} = D_0 + \Sigma^i \Sigma^{-1} l + \frac{1}{\gamma_A} \ln\left(-L - e^{-\gamma_A c_0^{Ai}}\right) + R\left(\mathcal{W}_0^{Pi} + M_A - c_0^{Ai}\right) \\
& \qquad -\frac{\gamma_P}{2}\|(\xi^{Pi})^\top + \Sigma^i\|^2 - \frac{\delta}{2}\mu_i^2 - \frac{\gamma_A}{2}\|\xi^{Ai}\|^2 + (\xi^{Pi} + \xi^{Ai})^\top \theta, \\
& \pi^{Ai} \in \arg\max_{\widetilde{\pi}^{Ai} \in K_i} \left\{-\frac{\gamma_A}{2}\|\beta^i + (\Sigma^S)^\top \widetilde{\pi}^{Ai}\|^2 + (\widetilde{\pi}^{Ai})^\top \Sigma^S \theta\right\}.
\end{aligned}$$

Now, the principal's problem is to choose her aggregate risk positions ξ^{Pi}, rather than choose π^{Pi} and β^i (or C_i) separately, subject to the agent portfolio choice constraint. We first solve this principal's problem ignoring the agent portfolio choice constraint, i.e., imposing no restrictions on K^i. Then, we show that in the first best, the constraint is not binding: That is, no conflicts of interests arise between the principal and agent over the agent's portfolio choice.[11]

The principal's certainty equivalent wealth (3.5) implies that the first-order conditions (FOCs) with respect to μ^i, ξ^{Pi} and ξ^{Ai}, $i = 1, \ldots, N$, to maximize V_0^{Pi} are as follows:

$$\delta \mu_i = f(I_i), \tag{4.1}$$

$$\theta - \gamma_P((\Sigma^i)^\top + \xi^{Pi}) = 0, \tag{4.2}$$

$$\theta - \gamma_A \xi^{Ai} = 0. \tag{4.3}$$

The FOC (4.3) describes the agent's choice of ξ^{Ai} that the principal wishes the agent to choose, and it coincides with the agent's optimal choice condition (3.2), or his FOC (3.4). Thus, in the first best, there does not arise conflicts of interest in the agent's portfolio choice, even when the agent is allowed to freely trade securities in financial markets.

4.1. Market Clearing

As the principal's original problem is transformed into that of choosing (ξ^{Ai}, ξ^{Pi}) among other things, it is necessary to express relevant market clearing conditions in

[11] In the first best, the agent is always committed to exert the effort level as desired/instructed by the principal, as long as the contract satisfies the participation constraint. No incentive or implementability issues, but only participation issues exist. Thus, in the first best, only a risk-sharing problem can remain. A linear contract can serve a risk-sharing role in the absence of financial markets. In the presence of financial markets, first-best contracts do not need to depend on the outcome at all, because the agent can always optimally choose his risk exposure in the financial markets. In particular, constant contracts as well as linear contracts can be optimal when the agent is allowed to trade in financial markets. As can be seen in Proposition 3.1, the contract sensitivity β in the first best can be any real number, as long as it satisfies equations (3.1) and (3.2).

terms of (ξ^{Ai}, ξ^{Pi}). Assuming that Σ_S is invertible, one can see that the market clearing condition $\sum_{i=1}^{N}(\pi^{Ai} + \pi^{Pi}) = \mathbf{0}$ is equivalently restated in terms of ξ's as follows:

$$\sum_{i=1}^{N}(\xi^{Ai} + \xi^{Pi}) = \mathbf{0}, \tag{4.4}$$

where $\mathbf{0}$ is an $(N+1)$-vector of zeros.

Given the principal's FOCs (4.2) and (4.3), the market clearing condition (4.4) immediately implies equilibrium market prices of risks θ are as follows:

PROPOSITION 4.1. *Assume Σ^S is invertible. The stock market uniquely clears when the vector of market prices of risks θ is given as follow:*

$$\theta_k = \begin{cases} \dfrac{f(I_k)\,\sigma_u}{N\tau}, & k \neq N+1 \\ \dfrac{\sigma_c}{N\tau}\displaystyle\sum_{i=1}^{N} f(I_i), & k = N+1. \end{cases} \tag{4.5}$$

REMARK. Using the symmetry and taking the limit as $N \to \infty$, we have

$$\theta_k = \begin{cases} 0, & k \neq N+1 \\ \dfrac{\sigma_c}{\tau} f(I), & k = N+1. \end{cases} \tag{4.6}$$

The zero market prices of risks for $k \neq N+1$ result from the fact that all unique risks are completely diversifiable in the limit as $N \to \infty$.

With θ_k given as in (4.5), the principal's certainty equivalent wealth can be further simplified as follows.

$$V_0^{Pi} = R\big(M_P^i + M_A^i - c_0^{Ai} - c_0^{Pi}\big) + \frac{1}{\gamma_A}\ln\big(-L - e^{-\gamma_A c_0^{Ai}}\big) + G_i^{FB}(I,\theta),$$

$$G_i^{FB}(I,\theta) = D_0 + \left(\frac{\theta^2}{2\gamma_P} + \frac{\theta^2}{2\gamma_A}\right) - RI_i + \frac{1}{2\delta}f^2(I_i) - [f(I_i)\sigma_u\theta_i + f(I_i)\sigma_c\theta_c],$$

and the principal's expected utility now becomes $-e^{-\gamma_P c_0^{Pi}} - e^{-\gamma_P V_0^{Pi}}$. The FOCs for the utility maximization are

$$\exp\left\{-\gamma_P\left(c_0^{Pi} - V_0^{Pi}\right)\right\} = R, \tag{4.7}$$

$$\exp\left\{-\gamma_A\left(c_0^{Ai} - V_0^{Ai}\right)\right\} = R, \tag{4.8}$$

$$\frac{1}{\delta}f(I_i)f'(I_i) = R + f'(I_i)[\sigma_u\theta_i + \sigma_c\theta_c]. \tag{4.9}$$

The first two FOCs simply confirm the well-known fact that the interest rate R in equilibrium is determined in such a way that the marginal rates of substitution between current and certainty equivalent of future consumption are equalized across all principals and agents. The last FOC also confirms that the marginal product equals (one plus) the marginal cost of capital. The RHS is the (one plus) marginal cost of capital which is equal to (one plus) the sum of risk-free rate and marginal risk premium.

Substituting (4.5) back into the FOC (4.9), we have

$$R = \alpha \kappa_F I_i^{2\alpha - 1}, \tag{4.10}$$

where

$$\kappa_F := \frac{1}{\delta} - \frac{\sigma_u^2}{N\tau} - \frac{\sigma_c^2}{\tau}. \tag{4.11}$$

The RHS of equation (4.10) is the (one plus) marginal certainty-equivalent dollar growth rate, which critically depends on parameter κ_F. Since $2\alpha - 1 < 0$, equation (4.10) confirms a simple intuition that the higher the interest rate, the lower the real investment level. This equation will turn out to be very useful in comparing the second-best to the first-best solutions.

However, in order to completely understand the real investment decision in equilibrium, we still need to know equilibrium interest rate R. Recall that market clearing at time 0 implies, by the symmetry, $M_P + M_A = I + c_0^P + c_0^A$ for all i. Then, the market clearing at time 0 and FOCs (4.7) to (4.9) imply that in equilibrium, interest rate R, and the market price of risk θ should be related to each other as in the following proposition:

PROPOSITION 4.2. *In equilibrium, the first-best pair* (R, θ) *satisfies the following market clearing condition:*

$$M_A + M_P - D_0 = I(R, \theta) - \tau \ln(R) + \frac{\kappa_F}{2} f^2(I(R, \theta)), \tag{4.12}$$

where $I(R, \theta) \in \arg\max_{\bar{I}} G^{FB}(\bar{I}; \theta)$*, and* θ *satisfies (4.5).*

Then, condition (4.12) turns out to be the same as that of the single-firm economy in Sung and Wan (2009), except that parameter κ_F is slightly modified. Therefore, implications for the single-firm economy still hold for our many-firms economy. Both equations (4.10) and (4.12) determine the principal's real investment decision in equilibrium. Later, these two equations will be extensively utilized to compare the second-best solutions to the first-best benchmarks. Next, we turn to the second-best case.

5. THE SECOND BEST

In the second best, the principal's certainty equivalent wealth (3.5) is subject to the incentive compatibility constraint $\mu_i = \frac{1}{\delta\sigma_u}\xi_{Aii}$, where ξ_{Aii} denotes the i-th component of ξ^{Ai}. Substituting this constraint into (3.5), we restate the principal's second-best problem as follows:

PROBLEM 2′. *Choose* $(c_0^{Pi}, c_0^{Ai}, I_i, \{\xi^{Pi}, \xi^{Ai}\}, K^i)$ *to*

$$\begin{aligned}
\max \quad & -\exp\left\{-\gamma_P c_0^{Pi}\right\} - \exp\left\{-\gamma_P V_0^{Pi}\right\} \\
s.t. \quad & V_0^{Pi} = D_0 + \Sigma^i \Sigma^{-1} l + \frac{1}{\gamma_A} \ln\left(-L - e^{-\gamma_A c_0^{Ai}}\right) + R\left(\mathcal{W}_0^{Pi} + M_A - c_0^{Ai}\right) \\
& \qquad -\frac{\gamma_P}{2}\|(\xi^{Pi})^\top + \Sigma^i\|^2 - \frac{1}{2\delta\sigma_u^2}(\xi_{Aii})^2 - \frac{\gamma_A}{2}\|\xi^{Ai}\|^2 + (\xi^{Pi} + \xi^{Ai})^\top \theta, \\
& \pi^{Ai} \in \arg\max_{\tilde{\pi}^{Ai} \in K_i} \left\{-\frac{\gamma_A}{2}\|\beta^i + (\Sigma^S)^\top \tilde{\pi}^{Ai}\|^2 + (\tilde{\pi}^{Ai})^\top \Sigma^S \theta\right\}.
\end{aligned}$$

As was the case with the first best, we first ignore the agent's portfolio choice problem. However, unlike the first-best case, we shall show that conflicts of interests arise between the principal and agent about optimal portfolios for the agent, that is, the constraint on the agent portfolio choice is binding. After maximizing V_0^{Pi} without the portfolio choice constraint, or condition (3.2), we find proper restrictions on $K^i(\subset \mathcal{R}^{N+1})$ under which the principal's optimum can be implemented/supported, i.e., under which the agent optimally chooses his portfolios the way the principal wishes him to choose.

The FOCs for the principal's problem 2′ with respect to ξ^{Pi} and ξ^{Ai} are as follows:

$$\xi_{Pik} = \begin{cases} \dfrac{\theta_k}{\gamma_P}, & k \neq i,\ N+1 \\ \dfrac{\theta_i}{\gamma_P} - \sigma_u f(I_i), & k = i \\ \dfrac{\theta_c}{\gamma_P} - \sigma_c f(I_i), & k = N+1, \end{cases} \tag{5.1}$$

$$\xi_{Aik} = \begin{cases} \dfrac{\theta_k}{\gamma_A}, & k \neq i \\ \dfrac{\delta\sigma_u^2}{1+\gamma_A\delta\sigma_u^2}\left(\theta_i + \dfrac{f(I_i)}{\delta\sigma_u}\right), & k = i, \end{cases} \tag{5.2}$$

where ξ_{Pik} is the principal's aggregate risk position on the k-th unique risk $\tilde{B}_k$, and ξ_{Aik} is the agent's aggregate risk position that the principal wants the agent to maintain on the kth unique risk source.

As compared with the agent's desired choice expressed in (3.4), the FOC (5.2) suggests that the principal's optimal choice of ξ_{Aik} is identical to that of the agent for all $k \neq i$, but that for $k = i$, agent i wishes to choose $\xi_{Aii} = \frac{\theta_i}{\gamma_A}$, whereas the principal wants the agent to choose $\xi_{Aii} = \frac{\delta\sigma_u^2}{1+\gamma_A\delta\sigma_u^2}(\theta_i + \frac{f(I_i)}{\delta\sigma_u})$. We shall show in Section 5.2 that the principal's desired ξ_{Aii} can be implemented by imposing a restriction on the agent's stock transactions, and thus that the agent's optimal effort is

$$\mu = \frac{1}{\delta\sigma_u}\xi_{Aii} = \frac{\sigma_u}{1+\gamma_A\delta\sigma_u^2}\left(\theta_i + \frac{f(I_i)}{\delta\sigma_u}\right). \tag{5.3}$$

The structure of the optimal effort shall be discussed in more detail in Section 5.2.2.

5.1. Market Clearing

Again, assuming that Σ^S is invertible, $\sum_{i=1}^N(\pi^{Ai} + \pi^{Pi}) = \mathbf{0}$ is equivalent to $\sum_{i=1}^N(\xi^{Pi} + \xi^{Ai}) = \mathbf{0}$. Applying this market clearing condition to FOCs (5.1) and (5.2), we have equilibrium market prices of risks as follows:

PROPOSITION 5.1. *Assume that Σ^S is invertible and that ξ^{Ai} satisfying (5.2) is implementable, i.e., the agent chooses his portfolio to satisfy (5.2). Then the stock market*

clearance implies that market prices of risks are uniquely given as follows:

$$\theta_k = \begin{cases} \dfrac{\gamma_A^2 \delta \sigma_u^3 f(I^k)}{N\tau\left(\gamma_A\left(1+\gamma_A\delta\sigma_u^2\right)\right)-1}, & k \le N \\ \dfrac{\sigma_c}{N\tau}\displaystyle\sum_{k=1}^{N} f(I^k) & k = N+1. \end{cases} \tag{5.4}$$

The proof is omitted as it is similar to that of Proposition 4.1. Note that the market price of *market risk* is independent of individual firms' productivity parameters. This independence results from our assumption that no individual firms affect the distribution (or the mean) of the market risk. Thus, given I, the market price of market risk becomes identical to that of the first best. Unlike that of the first best, however, the market price of each *unique risk* for finite N depends on corresponding individual firm's productivity parameters δ and $f(I)$, which is consistent with the single-firm economy in Sung and Wan (2009), where the single representative firm affects the risk of the economy. Nevertheless, as $N \to \infty$, all unique risks can be diversified away, and thus all market prices of those *unique risks* approach zero, just as those of the first best do. In this case, the optimal agent effort (5.3) converges to the same level as the one that is already known in the literature for the case where the principal is risk neutral and there are no financial markets, i.e.,

$$\mu_i = \frac{f(I_i)}{\delta\left(1+\gamma_A\delta\sigma_u^2\right)}.$$

We shall discuss the reason for this convergence in Section 5.2.2.

With equations (5.1), (5.2), and (5.4), the principal's certainty equivalent (remaining) wealth right after her initial consumption becomes

$$V_0^{Pi} = R\left(M_P^i + M_A^i - c_0^{Ai} - c_0^{Pi}\right) - V_0^{Ai} + G_i^{SB}(I_i, \theta),$$

where

$$G_i^{SB}(I_i,\theta) = D_0 - RI_i + \frac{\tau}{2}\theta^2 - \sigma_u f(I_i)\theta_i - \sigma_c f(I_i)\theta_c - \frac{\theta_i^2}{2\gamma_A} + \frac{\delta\sigma_u^2}{2\left(1+\gamma_A\delta\sigma_u^2\right)}\left(\theta_i + \frac{f(I_i)}{\delta\sigma_u}\right)^2.$$

Thus, the principal's expected utility at time 0 becomes $-e^{-\gamma_P c_0^{Pi}} - e^{-\gamma_P V_0^{Pi}}$, and the FOCs with respect to $(c_0^{Pi}, c_0^{Ai}, I_i)$ are

$$R = \exp\left\{-\gamma_P\left(c_0^{Pi} - V_0^{Pi}\right)\right\}, \tag{5.5}$$

$$R = \exp\left\{-\gamma_A\left(c_0^{Ai} - V_0^{Ai}\right)\right\}, \tag{5.6}$$

$$R + \sigma_u f'(I_i)\theta_i + \sigma_c f'(I_i)\theta_c = \frac{\sigma_u f'(I_i)}{\left(1+\gamma_A\delta\sigma_u^2\right)}\left(\theta_i + \frac{f(I_i)}{\delta\sigma_u}\right). \tag{5.7}$$

The FOCs (5.5) and (5.6) tell us that the interest rate is the MRS between current and certainty-equivalent future consumption and that the MRS's are equalized across all principals and agents in the economy. The LHS of FOC (5.7) is the (one plus) marginal

cost of capital and the RHS is the second-best marginal product of real investment. This FOC implies

$$R = \alpha \kappa_S I^{2\alpha - 1}, \tag{5.8}$$

where

$$\kappa_S := \frac{1}{\delta} - \frac{\gamma_A \sigma_u^2}{\left(1 + \gamma_A \delta \sigma_u^2\right)} \left[1 + \frac{\gamma_A^3 \delta^2 \sigma_u^6}{\gamma_A N \tau \left(1 + \gamma_A \delta \sigma_u^2\right) - 1} \right] - \frac{\sigma_c^2}{\tau}. \tag{5.9}$$

Parameter κ_S is a measure of the second-best marginal certainty-equivalent dollar growth rate of the principal's wealth.

In Proposition 5.1, market clearing over time $t \in (0, 1]$ implies (5.4). Now, recall that market clearing at time 0 implies, by the symmetry, $M_P + M_A = I + c_0^P + c_0^A$ for all i. Then, with FOCs (5.5) and (5.6), the market clearing at time 0 implies that in equilibrium, the interest rate R, and the market price of risk θ are related to each other as follows:

PROPOSITION 5.2. *In equilibrium, the second best* (R, θ) *satisfy the following market clearing equation:*

$$M_P + M_A - D_0 = I(R, \theta) - \tau \ln(R) + \frac{\kappa_S}{2} f^2(I(R, \theta)). \tag{5.10}$$

where $I(R, \theta) \in \arg\max_I G^{SB}(\bar{I}; \theta)$, *and* θ *satisfies (5.4).*

The principal's second-best real investment decision satisfies both (5.8) and (5.10). Later, we will compare these two equations with corresponding equations (4.10) and (4.12) from the first-best case, to investigate moral hazard effects on the economy.

A comment is in order. The above FOCs and Proposition 5.2 suggest that the equilibrium interest rate R is independent of the agent's reservation utility L. This independence can be somewhat striking because one might conjecture that a change in L might bring about changes in consumption plans of both the principal and agent, affecting the market clearing condition and thus the resulting interest rate in equilibrium. This conjecture is not true in our economy.

In order to obtain an intuition about the independence of the equilibrium interest rate from L, we start with that of (μ, I). Note that the reservation utility level L does not affect the CARA agent's effort choice μ, and thus the marginal productivity of the firm has to be independent of L. The optimal investment level is determined as the marginal productivity equals the marginal cost of capital. Since the marginal productivity depends on I, and the cost of capital is exogenous to both the principal and agent (in our price-taking competitive equilibrium), optimal investment level I has to be independent of L.

To see how R becomes independent of L, recall that given all equity premiums on all investments in the economy, all CARA investors invest fixed dollar amounts in risky assets, independently of L, and all remaining wealth after initial consumption is invested in the risk-free asset, which is determined as the market clears. However, the CARA preference again implies that the market clearing condition, $M_A + M_P = I + c_0^P + c_0^A$ is independent of L, as equations (4.12) and (5.10) suggest.

To see the independence of the market clearing condition, recall that individuals choose current and future certainty-equivalent consumption, satisfying, for $k = P, A$,

$$\exp\left\{-\gamma_k \left(c_0^k - V_0^k\right)\right\} = R.$$

Now, consider a transfer of time 0 wealth Δw from the principal to the agent. Then the agent increases his current and future consumption by Δc_0^A and ΔV_0^A, respectively, such that $\Delta w = \Delta c_0^A + (1/R)\Delta V_0^A$, and

$$\exp\left\{-\gamma_A\left(c_0^{A'} - V_0^{A'}\right)\right\} = R,$$

where $c_0^{A'} := c_0^A + \Delta c_0^A$ and $V_0^{A'} := V_0^A + \Delta V_0^A$. This implies that $\Delta c_0^A = \Delta V_0^A$ and thus $\Delta c_0^A = (R/(1+R))\Delta w$. Note that the wealth transfer leads to a change in L to L' where $L' = -\exp(-\gamma_A c_0^{A'}) - \exp(-\gamma_A V_0^{A'})$.

On the other hand, the principal decreases her wealth by Δw, and her consumption plan by $(\Delta c_0^P, \Delta V_0^A)$ such that $\Delta w = \Delta c_0^P + (1/R)\Delta V_0^P$, and

$$\exp\left\{-\gamma_P\left(c_0^{P'} - V_0^{P'}\right)\right\} = R,$$

where $c_0^{P'} := c_0^P - \Delta c_0^P$ and $V_0^{P'} := V_0^P - \Delta V_0^P$. This also implies $\Delta c_0^P = (R/(1+R))\Delta w$. Thus, the new market clearing condition after the wealth transfer is

$$M_A + M_P = I + c_0^{P'} + c_0^{A'} = I + c_0^P + c_0^A.$$

That is, a change in L does not affect the market clearing condition, and as a result, the risk-free rate becomes independent of L. Hence, we can manage to express all important decisions variables $(\mu, I, R, \theta, \Sigma^S)$ independently of L.[12]

5.2. Implementing the Principal's Optimum

Recall that the original problems 1 and 2 with a set of choice variables $(C_1^i, \mu^i, \pi^{Ai}, \pi^{Pi})$ have been simplified to equivalent problems 1′ and 2′ with a new set of choice variables $(\mu_i, \xi^{Ai}, \xi^{Pi})$. As a result, each principal and agent in the new problems resolve their risk-sharing and incentives issues by choosing (ξ^{Ai}, ξ^{Pi}), whereas they do the same by choosing $(C_1^i, \pi^{Ai}, \pi^{Pi})$ in the original problems. The reason why both formulations remain to be equivalent in spite of the reduction of the number of variables is that each principal and agent are only interested in their own aggregate risk positions resulting from not only their compensation contracts but financial and real asset portfolios.

Nevertheless, there still remains a question of how to implement ξ^{Ai} using original variables (π^{Ai}, β^i). Comparing FOC (5.2) to (3.4), one can see that both the principal and agent agree on the choice of ξ_{Aik}, for all $k \neq i$. The agreement implies that contract sensitivities $\{\beta_{ik}; \forall k, k \neq i\}$ except β_{ii} affect neither the principal nor the agent's expected utility. The reason is that given an arbitrary set of sensitivities $\{\beta_{ik}; \forall k, k \neq i\}$, the agent can always undo them by using his financial market transactions to achieve his desired $\{\xi_{Aik}, \forall k, k \neq i\}$, and the agent does so exactly the way the principal wishes the agent to do.

However, conflicts of interest occur for the agent's aggregate risk position in (the unique risk of) stock i, because the principal's desired position in ξ_{Aii} does not agree with that of the agent, as the comparison of FOC (5.2) to (3.4) suggests. Thus, principal i wishes to figure out a way to make sure her agent chooses ξ_{Aii} as desired by herself. The principal may implement her desired ξ_{Aii} by forcing the agent to choose a particular π^{Ai},

[12]It is well known in the asst pricing literature that θ and Σ^S do not depend on L from the beginning, as they only depend on Σ and risk aversion.

and assigning a contract with β_{ii} such that

$$\frac{\delta\sigma_u^2}{1+\gamma_A\delta\sigma_u^2}\left(\theta_i+\frac{f(I_i)}{\delta\sigma_u}\right)=[(\Sigma^S)^\top\pi^{Ai}]_i+\beta_{ii}.$$

However, in the real world, even when π^{Ai} is observable, forcing the agent to choose particular portfolio positions on all risky assets is practically impossible, partly because the agent financial portfolio is a private property, or partly because actual observation of the agent complete history of all stock transactions may require the principal to incur unjustifiably high monitoring costs.

The monitoring cost problem can be greatly simplified with the following two popular subclasses of compensation schemes for agent i: (1) RPE contracts that typically appear in the standard principal-agent literature, and (2) total outcome-based contracts which are more likely encountered in practice. For the first type,

$$\beta^i=\begin{cases}\beta_{ik}=0, & \text{for } k\neq i,\ N+1,\\ \beta_{ii}=\dfrac{\delta\sigma_u^2}{1+\gamma_A\delta\sigma_u^2}\left(\theta_i+\dfrac{f(I_i)}{\delta\sigma_u}\right).\end{cases}\tag{5.11}$$

Each contract in this class depends at most on $\tilde{B}_i$, the unique risk of his own stock, but neither on unique risks of other firms nor on the market risk. Since $\tilde{B}_i=\frac{1}{\sigma_u f(I)}\{D_1-\sigma_c f(I)\tilde{B}_c\}$, one can view $\tilde{B}_i$ as a measure of the agent performance after adjusting for relative performance. Thus, this type constitutes a class of RPE contracts.

For the second type,

$$\beta^i=\begin{cases}\beta_{ik}=0, & \text{for } k\neq i, \quad N+1,\\ \beta_{ii}=\dfrac{\delta\sigma_u^2}{1+\gamma_A\delta\sigma_u^2}\left(\theta_i+\dfrac{f(I_i)}{\delta\sigma_u}\right),\\ \beta_{ic}=\beta_{ii}\dfrac{\sigma_c}{\sigma_u}.\end{cases}\tag{5.12}$$

In this case, the incentive part of the contract is $\beta_{ii}\tilde{B}_i+\beta_{ic}\tilde{B}_c$ that is equal to $\frac{\beta_{ii}}{\sigma_u f(I_i)}D_1$. As a result, the performance measure is the total outcome D_1 alone, and there is no adjustment for relative performance.

We shall argue that both of the above two types of compensation schemes can be equally optimal as long as the principal properly chooses a pair of (π^{Ai},β^i) to implement the principal's optimal ξ^{Ai}. To describe right pairs (π^{Ai},β^i) for the two classes of compensation schemes, we need to understand the structure of Σ^S, the stock price volatility.

5.2.1. Equilibrium Stock Prices. In this subsection, we examine the stock price volatility when the incentive contract is given with its incentive part constrained to be $\beta_{ii}\tilde{B}_i+\beta_{ic}\tilde{B}_c$ where $\beta_{ii},\beta_{ic}\in\mathcal{R}$. This class encompasses the above-mentioned two subclasses of contracts as special cases. Then, it shall be seen that the structure of Σ^S, the stock-price volatility matrix, is greatly simplified.

PROPOSITION 5.3. *Given a class of contracts with their incentive parts represented in the form of* $\beta_{ii}\tilde{B}_i + \beta_{ic}\tilde{B}_c$ *where* $(\beta_{ii}, \beta_{ic}) \in \mathcal{R}^2$, *the market price of stock* i *is given by*

$$S_1^i = RS_0^i + (\sigma_u I^\alpha - \beta_{ii})\tilde{B}_i^\theta + (\sigma_c I^\alpha - \beta_{ic})\tilde{B}_c^\theta,$$

with the initial stock price being

(5.13)

$$S_0^i = \left(M_A^i - c_0^{Ai}\right) + R^{-1}\left[D_0 + \left(\mu_i^* - \sigma_u\theta_i - \sigma_c\theta_c\right)I^\alpha - V_0^{Ai} - \frac{\delta}{2}\left(\mu_i^*\right)^2 - \frac{\gamma_A}{2}\|\xi^{Ai}\|^2 + \xi^{Ai}\theta\right].$$

Thus, the matrix of volatilities of stocks is given as follows:

$$\Sigma^S = \begin{bmatrix} \sigma_{11}^S & 0 & \dots & 0 & \sigma_{1c}^S \\ 0 & \sigma_{22}^S & 0 & \dots & \sigma_{2c}^S \\ \dots. & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_{NN}^S & \sigma_{N,c}^S \\ 0 & 0 & 0 & \dots & \sigma_{cc}^S \end{bmatrix}, \tag{5.14}$$

where

$$\sigma_{ii}^S := (\sigma_u I^\alpha - \beta_{ii}), \quad \sigma_{ic}^S := (\sigma_c I^\alpha - \beta_{ic}), \quad \textit{and} \quad \sigma_{cc}^S = \sigma_Y.$$

Given a compensation contract with its incentive part given in the form of $\beta_{ii}\tilde{B}_i + \beta_{ic}\tilde{B}_c$ where $(\beta_{ii}, \beta_{ic}) \in \mathcal{R}^2$, the riskiness of residual claim $\tilde{D}_i - C_i$ depends only on two Brownian motions, $\tilde{B}_i$ and $\tilde{B}_c$. Since the stock price of each firm is the present value of its residual claim which is only affected by two risk sources, $\tilde{B}_i$ and $\tilde{B}_c$, it is intuitive that the stock price volatility depends only on the same two risk sources. Thus, the matrix of stock price volatilities is given in a diagonal form as that in (5.14).

5.2.2. Irrelevance of RPE Contracting in the Presence of Financial Markets. Given the characterization of the matrix of stock price volatilities as in Proposition 5.3, we are now ready to discuss how to implement the above-mentioned two types of compensation schemes preserving desired incentives implied by the schemes.

PROPOSITION 5.4. *Suppose a contract is given with its incentive part represented by* $\beta_{ii}\tilde{B}_i + \beta_{ic}\tilde{B}_c$, *where* $\beta_{ii} = \frac{\delta\sigma_u^2}{1+\gamma_A\delta\sigma_u^2}(\theta_i + \frac{f(I_i)}{\delta\sigma_u})$ *and* $\beta_{ic} \in \mathcal{R}$. *Then, in equilibrium, principal* i *can achieve her optimal* ξ^{Ai}, *if and only if* $\pi_{Aii} = 0$, *i.e., she forbids agent* i *to trade stock* i.[13]

REMARK. One can also show that the agent would optimally maintain $\pi_{Aii} = 0$, if he is forbidden to short-sell. Thus, in this paper, the trading restriction $\pi_{Aii} = 0$ is equivalent to the short-selling restriction. See Sung and Wan (2009).

[13]Alternatively, the principal can achieve her optimum by providing a flat contract and requiring the agent to maintain his financial portfolio to be consistent with her optimum, i.e., requiring the agent to hold at least as many shares of stock i as $\pi_{Aii} = \frac{1}{\sigma_u I^\alpha}\frac{\delta\sigma_u^2}{1+\gamma_A\delta\sigma_u^2}(\theta_i + \frac{f(I_i)}{\delta\sigma_u})$. However, forcing the agent to choose a particular portfolio may be realistically too costly.

Proposition 5.4 also implies that when agent i is forbidden to trade stock i, both compensation schemes given in (5.11) and (5.12) are optimal. That is, in the presence of complete financial market, as long as β_{ii}'s of the two schemes are the same, it does not matter to the principal's wealth whether the scheme is given in a RPE form as in (5.11), or just in a total outcome-based form as in (5.12). The reason is that, regardless of β_{ic}, the agent can always undo/restructure the β_{ic} (systematic-risk) part of the compensation contract to his preference by trading financial assets including the $N+1$-st asset. Strikingly, the agent's restructuring effort is also in the best interest of the principal. Thus, it does not matter if the contract depends on $\tilde{B}_c$ or not. This result is in contrast with a partial equilibrium result by Ou-Yang (2005), and provides an explanation of the empirical puzzle presented by Chiappori and Salanié (2003) with their statement that "firms do not seem to use relative performance evaluation of managers very much."

There is another striking implication coming out of Proposition 5.4. Note that the sensitivity β_{ii} in (5.11) and (5.12) can be decomposed into two components, one for risk sharing and the other for incentives. The first component is $\frac{\delta\sigma_u^2}{1+\gamma_A\delta\sigma_u^2}\theta_i$ for risk-sharing, and it directly depends on θ_i, the market price of unique risk of firm i, but not on γ_P, the principal's risk aversion. In the literature, the extent of risk sharing depends on the principal's risk aversion. In the presence of financial markets, however, risk sharing is affected not by principal's risk aversion, but by market prices of risks, because both the principal and agent can undo and restructure their risks at market prices.

The second component is $\frac{\sigma_u f(I_i)}{1+\gamma_A\delta\sigma_u^2}$ for work incentives which the principal would like the contract sensitivity to be, were she risk neutral and were not there financial markets. Note that in our economy, as $N\to\infty$, θ_i approaches zero and so does the first component of the sensitivity. Thus, in our very large economy, the optimal contract sensitivity reduces to the second component. That is, optimal contracting is, in the limit, only concerned with incentives, and the risk sharing role of a contract can be completely ignored. The reason is that in our infinite economy, all idiosyncratic risks are diversifiable, and the risk-sharing of diversifiable idiosyncratic risks in an economy with financial markets is not important.

This implication gives a justification of the frequent assumption in the standard principal-agent literature that the principal is risk neutral. In the literature, the risk-neutrality assumption is commonly defended by the argument, without a proof, that the principal is well diversified, and thus she does not care about idiosyncratic risks. Proposition 5.4 can serve as a proof for the practice.

To recapitulate, in the presence of financial markets, the optimal contract can ignore both important issues in contracting: (1) RPE and (2) risk sharing. The existence of complete financial markets obviates the necessity of not only RPE in optimal contracting, whether the economy is finite or infinite, but also the risk-sharing role of contracts, if the economy is large enough. In particular, β_{ic} affects neither incentives nor risk sharing.

5.2.3. Correlated Idiosyncratic Risks: A Remark. In the last section we have seen that the performance measure for each agent does not have to be of RPE if the outcome is correlated with traded assets. For this conclusion, we have assumed that all idiosyncratic risks are independent of each other. Then, what if idiosyncratic risks are correlated? In particular, what if the two firms i and j had $\tilde{B}_i$ and $\tilde{B}_j$ correlated?

Then we can think of two cases: (1) a case when asset k is traded and (2) the other case when the asset is not traded. In the former case, the same conclusion as in the paper still

applies. That is, the optimal performance measure does not have to be of RPE, because the agent voluntarily hedges the covariance risk in the financial markets. In the latter case, the optimal measure has to be of RPE, as the existing literature suggests.

Suppose that outcomes of the two firms j and k have their idiosyncratic risks $\tilde{B}_j$ and $\tilde{B}_k$ correlated and are given as follows:

$$\tilde{D}_j = \mu_j + \sigma_{jc}\tilde{B}_c + \sigma_{jj}\tilde{B}_j, \quad \text{and} \quad \tilde{D}_k = \mu_k + \sigma_{kc}\tilde{B}_c + \sigma_{kk}\tilde{B}_k,$$

where $\tilde{B}_c$ is the common market risk that is independent of the two idiosyncratic risks. Also suppose that there is a financial asset for $\tilde{B}_c$, but no financial assets are based on the covariance of the two assets. Principals are concerned with $\tilde{D}'_j$ and $\tilde{D}'_k$, where

$$\tilde{D}'_j = \tilde{D}_j - \sigma_{jc}\tilde{B}_c = \mu_j + \sigma_{jj}\tilde{B}_j, \quad \text{and} \quad \tilde{D}'_k = \tilde{D}_k - \sigma_{kc}\tilde{B}_c = \mu_k + \sigma_{kk}\tilde{B}_k.$$

In this case, the optimal performance measure for firm j is $\tilde{D}'_j - b\tilde{D}'_k$, where $b = \frac{\mathrm{COV}(\tilde{D}'_j, \tilde{D}'_k)}{\mathrm{VAR}(\tilde{D}'_k)}$, which is the variance minimizing performance measure. See Koo, Shim, and Sung (2008). Of course this measure is a RPE metric. If asset k is traded, then agent j will voluntarily minimize the compensation risk by taking an additional short position in βb shares of asset k, where β is the contract sensitivity to $\tilde{D}'_j$ or $\tilde{D}_j$. The reason is that that even under moral hazard, the risk averse agent is always interested in minimizing risks if he can do it at a zero NPV. See Sung (1995, proposition 3).

5.3. Equilibrium Portfolios

The next question is: given optimal contracts and proper restrictions on agents' trading activities, how would principals and agents manage in general their financial portfolios?

PROPOSITION 5.5. *Suppose that each contract i, $i = 1, \ldots, N$, is given with its incentive part represented by $\beta_{ii}\tilde{B}_i + \beta_{ic}\tilde{B}_c$, where $\beta_{ii} = \frac{\delta\sigma_u^2}{1+\gamma_A\delta\sigma_u^2}(\theta_i + \frac{f(I_i)}{\delta\sigma_u})$ and $\beta_{ic} \in \mathcal{R}$. Then, principal i's and agent i's optimal portfolios are as follows:*

$$\pi_{Pik} = \begin{cases} \dfrac{1}{\sigma_{kk}^S}\dfrac{\theta_k}{\gamma_P} & k \neq i, N+1 \\ \dfrac{1}{\sigma_{ii}^S}\dfrac{\theta_i}{\gamma_P} - 1 & k = i \\ \dfrac{1}{\sigma_Y^S}\left\{\left[\dfrac{\theta_c}{\gamma_P} - (\sigma_c f(I) - \beta_{ic})\right] - \displaystyle\sum_{j=1}^{N}\pi_{Pij}\sigma_{jc}^S\right\} & k = N+1, \end{cases}$$

$$\pi_{Aik} = \begin{cases} \dfrac{1}{\sigma_{kk}^S}\dfrac{\theta_k}{\gamma_A} & k \neq i, N+1 \\ 0 & k = i \\ \dfrac{1}{\sigma_Y^S}\left\{\left(\dfrac{\theta_c}{\gamma_A} - \beta_{ic}\right) - \displaystyle\sum_{j=1}^{N}\pi_{Aij}\sigma_{jc}^S\right\} & k = N+1. \end{cases}$$

Since the symmetry implies $\theta_i = \theta_k, i, k = 1, \ldots, N$, Proposition 5.5 tells us how both principals and agents manage their financial portfolios in financial markets. After signing an optimal compensation contract, principal i optimally liquidates her entire real-asset position and holds the market portfolio, i.e., an equally weighted portfolio of N stocks. Consequently, the presence of financial markets naturally induces dispersed ownership of public firms. In particular, while holding the market portfolio, principal i optimally balances her exposure to the market risk $\tilde{B}_c$ relative to exposures to unique risks $\{\tilde{B}_i; i = 1, \ldots, N\}$ by using the $(N+1)$-st asset, the pure financial asset in the economy.

On the other hand, agent i also uses the same financial asset to adjust his exposure to the aggregate market risk resulting not only from his stock portfolio but from his compensation contract. In particular, whatever β_{ic} is given, the agent always hedges away the risk related to β_{ic} in financial markets until he achieves the optimal level of exposure to the market risk. As a result, the agent's optimal exposure to the market risk ξ_{Aic} remains the same regardless of β_{ic}. This agent behavior is consistent with implications of Proposition 5.4, i.e., the irrelevance of the RPE in the presence of financial markets.

Moreover, the agent behavior on his portfolio formulation can be related to the literature on portfolio choice problems in the presence of unhedgable labor income risks. Our result on the agent's trading restriction is precisely consistent with the assumption by Heaton and Lucas (1996) that "agents cannot write contracts contingent on future labor income." Under their agents' trading restriction, Heaton and Lucas argue for "a sizable equity premium only if transactions costs are larger or the assumed quantity of tradable assets" is limited. In our model, the presence of labor income risk by itself gives no effect on the equity premium. Thus, one can guess that the sizable equity premium in Heaton and Lucas may have nothing to do with unhedgable labor income risks, and more to do with other assumptions in their model such as short-sale constraints and transactions costs.

6. COMPARISON OF FIRST- AND SECOND-BEST INTEREST RATES, EQUITY PREMIA, AND REAL INVESTMENT

Given our characterizations of principals' and agents' optimal consumption and investment decisions, market clearing conditions can shed light on popular issues in the literature: interest rates and equity premia. Sung and Wan (2009) investigate equilibrium effects of moral hazard on these financial market variables in a single-firm economy. In this section, we show that most of the results for the single-firm economy hold even for our multifirm economy, whether the economy is composed of a finite or infinite number of firms.

By Propositions 4.2 and 5.2, the first- and second-best interest rates and time 0 market clearing conditions lead to the following relationships: for $j = F,\ S$,

$$R^j = \alpha\kappa_j I_j^{2\alpha-1}, \tag{6.1}$$

$$M_P + M_A - D_0 = I_j - \tau \ln(R^j) + \frac{\kappa_j}{2} f^2(I_j). \tag{6.2}$$

Rewriting this market clearing condition in terms of R, we have

$$M - D_0 = \left(\frac{R}{\alpha\kappa}\right)^{\frac{1}{2\alpha-1}} \left[1 + \left(\frac{R}{2\alpha}\right)\right] - \tau \ln(R).$$

Differentiating R with respect to κ, one can immediately see that $\frac{\partial R}{\partial \kappa} > 0$. Since $\kappa_S < \kappa_F$, we have $R^S < R^F$. We summarize this result as below.

PROPOSITION 6.1. *In our multifirm economy, the second-best interest rate is lower than that of the first best.*

Note that the productivity of the second-best economy is lower than that of the first best. Thus, in anticipation of poor productivity, the marginal utility for future consumption in the second best becomes high and investors would like to save for the future, increasing the supply of capital which in turn drives the interest rate down.

On the other hand, the same market clearing condition in terms of I can be restated as follows:

$$M - D_0 + \tau \ln(\alpha\kappa) = I + \tau(1 - 2\alpha)\ln(I) + \frac{\kappa}{2} I^{2\alpha}. \tag{6.3}$$

This condition implies that the market clearing investment level I is a function of $(\kappa; \alpha, \tau, M, D_0)$ given (α, τ, M, D_0). Note that given (α, τ, M, D_0), function $I(\kappa; \alpha, \tau, M, D_0)$ is strictly concave in κ, and there exist a unique κ^* such that I achieves its maximum at $\kappa = \kappa^*$ where

$$\kappa^* := \frac{2\tau}{I^{2\alpha}(\kappa^*)}.$$

Since $\kappa_S < \kappa_F$, we have the following conclusion.

PROPOSITION 6.2. *In our multifirm economy, if $\kappa_F \le \kappa^*$, then $I_S < I_F$; and if $\kappa_S \ge \kappa^*$, then $I_S > I_F$.*

That is, moral hazard can result in underinvestment (overinvestment) problems if profitability of production opportunities is sufficiently low (high). It is striking that pure moral hazard problems can result in overinvestment, even in the absence of the well-known empire building motivation proposed by Jensen (1986). In this paper, note that the level of investment is affected not only by the profitability of production opportunities but by the cost of capital, and the second-best cost of capital can be lower than that of the first best. Overinvestment problems can occur in general equilibrium when the poor profitability can be overcome by the lower cost of capital.

Finally, in order to compare equity premia of market portfolios in the first- and second-best economies, note first that the expected terminal prices of stock i in both economies are given in the following form:

$$E^{\mu}[S_1] = RS_0 + [(\sigma_u I^{\alpha} - \beta_{ii})\theta_i + (\sigma_c I^{\alpha} - \beta_{ic})\theta_c].$$

Thus, the risk premium on stock i is

$$\frac{E^{\mu}[S_1]}{S_0} - R = \frac{(\sigma_u I^{\alpha} - \beta_{ii})\theta_i + (\sigma_c I^{\alpha} - \beta_{ic})\theta_c}{S_0}. \tag{6.4}$$

We assume that the NPV of the production is zero in equilibrium, i.e., $S_0 = I$. At time 0, the market portfolio is an equally weighted portfolio of N stocks. Suppose that all compensation contracts are of RPE such that $\beta_{ic} = 0, \forall i$. Then, in our infinite economy, the equity premium (on the market portfolio) denoted by v_0^M is, for $k = F, S$,

$$v_0^{M,k} = \sigma_c I^{\alpha-1}\theta_c = \frac{\gamma_P \gamma_A}{\gamma_P + \gamma_A} \sigma_c^2 I_k^{2\alpha-1}.$$

PROPOSITION 6.3. *Suppose that* $S_0 = I$, $\beta_{ic} = 0$, $\forall i$, *and* $N \to \infty$. *Then*

$$\nu_0^{M,S} > (=, <)\, \nu_0^{M,F}, \quad \text{if and only if} \quad I_S < (=, >) I_F.$$

Although firms in this paper can never directly affect the systematic risk, (collective) moral hazard occurring inside firms can contribute to an increase or a decrease in the equity premium of the whole economy, because moral hazard affects real investment levels and thus the aggregate volatility of the economy. In particular, *the equity premium in our infinite economy turns out to be equal to a risk premium on marginal market risk from real investment*. Since the marginal market risk $\alpha\sigma_c I^{\alpha-1}$ decreases with the amount of real investment, one can easily see that the second-best equity premium becomes higher (lower) than that of the first best if and only if there are underinvestment (overinvestment) problems in the second best.

6.1. On the Risk-free Rate and Equity Premium Puzzles

Note that there is a subtle difference between economic situations leading to the above equilibrium effects, and to the famous Mehra and Prescott's (1985) equity premium and Weil's (1989) risk-free rate puzzles. Recall that the puzzles occur because classical asset pricing models with reasonable levels of the representative agent's risk aversion are not able to justify observed high equity premia and low risk-free rates for the observed volatilities of the market portfolio returns. That is, authors in the literature are concerned with the level of the representative investor's risk aversion, given the equity premium, the risk-free rate and the volatility of the market portfolio, whereas our equilibrium effects in the last section are concerned with the equilibrium equity premium, risk-free rate and volatility, given levels of risk aversion of all economic agents, i.e., all principals and agents.

Recast in terms of economic settings in the literature on the two puzzles, our equilibrium effects imply that moral hazard can help explain the low risk-free rate puzzle, but it cannot contribute to the resolution of the equity premium puzzle. To see this, note that for $k = F, S$, as $N \to \infty$, $\nu_0^{M,k} = \sigma_M^k \theta_c^k$, where σ_S^k is the average return volatility of the market portfolio over the whole period $[0, 1]$ in the kth best economy which we define as follows:

$$\sigma_S^k \equiv \sqrt{\operatorname{var}\left(\frac{S_1^k - S_0^k}{S_0^k}\right)}.$$

Suppose that one observes σ_M, $S_0 (= I)$ and R, respectively, as the volatility of the market portfolio, the stock price, and the risk-free rate. Then by (6.2), for $j = F, S$, $R = \alpha\kappa_j I^{2\alpha-1}$. Thus, both the first- and second-best economies support the same observations (σ_M, I, R), then we must have $\kappa_F = \kappa_S$. In order for this equality to hold, equations (4.11) and (5.9) imply

$$\frac{\sigma_c^2}{\tau_F} = \frac{\gamma_A^S \sigma_u^2}{\left(1 + \gamma_A^S \delta \sigma_u^2\right)} + \frac{\sigma_c^2}{\tau_S}.$$

This equation immediately implies that the second-best aggregate risk tolerance is higher than that of the first best, i.e., $\tau_S > \tau_F$. (Note that the aggregate risk tolerance is analogous

to the reciprocal of the representative investor's risk aversion in the classical asset pricing literature.) That is, in order to justify observations of (σ_M, $S_0(= I)$, R), the second-best economy requires a lower level of risk aversion than the first-best economy does. This result suggests that *moral hazard can contribute to the resolution of the risk-free rate puzzle*.

Recall that the interest rate in equilibrium is the MRS between current and future certainty-equivalent consumption. However, holding other things constant, we know that the second-best risk-free rate is lower than the first best. Now suppose that both economies are identical except for the existence of moral hazard and the principal's risk aversion level, and that the risk-free rates of the first- and second-best economies are the same. Then, one can see that it is necessary to have the aggregate risk aversion level in the second best be lower than that of the first best, in order to equalize both economies in terms of the MRS. (Note that given a plan for current and certainty-equivalent future consumption, the higher the risk aversion, the lower the interest rate.)

In order to relate our equilibrium results to the equity premium puzzle, suppose that one observes σ_M, $S_0(= I)$ and ν_0, respectively, as the volatility of the market portfolio, the stock price, and the risk premium. If the economy is under moral hazard, then $\nu_0 = \sigma_M \theta_c^S$ and by equation (5.4),

$$\frac{\nu_0}{\sigma_M} = \frac{\gamma_A^S \gamma_P^S}{\gamma_A^S + \gamma_P^S} \sigma_c f(I),$$

where γ_P^S and γ_A^S are, respectively, the risk aversion levels of principals and agents supporting the second-best economy characterized by σ_M, $S_0(= I)$ and ν_0. If the economy is free from moral hazard, then by equation (4.5),

$$\frac{\nu_0}{\sigma_M} = \frac{\gamma_A^F \gamma_P^F}{\gamma_A^F + \gamma_P^F} \sigma_c f(I),$$

where γ_P^F and γ_A^F are supporting risk aversion levels of principals and agents in the first best.

Therefore, it is clear that the risk aversion levels supporting the observations must be the same for both the first- and second-best economies. That is, *moral hazard cannot contribute to the resolution of the equity premium puzzle*. The intuition behind this result is straightforward. If the volatilities of the market portfolio returns were the same for both the first- and second-best economies, i.e., $\sigma_M^F = \sigma_M^S = \sigma_M$, then investors would price the market risk in exactly the same way as they do in the first best, independently of moral hazard occurring in product markets.

Note that the inability of moral hazard alone to explain the equity premium puzzle is in contrast with results in the literature. See Kahn (1990) and Kocherlakota and Pistaferri (2009) for moral hazard resulting in a high equity premium, and Kocherlakota (1998) for moral hazard causing a low equity premium.

7. CONCLUSION

We have presented a general equilibrium model of an economy with many firms in the presence of moral hazard and financial markets. The economy distinguishes itself from

the single-firm economy of Sung and Wan (2009) in that each firm can exert zero influence on (the distribution of) market risks of the economy, although it affects that of a unique risk.

We have argued that in the presence of financial markets, RPE is irrelevant in optimal contracting, because given any compensation contracts, agents would optimally adjust through financial transactions their exposures to risks that are not related to their job performance, and their transactions do not affect principals' wealth.

Our model has enabled us to reexamine the fundamental nature of optimal contracting decisions. We have shown that in an economy with infinitely many firms, optimal contracting decisions are not tradeoffs between risk sharing and incentives, but all about incentives. That is, no consideration for risk sharing is necessary, because desired risk sharing is achieved by all individuals in the economy on their own accounts.

Moreover, in our multifirm economy, the second-best risk-free rate is lower than that of the first best. The second-best equity premium can be either higher or lower than that of the first best, because not only underinvestment but overinvestment problems can occur in the second best. These comparisons result from comparing a production economy with financial market in the absence of moral hazard with the same economy in the presence of moral hazard. (Note that with moral hazard introduced into the economy, equilibrium investment levels changes and so do equilibrium volatilities of assets.) These results are about effects of moral hazard given a production economy with financial markets, but they are not about the well-known risk-free rate and equity premium puzzles.

Recall that, for the equity premium puzzle in the literature, researchers take the two important economic factors as exogenously given, namely, the volatility of market portfolio returns and equity premium, and then they compute the risk aversion level of the representative agent to make the two exogenously specified factors justified in equilibrium. However, in our moral-hazard economy, all volatilities are endogenously determined. In order to relate our results to the puzzle, we have recast our moral-hazard economy into the economic setting considered in the literature for the puzzle. That is, we have considered the first- and second-best economies which yield the same (hypothetically) observed levels of equity premium and market-portfolio return volatility, and compared first-best and second-best levels of risk aversion which can support the observed equity premium and market-portfolio return volatility in equilibrium.

We have shown that the same observed equity premium and volatility can only be supported by the same level of the risk aversion, whether the economy is under moral hazard or not. The reason is that investors would price the market risk in the exactly the same way as they do in the first best, independently of moral hazard occurring in product markets. Thus, we have argued that moral hazard cannot help explain the equity premium puzzle, which is in contrast to what the literature suggests.

On the other hand, we have shown that moral hazard can contribute to the resolution of the risk-free rate puzzle, because given observations of the volatility and risk-free rate, the second-best economy requires a lower level of risk aversion to justify the observations in equilibrium than does the first-best counterpart. The reason is that holding other things constant including risk aversion, the marginal utility for future consumption in the second best is higher than that of the first best, because of the anticipation of poor productivity in the second best.

APPENDIX A: TABLE OF NOTATION

$\tilde{B}^n := (\tilde{B}^n_1, \ldots, \tilde{B}^n_N, \tilde{B}^n_c)^\top$, $n = 0, \mu, \theta$, $(N+1)$-vector of independent standard normal random variables under P^n.

C_i, managerial compensation of firm i.

$\tilde{D}_i$, accounting value of firm i at time 1.

I_i, initial investment in firm i.

K_i, agent trading restriction.

M_k, $k = P, A$, endowments.

P^0, P^μ, P^θ, probability measures under which $\tilde{B}^0$, $\tilde{B}^\mu$ and $\tilde{B}^\theta$, resp., are $(N+1)$-vectors of independent standard normal random variables.

R, one plus risk-free rate.

S^i_0, S^i_1, financial-market prices of stock i at time 0 and 1, resp.

$\mathcal{W}^{ki}_t$, the financial account balance of individual ki at time t, $k = P, A$, $i = 1, \ldots, N$.

c^{ki}_t, individual ki's consumption at time t, $k =, P, A$, $i = 1, \ldots, N$.

$l(\mu) := [f(I_1)\mu_1, \ldots, f(I_N)\mu_N, 0]^\top$.

α, production (decreasing) returns-to-scale parameter ($<\frac{1}{2}$).

γ_k, $k = P, A$, CARA coefficients for principals and agents, respectively.

$\tau := 1/\gamma_A + 1/\gamma_P$, the aggregate risk tolerance of each firm.

σ_u, an idiosyncratic-risk volatility parameter of each firm.

σ_c, a systematic-risk volatility parameter of each firm.

σ_Y, a systematic-risk volatility parameter of the pure financial asset Y.

$\Sigma^i = (0, \ldots \sigma_u I^\alpha_i, 0, \ldots, 0, \sigma_c I^\alpha_i)$, firm i's production volatility vector.

$\Sigma := [\Sigma^1, \ldots, \Sigma^{N+1}]^\top$.

σ^S_{ij}, stock i's price volatility atributable to the jth risk source, $\tilde{B}^{\mu^*}_j$.

Σ^S_i, Stock i's $(N+1)$-vector of price volatilities from $(N+1)$ risk sources,

$\Sigma^S = (\Sigma^S_1, \ldots, \Sigma^S_N, \Sigma^S_Y)^\top$, a $(N+1) \times (N+1)$ matrix.

μ_i, agent i's effort level.

δ, agent's effort efficiency parameter.

$\beta^i := (\beta_{i1}, \ldots., \beta_{iN}, \beta_{i,N+1})^\top$, $\beta_{i,N+1} \equiv \beta_{ic}$, contract sensitivity vector.

θ_j, the market price of risk for the jth risk source, $\tilde{B}^{\mu^*}_j$.

$\theta = (\theta_1, \ldots, \theta_N, \theta_c)^\top$.

$\pi^{ki} = (\pi_{ki1}, \ldots, \pi_{kiN}, \pi_{ki,N+1})$, $\pi_{ki,N+1} \equiv \pi_{kic}$, individual ki's financial portfolio position.

$\xi^{ki} = (\xi_{ki1}, \ldots, \xi_{kiN}, \xi_{kic})^\top$ $k = P, A$, individual k's aggregate risk exposure.

$\nu^{M,k}_0$ $k = F, S$, equity premium on the market portfolio. F: first best. S: second best.

APPENDIX B: PROOF OF PROPOSITION 3.1

By substitution,

$$-\exp\{-\gamma_A V_0^{Ai}\} \equiv E^{\mu}\left[-\exp\left\{-\gamma_A\left(R\big(M_A - c_0^{Ai}\big) + (\pi^{Ai})^{\top}\Sigma^S B^{\theta} + a_i + (\beta^i)^{\top}\tilde{B}^0 - \frac{\delta}{2}\mu_i^2\right)\right\}\right],$$

Note that

$$\begin{aligned} &R\big(M_A - c_0^{Ai}\big) + (\pi^{Ai})^{\top}\Sigma^S B^{\theta} + a_i + (\beta^i)^{\top}\tilde{B}^0 - \frac{\delta}{2}\mu_i^2 \\ &= R\big(M_A - c_0^{Ai}\big) + a_i - \frac{\delta}{2}\mu_i^2 + (\pi^{Ai})^{\top}\Sigma^S\big(\Sigma^{-1}(l - l^*) + \theta\big) \\ &\quad + (\beta^i)^{\top}\Sigma^{-1}l + ((\pi^{Ai})^{\top}\Sigma^S + (\beta^i)^{\top})B^{\mu}. \end{aligned}$$

Thus,

$$\begin{aligned} V_0^{Ai} = {} & R\big(M_A - c_0^{Ai}\big) + a_i - \frac{\delta}{2}\mu_i^2 + (\pi^{Ai})^{\top}\Sigma^S\{\Sigma^{-1}(l - l^*) + \theta\} \\ & + (\beta^i)^{\top}\Sigma^{-1}l - \frac{\gamma_A}{2}\|(\pi^{Ai})^{\top}\Sigma^S + (\beta^i)^{\top}\|^2. \end{aligned} \tag{B.1}$$

The agent chooses π^{Ai} in the first best, and (μ_i, π^{Ai}) in the second best, to maximize V_0^{Ai}. In particular, in the second best,

$$\begin{aligned} (\mu_i, \pi^{Ai}) \in \arg\max_{\tilde{\mu}, \tilde{\pi}^{Ai}} & -\frac{\delta}{2}\mu_i^2 + (\pi^{Ai})^{\top}\Sigma^S(\Sigma^{-1}(l - l^*) + \theta) \\ & + (\beta^i)^{\top}\Sigma^{-1}l - \frac{\gamma_A}{2}\|(\pi^{Ai})^{\top}\Sigma^S + (\beta^i)^{\top}\|^2, \end{aligned}$$

which is presented in expressions (3.2) and (3.3). On the other hand, since $l = l^*$ in equilibrium, the agent's certainty equivalent (B.1) implies that

$$a_i = V_0^{Ai} - R\big(M_A - c_0^{Ai}\big) + \frac{\delta}{2}\mu_i^2 - (\pi^{Ai})^{\top}\Sigma^S\theta - (\beta^i)^{\top}\Sigma^{-1}l + \frac{\gamma_A}{2}\|(\pi^{Ai})^{\top}\Sigma^S + (\beta^i)^{\top}\|^2,$$

and thus that C_i is given as in (3.1). □

APPENDIX C: PROOF OF PROPOSITION 4.1

By the market clearing condition (4.4), the FOCs (4.2) and (4.3) imply

$$N\left(\frac{1}{\gamma_A} + \frac{1}{\gamma_P}\right)\theta = \sum_{i=1}^{N}(\Sigma^i)^{\top} = \begin{bmatrix} f(I_1)\sigma_u \\ f(I_2)\sigma_u \\ \vdots \\ f(I_N)\sigma_u \\ \sigma_c\sum_{i=1}^{N} f(I_i) \end{bmatrix},$$

and thus the statement follows. □

APPENDIX D: PROOF OF PROPOSITION 5.3

Recall that with $\beta^{ik} = 0$ for all $k \neq i$, $N+1$, each contract can be represented in the following form:

$$C_i = V_0^{Ai} - R\big(M_A^i - c_0^{Ai}\big) + \beta_{ii}\tilde{B}_i + \beta_{ic}\tilde{B}_c + \frac{\delta}{2}\mu_i^2 + \frac{\gamma_A}{2}\|\xi^{Ai}\|^2 - \xi^{Ai}\theta + (\beta^i)^\top\theta - \frac{\beta_{ii}\mu_i}{\sigma_u},$$

where V_0^{Ai} is a constant satisfying the participation constraint. Note that by the definition of stock i, $S_1^i = D_1^i - C_1^i$ and that

$$\begin{aligned}
\tilde{D}_i - C_i &= D_0 + (\sigma_u f - \beta_{ii})\tilde{B}_i + (\sigma_c f - \beta_{ic})\tilde{B}_c - V_0^{Ai} + R\big(M_A^i - c_0^{Ai}\big) \\
&\quad - \left[\frac{\delta}{2}\mu_i^2 + \frac{\gamma_A}{2}\left\|\xi^{Ai}\right\|^2 - \xi^{Ai}\theta + (\beta^i)^\top\theta - \frac{\beta_{ii}\mu_i}{\sigma_u}\right], \\
&= D_0 + (\sigma_u f - \beta_{ii})\left(\tilde{B}_i^{\mu^*} + \frac{\mu_i^*}{\sigma_u}\right) + (\sigma_c f - \beta_{ic})\tilde{B}_c^{\mu^*} - V_0^{Ai} + R\big(M_A^i - c_0^{Ai}\big) \\
&\quad - \left[\frac{\delta}{2}\mu_i^2 + \frac{\gamma_A}{2}\|\xi^{Ai}\|^2 - \xi^{Ai}\theta + (\beta^i)^\top\theta - \frac{\beta_{ii}\mu_i}{\sigma_u}\right].
\end{aligned}$$

Comparing this with the no-arbitrage pricing rule (2.5), we must have $\sigma_{ij}^S \equiv 0$ for $j \neq i$, $N+1$, and

$$\sigma_{ii}^S = (\sigma_u I^\alpha - \beta_{ii}), \quad \text{and} \quad \sigma_{i,N+1}^S = (\sigma_c I^\alpha - \beta_{ic}).$$

Similarly, one can show that $\sigma_{N+1,N+1}^S = \sigma_Y$.

Moreover, by the pricing rule,

$$E^{\mu^*}\big[S_1^i\big] = RS_0^i + (\sigma_u I^\alpha - \beta_{ii})\theta_i + (\sigma_c I^\alpha - \beta_{ic})\theta_c.$$

However,

$$\begin{aligned}
E^{\mu^*}[\tilde{D}_i - C_i] &= D_0 + (\sigma_u f - \beta_{ii})\frac{\mu_i^*}{\sigma_u} - V_0^{Ai} + R\big(M_A^i - c_0^{Ai}\big) \\
&\quad - \left[\frac{\delta}{2}(\mu_i^*)^2 + \frac{\gamma_A}{2}\left\|\xi^{Ai}\right\|^2 - \xi^{Ai}\theta + (\beta^i)^\top\theta - \frac{\beta_{ii}(\mu_i^*)}{\sigma_u}\right].
\end{aligned}$$

Therefore, the initial price of stock i is given as in (5.13). □

APPENDIX E: PROOF OF PROPOSITION 5.4

FOC (5.2) implies that the principal wishes the agent to choose his portfolio π^{Ai} such that $\xi_{Aii} = \beta_{ii}$, whereas the agent FOC (3.4) implies the agent wants to choose his portfolio π^{Ai} such that $\xi_{Aii} = \frac{\theta_i}{\gamma_A}$. However, since $\beta_{ii} \neq \frac{\theta_i}{\gamma_A}$ in equilibrium, the conflict between the principal and agent must unavoidably arise, if the agent is allowed to freely trade all stocks. Thus, in order to achieve her optimum, it is necessary for the principal to require the agent to choose π^{Ai} such that $[(\Sigma^S)^\top\pi^{Ai}]_i = 0$, because $\xi_{Aii} = \beta_{ii} + [(\Sigma^S)^\top\pi^{Ai}]_i$. However, by Proposition 5.3, $[(\Sigma^S)^\top\pi^{Ai}]_i = (\sigma_u I^\alpha - \beta_{ii})\pi_{Aii}$. Thus, the principal wishes to have $(\sigma_u I^\alpha - \beta_{ii})\pi_{Aii} = 0$. Since $\sigma_u I^\alpha - \beta_{ii} > 0$ in equilibrium, the principal achieves the optimum if and only if $\pi_{Aii} = 0$.

APPENDIX F: PROOF OF PROPOSITION 5.5

Given the form of contracts as in the statement, the stock price volatility matrix turns out to be in the form of (5.14), and thus

$$(\Sigma^S)^\top \pi^{Ai} = \begin{bmatrix} \pi_{Ai1}\sigma_{11}^S \\ \vdots \\ \pi_{Aii}\sigma_{ii}^S \\ \vdots \\ \pi_{Ai,N+1}\sigma_Y^S + \sum_{j=1}^{N} \pi_{Aij}\sigma_{jc}^S \end{bmatrix}.$$

However, the definitions of ξ's and their optimal values imply

$$[(\Sigma^S)^\top \pi^{Pi}]_k = [\xi^{Pi} + \beta^i]_k = \begin{cases} \dfrac{\theta_k}{\gamma_P} & k \neq i,\ N+1, \\ \dfrac{\theta_i}{\gamma_P} - (\sigma_u f(I_i) - \beta_{ii}) & k = i, \\ \dfrac{\theta_c}{\gamma_P} - (\sigma_c f(I_i) - \beta_{ic}) & k = N+1, \end{cases}$$

$$[(\Sigma^S)^\top \pi^{Ai}]_k = [\xi^{Ai} - \beta^i]_k = \begin{cases} \dfrac{\theta_k}{\gamma_A} & k \neq i,\ N+1, \\ 0 & k = i, \\ \dfrac{\theta_c}{\gamma_A} - \beta_{ic} & k = N+1. \end{cases}$$

Thus, the statement follows.

APPENDIX G: THE LINEARITY RESULT WITH A CONTINUOUS-TIME MODEL

In this section, we introduce financial markets to the Holmstrom–Milgrom economy and show that the Holmstrom–Milgrom linearity result still holds even in the presence of financial markets. Since the first-best equilibrium is similar, we only consider the second-best case. Let the accounting value of firm i at time t denoted by $\tilde{D}_i(t)$, and its dynamics given by

$$d\tilde{D}_t^i = \Sigma^i d\tilde{B}_t^0, \quad \tilde{D}_0^i = D_0, \quad 0 \leq t \leq 1,$$

where $\tilde{B}_t^0$ is an $(N+1)$-vector of independent standard Brownian motions under measure P^0. Note that $\tilde{B}_1^0$ and $\tilde{D}_1^i$ correspond, respectively, to $\tilde{B}^0$ and $\tilde{D}_i$ in our discrete-time model. Also let all agents' effort $\{\mu = (\mu_t^1, \ldots, \mu_t^N)_{0\leq t\leq 1}\}$ jointly change the probability measure of $\{\tilde{D}_t^i, i = 1, \ldots, N\}$ from P^0 to P^μ such that

$$\frac{dP^\mu}{dP^0} = \exp\left\{-\frac{1}{2}\int_0^1 \|\Sigma^{-1} l_t\|^2 dt + \int_0^1 (\Sigma^{-1} l_t)^\top d\tilde{B}_t^0\right\},$$

where $l_t := (f(I_1)\mu_t^1, \ldots, f(I_N)\mu_t^N, 0)^\top$. However, P^μ is neither observable nor verifiable.

Each agent i can observe his own effort level μ_t^i, but cannot observe other agents' effort levels μ_t^{-i}, where $\mu_t^{-i} := (\mu_t^1, \ldots, \mu_t^{i-1}, \mu_t^{i+1}, \ldots, \mu_t^N)$. Under P^μ, $\tilde{B}_t^\mu = \tilde{B}_t^0 - \int_0^t \Sigma^{-1} l_s ds$ is an $(N+1)$-vector of independent standard Brownian motions. Thus, under measure P^μ,

$$\tilde{D}_i(t) = D_0 + \int_0^t f(I_i)\,\mu_s^i ds + \int_0^t \sigma_u d\tilde{B}_s^{\mu,i} + \int_0^t \sigma_c d\tilde{B}_t^c.$$

The market prices of stocks depend on investors' beliefs on agents' effort levels. Let their beliefs be $\{\mu^*(t)\}_{0\le t\le 1}$. Given the beliefs, the financial markets are complete under probability space $(\Omega, P^{\mu^*}, \mathcal{F}_t)$ with the risk-neutral measure Q such that

$$\frac{dQ}{dP^{\mu^*}} = \exp\left\{-\frac{1}{2}\int_0^1 \|\theta_t\|^2 dt - \int_0^1 \theta_t^\top dB_t^{\mu^*}\right\}.$$

Then, $\{\theta_t\}$ is an $\mathcal{F}_t$-adapted $(N+1)$-vector of real-valued market-price-of-risk processes. Let $R_t = e^{rt}$ where r is the short rate of the economy. Then $R_1 (= e^r = R)$ is one plus interest rate rate as defined in the text. Then the quantity $R^{-1}\frac{dQ}{dP^{\mu^*}}$ is called the pricing kernel which is well known (and actually turns out) to be principals' MRS between current (time 0) and future (time 1) state-dependent consumption in equilibrium. Then the martingale representation theorem implies that prices of all assets traded in financial markets can be represented in the following linear form:

$$S_t^i = R_t S_0^i + \int_0^t \sigma_i^S(s)\theta_s ds + \int_0^t \sigma_i^S(s)\, d\tilde{B}_s^{\mu^*}. \tag{G.1}$$

Assume θ_t is uniformly bounded. Under Q, again by the Girsanov theorem,

$$\tilde{B}_t^\theta = \tilde{B}_t^{\mu^*} + \int_0^t \theta_s ds$$

is an $(N+1)$-vector of standard Brownian motions. Asset prices can be written as

$$S_t^i = R_t S_0^i + \int_0^t \sigma_i^S(s)\, d\tilde{B}_s^\theta, \tag{G.2}$$

where $\sigma_i^S(s) = \left(\sigma_{i1}^S(t), \ldots, \sigma_{iN}^S(t), \sigma_{i,N+1}^S(t),\right)$.

Similarly, the principal's and the agent's wealth processes become

$$\mathcal{W}_t^{Pi} = R_t \mathcal{W}_0^{Pi} + \int_0^t \left(\pi_s^{Pi}\right)^\top \Sigma_s^S d\tilde{B}_s^\theta, \tag{G.3}$$

$$\mathcal{W}_t^{Ai} = R_t \mathcal{W}_0^{Ai} + \int_0^t \left(\pi_s^{Ai}\right)^\top \Sigma_s^S d\tilde{B}_s^\theta. \tag{G.4}$$

We assume that $\pi_t^{Aik} \in K_i^k \subset \mathcal{R}$, where $i = 1, \ldots, N$ and $k = 1, \ldots, N+1$, and that $E[\int_0^1 \|(\pi_t^{Pi})^\top \Sigma_t^S\|^2 dt] < \infty$ and $E[\int_0^1 \|(\pi_t^{Ai})^\top \Sigma_t^S\|^2 dt] < \infty$, excluding the well-known doubling strategies in financial markets. We allow principal i to choose sets K_i^k's, $k = 1, \ldots, N+1$, in order to impose restrictions on agent i's risky-asset transactions. For example, if $K_i^k \ge 0$, then shortselling stock k by agent i is prohibited.

For the rest of the analysis of this section, we rely on the following Lemma.

LEMMA G.1 (Representation of the certainty equivalent wealth). *Let $V_t^{\mu,\pi}$ be the certainty equivalent wealth at time t of the following conditional expected utility function:*

$$-\exp\{-\gamma V_t^{\mu,\pi}\} = E^{\mu}\left[-\exp\left\{-\gamma\left(\eta + \int_t^1 H(\mu_s,\pi_s)ds + \int_t^1 v^{\top}(\mu_s,\pi_s)d\tilde{B}_s^0\right)\right\}\middle|\ \mathcal{F}_t\right],$$

where η is $\mathcal{F}_1$-measurable, and H and v are, respectively, scalar-valued and $(N+1)$-vector-valued functions. Assume that the expectation in this definition is well defined, i.e., $|V_t^{\mu,\pi}| < \infty$. Then, there exists a unique $\mathcal{F}_t$-predictable and square integrable $(N+1)$-vector processes $\{Z_t\}$ such that $V_t^{\mu,\pi}$ can be represented in the following form:

$$V_t^{\mu,\pi} = \eta - \int_t^1 \frac{Z_s^{\top}}{\gamma} d\tilde{B}_s^0 + \int_t^1 \left(H(\mu_s,\pi_s) + \left(\frac{Z_s}{\gamma} + v(\mu_s,\pi_s)\right)^{\top}\Sigma^{-1}l - \frac{\gamma}{2}\left\|\frac{Z_s}{\gamma} + v(\mu_s,\pi_s)\right\|^2\right)ds.$$

We utilize this Lemma to examine the necessary structure of optimal compensation contract C_1^i. Note that agent i's expected utility at time t can be written as follows:

$$-\exp\left(-\gamma_A V_t^{Ai}\right) := E^{\mu}\left[-\exp\left\{-\gamma_A\left(C_1^i + R\mathcal{W}_0^{Ai} + \int_t^1 \left(\pi_s^{Ai}\right)^{\top}\Sigma_s^S d\tilde{B}_s^{\theta} - \int_t^1 \frac{\delta}{2}\mu_i^2(s)\,ds\right)\right\}\middle|\ \mathcal{F}_t\right].$$

By Lemma G.1, the certainty equivalent wealth is

$$V_t^{Ai} = R\mathcal{W}_0^{Ai} + C_1^i - \int_t^1 \left(\beta_s^i\right)^{\top} d\tilde{B}_s^0 + \int_t^1 \left\{\left(\pi_s^{Ai}\right)^{\top}\Sigma_s^S\left(\theta_s - \Sigma^{-1}l_s^* + \Sigma^{-1}l_s\right) - \frac{\delta}{2}\mu_i^2(s) + \left(\beta_s^i\right)^{\top}\Sigma^{-1}l_s - \frac{\gamma_A}{2}\left\|\beta_s^i + \left(\Sigma_s^S\right)^{\top}\pi_s^{Ai}\right\|^2\right\}ds.$$

Note that, in equilibrium, since the market correctly predicts the manager's effort, we have $l_t^* = l_t$. Thus, in equilibrium, if the agent chooses $\mu_i(t)$ given C_1^i, then C_1^i can be represented as follows:

$$C_1^i = V_0^{Ai} - R\mathcal{W}_0^{Ai} + \int_0^1 \left(\beta_t^i\right)^{\top} d\tilde{B}_t^0 + \int_0^1 \left[\frac{\delta}{2}\mu_i^2(t) + \frac{\gamma_A}{2}\|\beta_t^i + \left(\Sigma_t^S\right)^{\top}\pi_t^{Ai}\|^2 - \left(\pi_t^{Ai}\right)^{\top}\Sigma_t^S\theta_t - \left(\beta_t^i\right)^{\top}\Sigma^{-1}l_t\right]dt. \tag{G.5}$$

This equation gives a necessary condition for contracts to be optimal in the presence of financial markets satisfying the participation constraint with certainty equivalent V_0^{Ai}. However, this condition does not mean sufficiency for the optimality, as given C_1^i in the form of (G.5), the agent may not necessarily choose $\mu_i(t)$. In other words, not all C_1^i's of the form (G.5) are implementable.

The next lemma is about the implementability.

LEMMA G.2. *Suppose a compensation contract C_1^i is given to the agent in the form of (G.5) with $\{\mu_t^i\}$. Then the agent optimally chooses his effort level to be $\{\mu_t^i\}$ if and only if*

μ_t^i in (G.5) for C_1^i satisfies, for $0 \le t \le 1$ almost surely

$$\mu_t^i = \frac{1}{\delta\sigma_u}\left[\beta_t^i + \left(\Sigma_t^S\right)^\top \pi_t^{Ai}\right]_i, \tag{G.6}$$

or equivalently

$$\mu_t^i \in \arg\max_{\hat{\mu}} \left(\beta_t^i + \left(\Sigma_t^S\right)^\top \pi_t^{Ai}\right)\Sigma^{-1}\hat{l} - \frac{\delta}{2}\hat{\mu}_t^2.$$

Proof. The statement and its proof are similar to those of Schättler and Sung (1993, theorem 4.2) except that this lemma has financial-market decision variable π_t^{Ai} added. Suppose that given C_1^i in the form of (G.5) with $\{\mu_t^i\}$, the agent chooses $\{\hat{\mu}_t\}_{0\le t\le 1}$. Then, the agent's expected utility is

$$\begin{aligned}
-\exp\{-\gamma\,\hat{V}_0\} = E^{\hat{\mu}}\Bigg[-\exp\Bigg\{-\gamma_A\Bigg(V_0^{Ai} + \int_0^1 \frac{\gamma_A}{2}\|\beta_t^i + \left(\Sigma_t^S\right)^\top \pi_t^{Ai}\|^2 dt \\
+ \int_0^1 \left(\beta_t^i + \left(\Sigma_t^S\right)^\top \pi_t^{Ai}\right)^\top d\tilde{B}_t^{\hat{\mu}} + \int_0^1 \big[H\big(\hat{\mu}_t^i; \beta_t^i, \pi_t^{Ai}; \Sigma_t^S, \sigma_u, \mu_t^{-i}, \delta\big) \\
- H\big(\mu_i; \beta, \pi^{Ai}; \Sigma^S, \sigma_u, \mu^{-i}, \delta\big)\big]dt\Bigg)\Bigg\}\Bigg],
\end{aligned}$$

where

$$H(\mu^i; \beta, \pi^{Ai}; \Sigma^S, \sigma_u, \mu^{-i}, \delta) := (\beta^i + (\Sigma^S)^\top \pi^{Ai})\Sigma^{-1} l - \frac{\delta}{2}(\mu^i)^2.$$

Note that μ_t^i satisfying (G.6) uniquely maximizes $H(\mu^i; \beta_t, \pi_t^{Ai}; \Sigma_t^S, \sigma_u, \mu_t^{-i}, \delta)$ over μ^i. Thus we have

$$\begin{aligned}
-\exp\{-\gamma\,\hat{V}_0\} &\le E^{\hat{\mu}}\Bigg[-\exp\Bigg\{-\gamma\left(V_0^{Ai} + \int_0^1 \frac{\gamma_A}{2}\left\|\beta_t^i + \left(\Sigma_t^S\right)^\top \pi_t^{Ai}\right\|^2 dt \right.\\
&\qquad \left. + \int_0^1 \left(\beta_t^i + \left(\Sigma_t^S\right)^\top \pi_t^{Ai}\right)^\top d\tilde{B}_t^{\hat{\mu}}\right)\Bigg\}\Bigg] \\
&= -\exp\left(-\gamma\, V_0^{Ai}\right).
\end{aligned}$$

Thus, the agent's expected utility $-\exp\{-\gamma\,\hat{V}_0\}$ is maximized if and only if $\hat{\mu}_t^i = \mu_t^i$ almost surely. □

The representation (G.5) of compensation contracts, together with the implementability condition (G.6), completely describes the class of all implementable contracts. Thus, the principal only needs to choose a contract from this class. Suppose that the principal chooses a contract C_1^i from the class with $(\pi_t^{Pi}, \pi_t^{Ai}, \beta_t^i)$. Let

$$\xi_t^{Pi} = \left(\Sigma_t^S\right)^\top \pi_t^{Pi} - \beta_t^i, \quad \xi_t^{Ai} = \left(\Sigma_t^S\right)^\top \pi_t^{Ai} + \beta_t^i.$$

Recall that $\mu_t^i = \frac{1}{\delta\sigma_u}\xi_t^{Aii}$, where ξ_t^{Aii} is the ith component of ξ_t^{Ai}. Then, the principal's expected utility for the second period consumption can be expressed in terms of ξ^{Pi} and

ξ^{Ai}, and it becomes $E^{\mu}[-\exp\{-\gamma_P V_1^{Pi}\}]$, where

$$\begin{aligned} V_1^{Pi} &= \mathcal{W}_1^{Pi} + \tilde{D}_1^i - C_1^i, \\ &= R\big(\mathcal{W}_0^{Pi} + \mathcal{W}_0^{Ai}\big) + D_0 - V_0^{Ai} + \int_0^1 \big(\big(\xi_t^{Pi}\big)^{\top} + \Sigma^i\big) d\tilde{B}_t^{\mu} \\ &\quad + \int_0^1 \Big[\big(\xi_t^{Pi}\big) + \xi_t^{Ai}\big)^{\top}\theta_t - \frac{1}{2\delta\sigma_u^2}\big(\xi_t^{Aii}\big)^2 - \frac{\gamma_A}{2}\big\|\xi_t^{Ai}\big\|^2 \\ &\quad + \big(\xi_t^{Pi} + \beta_t^i\big)^{\top}\Sigma^{-1}\big(l_t - l_t^*\big) + \frac{f(I_i)}{\delta\sigma_u}\xi_t^{Aii}\Big] dt. \end{aligned}$$

Thus, the principal's problem can be restated as a problem of choosing (ξ_t^{Pi}, ξ_t^{Ai}) to maximize her expected utility.

Suppose (ξ_t^{Pi}, ξ_t^{Ai}) are the optimal choice of the principal. Let V_0^{Pi} be the principal's certainty equivalent such that $-\exp\{-\gamma_P V_0^{Pi}\} = E^{\mu}[-\exp\{-\gamma_P V_1^{Pi}\}]$. Then, by Lemma G.1, there exists a unique $\mathcal{F}_t$-predictable and square integrable $(N+1)$-vector process $\{\alpha_t^i\}_{0\le t\le 1}$ such that V_0^{Pi} has the following representation

$$\begin{aligned} V_0^{Pi} &= R\big(\mathcal{W}_0^{Ai} + \mathcal{W}_0^{Pi}\big) + D_0 - V_0^{Ai} - \int_0^1 \big(\alpha_t^i\big)^{\top} d\tilde{B}_t^0 \\ &\quad + \int_0^1 \Big[\big(\xi_t^{Pi} + \xi_t^{Ai}\big)^{\top}\theta_t - \frac{\gamma_P}{2}\big\|\alpha_t^i + (\Sigma^i)^{\top} + \xi_t^{Pi}\big\|^2 \\ &\qquad - \frac{\gamma_A}{2}\big\|\xi_t^{Ai}\big\|^2 - \frac{1}{2\delta\sigma_u^2}\big(\xi_t^{Aii}\big)^2 + \big(\big(\alpha_t^i\big)^{\top} + \Sigma^i\big)\Sigma^{-1} l_t\Big] dt. \end{aligned} \tag{G.7}$$

LEMMA G.3. *In equilibrium, given the representation (G.7) for V_0^{Pi}, $\big(\xi_t^{Pi}, \xi_t^{Ai}\big)$ is optimal if and only if, almost surely,*

$$\xi_t^{Pik} = \begin{cases} \dfrac{\theta_t^k}{\gamma_P} - \alpha_t^{ik}, & k \neq i,\ N+1 \\ \dfrac{\theta_t^i}{\gamma_P} - \alpha_t^{ii} - \sigma_u f(I_i), & k = i \\ \dfrac{\theta_c}{\gamma_P} - \alpha_t^{i,N+1} - \sigma_c f(I_i), & k = N+1, \end{cases} \quad \text{or} \quad \alpha_t^i + \Sigma_i^{\top} + \xi_t^{Pi} = \frac{\theta}{\gamma_P} \tag{G.8}$$

$$\xi_t^{Aik} = \begin{cases} \dfrac{\theta_t^k}{\gamma_A}, & k \neq i \\ \dfrac{\delta\sigma_u^2}{1+\gamma_A\delta\sigma_u^2}\left(\theta_t^i + \big(\alpha_t^{ii} + \sigma_u f(I_i)\big)\dfrac{1}{\delta\sigma_u^2}\right), & k = i, \end{cases} \tag{G.9}$$

$$\mu_t^{i*} = \frac{\sigma_u}{1+\gamma_A\delta\sigma_u^2}\left(\theta_t^i + \big(\alpha_t^{ii} + \sigma_u f(I_i)\big)\frac{1}{\delta\sigma_u^2}\right). \tag{G.10}$$

Proof. Consider an arbitrary admissible controls $(\hat{\xi}_t^{Pi}, \hat{\xi}_t^{Ai})$. In equilibrium where $l^* = l$, the wealth at time 1 is

$$\hat{V}_1^{Pi} = R(\mathcal{W}_0^{Pi} + \mathcal{W}_0^{Ai}) + D_0 - V_0^{Ai} + \int_0^1 \left((\hat{\xi}_t^{Pi})^\top + \Sigma^i\right) d\tilde{B}_t^{\hat{\mu}} + \int_0^1 \left[\left(\hat{\xi}_t^{Pi} + \hat{\xi}_t^{Ai}\right)^\top \theta_t - \frac{1}{2\delta\sigma_u^2}\left(\hat{\xi}_t^{Aii}\right)^2 - \frac{\gamma_A}{2}\left\|\hat{\xi}_t^{Ai}\right\|^2 + \Sigma^i \Sigma^{-1} \hat{l} \right] dt.$$

Eliminating $R(\mathcal{W}_0^{Ai} + \mathcal{W}_0^{Pi}) + D_0 - V_0^{Ai}$ from this equation using (G.7), we have

$$\begin{aligned}\hat{V}_1^{Pi} = {} & V_0^{Pi} + \int_0^1 \left((\alpha_t^i)^\top + \Sigma^i + (\hat{\xi}_t^{Pi})^\top\right) d\tilde{B}_t^{\hat{\mu}} \\ & + \int_0^1 \left[\left(\hat{\xi}_t^{Pi} + \hat{\xi}_t^{Ai}\right)^\top \theta_t - \frac{1}{2\delta\sigma_u^2}\left(\hat{\xi}_t^{Aii}\right)^2 - \frac{\gamma_A}{2}\left\|\hat{\xi}_t^{Ai}\right\|^2 + \left((\alpha_t^i)^\top + \Sigma^i\right)\Sigma^{-1}\hat{l} \right] dt \\ & - \int_0^1 \left[\left(\xi_t^{Pi} + \xi_t^{Ai}\right)^\top \theta_t - \frac{\gamma_P}{2}\left\|\alpha_t^i + (\Sigma^i)^\top + \xi_t^{Pi}\right\|^2 \right. \\ & \left. - \frac{\gamma_A}{2}\left\|\xi_t^{Ai}\right\|^2 - \frac{1}{2\delta\sigma_u^2}\left(\xi_t^{Aii}\right)^2 + \left((\alpha_t^i)^\top + \Sigma^i\right)\Sigma^{-1} l_t \right] dt.\end{aligned}$$

Define

$$\begin{aligned}H(\alpha, \theta, \xi^P, \xi^A) = {} & (\xi^P + \xi^A)^\top \theta - \frac{\gamma_P}{2}\|\alpha + (\Sigma^i)^\top + \xi^P\|^2 - \frac{\gamma_A}{2}\|\xi^A\|^2 \\ & - \frac{1}{2\delta\sigma_u^2}(\xi^{Aii})^2 + ((\alpha^i)^\top + \Sigma^i)\Sigma^{-1} l.\end{aligned}$$

Then, we have

$$\begin{aligned}\hat{V}_1^{Pi} = {} & V_0^{Pi} + \int_0^1 \left[H\left(\alpha_t^i, \theta_t, \hat{\xi}_t^{Pi}, \hat{\xi}_t^{Ai}\right) - H\left(\alpha_t^i, \theta_t, \xi_t^{Pi}, \xi_t^{Ai}\right)\right] dt \\ & + \int_0^1 \frac{\gamma_P}{2}\left\|\alpha_t^i + (\Sigma^i)^\top + \hat{\xi}_t^{Pi}\right\|^2 dt + \int_0^1 \left(\alpha_t^i + (\Sigma^i)^\top + \hat{\xi}_t^{Pi}\right) dB_t^{\hat{\mu}}.\end{aligned}$$

Suppose that the pair (ξ_t^{Pi}, ξ_t^{Ai}) maximizes $H(\alpha_t^i, \theta_t, \bar{\xi}_t^{Pi}, \bar{\xi}_t^{Ai})$ with respect to $(\bar{\xi}_t^{Pi}, \bar{\xi}_t^{Ai})$. Then, $E[-\exp\{-\gamma_P \hat{V}_1^{Pi}\}] \le E[-\exp\{-\gamma_P V_1^{Pi}\}] = -\exp(-\gamma_P V_0^{Pi})$, and the equality holds when $(\hat{\xi}_t^{Pi}, \hat{\xi}_t^{Ai}) = (\xi_t^{Pi}, \xi_t^{Ai})$ almost surely. Therefore, (ξ_t^{Pi}, ξ_t^{Ai}) maximizes the principal's expected utility, and has to satisfy equations (G.8) to (G.9), which are the first-order necessary conditions to maximize $H(\alpha_t^i, \theta_t, \bar{\xi}_t^{Pi}, \bar{\xi}_t^{Ai})$. The condition (G.10) is the incentive compatibility condition for the principal's second-best problem.

If there exists $(V_0^{Pi}, \alpha^i, \xi^{Pi}, \xi^{Ai})$ satisfying equations (G.7) to (G.10), then the optimal expected utility of the principal is V_0^{Pi} and any other admissible choice of $(\hat{\xi}^{Pi}, \hat{\xi}^{Ai})$ cannot result in a larger expected utility at time 0 for the principal. Hence we prove the sufficiency. We leave the existence and uniqueness of $(V_0^{Pi}, \alpha^i(s))$ to the following Lemma G.4. □

LEMMA G.4. *Given $(\theta(s))_{0\le s\le 1}$, there exists unique $(V_0^{Pi}, \alpha^i, \xi^{Pi}, \xi^{Ai})$ satisfying (G.8) to (G.10). Moreover, the unique α^i is equal to* $\mathbf{0}$ *in equilibrium, where* $\mathbf{0}$ *is an* $(N+1)$*-vector of zeros.*

Proof. Plugging equations (G.8) to (G.10) into (G.7), after a tedious computation, we have (G.8) reduced to

(G.11)

$$R\left(\mathcal{W}_0^{Ai}+\mathcal{W}_0^{Pi}\right)+D_0-V^{Ai}=V_0^{Pi}+\int_0^1 H_0(\theta_t)dt + \int_0^1\left[H_1(\theta_t)\alpha_t^{ii}+\int_0^1 q\left(\alpha_t^{ii}\right)^2\right]dt-\int_0^1\alpha_t^i d\tilde{B}_t^0$$

where q is a strictly positive constant (in fact, $q=(\delta\sigma_u)/(2(1+\gamma_A\delta\sigma_u^2))$), and $H_1(\theta)$ and $H_0(\theta)$ are deterministic functions of θ and independent of α^i. We only need to show that there exists a unique pair of (V^{Pi},α_s^i) given any uniformly bounded $\{\theta_s\}_{0\le s\le 1}$. Define $B_t'=B_t^0-\int_0^t H_1(\theta_s)\,ds$. From the boundedness of θ_s, we know that B_t' is a Brownian motion under measure P'

$$\frac{dP'}{dP}=\exp\left\{-\frac{1}{2}\int_0^1 H_1^2(\theta_t)\,dt+\int_0^1 H_1(\theta_s)d\tilde{B}_t^0\right\}.$$

Define

$$K_t=V_0^{Pi}+\int_0^t q\left(\alpha_s^{ii}\right)^2 ds-\int_0^t\alpha_s^i d\tilde{B}_s'.$$

Then $K_0=V_0^{Pi}$, and by (G.11), $K_1=R(\mathcal{W}_0^{Ai}+\mathcal{W}_0^{Pi})+D_0-V^{Ai}(0)+\int_0^1 H_0(\theta_s)ds$. Applying Ito's formula, we have $de^{-qK_s}=qe^{-qK_t}\alpha_t^i d\tilde{B}_t'$. Hence, e^{-qK_t} is a martingale under measure P'. Then, we have $E[e^{-qK_1}\frac{dP'}{dP}]=e^{-qV^{Pi}}$, which implies

$$V_0^{Pi}=-\frac{1}{q}\log\left\{E^0\left[\exp\left\{-\frac{1}{2}\int_0^1 H_1^2(\theta_t)\,dt+\int_0^1 H_1(\theta_t)\,d\tilde{B}_t^0 -q\left[R\left(\mathcal{W}_0^{Ai}+\mathcal{W}_0^{Pi}\right)+D_0-V_0^{Ai}-\int_0^1 H_0(\theta_s)ds\right]\right\}\right]\right\}.$$

To see the uniqueness of α_s^i, note that by the martingale representation theorem, there exists a unique Λ_t such that

$$e^{-qK_t}=e^{-qV^{Pi}}+\int_0^t\Lambda_s dB_s'=E_t'[e^{-qK_1}].$$

Since $de^{-qK_t}=qe^{-qK_t}\alpha_t^i dB_t'$, the uniqueness of Λ_t implies that we must uniquely have

$$\Lambda_t=\alpha_t^i qe^{-qK_t}.$$

This also implies the uniqueness of α_s^i. Thus, for any uniformly bounded $\{\theta_t\}$, there exists a unique pair of $(V^{Pi},\{\alpha_t^i\})$.

For the second part of the statement, we first conjecture that in equilibrium, such unique α_t^i is $\mathbf{0}$ for all i, and then show this conjecture is true. From the market equilibrium condition, $\sum_i(\xi_t^{Ai}+\xi_t^{Pi})=\mathbf{0}$, and from equations (G.8) to (G.10), we know θ_t is a deterministic function of α_t which we denote by $h(\alpha_t)$. Then $h(\alpha_t)=h(\mathbf{0})$ under the conjecture. By (G.11), V_0^{Pi} is given by

(G.12)
$$V^{Pi}=R\left(\mathcal{W}_0^{Ai}+\mathcal{W}_0^{Pi}\right)+D_0-V^{Ai}-\int_0^1 H_0(h(0))ds.$$

Hence, V^{Pi} with $\alpha_t^i = \mathbf{0}$ exists. Moreover, when $\alpha_t^i = \mathbf{0}$, equations (G.8) to (G.10) imply that optimal $(\xi_t^{Pi}, \xi_t^{Ai}, \mu_t^i)$ exist and are constant over time in equilibrium. That is, when $\alpha_t^i = \mathbf{0}$, $(V^{Pi}, \xi_t^{Pi}, \xi_t^{Ai}, \mu_t^i)$ constitute a solution to necessary conditions in equations (G.8) to (G.10). Since the solution has to be unique by Lemma G.4, we have $\alpha_t^i = \mathbf{0}$ in equilibrium, i.e., our conjecture is verified to be true. □

Finally, recall that given a contract C_1^i, the agent is only concerned with aggregate risk position ξ_t^{Ai}. Since $\xi_t^{Ai} = (\Sigma_t^S)^\top \pi_t^{Ai} + \beta_t^i$, there is a degree of freedom in choosing optimal (π_t^{Ai}, β_t^i) given optimal ξ_t^{Ai}. Thus, without loss of generality, we may let β_s be constant over time. Then, the compensation (G.5) is simplified to

$$C_1^i = V^{Ai} - R\mathcal{W}_0^{Ai} + \frac{\delta}{2}\mu_i^2 + \frac{\gamma_A}{2}\|\xi^{Ai}\|^2 - \xi^{Ai}\theta - (\beta^i)^\top[\Sigma^{-1}l - \theta] + \beta^i \tilde{B}_1^0,$$

which is identical to the linear compensation contract under our discrete-time setup.

To see this linearity result intuitively, let us suppose that the market prices of risks are constant over time. Then the principal's time 1 utility in continuous-time model becomes

$$\begin{aligned} &E^\mu\left[-\exp\left\{\mathcal{W}_1^{Pi} + D_i - C_i\right\}\right] \\ &= E^\mu\left[-\exp\left\{\mathcal{W}_0^{Pi}R + \int_0^1 \left(\pi_t^{Pi}\right)^\top \Sigma^S d\tilde{B}^\theta + D_i - C_i\right\}\right], \\ &= E^\mu\left[-\exp\left\{\mathcal{W}_0^{Pi}R + \int_0^1 \left(\pi_t^{Pi}\right)^\top \Sigma^S\left(d\tilde{B}^{\mu^*} + \theta dt\right) + D_i - C_i\right\}\right]. \end{aligned}$$

The first equality utilizes the self-financing condition. Note that $\mu = \mu^*$ in equilibrium, and that there are extra terms in the exponent of the utility, in addition to the Holmstrom and Milgrom's term $D_i - C_i$. When θ is a constant vector, these extra terms are simply an arithmetic Brownian motion. Then our principal's problem basically turns into the same problem as in Holmstrom and Milgrom except that the principal chooses not only (μ, C_i) but π^{Pi} and implicitly π^{Ai}. Here, π^{Pi} affects both the drift and volatility of the exponent, which is similar to the case of observable volatility analyzed in Sung (1995). The exponent of agent's utility can also be expressed in terms of a arithmetic Brownian motion with choice variables (μ, π^{Ai}). Since π^{Ai} is observable, the principal can choose it for the agent. Thus, under the assumption of constant market prices of risks and stock volatilities, our contracting problem turns out to be a variation of the model in Sung (1995) and thus we have the linearity of optimal contract again.

REFERENCES

BISIN, A., and P. GOTTARDI (1999): General Competitive Analysis with Asymmetric Information, *J. Econ. Theory* 87, 1–48.

CITANNA, A., and A. VILLANACCI (2002): Competitive Equilibrium with Moral Hazard in Economies with Multiple Commodities, *J. Math. Econ.*, 38, 117–147.

CIVITANIĆ, J., X. WAN, and J. ZHANG (2007): Continuous-Time Principal-Agent Problems with Hidden Action and Lump-Sum Payment, Working Paper, Caltech.

CHIAPPORI, P.-A., and B. SALANIÉ (2003): Testing Contract Theory: A Survey of Some Recent Work, in *Advances in Economics and Econometrics*, M. Dewatripont, L. Hansen, and S. Turnovsky, eds., Cambridge University Press, pp. 115–149.

CONSTANTINIDES, G. M. (1990): Habit Formation: A Resolution of the Equity Premium Puzzle, *J. Polit. Econ.* 93, 519–543.

DANTHINE, J.-P., and J. B. DONALDSON (2007): Executive Compensation: The View from General Equilibrium, Working Paper, University of Lausanne.

GARVEY, G., and T. MILBOURN (2003): Incentive Compensation When Executives Can Hedge the Market: Evidence of Relative Performance Evaluation in the Cross Section, *J. Financ.* LVIII, 1557–1581.

GREEN, J., and N. STOKEY (1983): A Comparison of Tournaments and Contracts, *J. Polit. Econ.* 91(3), 349–64.

HEATON, J., and D. J. LUCAS (1996): Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Pricing, *J. Polit. Econ.* 104, 443–487.

HOLMSTROM, B. (1982): Moral Hazard in Teams, *Bell J. Econ.* 13, 324–340.

HOLMSTROM, B., and P. MILGROM (1987): Aggregation and Linearity in the Provision of Intertemporal Incentives, *Econometrica* 55, 303–328.

JENSEN, M. (1986): The Agency Costs of Free Cash Flow: Corporate Finance and Takeovers, *Am. Econ. Rev.* 76, 323–330.

KAHN, J. (1990): Moral Hazard, Imperfect Risk-Sharing and the Behavior of Asset Returns, *J. Monetary Econ.* 26, 27–44.

KOCHERLAKOTA, N. R. (1998): The Effects of Moral Hazard on Asset Prices When Financial Markets Are Complete, *J. Monetary Econ.* 41, 39–56.

KOCHERLAKOTA, N. R., and L. PISTAFERRI (2009): Asset Pricing Implications of Pareto Optimality with Private Information, *J. Polit. Econ.* 117, 555–590.

KOO, H. K., G. SHIM, and J. SUNG (2008): Optimal Multi-agent Performance Measures for Team Contracts, *Math. Financ.* 18, 649–667.

KYLE, A. (1985): Continuous Auctions and Insider Trading, *Econometrica* 47, 331–359.

LAZEAR, E., and S. ROSEN (1981): Rank Order Tournaments as Optimum Labor Contracts, *J. Polit. Econ.* 89, 841–64.

MAGILL, M., and M. QUINZII (2005): Normative Properties of Stock Market Equilibrium with Moral Hazard, Working Paper, University of Southern California.

MAUG, E. (2000): The Relative Performance Puzzle, *Schmalenbach Bus. Rev.* 52, 3–24.

MEHRA, R., and E. PRESCOTT (1985): The Equity Premium: A Puzzle, *J. Monetary Econ.* 15, 145–161.

NALEBUFF, B., and J. STIGLITZ (1983): Prizes and Incentives: Toward a General Theory of Compensation and Competition, *Bell J. Econ.* 14, 21–43.

OU-YANG, H. (2005): An Equilibrium Model of Asset Pricing and Moral Hazard, *Rev. Financ. Stud.* 18, 1253–1303.

PRESCOTT, E. C., and R. M. TOWNSEND (1984): Pareto Optima and Competitive Equilibria with Adverse Selection and Moral Hazard, *Econometrica* 52, 21–45.

ROSS, S. (1976): The Arbitrage Theory of Capital Asset Pricing, *J. Econ. Theory* 13(3), 341–360.

SCHÄTTLER, H., and J. SUNG (1993): The First-Order Approach to Continuous-Time Principal-Agent Problem with Exponential Utility, *J. Econ. Theory* 61, 331–371.

SHAVELL, S. (1979): Risk Sharing and Incentives in the Principal and Agent Relationship, *Bell J. Econ.* 10, 55–73.

STIGLITZ, J. (1974): Incentives and Risk Sharing in Sharecropping, *Rev. Econ. Stud.* 41, 219–255.

SUNG, J. (1995): Linearity with Project Selection and Controllable Diffusion Rate in Continuous-Time Principal-Agent Problems, *RAND J. Econ.* 26, 720–743.

SUNG, J., and X. WAN (2009): Equilibrium Equity Premium, Interest Rate and the Cost of Capital in a Moral Hazard Economy, Working Paper, Ajou University and Hong Kong University of Science and Technology.

WANG, J. (1993): A Model of Intertemporal Asset Prices under Asymmetric Information, *Rev. Econ. Stud.* 60, 249–282.

WEIL, P. (1989): The Equity Premium Puzzle and the Riskfree Rate Puzzle, *J. Monetary Econ.* 24, 401–422.

WILLIAMS, N. (2006): On Dynamic Principal-Agent Problems in Continuous Time, Working Paper, Princeton University.

ZAME, W. R. (2007): Incentives, Contracts and Markets: A General Equilibrium Theory of Firms, *Econometrica* 75(5), 1453–1500.

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

Mathematical Finance, Vol. 25, No. 4 (October 2015), 869–889

MARKETS FOR INFLATION-INDEXED BONDS AS MECHANISMS FOR EFFICIENT MONETARY POLICY

CHRISTIAN-OLIVER EWALD

Adam Smith Business School—Economics, University of Glasgow, UK

JOHANNES GEISSLER

Center for Dynamic Macroeconomic Analysis, School of Economics and Finance, University of St. Andrews, UK

We consider a continuous-time framework featuring a central bank, private agents, and a financial market. The central bank's objective is to maximize a functional, which measures the classical trade-off between output and inflation over time plus income from the sales of inflation-indexed bonds minus payments for the liabilities that the inflation-indexed bonds produce at maturity. Private agents are assumed to have adaptive expectations. The financial market is modeled in continuous-time Black–Scholes–Merton style and financial agents are averse against inflation risk, attaching an inflation risk premium to nominal bonds. Following this route, we explain demand for inflation-indexed securities on the financial market from a no-arbitrage assumption and derive pricing formulas for inflation-linked bonds and calls, which lead to a supply-demand equilibrium. Furthermore, we study the consequences that the sales of inflation-indexed securities have on the observed inflation rate and price level. Similar to the study of Walsh, we find that the inflationary bias is significantly reduced, and hence that markets for inflation-indexed bonds provide a mechanism to reduce inflationary bias and increase central bank's credibility.

KEY WORDS: monetary policy, inflationary bias, mechanisms, inflation-indexed securities.

1. INTRODUCTION

The first one-billion pounds 2% inflation-indexed UK Treasury bond was issued at par on March 27, 1981, maturing in 1996. Today, inflation-indexed gilts comprise 27% of UK government gilts outstanding by market value. The United States and France followed the British example in 1997 and 1998, respectively, and in many other countries, including Australia, Sweden, and Canada, inflation-indexed securities are now traded.[1]

The financial literature on pricing of inflation-indexed securities generally takes inflation as exogenously given, and applies standard Black–Scholes-type theory or the Heath–Jarrow–Morton approach to term structure modeling to develop pricing formulas; see, for example, Jarrow and Yilderim (2003). These models do not fully

Both authors are thankful to Prof. Charles Nolan for useful comments and suggestions. Christian-Oliver Ewald gratefully acknowledges funding through the Australian Research Council's Discovery Grant DP1095969.

Manuscript received October 2012; final revision received December 2012.

Address correspondence to Christian-Oliver Ewald, Adam Smith Business School, University of Glasgow, Glasgow, G12 8QQ, United Kingdom; e-mail: christian.ewald@glasgow.ac.uk.

[1] Freely quoted from Sir Edward George, former Governor of the Bank of England, in the foreword of Deacon, Derry, and Mirfedereski (2004).

DOI: 10.1111/mafi.12039

recognize the central bank's role in issuing inflation-indexed securities and the feedback effects on the central bank's optimal monetary policy. Dynamic models that focus on the central bank's trade-off between output and inflation costs without presence of inflation-indexed securities have been developed in the economic literature, since Barro and Gordon (1983a, b) and Kydland and Prescott (1977).

It has been recognized that governments can reduce their borrowing costs by issuing inflation-indexed securities and in this way avoid paying a costly inflation risk premium. Gong and Remolona (1996) use US data to measure the inflation risk premium, obtaining a figure of 100–300 basis points for a 5-year bond and 50–150 basis points for a 10-year bond. Similar observations are made by Campbell and Shiller (1996) and Campbell, Shiller, and Viceira (2009).

Further, it is generally agreed that the existence of indexed debt will help a government to stabilize its real financing cost and thereby reduce the need for tax changes. Barro (1996) argues that for an economy with no random fluctuations in government financing, the ideal form of government debt would be inflation-indexed perpetual bonds, which would provide a uniform stream of real coupons ad infinitum.

Finally, it has been recognized that the existence of indexed debt removes one of the incentives for a government to adopt inflationary policies, namely, the possibility of "cheating" the lender and reducing the real value of its outstanding liabilities. Indeed, Hanke and Walters (1994) claim that one of the reasons why the US Treasury was rather slow in issuing indexed bonds was to keep open the possibility of reducing the real value of the Federal debt via inflation.[2]

The question of what effect inflation-indexed securities have on the trade-off between inflation and output under a New Classical Phillips curve approach, however, seems to be completely missing from the literature. The objective in this paper is to present and analyze a rather simple framework, which captures the effect of introducing inflation-indexed securities within a classical-game-theoretic framework featuring a central bank and private agents, where the trade-off between output and inflation is characterized by a New Classical Phillips curve. A motivation for our work has been given by Walsh (1995), who suggested that in order to escape the socially inefficient high inflation equilibrium within a repeated Barro and Gordon type framework, the central bank should be paid a transfer payment by the government after each round. Walsh constructs the payment in such a way, as to make it optimal for the central bank to only respond to the economy's shocks. He shows that the transfer payment $p(\pi) = t_0 - \lambda k\pi$ gives the central bank an appropriate incentive.

In our opinion, it is difficult to implement Walsh's mechanism in reality. A binary version of Walsh's mechanism, which releases a central banker from his post if an inflation target is not reached, may be implementable, but would likely be far from optimal. In addition, we find that with regards to central bank independence, transfer payments from the government to the central bank could be seen in a critical light. In this paper, we show that markets for inflation-indexed securities can effectively fill the gap for a functional and practically implementable Walsh-like mechanism.[3] We show how inflation-indexed

[2]A different aspect of cheating speculation has been pointed out by Deacon et al. "An unscrupulous finance minister might be tempted to shift the fiscal burden from sales taxes to income tax prior to an indexed bond's redemption date, solely to induce a fall in the CPI and thereby reduce the government's borrowing costs."

[3]It follows from the previous paragraphs that in many countries, this mechanism is already in place. However, its "mechanics" are not well understood and there seems to be plenty of scope in optimizing the mechanism. This paper is a first step toward a better understanding of it.

securities sold by the central bank on the financial market can create a similar transfer payment as Walsh suggests and that the introduction of a market for inflation-indexed securities to the classical modeling framework leaves the central bank, the private agents, and the financial market participants better off. A Pareto improvement is obtained. To make this analysis more interesting and to better meld with the existing financial literature, we choose a continuous-time Black–Scholes–Merton-like framework in which shocks are produced by a white noise process, i.e., the increments of a Brownian motion. This enables us to apply classical stochastic optimal control theory and for a given number N of inflation-indexed securities of a specific type issued, to derive analytic formulas for the central bank's price $p(N)$, the central bank's optimal monetary policy $\pi(N)$, and the private agent's expected inflation level $\pi^e(N)$. The number N^* of inflation-indexed securities actually issued by the central bank is then determined from fitting the central banks price to the arbitrage free price on the financial market. By doing this, it is guaranteed that supply for inflation-indexed securities equals demand. As such a partial equilibrium $(p(N^*), N^*)$ is obtained.

We assume that private agents' expectations follow an adaptive stochastic dynamics. An alternative setup, featuring rational expectations but a dynamically more constrained central bank, is presented in Ewald and Geissler (2012). In the case presented here, the central bank is able to use incoming information at any time to adjust the inflation rate fully dynamically, and given the limited speed at which the private agents can process new incoming information, and the apparent information advantage of the central bank, we feel that an adaptive expectations approach is justified and appropriate; see, for example, Evans and Honkapohja (2001). At least heuristically, we can interpret the adaptive dynamics as rational expectation obtained from a filtering argument, with private agents having partial information.[4]

Adding to the financial literature, we note that to the best of our knowledge, this paper presents the first pricing formulas for inflation-indexed securities in which monetary policy is integrated.[5]

The remainder of this paper is organized as follows. The theoretical model is set up and solved throughout Sections 2–6 while in Section 7, we present a number of numerical results. The main conclusions are summarized in Section 8.

2. THE CENTRAL BANK AND ILBs

The New Classical Phillips Curve approach provides the relationship

$$y_t = y_t^n + a\left(\pi_t - \pi_t^e\right) \tag{2.1}$$

among real output growth y_t, natural real output growth y_t^n, actual inflation π_t, and expected inflation π_t^e. The central bank is tempted to increase inflation π_t in order to generate more output in the economy. In a first-order approximation, the central bank's accumulated gains (in real terms) over the time interval $[t, T]$ following the policy π_t are

[4]In discrete time, Pearlman, Currie, Levine (1986) provide a rigorous analysis that supports this interpretation; however, no results for the continuous-time Brownian-motion-based case seem to appear in the literature.

[5]The role of interest rates futures and swap markets in the context of monetary policy and financial stability has been discussed in Driffil et al. (2006). These authors also start from a New Classical Phillips Curve approach, but the model developed from then is substantially different. The type of securities considered is different as well.

given by

$$Y_t^n \cdot \int_t^T a(\pi_u - \pi_u^e)du,$$

where Y_t^n is (absolute) natural real output. In the following, we assume for simplicity that Y_t^n is constant. Alternatively, one can think of the central bank optimizing relative to natural output. For notational simplicity and in order to avoid introducing an extra parameter, we also keep writing a for the expression $Y_t^n \cdot a$, i.e., by default, we multiply the Phillips curve slope parameter with the natural real output at a reference time.

On the other hand, high inflation causes social costs, which we assume to be quadratic of type $\frac{\tilde{\lambda}}{2}\pi_t^2$. These costs are measured in real monetary terms too. The central bank's instantaneous benefit function is then derived from (2.1) subtracting social costs:

$$a(\pi_t - \pi_t^e) - \frac{\tilde{\lambda}}{2}\pi_t^2 = a\left(\pi_t - \pi_t^e - \frac{\lambda}{2}\pi_t^2\right), \tag{2.2}$$

with $\lambda := \frac{\tilde{\lambda}}{a}$. This type of benefit function is used in many of the now classical models, see Barro and Gordon (1983a,b) for example.[6] Note that in our setup, expression (2.2) can be identified with an instantaneous monetary payoff in real terms. We take this point of view in the remainder of this paper.

We assume that private agents build their expectation on the actual inflation rate, following an adaptive learning process

$$d\pi_t^e = \gamma(\pi_t - \pi_t^e)dt, \tag{2.3}$$

where the parameter γ is a measure for the speed of private agents adjustment from expected inflation to actual inflation. Models of this form are well known in the literature of learning and have been used in Phelps (1967), for example. For a general discussion, see also Evans and Honkapohja (2001). More precisely, we assume that actual inflation is given by

$$\pi_t = u_t + \sigma \dot{W}_t, \tag{2.4}$$

where u_t is the inflation rate chosen by the central bank and $\dot{W}_t$ is a generalized white noise process, i.e., the continuous analogue to an i.i.d. sequence of independent normal distributed random variables in a discrete time model. The generalized white noise process $\dot{W}_t$ satisfies the relationship $\int_0^t \dot{W}_s ds = W_t$ with W_t a standard Brownian motion, which will lead us into a Brownian-motion-based framework.[7] Using this, substitution of (2.4) into (2.3) leads to

$$d\pi_t^e = \gamma(u_t - \pi_t^e)dt + \gamma\sigma\, dW_t. \tag{2.5}$$

The stochastic learning dynamics (2.5) may also be interpreted as the result of a filtering process and interpreted in this way, it is related to rational expectations under partial information. Within a discrete time framework, Pearlman, Currie, and Levine

[6]It is still used in much of the more recent literature, where (2.2) is interpreted as a linear quadratic approximation of the equilibrium conditions under a deterministic steady state; compare Driffil et al. (2006).

[7]See, for example, Oksendal (2000, p. 21) or Hida (1980).

(1986) show how a discrete time version of (2.5) can be obtained from the assumption of rational expectations under partial information. In continuous time under Brownian uncertainty, this is far more complex, and up to this day not fully understood. As Yannacopoulos (2008) demonstrates, the mathematical correct setup for continuous-time rational expectation models under Brownian uncertainty is the framework of forward–backward stochastic differential equations (FBSDEs). The problem considered in the current paper, however, also involves a stochastic optimal control problem, an element that is missing in Yannacopoulos (2008), and that significantly complicates things here. The framework likely to evolve from a full rational expectation assumption is that of a stochastic control problem for FBSDEs. It is not clear whether analytic solutions as in the current paper could be obtained from such an assumption. We leave this interesting problem open for future research.

For $\gamma \to \infty$, the dynamics (2.5) formally approaches rational expectations[8] and further for finite γ, we obtain that $\lim_{t\to\infty} \mathbb{E}(\pi_t^e) = \lim_{t\to\infty} \mathbb{E}(u_t)$ if any of the two limits exist. Additionally, we can assume that $\pi_0^e = u_0$, i.e., we have rational expectations for the initial inflation estimate. Taking all of the above into account, the proposed model is hence not too far away from a full rational expectation model. Finally, Ewald and Geissler (2012) present a full rational expectation framework for the case where the central bank's inflation policy is dynamically constrained and inflation-linked bonds (ILBs) are replaced by inflation-linked calls. The results obtained there confirm the results obtained in the following sections of the current paper under a full rational expectation hypothesis.

We note that the variance of $\dot{W}_t$ is infinite, and therefore in order to guarantee that the instantaneous payoff function (2.2) has finite expectation, we have to replace π_t^2, which would have been a more natural candidate otherwise, with u_t^2. This is justified, noticing that the infinite variance part of π_t does not depend on u_t and would affect the instantaneous payoff functional (2.2) always in the same way, independent of what value for u_t is chosen. Hence, we do not really change the central bank's optimization problem by replacing π_t^2 with u_t^2.[9]

We denote with r the central bank's time preference parameter for measuring future against current objectives. Note that, in general, r does not have to coincide with the nominal market interest rate, which we will later denote by r_i.

[8]Note though that an equilibrium obtained under (2.5) does not necessarily converge to a rational expectation equilibrium, if $\gamma \to \infty$.

[9]More precisely, for a small but not infinitesimally small time interval Δt, in approximation, we have $\pi_t = u_t + \sigma \frac{\Delta W_t}{\Delta t}$ and substitution in (2.2) as well as taking expectations gives

$$\begin{aligned}
\mathbb{E}\left(a\left(\pi_t - \pi_t^e - \frac{\lambda}{2}\pi_t^2\right)\right) &= a\mathbb{E}\left(u_t + \sigma\frac{\Delta W_t}{\Delta t} - \pi_t^e - \frac{\lambda}{2}\left(u_t + \sigma\frac{\Delta W_t}{\Delta t}\right)^2\right) \\
&= a\mathbb{E}\left(u_t - \pi_t^e - \frac{\lambda}{2}u_t^2\right) + a\sigma\mathbb{E}\left(\frac{\Delta W_t}{\Delta t}\right) \\
&\quad - a\sigma\lambda\mathbb{E}\left(u_t\frac{\Delta W_t}{\Delta t}\right) - a\sigma^2\frac{\lambda}{2}\mathbb{E}\left(\left(\frac{\Delta W_t}{\Delta t}\right)^2\right) \\
&= \mathbb{E}\left(a\left(u_t - \pi_t^e - \frac{\lambda}{2}u_t^2\right)\right) - a\sigma^2\frac{\lambda}{2}\frac{1}{\Delta t},
\end{aligned}$$

as $\mathbb{E}(\frac{\Delta W_t}{\Delta t}) = 0$ and $\mathbb{E}(u_t\frac{\Delta W_t}{\Delta t}) = \mathbb{E}(u_t\mathbb{E}_t(\frac{\Delta W_t}{\Delta t})) = 0$. While the second term in the last line of equations above tends to infinity for $\Delta t \to 0$, for each fixed Δt, it is finite and independent of u_t and hence does not affect the optimization.

Corresponding to the inflation rate π_t is the price level P_t, which follows the dynamic

$$dP_t = P_t\pi_t dt = P_t(u_t dt + \sigma\, dW_t), \tag{2.6}$$

with $P_0 = 1$, i.e., the initial price level is normalized to 1. In this paper, we consider the following inflation-indexed securities:

DEFINITION 2.1. An ILB with maturity time T issued at time $s \in [0, T]$ is a financial contract that pays off $\frac{P_T}{P_s}$ Dollar at time T.

The contract in Definition 2.1 is idealized. Products traded in reality feature a slightly more complicated payoff structure. For example, the final payment at maturity of a Treasury Inflation Protected Security (TIPS) issued in the United States corresponds to an inflation-linked call,[10] while the payments prior to maturity, i.e., the coupons, correspond to ILBs.[11] For further details on the specifics of these contracts, we refer to the website of the US Treasury.[12] In terms of pricing of TIPS, it is standard to "strip" the contracts into its individual components (i.e., bonds and calls) and price each of them individually; see, for example, Jarrow and Yilderim (2003). Hence, treating ILBs and calls as individual entities bears some justification. For a comprehensive overview about what contracts are traded, we refer to Deacon et al. (2004).

Assuming that the central bank sells a number N of ILBs on the primary market, which for simplicity we assume are all of the same maturity T and issued at the same time $s \in [0, T]$, the central bank enters a liability in terms of a dollar payment of $N \cdot \frac{P_T}{P_s}$ at time T. At time s, any of the individual ILBs can be structured into a zero coupon bond B_s with face value 1 dollar and the inflation compensation $\frac{P_T - P_s}{P_s}$. We assume that the market interest rate r_i is deterministic, and hence an ordinary zero coupon bond is valued at $B_s = e^{-r_i(T-s)}$. We further assume that the central bank buys these zero coupon bonds from private banks, which will take them out of the balance sheet of the central bank. The inflation compensation of $\frac{P_T - P_s}{P_s}$ Dollar, however, remains as the central bank's liability.

The central bank optimizes in real terms with reference to time s, which is when the central bank makes the decision to sell a number N of ILBs on the market. The liability in real terms with reference to time s for the central bank from selling a number N of ILBs then becomes

$$-N\left(\frac{P_T - P_s}{P_s}\right) \cdot \left(\frac{P_T}{P_s}\right)^{-1} = -N\frac{P_T - P_s}{P_T} = -N\left(1 - \frac{P_s}{P_T}\right). \tag{2.7}$$

The central bank's instantaneous real term payoff (which is also in reference to time s) is given by (2.2). This has to be accumulated over time and measured against the terminal real term payoff from above. Its objective in this setup can hence be formulated as

[10] Which like a European call option is capped from below, keeping the upside potential.

[11] Other countries including the United Kingdom, Germany, France, Australia, and Sweden sell similar products.

[12] Website of the US treasury at http://www.treasurydirect.gov/indiv/research/indepth/tips/res_tips_rates.htm.

$$\hat{V}(t, \pi^e, P; N) := \max_{u_v} \mathbb{E}\left(a \int_t^T e^{-r(v-t)} \left(u_v - \pi_v^e - \frac{\lambda}{2} u_v^2 \right) dv \right.$$
$$\left. - e^{-r(T-t)} N \left(1 - \frac{P_s}{P_T} \right) \middle| \pi_t^e = \pi^e, P_t = P \right) \tag{2.8}$$

subject to (2.5) and (2.6). We assume for the moment that the number of ILBs N, which appears as a parameter in the value function (2.8), is given. Nevertheless, in the following sections, N will become a choice variable.[13] Note that the case $N = 0$ corresponds to the case where no ILBs are sold. Also, note that the discount factor $e^{-r(T-t)}$ in (2.8) comes from the central bank's time preference, and is not a priori a financial discount factor. Mathematically, problem (2.8) represents a stochastic optimal control problem, which we will approach by classical Hamilton–Jacobi–Bellman theory. This, however, is only possible if $t \geq s$ as otherwise the stochastic state variable P will have to be evaluated at two different times in the future, s and T, a constellation that does not adhere to the standard HJB-framework and would be more difficult to handle. Nevertheless, to solve the central bank's decision problem at time s, we may assume without loss of generality that $t \geq s$, and hence P_s is a known constant at time t.

In order to obtain a compact analytical solution, we make the following simplifying assumption:

$$1 - \frac{P_s}{P_T} \approx \log\left(\frac{P_T}{P_s} \right). \tag{2.9}$$

This approximation is valid for reasonable inflation rates of up to 8%. Using this approximation, the optimal control problem of the central bank becomes

$$V(t, \pi^e, P; N) := \max_{u_v} \mathbb{E}\left(a \int_t^T e^{-r(v-t)} \left(u_v - \pi_v^e - \frac{\lambda}{2} u_v^2 \right) dv \right.$$
$$\left. - e^{-r(T-t)} N \log\left(\frac{P_T}{P_s} \right) \middle| \pi_t^e = \pi^e, P_t = P \right) \tag{2.10}$$

subject to (2.5) and (2.6). Note that V is still in terms of real money at time s and that eventually we will be concerned with evaluating (2.10) at time $t = s$. In addition, note that $V(t, \pi^e, P; 0) = \hat{V}(t, \pi^e, P; 0)$ and the quality of the approximation also depends on the number N of ILBs sold.

For $N > 0$, in general, we have that $V(s, \pi^e, P; N) \neq V(s, \pi^e, P; 0)$. Clearly, the central bank would want to be appropriately compensated, and therefore charge a price for the ILBs reflecting the difference $V(s, \pi^e, P; 0) - V(s, \pi^e, P; N)$ as well as the value of the zero coupon bond component. As V is in real terms, this then has to be converted via the actual price level P_s at time s into nominal terms. We follow up on this in Section 4 after discussing the effect of selling ILBs on the central banks optimal monetary policy.

[13] To emphasize that N is not a state variable, we have separated it from the other variables with a semicolon.

3. OPTIMAL MONETARY POLICY WITH ILBs

In this section, we attempt to solve the central bank's stochastic optimal control problem (2.10). For convenience, we divide the value function V from (2.10) by the parameter a:

(3.1)
$$\begin{aligned}\tilde{V}(t,\pi^e,P;N) &:= \frac{1}{a}V(t,\pi^e,P,N)\\ &= \max_{u_v}\mathbb{E}\left(\int_t^T e^{-r(v-t)}\left(u_v-\pi_v^e-\frac{\lambda}{2}u_v^2\right)dv - e^{-rT}\frac{N}{a}\log\left(\frac{P_T}{P_s}\right)\middle|\,\pi_t^e=\pi^e,\,P_t=P\right)\end{aligned}$$

subject to (2.5) and (2.6). The Hamilton–Jacobi–Bellman equation is then given by

$$\begin{aligned}0 = \max_u\Big(&u-\pi^e-\frac{\lambda}{2}u^2+u\left(\gamma\tilde{V}_{\pi^e}+\tilde{V}_P P\right)-\gamma\tilde{V}_{\pi^e}\pi^e\Big)\\ &+\frac{1}{2}\sigma^2\left(\tilde{V}_{PP}P^2+\tilde{V}_{\pi^e\pi^e}\gamma^2+2\tilde{V}_{\pi^eP}P\gamma\right)+\tilde{V}_t-r\tilde{V},\end{aligned}$$

and the optimal policy u^* is obtained from the first-order condition

$$u^* = \frac{1}{\lambda}(1+\gamma\tilde{V}_{\pi^e}+\tilde{V}_P P).$$

Substituting this into the PDE for the value function, we get

(3.2)
$$\begin{aligned}0 &= \frac{1}{\lambda}(1+\gamma\tilde{V}_{\pi^e}+\tilde{V}_P P)-\pi^e-\frac{1}{2\lambda}(1+\gamma\tilde{V}_{\pi^e}+\tilde{V}_P P)^2\\ &\quad+\frac{1}{\lambda}(1+\gamma\tilde{V}_{\pi^e}+\tilde{V}_P P)(\gamma\tilde{V}_{\pi^e}+\tilde{V}_P P)\\ &\quad-\gamma\tilde{V}_{\pi^e}\pi^e+\frac{1}{2}\sigma^2(\tilde{V}_{PP}P^2+\tilde{V}_{\pi^e\pi^e}\gamma^2+2\tilde{V}_{\pi^eP}P\gamma)+\tilde{V}_t-r\tilde{V}\\ &= \frac{1}{2\lambda}(1+\gamma\tilde{V}_{\pi^e}+\tilde{V}_P P)^2-\pi^e(1+\gamma\tilde{V}_{\pi^e})\\ &\quad+\frac{1}{2}\sigma^2(\tilde{V}_{PP}P^2+\tilde{V}_{\pi^e\pi^e}\gamma^2+2\tilde{V}_{\pi^eP}P\gamma)+\tilde{V}_t-r\tilde{V}.\end{aligned}$$

The value function $\tilde{V}$ has to satisfy (3.2) subject to the terminal condition

(3.3)
$$\tilde{V}(T,\pi^e,P;N) = -\frac{N}{a}\log\left(\frac{P}{P_s}\right).$$

It is not hard to verify that a solution is given by

(3.4)
$$\tilde{V}(t,\pi^e,P;N) = -\frac{N}{a}\log\left(\frac{P}{P_s}\right)e^{-r(T-t)}+A_t\pi^e+C_t+D_t(N)$$

with

(3.5)
$$A_t = \frac{1}{\gamma+r}(e^{-(\gamma+r)(T-t)}-1)$$

$$C_t = \frac{1}{\lambda(\gamma+r)^2}\Big[\frac{r}{2}(1+e^{-r(T-t)}) - re^{-(\gamma+r)(T-t)} + \frac{\gamma^2}{2(2\gamma+r)}(e^{-r(T-t)} - e^{-2(\gamma+r)(T-t)})\Big] \tag{3.6}$$

$$D_t(N) = e^{-r(T-t)}\Big[\frac{N^2}{2a^2\lambda r}(1-e^{-r(T-t)}) + \frac{N\sigma^2}{2a}(T-t) - \frac{N}{a\lambda(\gamma+r)^2}(r(\gamma+r)(T-t)+\gamma(1-e^{-(\gamma+r)(T-t)}))\Big]. \tag{3.7}$$

For the optimal monetary policy given that N ILBs are sold, we therefore conclude

$$u_t^*(N) = \frac{1}{\lambda(\gamma+r)}\cdot(\gamma e^{-(\gamma+r)(T-t)}+r) - \frac{Ne^{-r(T-t)}}{a\lambda}, \tag{3.8}$$

for $t \geq s$. Note that we do not look at how optimal monetary policy is affected before the sale takes place, in other words, when the central bank has the (real) option to sell ILBs at some given point in the future. We consider this and related problems in future work.

The case $N = 0$ corresponds to the case where the central bank does not sell any ILBs; hence, we clearly see from (3.8) that the sale of ILBs causes the central bank to aim for lower inflation. It is also noteworthy to say that the optimal strategy above is deterministic and does not depend on feedback from π_t^e or P_t, no matter whether ILBs are sold or not. On the other hand, it crucially depends on the learning rate γ.

In the following, we denote the corresponding processes of inflation, expected inflation, price level, etc., under the optimal monetary policy u_t^* with the superscript *. For particular levels π_s^e and P_s at time s, we have[14]

$$\pi_t^*(N) = \frac{1}{\lambda(\gamma+r)} + (\gamma e^{-(\gamma+r)(T-t)}+r) - \frac{Ne^{-r(T-t)}}{a\lambda} + \sigma\dot{W}_t, \tag{3.9}$$

(3.10)

$$\pi_t^{e^*}(N) = \pi_s^e e^{-\gamma(t-s)} + \frac{\gamma^2 e^{-(\gamma+r)T}e^{-\gamma(t-s)}}{\lambda(\gamma+r)(2\gamma+r)}(e^{(2\gamma+r)t} - e^{(2\gamma+r)s}) + \frac{r}{\lambda(\gamma+r)}(e^{\gamma t} - e^{\gamma s}) - \frac{N\gamma e^{-rT}e^{\gamma(t-s)}}{a\lambda(r-\gamma)}(e^{(r-\gamma)t} - e^{(r-\gamma)s}) + \gamma\sigma e^{-\gamma(t-s)}\int_s^t e^{\gamma v}dW_v,$$

and

$$\begin{aligned}\log P_t^*(N) &= \log P_s + \int_s^t \pi_v dv - \frac{1}{2}\sigma^2(t-s)\\ &= \log P_s + \frac{\gamma}{\lambda(\gamma+r)^2}(e^{-(\gamma+r)(T-t)} - e^{-(\gamma+r)(T-s)}) + \frac{r(t-s)}{\lambda(\gamma+r)}\\ &\quad - \frac{N}{a\lambda r}(e^{-r(T-t)} - e^{-r(T-s)}) - \frac{1}{2}\sigma^2(t-s) + \sigma(W_t - W_s),\end{aligned} \tag{3.11}$$

with $s \leq t \leq T$.

[14]For $r = \gamma$, the expression $\frac{N\gamma e^{-rT}e^{\gamma(t-s)}}{a\lambda(r-\gamma)}(e^{(r-\gamma)t} - e^{(r-\gamma)s})$ in $\pi_t^{e^*}(N)$ needs to be replaced by $\frac{N\gamma e^{-rT}e^{\gamma(t-s)}}{a}(t-s)$.

We conclude from (3.9), (3.10), and (3.11) that inflation, expected inflation, and price level are all lower, when the central bank sells ILBs. Note that this inflation reduction does not come at a cost to the central bank, as long as it charges a real price for the ILBs' inflation compensation component that is larger than $\frac{V(s,\pi^e,P_s,0)-V(s,\pi^e,P_s,N)}{N}$ per ILB. The actual amount N of ILBs issued by the central bank will be determined in Section 6 from a supply-demand equilibrium. To do so, we first consider the supply side and the demand side individually in the next two sections.

4. CENTRAL BANK'S PRICING OF ILBs

Within the setup developed in the previous sections, we now derive a pricing formula for ILBs, when demand N is given. This relationship will provide us with the supply curve for ILBs. Let us first note that

$$\tilde{V}(t,\pi^e,P;N)-\tilde{V}(t,\pi^e,P;0)=D_t(N)-\frac{N}{a}e^{-r(T-t)}\log\left(\frac{P}{P_s}\right),$$

with $D_t(N)$ as defined in (3.7) and hence at time $t=s$ when the current price level is $P=P_s$

$$\tilde{V}(s,\pi^e,P_s;N)-\tilde{V}(s,\pi^e,P_s;0)=D_s(N). \tag{4.1}$$

To exclude expected deflation, which can hardly be a central bank's objective, we need

$$\log\left(\frac{\mathbb{E}_s\left(P_T^*(N)\right)}{P_s}\right)\geq 0. \tag{4.2}$$

From (3.11), it can be easily verified that

$$\log\left(\frac{\mathbb{E}_s\left(P_T^*(N)\right)}{P_s}\right)=\frac{\gamma}{\lambda(\gamma+r)^2}(1-e^{-(\gamma+r)(T-s)})+\frac{r(T-s)}{\lambda(\gamma+r)}-\frac{N}{a\lambda r}(1-e^{-r(T-s)}),$$

and therefore, we require

$$0\leq N\leq\frac{ar}{(1-e^{-r(T-s)})(\gamma+r)^2}(r(\gamma+r)(T-s)+\gamma(1-e^{-(\gamma+r)(T-s)})). \tag{4.3}$$

This is a condition on the maximum number of ILBs that the central bank can issue, without intentionally causing deflation. We assume that condition (4.3) holds for the remainder of this paper.

The loss in real terms per ILB issued by the central bank is given by

$$\frac{V(s,\pi^e,P;0)-V(s,\pi^e,P;N)}{N}=-a\frac{D_s(N)}{N},$$

and therefore, the central bank's (indifference) price in terms of nominal money at time s for this component yields

$$P_s\frac{V(s,\pi^e,P_s,0)-V(s,\pi^e,P_s,N)}{N}=-P_s\left(a\frac{D_s(N)}{N}\right). \tag{4.4}$$

Expression (4.4) can be positive or negative. On top of this, however, the central bank has to charge the price of the zero coupon bond component, leading to a minimum

acceptable price for one ILB, given that N ILBs are issued by the central bank, of

$$p_s(N) = e^{-r_i(T-s)} - P_s\left(a\frac{D_s(N)}{N}\right).$$

In reality, inflation-indexed securities are mostly sold in the form of auctions, single or multiple price, where the central bank is setting a minimum acceptable price as above; compare Deacon et al. (2004) or Price (1997). An auction framework in the context of this paper, however, would turn out to be far too complex and difficult to handle, which is why we assume here that the central bank earns normal profits only.[15] The expression for $p_s(N)$ above can be easily computed with help of equation (3.7), leading to

(4.5)
$$p_s(N) = P_s\Big[e^{-r(T-s)}\left(\frac{1}{\lambda(\gamma+r)^2}[r(\gamma+r)(T-s)+\gamma(1-e^{-(\gamma+r)(T-s)})] - \frac{\sigma^2}{2}(T-s)\right) - \frac{Ne^{-r(T-s)}}{2a\lambda r}(1-e^{-r(T-s)})\Big] + e^{-r_i(T-t)}.$$

The price of one ILB is therefore an affine linear function that is decreasing in the number N of ILBs issued. The expression (4.5) can also be interpreted as the central banks average cost curve. It is noteworthy that the price derived in (4.5) does not depend on the current (expected) inflation rate. The reason for this is that in (3.4), the factor A_t does not depend on N and hence cancels out in (2.2). However, it crucially depends on private agents learning rate γ. This is intuitively clear since this parameter rather than the actual level of expected inflation has an impact on the central bank's policy. The price also depends on the price level P_s at the time of the sale.

5. THE FINANCIAL MARKET

In the previous section, we have seen how the central bank would set the price for an ILB depending on the number of ILBs issued. We now look at the demand side and consider what agents acting on financial markets would be prepared to pay for the ILBs. To do so, we need to set up a model for an underlying financial market which we assume to exist, no matter whether ILBs are traded or not and sits on top of what has been introduced in Section 2

We assume that nominal bonds are traded as primary assets and financial agents are averse against inflation risk. Due to the stochastic price level (2.6), nominal bonds have risky real returns and therefore carry an inflation risk premium. Let r_i denote the nominal interest rate and B_t the nominal value of the nominal bond, then the financial market in primary assets is described by the two dynamics

(5.1)
$$dB_t = B_t r_i dt,$$
$$dP_t = P_t(u_t dt + \sigma dW_t).$$

The inflation risk premium IRP is then characterized by the relationship

(5.2)
$$IRP = \mathbb{E}_t\left(\frac{d(B_t/P_t)}{B_t/P_t}\right) - \text{real return of ILB}.$$

[15]It could be argued as well that any excessive supernormal profits for the central bank resulting from excessively high bids in single price auctions, from the perspective of the central bank, are random, and hence cannot be accounted for strategically. In particular, the central bank does not act like a monopolist.

Empirically, this inflation risk premium has been determined by Gong and Remolona (1996) as well as Campbell et al. (2009) and lies in the range of 50–150 basis points for 5.1 year bonds. We will determine the price of the ILBs in a way that is consistent with this.

Note that since the price level P_t is not a tradable asset, (5.1) as such describes an incomplete market. Without ILBs, inflation risk cannot be fully hedged, which given that financial agents are averse toward inflation risk is in principle enough to explain the demand for ILBs. In an arbitrage-free framework, the actual level of ILBs demanded is then determined by the equivalent martingale measure[16] that financial agents impose. As the market is incomplete, there is no unique such measure, but as we will see, a choice of an equivalent martingale measure will correspond to a choice of an inflation risk premium. For this, let

$$\tilde{W}_t = W_t + \theta t$$

be a Brownian motion under the chosen equivalent martingale measure $\tilde{\mathbb{P}}$.[17] Then,

$$dP_t = P_t((u_t - \sigma\theta)dt + \sigma d\tilde{W}_t).$$

Given optimal behavior of the central bank when N ILBs are sold, the price-level process (2.6) under the pricing measure appears as

$$P_T^*(N) = P_s e^{\int_s^T (u_v^*(N) - \sigma\theta - \frac{1}{2}\sigma^2)dv + \int_s^T \sigma d\tilde{W}_v}.$$

Hence, denoting with $\tilde{\mathbb{E}}_s$ the expectation under the pricing measure $\tilde{\mathbb{P}}$, conditioned on information available at time s, the arbitrage-free price of one ILB sold at time s is given by

$$\begin{aligned}\tilde{p}_s(N) &= e^{-r_i(T-s)}\tilde{\mathbb{E}}_s(P_T^*(N)) = P_s e^{-r_i(T-s)}\tilde{\mathbb{E}}_s\left(e^{\int_s^T (u_v^*(N) - \sigma\theta - \frac{1}{2}\sigma^2)dv + \int_s^T \sigma d\tilde{W}_v}\right)\\ &= P_s e^{-(r_i+\sigma\theta)(T-s)} e^{\int_s^T u_v^*(N)dv}.\end{aligned}$$

Note that we used here that the optimal policy u^* is deterministic, see (3.8). The integral in the exponent can be easily computed as

$$\begin{aligned}\int_s^T u_v^*(N)dv = &\frac{1}{\lambda(\gamma + r)^2}[r(\gamma + r)(T - s) + \gamma(1 - e^{-(\gamma+r)(T-s)})]\\ &- \frac{N}{a\lambda r}(1 - e^{-r(T-s)}).\end{aligned}$$

The last formula covers the demand side, once the inflation risk premium is identified. It is not difficult now to show that the (expected) real returns of the nominal bond resp. the ILB are given by

$$\begin{aligned}\mathbb{E}_t\left(\frac{d(B_t/P_t)}{B_t/P_t}\right) &= r_i - u_t^*(N) + \sigma^2,\\ \frac{d(p_t(N)/P_t)}{(p_t(N)/P_t)} &= r_i - u_t^*(N) + \sigma\theta,\end{aligned}$$

[16] In other words, the pricing kernel.

[17] The measure $\tilde{\mathbb{P}}$ or more precisely its density is obtained by the so-called Girsanov transformation.

and therefore by using (5.2) that the inflation risk premium is given by

$$IRP = \sigma^2 - \sigma\theta. \tag{5.3}$$

Adding ILBs with the price functional above completes the financial market, which then guarantees that the secondary market is in a supply-demand equilibrium, compare Karatzas and Shreve (1998, corollary 5.4, p. 173)[18] and Karatzas (1997, proposition 3.4.1, p. 65).

We now have formulas for the central bank's price $p_s(N)$ (supply side) as well as the private agent's price $\tilde{p}_s(N)$ (demand side) and are ready to investigate for what N market equilibrium in the primary market for ILBs is obtained, i.e., $p_t(N) = \tilde{p}_t(N)$.

6. EQUILIBRIUM IN THE MARKET FOR ILBs

Concluding from the previous two sections, there will be excess demand from the private sector for ILBs as long as $p_s(N) \leq \tilde{p}_s(N)$. The latter is equivalent to

$$\begin{aligned}\left(e^{-r_i(T-s)} - a P_s \frac{D_s(N)}{N}\right) &\leq e^{-r_i(T-s)}\tilde{\mathbb{E}}_s P_T^*(N) \\ &= e^{-r_i(T-s)} e^{-\sigma \frac{\mu - r_i}{\sigma_S}(T-s)} \mathbb{E}_s P_T^*(N),\end{aligned}$$

which, in turn, is equivalent to

$$0 \leq e^{-r_i(T-s)}\left(\tilde{\mathbb{E}}_s P_T^*(N) - 1\right) + P_s\left(\frac{a D_s(N)}{N}\right). \tag{6.1}$$

Previously, we have seen that $\tilde{\mathbb{E}}_s P_T^*(N) = P_s e^{\int_s^T u_v^*(N) - \sigma\theta dv}$ and

$$\int_s^T u_v^*(N) dv = \eta_{T-s} - x_{T-s}, \tag{6.2}$$

with

$$\begin{aligned}\eta_{T-s} &:= \frac{1}{\lambda(\gamma + r)^2}[r(\gamma + r)(T-s) + \gamma(1 - e^{-(\gamma+r)(T-s)})] \\ x_{T-s} &:= \frac{N}{a\lambda r}(1 - e^{-r(T-s)}).\end{aligned}$$

Further using (2.7) and the notation of η_{T-s} and x_{T-s} introduced above, we can write

$$a\frac{D_s(N)}{N} = e^{-r(T-s)}\left[\frac{1}{2}x_{T-s} + \frac{\sigma^2}{2}(T-s) - \eta_{T-s}\right],$$

and hence equality in equation (6.1) is equivalent to

$$0 = P_s e^{-(r_i + \sigma\rho)(T-s)} e^{\eta_{T-s}} e^{-x_{T-s}} - e^{-r_i(T-s)} + P_s e^{-r(T-s)}\left(\frac{1}{2}x_{T-s} + \frac{\sigma^2}{2}(T-s) - \eta_{T-s}\right). \tag{6.3}$$

[18] First Edition, Chapter 4: Equilibrium in a Complete Market.

We write (6.3) as

$$0 = e^{-x_{T-s}} + bx_{T-s} - c, \tag{6.4}$$

with

$$b = \frac{e^{-r(T-s)}}{2e^{\eta_{T-s}}e^{-(r_i+\sigma\rho)(T-s)}},$$
$$c = \frac{1}{e^{\eta_{T-s}}e^{-(r_i+\sigma\rho)(T-s)}}\left[\frac{e^{-r_i(T-s)}}{P_s} + e^{-r(T-s)}\left(\eta_{T-s} - \frac{\sigma^2}{2}(T-s)\right)\right].$$

A solution of equation (6.4) is then given by

$$x_{1,2} = \frac{c}{b} + \mathbb{W}\left(-\frac{e^{-\frac{c}{b}}}{b}\right) = \frac{2}{P_s}e^{(r-r_i)(T-s)} + 2\eta_{T-s} - \sigma^2(T-s) + \mathbb{W}\big(-2e^{\eta_{T-s}}e^{(r-r_i-\sigma\rho)(T-s)}e^{\sigma^2(T-s)-\frac{2}{P_s}e^{(r-r_i)(T-s)}-2\eta_{T-s}}\big), \tag{6.5}$$

where $\mathbb{W}$ denotes the Lambert W-function; see Abramowitz and Stegun (1965). The subindices in $x_{1,2}$ indicate that there are possibly two solutions. Mathematically, the reason for this is that for arguments $-\exp(-1) < x < 0$, the Lambert function delivers two values instead of one. The two possible values for x correspond to two possible values for N via the relationship $-\frac{N}{a\lambda r}(1 - e^{-r(T-s)}) = x$ introduced earlier. We argue in the following that economically, only the smaller of the two possible values is relevant. For this, let $f(x)$ denote the right-hand side of (6.4). Obviously, f is a convex function in x with exactly one local minimum which we denote with x^*. Hence, there are solutions to (6.4) if and only if $f(x^*) \le 0$. Furthermore, f is obviously increasing in x for all $x \ge x^*$ and decreasing for all $x \le x^*$. If this is the case, let $x_1 \le x^* \le x_2$ denote the roots of f as given above. On the other hand (as seen before), there is always demand if $f(x) \ge 0$ and hence once $x \ge x_2$ for any corresponding N, there will always be demand. However, there will be no demand if N is such that $x_1 < x < x_2$, while there will be demand if $x \le x_1$. In that sense, the central bank can start issuing ILBs until N is such that $x = x_1$ and a locally stable equilibrium is attained—i.e., as long as N does not jump to $x = x_2$, the bank cannot issue more ILBs. From this analysis, we conclude the following:

PROPOSITION 6.1. *Let us denote with x_1 the smaller of the two values in equation (3.5). If $x_1 > 0$, then the equilibrium quantity of ILBs (if issued at time s) is given by*

$$N^* = \frac{a\lambda r}{(1-e^{-rT})}\left[\frac{2}{P_s}e^{(r-r_i)(T-s)} + 2\eta_{T-s} - \sigma^2(T-s) + \mathbb{W}\big(-2e^{\eta_{T-s}}e^{(r-r_i-\sigma\rho)(T-s)}e^{\sigma^2(T-s)-\frac{2}{P_s}e^{(r-r_i)(T-s)}-2\eta_{T-s}}\big)\right], \tag{6.6}$$

where the value of the Lambert function is chosen as the smaller of the possibly two values. The equilibrium price is determined by equation (4.5) as

$$p_s(N^*) = P_s e^{-(r_i+\sigma\theta)(T-s)+\eta_{T-s}-x_{T-s}}. \tag{6.7}$$

In general, it is not guaranteed that $x_1 \ge 0$. This depends on the parameters—it also depends on the time to maturity T, which suggests that the central bank would need to dynamically control the ILB supply. Note that while in our analysis, the time s at which

the central bank can issue ILBs is arbitrary, the central bank can issue ILBs only once. This, of course, does not mean that trade of ILBs terminates after the central bank has issued the ILBs. To the contrary, ILBs will be traded on the secondary market among private agents and priced within the arbitrage-free framework developed in Section 5. These trades, however, will not feed back into inflation, and therefore do not affect monetary policy.[19] In an extension of the model presented here, one could, however, allow the central bank to issue ILBs at different times consecutively and, in fact, let the central bank choose these times. We leave this far more complicated analysis for future research.

Compared to the classical case (i.e., $N = 0$), the equilibrium with ILBs features lower inflation levels (equations (3.9) and (3.10)), while the central bank is fully compensated (in real terms) through the income from the sales of ILBs (equation (4.4)). As there is demand for ILBs on the financial market (i.e., $\tilde{p}_s(N) > 0$), we observe that all three parties, i.e., central bank, private agents, and financial agents, benefit from the existence of a market for ILBs and that, in fact, a Pareto improvement as compared to the case without a market for ILBs is obtained.

7. NUMERICAL RESULTS

Figure 7.1 shows the US Consumer Price Index (CPI-U), monthly from January 2009 until December 2011. From these data, we estimated the annual volatility of the price level as $\sigma = 0.0119$. Assuming a nominal interest rate of $r_i = 0.03$ and $\theta = -0.41$, we obtain an inflation risk premium of 50 basis points, comparing well with Gong and Remolona (1996) and Campbell et al. (2009).

In order to compute the optimal inflation policy for the central bank and in conclusion determine the equilibrium price and quantity of ILBs traded on the market, we choose the following set of parameters:[20] $\lambda = 20$, $a = 10^{11}$,[21] $r = 0.1$, and $\gamma = 0.1$. We are considering ILBs, maturing at time $T = 10$ (years), which are sold by the central bank at a time $s \in [0, T]$. For ease of notation, the price level at reference time s is normalized to one. The following Figure 7.2 shows the central banks' optimal inflation policy over the 10-year time period, after issuing the quantity N of ILBs at time $s = 0$.

We can clearly see that with an increasing quantity N of ILBs issued, the optimal inflation rate is reduced. This effect is less prominent at the beginning of the period, from

[19] This does not mean that the secondary private market for inflation-indexed securities should be ignored in general, neither has it been here. Financial intermediaries provide an essential service to the public in selling inflation-indexed products. In fact, in the United Kingdom, the 1995 Pensions Act requires defined benefit pension schemes to index pensions entitlements to retail price index inflation. These services are indirectly accounted for in our model via the ILB pricing formulas (4.4) and (6.3), which prices inflation risk, and hence also these services. Note, further, that most secondary issuers of inflation-indexed securities buy these from the central bank in the first place. Quoting Deacon et al. (2004, p. 69): "Wojnilower (1997) argues that the general lack of private sector issuance is because no company can make its indexing promise credible; only governments can guarantee to be able to honor such promises." We argue therefore that any feedback effects from private sales of inflation-indexed securities on the central banks monetary policy can only be marginally.

[20] See Section 2 for their interpretation.

[21] Note that the parameter a in our setup is the product of the slope of the Phillips curve and natural real output. The slope of the New Keynesian Phillips Curve has been estimated in Schorfheide (2008) to be in the range of 0.005–0.135, while UK GDP in 2011 lies at about 2.25×10^{12} in purchasing power parity adjusted US Dollars. Picking a slope of 0.05, say, and multiplying it with 2.25×10^{12} will then give a value of a in our model of about 10^{11}.

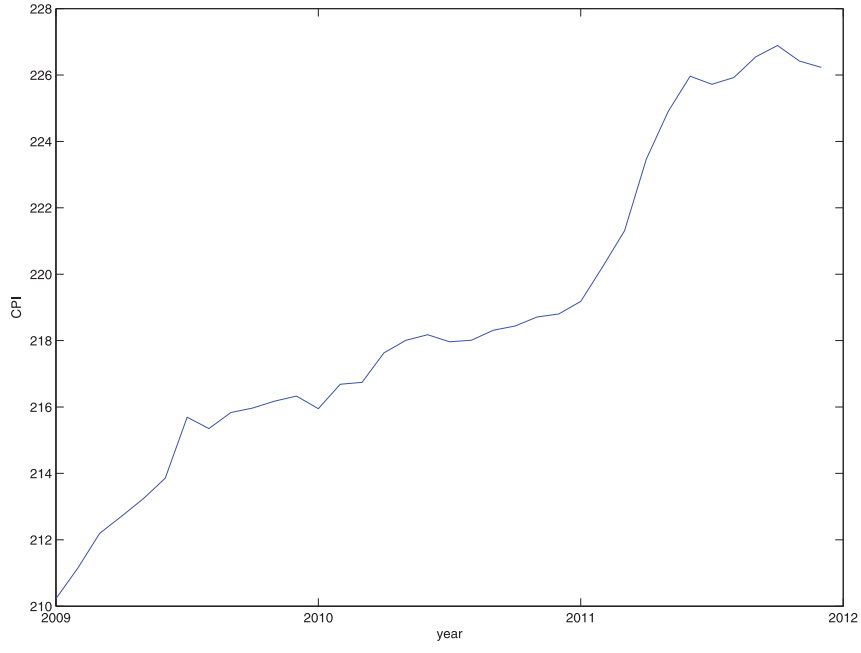

FIGURE 7.1. CPI-U, monthly from 2009 to 2011.

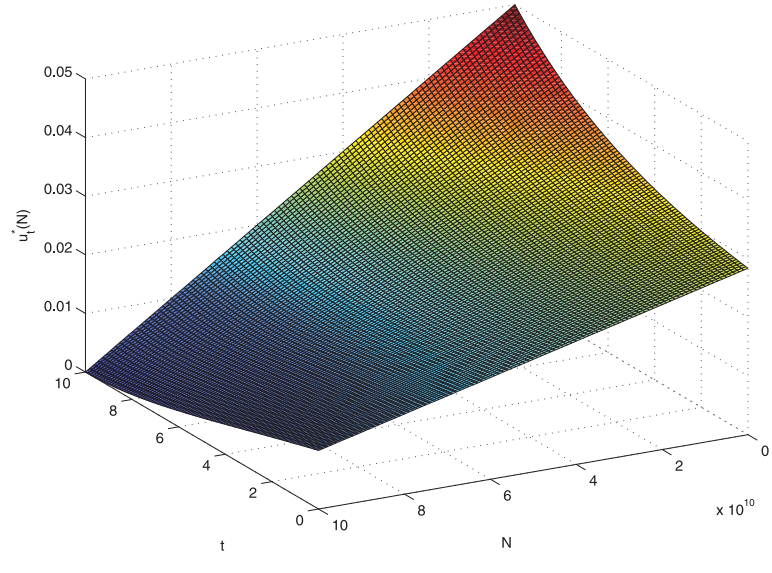

FIGURE 7.2. Optimal inflation policy $u_t^*(N)$ over time t, given that N ILBs have been sold.

about 3% to 1% at time $t = 0$ (in the case that $N = 10^{11}$ ILBs are issued), but increases in magnitude toward the end of the period, from about 5% to close to 0% at time $t = 10$. The question of how many ILBs the central bank is able to sell depends on the demand from

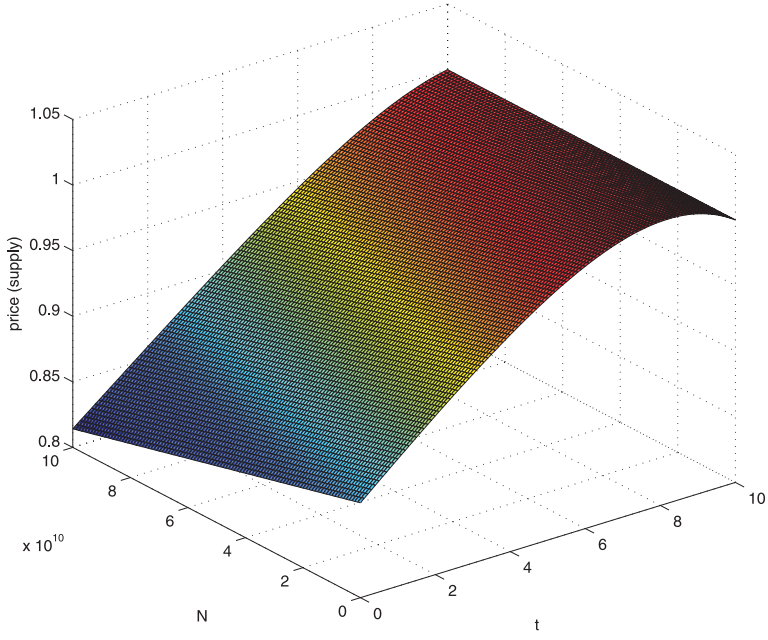

FIGURE 7.3. Supply-price function over time t for ILBs sold by the central bank. Price level normalized to one.

financial agents of course. Figures 7.3 and 7.4 represent the supply- and demand-price functions for ILBs as determined in Sections 4 and 5.

Note that the price of an ILB can, in principle, be higher than one, even if no coupons are paid, if the expected inflation rate is larger than the nominal interest rate (i.e., negative real interest rate, as observed in many countries at the moment). We will see that in our model in equilibrium, this will only occur when the ILB is issued at a very late stage, i.e., later than $t = 8.5$ years. Figures 7.5 and 7.6 show, respectively, the equilibrium price and equilibrium quantity of ILBs. If issued at time $s = 0$, the central bank would sell about 10×10^{10} for a price of 0.83 US dollars, a total monetary value of 83.4 billion US dollars. If issued at a later time, the central bank will sell less ILBs for higher and higher prices until the price trend reverses at about $s = 9$ years. This makes sense, of course, as there is little to lose from inflation over short periods of time. Though note that as s tends toward T, the quantity of ILBs issued does not approach 0, but a limit of roughly 2×10^{10}. Figure 7.5 also shows that if the central bank issues the ILBs at time $t = 8.5$ or later, their price is larger than 1. The reason for this can be identified from Figure 7.7: The equilibrium rate of inflation when ILBs are issued that late raises above the nominal interest rate during the remainder of the period (second and third curve from top), causing the real rate of interest to be negative. The top red curve in Figure 7.7 shows the central bank's optimal inflation policy, when the central bank is not given the opportunity to sell ILBs (i.e., no markets for ILBs). Comparing this to the bottom blue curve, which shows the optimal inflation policy, given that the equilibrium quantity of ILBs has been issued at time $s = 0$, we see that inflation is reduced from about 2.75% to 1% at the beginning of the 10-year period and from about 5%–0.1% at the end of the 10-year period.

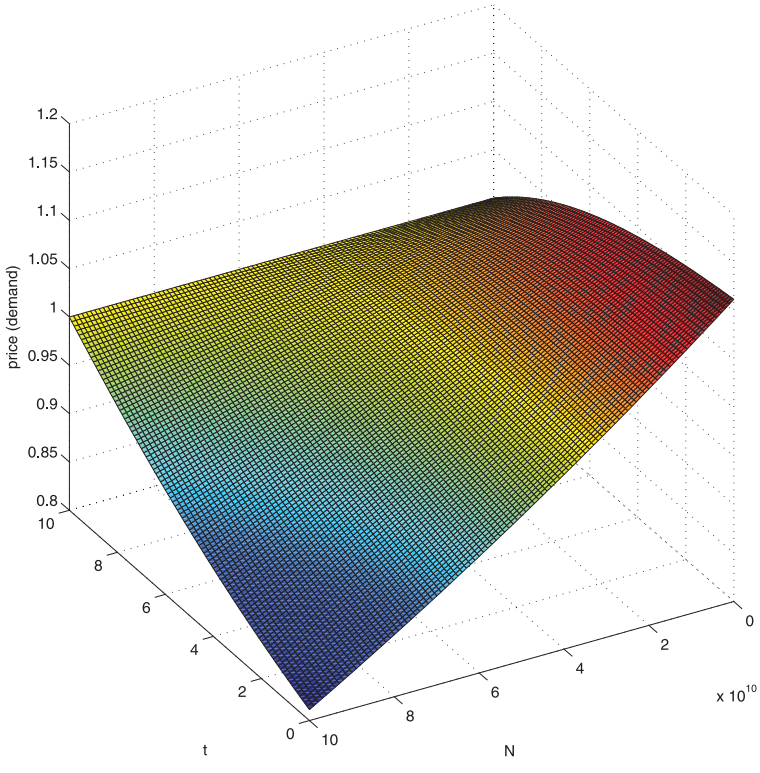

FIGURE 7.4. Demand-price function over time t for ILBs, for financial agents who attach an inflation risk premium of 50 bp to nominal bonds. Price level normalized to one.

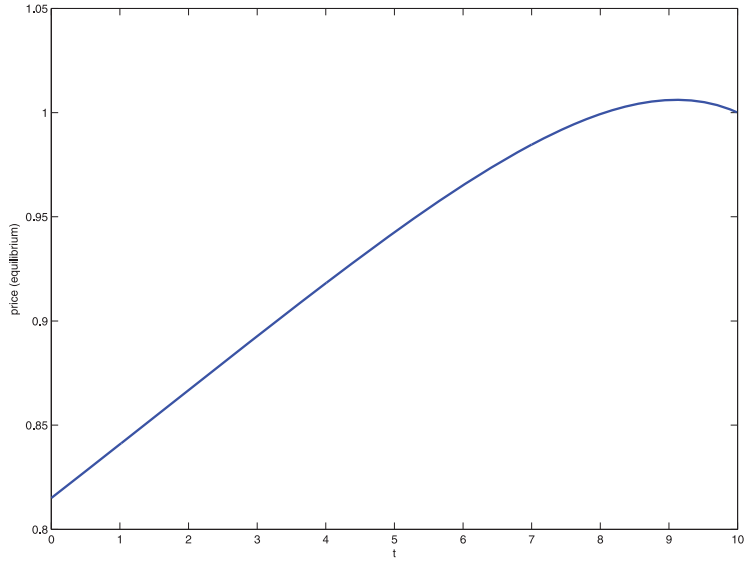

FIGURE 7.5. Equilibrium price function over time t for ILBs. Price level normalized to one.

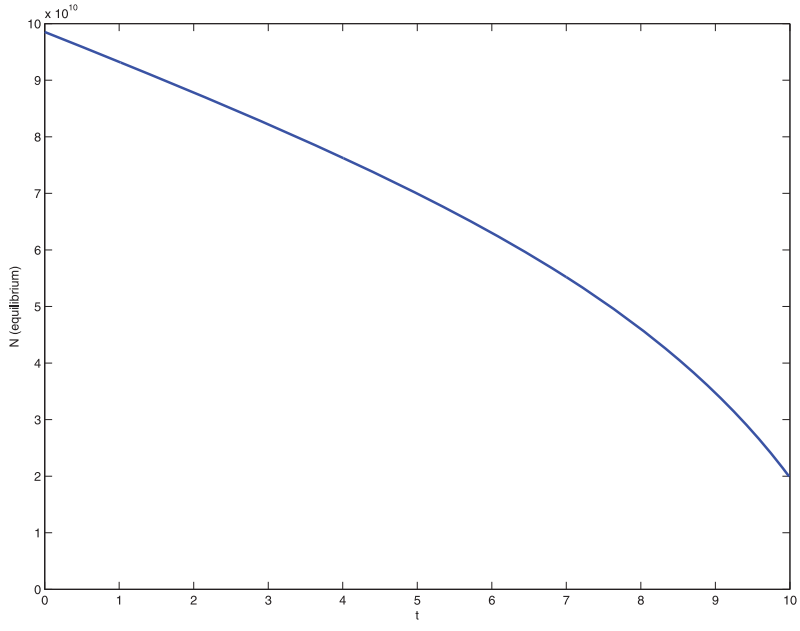

FIGURE 7.6. Equilibrium quantity of ILBs sold as a function over time t.

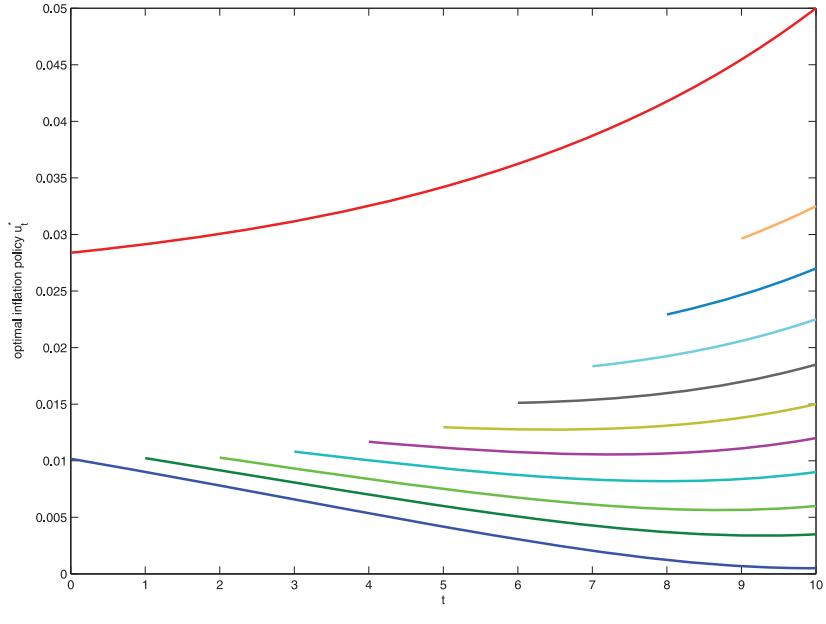

FIGURE 7.7. Inflation in equilibrium over time when ILBs are issued at time $t = 0$ (lower blue curve), $t = 1$ (lower green curve), ..., $t = 9$ (short upper orange curve), and without issuing any ILBs (long upper red curve).

The effect of ILBs on output gain from unexpected inflation over this period is measured by the term $\mathbb{E}\left(\int_0^T a(\pi_t - \pi_t^e)dt\right)$ that is part of the central banks utility function.[22] This term can be computed analytically for the equilibrium policies, but the expression is extremely long, so we omit it here. Numerically, using the previous set of parameters, we obtain for the output gain without ILBs a value of $(1.496 - 9.999 \times \pi_0^{e^*}) \times 10^{10}$, compared with a value of $(1.364 - 9.999 \times \pi_0^{e^*}) \times 10^{10}$ with ILBs. The output gain depends, in fact, on the initial inflation expectation $\pi_0^{e^*}$. If we assume that in both cases, with or without ILBs, $\pi_0^{e^*} = 0.02$, the inflation target of the Bank of England, than output gain without ILBs accumulates to 1.296×10^{10} compared to 1.164×10^{10} with ILBs. Hence, output gain from unexpected inflation over the 10-year period is reduced by 11% if ILBs are issued.[23] However, as indicated in Section 2 our rational expectation interpretation of the initial inflation estimates $\pi_0^{e^*}$ indicates that we need to use different values for $\pi_0^{e^*}$, depending on whether there are markets for ILBs or not.[24] Assuming that $\pi_0^{e^*} = u_0^*(0) = 0.0275$ without ILBs and $\pi_0^{e^*} = u_0^*(N^*) = 0.01$ with ILBs (compare Figure 7.7), we obtain for the output gain without ILBs the value of 1.221×10^{10} compared with 1.264×10^{10} with ILBs. In this case therefore not only is inflation significantly reduced with ILBs, but in addition, output gain is higher than without ILBs. The latter is intuitive. As inflation expectations are generally lower with ILBs, potentially more can be gained from unexpected inflation. In the latter case, we have a double benefit in welfare terms from the existence of markets for ILBs, higher output gain and lower inflation.

8. CONCLUSIONS

We have studied the effect of markets for inflation-indexed bonds on the central bank's monetary policy. We presented a continuous-time model featuring a central bank, private agents, and a financial market, in which the central bank can adjust inflation and in addition can issue inflation-indexed securities, which it sells on the financial market. Within this model, we have derived equilibrium prices and quantity for the inflation-indexed bonds issued. We have shown that the introduction of inflation-indexed securities sold by the central bank can reduce central bank's inflationary bias, and that central bank, private agents, as well as the financial agents in our model are better off with inflation-indexed bonds than without. In this way, inflation-indexed bonds should be seen as effective and powerful monetary policy instruments. From the financial derivatives point of view, we have derived the first pricing formulas for inflation-indexed bonds within an integrated framework, which includes the central bank as a significant component.

REFERENCES

ABRAMOWITZ, M., and I. STEGUN (1965): *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Mineola, NY: Dover Publications.

[22] Note once more that the Phillips curve slope parameter a in this paper by default includes natural real output Y_t^n as setup at the beginning of Section 2

[23] Note that this is not the same as saying that GDP is 11% lower over the same period.

[24] Note that in a rational expectation context, the situation where there are no markets for ILBs, and hence the central bank cannot offer ILBs, is different from the situation where there are markets for ILBs but the central bank decides not to offer any, as considered in the pricing of ILBs.

BARRO, R. J. (1996): Optimal Funding and Indexed Bonds. Index-Linked Debt, Papers presented at the Bank of England Conference, September 1995, pp. 39–44.

BARRO, R. J., and D. B. GORDON (1983a): A Positive Theory of Monetary Policy in a Nature Rate Model, 91(4), 589–610.

BARRO, R. J., and D. B. GORDON (1983b): Rules, Discretion and Reputation in a Model of Monetary Policy, *J. Monetary Econ.* 12, 101–121.

CAMPBELL, J. Y., and R. J. SHILLER (1996): A Scorecard for Indexed Government Debt, Harvard Institute of Economic Research Discussion Paper No. 1758.

CAMPBELL, J., R. J. SHILLER, and L. M. VICEIRA (2009): Understanding Inflation-Indexed Bond Markets, Brookings Papers on Economic Activity 79–120.

DEACON, M., A. DERRY, and D. MIRFEDERESKI (2004): *Inflation-Indexed Securities: Bonds, Swaps and Other Derivatives*, Chichester, UK: Wiley Finance.

DRIFFIL, J., R. ZENO, P. SAVONA, and C. ZAZZARA (2006): Monetary Policy and Financial Stability: What Role for the Futures Market?, *J. Financ. Stab.* 2, 95–112.

EVANS, G., and S. HONKAPOHJA (2001): *Learning and Expectations in Macroeconomics*, Princeton, NJ; Oxford: Princeton University Press.

EWALD, C.-O., and J. GEISSLER (2012): Optimal Contracts for Central Bankers: Calls on Inflation, Working Paper. Available at SSRN: http://ssrn.com/abstract=1633894.

GONG, F. F., and E. M. REMOLONA (1996): Inflation Risk in the US Yield Curve: The Usefulness of Indexed Bonds. Federal Reserve Bank of New York Research Paper No. 9637, November.

HANKE, S. H., and A. WALTERS (1994): Greenspan Bonds. Forbes, 12 September.

JARROW, R., and Y. YILDIRIM (2003): Pricing Treasury Inflation Protected Securities and Related Derivatives Using an HJM Model, *J. Financ. Quantit. Anal.* 38(2), 409–430.

KARATZAS, I. (1997): Lectures on the Mathematics of Finance, *CRM Monograph Series 8*, p. 148, Providence, RI: American Mathematical Society.

KARATZAS, I., and S. SHREVE (1998): *Methods of Mathematical Finance*, New York: Springer.

KYDLAND, F. E., and E. C. PRESCOTT (1977): Rules Rather Than Discretion: The Inconsistency of Optimal Plans, *J. Polit. Econ.* 85(3), 473–491.

PEARLMAN, J., D. CURRIE, and P. LEVINE (1986): Rational Expectations Models with Partial Information, *Econ. Model.* 3(2), 90–105.

PHELPS, E.-S. (1967): Money-Wage Dynamics and Labor-Market Equilibrium, *J. Polit. Econ.* 76(4), 678–711, Part 2: Issues in Monetary Research.

PRICE, R. (1997): The Rationale and Design of Inflation-Indexed Bonds. International Monetary Fund Working Paper No. 97/12, January.

SCHORFHEIDE, F. (2008): DSGE Model-Based Estimation of the New Keynesian Phillips Curve, *Econ. Quarterly* 94(4), 397–433.

WALSH, C. E. (1995): Optimal Contracts for Central Banks, *Am. Econ. Rev.* 85(1), 150–167.

WALSH, C. E. (2003): *Monetary Theory and Policy*, 2nd ed., Cambridge, MA: The MIT Press.

WOJNILOWER, A. M. (1997): *Inflation-Indexed Bonds: Promising the Moon*, New York: The Clipper Group.

YANNACOPOULOS, A. (2008). Rational Expectations Models: An Approach Using Forward-Backward Stochastic Differential Equations, *J. Math. Econ.* 44(3–4), 251–276.