The Limits of Leverage*

Paolo Guasoni†  Eberhard Mayerhofer‡

June 8, 2015

Abstract

When trading incurs proportional costs, leverage can scale an asset’s return only up to a maximum multiple, which is sensitive to its volatility and liquidity. In a model with one safe and one risky asset, with constant investment opportunities and proportional costs, we find strategies that maximize long term return given average volatility. As leverage increases, rising rebalancing costs imply declining Sharpe ratios. Beyond a critical level, even returns decline. Holding the Sharpe ratio constant, higher volatility leads to superior returns through lower costs. For funds replicating benchmark multiples, such as leveraged ETFs, we identify the strategies that optimally trade off alpha against tracking error, and find that they depend on the target multiple and the benchmark’s liquidity, but not its volatility.

JEL: G11, G12.


Keywords: leverage, transaction costs, portfolio choice, performance evaluation, ETFs.

*For helpful comments, we thank Jakša Cvitanić, Gur Huberman, Tim Leung, Johannes Muhle-Karbe, Ronnie Sircar, Matt Spiegel, René Stulz, Peter Tankov, Walter Schachermayer, and seminar participants at ETH Zürich, Vienna Graduate School of Finance, Quant Europe, Quant USA, Banff Workshop on Arbitrage and Portfolio Optimization, Institute of Finance at USI (Lugano).

†Boston University, Department of Mathematics and Statistics, 111 Cummington Street, Boston, MA 02215, USA, and Dublin City University, School of Mathematical Sciences, Glasnevin, Dublin 9, Ireland, email guasoni@bu.edu. Partially supported by the ERC (278295), NSF (DMS-1412529), SFI (07/MI/008, 07/SK/M1189, 08/SRC/FMC1389), and FP7 (RG-248896).

‡Dublin City University, School of Mathematical Sciences, Glasnevin, Dublin 9, Ireland, email eberhard.mayerhofer@dcu.ie.
1 Introduction

If trading is costless, leverage can scale returns without limits. Using the words of Sharpe (2011):

“If an investor can borrow or lend as desired, any portfolio can be leveraged up or down. A combination with a proportion $k$ invested in a risky portfolio and $1 - k$ in the riskless asset will have an expected excess return of $k$ and a standard deviation equal to $k$ times the standard deviation of the risky portfolio. Importantly, the Sharpe Ratio of the combination will be the same as that of the risky portfolio.”

In theory, this insight implies that the efficient frontier is linear, that efficient portfolios are identified by their common maximum Sharpe ratio, and that any of them spans all the other ones. Also, if leverage can deliver any expected returns, then risk-neutral portfolio choice is meaningless, as it leads to infinite leverage.

In practice, hedge funds and high-frequency trading firms employ leverage to obtain high returns from small relative mispricing of assets. Recent financial products such as leveraged mutual funds and exchange traded funds (ETFs) closely follow the strategy described by Sharpe, rebalancing their exposure to an underlying asset, with the aim of replicating a multiple of its daily return.

This paper shows that trading costs undermine these classical properties of leverage and set sharp theoretical limits to its applications. We start by characterizing the set of portfolios that maximize long term expected returns for given average volatility, extending the familiar efficient frontier to a market with one safe and one risky asset, where both investment opportunities and relative bid-ask spreads are constant. Figure 1 plots this frontier: expectedly, trading costs decrease returns, with the exception of a full safe investment (the axes origin) or a full risky investment (the attachment point with unit coordinates), which lead to static portfolios without trading, and hence earn their frictionless return.

But trading costs do not merely reduce expected returns below their frictionless benchmarks. Unexpectedly, in the leverage regime (the right of the full-investment point) rebalancing costs rise so quickly with volatility that returns cannot increase beyond a critical factor, the leverage multiplier or, briefly, the multiplier. The multiplier depends on the relative bid-ask spread $\varepsilon$, the expected excess return $\mu$ and volatility $\sigma$, and approximately equals

$$0.3815 \left( \frac{\mu}{\sigma^2} \right)^{1/2} \varepsilon^{-1/2}. \tag{1.1}$$

Table 1 shows that even a modest bid-ask spread of 0.10% implies a multiplier of 23 for an asset with 10% volatility and 5% expected return (similar to a long-term bond), while the multiplier declines to 10 for an asset with equal Sharpe ratio, but with a volatility of 50% (similar to an individual stock). Leverage opportunities are much more limited for more illiquid assets with a spread of 1%, from less than 8 for 10% volatility to less than 4 for 50% volatility. Importantly, these limits on leverage hold even allowing for continuous trading, infinite market depth (any quantity trades at the bid or ask price), and zero capital requirements.

Our results have three broad implications. First, with a positive bid-ask spread even a risk-neutral investor who seeks to maximize expected long-run returns will take finite leverage, and in

---

1 A famous example is Long Term Capital Management, which used leverage of up to 30 to 40 times to increase returns from convergence trades between on-the-run and off-the-run treasury bonds, see Edwards (1999).

2 As we focus on long term investments, we neglect the one-off costs of set up and liquidation, which are negligible over a long holding period.
Table 1: Leverage multiplier (maximum factor by which a risky asset’s return can be scaled) for different asset volatilities and bid-ask spreads, holding the Sharpe ratio at the constant level of 0.5. Multipliers are obtained from numerical solutions of (3.1), while their approximations from (1.1) are in brackets.

<table>
<thead>
<tr>
<th>Volatility ($\sigma$)</th>
<th>0.01%</th>
<th>0.10%</th>
<th>1.00%</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>71.85 (71.22)</td>
<td>23.15 (22.58)</td>
<td>7.72 (7.12)</td>
</tr>
<tr>
<td>20%</td>
<td>50.88 (50.36)</td>
<td>16.45 (15.92)</td>
<td>5.56 (5.04)</td>
</tr>
<tr>
<td>50%</td>
<td>32.30 (31.85)</td>
<td>10.54 (10.07)</td>
<td>3.66 (3.18)</td>
</tr>
</tbody>
</table>

Figure 1: Efficient Frontier with trading costs, as expected excess return (vertical axis, in multiples of the asset’s return) against standard deviation (horizontal axis, in multiples of the asset’s volatility). The asset has expected excess return $\mu = 8\%$, volatility $\sigma = 16\%$, and bid-ask spread of 1%. The upper line denotes the classical efficient frontier, with no transaction costs. The maximum height of the curve corresponds to the leverage multiplier.
fact a rather low leverage ratio in an illiquid market – risk-neutral portfolio choice is meaningful. The resulting multiplier sets an endogenous level of risk that the investor chooses not to exceed regardless of risk aversion, simply to avoid reducing returns with trading costs. In this context, margin requirements based on volatility (such as value at risk and its variations) are binding only when they reduce leverage below the multiplier, and are otherwise redundant. In addition, the multiplier shows that an exogenous increase in trading costs, such as a proportional Tobin tax on financial transactions, implicitly reduces the maximum leverage that any investor who seeks return is willing to take, regardless of risk attitudes.

Second, two assets with the same Sharpe ratio do not generate the same efficient frontier with trading costs, and more volatility leads to a superior frontier. For example (Table 1) with a 1% spread the maximum leveraged return on an asset with 10% volatility and 5% return is $7.72 \times 5\% \approx 39\%$. By contrast, an asset with 50% volatility and 25% return (equivalent to the previous one from a classical viewpoint, since it has the same Sharpe ratio 0.5), leads to a maximum leveraged return of $3.66 \times 25\% \approx 92\%$. The reason is that a more volatile asset requires a lower leverage ratio (hence lower rebalancing costs) to reach a certain return. Thus, an asset with higher volatility spans an efficient frontier that achieves higher returns through lower costs.

Third, our analysis delivers the first treatment of optimal replication of leveraged ETF on an illiquid benchmark. We obtain optimal trading policies, their performance, and the theoretical bounds on the potential returns of leveraged ETFs. In particular, we derive a testable restriction between the resulting alpha and the tracking error of an optimally replicated fund. In a frictionless setting, an ETF can perfectly scale returns by any factor, without any tracking error: alpha is zero and the fund’s returns are perfectly correlated with the benchmark’s. In reality, leveraged ETFs, which have been introduced only since 2006, currently have leverage factors of up to three (minus three for inverse funds), and funds on less liquid assets have significant tracking error.

Under optimal replication, we obtain the following relation between the intercept $\bar{\alpha}$ and the tracking error $\text{TrE}$ in the regression of the ETF return (net of management fees) on the benchmark’s return

$$\bar{\alpha} \approx -\frac{\sqrt{3}}{12} \sigma^2 \Lambda^2 (1 - \Lambda)^2 \frac{\varepsilon}{\text{TrE}}, \quad (1.2)$$

where $\Lambda$ is the target benchmark multiple, $\sigma$ is the benchmark’s volatility, and $\varepsilon$ is its relative spread. The equation makes the optimal replication trade-off clear: a lower tracking error leads to a more negative alpha through higher costs, and vice versa. More importantly, the equation offers a testable relationship among observable quantities, without involving the expected excess return $\mu$, which is notoriously hard to estimate with precision.

This paper bears on the established literature on portfolio choice with frictions and on the nascent literature on leveraged ETFs. The effect of transaction costs on portfolio choice is first studied by Magill and Constantinides (1976), Constantinides (1986), and Davis and Norman (1990), who identify a wide no-trade region, and derive the optimal trading boundaries through numerical procedures. While these papers focus on the maximization of expected utility from intertemporal consumption on an infinite horizon, Taksar, Klass and Assaf (1988), and Dumas and Luciano (1991) show that similar strategies are obtained in a model with terminal wealth and a long horizon – time preference has negligible effects on trading policies. This paper adopts the same approach of a long horizon, both for the sake of tractability, and because it focuses on the trade-off between return, risk, and costs, rather than consumption.

Our asymptotic results for positive risk aversion are similar in spirit to the ones derived by Shreve and Soner (1994), Rogers (2004), Gerhold, Guasoni, Muhle-Karbe and Schachermayer (2014), and
Kallsen and Muhle-Karbe (2013), whereby transaction costs imply a no-trade region with width of order $O(\varepsilon^{1/3})$ and a welfare effect of order $O(\varepsilon^{2/3})$. We also find that the trading boundaries obtained from a local mean-variance criterion are equivalent at the first order to the ones obtained from power utility. The risk-neutral expansions and the limits of leverage of order $O(\varepsilon^{-1/2})$ are new, and are qualitatively different from the risk-averse case. These results are not regular perturbations of a frictionless analogue, which is ill-posed. They are rather singular perturbations, which display the speed at which the frictionless problem becomes ill-posed as the crucial friction parameter vanishes.

Our paper also contributes to the literature on leveraged ETFs. Tang and Xu (2013) observe that leveraged funds deviate significantly from their benchmarks even after management fees, and separate tracking error into a compounding component, due to the convexity of leveraged returns and a rebalancing component, due to trading frictions (cf. Jarrow (2010); Lu et al. (2012); Avellaneda and Zhang (2010); Cheng and Madhavan (2009)). Jiang and Yan (2012), Avellaneda and Dobi (2012), and Guo and Leung (2014) report that ETFs significantly underperform their benchmarks even at daily frequencies, and Wagalath (2014) derives an asymptotic expression for the slippage that results from rebalancing at fixed intervals. We incorporate trading costs explicitly in the model, and derive optimal replication policies that trade off alpha against tracking error. Expectedly, such strategies entail buy and sell boundaries that depend on the benchmark’s liquidity and on the relative importance of alpha and tracking error. Unexpectedly, such boundaries do not depend on the benchmark’s volatility, suggesting that they may be robust to stochastic volatility.

Finally, this paper connects to the recent work of Frazzini and Pedersen (2012) on embedded leverage. If different investors face different leverage constraints, they find that in equilibrium assets with higher factor exposures trade at a premium, thereby earning a lower return. Frazzini and Pedersen (2014) confirm this prediction across a range of markets and asset classes, and Asness et al. (2012) use it to explain the performance risk-parity strategies. With exogenous asset prices, we find that assets with higher volatility generate a superior efficient frontier by requiring lower rebalancing costs for the same return. This observation suggests that the embedded leverage premium may be induced by rebalancing costs in addition to leverage constraints, and should be higher for more illiquid assets.

The paper is organized as follows: section 2 introduces the model and the optimization problem. Section 3 contains the main results, which characterize the efficient frontier in the risk-averse (Theorem 3.1) and risk-neutral (Theorem 3.3) cases, nesting optimal leverage replication as a special case (Theorem 3.2). Section 4 discusses the implications of these results for the efficient frontier, the trading boundaries of optimal policies, the embedded leverage effect, and leveraged ETF replication. The section concludes with two supporting results, which show that the risk-neutral solutions arise as limits of their risk-averse counterparts for low risk-aversion, and that the risk-neutral solutions are not constrained by the solvency condition. Section 5 offers a derivation of the main free-boundary problems from heuristic control arguments, and concluding remarks are in section 6. All proofs are in the appendix.
2 Model

The market includes one safe asset earning a constant interest rate of \( r \geq 0 \) and a risky asset with ask (buying) price \( S_t \) that follows
\[
\frac{dS_t}{S_t} = (\mu + r)dt + \sigma dB_t, \quad S_0, \sigma, \mu \geq 0,
\]
where \( B \) is a standard Brownian motion. The risky asset’s bid (selling) price is \((1 - \varepsilon)S_t\), which implies a constant relative bid-ask spread of \( \varepsilon > 0 \), or, equivalently, constant proportional transaction costs. A self-financing trading strategy is summarized by its initial capital \( x \) and the number of shares \( \varphi_t \) of the risky asset held at time \( t \). Denote by \( w_t \) the fund’s wealth at time \( t \), which is the sum of the safe position \( x - \int_0^t S_s d\varphi_s - \varepsilon \int_0^t S_s d\varphi_s^\downarrow \) and the risky position \( S_t \varphi_t \) evaluated at the ask price\(^3\):
\[
w_t = x - \int_0^t S_s d\varphi_s - \varepsilon \int_0^t S_s d\varphi_s^\downarrow + S_t \varphi_t. \tag{2.1}
\]

We further require a strategy \( \varphi \) to be solvent, in that its corresponding wealth \( w_t \) is strictly positive at all times. (Admissible strategies are formally described in Definition A.1 below.)

Our objective function trades off a fund’s average return against its realized variance relative to a benchmark. The portfolio return \( r_t \) over the time-interval \([t - \Delta t, t] \) is
\[
r_t = \frac{w_t - w_{t-\Delta t}}{w_{t-\Delta t}}, \tag{2.2}
\]
while the annualized average return has the continuous-time approximation\(^4\) \((\Delta t = T/n)\)
\[
\bar{r}_T = \frac{1}{n\Delta t} \sum_{k=1}^n r_{k\Delta t} \approx \frac{1}{T} \int_0^T \frac{dw_t}{w_t}. \tag{2.3}
\]

In the familiar setting of no trading costs, \( \frac{1}{T} \int_0^T \frac{dw_t}{w_t} = r + \frac{1}{T} \int_0^T \mu \pi_t dt + \frac{1}{T} \int_0^T \sigma \pi_t dB_t \), where \( \pi_t = \varphi_t S_t/w_t \) is the portfolio weight of the risky asset, hence the average return equals the average risky exposure times its excess return, plus the safe rate.

Likewise, the average squared volatility on \([0, T] \) is obtained by the usual variance estimator applied to returns, and has the continuous-time approximation
\[
\frac{1}{n\Delta t} \sum_{k=1}^n r_{k\Delta t}^2 \approx \frac{1}{T} \int_0^T \frac{d\langle w \rangle_t}{w_t^2} = \frac{\sigma^2}{T} \int_0^T \pi_t^2 dt.
\]
(The last equality holds because the trading cost term \( \varepsilon \int_0^t S_s d\varphi_s^\downarrow \) in (2.1) is increasing and continuous, hence its total sum of squares is negligible on a fine grid.) More generally, the average squared tracking error with respect to a multiple \( \Lambda \) of the asset’s return \( r_t^S \) is approximated by
\[
\frac{1}{T} \sum_{k=1}^n (r_{k\Delta t} - \Lambda r_{k\Delta t}^S)^2 \approx \frac{1}{T} \left\langle \int_0^T \left( \frac{dw_t}{w_t} - \Lambda \frac{dS_t}{S_t} \right) \right\rangle T = \frac{\sigma^2}{T} \int_0^T (\pi_t - \Lambda)^2 dt, \tag{2.3}
\]

\(^3\)The convention of evaluating the risky position at the ask price is inconsequential. Using the bid price instead leads to the same results up to a change of notation.

\(^4\)All finite statistics on this page converge in probability to their continuous-time counterparts.
and coincides with average squared volatility for \( \Lambda = 0 \). With these definitions, the risk-return trade-off is captured by maximizing

\[
\max_{\varphi} \frac{1}{T} \mathbb{E} \left[ \int_0^T \frac{dw_t}{w_t} - \frac{\gamma}{2} \left( \int_0^T \left( \frac{dw_t}{w_t} - \Lambda \frac{dS_t}{S_t} \right) \right) \right],
\]

(2.4)

where the parameter \( \gamma > 0 \) is interpreted as a proxy for risk-aversion.

This objective nests several familiar problems. With \( \Lambda = 0 \) and without trading costs it reduces to

\[
\max_{\pi} \frac{1}{T} \mathbb{E} \left[ \int_0^T \left( \mu_\pi - \frac{\gamma}{2} \sigma^2 \pi^2 \right) dt \right]
\]

(2.5)

which leads to the optimal constant-proportion portfolio \( \pi = \frac{\mu_\pi}{\gamma \sigma^2} \) dating back to Markowitz and Merton, and confirms that in a geometric Brownian motion market with costless trading, the objective considered here is equivalent to utility-maximization with constant relative risk aversion. With or without transaction costs, the risk-neutral objective \( \Lambda = \gamma = 0 \) boils down to the average annualized return over a long horizon, while \( \Lambda = 0, \gamma = 1 \) reduces to logarithmic utility.

With \( \Lambda > 0 \) the objective (2.4) maximizes average return for given tracking error, which is relevant for funds that aim at replicating multiples of a benchmark’s return. In general, alpha arises from the difference between the exposure to the benchmark in excess of \( \Lambda \) and average trading costs, while the tracking error results from the departure of the fund from the target exposure \( \Lambda \). In practice, managers of leveraged ETFs do not attempt to outperform their benchmarks through over- or under-exposure, hence that their typical objective is summarized by \( \Lambda > 0, \mu = 0 \).

To proceed further, note first that the objective function (2.4) has a more concrete expression (see Section A.1).

**Lemma 2.1.** For any \( T > 0 \) and for any admissible trading strategy \( \varphi \),

\[
F_T(\varphi) := \frac{1}{T} \mathbb{E} \left[ \int_0^T \frac{dw_t}{w_t} - \frac{\gamma}{2} \left( \int_0^T \left( \frac{dw_t}{w_t} - \Lambda \frac{dS_t}{S_t} \right) \right) \right]
\]

(2.6)

\[
= r + \frac{1}{T} \mathbb{E} \left[ \int_0^T \left( \mu_\pi - \frac{\gamma}{2} \sigma^2 \pi^2 \right) dt - \varepsilon \int_0^T \pi_t \frac{d\varphi_t}{\varphi_t} \right].
\]

(2.7)

The final term in (2.7) represents trading costs, which hinder continuous portfolio rebalancing and makes constant-proportion strategies unfeasible. The reason is that it is costly both to keep the exposure to the risky asset high enough to achieve the desired return, and to keep it low enough to limit the level of risk – trading costs reduce returns and increase risk.

To neglect the spurious, non-recurring effects of portfolio set-up and liquidation, we focus on the Equivalent Safe Rate\(^5\)

\[
\text{ESR}(\varphi) := \lim_{T \to \infty} \sup T F_T(\varphi)
\]

(2.8)

which is akin to the one used by Dumas and Luciano (1991) in the context of utility maximization.

---

\(^5\)In this equation the \( \lim \sup \) is used merely to guarantee a good-definition a priori. A posteriori, we show that optimal strategies exist in which the limit superior is a limit, hence the similar problem defined in terms of \( \lim \inf \) leads to the same solution.
3 Main Results

The first result characterizes the optimal solution to the main objective in (2.8) in the usual case of a positive aversion to risk \((\gamma > 0)\). In this setting, the next theorem shows that trading costs create a no-trade region around the frictionless portfolio \(\pi^* = \frac{\mu}{\gamma \sigma^2}\), and states the asymptotic expansions of the resulting average return and standard deviation\(^6\), thereby extending the familiar efficient frontier to account for trading costs.

**Theorem 3.1 (Risk Aversion and Efficient Frontier).** Let \(\gamma \neq \mu/\sigma^2\).

(i) For any \(\gamma > 0\) there exists \(\varepsilon_0 > 0\) such that for all \(\varepsilon < \varepsilon_0\), the free boundary problem

\[
\frac{1}{2} \sigma^2 \zeta^2 W''(\zeta) + (\sigma^2 + \mu) \zeta W'(\zeta) + \mu W(\zeta) - \frac{1}{(1 + \zeta)^2} \left( \mu + \gamma \sigma^2 \Lambda - \gamma \sigma^2 \frac{\zeta}{1 + \zeta} \right) = 0, \tag{3.1}
\]

\[
W(\zeta^-) = 0, \tag{3.2}
\]

\[
W'(\zeta^-) = 0, \tag{3.3}
\]

\[
W(\zeta^+) = \frac{\varepsilon}{(1 + \zeta^+)(1 + (1 - \varepsilon)\zeta^+)}, \tag{3.4}
\]

\[
W'(\zeta^+) = \frac{\varepsilon(\varepsilon - 2(1 - \varepsilon)\zeta^+ - 2)}{(1 + \zeta^+)^2(1 + (1 - \varepsilon)\zeta^+)^2} \tag{3.5}
\]

has a unique solution \((W, \zeta^-, \zeta^+))\) for which \(\zeta^- < \zeta^+\).

(ii) The trading strategy \(\hat{\phi}\) that buys at \(\pi_- := \zeta^-/(1 + \zeta^-)\) and sells at \(\pi_+ := \zeta^+/(1 + \zeta^+)\) as little as to keep the risky weight \(\pi_t = \zeta_t/(1 + \zeta_t)\) within the interval \([\pi_-, \pi_+]\), is optimal.

(iii) The maximum performance is

\[
\max_{\varphi \in \Phi} \lim_{T \to \infty} \frac{1}{T} \mathbb{E} \left[ \int_0^T \left( \mu \pi_t - \frac{\gamma \sigma^2}{2} (\pi_t - \Lambda)^2 \right) dt - \varepsilon \int_0^T \frac{d\varphi_t}{\varphi_t} \right] = \mu \pi_+ - \frac{\gamma \sigma^2}{2} (\pi_+ - \Lambda)^2, \tag{3.6}
\]

where \(\Phi\) is the set of admissible strategies in Definition A.1.

(iv) Let \(\theta_* := \pi_* + \Lambda\). The trading boundaries \(\pi_-\) and \(\pi_+\) have the asymptotic expansions

\[
\pi_\pm = \theta_* \pm \left( \frac{3}{4\gamma}(\theta_*)^2(\theta_* - 1)^2 \right)^{1/3} \varepsilon^{1/3} - \left( \frac{(1 - \gamma)\pi_* + \Lambda}{\gamma} \right) \left( \frac{\gamma \theta_*(\theta_* - 1)}{6} \right)^{1/3} \varepsilon^{2/3} + O(\varepsilon). \tag{3.7}
\]

The long-run mean \((\hat{m})\), standard deviation \((\hat{s})\), average trading costs \((\text{ATC})\) and equivalent safe rate \((\text{ESR})\) have expansions (using the convention \(a^{1/n} = \text{sign}(a)|a|^{1/n}\) for any \(a \in \mathbb{R}\))

\(^6\)The exact formulae for average return, standard deviation, and average trading costs are in Appendix C.
and odd integer \( n \), and \( a^{2/n} = (a^2)^{1/n} \)

\[
\hat{m} := \lim_{T \to \infty} \frac{1}{T} \int_0^T \frac{dw_t}{w_t} = r + \mu \Lambda + \frac{\mu^2}{\gamma \sigma^2} - \frac{\sigma^2}{2\gamma} \left( \Lambda^2 + \pi_*(5\pi_* - 3) + \Lambda(6\pi_* - 1) \right) \frac{\left( \frac{\gamma \theta_*(\theta_* - 1)}{6} \right)^{1/3}}{\varepsilon^{2/3} + O(\varepsilon)},
\]

(3.8)

\[
\hat{s} := \lim_{T \to \infty} \sqrt{\frac{1}{T} \left( \int_0^T \frac{dw_t}{w_t} \right)} = \frac{\mu(1 + \Lambda/\pi_*)}{\gamma \sigma} - \frac{\sigma \left( \frac{7\theta_* - 3}{4\gamma} \right)}{\left( \frac{3\theta_*^2(\theta_* - 1)}{4} \right)^{2/3}} \varepsilon^{2/3} + O(\varepsilon),
\]

(3.9)

\[
ATC := \lim_{T \to \infty} \frac{1}{T} \int_0^T \pi_t \frac{d\varphi_t}{\varphi_t} = \frac{3\sigma^2}{\gamma} \left( \frac{\gamma \theta_*(\theta_* - 1)}{6} \right)^{4/3} \varepsilon^{2/3} + O(\varepsilon),
\]

(3.10)

\[
ESR = r + \frac{\gamma \sigma^2}{2} \left( \left( \pi_* + \Lambda \right)^2 - \Lambda^2 \right) - \frac{\gamma \sigma^2}{2} \left( 3 \frac{\theta_*^2(\theta_* - 1)}{4\gamma} \right)^{2/3} \varepsilon^{2/3} + O(\varepsilon).
\]

(3.11)

\[
\text{Proof.} \quad \text{The proof of the main part of this theorem is divided into Propositions B.1, B.4 and B.6 in Appendix B. The proof of the asymptotical results is provided in Section C.2.}
\]

The case \( \mu = 0 \) is noteworthy because it summarizes the objective of the manager of a leveraged fund. In addition, in this important case the result simplifies considerably, as it is independent of the benchmark’s volatility \( \sigma \), therefore we report its statement in detail:

**Theorem 3.2** (Optimal Replication of Leveraged Benchmark). Assume \( \Lambda \neq 0,1 \) and \( \mu = 0 \neq \sigma \), which leads to the HJB equation

\[
\frac{1}{2} \xi^2 W''(\zeta) + \zeta W'(\zeta) - \frac{\gamma}{(1 + \zeta)^2} \left( \Lambda - \frac{\zeta}{1 + \zeta} \right) = 0.
\]

(3.13)

(i) The maximum performance is

\[
\max_{\varphi \in \Phi} \lim_{T \to \infty} \frac{1}{T} \mathbb{E} \left[ -\frac{\gamma \sigma^2}{2} \int_0^T (\pi_t - \Lambda)^2 dt - \varepsilon \int_0^T \pi_t \frac{d\varphi_t}{\varphi_t} \right] = -\frac{\gamma \sigma^2}{2} (\pi_* - \Lambda)^2,
\]

where \( \Phi \) is the set of admissible strategies in Definition A.1.

(ii) Average Trading costs (ATC), long-run mean \( \hat{m} \), and standard deviation \( \hat{s} \) are

\[
ATC = \frac{\sigma^2}{2} \frac{\pi_- \pi_+ (\pi_+ - 1)^2}{(\pi_+ - \pi_-)(1/\varepsilon - \pi_+)},
\]

(3.14)

\[
\hat{m} = r - ATC,
\]

(3.15)

\[
\hat{s} = \sigma \sqrt{\pi_- \pi_+}.
\]

(3.16)

(iii) Excess portfolio returns satisfy

\[
\lim_{T \to \infty} \left( \frac{1}{T} \int_0^T \frac{dw_t}{w_t} - r \right) = \bar{\alpha} + \bar{\beta} \lim_{T \to \infty} \left( \frac{1}{T} \int_0^T \frac{dS_t}{S_t} - r \right) \quad \text{a.s.}
\]
The realized alpha, multiplier $\bar{\beta}$, and $R^2$ of this regression are

$$\bar{\alpha} = -\text{ATC},$$

$$\bar{\beta} := \lim_{T \to \infty} \frac{1}{T} \int_0^T \pi_t dt = \log(\pi_+/\pi_-) \frac{\pi_+ - \pi_-}{\pi_+ - \pi_-},$$

$$R^2 := \lim_{T \to \infty} \left( \frac{1}{T} \int_0^T \pi_t dt \right)^2 = \pi_- \pi_+ \left( \frac{\log(\pi_+)}{\pi_+ - \pi_-} \right)^2.$$ (3.19)

(iv) The tracking error is

$$\text{TrE} = \sqrt{\lim_{T \to \infty} \frac{1}{T} \left( \int_0^T \frac{dS_t}{S_t} - \Lambda \int_0^T \frac{dS_t}{S_t} \right)^2} = \sigma \sqrt{\pi_- \pi_+ + \Lambda(\Lambda - 2\bar{\beta})}.$$ (3.20)

(v) The trading boundaries $\pi_-$ and $\pi_+$ have the asymptotic expansions

$$\pi_{\pm} = \Lambda \pm \left( \frac{3}{4\gamma} \Lambda^2 (\Lambda - 1)^2 \right)^{1/3} \pm \frac{\Lambda}{\gamma} \left( \frac{\gamma \Lambda (\Lambda - 1)}{6} \right)^{1/3} \varepsilon^{2/3} + O(\varepsilon).$$ (3.21)

The long-run mean ($\bar{m}$), standard deviation ($\bar{s}$), average trading costs (ATC) and Equivalent Safe Rate (ESR)

$$\bar{m} := \lim_{T \to \infty} \frac{1}{T} \int_0^T \frac{dS_t}{S_t} = r - \frac{\sigma^2}{2\gamma} \Lambda (\Lambda - 1) \left( \frac{\gamma \Lambda (\Lambda - 1)}{6} \right)^{1/3} \varepsilon^{2/3} + O(\varepsilon),$$

$$\bar{s} := \lim_{T \to \infty} \sqrt{\frac{1}{T} \left( \int_0^T \frac{dS_t}{S_t} \right)^2} = \gamma \sigma^2 \Lambda - \frac{\sigma(7\Lambda - 3)}{4\gamma} \left( \frac{\gamma \Lambda (\Lambda - 1)}{6} \right)^{1/3} \varepsilon^{2/3} + O(\varepsilon),$$

$$\text{ATC} := \lim_{T \to \infty} \frac{1}{T} \int_0^T \pi_t dw_t = 3\gamma \sigma^2 \left( \frac{\gamma \Lambda (\Lambda - 1)}{6} \right)^{1/3} \varepsilon^{2/3} + O(\varepsilon),$$

$$\text{ESR} = r - \frac{\gamma \sigma^2}{2} \left( \frac{3}{4\gamma} \Lambda^2 (\Lambda - 1)^2 \right)^{2/3} \varepsilon^{2/3} + O(\varepsilon).$$

Proof. See Section D. □

In contrast to the risk-averse objectives considered above, the risk-neutral objective leads to a solution which does not have a frictionless analogue: for small trading costs, both the optimal policy and its performance become unbounded as the optimal leverage increases arbitrarily. The next result describes the solution to the risk-neutral problem, identifying the approximate dependence of the leverage multiplier and its performance on the asset’s risk, return, and liquidity.

**Theorem 3.3** (Risk Neutrality and Limits of Leverage). Let $\Lambda = \gamma = 0$.

(i) There exists $\varepsilon_0 > 0$ such that for all $\varepsilon < \varepsilon_0$, the free boundary problem (3.1)–(3.5) has a unique solution $(W, \zeta_-, \zeta_+)$ with $\zeta_- < \zeta_+$.

(ii) The trading strategy $\dot{\varphi}$ that buys at $\pi_- := \zeta_-/(1 + \zeta_-)$ and sells at $\pi_+ := \zeta_+/(1 + \zeta_+)$ to keep the risky weight $\pi_t = \zeta_t/(1 + \zeta_t)$ within the interval $[\pi_-, \pi_+]$ is optimal.
The maximum expected return is
\[
\max_{\varphi \in \Phi} \lim_{T \to \infty} \frac{1}{T} \int_0^T \frac{dw_t}{w_t} = r + \mu \pi_-.
\] (3.22)

The trading boundaries have the series expansions
\[
\pi_- = (1 - \kappa) \kappa^{1/2} \left( \frac{\mu}{\sigma^2} \right)^{1/2} \varepsilon^{-1/2} + 1 + O(\varepsilon^{1/2}),
\] (3.23)
\[
\pi_+ = \kappa^{1/2} \left( \frac{\mu}{\sigma^2} \right)^{1/2} \varepsilon^{-1/2} + 1 + O(\varepsilon^{1/2}),
\] (3.24)

where \( \kappa \approx 0.5828 \) is the unique solution of
\[
\frac{3}{2} \xi + \log(1 - \xi) = 0, \quad \xi \in (0, 1).
\]

Proof. See Section E.

The next section discusses how these results modify the familiar intuition about risk, return, and performance evaluation in the context of trading costs.

4 Implications and Applications

4.1 Efficient frontier \((\Lambda = 0)\)

Theorem 3.1 extends the familiar efficient frontier to account for trading costs. Compared to the linear frictionless frontier, average returns decline because of rebalancing losses. Average volatility increases because more risk becomes necessary to obtain a given return net of trading costs.

To better understand the effect of trading costs on return and volatility, consider the dynamics of the portfolio weight in the absence of trading, which is
\[
d\pi_t = \pi_t(1 - \pi_t)(\mu - \sigma^2 \pi_t)dt + \sigma \pi_t(1 - \pi_t)dB_t.
\] (4.1)

The central quantity here is the portfolio weight volatility \(\sigma \pi_t(1 - \pi_t)\), which vanishes for the single-asset portfolios \(\pi_t = 0\) or \(\pi_t = 1\), remains bounded above by \(\sigma/4\) in the long-only case \(\pi_t \in [0, 1]\), and rises quickly with leverage \((\pi_t > 1)\). This quantity is important because it measures the extent to which a portfolio, left to itself, strays from its initial composition in response to market shocks and, by reflection, the quantity of trading that is necessary to keep it within some region. In the long-only case, the portfolio weight volatility decreases as the no-trade region widens to span \([0, 1]\), which means that a portfolio tends to spend more time near the boundaries. By contrast, with leverage portfolio weight volatility increases, which means that a wider boundary does not necessarily mitigate trading costs.

Consistent with this intuition, equations (3.8), (3.10) show that the impact of trading costs is smaller on long-only portfolios, but rises quickly with leverage. Small trading costs reduce returns and increase volatility at the order of \(\varepsilon^{2/3}\) but, crucially, as leverage increases the error of this approximation also increases, and lower values of \(\gamma\) make it precise for ever smaller values of \(\varepsilon\).
Figure 2: Efficient Frontier with trading costs, as expected excess return (vertical axis, in multiples of the asset’s expected excess return) against standard deviation (horizontal axis, in multiples of the asset’s volatility). The asset has expected excess return $\mu = 8\%$, volatility $\sigma = 16\%$, and bid-ask spread of 0.1%, 0.5%, 1%. The upper line is the frictionless efficient frontier. The maximum of each curve is the leverage multiplier.

The performance (3.12) coincides at the first order with the equivalent safe rate from utility maximization with constant relative risk aversion $\gamma$ (Gerhold et al., 2014, Equation (2.4)), supporting the interpretation of $\gamma$ as a risk-aversion parameter, and confirming that, for asymptotically small costs, the efficient frontier captures the risk-return trade-off faced by a utility maximizer.

Figure 2 displays the effect of trading costs on the efficient frontier. As the bid-ask spread declines, the frontier increases to the linear frictionless frontier, and the asymptotic results in the theorem become more accurate. However, if the spread is held constant as leverage (hence volatility) increases, the asymptotic expansions become inaccurate, and in fact the efficient frontier ceases to increase at all after the leverage multiplier is reached.
4.2 Trading Boundaries ($\Lambda = 0$)

Each point in the efficient frontier corresponds to a rebalancing strategy that is optimal for some value of the risk-aversion parameter $\gamma$. For small trading costs, equation (3.7) implies that the trading boundaries corresponding to the efficient frontier depart from the ones arising in utility maximization, which are (Gerhold et al., 2014)

$$
\pi_{\pm} = \pi_s \pm \left( \frac{3}{4\gamma} (\pi_s)^2 (1 - \pi_s)^2 \right)^{1/3} \varepsilon^{1/3} + O(\varepsilon). \tag{4.2}
$$

The term of order $\varepsilon^{2/3}$ vanishes for $\gamma = 1$ because this case coincides with the maximization of logarithmic utility. For high levels of leverage ($\gamma < 1$ and $\pi_s > 1$), this term implies that the trading boundaries that generate the efficient frontier are lower than the trading boundaries that maximize utility. In Figure 3, $\gamma \to \infty$ corresponds to the safe portfolio in the origin $(0,0)$, while $\gamma = \mu/\sigma^2$ to the risky investment $(1,1)$, which has by definition the same volatility and return.

Figure 3: Trading boundaries $\pi_{\pm}$ (vertical axis, outer curves, as multiples of wealth in risky assets) and implied Merton fraction (middle curve) against average portfolio volatility (horizontal axis, as multiples of $\sigma$). $\mu = 8\%$, $\sigma = 16\%$, and $\varepsilon = 1\%$. 
as the risky asset. As $\gamma$ declines to zero, the trading boundaries converge to the right endpoints, which correspond to the strategy that maximizes average return with no regard for risk, thereby achieving the multiplier.

Observe that (Figure 3), as leverage increases, the sell boundary rises more quickly than the buy boundary. For example, the risk-neutral portfolio tolerates leverage fluctuations from approximately 6, below which it will increase the risky position, up to approximately 14, above which it will reduce it. The locations of these boundaries trade off the need to keep exposure to the risky asset high to maximize return while also keeping rebalancing costs low. Risk aversion makes boundaries closer to each other by penalizing the high realized variance generated by the wide risk-neutral boundaries.

Importantly, these boundaries remain finite even as the frictionless Merton portfolio $\mu/(\gamma\sigma^2)$ diverges to infinity as $\gamma$ declines to zero. Thus the no-trade region is obviously not symmetric around the frictionless portfolio, in contrast to the boundaries arising from utility maximization (Gerhold et al., 2014), which are always symmetric, and hence diverge when $\gamma$ is low. The difference is that here the risk-neutral objective is to maximize the expected return of the portfolio, while a risk-neutral utility maximizer focuses on expected wealth. In a frictionless setting this distinction is irrelevant, and an investor can use a return-maximizing policy to maximize wealth instead. But trading costs drive a wedge between these two ostensibly equivalent risk-neutral criteria – maximizing expected return is not the same as maximizing expected wealth.

Theorem 3.3 (iv) describes in the risk-neutral case the optimal trading boundaries, which satisfy the approximate relation

$$\frac{\pi_-}{\pi_+} \approx 0.4172$$

which is universal in that it holds for any asset, regardless of risk, return and liquidity. This relation means that an optimal risk-neutral rebalancing strategy should always tolerate wide variations in leverage over time, and that the maximum allowed leverage should be approximately 2.5 times the minimum. More frequent rebalancing cannot achieve the maximum return: it can be explained either by risk aversion or by elements that lie outside the model, such as price jumps.

The liquidation constraint (A.1) implies that

$$\pi_t < \frac{1}{\epsilon}$$

for every admissible trading strategy. Since $\pi_t \leq \pi_+$ for the optimal trading strategy in Theorem 3.1 and Theorem 3.3, the upper bound (4.4) is never binding for realistic bid-ask-spreads.

## 4.3 Embedded leverage

In frictionless markets, two perfectly correlated assets with equal Sharpe ratio generate the same efficient frontier, and in fact the same payoff space. This equivalence fails in the presence of trading costs: the more volatile asset is superior, in that it generates an efficient frontier that dominates the one generated by the less volatile asset. Figure 4 (top of the three curves) displays this phenomenon: for example, a portfolio with an average return of 50% net of trading costs is obtained from an asset with 25% return and 50% volatility at a small cost, as an average leverage factor of 2 entails moderate rebalancing.

Achieving the same 50% return from an asset with 20% volatility (and 10% return) is more onerous: trading costs require leverage higher than 5, which in turn increases trading costs. Overall,
Figure 4: Efficient Frontier, as average expected excess return (vertical axis) against volatility (horizontal axis), for an asset with Sharpe ratio \( \mu/\sigma = 0.5 \), for various levels of asset volatility, from 10% (bottom), 20%, to 50% (top), for a bid-ask spread \( \varepsilon = 1\% \). The straight line is the frictionless frontier.

the resulting portfolio needs about 120% rather than 100% volatility to achieve the desired 50% average return (middle curve in Figure 4).

From an asset with 10% volatility (and 5% return), obtaining a 50% return net of trading costs is impossible (bottom curve in Figure 4), because the leverage multiplier is less than 8 (Table 1, top right), and therefore the return can be scaled to less than 40%. The intuition is clear: increasing leverage also increases trading costs, calling in turn for more leverage to increase return, but also further increasing costs. At some point, the marginal net return from more leverage becomes zero, and increasing it does more harm than good.

Because an asset with higher volatility is superior to another one, perfectly correlated and with equal Sharpe ratio, but with lower volatility, the model suggests that in equilibrium they cannot coexist, and that the asset with lower volatility should offer a higher return to be held by investors. Indeed, Frazzini and Pedersen (2012, 2014) document significant negative excess returns in assets with embedded leverage (higher volatility), and offer a theoretical explanation based on
Figure 5: Tracking error (vertical axis) against $-\bar{\alpha}$ (horizontal axis), for leveraged (solid) and inverse (dashed) funds, for -3, +4 (top), -2, +3 (middle), -1, +2 (bottom). Risk aversion $\gamma$ increases from zero (left) to $\infty$ (right). A $k+1$-leveraged fund is akin to a $-k$ inverse one, as the respective curves (same color) approach for low and high risk aversion. $\epsilon = 1\%$, $\mu = 0$, $\sigma = 16\%$.

heterogeneous leverage constraints, which lead more constrained investors to bid up prices (and hence lower returns) of more volatile assets. This paper suggests that the same phenomenon may arise even in the absence of constraints, as a result of rebalancing costs. In contrast to constraints-based explanations, our model suggests that the premium for embedded leverage should be higher for more illiquid assets.

4.4 Optimal Replication of a Leveraged Benchmark

Leveraged and inverse ETFs seek to replicate a multiple of the daily return on an index by frequently rebalancing their portfolio to keep a constant leverage ratio, which typically varies between -3 for inverse funds to +3 for leveraged funds.

Portfolio performance measures based on the regression of a fund’s return against its benchmark’s return are ubiquitous and, as a result, are closely monitored by managers who are evaluated
with such performance measures. In theory, in a frictionless market continuous rebalancing yields a perfect replication of a leveraged benchmark, i.e., zero alpha and tracking error. In practice, trading costs create a trade-off between the frequent rebalancing that generates low tracking error and the low trading costs that prevent alpha from becoming too negative. This trade-off becomes especially relevant for funds that seek to replicate large multiples of illiquid benchmarks.

Theorem 3.2 offers the optimal trading policies and their performance for the replication of a benchmark with zero excess return ($\mu = 0$). This assumption is substantively appropriate, as managers of leveraged funds do not seek to outperform their targets by earning a risk premium, which would become optimal for $\mu \neq 0$. Accordingly, the risk aversion parameter $\gamma$ is interpreted as the manager’s aversion to tracking error rather than negative alpha.

The next proposition describes the trade-off between alpha and tracking error for small trading costs.

**Proposition 4.1.** Recall that $\theta_* = \pi_* + \Lambda = \frac{\mu}{\gamma \sigma^2} + \Lambda$. Alpha, beta, and the tracking error have the asymptotic expansions:

\[
\bar{\alpha} = -\text{ATC} = -\frac{3\sigma^2}{\gamma} \left( \frac{\gamma \theta_* (\theta_* - 1)}{6} \right)^{4/3} \varepsilon^{2/3} + O(\varepsilon), \tag{4.5}
\]

\[
\bar{\beta} = \theta_* - \frac{2\theta_* - 1}{\gamma} \left( \frac{\gamma \theta_* (\theta_* - 1)}{6} \right)^{1/3} \varepsilon^{2/3} + O(\varepsilon), \tag{4.6}
\]

\[
\text{TrE} = \sigma \sqrt{3} \left( \frac{\theta_* (\theta_* - 1)}{6 \sqrt{\gamma}} \right)^{2/3} \varepsilon^{1/3} + O(\varepsilon), \tag{4.7}
\]

whence

\[
\bar{\alpha} = -\frac{\sqrt{3}}{12} \sigma^2 \theta_*^2 (\theta_* - 1)^2 \frac{\varepsilon}{\text{TrE}} + O(\varepsilon^{4/3}). \tag{4.8}
\]

This result shows that the alpha of a leveraged portfolio, abstracting from management fees, equals minus the expected costs. The tracking error departs from zero as the spread $\varepsilon$ increases, and as the target leverage $\theta_* = \Lambda$ rises above the buy-and-hold level of one. Other things equal, a fund that seeks to replicate a larger multiple $\Lambda$ of a benchmark’s return has a higher tracking error and a more negative alpha. Thus, it is misleading to compare two funds with different targets, and to conclude that one is better managed than the other merely because its tracking error is lower, or because its alpha is less negative. Even two funds with the same target may be optimally managed, as one may seek lower tracking error at the expense of more negative alpha.

Equation (4.8) offers an approximate relation in terms of observable quantities only, and can be used as a measure of replication performance that controls for the effects of trading costs, volatility, leverage, and tracking error. Thus, subtracting from the realized alpha the fund’s expense ratio and the right-hand side of (4.8) yields the amount of alpha that is unexplained by the model, and hence can be plausibly attributed to managerial skill – or lack thereof.

Consistent with the intuition underlying the previous results, high tracking error is not necessarily evidence of poor manager performance if the underlying asset is illiquid. On the contrary, a savvy management strategy must accept higher tracking error to achieve higher alpha, and low tracking error is consistent only with more negative alpha.

Figure 5 displays this trade-off in logarithmic scale. The small costs regime corresponds to the right side of the figure, in which the inverse relation (negative linear in logarithmic scale) between alpha and tracking error is clear. Consistent with the term $\theta_*^2 (\theta_* - 1)^2$ in equation (4.8), a fund
that tracks $\Lambda > 0$ times the return of a benchmark faces an approximately similar trade-off as an inverse ETF that replicates $-(\Lambda - 1) < 0$ times the return on the benchmark. In short, a $3\times$ leveraged fund is as difficult to manage as a $-2\times$ inverse fund.

The left side of the picture displays the low risk aversion regime, in which tracking error becomes insensitive to alpha, because the manager substantially strays from the benchmark in order to avoid trading costs. To understand this regime it is useful to consider Figure 6, which displays the buy and sell boundaries for optimal replication. In contrast to their similar performance, leveraged and inverse funds are replicated by rather different policies: as risk aversion $\gamma$ declines to zero, the trading boundaries for inverse funds steadily widen ($\lim_{\gamma \downarrow 0} \pi^-(\gamma) = -\infty$ and $\lim_{\gamma \downarrow 0} \pi^+(\gamma) = 0$), as the manager brings alpha closer to zero by reducing trading costs, consistent with intuition.

Less intuitively, as risk aversion $\gamma$ declines to zero the trading boundaries for leveraged funds first widen, then collapse to one ($\lim_{\gamma \downarrow 0} \pi^\pm = 1$). The explanation of this asymmetric pattern is that, unlike negative multiples, the unit multiple entails zero replication cost, which is an attractive alternative when the emphasis is on minimizing costs, at the expense of departing significantly from the target exposure. Accordingly, the resulting tracking error satisfies $\lim_{\gamma \downarrow 0} \text{TrE} = \sigma(\Lambda - 1)$ for $\Lambda > 1$. By contrast, an inverse fund can only allow the exposure to wander in the negative domain, as the closest no-cost exposure is zero, whence $\lim_{\gamma \downarrow 0} \text{TrE} = \sigma|\Lambda|$ for $\Lambda < 0$.

A salient feature of equation 3.21 is that the optimal replication boundaries depend on the risk aversion $\gamma$ and the spread $\varepsilon$, but not on the benchmark’s volatility $\sigma$, which is in fact absent from equation (3.13), thereby suggesting that these boundaries may be optimal even in a stochastic volatility setting. Although the optimal replication strategy is insensitive to the asset’s volatility, note that both its alpha and tracking error worsen as volatility increases. In other words, a rise in volatility does not change the portfolio weights at which it is optimal to buy or sell, but leads to more fluctuations and hence trading, which in turn generate higher tracking error and a more negative alpha.

### 4.5 From risk aversion to risk neutrality

Theorems 3.1 and 3.3 are qualitatively different: while Theorem 3.1 with positive risk aversion leads to a regular perturbation of the Markowitz-Merton solution, Theorem 3.3 with risk-neutrality leads to a novel result with no meaningful analogue in the frictionless setting – a singular perturbation. Furthermore, a close reading of the statement of Theorem 3.1 shows that the existence of a solution to the free-boundary problem, and the asymptotic expansions, hold for $\varepsilon$ less than some threshold $\bar{\varepsilon}(\gamma)$ that depends on the risk aversion $\gamma$. In particular, if $\gamma$ approaches zero while $\varepsilon$ is held constant, Theorem 3.1 does not offer any conclusions on the convergence of the risk-averse to the risk-neutral solution. Still, if the risk-neutral result is to be accepted as a genuine phenomenon rather than an artifact, it should be clarified whether the risk-averse trading policy and its performance converge to their risk neutral counterparts as risk aversion vanishes. The next result resolves this point under some parametric restrictions.

We first introduce the functions

$$G(\zeta) := \frac{\varepsilon}{(1 + \zeta)(1 + (1 - \varepsilon)\zeta)}, \quad h(\zeta) = (\mu + \gamma\sigma^2\Lambda)\left(\frac{\zeta}{1 + \zeta}\right) - \frac{\gamma\sigma^2}{2}\left(\frac{\zeta}{1 + \zeta}\right)^2.$$

We further associate to any solution $(W(\cdot; \gamma), \zeta^-(\gamma), \zeta^+(\gamma))$ of the free boundary problem (3.1) the
Figure 6: Trading boundaries (vertical axis) versus tracking error (horizontal axis) for leveraged (solid) and inverse (dashed) funds, with multipliers 4 (top), 3, 2, −1, −2, −3 (bottom). As risk aversion $\gamma$ decreases from $\infty$ (left) to 0 (right), for inverse funds the trading boundaries widen around the target, whereas for leveraged funds they first widen and then collapse to one. $\varepsilon = 1\%$, $\mu = 0\%$, $\sigma = 16\%$.

Theorem 4.2. Let $\mu > \sigma^2$, $\varepsilon > 0$, and $\bar{\gamma} > 0$, and Assume that, for any $\gamma \in [0, \bar{\gamma}]$ the free boundary problem (3.1) has a unique solution $(W; \zeta_-, \zeta_+)$ satisfying

$$\zeta_+ < -\frac{1}{1-\varepsilon} \quad (4.9)$$
and that the function $\hat{W}$ satisfies, for each $\gamma \in (0, \bar{\gamma})$, the HJB equation

$$\min \left( \frac{\sigma^2}{2} \hat{W}' + \mu \hat{W} - h(\zeta) + h(\zeta_-), G(\zeta) - \hat{W}, \hat{W} \right) = 0. \tag{4.10}$$

Then, (4.10) is satisfied also for $\gamma = 0$, and for each $\gamma \in [0, \bar{\gamma}]$, the trading strategy that buys at $\pi_- (\gamma) = \frac{\zeta_- (\gamma)}{1 + \zeta_- (\gamma)}$ and sells at $\pi_+ (\gamma) = \frac{\zeta_+ (\gamma)}{1 + \zeta_+ (\gamma)}$ to keep the risky weight $\pi_t = \zeta_t / (1 + \zeta_t)$ within the interval $[\pi_-(\gamma), \pi_+(\gamma)]$ is optimal. Furthermore, $\zeta_\pm (\gamma) \to \zeta_\pm (0)$ and $\hat{W} (\zeta; \gamma) \to \hat{W} (\zeta; 0)$ as $\gamma \downarrow 0$, each $\zeta \in \mathbb{R}$.

In summary, this result confirms that, as the risk-aversion parameter $\gamma$ declines to zero, the risk-averse policy in Theorem 3.1 can only converge to the risk-neutral policy in Theorem 3.3, and that the corresponding mean-variance objective in Theorem 3.1 converges to the average return in Theorem 3.3.

## 5 Heuristic Solution

This section offers a heuristic derivation of the HJB equation. Consider the finite-horizon objective

$$\max_{\varphi \in \Phi} \mathbb{E} \left[ \int_0^T \left( (\mu + \gamma \sigma^2 \Lambda) \pi_t - \frac{\gamma \sigma^2}{2} \pi_t^2 \right) dt - \varepsilon \int_0^T \pi_t \text{d} \varphi_t \right] \tag{5.1}$$

From the outset, it is clear that this objective is scale-invariant, because doubling the initial number of risky shares and safe units, and also doubling the number of shares $\varphi_t$ held at time $t$ has the effect of keeping the objective functional constant. Thus, we conjecture that the residual value function $V$ depends on the calendar time $t$ and on the variable $\zeta_t = \pi_t / (1 - \pi_t)$, which denotes the number of shares held for each unit of the safe asset. In terms of this variable, the conditional value of the above objective at time $t$ becomes:

$$F^\varphi (t) = \int_0^t \left( (\mu + \gamma \sigma^2 \Lambda) \frac{\zeta_s}{1 + \zeta_s} - \frac{\gamma \sigma^2}{2} \frac{\zeta_s^2}{(1 + \zeta_s)^2} \right) ds - \varepsilon \int_0^t \frac{\zeta_s}{1 + \zeta_s} \frac{d \varphi_s}{\varphi_s} + V(t, \zeta_t). \tag{5.2}$$

By Itô’s formula, the dynamics of $F^\varphi$ is

$$dF^\varphi (t) = \left( (\mu + \gamma \sigma^2 \Lambda) \frac{\zeta_t}{1 + \zeta_t} - \frac{\gamma \sigma^2}{2} \frac{\zeta_t^2}{(1 + \zeta_t)^2} \right) dt - \varepsilon \frac{\zeta_t}{1 + \zeta_t} \frac{d \varphi_t}{\varphi_t} + V_t (t, \zeta_t) dt + V_c (t, \zeta_t) d \zeta_t + \frac{1}{2} V_{\zeta \zeta} (t, \zeta_t) d \langle \zeta \rangle_t,$$

where subscripts of $V$ denote respective partial derivatives. The self-financing condition (2.1) implies that

$$\frac{d \zeta_t}{\zeta_t} = \mu dt + \sigma dW_t + (1 + \zeta_t) \frac{d \varphi_t}{\varphi_t} + \varepsilon \frac{d \varphi^\uparrow_t}{\varphi_t}, \tag{5.3}$$

which in turn allows to simplify the dynamics of $F^\varphi$ to (henceforth the arguments of $V$ are omitted for brevity)

$$dF^\varphi (t) = \left( (\mu + \gamma \sigma^2 \Lambda) \frac{\zeta_t}{1 + \zeta_t} - \frac{\gamma \sigma^2}{2} \frac{\zeta_t^2}{(1 + \zeta_t)^2} + V_t + \frac{\sigma^2}{2} \zeta_t^2 V_{\zeta \zeta} + \mu \zeta_t V_c \right) dt - \zeta_t \left( V_{\zeta} (1 + (1 - \varepsilon) \zeta_t) + \frac{\varepsilon}{1 + \zeta_t} \right) \frac{d \varphi^\uparrow_t}{\varphi_t} + \zeta_t (1 + \zeta_t) V_{\zeta} \frac{d \varphi^\uparrow_t}{\varphi_t} + \sigma \zeta_t V_c dW_t. \tag{5.4}$$
Now, by the martingale principle of optimal control (Davis and Varaiya, 1973) the process $F^\varphi(t)$ above needs to be a supermartingale for any trading policy $\varphi$, and a martingale for the optimal policy. Since $\varphi^\uparrow$ and $\varphi^\downarrow$ are increasing processes, the supermartingale condition implies the inequalities

$$-\frac{\varepsilon}{(1+\zeta)(1+(1-\varepsilon)\zeta)} \leq V_\zeta \leq 0, \quad (5.6)$$

and the martingale condition prescribes that the left (respectively, right) inequality becomes an equality at the points of increase of $\varphi^\downarrow$ (resp. $\varphi^\uparrow$). Likewise, it follows that

$$(\mu + \gamma \sigma^2 \Lambda) \frac{\zeta}{1+\zeta} - \gamma \sigma^2 \frac{\zeta^2}{2(1+\zeta)^2} + V_\zeta + \frac{\sigma^2}{2} \zeta V_{\zeta\zeta} + \mu V_\zeta \leq 0 \quad (5.7)$$

with the inequality holding as an equality whenever both inequalities in (5.6) are strict. To achieve a stationary (that is, time-homogeneous) system, suppose that the residual value function is of the form $V(t, \zeta) = \lambda(T-t) - \int^\zeta W(z) dz$ for some $\lambda$ to be determined, which represents the average optimal performance over a long period of time. Replacing this parametric form of the solution, the above inequalities become

$$0 \leq W(\zeta) \leq \frac{\varepsilon}{(1+\zeta)(1+(1-\varepsilon)\zeta)}, \quad (5.8)$$

$$(\mu + \gamma \sigma^2 \Lambda) \frac{\zeta}{1+\zeta} - \gamma \sigma^2 \frac{\zeta^2}{2(1+\zeta)^2} - \lambda - \frac{\sigma^2}{2} \zeta W'(\zeta) - \mu \zeta W(\zeta) \leq 0, \quad (5.9)$$

Assuming further that the the first inequality holds over some interval $[\zeta-, \zeta+]$, with each inequality reducing to an equality at the respective endpoint, the optimality conditions become

$$\frac{\sigma^2}{2} \zeta^2 W'(\zeta) + \mu \zeta W(\zeta) - (\mu + \gamma \sigma^2 \Lambda) \frac{\zeta}{1+\zeta} + \frac{\gamma \sigma^2}{2} \frac{\zeta^2}{(1+\zeta)^2} + \lambda = 0 \quad \text{for } \zeta \in [\zeta-, \zeta+], \quad (5.10)$$

$$W'(\zeta-) = 0, \quad (5.11)$$

$$W'(\zeta+) = \frac{\varepsilon}{(\zeta+1)(1+(1-\varepsilon)\zeta+)}, \quad (5.12)$$

which lead to a family of candidate value functions, each of them corresponding to a pair or boundaries $[\zeta-, \zeta+]$. The optimal boundaries are identified by the smooth-pasting conditions, formally derived by differentiating (5.11) and (5.12) with respect to their boundaries

$$W'(\zeta-) = 0, \quad (5.13)$$

$$W'(\zeta+) = \frac{\varepsilon(\varepsilon-2(1-\varepsilon)\zeta_+ - 2)}{(1+\zeta_+)^2(1+(1-\varepsilon)\zeta_+)^2}. \quad (5.14)$$

These conditions allow to identify the value function. The four unknowns are the free parameter in the general solution to the ordinary differential equation (5.10), the free boundaries $\zeta-$ and $\zeta+$, and the optimal rate $\lambda$. These quantities are identified by the boundary and smooth-pasting conditions (5.11)–(5.14).
6 Conclusion

The costs of rebalancing a leveraged portfolio are substantial, and detract from its ostensible frictionless return. As leverage increases, such costs rise faster than the frictionless return, making it impossible for an investor to lever an asset’s return beyond a certain multiple, net of trading costs.

In contrast to the frictionless theory, trading costs make the risk-return trade-off nonlinear. An investor who seeks high return prefers an asset with high volatility to another one with equal Sharpe ratio but lower volatility, because higher volatility makes leverage cheaper to realize. A risk-neutral, return-maximizing investor does not take infinite leverage, but rather keeps it within a band that balances high exposure with low rebalancing costs.

These findings have broad implications in portfolio choice, asset pricing, and financial intermediation. For example, a bank that extends risky loans is akin to an investor trading in an illiquid risky asset: in contrast to the frictionless common wisdom, our results imply that such a bank will not increase its balance sheet without bounds, even if it is neutral to risk and regulatory capital requirements are absent. However, the endogenous finite leverage is sensitive to the volatility and the liquidity of the loans, suggesting that attempts to encourage or discourage bank lending should address these factors.

A direct application of these results is the optimal replication and performance evaluation of leveraged funds. As replication strategies face a trade-off between low tracking error and lower alpha, we derive a testable restriction that any optimal replication policy must satisfy.

A Admissible Strategies

A strategy is admissible if it is nonanticipative and solvent, up to a small increase in the spread:

Definition A.1. Let $x > 0$ (the initial capital) and let $(\varphi^\uparrow_t)_{t \geq 0}$ and $(\varphi^\downarrow_t)_{t \geq 0}$ (the cumulative number of shares bought and sold, respectively) be continuous, increasing processes, adapted to the augmented natural filtration of $B$. $(x, \varphi_t = \varphi^\uparrow_t - \varphi^\downarrow_t)$ is an admissible strategy if its liquidation value is strictly positive at all times: There exists $\varepsilon' > \varepsilon$ such that

$$x - \int_0^t S_s d\varphi_s + S_t \varphi_t - \varepsilon' \int_0^t S_s d\varphi^\uparrow_s - \varepsilon' \varphi^\downarrow_t S_t > 0 \quad \text{a.s. for all } t \geq 0. \quad (A.1)$$

The family of admissible trading strategies is denoted by $\Phi$.

The following lemma shows that, without loss of generality, it is safe to exclude trading strategies that involve short selling or violate certain integrability conditions, because each such strategy cannot be optimal.

Lemma A.2. Let $\varphi \in \Phi$ be optimal for (2.8). Then:

(i) the strategy $\hat{\varphi}_t := \varphi_t 1_{\{\varphi_t \geq 0\}}$ is also optimal; and

(ii) the following integrability conditions hold$^7$

$$\int_0^t \pi_u^2 du < \infty, \quad \int_0^t \pi_u \frac{|\varphi_u|}{\varphi_u} < \infty \quad \text{a.s. for all } t \geq 0, \quad (A.2)$$

$^7$Note that $\frac{d}{\varphi_t} = \frac{S_t}{\varphi_t}$, therefore on the set $\{(\omega, t) : \varphi_t = 0\}$ the quantity $\frac{d}{\varphi_t}$ is well-defined.
where \(\|\varphi_t\|\) denotes the total variation of \(\varphi\) on \([0,t]\).

Proof. Proof of (i): It is clear that \(\hat{\varphi}_t\) is an admissible trading strategy if \(\varphi\) is. Furthermore \(\hat{\pi}_t \geq \pi_t\) at all times \(t\), and \(\hat{\pi}_t = 0\) whenever \(\varphi_t < 0\), whence \(F_T(\hat{\varphi}) \geq F_T(\varphi)\), each \(T > 0\).

Proof of (ii): The first integrability condition in (A.2) is trivially satisfied, since it is a direct consequence of (A.1), which in turn implies \(\pi_t \leq 1/\varepsilon\), for all \(t\), a.s.. To check the second one, suppose that some continuous, finite variation trading strategy \(\varphi\) satisfying \(\pi_t \leq 1/\varepsilon\) exists such that for some \(t > 0\)

\[
L_t := \int_0^t \pi_u \frac{d|\varphi_u|}{\varphi_u} = \infty
\]

with positive probability. Then on the same event one of the following integrals must be infinite, i.e.

\[
L_1^1 := \int_0^t \pi_u \frac{d\varphi^\uparrow_u}{\varphi_u} = \infty, \quad \text{or} \quad L_1^2 := \int_0^t \pi_u \frac{d\varphi^\downarrow_u}{\varphi_u} = \infty.
\]

The former case leads to infinite average transaction costs (the third term times \(-1/\varepsilon\) in (2.7)), hence the objective functional equals \(-\infty\), which is outperformed by any buy-and-hold strategy. Likewise, if \(L_2^1 = \infty\), then

\[
\log(\varphi_t) - \log(\varphi_0) = \int_0^t \frac{d\varphi_u}{\varphi_u} = \infty
\]

which is impossible, since \(\varphi_t\) must be finite at all times due to proportional transaction costs. \(\square\)

The following lemma describes the dynamics of a the wealth process \(w_t\), the risky weight \(\pi_t\), and the risky/safe ratio \(\zeta_t\).

**Lemma A.3.** For any admissible trading strategy \(\varphi\), \(^8\)

\[
\frac{d\zeta_t}{\zeta_t} = \mu dt + \sigma dB_t + (1 + \zeta_t) \frac{d\varphi^\uparrow_t}{\varphi_t} - (1 + (1 - \varepsilon)\zeta_t) \frac{d\varphi^\downarrow_t}{\varphi_t}, \tag{A.3}
\]

\[
\frac{dw_t}{w_t} = rd t + \pi_t(\mu dt + \sigma dB_t - \varepsilon \frac{d\varphi^\uparrow_t}{\varphi_t}), \tag{A.4}
\]

\[
\frac{d\pi_t}{\pi_t} = (1 - \pi_t)(\mu dt + \sigma dB_t) - \pi_t(1 - \pi_t) \sigma^2 dt + \frac{d\varphi^\uparrow_t}{\varphi_t} - (1 - \varepsilon \pi_t) \frac{d\varphi^\downarrow_t}{\varphi_t}. \tag{A.5}
\]

For any such strategy, the functional

\[
F_T(\varphi) := \frac{1}{T} \mathbb{E} \left[ \int_0^T \frac{dw_t}{w_t} - \gamma - \frac{1}{2} \int_0^T \left\langle \frac{dw_t}{w_t} - \Lambda \frac{dS_t}{S_t} \right\rangle dt \right] \tag{A.6}
\]

equals to

\[
F_T(\varphi) = r + \frac{1}{T} \mathbb{E} \left[ \int_0^T \left( \mu \pi_t - \gamma \frac{\sigma^2}{2} (\pi_t - \Lambda)^2 \right) dt - \varepsilon \int_0^T \pi_t \frac{d\varphi^\downarrow_t}{\varphi_t} \right]. \tag{A.7}
\]

\(^8\)The notation \(\frac{dX_t}{X_t} = dY_t\) means \(X_t = X_0 + \int_0^t X_s dY_s\), hence the SDEs are well defined even for zero \(X_t\).
Proof. Denoting by $X_t$ and $Y_t$ the wealth in the safe and risky positions respectively, the self-financing condition boils down to

$$dX_t = rX_t dt - S_t d\varphi_t^+ + (1 - \varepsilon) S_t d\varphi_t^-, \quad (A.8)$$

$$dY_t = S_t d\varphi_t^- - S_t d\varphi_t^+ + \varphi_t S_t dt. \quad (A.9)$$

and hence

$$\frac{dX_t}{X_t} = rd - \zeta d\varphi_t^+ + (1 - \varepsilon) \zeta d\varphi_t^-, \quad (A.10)$$

$$\frac{dY_t}{Y_t} = \frac{d\varphi_t^-}{\varphi_t} - \frac{d\varphi_t^+}{\varphi_t} + \frac{dS_t}{S_t}, \quad (A.11)$$

$$\frac{d(Y_t/X_t)}{Y_t/X_t} = \frac{dY_t}{Y_t} - \frac{dX_t}{X_t} + \frac{d(X)_t}{X_t} - \frac{d(X,Y)_t}{X_t Y_t} = \frac{dY_t}{Y_t} - \frac{dX_t}{X_t}. \quad (A.12)$$

Equation (A.3) follows from the last equation, and (A.4) holds in view of equation (A.10) and (A.11). For the derivation of equation (A.5), one uses the identity $\pi_t = 1 - \frac{1}{1 + \zeta}$ and (A.3). The expression in (A.7) for the objective functional follows from equation (A.4).

Remark A.4. Throughout the appendix, we shall focus on the following objective functional, which is tantamount to maximizing the Equivalent Safe Rate (see equation (2.8)),

$$F_\infty(\varphi) = \lim_{T \to \infty} F_T(\varphi) - \frac{\gamma \sigma^2 \Lambda^2}{2}$$

$$= r + \frac{1}{T} \lim_{T \to \infty} \frac{1}{T} \mathbb{E} \left[ \int_0^T \left( (\mu + \gamma \sigma^2 \Lambda) \pi_t - \frac{\gamma \sigma^2}{2} \pi_t^2 \right) dt - \varepsilon \int_0^T \pi_t \frac{d\varphi_t^+}{\varphi_t} \right].$$

A.1 Proof of Lemma 2.1

Proof of Lemma 2.1. See the second part of Lemma A.3.

□

B Proof of Theorem 3.1

This section contains a series of propositions, which lead to the proof of Theorem 3.1 (i)–(iii). Part (iv) of the theorem is postponed to Appendix C. Set

$$G(\zeta) := \frac{\varepsilon}{(1 + \zeta)(1 + (1 - \varepsilon) \zeta)} \quad \text{and} \quad h(\zeta) := (\mu + \gamma \sigma^2 \Lambda) \left( \frac{\zeta}{1 + \zeta} \right) - \frac{\gamma \sigma^2}{2} \left( \frac{\zeta}{1 + \zeta} \right)^2. \quad (B.1)$$

Defining $H := h'$, the free boundary problem (3.1)–(3.5) reduces to

$$\frac{1}{2} \sigma^2 \zeta^2 W''(\zeta) + (\sigma^2 + \mu) \zeta W'(\zeta) + \mu W(\zeta) - H(\zeta) = 0, \quad (B.2)$$

$$W(\zeta_-) = 0, \quad (B.3)$$

$$W'(\zeta_-) = 0, \quad (B.4)$$

$$W(\zeta_+) = G(\zeta_+), \quad (B.5)$$

$$W'(\zeta_+) = G'(\zeta_+). \quad (B.6)$$
For the Merton fraction shifted by the target multiplier, the notation

\[ \theta_* = \pi_* + \Delta \]

is used.

**Proposition B.1.** Let \( \gamma > 0 \). For sufficiently small \( \varepsilon \), the free boundary problem (B.2)–(B.6) has a unique solution \((W, \zeta_-, \zeta_+)\), with \( \zeta_- < \zeta_+ \). The free boundaries have the asymptotic expansion

\[
\zeta_\pm = \frac{\theta_*}{1 - \theta_*} \pm \left( \frac{3}{4\gamma} \right)^{1/3} \left( \frac{\theta_*}{(\theta_* - 1)^2} \right)^{2/3} \varepsilon^{1/3} - \frac{2\gamma\Delta + (5 - 2\gamma)\theta_*}{2\gamma(\theta_* - 1)^2} \left( \frac{\gamma\theta_*(\theta_* - 1)}{6} \right)^{1/3} \varepsilon^{2/3} + O(\varepsilon).
\]

(B.7)

**Proof of Proposition B.1.** Since \( \zeta_- \notin \{-1, 0\} \), any solution of the initial value problem (B.2)–(B.4) is of the form

\[
W(\zeta_-, \zeta) = \frac{2}{(\sigma\zeta)^2} \int_{\zeta_-}^{\zeta} (h(y) - h(\zeta_-)) \left( \frac{y}{\zeta} \right)^{2\gamma\pi_* - 2} dy.
\]

(B.8)

Suppose \((W, \zeta_-, \zeta_+)\) is a solution of (B.2)–(B.6). Due to (B.8) \( W(\cdot) \equiv W(\zeta_-, \cdot) \). Let

\[
J(\zeta_-, \zeta) := \frac{\sigma^2 \varepsilon^{2\gamma\pi_*}}{2} W(\zeta_-, \zeta).
\]

(B.9)

By the terminal conditions (B.5)–(B.6) at \( \zeta_+ \), and setting \( \delta = \varepsilon^{1/3} \), \((\zeta_-, \zeta_+)\) satisfy the following system of non-linear equations,

\[
\Psi_1(\zeta_-, \zeta_+) := W(\zeta_-, \zeta_+) - \frac{\delta^3}{(1 + \zeta_+)(1 + (1 - \delta^3)\zeta_+)} = 0,
\]

(B.10)

\[
\Psi_2(\zeta_-, \zeta_+) := \frac{2(h(\zeta_+) - h(\zeta_-))}{\sigma^2 \zeta_+^2} W(\zeta_-, \zeta_+) - \frac{2\gamma\pi_*}{\zeta_+} W(\zeta_-, \zeta_+) - \frac{(1 - \delta^3)(1 + (1 - \delta^3)\zeta_+)^2}{2 + (1 + (1 - \delta^3)\zeta_+)^2} = 0.
\]

(B.11)

Conversely, if \((\zeta_-, \zeta_+)\) solve (B.10)–(B.11), then the triplet \((W(\cdot; \zeta_-), \zeta_-, \zeta_+)\) provides a solution to the free boundary problem (B.2)–(B.6). Therefore, to provide a unique solution of the free boundary problem, it suffices to provide a unique solution of (B.10)–(B.11).

To obtain a guess for the asymptotic expansions of \( \zeta_\pm \), develop \( \Psi_{1,2} \) around

\[
\zeta_\mp = \zeta_* + B_{1,2} \delta + O(\delta^2), \quad \zeta_* = \frac{\theta_*}{1 - \theta_*},
\]

which yields

\[
\Psi_1(\zeta_\pm(\delta)) = -\frac{\gamma(1 - \theta_*)^6}{3\theta_*^2} \left( 2B_1^3 - 3B_1^2B_2 + B_2^3 + \frac{3\theta_*^2}{\gamma(1 - \theta_*)^4} \right) \delta^3 + O(\delta^4),
\]

(B.12)

\[
\Psi_2(\zeta_\pm(\delta)) = \frac{(B_1 - B_2)(B_1 + B_2)\gamma(\theta_* - 1)^6}{\theta_*^2} \delta^2 + O(\delta^3).
\]

(B.13)

By equating the coefficients of the leading order terms to zero, one arrives at the system,

\[
2B_1^3 - 3B_1^2B_2 + B_2^3 + \frac{3\theta_*^2}{\gamma(1 - \theta_*)^4} = 0,
\]

(B.14)

\[
B_1 + B_2 = 0,
\]

(B.15)
which implies $B_1 = -B_2$ solves
\[
B_1^2 = -\frac{3}{4\gamma} \frac{\theta_*^2}{(1 - \theta_*)^2} = 0,
\]
which is an equation with a single, real-valued solution, namely
\[
B_1 = -\left(\frac{3}{4\gamma}\right)^{1/3} \frac{\theta_*}{(1 - \theta_*)^{2/3}}.
\] (B.16)

Claim: For sufficiently small $\delta$ the system (B.10)–(B.11) has a unique analytic solution around
\[
\zeta_{0,\pm} := \frac{\theta_*}{1 - \theta_*} \pm \left(\frac{3}{4\gamma}\right)^{1/3} \frac{\theta_*}{(1 - \theta_*)^{2/3}} \delta.
\]
This is equivalent to claiming that the corresponding system of equations $\Phi = (\Phi_1, \Phi_2) = 0$ for $(\eta_-, \eta_+)$, around $(B_1, B_2)$ has a unique solution, where
\[
\eta_{\pm} := \frac{\zeta_{\pm} - \frac{\theta_*}{1 - \theta_*}}{\delta}
\]
and
\[
\Phi_1 := \Psi_1(\zeta_-(\eta_-), \zeta_+(\eta_+)) \quad \Phi_2 := \Psi_2(\zeta_-(\eta_-), \zeta_+(\eta_+)).
\] (B.17)

By Proposition B.2, there exists a unique solution for sufficiently small $\delta > 0$, which is analytic in $\delta$. Hence, also the original system $\Psi(\zeta_-, \zeta_+) = 0$ has a unique solution $(\zeta_-, \zeta_+)$ around $\frac{\theta_*}{1 - \theta_*}$. As a consequence, the free boundary problem (B.2)–(B.6) has a unique solution for sufficiently small $\varepsilon$.

To derive the higher order terms of (B.7), it is useful to rewrite the integral (B.9) as
\[
J(\zeta_-, \zeta_+) = \frac{h(\zeta_-(\eta_-), \zeta_+(\eta_+))}{2\gamma \pi_* - 1} + \int_{\zeta_-}^{\zeta_+} h(y) y^{2\gamma \pi_* - 2} dy.
\] (B.18)

The derivative of $I_2$ with respect to $\delta$ equals
\[
\frac{dI_2}{d\delta} = h(\zeta_+(\eta_+)) \frac{2\gamma \pi_* - 2 d\zeta_+}{d\delta} - h(\zeta_-(\eta_-)) \frac{2\gamma \pi_* - 2 d\zeta_-}{d\delta},
\] (B.19)
and shall be expanded as a power series in $\delta$. Integration with respect to $\delta$ then yields an asymptotic expansion of $I_2$.

To obtain these expansions, guess a solution of equations (B.10)–(B.11) of the form
\[
\zeta_{\pm} = \frac{\theta_*}{1 - \theta_*} \pm \left(\frac{3}{4\gamma}\right)^{1/3} \frac{\theta_*}{(1 - \theta_*)^{2/3}} \delta + A_{\pm} \delta^2 + O(\delta^3),
\]
for some unknowns $A_{\pm}$, and substitute it into equations (B.10)–(B.11), using thereby (B.18) and (B.19). Comparing the coefficients in the asymptotic expansion of the two equations reveals that
\[
A_- = A_+ = \left(\frac{2\gamma \Lambda + (5 - 2\gamma) \theta_*}{2\gamma(1 - \theta_*)^2}\right) \left(\frac{\gamma \theta_*(1 - \theta_*)}{6}\right)^{1/3},
\]
and thus (B.7) is derived. \[ \Box \]
Proposition B.2. For sufficiently small $\delta > 0$, the system $\Phi(\eta_{\pm}(\delta), \delta) = 0$ defined by (B.17), has a unique solution $(\eta_-(\delta), \eta_+(\delta))$ near $(B_1, B_2)$, which is analytic in $\delta$.

Proof. Denote by $D\Phi$ the Frechet differential of $\Phi$. Claim: the Jacobian satisfies

$$\det(D\Phi)(\eta_- = B_1, \eta_+ = B_2, \delta = 0) = \frac{6\gamma(1 - \theta_*)^3(2\gamma\pi_* - 1)}{\theta_*^2} \neq 0,$$

(B.20)

hence the implicit function theorem for analytic functions (Gunning and Rossi, 2009, Theorem I.B.4) ensures that for sufficiently small $\delta$ there exists a unique solution $(\eta_-, \eta_+)$ of $\Phi(\eta_-, \eta_+)=0$ around $(B_1, B_2)$ which is analytic in $\delta$.

So it remains to verify the identity in (B.20). To this end, note that, by construction,

$$\Psi_2(\zeta_-, \zeta_+) := \frac{\partial \Psi_1(\zeta_-, \zeta_+)}{\partial \zeta_+},$$

whence

$$\frac{\partial \Phi_1(\eta_-, \eta_+)}{\partial \eta_+} |_{(B_1, B_2, 0)} = 0.$$

Thus

$$\det(D\Phi)(B_1, B_2, 0) = \frac{\partial \Phi_1(\eta_-, \eta_+)}{\partial \eta_-} |_{(B_1, B_2, 0)} \times \frac{\partial \Phi_2(\eta_-, \eta_+)}{\partial \eta_+} |_{(B_1, B_2, 0)}.$$

Since

$$\frac{\partial \psi_1}{\partial \zeta_-} = -\frac{2h'(\zeta_-)}{2\mu/\sigma^2-1} \left(\frac{\zeta_+^{2\mu/\sigma^2-2}}{2\mu/\sigma^2-1} - \frac{\zeta_-^{2\mu/\sigma^2-2}}{2\mu/\sigma^2-1}\right)$$

and since by the chain rule

$$\frac{\partial \Phi_1(\eta_-, \eta_+)}{\partial \eta_-} = \frac{1}{\delta_3} \frac{\partial \psi_1}{\partial \zeta_-} \times \delta$$

it follows that

$$\frac{\partial \Phi_1(\eta_-, \eta_+)}{\partial \eta_-} |_{(B_1, B_2, 0)} = \frac{6^{2/3}(1 - \theta_*)^3(\gamma\theta_*(1 - \theta_*) + 1)^{1/3}(1 - 2\gamma\pi_*)}{\theta_*}.$$

Similarly,

$$\frac{\partial \Phi_2(\eta_-, \eta_+)}{\partial \eta_+} |_{(B_1, B_2, 0)} = -\frac{6^{1/3}(1 - \theta_*)^4(\gamma(1 - \theta_*)\theta_*)^{2/3}}{\theta_*^2},$$

from which (B.20), and hence the assertion, follows. \qed

In the following, $C^2(A)$ denotes the space of twice continuously differentiable functions on an open set $A \subset \mathbb{R}$.

Definition B.3. A solution of the HJB equation is a pair $(V, \lambda)$, where $V \in C^2$ and $\lambda \in \mathbb{R}$, which satisfies

$$\min(\mathcal{A}V(x) - h(x) + \lambda, G(x) - V'(x), V'(x)) = 0, \quad x \in \left(-\infty, -\frac{1}{1-\varepsilon}\right) \cup (0, \infty),$$

(B.21)

where $\mathcal{A} : C^2(\mathbb{R}) \mapsto C^2(\mathbb{R})$ is the differential operator

$$\mathcal{A}f(x) := \frac{\sigma^2}{2} x^2 f''(x) + \mu x f'(x).$$
Proposition B.4. Let \((W, \zeta-, \zeta+)\) be the solution of the free boundary problem (B.5)–(B.6) (provided by Proposition B.1) with asymptotic expansion (B.7). For sufficiently small \(\varepsilon\), the pair

\[ V(\cdot) := \int_0^\zeta \hat{W}(\zeta)d\zeta, \quad \lambda := h(\zeta_), \]

where

\[ \hat{W}(\zeta) := \begin{cases} 0 & \text{for } \zeta < \zeta_-, \\ W(\zeta) & \text{for } \zeta \in [\zeta-, \zeta_+], \\ G(\zeta) & \text{for } \zeta \geq \zeta_+, \end{cases} \tag{B.22} \]

is a solution of the HJB equation (B.21).

Proof of Proposition B.4. To check that \((V, \lambda)\) solves the HJB equation (B.21), consider separately the domains \([\zeta-, \zeta_+], \zeta < \zeta_\) and \(\zeta > \zeta_+\). In the following, the decompositions

\[ G(\zeta) = \frac{1}{1 + \zeta} - \frac{1 - \varepsilon}{1 + (1 - \varepsilon)\zeta} \quad \text{and} \quad G'(\zeta) = \left( \frac{1 - \varepsilon}{1 + (1 - \varepsilon)\zeta} \right)^2 - \frac{1}{(1 + \zeta)^2} \]

are used. First, note that on \([\zeta-, \zeta_+],\) by construction it holds that

\[ (\mathcal{A}V(\zeta) - h(\zeta) + h(\zeta_+))^\prime = \frac{1}{2}\sigma^2\zeta^2W''(\zeta) + (\sigma^2 + \mu)\zeta W'(\zeta) + \mu W(\zeta) - H(\zeta) = 0. \]

Furthermore, in view of the initial conditions (B.3)–(B.4),

\[ (\mathcal{A}V(\zeta) - h(\zeta) + h(\zeta_+)) \big|_{\zeta=\zeta_-} = \mathcal{A}V(\zeta) \big|_{\zeta=\zeta_-} = 0, \]

whence

\[ \mathcal{A}V(\zeta) - h(\zeta) + h(\zeta_-) \equiv 0, \quad \zeta \in [\zeta-, \zeta_+]. \]

To see that \(0 \leq V' \leq G\) on all of \([\zeta-, \zeta_+],\) observe that

\[ (h(\zeta) - h(\zeta_-))^\prime = h'(\zeta) = H(\zeta) = \frac{\mu}{\pi_*(1 + \zeta)^2} \left( \pi_* - \frac{(1 - \Lambda)\zeta - \Lambda}{1 + \zeta} \right). \tag{B.23} \]

Note that for \(\zeta_- < \zeta \leq \zeta^\ast,\) where \(\zeta^\ast/(1 + \zeta^\ast) = \theta_*,\) \(V'(\zeta) = W'(\zeta) > 0.\) It is shown that also \(W(\cdot) \geq 0\) on all of \([\zeta-, \zeta_+].\) This is equivalent to showing non-negativity of

\[ w(\zeta) := 2\sigma^2\zeta^2\pi^\ast W(\zeta) = \int_{\zeta_-}^\zeta (h(x) - h(\zeta_-))x^{2\gamma\pi^\ast - 2}dx. \tag{B.24} \]

Now \(w'(\zeta) = (h(\zeta) - h(\zeta_-))\zeta^{2\gamma\pi^\ast - 2} = 0\) if and only if \(h(\zeta_-) = h(\zeta).\) Hence, either \(\zeta = \zeta_-\) or \(\zeta\) satisfies the implicit equation

\[ \pi(\zeta) = \frac{\zeta}{1 + \zeta} = 2(\pi_* + \Lambda) - \pi_. \]

By the first-order asymptotics of (B.7), one thus obtains \(\tilde{\zeta} \notin [\zeta-, \zeta_+]\) for sufficiently small \(\varepsilon.\) Therefore \(w' > 0\) on \((\zeta_, \zeta_+],\) and by (B.24) it follows that \(V' \geq 0\) on all of \([\zeta-, \zeta_+].\) To conclude the validity of the HJB equation on \([\zeta-, \zeta_+],\) it only remains to show the inequality \(V' \leq G.\) To
this end, notice that \( \Psi_1(\zeta) = W(\zeta) - G(\zeta) \), (this is the function defined in (B.10), with fixed \( \zeta_- \)) satisfies
\[
\Psi_1(\zeta_-) = -G(\zeta_-) = -\frac{\epsilon}{(1 + \zeta_-(1 + (1 - \epsilon)\zeta_-)} = -(1 - \theta_*)^2\epsilon + O(\epsilon^{4/3}),
\]
hence for sufficiently small \( \epsilon \), \( \Psi_1(\zeta) < 0 \) on some interval \( [\zeta_-, \bar{\zeta}] \), and \( \Psi_1(\bar{\zeta}) = 0 \). Therefore, \( \bar{\zeta} \leq \zeta_+ \). Since \( \Psi_1(\zeta_+) = 0 \) by construction, it suffices to show that \( \bar{\zeta} = \zeta_+ \) to prove non-negativity of \( V' \) on \( [\zeta_-, \zeta_+] \). Assume, for a contradiction, there exists a sequence \( \delta_k \downarrow 0 \) such that for each \( k \in \mathbb{N} \), \( \Psi_1(\zeta(\delta_k)) = 0 \), and that \( \zeta_-(\delta_k) < \zeta(\delta_k) < \zeta_+(\delta_k) \). Changing to the variable \( u = \frac{\zeta - \zeta_-}{\delta} \), and introducing the notation \( u_\pm = \frac{\zeta_\pm - \zeta_-}{\delta} \), \( \bar{u} = \frac{\bar{\zeta} - \zeta_-}{\delta} \) shall prove convenient. By selecting, if necessary, a subsequence, one may without loss of generality that \( \bar{u}(\delta_k) \) converges, hence it must satisfy
\[
\lim_{k \to \infty} \bar{u}(\delta_k) =: B_0 \in [B_1, B_2],
\]
where \( B_1 \) is defined in (B.16), and \( B_2 = -B_1 \). The calculations leading to (B.16) therefore entail that \( B_0 \) must satisfy (B.14) in place of \( B_2 \), i.e.
\[
2B_1^3 - 3B_1^2B_0 + B_0^3 + \frac{3\theta_*^2}{\gamma(1 - \theta_*)^4} = 0.
\]
With \( B_1 \) from (B.16) and the change of variable \( \xi = -B_0/B_1 \) implies
\[
2 - 3\xi + \xi^3 = 0
\]
which has the only solutions 1 and -2. Therefore, (B.25) has the only relevant solution
\[
B_0 = -B_1 = B_2.
\]
By intertwining \( u_+(\delta) \) and \( \bar{u}(\delta_k) \), one can introduce
\[
\bar{u}^*(\delta) = \begin{cases} 
\bar{u}(\delta_k), & k \in \mathbb{N} \\
u_+(\delta), & \text{otherwise} 
\end{cases}
\]
Hence \( (u_-(\delta), u^*(\delta)) \) satisfies \( \Phi(u_-, u^*) = 0 \) near \( (B_1, B_2) \), for sufficiently small \( \delta \). By Proposition B.2, \( u^*(\delta) = u_+(\delta) \), which contradicts our assumption \( \bar{\zeta} \neq \zeta_+ \).

Consider now \( \zeta \leq \zeta_- \). \( V \) solves the HJB equation, if
\[
AV - h(\zeta) + h(\zeta_-) = h(\zeta_-) - h(\zeta) \geq 0, \quad G(\zeta) \geq 0.
\]
Since \( h(\zeta) - h(\zeta_-) = 0 \) for \( \zeta = \zeta_- \), to obtain the first inequality it suffices to show that (B.23) is non-negative. Now for small \( \epsilon \) clearly \( \pi_- < \theta_* \), hence for \( \zeta = \zeta_- \) (B.23) is indeed strictly positive. To settle the second inequality, recall that either \( \zeta < -1/(1 - \epsilon) \) or \( \zeta > 0 \). On these domains, \( G \) is clearly a strictly positive function. Hence it is proved that \( V \) satisfies the HJB equation for \( \zeta \leq \zeta_- \).

Finally, consider \( \zeta \geq \zeta_+ \). Since \( G = W \), it suffices to show
\[
L(\zeta) := AV(\zeta) - h(\zeta) + h(\zeta_-) \geq 0, \quad G(\zeta) \geq 0.
\]
The second estimate is straightforward: Let \( \zeta_+ > -1 \), then \( \zeta > -1 \), and for sufficiently small \( \epsilon \), \( (1 - \epsilon)\zeta > -1 \), hence \( G(\zeta) > 0 \). The case \( \zeta_+ < -1 \) can be dealt with similarly. For the first inequality in (B.26), note that
\[
L(\zeta) = \frac{\sigma^2 \zeta^2}{2}G'(\zeta) + \mu \zeta G(\zeta) - h(\zeta) + h(\zeta_-) =: \kappa(\zeta)
\]
is a rational function and of course \( \kappa(\zeta) = 0 \). Therefore it suffices to show \( \kappa \) has no zeros on \([\zeta_+, -1/(1-\varepsilon)]\), besides \( \zeta_+ \).

The case \( \gamma = 1 \) is simple as \( \kappa(\zeta) = 0 \) can be reduced to solving a quadratic equation (see also Takacs, Klass and Assaf (1988)). All other cases require investigating a fourth-order polynomial, as seen below. However, to demonstrate the strength and clarity of the asymptotic approach of this paper, the case \( \gamma = 1 \) is discussed first. The transformation \( z = \frac{\zeta}{1-\varepsilon} \) leads to

\[
\frac{(1-\varepsilon)\zeta}{1+(1-\varepsilon)\zeta} = \frac{(1-\varepsilon)z}{1-\varepsilon z}
\]

and thus one can rewrite \( \kappa \) in terms of \( z \), denoting it by

\[
F(z, \varepsilon) = \kappa(z(\zeta))
\]

It is proved next that \( F \) has no zeros on \((\pi_+, 1/\varepsilon)\). Since \( F(\pi_+) = 0 \), polynomial division by \((z - \pi_+)\) yields

\[
F(z, \varepsilon) = \frac{(z - \pi_+)}{(1-\varepsilon z)^2} g(z), 
\]

where \( g(z) \) is a linear factor, and the following asymptotic expansions hold

\[
g(\pi_+) = \sigma^2 \left( \frac{3}{4\gamma} \left( \frac{\mu}{\sigma^2} + \Lambda \right) \left( 1 - \left( \frac{\mu}{\sigma^2} + \Lambda \right) \right)^2 \right)^{1/3} \varepsilon^{1/3} + O(\varepsilon^{2/3}),
\]

\[
g(1/\varepsilon) = \frac{\sigma^2}{2\varepsilon} + O(1).
\]

It follows that \( g \) has no zeros on \([\pi_+, 1/\varepsilon]\), for sufficiently small \( \varepsilon \). Hence \( F(z) > 0 \) for \( z \in (\pi_+, 1/\varepsilon) \).

For the remainder of the proof, suppose \( \gamma \neq 1 \). Using the transformation \( z = \frac{\zeta}{1-\varepsilon} \) one can rewrite, similarly as in the \( \gamma = 1 \) case, \( \kappa \) in terms of \( z \), and one gets

\[
F(z, \varepsilon) = \kappa(z)
\]

It is proved next that \( F \) has no zeros on \((\pi_+, 1/\varepsilon)\).

Since \( F(\pi_+) = 0 \), polynomial division by \((z - \pi_+)\) yields (B.27), where the third order polynomial \( g \) has derivative

\[
g' = a_0 + a_1 z + a_2 z^2,
\]

where the coefficients \( a_0, a_1 \) and \( a_2 \) are complicated, yet explicit, functions of the parameters and the relative bid-ask spread \( \varepsilon \).

In view of (B.27), it is enough to show that \( g \) has no zeros on \([\pi_+, 1/\varepsilon]\). First, note the following asymptotic expansions,

\[
g(\pi_+) = \left( \frac{3}{4\gamma} \theta^2_0(\theta_0 - 1)^2 \right)^{1/3} \varepsilon^{1/3} + O(\varepsilon^{2/3}), 
\]

\[
g(1/\varepsilon) = \frac{\sigma^2}{2\varepsilon} + O(1).
\]
Therefore, for sufficiently small $\varepsilon$, $g > 0$ on both endpoints of $[\pi+, 1/\varepsilon]$. It remains to show that any local minimum of $g$ in $[\pi+, 1/\varepsilon]$ is non-negative. In searching for local extrema, one obtains complex numbers $z_\pm$ where $g'(z_\pm) = 0$. The asymptotic expansions of $z_\pm$ are

$$z_\pm = \frac{2}{3\varepsilon} \pm \frac{1}{3\varepsilon} \sqrt{\frac{\gamma - 4}{\gamma - 1}} + O(1).$$

Obviously, there are no local extrema in $[\pi+, 1/\varepsilon]$ whenever $\gamma \in [1, 4)$. Therefore $g > 0$ on all of $[\pi+, 1/\varepsilon]$, and thus $F(z) \geq 0$ on $[\pi+, 1/\varepsilon]$. The non-trivial case $\gamma \notin [1, 4)$ remains:

For $0 < \gamma < 1$ it holds that $\frac{2\gamma - 1}{\gamma - 1} > 2$, hence $z_\pm \notin [\pi+, 1/\varepsilon]$. It follows that $g'$ has no zeros in this interval and thus $g > 0$ on $[\pi+, 1/\varepsilon]$.

Next, consider $\gamma \geq 4$: The local minimum $z_-$ of a third order polynomial with negative leading coefficient satisfies $z_- < z_+$ and $g(z_-) < g(z_+)$. In view of (B.28) and (B.29), it remains to show $g(z_-) > 0$. It holds that

$$g(z_-) = \frac{-\sqrt{\gamma^2 - 2\gamma + 4} + 2\gamma - 2}{2\sqrt{\gamma^2 - 2\gamma + 4} + 2\gamma + 2} \frac{\sigma^2}{27(\gamma - 1)} + O(1)$$

$$= \frac{3\gamma + (\gamma - 4)(2 + \gamma + \sqrt{(\gamma - 4)(\gamma - 1)})}{27(\gamma - 1)} + O(1),$$

whence $g(z_-) > 0$ for sufficiently small $\varepsilon$. Hence $g > 0$ on $[\pi+, 1/\varepsilon]$ is shown.

Summarizing, $\kappa(\zeta) \geq 0$ on $\zeta \geq \zeta_+$, which proves that the HJB equation (B.21) holds.

$\square$

**Lemma B.5.** Let $\eta_- < \eta_+$ be such that either $\eta_+ < -1/(1-\varepsilon)$ or $\eta_- > 0$. Then there exists an admissible trading strategy $\hat{\varphi}$ such that the risky-safe ratio $\eta_t$ satisfies SDE (A.3). Moreover, $(\eta_t, \hat{\varphi}_t^1, \hat{\varphi}_t^2)$ is a reflected diffusion on the interval $[\eta_-, \eta_+]$. In particular, $\eta_t$ has stationary density equals

$$\nu(\eta) := \frac{2\mu - 1}{\eta_+^2 - 1} \eta_+^{2\sigma^2 - 2}, \quad \eta \in [\eta_-, \eta_+],$$

when $\eta_- > 0$, and otherwise equals

$$\nu(\eta) := \frac{2\mu - 1}{|\eta_-|^2 - 1} |\eta_-|^2 + 2 - 2, \quad \eta \in [\eta_-, \eta_+].$$

**Proof.** By the solution of the Skorohod problem for two reflecting boundaries Kruk et al. (2007), there exists a well-defined reflected diffusion $(\eta_t, L_t, U_t)$ satisfying

$$\frac{d\eta_t}{\eta_t} = \mu dt + \sigma dB_t + dL_t - dU_t,$$

where $W$ is a standard Brownian motion, and $L$ (resp. $U$) is a non-decreasing processes which increases only on the set $\{\eta = \eta_+\}$ (resp. $\{\eta = \eta_-\}$). Also, $\eta_- > 0$ or $\eta_+ < -1/(1-\varepsilon)$ implies that $\eta_t > 0$ or $\eta_t < -1/(1-\varepsilon)$ for all $t$, almost surely. Hence for each $t > 0$,

$$(1 + (1-\varepsilon)\eta_t), \quad (1 + \eta_t)$$
are invertible, almost surely. Define the increasing processes \((\hat{\varphi}^\uparrow), \hat{\varphi}^\downarrow\) by
\[
\frac{d\hat{\varphi}^\uparrow_t}{\hat{\varphi}_t} = (1 + \eta_t)^{-1}dL_t
\]
and
\[
\frac{d\hat{\varphi}^\downarrow_t}{\hat{\varphi}_t} = (1 + (1 - \varepsilon)\eta_t)^{-1}dU_t.
\]
By construction, the associated measures \(d\hat{\varphi}^\uparrow, d\hat{\varphi}^\downarrow\) are supported on \(\eta_t = \eta_-\) and \(\eta_t = \eta_+\), respectively. Hence \(\hat{\varphi}\) is a trading strategy, which by Lemma A.3 yields a risky/safe satisfying precisely the stochastic differential equation (A.3).

The admissibility of the trading strategy is clear, as \(\hat{\varphi}\) is a continuous, finite variation trading strategy, and since it satisfies \(\pi^\ast < 1/\varepsilon\), which implies that there exists \(\varepsilon' > \varepsilon\) such that \(\pi_t < 1/\varepsilon'\), for all \(t > 0\), a.s.. Finally, the form of the stationary density \(\nu(\eta)\), follows from the stationary Focker-Planck equation: The infinitesimal generator of \(\zeta_t\) is
\[
Af(\zeta) = \frac{\sigma^2}{2} \zeta^2 f''(\zeta) + \mu \zeta f'(\zeta) =: a(\zeta)/2f''(\zeta) + b(\zeta)\zeta'(\zeta).
\]
The invariant density \(\nu\) solves the adjoint differential equation
\[
A^\ast \nu(\eta) = (a(\eta)\nu(\eta))' - 2b(\eta)\nu(\eta) = 0
\]
and therefore equals
\[
\nu(\eta) = \frac{c}{a(\eta)} \exp \left( \int \frac{2b(\eta)}{a(\eta)} d\eta \right), \quad \text{(B.32)}
\]
where the constant \(c > 0\) depends on the boundaries \(\zeta_-, \zeta_+\). By integration, and distinguishing the cases \(\eta_+ < 0\) or \(\eta_- > 0\), the explicity expressions (B.30) and (B.31) are obtained.

The following constitutes the verification of optimality of the trading strategy of Lemma B.5 with the trading boundaries in Proposition B.1:

**Proposition B.6.** Let \(\zeta_\pm\) be the free boundaries as derived in Proposition B.1, and denote by \(\hat{\varphi}\) the trading strategy of Lemma B.5 associated with these free boundaries. Set
\[
\pi_\pm := \zeta_\pm / (1 + \zeta_\pm).
\]
Then for all \(t > 0\), the fraction of wealth \(\pi_t\) invested in the risky asset lies in the interval \([\pi_-, \pi_+]\), almost surely, entails no trading whenever \(\pi \in (\pi_-, \pi_+)\) (the no-trade region) and engages in trading only at the boundaries \(\pi_\pm\). For sufficiently small \(\varepsilon\), \(\hat{\varphi}\) is optimal, and the value function is
\[
F_\infty(\hat{\varphi}) = r + \max_{\varphi \in \Phi} \lim_{T \to \infty} \frac{1}{T} \mathbb{E} \left[ \int_0^T \left( (\mu + \gamma \sigma^2 \Lambda)\pi_t - \frac{\gamma \sigma^2}{2} \pi_t^2 \right) dt - \varepsilon \int_0^T \pi_t^2 \frac{d\varphi^\downarrow_t}{\varphi_t} \right] = r + (\mu + \gamma \sigma^2 \Lambda)\pi_- - \frac{\gamma \sigma^2}{2} \pi_-^2. \quad \text{(B.33)}
\]
Proof of Proposition B.6. Recall from Proposition B.4 that $\lambda = h(\zeta)$ and $(V, \lambda)$, defined from the unique solution of the free boundary problem, is a solution of the HJB equation (B.21). For the verification, the proportion $\pi_t$ of wealth in the risky asset is used, instead of the risky/safe ratio $\zeta_t$. The change of variables

$$\zeta = -1 + \frac{1}{1 - \pi}$$

amounts to a compactification of the real line, such that the two intervals $[-\infty, -1/(1 - \varepsilon))$ and $(0, \infty]$ are mapped onto the connected interval $[0, 1/\varepsilon)$. Denote by $L$ the differential operator

$$(L f)(\pi) := \frac{\sigma^2}{2} f''(\pi) \pi^2 (1 - \pi)^2 + f'(\pi)(\mu - \sigma^2 \pi) \pi (1 - \pi).$$

The function $\hat{V}(\pi) := V(\zeta(\pi))$ satisfies the HJB equation

$$\min(\mathcal{L}(\hat{V}(\pi) - \hat{h}(\pi) + \lambda, \hat{V}'(\pi), \hat{V}'(\pi) - \frac{\varepsilon}{1 - \varepsilon \pi}) = 0,$$

for $0 \leq \pi < 1/\varepsilon$, where $\hat{h}(\pi) = h(\zeta(\pi)) = (\mu + \gamma \sigma^2 \Lambda) \pi - \frac{\gamma \sigma^2}{2} \pi^2$.

First, it is shown that $F_{\infty}(\varphi) \leq \lambda + r$, for any admissible trading strategy $\varphi$. By Lemma A.2 one may without loss of generality assume $\pi_t \geq 0$, almost surely, for all $t \geq 0$. An application of Itô’s formula to the stochastic process $\hat{V}(\pi_t)$, where $\hat{V}$ is the solution of the HJB equation (B.34), yields

$$\hat{V}(\pi_T) - \hat{V}(\pi_0) = \int_0^T \hat{V}'(\pi_t)d\pi_t + \frac{1}{2} \hat{V}''(\pi_t)d\langle \pi \rangle_t$$

$$= \int_0^T \left(\mathcal{L}(\hat{V}(\pi) - \hat{h}(\pi) + \lambda) \right) dt + \int_0^T (\hat{h}(\pi_t) - \lambda) dt$$

$$+ \int_0^T \hat{V}'(\pi_t)\pi_t(1 - \pi_t)\sigma dB_t$$

$$- \int_0^T \hat{V}'(\pi_t)(1 - \varepsilon \pi_t)\pi_t d\varphi_t^l$$

$$+ \int_0^T \hat{V}'(\pi_t)\pi_t d\varphi_t^u.$$  

(B.35)

(B.36)

(B.37)

(B.38)

(B.39)

The first term in line (B.36) is non-negative, due to (B.34). Furthermore, (A.1) implies the existence of $\varepsilon' > \varepsilon$ such that $\pi_t < 1/\varepsilon' < 1/\varepsilon$, for all $t$, a.s.. Using (B.34) one thus obtains

$$\hat{V}'(\pi_t) \leq \frac{\varepsilon\varepsilon'}{\varepsilon' - \varepsilon}, \quad \text{a.s. for all } t \geq 0.$$  

(B.40)

Hence (B.37) is a martingale with zero expectation. Again by (B.34) one has that

$$\hat{V}'(\pi_t)\pi_t(1 - \varepsilon \pi_t) \leq \varepsilon \pi_t,$$

which implies that for (B.38) one has

$$- \int_0^T \hat{V}'(\pi_t)(1 - \varepsilon \pi_t)\pi_t d\varphi_t^l \geq -\varepsilon \int_0^T \pi_t d\varphi_t^l.$$  

(B.41)
Finally, (B.39) is non-negative, because $\hat{V}' \geq 0$ due to (B.34). Thus, taking the expectation of (B.35) the estimate,

$$\frac{1}{T} \mathbb{E} [\hat{V}(\pi_T) - \hat{V}(\pi_0)] \geq -\lambda + \frac{1}{T} \mathbb{E} \left[ \int_0^T \hat{h}(\pi_t) dt \right] - \varepsilon \frac{1}{T} \int_0^T \pi_t \frac{d\varphi^\perp_t}{\varphi_t}$$  \hspace{1cm} (B.41)

follows. By (B.40)

$$|\hat{V}(\pi_t) - \hat{V}(\pi_0)| \leq |\pi_T - \pi_0| \sup_{0 < u \leq 1/c'} |\hat{V}'(u)| \leq \frac{\varepsilon}{c' - \varepsilon},$$

therefore

$$\lim_{T \to \infty} \frac{1}{T} \mathbb{E} [\hat{V}(\pi_T) - \hat{V}(\pi_0)] = 0.$$  \hspace{1cm} (B.42)

Finally, it is shown that the bound $\lambda + r$ is attained by the admissible trading strategy $\hat{\varphi}$ defined by Lemma (B.5) in terms of the free boundaries $(\zeta_-, \zeta_+)$. Let $\zeta_t$ be the corresponding risky/safe ratio. Using Itô's formula, one has

$$dV(\zeta_t) = V'(\zeta_t) \zeta_t \sigma dB_t + 0 - \varepsilon \pi_t \frac{d\varphi^\perp_t}{\varphi_t} + (h(\zeta_t) - \lambda) dt.$$  \hspace{1cm} (C.1)

Integration with respect to $t$ and division by $T$ yields, in view of (A.7),

$$\frac{1}{T} \mathbb{E} \left[ \int_0^T \left( (\mu + \gamma \sigma^2 \Lambda) \pi_t - \frac{\gamma}{2} \sigma^2 \pi_t^2 \right) dt - \varepsilon \int_0^T \pi_t \frac{d\varphi^\perp_t}{\varphi_t} \right] = \lambda + \frac{1}{T} \mathbb{E} [\hat{V}(\pi_T) - \hat{V}(\pi_0)].$$

Letting $T \to \infty$, one obtains $F_\infty(\hat{\varphi}) = \lambda + r$. Due to (B.42), $\hat{\varphi}$ is an optimal trading strategy.

\[\Box\]

**B.1 Proof of Theorem 3.1 (i)–(iii)**

Theorem 3.1 (i) is proved in Proposition B.1, and Theorem 3.1 (ii) & (iii) are proved in Proposition B.6.

**C Ergodic results**

In this section, ergodicity is utilized to derive closed-form expressions for average trading costs (ATC) and long-run mean and long-run variance of the optimal trading strategy. These formulas are then used to prove the asymptotic expansions of Theorem 3.1 (iv).

Let $\zeta_-, \zeta_+$ be the free boundaries obtained in Proposition B.1. Without loss of generality, assume that either $\zeta_- < \zeta_+ < -1$ (leveraged case) or $\zeta_- > \zeta_+ > 0$ throughout (non-leveraged case), and define the integral

$$I := \frac{1}{c} \int_{\zeta_-}^{\zeta_+} h(\zeta) |\zeta|^{2\gamma \pi_s - 2} d\zeta,$$

where the normalizing constant is

$$c := \int_{\zeta_-}^{\zeta_+} |\zeta|^{2\gamma \pi_s - 2} d\zeta = \text{sgn}(\zeta_-) \frac{|\zeta_+|^{2\gamma \pi_s - 1} - |\zeta_-|^{2\gamma \pi_s - 1}}{2\gamma \pi_s - 1}.$$  \hspace{1cm} (C.2)
The objective functional

Lemma C.1.

\[ I = h(\zeta_{-}) + \frac{\sigma^2(2\gamma\pi_{s} - 1)}{2} \left( \frac{G(\zeta_{+})\zeta_{+}}{1 - \left( \frac{\zeta_{-}}{\zeta_{+}} \right)^{2\gamma\pi_{s} - 1}} \right). \quad (C.3) \]

Proof. From equations (B.8) and (B.10) it follows that

\[ \int_{\zeta_{-}}^{\zeta_{+}} h(\zeta) |\zeta|^{2\gamma\pi_{s} - 2} d\zeta = h(\zeta_{-}) \text{sgn}(\zeta_{-}) \frac{|\zeta_{+}|^{2\gamma\pi_{s} - 1} - |\zeta_{-}|^{2\gamma\pi_{s} - 1}}{2\gamma\pi_{s} - 1} + \frac{\sigma^2\zeta_{+}^{2\gamma\pi_{s}}}{2} G(\zeta_{+}). \]

By normalizing, (C.3) follows. \qed

Let now \( \zeta \) be a geometric Brownian motion with parameters \((\mu, \sigma)\), reflected at \( \zeta_{-}, \zeta_{+} \) respectively, as in Lemma B.5. Recall the following ergodic result (Gerhold et al., 2014, Lemma C.1):

Lemma C.2. Let \( \eta_t \) be a diffusion on an interval \([l, u] \), \( 0 < l < u \), reflected at the boundaries, i.e. \[ d\eta_t = b(\eta_t) dt + a(\eta_t)^{1/2} dB_t + dL_t - dU_t, \]

where the mappings \( a(\eta) > 0 \) and \( b(\eta) \) are both continuous, and the continuous, non-decreasing processes \( L_t \) and \( U_t \) satisfy \( L_0 = U_0 = 0 \) and increase only on \( \{L_t = l\} \) and \( \{U_t = u\} \), respectively. Denoting by \( \nu(\eta) \) the invariant density of \( \eta_t \), the following almost sure limits hold:

\[ \lim_{T \to \infty} \frac{L_T}{T} = \frac{a(l)\nu(l)}{2}, \quad \lim_{T \to \infty} \frac{U_T}{T} = \frac{a(u)\nu(u)}{2}. \]

The next formula evaluates trading costs.

Lemma C.3. The average trading costs for the optimal trading policy are

\[ \text{ATC} := \varepsilon \lim_{T \to \infty} \frac{1}{T} \int_0^T \frac{d\varphi_t^\perp}{\varphi_t} = \frac{\sigma^2(2\gamma\pi_{s} - 1)}{2} \left( \frac{G(\zeta_{+})\zeta_{+}}{1 - \left( \frac{\zeta_{-}}{\zeta_{+}} \right)^{2\gamma\pi_{s} - 1}} \right). \quad (C.4) \]

Proof. Note that

\[ \varepsilon \int_0^T \frac{d\varphi_t^\perp}{\varphi_t} = G(\zeta_{+}) \frac{U_T}{T}. \]

Applying Lemma C.2 to \( \eta := \zeta \) and \( u = \zeta_{+} \), and using the stationary density of \( \zeta_t \) (Lemma B.5) which equals

\[ \nu(\zeta) := \text{sgn}(\zeta_{-}) \frac{2\gamma\pi_{s} - 1}{|\zeta_{+}|^{2\gamma\pi_{s} - 1} - |\zeta_{-}|^{2\gamma\pi_{s} - 1}} |\zeta|^{2\gamma\pi_{s} - 2}, \quad \zeta \in [\zeta_{-}, \zeta_{+}], \]

(C.4) is obtained. \qed
Remark C.4. An alternative proof provides a consistency check for the theory provided so far: By Lemma 2.1 one can rewrite the objective functional as

$$F_\infty(\varphi) = r + \lim_{T \to \infty} \frac{1}{T} \int_0^T h(\zeta_t)dt - \text{ATC}.$$ 

Now by the ergodic theorem (Borodin and Salminen, 2002, II.35 and II.36),

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T h(\zeta_t)dt = I,$$

hence using Lemma C.1 it follows that

$$F_\infty(\varphi) = r + h(\zeta_-) + \text{ATC} - \text{ATC} = r + h(\zeta_-)$$

which is in agreement with the formula in Proposition B.6.

C.1 Long-run mean and variance

Define

$$I_\mu := \int_{\zeta_-}^{\zeta_+} \frac{\zeta}{1 + \zeta} |\zeta|^{2\gamma \pi_*} d\zeta, \quad I_{s^2} := \int_{\zeta_-}^{\zeta_+} \left( \frac{\zeta}{1 + \zeta} \right)^2 |\zeta|^{2\gamma \pi_*} d\zeta.$$ 

Since the long-run mean and long-run variance are

$$\hat{m} := \lim_{T \to \infty} \frac{1}{T} \mathbb{E}[R_T] = r + \mu \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[ \int_0^T \pi_t dt \right] - \text{ATC}$$

$$\hat{s}^2 := \lim_{T \to \infty} \frac{1}{T} \mathbb{E}[(R)_T] = \sigma^2 \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[ \int_0^T \pi_t^2 dt \right]$$

$$= \frac{\sigma^2}{c} I_{s^2},$$

the following decomposition holds in view of the ergodic theorem (Borodin and Salminen, 2002, II.35 and II.36):

$$I = \frac{1}{c} \left( (\mu + \gamma \sigma^2 \Lambda) I_\mu - \frac{\gamma \sigma^2}{2} I_{s^2} \right) = \frac{\theta_*}{\pi_*} (\hat{m} - r + \text{ATC}) - \frac{\gamma}{2} \hat{s}^2$$

(C.5)

$$= h(\zeta_-) + \text{ATC}.$$ 

Integration by parts yields

$$I_\mu = \int_{\zeta_-}^{\zeta_+} \frac{\zeta}{1 + \zeta} |\zeta|^{2\gamma \pi_*} d\zeta = \frac{|\zeta_+|^{2\gamma \pi_*}}{2\gamma \pi_* (1 + \zeta_+)} - \frac{|\zeta_-|^{2\gamma \pi_*}}{2\gamma \pi_* (1 + \zeta_-)} + \frac{I_{s^2}}{2\gamma \pi_*}. \quad \text{(C.6)}$$

An application of this identity to (C.5) yields

$$I = \frac{\sigma^2}{2c} \frac{\theta_*}{\pi_*} \left( \frac{|\zeta_+|^{2\gamma \pi_*}}{1 + \zeta_+} - \frac{|\zeta_-|^{2\gamma \pi_*}}{1 + \zeta_-} + (1 - \frac{\gamma \pi_*}{\theta_*}) I_{s^2} \right).$$

Except for the singular case $\gamma = \theta_*/\pi_*$, one can extract $I_{s^2}$, and thus (C.6) and (C.4) yield a formula for $\hat{s}^2$. Therefore, the right side of equation (C.5) gives a formula for $\hat{m}$ in terms of $\hat{s}$.
Lemma C.5. When $\gamma \neq \theta_*/\pi_*$, the following identities hold:

$$\hat{s}^2 = \frac{2\pi_*}{\theta_* - \gamma\pi_*} (h(\zeta) + \text{ATC}) - \frac{\sigma^2}{c} \frac{\theta_*}{\theta_* - \gamma\pi_*} \left( \frac{|\zeta_+|^{2\gamma\pi_*}}{1 + \zeta_+} - \frac{|\zeta_-|^{2\gamma\pi_*}}{1 + \zeta_-} \right), \quad (C.7)$$

$$\hat{m} = r + \frac{\pi_*}{\theta_*} \left( \frac{\gamma}{2} \hat{s}^2 + h(\zeta) \right) - \frac{\Lambda}{\theta_*} \text{ATC}. \quad (C.8)$$

C.2 Proof of Theorem 3.1 (iv)

Proof. The asymptotic expansion (3.7) for the trading boundaries $\pi_{\pm}$ can be derived by developing $\frac{\zeta_{\pm}}{1 + \zeta_{\pm}}$ into a power series, thereby using the asymptotic expansions (B.7) of $\zeta_{\pm}$.

Long-run mean $\hat{m}$ and long-run variance $\hat{s}^2$, as well as average trading costs ATC and the value function $\lambda$ have closed form expressions in terms of the free boundaries $\zeta_{\pm}$ (see equations (C.8), (C.7), and equations (C.4) and (B.33)). Using these formulas in combination with the asymptotic expansions (B.7) of the free boundaries, the assertion follows. $\Box$

D Proof of Theorem 3.2

Parts (i) and (v) are special cases of Theorem 3.1 (iii) and (iv). The explicit formulas for ATC, $\hat{m}$ and $\hat{s}$ are derived from Lemma C.3 and Lemma C.5.

It thus remains to derive the formulas for $\bar{\alpha}$, $\bar{\beta}$, $R^2$ and Tracking Error TrE.

For a fixed time horizon, the regression

$$\frac{1}{T} \int_0^T \frac{dw_t}{w_t} - r = \bar{\alpha}_T + \bar{\beta}_T \left( \frac{1}{T} \int_0^T \frac{dw_t}{w_t} - r \right) \quad (D.1)$$

leads to the estimated slope, or beta, of this regression with the continuous-time approximation

$$\bar{\beta}_T \approx \frac{\left( \int_0^T \frac{dw}{w}, \int_0^T \frac{dS}{S} \right)_T}{\left( \int_0^T \frac{dS}{S} \right)_T} = \frac{\int_0^T \pi_t \sigma^2 dt}{\sigma^2 T} = \frac{1}{T} \int_0^T \pi_t dt. \quad (D.2)$$

As a result, the $\bar{\beta}_T$ converges to

$$\bar{\beta} = \lim_{T \to \infty} \frac{1}{T} \int_0^T \pi_t dt.$$  

According to Section C, ergodicity of the risky/safe ratio $\zeta_t$ may be invoked to obtain

$$\bar{\beta} = \frac{1}{c} \int_{\zeta_-}^{\zeta_+} \frac{u}{1 + u} u^{-2} du,$$

with the normalizing constant, due to (C.2), equals

$$c = (-\zeta_+)^{-1} - (\zeta_-)^{-1} = \frac{1}{\zeta_-} - \frac{1}{\zeta_+} > 0.$$  

Direct integration gives formula (3.18). By taking expectations in (D.1) and letting $T \to \infty$ equation (3.17) for $\bar{\alpha}$ is obtained.
Similarly, the long horizon \( R^2 \) of the regression, defined as the ratio between the variance of the predicted return and the variance of the realized return, is

\[
R^2 = \lim_{T \to \infty} \frac{\left( \frac{1}{T} \int_0^T \pi_t dt \right)^2}{\frac{1}{T} \int_0^T \pi_t^2 dt}.
\]  

Using

\[
\lim_{T \to \infty} \frac{1}{T} \int_0^T \pi_t^2 dt = \frac{1}{c} \int_{\zeta_-}^{\zeta_+} \left( \frac{u}{1 + u} \right)^2 u^{-2} du = -\frac{1}{c} \frac{1}{1 + \frac{1}{\zeta_-} - 1/\zeta_+} \left( \frac{1}{1 + \zeta_-} - \frac{1}{1 + \zeta_+} \right),
\]

and equation (3.18) yields equation (3.19). The tracking error (3.20) can be obtained quite similarly, because

\[
\text{TrE} = \sqrt{\lim_{T \to \infty} \left( \frac{1}{T} \int_0^T dw_t - \frac{\Lambda}{T} \int_0^T dS_t \right)} = \sigma \sqrt{\left( \lim_{T \to \infty} \frac{1}{T} \int_0^T \pi_t^2 dt - 2\Lambda \lim_{T \to \infty} \frac{1}{T} \int_0^T \pi_t dt + \Lambda^2 \right)}.
\]  

(E.4)

E Proof of Theorem 3.3

In this section the free boundary problem (3.1)–(3.5) for \( \gamma = 0 \) is solved for sufficiently small \( \varepsilon \), it is shown that the solution \((W, \zeta_-, \zeta_+)\) allows to construct a solution of the corresponding HJB equation, and similarly to the case \( \gamma > 0 \), a verification argument reveals an optimal trading strategy.

Numerical experiments using \( \gamma > 0 \) indicate that the trading boundaries \( \pi_\pm \) (hence the leverage multiplier) satisfy

\[
\lim_{\varepsilon \downarrow 0} \varepsilon^{1/2} \pi_\pm = 1/A_\pm
\]

for two constants \( A_- > A_+ > 0 \). This entails that the free boundaries obey the approximation

\[
\zeta_\pm \approx -1 - A_\pm \varepsilon^{1/2}
\]

for sufficiently small \( \varepsilon \). This insight lets us conjecture that \( \zeta_\pm \) are analytic in \( \delta := \varepsilon^{1/2} \).

The system (B.10)–(B.11) can be rewritten by using the new parameter \( \delta := \varepsilon^{1/2} \) and by multiplying the second equation by \( \delta \):

\[
W(\zeta_-, \zeta_+) - \frac{\delta^2}{(1 + \zeta_-)(1 + (1 - \delta^2)\zeta_+)} = 0,
\]

\[
\delta \left( \frac{2(h(\zeta_+) - h(\zeta_-))}{\sigma^2 \zeta_+^2} - \frac{2\mu/\sigma^2}{\zeta_+} W(\zeta_-, \zeta_+) - \frac{(1 - \delta^2)^2}{(1 + (1 - \delta^2)\zeta_+)^2} + \frac{1}{(1 + \zeta_+)^2} \right) = 0.
\]  

(E.1)

(E.2)

Using the transformation \( u = -\frac{1 - \xi}{\delta} \) and noting that \( |\xi| = 1 + \delta u \), one obtains

\[
\Xi(u_-, u) := W(-1 - u_- \delta, -1 - u\delta) = \frac{2\mu}{\sigma^2 (1 + u\delta)^2} \int_{u_-}^u \left( \frac{1}{u_- - \xi} - \frac{1}{1 + u\delta} \right) \left( \frac{1 + \xi\delta}{1 + u\delta} \right)^{2\mu/\sigma^2 - 2} d\xi.
\]

38
Accordingly, the system (E.1)–(E.2) transforms into

\[ \Xi(u_-, u_+) - \frac{1}{u_+((1 - \delta^2)u_+ - \delta)} = 0, \]  

(E.3)

\[ \frac{2\mu}{\sigma^2} \left( \frac{1}{u_+} - \frac{1}{u_-} + \frac{\delta}{1 + u_+ \delta} \Xi(u_-, u_+) \right) - \frac{2(1 - \delta^2)u_+ - \delta}{u_+^2(\delta + (\delta^2 - 1)u_+)^2} = 0. \]  

(E.4)

Letting \( \delta \to 0 \) in (E.3)–(E.4), one obtains an equation for, say \((A_-, A_+)\),

\[ 2\mu \sigma^2 \left( \log(\frac{A_-}{A_+}) - \frac{A_- - A_+}{A_-} \right) - \frac{1}{A_+^3} = 0, \]  

(E.5)

\[ \frac{\mu}{\sigma^2} \left( \frac{1}{A_+} - \frac{1}{A_-} \right) - \frac{1}{A_-^3} = 0. \]  

(E.6)

**Lemma E.1.** The unique solution \((A_-, A_+)\) of the system (E.5)–(E.6) is

\[ A_- = \kappa^{-1/2} \sqrt{\frac{\sigma^2}{\mu}}, \quad A_+ = \kappa^{-1/2} \sqrt{\frac{\sigma^2}{\mu}}, \]  

(E.7)

where \( \kappa \approx 0.5828 \) is the unique solution of

\[ f(\xi) := \frac{3}{2} \xi + \log(1 - \xi) = 0, \quad \xi \in (0, 1). \]  

(E.8)

**Proof.** Equation (E.6) gives

\[ A_- = \frac{\mu A_+^3}{\mu A_+^3 - \sigma^2}. \]  

(E.9)

Hence substituting (E.9) into (E.5) gives the well-posed transcendental equation

\[ -3A_+^2 + \frac{2\mu \log \left( \frac{\mu A_+^2}{A_+^2 - \sigma^2} \right)}{\sigma^2} = 0, \quad A_+ > 0. \]  

(E.10)

Therefore it is enough to establish that the unique solution of (E.10) is as in the second equation in line (E.7); the formula for \( A_- \) then follows from (E.9). To this end, substitute

\[ \xi := \frac{\sigma^2}{\mu A_+^2} \]

into (E.10) to obtain equation (E.8). Note that \( f(0) = 0, f' > 0 \) on \((0, 1/3)\) and \( f' < 0 \) on \((1/3, 1)\), while \( f(\xi) \downarrow -\infty \) as \( \xi \to 1 \). This implies that \( f \) has a single zero \( \kappa \) on \((1/3, 1)\) and thus the claim concerning \( A_+ \) is proved.

**Proposition E.2.** For sufficiently small \( \delta \), there exists a unique solution \((u_+, u_-)\) of (E.3)–(E.4) near \((A_-, A_+)\). This solution is analytic in \( \delta \) and satisfies the asymptotic expansion \( u_\pm = A_\pm + O(\delta) \), where \( A_\pm \) are in (E.7).
Proof. Denote the left sides of (E.3)–(E.4), by \( F_i((u_-,u_+),\delta), i = 1,2 \) and \( F = (F_1,F_2) \). By Lemma E.1, \( F((A_-,A_+),0) = 0 \). Since
\[
\frac{\partial \Xi}{\partial u_-}(A_-,A_+,0) = \frac{2\mu}{\sigma^2} \left( \frac{A_- - A_+}{A_-^2} \right), \quad \frac{\partial \Xi}{\partial u_+}(A_-,A_+,0) = \frac{2\mu}{\sigma^2} \left( \frac{A_+ - A_-}{A_- A_+} \right), 
\]
one obtains
\[
\frac{\partial F_1}{\partial u_-}(A_-,A_+,0) = \frac{2\mu}{\sigma^2} \left( \frac{A_- - A_+}{A_-^2} \right), \\
\frac{\partial F_1}{\partial u_+}(A_-,A_+,0) = \frac{2}{A_+^3} + \frac{2\mu}{\sigma^2} \left( \frac{A_+ - A_-}{A_- A_+} \right) = 0, \\
\frac{\partial F_2}{\partial u_+}(A_-,A_+,0) = \frac{6}{A_+^4} - \frac{2\mu}{\sigma^2} \left( \frac{1}{A_+^2} \right), 
\]
where the second line vanishes due to (E.6), and therefore, the Jacobian \( DF \) of \( F \) satisfies
\[
\det(DF)((A_-,A_+),0) = \frac{\partial F_1}{\partial u_-}((A_-,A_+),0) \times \frac{\partial F_2}{\partial u_+}((A_-,A_+),0) \\
= -4(\mu/\sigma^2)^{7/2}(\kappa - 1)^{5/2}(3\kappa - 1) \neq 0, 
\]
because \( \kappa \in (1/3,1) \). Hence by the implicit function theorem for analytic functions (Gunning and Rossi, 2009, Theorem I.B.4) the assertion follows. \( \square \)

**Lemma E.3.** Let \( \kappa \) be the solution of (E.8), and \( \theta \in [0,1] \). If
\[
f(\theta) = \log(1 - \kappa(1-\theta)) + (1-\theta)\kappa + \frac{1}{2} \kappa(1-\kappa)^2 = 0 \tag{E.11}
\]
then \( \theta = 0 \).

Proof. Clearly \( f(0) = 0 \) and also \( f(1) = 1/2\kappa(1-\kappa)^2 > 0 \). There is a single local extremum of \( f \), in \((0,1)\), namely,
\[
\theta_1 = \frac{0.5}{\kappa^2} \left( 3\kappa^2 + \sqrt{4\kappa^4 - 3\kappa^4 - 2\kappa} \right) \approx 0.7669, 
\]
but since \( f'(0) = 0 \), and
\[
f''(0) = \frac{\kappa^2 \left( \kappa(3\kappa^2 - 7\kappa + 5) - 1 \right)}{(1-\kappa)^4} > 0 
\]
\( \theta_1 \) must be the global maximum. Hence \( f > 0 \) on \([0,1]\), whence \( \theta = 0 \), as claimed. \( \square \)

**Lemma E.4.** Let \( A_- \) be as in (E.7). The only solution of
\[
\frac{2\mu}{\sigma^2} \left( \log(A_-/\xi) - \frac{A_- - \xi}{A_-} \right) - \frac{1}{\xi^2} = 0 \tag{E.12}
\]
on \([A_+,A_-] \) is \( \xi = A_+ \).

Proof. Let \( \xi \) be a solution of (E.12). There exists \( \theta \in [0,1] \) such that
\[
\xi = \theta A_- + (1-\theta)A_+ = A_+ \left( \frac{1 + \kappa(1-\theta)}{1 - \kappa} \right). 
\]
Hence \( A_+^*/A_- = 1 + \kappa(\theta - 1) \), and therefore (E.12) can be rewritten as (E.11). An application of Lemma E.3 yields \( \xi = A_+ \). \( \square \)
E.1 Proof of Theorem 3.3

Proof. Arguing similarly as in the Proof of Proposition B.1 for the case $\gamma > 0$, the solvability of the free boundary problem (3.1)–(3.5) for $\gamma = 0$ is equivalent to solvability of the non-linear system (E.1)–(E.2). This, in turn, is equivalent to solving (E.3)–(E.4) for $(u_+ (\delta), u_- (\delta))$. A unique solutions of the transformed system (E.3)–(E.4) near $(A_+, A_-)$ is provided by Proposition E.2, and one has $\zeta_\pm = -1 - u_\pm \delta$. In particular, one obtains

$$\zeta_\pm = -1 - A_\pm \varepsilon^{1/2} + O(1). \tag{E.13}$$

The solution of (3.1)–(3.5) is

$$W(\zeta) := \frac{2\mu}{\sigma^2 |\zeta|^{\frac{\sigma^2}{2}}} \int_{\zeta_-}^{\zeta} \left( \frac{y}{1 + y} - \frac{\zeta_-}{1 + \zeta_-} \right) |y|^{2\mu/\sigma^2 - 2} dy. \tag{E.14}$$

One defines exactly as in (B.22) a candidate solution $(V, \lambda)$ of the HJB equation (B.21). Next it is shown that $(V, \lambda)$ solves the HJB equation (B.21) for the intervals $[\zeta_-, \zeta_+]$, $(-\infty, \zeta_-]$ and finally for $[\zeta_+, \infty)$. In fact, the interval $[-1/(1 + \varepsilon), 0)$ is excluded.

On $[\zeta_-, \zeta_+]$,

$$(AV(\zeta) - h(\zeta) + h(\zeta_+))' = \frac{1}{2} \sigma^2 \zeta^2 W''(\zeta) + (\sigma^2 + \mu) \zeta W'(\zeta) + \mu W(\zeta) - \frac{\mu}{(1 + \zeta)^2} = 0$$

by construction. Because of the initial conditions (3.2)–(3.3),

$$(AV(\zeta) - h(\zeta) + h(\zeta_-)) |_{\zeta=\zeta_-} = AV(\zeta) |_{\zeta=\zeta_-} = 0$$

and thus

$$AV(\zeta) - h(\zeta) + h(\zeta_-) \equiv 0, \quad \zeta \in [\zeta_-, \zeta_+].$$

Next it is shown that $0 \leq V' \leq G$ on all of $[\zeta_-, \zeta_+]$. Since

$$(h(\zeta) - h(\zeta_-))' = h'(\zeta) = \frac{\mu}{(1 + \zeta)^2} \tag{E.15}$$

is strictly positive, $h(\zeta) - h(\zeta_-) > 0$ for $\zeta \in (\zeta_-, \zeta_+].$ From the explicit formula (E.14) one therefore may conclude that $V' = W \geq 0$ for $\zeta \in [\zeta_-, \zeta_+]$. It remains to show $V' \leq G$. Since $V'(\zeta_+) - G(\zeta_+) = 0$, and since $V'(\zeta_-) - G(\zeta_-) = -G(\zeta_-) < 0$, it suffices to rule out any zero $\zeta^*_+$ of $V'(\zeta) - G(\zeta)$ on $(\zeta_-, \zeta_+)$, for sufficiently small $\varepsilon$. This is equivalent to ruling out any zeros of

$$\kappa(u, \delta) := V'(\zeta(u)) - G(\zeta(u)), \quad u \in (u_+ (\delta), u_- (\delta)),$$

where $\zeta(u) = -1 - u \delta$, for sufficiently small $\delta$. Recall that $u_\pm (\delta)$ is implicitly defined by $\zeta_\pm = -1 - u_\pm (\delta) \delta$, $\lim_{\delta \to 0} u_\pm (\delta) = A_\pm$. Assume, for a contradiction, there exists $\delta_k \downarrow 0$ and a sequence $u_+ (\delta_k)$ satisfying $u_-(\delta_k) < u_+^*(\delta_k) < u_+(\delta_k)$ which is a solution of $\kappa(u_+^*(\delta_k), \delta_k) = 0$ for each $k \in \mathbb{N}$.

By taking a subsequence, if necessary, one may without loss of generality assume $u_+^*(\delta_k) \to A_+^* \in [A_+, A_-]$ as $k \to \infty$. Suppose first that $A_+^* = A_+$ and define the map $\delta \mapsto u^*(\delta)$ by intertwining $u_+$ and $u_+^*$ as follows:

$$u_+^*(\delta) = \begin{cases} u_+^*(\delta_k), \quad k \in \mathbb{N} \\ u_+(\delta), \quad \delta \neq \delta_k \end{cases}.$$
Then for sufficiently small \( \delta \), the pair \((u_-(\delta), u_+^*(\delta))\) solves (E.3)–(E.4) near \((A_-, A_+)\), hence by Proposition E.2, \( u_+^* = u_+ \), a contradiction to our previous assumption \( \zeta_+^* \in (\zeta_-, \zeta_+) \). Second, consider the possibility \( A_+^* \in (A_+, A_-) \): By equation (E.3)

\[
\frac{2\mu}{\sigma^2} \left( \log \left( \frac{A_+/A_-^*}{A_-} \right) - \frac{A_- - A_-^*}{A_-^*} \right) - \frac{1}{(A_-^*)^2} = 0.
\]

Lemma E.4 states \( A_+^* = A_+ \), which is also impossible. Hence \( V'(\zeta) - G(\zeta) \) has no zeroes on \((\zeta_-, \zeta_+)\), and thus \( V \) solves the HJB equation on \([\zeta_-, \zeta_+]\).

Consider now \( \zeta \leq \zeta_- \). \( V \) solves the HJB equation, if

\[
A V - h(\zeta) + h(\zeta_-) = h(\zeta_-) - h(\zeta) \geq 0, \quad G(\zeta) \geq 0.
\]

The first inequality is clearly fulfilled. Also, since \( \zeta < -1/(1-\varepsilon) \) or \( \zeta > 0 \), \( G \) is a strictly positive function on \([-\infty, \zeta_-] \), which finishes the proof for \( \zeta \leq \zeta_- \).

Finally, consider \( \zeta \geq \zeta_+ \). Since \( G = W \), it suffices to show that

\[
L(\zeta) := A V - h(\zeta) + h(\zeta_-) \geq 0, \quad G(\zeta) \geq 0.
\]

(E.16)

The second inequality has just been proved. So only the first inequality in (E.16) needs to be established. Setting

\[
h_1(\zeta) = \mu \frac{\zeta}{1 + \zeta} - \frac{\sigma^2}{2} \left( \frac{\zeta}{1 + \zeta} \right)^2
\]

one can write

\[
L(\zeta) = \frac{\sigma^2 \zeta^2}{2} G'(\zeta) + \mu \zeta G(\zeta) - h(\zeta) + h(\zeta_-)
= h(\zeta_-) - h_1((1-\varepsilon)\zeta) - \frac{\sigma^2}{2} \left( \frac{\zeta}{1 + \zeta} \right)^2.
\]

Therefore, by the boundary conditions at \( \zeta_+ \),

\[
L(\zeta_+) = \frac{\sigma^2 \zeta^2}{2} W'(\zeta_+) + \mu \zeta W(\zeta_+) + h(\zeta_-) - h(\zeta_+) = 0.
\]

The last equality follows from our knowledge concerning the HJB equation on \([\zeta_-, \zeta_+]\).

To show that \( L(\zeta) \geq 0 \) for all \( \zeta \), it suffices to show that there are no solutions of the equation

\[
\kappa(\zeta) := h(\zeta_-) - h_1((1-\varepsilon)\zeta) - \frac{\sigma^2}{2} \left( \frac{\zeta}{1 + \zeta} \right)^2 = 0
\]

(E.17)

on \( \zeta \geq \zeta_+ \) except \( \zeta_+ \). The transformation \( z = \frac{\zeta}{1 + \zeta} \) yields

\[
\frac{(1-\varepsilon)\zeta}{1 + (1-\varepsilon)\zeta} = \frac{(1-\varepsilon)z}{1 - \varepsilon z}
\]

and thus one can rewrite (E.17) in terms of \( z \) and redefined as

\[
F(z, \varepsilon) := \mu \pi_- - \mu \left( \frac{(1-\varepsilon)z}{1 - \varepsilon z} \right)^2 + \frac{\sigma^2}{2} \left( \frac{(1-\varepsilon)z}{1 - \varepsilon z} \right)^2 - \frac{\sigma^2}{2} z^2.
\]
It is proved next that $F$ has no zeros on $(\pi_+, 1/\varepsilon)$: Since $F(\pi_+) = 0$, polynomial division by $(z - \pi_+)$ yields

$$F(z, \varepsilon) = \frac{(z - \pi_+)}{(1 - \varepsilon z)^2} g(z),$$  \hspace{1cm} (E.18)

where the third order polynomial $g$ has derivative

$$g' = a_0 + a_1 z + a_2 z^2,$$

with certain, relatively complicated but explicit, coefficients $a_0, a_1, a_2$. By the second formula of (E.7)

$$g(\pi_+) = -\mu + \frac{3\sigma^2}{A_+^2} + O(\varepsilon^{1/2})$$  \hspace{1cm} (E.19)

is strictly positive for sufficiently small $\varepsilon$, since $\kappa > 1/3$. The solutions $z_\pm$ of the equation

$$g'(z) = 0$$

are

$$z_- = -\frac{1}{2A_+ \varepsilon^{1/2}} + O(1), \quad z_+ = \frac{4}{3\varepsilon} + O(1).$$

The first one is negative for sufficiently small $\varepsilon$, hence irrelevant, and the second is larger than $1/\varepsilon$ for sufficiently small $\varepsilon$, hence also irrelevant. Since

$$g'(1/\varepsilon) = \frac{\sigma^2}{2} + O(\varepsilon^{1/2})$$

it follows that $g'(z) > 0$ on all of $[\pi_+, 1/\varepsilon]$. Together with (E.19) it follows that $g > 0$ on $[\pi_+, 1/\varepsilon]$. Hence $F(z) > 0$ for all $z > \pi_+$ which proves that $(V, \lambda)$ solves the HJB equation (B.21).

Using the proof of Proposition B.6, one can obtain assertion (ii) and (iii). Finally, the expansions of the trading boundaries claimed in (iv) follow from the asymptotic expansions of the free boundaries $\zeta_-, \zeta_+$ in (E.13).

\[\Box\]

### E.2 Proof of Proposition 4.1

**Proof of Proposition 4.1.** The formulas can be obtained by using the asymptotic expansions provided by Theorem 3.1 (iv), along the lines of the proof of Theorem 3.2 in Section D. To this end, use the first identities of each of the equations (3.17), (3.18), and also the identity (D.4).

\[\Box\]

### F Convergence

**Lemma F.1.** Let $\mu > \sigma^2$. There exists $\delta_0 > 0$ such that for all $0 \leq \gamma < \gamma_0 := \frac{\mu}{\sigma^2}$, $\delta \leq \delta_0$ the objective functional for a trading strategy $\varphi$ which only engages in buying at $\pi_- = 1 + \delta$ and selling at $\pi_+ = (1 - \delta)/\varepsilon > \pi_-$ outperforms a buy and hold strategy. More precisely, for all $\gamma < \gamma_0$, $\delta \leq \delta_0$

$$F_\infty(\varphi) \geq r + \mu(1 + \Lambda/\pi_*) - \frac{\gamma}{2} \sigma^2 + \left(\frac{\mu - \gamma \sigma^2}{2}\right) \delta > r + \mu(1 + \Lambda/\pi_*) - \frac{\gamma \sigma^2}{2}.$$
Proof. Using the stationary density $\rho(d\pi)$ of $\pi_t$ on $[\pi_-, \pi_+]$ (which can be derived from Lemma B.5), one obtains
\[
F_\infty(\varphi) = r + \int_{\pi_-}^{\pi_+} \left((\mu + \gamma\sigma^2\Lambda)\pi - \frac{\gamma\sigma^2}{2}\pi^2\right)\rho(d\pi) - \text{ATC}
\]
\[
\geq r + (\mu + \gamma\sigma^2\Lambda)(1 + \delta) - \frac{\gamma\sigma^2}{2}(1 + \delta)^2 - \frac{(\delta + 1)(2\epsilon - 1)^3(2\mu - \sigma^2)}{4\epsilon(\delta + 2(\delta + 1)\epsilon + \delta + 1)\left(\frac{\rho}{\sigma^2}\right)^2 + (\delta + 1)(2\epsilon - 1)}
\]
\[
\geq r + (\mu + \gamma\sigma^2\Lambda) - \frac{\gamma\sigma^2}{2} + (\mu + \gamma\sigma^2\Lambda - \gamma\sigma^2)\delta - O(\delta\min(2\frac{\rho}{\sigma^2} - 1)),
\]
where Lemma C.3 has been invoked to calculate and estimate the average trading costs ATC. The asymptotic expansion holds for sufficiently small $\delta$. Since $\mu > \gamma\sigma^2$, the claim follows. \qed

F.1 Proof of Theorem 4.2

Proof. By equation (4.9), the curves $(0, \gamma] \to \mathbb{R} : \gamma \mapsto \pi_\pm(\gamma)$ range in a relatively compact set, namely $[1, \frac{1}{\epsilon}]$. Consider therefore a sequence $\gamma_k, k = 1, 2, \ldots$ which satisfies
\[
1 \leq \pi_0^k := \lim_{i \to \infty} \pi_-(\gamma_k) \leq \lim_{i \to \infty} \pi_+(\gamma_k) =: \pi_+^k \leq 1/\epsilon.
\]
Set $c_k^\pm := \frac{\pi_\pm(\gamma_k)}{1 - \pi_\pm(\gamma_k)}$, for $k = 0, 1, 2, \ldots$ and note that $-\infty \leq c_0^k \leq 0 \leq -\frac{1}{1 - \epsilon}$.

For each $k, k = 1, 2, \ldots$, by assumption the HJB equation (B.21) is satisfied with $\lambda = \lambda_k := h(\zeta_k)$. Using the verification arguments of the proof of Proposition B.6 it follows that the trading strategies associated with the intervals $[\pi_-(\gamma_k), \pi_+(\gamma_k)]$ are optimal.

Next, three elementary facts are proved:

(i) $\pi_0^k > 1$, which is equivalent to $\zeta_0^k > -\infty$. Assume, for a contradiction, $\pi_0^k = 1$. Then $\pi_-(\gamma_k) \to 1$ and thus $\lambda_k \to \mu$, as $k \to \infty$. Hence, the objective functional eventually minorizes the uniform bound provided by Lemma F.1, a mere impossibility to optimality. Hence $\pi_0^k > 1$.

(ii) $\pi_0^k < \pi_+^k$: This holds due to the fact that, by observing limits for the initial and terminal conditions of zero order in (3.1),
\[
W(\zeta_0^k) = 0 < G(\zeta_0^k).
\]

(iii) Also, $\pi_0^k < \frac{1}{\epsilon}$. Assume, for a contradiction, that $\pi_0^k = \frac{1}{\epsilon}$. Then $G(\zeta_0^k) \to \infty$, as $k \to \infty$, and, since $\zeta_0^k < \pi_+^0$, the average trading costs corresponding to $\gamma_k$ satisfy (by Lemma C.3)
\[
\text{ATC}(k) := \frac{\sigma^2}{2} \left(\frac{2\mu}{\sigma^2} - 1\right) \frac{G(\zeta_k^\pm)\zeta_k^\pm}{1 - (\frac{\zeta_k^\pm}{\pi_+^0})^{2\mu/\sigma^2 - 1}} \to \infty,
\]
as $k \to \infty$. Denote by $\varphi^k$ the trading strategy which only buys (resp. sells) at $\pi_-(\gamma_k)$ (resp. $\pi_+(\gamma_k)$). By the results of Appendix C the value function satisfies for each $k$
\[
F_\infty(\varphi^k) = \int_{\pi_-^k(\gamma_k)}^{\pi_+(\gamma_k)} ((\mu + \gamma_k\sigma^2\Lambda)\pi - \frac{\gamma_k\sigma^2}{2}\pi^2)\rho(d\pi) - \text{ATC}(k) \leq \frac{\mu + \gamma_k\sigma^2\Lambda}{\epsilon} - \text{ATC}(k) \to -\infty
\]
as $k \to \infty$. In particular, for sufficiently large $k \geq k_0$, a buy-and-hold strategy $\varphi$ satisfies

$$F_\infty(\varphi) = \mu + \gamma_k \sigma^2 \Lambda - \frac{\gamma_k \sigma^2}{2} > F_\infty(\hat{\varphi}^k),$$

which contradicts the assumption concerning optimality of the trading strategy $[\pi_-(\gamma_k), \pi_+(\gamma_k)]$. Hence $\pi_0^+ < 1/\varepsilon$.

Since the sequence $\zeta^k$ converges, by (Keller-Ressel et al., 2010, Lemma 9) the solutions of the initial value problem associated with (3.1) and $\gamma_k$, namely $W(\zeta; \zeta^k)$, converge to the solution of the initial value problem (3.1) (for $\gamma = 0$),

$$W^0(\zeta) = -\frac{2}{\sigma^2 \zeta^2} \int_{\zeta^0}^\zeta (\mu - \frac{\zeta^0}{1 + \zeta^0} - \mu \frac{\zeta^0}{1 + \zeta^0}) (\zeta/\zeta^0)^{2\mu/\sigma^2 - 2} d\zeta.$$

The terminal conditions are met by $W^0$, because $G$ is continuous on $(-\infty, -1/\varepsilon)$. Also, for each $k, k = 1, 2, \ldots$, by assumption the HJB equation (B.21) is satisfied. Non-negativity is preserved by taking limits, hence, $(W(\zeta; 0), \lambda_0)$ satisfies the HJB equation as well. Using the verification arguments of the proof of Proposition B.6 it follows that the trading strategies associated with the intervals $[\pi_-(\gamma), \pi_+(\gamma)]$ are not only optimal for risk-aversion levels $\gamma \in [0, \bar{\gamma}]$, but also $[\pi_0^-, \pi_0^+]$ is optimal for a risk-neutral investor.

$\zeta^-(\gamma)$ can have only one accumulation point for $\gamma \downarrow 0$, because $\lambda_0 = h(\zeta^0)$ is the value function. Uniqueness of $\zeta^0$ is therefore clear and it follows that $\zeta^0 = \zeta^-(0)$. By assumption, the free boundary problem has a unique solution, hence it follows that $\pi_+(0) = \pi_0^+$. In particular, the curves $(0, \bar{\gamma}] \to \mathbb{R} : \gamma \mapsto \pi_\pm(\gamma)$ each have a unique limit $\pi_0^\pm$ as $\gamma \downarrow 0$, which equals $\pi_\pm(0)$, the solution of the free boundary problem.

References


