Portfolio selection with frictions
Abstract

This thesis studies several extensions of the utility maximization problem with (small) proportional transaction costs. We analyze how the optimal investment policy of a risk-averse investor is affected by the interplay between proportional transaction costs and i) binding portfolio constraints, ii) linear price impact, iii) continuous dividends and iv) the presence of multiple risky assets.

In the first part of the thesis, we consider the one-dimensional Black-Scholes model, i.e., the financial market consists of one safe asset with constant interest rate and one risky asset following a geometric Brownian motion. In the presence of (small) proportional transaction costs, an investor with constant relative risk aversion maximizes the long-term growth rate of her expected utility from terminal wealth. We incorporate additional market frictions such as i) binding portfolio constraints, ii) linear price impact, and iii) continuous dividends into the optimization problem. For each extension we characterize the optimal trading policy and discuss possible applications.

In the second part of the thesis, we consider a general multidimensional diffusion model in which trades are subject to small proportional transaction costs. In concrete terms, we assume that an investor with constant relative risk aversion maximizes her expected (risk-adjusted) profits. Since efficient algorithms for computing the optimal no-trade region in high dimensions are not available, we track the frictionless solution and optimize over suitable trading times. We provide explicit formulas for the optimal trading frequency and the associated welfare loss due to small transaction costs. Using Monte-Carlo simulations we numerically assess the performance of the trading frequency based strategy and compare it to several benchmarks.
Kurzfassung


Im zweiten Teil der Arbeit betrachten wir ein allgemeines, multidimensionales Diffusionsmodell, in welchem Aktienerstkäufe beziehungsweise -verkäufe kleine proportionale Transaktionskosten nach sich ziehen. Konkret nehmen wir an, dass ein Investor mit konstanter Risikoversion die erwarteten (nach Risikogewichteten) Profite maximiert. Da in höheren Dimensionen kein effizienter Algorithmus zur Berechnung der No-Trade Region bekannt ist, nehmen wir die fraktionslose Lösung als Basis und optimieren die zugehörigen Handelszeiten. Wir leiten explizite Formeln für die optimale Handelsfrequenz und den assoziierten Nutzenverlust durch kleine Transaktionskosten her. Mit Hilfe von Monte-Carlo Simulationen testen wir die Performance unserer Lösung und vergleichen diese mit verschiedenen Alternativstrategien.
vi
Acknowledgements

First and foremost, I would like to express my sincere gratitude to Prof. Johannes Muhle-Karbe for giving me the opportunity to write a PhD thesis under his supervision. I have greatly benefited from his immense knowledge on the subject and would like to thank him for his support, encouragement and patience during the past years. Moreover, I am deeply grateful to Prof. H. Mete Soner and Prof. Steven E. Shreve for agreeing to be on my thesis committee. In particular, I would like to thank Prof. Steven E. Shreve for his detailed suggestions.

Parts of the thesis are based on fruitful collaborations with Prof. Paolo Guasoni, Dr. Marko Weber and Dr. Ibrahim Ekren. I was very fortunate to work with these excellent researchers and have benefited a lot from many insightful discussions with them. My sincere thanks also go to Max Reppen for generously sharing his research results.

I wish to thank my friends and colleagues at Group 3 for countless funny moments. In particular, I would like to express my appreciation to Martin Herdegen for sharing a lot of helpful tips with me. Special thanks also go to Thomas Cayé for many enjoyable lunch hours. Albert Altarovici, Erdinç Akyildirim, Kai Du, Matti Kiiski, Mario Šikić and Danijel Zivoi – my office mates during the past years – deserve a lot of credit for putting up with me. Furthermore, I will certainly miss the exciting tennis matches with Kai Du, Ludovic Moreau, Max Reppen and Ariel Neufeld as well as the crazy ping pong duels with Albert Altarovici.

Last but not least, I would like to thank my parents for their love and support during the past years. This thesis would not have been possible without my supportive girlfriend Tamara Hediger who has been a constant source of love and inspiration. This thesis is dedicated to her.
Contents

1 Introduction .......................... 1
   1.1 Outline of the Thesis ............... 6

2 Transaction Costs and Binding Portfolio Constraints 9
   2.1 Introduction ....................... 9
   2.2 Model and Main Results ............. 11
   2.3 Implications ....................... 14
       2.3.1 Asset Demand and Welfare .......... 14
       2.3.2 Trading Volume ..................... 16
       2.3.3 Turnover, Spreads, and Liquidity Premia .......... 17
       2.3.4 Applications ..................... 19
   2.4 Heuristics .......................... 23
   2.5 Proofs of Theorem 2.2.3 and Theorem 2.2.4 ..... 25
       2.5.1 Proof of Theorem 2.2.3 ............. 25
           2.5.1.1 Explicit Formulae and their Properties .... 25
           2.5.1.2 Shadow Prices and Verification ......... 27
           2.5.1.3 Trading Volume ..................... 33
       2.5.2 Proof of Theorem 2.2.4 ............. 33

3 Transaction Costs and Linear Price Impact .... 35
   3.1 Introduction ....................... 35
   3.2 Model ............................. 37
   3.3 Main Results ....................... 39
   3.4 Numerical Examples .................. 40
   3.5 Heuristics .......................... 42
   3.6 Formal Asymptotics ................. 44
   3.7 Proofs of Theorem 3.3.1 and Theorem 3.6.1 .... 53
       3.7.1 Proof of Theorem 3.3.1 ............. 53
       3.7.2 Proof of Proposition 3.6.1 .......... 60

4 Transaction Costs and Continuous Dividends .... 63
   4.1 Introduction ....................... 63
   4.2 Model and Main Results ............. 66
       4.2.1 Model .......................... 66
       4.2.2 Main Results ..................... 67
       4.2.3 Statement of the Results .......... 73
Chapter 1

Introduction

Spearheaded by Merton in his seminal papers [98, 99], the continuous-time\(^1\) optimal investment problem has attracted a lot of attention in the field of Mathematical Finance. In the simplest setting, the problem considers a rational investor with constant relative risk aversion, trading in a financial market consisting of one safe asset and some risky ones. For a given planning horizon and a fixed initial endowment, the optimal investment problem deals with the following question: \textit{How should wealth be allocated between the safe and risky assets such that the investor's expected utility}\(^2\) of terminal wealth is maximized?

Using the tools of stochastic dynamic programming a first answer was provided by Merton in the frictionless Black-Scholes model, where the market consists of one safe asset with constant interest rate and multiple risky assets following geometric Brownian motions. Merton showed that, independent of the time horizon, the optimal strategy is to keep a constant fraction of wealth in each asset. Furthermore, he worked out an \textit{explicit} formula for the optimal fraction. Often, the related optimization problem and the associated solution are referred to as the \textit{Merton problem} and the \textit{Merton solution}, respectively.\(^3\) Since the groundbreaking results of Merton, numerous generalizations were studied to relax the original (unrealistic) model assumptions cf. e.g., [73, 74, 112, 75].

In this thesis we are interested in the impact of market frictions on the solution of the Merton problem. More specifically, we investigate how the optimal strategy is affected by the interplay between proportional transaction costs\(^4\) and other market features. Even in the presence of tiny proportional transaction costs

\(^{1}\)The discrete-time model was first introduced by Markowitz [93]. Further studies include [83, 102, 110].

\(^{2}\)In this thesis, we assume the investor has an iso-elastic utility function, i.e., \(U(x) = \frac{x^{1-\gamma}}{1-\gamma}\) for \(0 < \gamma \neq 1\).

\(^{3}\)Merton [98, 99] studied the general optimal consumption-investment problem, i.e., the investor maximizes the integrated utility of continuous consumption. Thus, both the investment decision and the consumption rule need to be determined. In this thesis, we assume that the investor only consumes at the terminal time. That is, by postponing the time horizon to infinity we solely focus on the optimal investment policy and neglect the corresponding consumption rule.

\(^{4}\)Proportional trading fees are omnipresent in our daily routine in the form of bid-ask spreads, e.g., at currency exchanges.
the Merton solution is no longer feasible, since the optimal wealth trajectory associated with the Merton solution is generically a diffusion process. Thus, following the Merton solution would require incessant trading and is therefore ruinously expensive in the presence of transaction costs. In the following, we shall catch a glimpse of the vast and rapidly growing literature on the Merton problem with proportional transaction costs.

The one-dimensional Merton problem with proportional transaction costs was first studied by Magill and Constantinides [91] and Constantinides [22]. Their (heuristic) insights suggested that the optimal strategy is characterized by a so-called no-trade region. In concrete terms, Magill and Constantinides [91] realized that "...the investor trades in securities when the variation in the underlying security prices forces his portfolio proportions outside a certain region about the optimal proportions in the absence of transaction costs". That is, a rational investor would trade the risky asset only if the risky weight (i.e., the fraction of wealth invested in the risky asset) is outside a certain region. For small proportional transaction costs the no-trade region contains the Merton proportion cf., e.g., [62]. Meaning, the investor only trades if the current risky weight significantly deviates from the optimal frictionless solution. Constantinides [22] further continued with the analysis and observed that: "...investors accommodate large transaction costs by drastically reducing the frequency and volume of trade". However, Constantinides [22] also reported that the welfare effect of transaction costs is typically small, since "...a small liquidity premium is sufficient to compensate an investor for deviating significantly from the target portfolio proportions".

Mathematically, the Merton problem with proportional transaction costs has been successfully studied using tools both from stochastic control theory and duality theory. In their landmark papers, Taksar, Klass and Assaf [119] as well as Davis and Norman [31] introduced the tools of (singular) stochastic control theory to deal with proportional transaction costs. While Davis and Norman [31] put the infinite horizon investment-consumption problem on a firm theoretical footing by rigorously setting up the associated stochastic control problem, Taksar, Klass and Assaf [119] studied the Merton problem with transaction costs in

---

5The only exception is if the Merton solution prescribes to buy-and-hold a single risky asset.

6While proportional transaction costs are common practice for large trades, up to a certain trading amount, private investors typically pay a fixed commission per trade — regardless of the trading sizes. Using the machinery of stochastic impulse control the Merton problem with fixed transaction costs was studied e.g., in [78, 5]. However, for institutional investors, these fixed transaction costs become negligible compared to their proportional counterparts and are typically waived. Thus, we disregard fixed transaction costs in the present thesis. For large portfolios, price impact is an additional key concern as we shall see in Chapter 3.

7The dimension refers to the number of risky assets considered in the Merton problem.

8While the infinite horizon problem leads to stationary policies, in a finite time horizon model the optimal no-trade region is time-dependent. For instance, the investor might refrain from trading (even in the presence of a positive equity premium) if the time to maturity is so short that potential gains from trading are insufficient to compensate the required transaction costs, cf., e.g., [85, 29, 14].
the context of maximizing the expected growth rate of wealth.\textsuperscript{9} Both studies showed that it is optimal to keep the risky weight (i.e., the ratio of funds in risky and safe assets) inside a no-trade region by reflecting off its boundaries (cf. also Dumas and Luciano [37]). That is, the investor trades just enough for the risky weight to stay inside the no-trade region. Since the fraction of wealth invested in the risky asset is in general a diffusion process, singular continuous processes, i.e., local time processes are needed to describe the optimal trading behavior.\textsuperscript{10} To determine the boundaries of the no-trade region, Davis and Norman [31] provided a numerical algorithm to solve the associated free-boundary problem. By means of viscosity solutions, Shreve and Soner [113] removed several assumptions from Davis and Norman [31]. For small transaction costs Shreve and Soner [113] and later Janeček and Shreve [62] also performed a sensitivity analysis of the value function, i.e., they derived the asymptotic expansion of the value function in terms of fractional powers of the transaction costs – confirming Constantinides’ observation about the small welfare effect of transaction costs (cf. [108, 43, 48, 63, 14]). Using homogenization techniques, Soner and Touzi [117] and Possamai, Soner and Touzi [107] successfully extended the analysis by allowing for general utility functions and stochastic dynamics in the one-dimensional and multidimensional setting,\textsuperscript{11} respectively.

On the other hand, starting with the pioneering papers of Jouini and Kallal [65] and Cvitanic and Karatzas [24], martingale methods have been introduced to deal with proportional transaction costs.\textsuperscript{12} In a finite horizon model, in which utility from terminal wealth is maximized, Cvitanic and Karatzas [24] proved the existence of an optimal strategy under the assumption that the corresponding dual problem admits a solution.\textsuperscript{13} Later, Kallsen and Muhle-Karbe [69] considered the infinite horizon Merton problem with proportional transaction costs and logarithmic utility. They constructed a fictitious shadow price process to which the powerful martingale theory was applied and solved this concrete problem. Herzegh and Prokaj [60] as well as Choi, Sirbu and Žitković [20] studied the construction of the shadow price process for power utility. The idea of Herzegh and Prokaj [60] is based on a result of Loewenstein [88, Theorem 1] suggesting that the shadow price process can be constructed as the marginal rate of substitu-

\textsuperscript{9}There is also another strand of literature which covers option pricing problems in the presence of proportional transaction costs, cf., e.g., [82, 61, 32, 25, 122, 8].

\textsuperscript{10}The Merton problem with proportional transaction costs is therefore also referred to as a singular stochastic control problem.

\textsuperscript{11}Akian, Menaldi and Sulem [1] extended the Davis and Norman model [31] by allowing for multiple uncorrelated risky assets and provided a characterization of the associated value function. For correlated assets Muthuraman and Kumar [103] performed a detailed numerical analysis of the optimal no-trade region when transaction costs are sufficiently small. Bichuch and Shreve [15] pointed out that in case of multiple correlated risky assets the assumption that the value function is twice continuously differentiable across the boundaries is questionable. In contrast, they provided a new methodology to derive asymptotic expansions. Compare also Liu [84] as well as Goodman and Ostrov [48] for other results in the multidimensional case.

\textsuperscript{12}Similar results in discrete time were obtained by Kusuoka [80].

\textsuperscript{13}The existence of the solution is addressed by Cvitanic and Wang [26]. Further work in this vein include, e.g., [66, 33].
In introduction of risky for safe assets for the optimal investor. The shadow price process (lying between the bid-ask spread) and the original price process leads to the same investment decisions in terms of optimal strategy and utility.\footnote{The shadow price approach has been used in various contexts, cf. e.g., \cite{81, 88, 26, 54, 55}.} Applying this shadow price approach, Gerhold, Muhle-Karbe and Schachermayer \cite{45, 44} revisited the models of Taksar, Klass and Assaf \cite{119} as well as Davis and Norman \cite{31} and derived asymptotic expansions for the optimal strategy and the optimal growth rate in case of logarithmic utility. Similar results were obtained by Gerhold, Guasoni, Muhle-Karbe and Schachermayer \cite{43} for power utility. Combined with the long-run analysis of Guasoni and Robertson \cite{56} the shadow price approach allows us to derive a duality-based verification theorem. For small proportional transaction costs, Kallsen and Muhle-Karbe established formal asymptotics for the optimal policy both in the context of option pricing \cite{70} and portfolio selection \cite{71} with general continuous dynamics.\footnote{Further issues regarding financial applications are discussed in Martin and Schöneborn \cite{96} and Martin \cite{95}.} Recently, Kallsen and Li \cite{68} have turned the results of Kallsen and Muhle-Karbe \cite{70} into rigorous proofs.

The main goal of this thesis is to study extensions of the Merton problem with proportional transaction costs by including additional market frictions. In the first part of the thesis we focus on the one-dimensional case, i.e., the market consists of one safe asset with constant interest rate and one risky asset following a geometric Brownian motion. For tractability reasons, we assume that the investor has constant relative risk aversion and postpone the horizon to infinity as in \cite{37, 49, 50}. The objective function is then to maximize the so-called equivalent safe rate. By definition, a full safe investment at the equivalent safe rate yields the same performance (in terms of utility) as trading in the original market.

Building on this baseline model, we discuss how incorporating binding portfolio constraints, a linear price impact and continuous dividends affect the optimal no-trade region. These additional market frictions are motivated by the following questions:

(i) \textit{binding portfolio constraints:} In addition to proportional transaction costs many mutual funds also face regulatory constraints, for example, a certain leverage ratio may not be exceeded. In this case, we are interested how the no-trade region is affected by these constraints.

(ii) \textit{linear price impact:} Whenever a large trading order is placed, the price of the risky asset tends to move against the investor (price impact), i.e., a large buy order (or sell order) drives up (or drives down) the price of the risky asset. For a large institutional trader both proportional transaction costs and price impact are key concerns. We attempt to deduce the optimal strategy when an investor is facing both market frictions.

(iii) \textit{continuous dividends:} The optimal strategy characterized by the no-trade region does not match the suggestion of many financial advisers to buy-and-hold the risky asset if the planning horizon is sufficiently long. Assuming
that the risky asset also pays continuous dividends we analyze the corresponding shift of the no-trade region. In particular, we address the question whether there are parameter constellations such that the optimal strategy is to never sell the risky asset.

For each extension, we derive the optimal strategy and discuss possible applications. To tackle these problems mathematically, we apply a two-step procedure combining tools from both stochastic control theory and duality theory. That is, we first use results from stochastic control theory to heuristically derive the associated Hamilton-Jacobi-Bellman (henceforth HJB) equation with the corresponding (free) boundary conditions. A candidate optimal strategy is then established by performing a detailed analysis of the HJB equation. With the candidate solution at hand, we prove a rigorous verification theorem by means of convex duality in a second step.

Since the bid-ask spread for a liquid stock is typically small,\(^{16}\) we put special focus on the limiting case when transaction costs are approaching zero. With additional frictions included into the Merton problem the optimal solution in general will not admit explicit formulas. The assumption of small transaction costs enables us to study trading costs as a small perturbation of the classical benchmark and hence to obtain analytic approximation formulas for the optimal strategy. However, while asymptotics are a crucial analytical tool, the qualities of the resulting approximation formulas depend on the underlying parameter values. Thus, it is important to know the limits of the asymptotics and recognize potential regime shifts as we shall see in Chapter 4.

In the second part of the thesis, we consider the multidimensional Merton problem with small proportional transaction costs. The optimal strategy is then given by a multidimensional no-trade region which in turn is characterized by the solution of a multidimensional free boundary problem, cf., e.g., [107]. However, asymptotic expansions are no longer available in closed form, and numerical computations\(^{17}\) of the trading boundaries are quite cumbersome and involved (even in the limit of small transaction costs). Thus, rather than computing an asymptotically optimal no-trade region, we track the frictionless solution and optimize over suitable trading times. That is, we are interested in the optimal trading frequency which strikes the optimal balance between limiting transaction costs and minimizing the associated tracking error.

To be more precise, we consider a market which consists of one safe asset with constant interest rate and multiple (correlated) risky assets following general Itô processes. As in [67, 96, 95, 42, 52] we maximize the expected relative returns and penalize the corresponding variances. The assumption of small transaction costs allows us to give a precise formulation of the associated optimization problem by means of asymptotic expansions. Using results from stochastic calculus we determine explicit formulas for the optimal trading frequency and the associated

\(^{16}\)For a stock with market capitalization rank less than 200, the bid-ask spread is about 25 basis points, cf., e.g., Novy-Marx and Velikov [104, Figure 1].

\(^{17}\)For instance, Altaracovici, Reppen and Soner [6] used a policy iteration scheme to compute the asymptotically optimal no-trade region, cf. also [107, 103].
welfare loss in the multidimensional setting. In addition, we conduct a Monte-Carlo simulation study to test the performance of the trading frequency based rebalancing rule and compare it to several alternatives.

1.1 Outline of the Thesis

This thesis is based on a series of research papers [86, 87, 51, 40] and the chapters are therefore self-contained. It is divided into two parts and consists of five chapters. Part I (i.e., Chapters 2, 3 and 4) contains the single stock analysis while Part II (i.e., Chapter 5) deals with the problem of multiple risky assets. For each chapter we now briefly describe the problem considered, state some of the most important results, and sketch the main ideas of the proof. For a more detailed exposition of the problem and a comprehensive overview of the related literature we refer to the separate introduction sections at the beginning of each chapter.

Chapter 2 is based on Liu and Muhle-Karbe [86] and studies the optimal investment problem in the one-dimensional Black-Scholes model where the investor is facing both small proportional transaction costs and binding exogenous portfolio constraints, i.e., an upper bound on the risky weight. In the infinite horizon model of Dumas and Luciano [37] we derive the optimal no-trade region and the associated welfare. All formulas are explicit in terms of the model parameters and a gap parameter which is the solution of a one-dimensional scalar equation. As transaction costs approach zero, explicit asymptotic expansions for the optimal strategy and the associated welfare are obtained. In particular, we find that in the presence of binding portfolio constraints the effects of transaction costs on portfolio composition and welfare are of the same order. As an application, we discuss how regulatory changes might affect deposit rates offered for illiquid loans.

For the proof, we first use partially heuristic arguments from stochastic control theory to derive the associated free boundary problem. Applying suitable variable transformations, we can solve the HJB equation explicitly, and this leads to a candidate optimal strategy. Next, we construct a fictitious shadow price process trading without transaction costs and portfolio constraints, which is equivalent to the original market both in terms of optimal strategy and utility. To conclude, we apply a verification theorem in the spirit of Guasoni and Robertson [56] to the shadow price process.

In Chapter 3, which stems from Liu, Muhle-Karbe and Weber [87], we consider the optimal investment problem of a large institutional trader, i.e., we combine the price impact model of Guasoni and Weber [57] with proportional transaction costs. Considering both market frictions, we show that the optimal strategy is again determined by a no-trade region. However, since singular controls are not compatible with price impact, the optimal strategy is to steer the risky weight process back into the no-trade region at a finite, absolutely continuous rate. As a compensation for the finite trading speed, the width of the no-trade region decreases compared to the pure transaction costs case. Moreover, for small transaction costs and a small price impact we perform a formal asymptotic analysis,
which reduces the complexity of the optimal solution to finding the root of a scalar equation.

As in Chapter 2, the proof again follows the two-step procedure. With additional linear price impact, we directly attack the associated HJB equation with the corresponding boundary conditions. A qualitative analysis of the resulting ordinary differential equation shows the existence and uniqueness of the solution associated to the HJB equation. This in turns allows us to perform a verification theorem similar to the one of Guasoni and Weber [57] to conclude optimality.

Chapter 4 is based on Guasoni, Liu and Muhle-Karbe [51] and studies the joint impact of proportional transaction costs and continuous dividends on the optimal investment policy. That is, we work with the infinite horizon model of Dumas and Luciano [37]. The dividend payments are proportional to the dollar amount invested in the risky asset and hence more or less cash is transferred into the bank account when the risky weight is high or low, respectively. This mean-reverting effect of the risky weight pushes the selling boundary upwards, i.e., the investor is more reluctant to sell the risky asset in the presence of continuous dividends. For sufficiently large dividends this upward trend results in the never-sell optimality, i.e., we show that for a realistic range of model parameters the investor should buy the stock when the risky weight is low, and otherwise hold it. Even when the never-sell strategy is not optimal, the loss of ignoring the selling signal is comparably smaller than the loss of ignoring the buying signal – a result that we derive using Monte-Carlo simulations. The presence of capital gains taxes further increases the attractiveness of the never-sell strategy. As an application, we discuss potential investment implications for retirement planning.

For the proof, we first heuristically derive the HJB equation with the corresponding free boundary conditions. Next, we apply an elegant variable transformation from Jang [64] to derive analytical solutions of the HJB equation in terms of special functions. This in turn leads to a candidate optimal strategy. We then construct a fictitious shadow price process which pays the same dividend payments as the original market. Finally, a verification theorem in spirit of Guasoni and Robertson [56] applied to the shadow price process allows us to conclude optimality.

In Chapter 5, which stems from Ekren, Liu and Muhle-Karbe [40], we consider the multidimensional Merton problem with small proportional transaction costs. In concrete terms, we assume that the market consists of one safe asset with constant interest rate and multiple (correlated) risky assets following general Itô processes. To circumvent the cumbersome task of numerically computing asymptotically optimal trading boundaries, we only consider a subclass of admissible strategies. More specifically, we focus on strategies which trade back directly to the Merton solution at trading times, which are already specified at the last transaction date and refrain from portfolio rebalancing in between. We derive explicit formulas for the optimal trading frequency and the associated welfare in arbitrarily high dimensions. Using Monte-Carlo simulations, we find that with constant investment opportunities our tractable solution yields almost optimal performance in arbitrary dimensions. However, when the investment op-
opportunities are stochastic, e.g., if the frictionless optimal strategies are of trend following type, the effect of transaction costs is more pronounced. In the one-dimensional case, the “time-base” rebalancing rule is considerably outperformed by the optimal strategy. In the case of multiple risky assets, even strategies with an ad-hoc no-trade region, i.e., an inactivity region obtained by concatenating optimal univariate no-trade intervals, achieve better performance than the trading frequency based competitor.

For the derivation, we first show that for small transaction costs the objective function can be decomposed into its frictionless counterpart as well as losses accrued due to transaction costs and discretization error. Applying tools from stochastic calculus, we explicitly determine the rates of convergence of both losses. We then perform a detailed asymptotic analysis to find the trading frequency which strikes the optimal balance between limiting transaction costs and minimizing the associated discretization error.

Finally, an overview of notations can be found in Appendix A.
Chapter 2
Transaction Costs and Binding Portfolio Constraints

2.1 Introduction

Transaction costs and trading constraints are two central frictions in financial markets, but most of the literature focuses either on transaction costs [91, 22, 31, 37, 113] or on portfolio constraints [121, 49, 23] – separately – with the exception of the paper by Dai, Jin, and Liu [27]. They describe the value function of a finite-horizon model with both frictions by means of a double-obstacle problem, obtain some monotonicity properties of the optimal trading boundaries, and conduct an extensive numerical analysis.

Our goal is to obtain more tractable results. To this end, we work with the infinite-horizon model of Dumas and Luciano [37]: an investor with constant relative risk aversion trades to maximize the long-term growth rate of her utility, in a market with one risky asset following geometric Brownian motion and one safe asset with constant interest rate. In the presence of a binding exogenous portfolio constraint, i.e., an upper bound on the risky weight, and proportional transaction costs, we derive the optimal trading strategy as well as the associated welfare and trading volume. As in the unconstrained case [43], all formulas are explicit in terms of the market and preference parameters as well as an additional transaction cost gap, characterized as the root of a scalar equation. As the spread becomes small, all quantities admit explicit asymptotic expansions, in terms of model and preference parameters only.

These results help to clarify the joint impact of transaction costs and constraints.

Constantinides [22] observed that – in the absence of constraints – “transaction costs have a first order effect on asset demand”. On the other hand, he found that their welfare impact is typically less pronounced, since “a small liquidity premium is sufficient to compensate an investor for deviating significantly from the target portfolio proportions”. Starting with Shreve and Soner [113], these numerical results have been made precise in an asymptotic manner: deviations from the

---

1This chapter is based on Liu and Muhle-Karbe [86].
optimal frictionless portfolio are of order $\varepsilon^{1/3}$ as the spread $\varepsilon$ becomes small [62], but the corresponding liquidity premium and utility loss are only of order $\varepsilon^{2/3}$ [113]. These results are robust within the class of diffusion models [117], but break down in the presence of binding constraints. Here, we find that the effects on portfolio composition and welfare are of the same order $\varepsilon^{1/2}$, in line with the numerical observation of [27] that “transaction costs can have a first-order effect” on welfare.

In the absence of transaction costs, the impact of binding portfolio constraints is easily quantified: harder constraints simply reduce the investor’s risky weight. Considering them jointly with trading costs additionally allows us to assess their impact on the share turnover the investor’s rebalancing generates. The latter turns out to be of order $\varepsilon^{-1/2}$, and therefore dominates its frictionless counterpart of order $\varepsilon^{-1/3}$ for sufficiently small spreads. The corresponding comparative statics strongly depend on whether the investor’s position is leveraged or not. In the absence of leverage, tighter constraints increase trading volume, and constrained turnover dominates its unconstrained counterpart even for large transaction costs. On the contrary, if the optimal policy prescribes a leveraged position in the risky asset, then sufficiently tight constraints reduce turnover, and unconstrained turnover need only be dominated for very small spreads.

Even though all individual asymptotic rates change compared to the unconstrained case, one key observation from the unconstrained case [43] remains valid: The welfare impact of transaction costs equals implied trading volume (measured consistently) times the spread, times a constant.

To illustrate the implications of our results, we discuss two applications. First, we consider an investor choosing which prime broker to use to buy a leveraged risky position on margin. Each broker is prepared to let the investor borrow at a specific rate and up to a given leverage constraint, and our results allow us to quantify the attractiveness of each combination. We find that among those brokers that are equally attractive in the absence of transaction costs, the investor typically prefers those with harder constraints (i.e., high margin requirements) and lower lending rates, because transaction costs make highly leveraged portfolios less attractive than they appear to be in frictionless markets. As a second application, we consider a bank that can borrow from its depositors at the safe rate to provide long-term loans, whose book values are assumed to follow geometric Brownian motion. Then, the bank’s optimization problem is precisely of the type considered above, and the portfolio constraints correspond to the minimum capital requirements imposed by regulatory authorities. This model can in turn be used to assess the impact of tighter regulatory constraints: Assuming that the bank aims to achieve the same performance, it will have to decrease its deposit rate to compensate for the negative welfare effect of harder constraints. The size of this effect crucially depends on the liquidity of the long-term loans the bank is providing, i.e., on the transaction costs incurred when prematurely liquidating them. Since leveraged positions are less attractive with transaction costs, the decrease in the banks deposit rate is most pronounced in the perfectly liquid case, and is diminished substantially if illiquidity is accounted for.
On a mathematical level, the proof of our verification theorem is based on applying a variant of the argument of Guasoni and Robertson [56] to a fictitious shadow price trading without transaction costs, which is equivalent to the original market with transaction costs both in terms of the optimal strategy and utility. In contrast to the unconstrained case [43], the dynamics of the shadow price involve a singular component, such that the corresponding frictionless market is not arbitrage-free. The constraints, however, prevent the optimal strategy from exploiting these arbitrage opportunities.

The remainder of this chapter is organized as follows: Section 2.2 describes the setting and states the main results, whose implications are discussed in Section 2.3. The main results are derived heuristically in Section 2.4, and proved rigorously in Section 2.5.

### 2.2 Model and Main Results

Consider a market with one safe asset \( S^0_t = e^{rt}, r > 0 \), and one risky asset, whose ask (buying) price \( S_t \) follows geometric Brownian motion:

\[
    dS_t/S_t = (\mu + r)dt + \sigma dW_t, \quad S_0 \in (0, \infty).
\]

Here, \( \mu > 0 \) is the expected excess return, \( \sigma > 0 \) the volatility, and \((W_t)_{t \geq 0}\) is a Brownian motion. The corresponding bid (selling) price is \((1 - \varepsilon)S_t\), where \( \varepsilon \in (0, 1) \) represents the width of the relative bid-ask spread.\(^2\)

A self-financing trading strategy is an \( \mathbb{R}^2 \)-valued, predictable process \((\phi^0, \phi)\) of finite variation: \((\phi^0_t, \phi_0) = (\xi^0, \xi) \in \mathbb{R}^2\) denotes the initial positions (in units) in the safe and risky asset, and \((\phi^0_t, \phi_t)\) denotes the positions held at time \( t \). Writing \( \phi_t = \phi^+_t - \phi^-_t \) as the difference between the cumulative number of shares bought \((\phi^+_t)\) and sold \((\phi^-_t)\) by time \( t \), the self-financing condition states that the safe position only changes due to trades in the risky asset:

\[
    S^0_t d\phi^0_t = -S_t d\phi^+_t + (1 - \varepsilon)S_t d\phi^-_t, \quad \forall t \geq 0. \tag{2.2.1}
\]

As is customary, attention is restricted to strategies that remain solvent at all times. In addition, we focus on strategies that satisfy an exogenous portfolio constraint, i.e., whose risky weight is uniformly bounded from above.

**Definition 2.2.1.** A self-financing trading strategy \((\phi^0, \phi)\) is called admissible, if its liquidation value is positive at all times,

\[
    \Xi_t^\phi := \phi^0_t S^0_t + (1 - \varepsilon)S_t \phi^+_t - S_t \phi^-_t \geq 0, \quad \forall t \geq 0,
\]

and it satisfies the portfolio constraint

\[
    \pi_t := \frac{\phi^+_t S_t}{\phi^0_t S^0_t + \phi^-_t S_t} \leq \pi_{\text{max}}, \quad \forall t \geq 0. \tag{2.2.2}
\]

\(^2\)This notation is equivalent to the usual setup with the same constant proportional transaction costs for purchases and sales [31, 62, 113]. Indeed, set \( \hat{S}_t = \frac{2\varepsilon}{1 - \varepsilon} S_t \) and \( \hat{\varepsilon} = \frac{\varepsilon}{1 - \varepsilon} \).

Then \(((1 - \varepsilon)\hat{S}_t, \hat{S}_t)\) coincides with \(((1 - \hat{\varepsilon})\hat{S}_t, (1 + \hat{\varepsilon})\hat{S}_t)\). Conversely, any bid-ask process \(((1 - \varepsilon)S_t, (1 + \varepsilon)S_t)\) with \( \varepsilon \in (0, 1) \) equals \(((1 - \hat{\varepsilon})S_t, \hat{S}_t)\) for \( S_t = (1 + \hat{\varepsilon})\hat{S}_t \) and \( \varepsilon = \frac{\hat{\varepsilon}}{1 - \hat{\varepsilon}} \).
It will sometimes be notationally convenient to decompose the constraint $\pi_{\text{max}}$ in a multiplicative manner as

$$\pi_{\text{max}} = \kappa \pi^*,$$

where

$$\pi^* = \frac{\mu}{(\gamma \sigma^2)}$$

is the frictionless optimal Merton proportion and $\kappa$ is the relative constraint. Throughout, the constraints are assumed to be binding, i.e., $\pi_{\text{max}} < \pi^*$, or equivalently, $\kappa < 1$. Otherwise, they have no effect as in the frictionless case if the transaction costs $\varepsilon$ are sufficiently small.

As in [37, 49, 50], the investor has constant relative risk aversion $0 < \gamma \neq 1$, an infinite planning horizon and maximizes the growth rate of her expected utility from terminal wealth over a long horizon.

**Definition 2.2.2.** An admissible strategy $(\phi^0, \phi)$ is called long-run optimal, if it maximizes the equivalent safe rate

$$\liminf_{T \to \infty} \frac{1}{T} \log \mathbb{E} \left[ (\Xi_T^{\phi})^{1-\gamma} \right]^{1-\gamma}$$

(2.2.3)

over all admissible strategies, where $0 < \gamma \neq 1$ denotes the investor’s relative risk aversion.

Our main results can be summarized as follows:

**Theorem 2.2.3.** Assume an investor with constant relative risk aversion $0 < \gamma \neq 1$ trades to maximize the equivalent safe rate in the presence of a binding portfolio constraint $0 < \pi_{\text{max}} = \kappa \pi^* \neq 1$.

Then, for small transaction costs $\varepsilon > 0$:

i) (Equivalent Safe Rate)

For the investor, trading the risky asset with transaction costs and portfolio constraints is equivalent to leaving all wealth in a hypothetical safe asset, which pays the higher equivalent safe rate

$$\text{ESR} = r + \frac{\mu^2}{2 \gamma \sigma^2} \left( 2 \kappa/(1 - \lambda) - \kappa^2 (1 - \lambda) \right)^2,$$

where the gap $\lambda$ is defined in Item iv) below.

ii) (Liquidity Premium)

Trading the risky asset with transaction costs and constraints is equivalent to trading a hypothetical asset with no transaction cost and no constraint, with the same volatility $\sigma$, but with lower expected excess return $\mu \sqrt{2 \kappa/(1 - \lambda) - \kappa^2 (1 - \lambda)}$. Thus, the liquidity premium is

$$\text{LiPr} = \mu - \mu \sqrt{2 \kappa/(1 - \lambda) - \kappa^2 (1 - \lambda)}.$$

---

3 That is, a full safe investment at the equivalent safe rate yields the same utility as investing optimally in the original market cf. Theorem 2.2.3 i).

4 In the degenerate case $\pi_{\text{max}} = \kappa \pi^* = 1$, an optimal strategy is to fully invest into the risky asset initially and never trade again afterwards. Both share and wealth turnover vanish in this case, and the equivalent safe rate and liquidity premium coincide with their counterparts in the absence of transaction costs.
iii) (Trading Policy)
It is optimal to keep the risky weight (in terms of the ask price) between the buying and selling boundaries

\[ \pi_- = (1 - \lambda)\pi_{\text{max}}, \quad \pi_+ = \pi_{\text{max}}. \]

iv) (Gap)
\( \lambda \) is the unique value for which the solution of the initial value problem

\[ 0 = \frac{\partial w}{\partial x} + (1 - \gamma)w(x)^2 + \left( \frac{2\mu}{\sigma^2} - 1 \right)w(x) - \frac{\mu^2}{\gamma \sigma^2}(1 - (1 - \kappa(1 - \lambda))^2), \]
\[ w(0) = (1 - \lambda)\pi_{\text{max}}, \]
also satisfies the terminal value condition

\[ w \left( \log \left( \frac{u}{l(\lambda)} \right) \right) = \frac{\pi_{\text{max}}(1 - \epsilon)}{(1 - \pi_{\text{max}}) + \pi_{\text{max}}(1 - \epsilon)} =: w_+, \]
where

\[ \frac{u}{l(\lambda)} = \frac{\pi_{\text{max}}/(1 - \pi_{\text{max}})}{(1 - \lambda)\pi_{\text{max}}/(1 - (1 - \lambda)\pi_{\text{max}})}. \]

v) (Share Turnover)
Share Turnover, defined as shares traded \( d\|\varphi\|_t = d\varphi_t^\uparrow + d\varphi_t^\downarrow \) divided by shares held \( |\varphi_t| \), has the long-term average:

\[ \text{ShTu} := \lim_{T \to \infty} \frac{1}{T} \int_0^T \frac{d\|\varphi\|_t}{|\varphi_t|} = \begin{cases} \frac{\sigma^2}{2} \left( \frac{2\mu}{\sigma^2} - 1 \right) \left( \frac{1 - \pi_-}{\frac{u}{l(\lambda)} \frac{2\mu}{\sigma^2} - 1} - \frac{1 - w_+}{\frac{u}{l(\lambda)} \frac{2\mu}{\sigma^2} - 1} \right), & \text{if } \mu \neq \frac{\sigma^2}{2}, \\ \frac{\sigma^2}{2\log \left( \frac{u}{l(\lambda)} \right)} (1 - \pi_- + 1 - w_+), & \text{if } \mu = \frac{\sigma^2}{2}. \end{cases} \]

vi) (Wealth Turnover)
Wealth Turnover, defined as wealth traded divided by wealth held, has the long-term average:

\[ \text{WeTu} := \lim_{T \to \infty} \frac{1}{T} \left( \int_0^T \frac{S_t d\varphi_t^\uparrow}{\varphi_t^0 S_t^0 + \varphi_t S_t} + \int_0^T \frac{(1 - \epsilon)S_t d\varphi_t^\downarrow}{\varphi_t^0 S_t^0 + \varphi_t(1 - \epsilon)S_t} \right) = \begin{cases} \frac{\sigma^2}{2} \left( \frac{2\mu}{\sigma^2} - 1 \right) \left( \frac{\pi_- (1 - \pi_-)}{(u/l(\lambda)) \frac{2\mu}{\sigma^2} - 1} - \frac{w_+(1-w_+)}{(u/l(\lambda)) \frac{2\mu}{\sigma^2} - 1} \right), & \text{if } \mu \neq \frac{\sigma^2}{2}, \\ \frac{\sigma^2}{2\log \left( \frac{u}{l(\lambda)} \right)} (\pi_- (1 - \pi_-) + w_+(1 - w_+)), & \text{if } \mu = \frac{\sigma^2}{2}. \end{cases} \]
vii) (Asymptotics)

The following asymptotic expansions hold true as $\varepsilon \downarrow 0$:

\[
\text{ESR} = r + \frac{\mu^2}{2\gamma \sigma^2} \left( \frac{2\pi_{\text{max}}}{\pi_*} - \left( \frac{\pi_{\text{max}}}{\pi_*} \right)^2 \right)
- \frac{\sigma^2}{\gamma} \left( \frac{\pi_{\text{max}}}{\gamma} \right) \left( 1 - \pi_{\text{max}} \right) \left( 1 - \pi_{\text{max}} \right)^2 \varepsilon^{1/2} + O(\varepsilon),
\]
\[
\text{LiPr} = \mu \left( 1 - \frac{2\pi_{\text{max}}}{\pi_*} \right) \left( \frac{\pi_{\text{max}}}{\gamma} \right) \left( 1 - \pi_{\text{max}} \right) \left( 1 - \pi_{\text{max}} \right)^2 \varepsilon^{1/2} + O(\varepsilon),
\]
\[
\pi_- = \pi_{\text{max}} - \left( \frac{1}{\gamma} \frac{\pi_{\text{max}}}{\pi_*} \left( 1 - \pi_{\text{max}} \right)^2 \right)^{1/2} \varepsilon^{1/2} + O(1),
\]
\[
\lambda = \left( \frac{1}{\gamma} \left( 1 - \pi_{\text{max}} \right)^2 \varepsilon^{1/2} + O(\varepsilon),
\]
\[
\text{ShTu} = \gamma \sigma^2 \left( \frac{\pi_{\text{max}}}{\gamma} \right) \left( 1 - \pi_{\text{max}} \right) \left( 1 - \pi_{\text{max}} \right)^2 \varepsilon^{-1/2} + O(1),
\]
\[
\text{WeTu} = \gamma \sigma^2 \left( \frac{\pi_{\text{max}}}{\gamma} \right) \left( 1 - \pi_{\text{max}} \right) \left( 1 - \pi_{\text{max}} \right)^2 \varepsilon^{-1/2} + O(1).
\]

As in the absence of constraints, the stationary policy from Theorem 2.2.3 is also approximately optimal - at the leading order $\varepsilon^{1/2}$ for small transaction costs $\varepsilon$ - for any finite time horizon:

**Theorem 2.2.4.** Fix a time horizon $T > 0$. Then, the finite-horizon equivalent safe rate of any admissible strategy $(\psi^0, \psi)$ satisfies the upper bound

\[
\frac{1}{T} \log E[(\Xi_T^{\psi})^{1-\gamma}]^{1-\gamma} \leq \text{ESR} + \frac{1}{T} \log (\xi_0^- + \xi_0 - S_0) + O(\varepsilon).
\]

The finite-horizon equivalent safe rate of the long-run optimal strategy $(\varphi^0, \varphi)$ from Theorem 2.2.3 satisfies the lower bound

\[
\frac{1}{T} \log E[(\Xi_T^{\varphi})^{1-\gamma}]^{1-\gamma} \geq \text{ESR} + \frac{1}{T} \log (\xi_0^- + \xi_0 - S_0) + O(\varepsilon).
\]

### 2.3 Implications

Let us now discuss some of the implications of our main result.

#### 2.3.1 Asset Demand and Welfare

In the absence of constraints, the impact of transaction costs on the optimal policy is large ($\sim \varepsilon^{1/3}$, cf. [62]) whereas their effect on welfare is small ($\sim \varepsilon^{2/3}$, cf.
2.3 Implications

With binding constraints, this no longer holds true: Both effects turn out to be proportional to the square-root of the spread $\varepsilon^{1/2}$, making precise the observation of [27] that transaction costs can have a first-order effect on welfare. The reason is that investors, if unconstrained, would like to hold larger risky positions than they are allowed to and therefore are more reluctant to tolerate downward swings of their risky position. This leads to a smaller no-trade region and hence, due to larger trading costs, to a bigger welfare impact.

More formally, this can also be rationalized by adapting the argument of Rogers [108]. He argued that the losses per unit time due to transaction costs are proportional to the local time of a reflected Brownian motion and hence of order $\varepsilon/x$, where $x$ denotes the length of the no-trade region. This reasoning only depends on the width of the no-trade region and hence remains valid also in the presence of constraints. What changes is the displacement loss due to deviating from the frictionless optimal policy. In the absence of constraints, the optimal frictionless Merton proportion $\pi_*$ lies in the no-trade region $[\pi_-, \pi_+]$. Consequently, the value function behaves like $(\pi - \pi_*)^2$ near the maximum $\pi_*$ according to Taylor’s theorem, leading to losses of order $x^2$ per unit time due to suboptimal portfolio composition. Minimizing the total loss then leads to $x \sim \varepsilon^{1/3}$ and a welfare loss $\sim \varepsilon^{2/3}$. In the presence of binding constraints, the frictionless minimizer no longer lies in the no-trade region; consequently, the value function behaves linearly leading to losses of order $x$. Again minimizing the total loss gives $x \sim \varepsilon^{1/2}$ and a welfare loss of the same order.

For small transaction costs, it is possible to determine the comparative stat-

---

Note that this need not hold for leveraged positions if transaction costs are sufficiently large [113, p. 675], but is always true if the transaction costs are small enough (compare [62, p. 182]).
ics of the width of the no-trade region for varying constraint levels by analyzing the leading term in the asymptotic expansion. The above argument that investors constrained from above only tolerate smaller downward moves of their risky weight suggests that harder constraints should lead to a smaller no-trade region. This intuitive reasoning is indeed true almost generically, i.e., unless the unconstrained frictionless risky weight lies too close to unity.\footnote{In the degenerate case \( \pi_* = 1 \), it is optimal to invest in a full risky position initially, and then hold the latter without further trades. Hence, the unconstrained no-trade region vanishes, but its width is increased by any non-trivial constraints. For unconstrained frictionless weights close enough to one (\( \pi_* \in [0.93, 1.25] \)) imposing constraints can similarly still increase the width of the no-trade region.} In all other cases the intuition that harder constraints lead to a smaller no-trade region indeed holds true at the leading order. Using the exact solution for \( \lambda \) Figure 2.1 depicts the length of the no-trade region against the bid-ask spread and confirms the asymptotic result.

## 2.3.2 Trading Volume

In the frictionless case, the impact of binding constraints is simple: They reduce the investor’s risky position. The present setting with transaction costs also allows us to quantify their effect on the implied trading volume, which shows a more involved picture. Indeed, two competing effects are at work here: On the one hand, the constraints reduce the investor’s risky position, thereby also reducing the amount of rebalancing necessary to keep the risky weight in the no-trade region. On the other hand, additional trading is required as the constraints (typically) lead to a smaller no-trade region. Our asymptotic results show that as the spread becomes small the latter effect prevails: The smaller no-trade region leads to share and wealth turnovers of order \( \varepsilon^{-1/2} \), in contrast to the rate \( \varepsilon^{-1/3} \) observed in the unconstrained case \([43]\). That is, for sufficiently small spreads, investors trade more in the presence of binding constraints.

The comparative statics of share turnover for small costs can again be analyzed by means of the leading term in the asymptotic expansion. In the no-leverage case \( \pi_* \in (0, 1) \), harder constraints not only (typically) decrease the width of the no-trade region but also move the risky weight away from the degenerate buy-and-hold strategy obtained for \( \pi_* = 1 \). Consequently, at the leading order \( \varepsilon^{-1/2} \), harder constraints always imply increased share turnover in this case. In contrast, the situation is ambiguous in the leverage case. On the one hand, harder constraints decrease the width of the no-trade region, thereby increasing turnover. On the other hand, however, they decrease turnover by pulling the risky weight closer to the buy-and-hold level. For sufficiently hard constraints (\( \pi_{\text{max}} \in [1, \frac{1+2\pi_*}{3}] \)) the latter effect prevails, decreasing turnover, and vice versa for \( \pi_{\text{max}} > \frac{1+2\pi_*}{3} \).

Since all of these results only hold true asymptotically as the spread \( \varepsilon \) becomes small, it is interesting to compare them with their exact counterparts, obtained by numerically solving for the gap \( \lambda \). The results are reported in Figure 2.2. In the absence of leverage for the frictionless weight \( \pi_* \), there is perfect agreement with
2.3 Implications

Figure 2.2: Left panel: Share turnover (vertical axis, annual fractions traded) plotted against the bid-ask spread $\varepsilon$ (horizontal axis) for an unconstrained weight $\pi^* = 62.5\%$ without leverage (solid), and with constraints $\pi_{\text{max}} = 50\%$ (dashed) and $40\%$ (dotted). Right panel: Share turnover against the spread for an unconstrained weight of $\pi^* = 390\%$ with leverage (solid), and with constraints $\pi_{\text{max}} = 225\%$ (dashed) and $175\%$ (dotted). Model parameters are $\mu = 8\%$, $\sigma = 16\%$ and risk aversion is $\gamma = 5$ (left panel) or $\gamma = 0.8$ (right panel).

the asymptotic results: Turnover is increasing with harder constraints, and larger than in the unconstrained case. In the leverage case, however, the situation is less clear-cut. The comparative statics again match the asymptotic results: Both constraints lie in the domain where harder constraints decrease turnover asymptotically, matching the numerical results. Compared to the unconstrained case, however, we observe that very low levels of transaction costs may be needed for the constrained turnover to surpass its unconstrained counterpart, in particular, for tight constraints.

The corresponding results for wealth turnover are more involved and are omitted for brevity.

2.3.3 Turnover, Spreads, and Liquidity Premia

In a model with transaction costs but without constraints, Gerhold et al. [43] pointed out the following connection between the welfare impact of small transaction costs and the turnover implied by the optimal strategy:\footnote{Here, both quantities have to be measured consistently, either focusing on the risky asset (liquidity premium and share turnover) or on the whole market (equivalent safe rate and wealth turnover).}

\[
\left( r + \frac{\mu^2}{2\gamma\sigma^2} \right) - \text{ESR} \sim \frac{3}{4} \varepsilon \text{WeTu} \quad \text{and} \quad \text{LiPr} \sim \frac{3}{4} \varepsilon \text{ShTu}. \tag{2.3.1}
\]

The interpretation is that the unobservable welfare effect of small transaction costs is approximately equal to a product of observables: trading volume, times the bid-ask spread, times a universal constant. This shows that the comparative statics of both quantities coincide, and allows us to estimate liquidity premia from data on trading volume.
In the presence of binding constraints, the asymptotic rates of all involved quantities change. On the contrary, the link (2.3.1) between them remains valid, up to changing the constants. To see this, first note that in the presence of the constraint but without transaction costs, the equivalent safe rate is given by

\[ r + \frac{\mu^2}{2\gamma \sigma^2} \left( \frac{2\pi_{\text{max}}}{\pi_s} - \left( \frac{\pi_{\text{max}}}{\pi_s} \right)^2 \right) - \text{ESR} \]

\[ \sim \gamma \sigma^2 \left( \frac{1}{\gamma} \left( \pi_s - \pi_{\text{max}} \right)^2 \left( 1 - \pi_{\text{max}} \right)^2 \pi_{\text{max}}^2 \right)^{1/2} \varepsilon^{1/2}. \]

In view of the asymptotic expansion for wealth turnover, this shows that the extra impact of transaction costs on the equivalent safe rate remains proportional to wealth turnover times the spread, also in the constrained case:

\[ r + \frac{\mu^2}{2\gamma \sigma^2} \left( \frac{2\pi_{\text{max}}}{\pi_s} - \left( \frac{\pi_{\text{max}}}{\pi_s} \right)^2 \right) - \text{ESR} \sim \varepsilon \text{WeTu}. \]

Similarly, notice that the total liquidity premium \( \text{LiPr} \) in Theorem 2.2.3 can be asymptotically decomposed into the liquidity premia \( \text{LiPr}^C \) and \( \text{LiPr}^T \) required to compensate for the constraints alone and the additional effect of the transaction costs:

\[ \text{LiPr} \sim \mu \left( 1 - \sqrt{\frac{2\pi_{\text{max}}}{\pi_s} - \left( \frac{\pi_{\text{max}}}{\pi_s} \right)^2} + \gamma \sigma^2 \left( \frac{1}{\gamma} \pi_{\text{max}} \left( \pi_s - \pi_{\text{max}} \right)^2 \left( 1 - \pi_{\text{max}} \right)^2 \pi_{\text{max}}^2 \right)^{1/2} \varepsilon^{1/2} \right) \]

\[ := \text{LiPr}^C + \text{LiPr}^T. \]

With this notation, we obtain the following analogue of the second relation in (2.3.1):

\[ \text{LiPr}^T \sim \left( \frac{\pi_{\text{max}}}{2\pi_s - \pi_{\text{max}}} \right)^{1/2} \varepsilon \text{ShTu}. \]

Hence, this result is also robust to the additional portfolio constraints, up to one important caveat. Unlike for wealth turnover and the equivalent safe rate above, the constant linking share turnover and the liquidity premium accrued due to transaction costs depends on the constraints, thereby leading to different comparative statics. Indeed, whereas the leading term of share turnover is always increasing with harder constraints in the absence of leverage, \( \pi_s \in (0, 1) \), the effect on the liquidity premium can be ambiguous due the presence of the extra factor \( \left[ \frac{\pi_{\text{max}}}{2\pi_s - \pi_{\text{max}}} \right]^{1/2} \), which is decreasing with harder constraints. This is in line with the numerical observation of [27] that “the liquidity premium can be higher even though position limits are less binding”. Whereas this may or may not be the case in the absence of leverage, it is in fact the generic situation for a leveraged position \( \pi_s > 1 \) and constraints \( \pi_{\text{max}} \in \left[ 1, \frac{1+2\pi_s}{3} \right) \). These results are illustrated in Figure 2.3, where the liquidity premia \( \text{LiPr}^T \) due to transaction costs are plotted against the spread in the unconstrained case and for two
binding constraints. In the no-leverage regime depicted in the left panel, the constrained liquidity premia dominate their unconstrained counterparts for all levels of transaction costs, in line with the larger asymptotic rate. Moreover, the liquidity premium increases with tighter constraints, in accordance the asymptotic comparative statics. These also predict the correct effect in the leverage case reported in the right panel. Here, $\pi_{\text{max}} \in [1, \frac{1+2\pi^*}{3})$ such that the leading-order terms of the liquidity premia are decreasing with tighter constraints, which matches the numerical results. Analogously as for share turnover, however, the constrained liquidity premium only dominates the unconstrained one for sufficiently low transaction costs and not for arbitrary levels like in the unleveraged case.

Since the total liquidity premium $\text{LiPr}$ measures the joint impact of constraints and transaction costs, it is interesting to discriminate between the relative contributions of these two market frictions. Figure 2.4 depicts the total liquidity premia for the same examples as before. Evidently, the impact of the transaction costs is larger in the leverage case, as in the absence of constraints [43], but the effect is diminished severely as the constraints bind harder. In contrast, the transaction costs have a much bigger influence than in the unconstrained case in the absence of leverage.

### 2.3.4 Applications

#### Selection of Prime Brokers

Consider an investor choosing which prime broker to use to buy a leveraged risky position on margin. Each broker is willing let the investor borrow from him at a
lending rate \( r > 0 \) and up to a leverage constraint \( \pi_{\text{max}} > 1 \),\(^8\) thereby allowing the investor to trade in a market with safe rate \( r \), excess return \( \tilde{\mu} - r \) of the risky asset (\( \tilde{\mu} \) denotes the total return), and leverage constraint \( \pi_{\text{max}} \).

Without transaction costs, the same equivalent safe rate \( \text{esr} \) can be achieved if

\[
 r = \frac{2\pi_{\text{max}}\tilde{\mu} - 2\text{esr} - \pi_{\text{max}}^2 \gamma \sigma^2}{2(\pi_{\text{max}} - 1)}. \tag{2.3.2}
\]

That is, all pairs \((r, \pi_{\text{max}})\) satisfying this relation lie on the same iso-utility curve for the investor, since the effect of having to pay a higher lending rate is precisely offset by the opportunity to borrow a larger amount for investing. This is illustrated in Figure 2.5.

With transaction costs, the investor is no longer indifferent between these different brokers, and our model allows us to assess which combinations are more attractive if the respective rebalancing costs are taken into account. Indeed, for pairs \((r, \pi_{\text{max}})\) satisfying (2.3.2), the equivalent safe rate with transaction costs is given by

\[
 \text{esr} - \left( \frac{\sigma^2}{2} \pi_{\text{max}}^2 (\pi_{\text{max}} - 1)(2\text{esr} - 2\tilde{\mu} - (\pi_{\text{max}} - 2)\pi_{\text{max}} \gamma \sigma^2) \right)^{1/2} \varepsilon^{1/2} + O(\varepsilon).
\]

If \( \pi_{\text{max}} = 1 \), the investor follows a buy-and-hold strategy and the transaction costs have no impact. As the constraint becomes softer, the effect of transaction costs increases, reaches its maximum at a critical level \( \pi_{\text{max}}^c \), and then decreases again, vanishing at the leading order \( \varepsilon^{1/2} \) as the constraint \( \pi_{\text{max}} \) tends towards the

---

\(^8\)Put differently, the broker’s margin requirement is \( 1/\pi_{\text{max}} \).
2.3 Implications

Figure 2.5: Left panel: 10% equivalent safe rate indifference curve without transaction cost plotted against the interest rate $r$ (vertical axis) and the constrained weight $\pi_{max}$ (horizontal axis, the corresponding unconstrained weight is 271%). Right panel: Leading-order loss in equivalent safe rate due to transaction costs $\varepsilon = 1\%$ plotted against the constrained weight $\pi_{max}$ with an interest rate $r$ such that the equivalent safe rate without transaction cost is always 10%. Model parameters are $\bar{\mu} = 8\%$, $\sigma = 16\%$, and risk aversion is $\gamma = 0.8$.

frictionless Merton proportion $\pi_* = (\bar{\mu} - r)/\gamma\sigma^2$. Consequently, for sufficiently hard constraints, the investor prefers tighter constraints and lower lending rates to softer constraints and higher lending rates, but the picture is reversed for sufficiently soft constraints, as illustrated in Figure 2.5.

The above leading-order analysis neglects that the effect of transaction costs does not vanish exactly as the constraint approaches the Merton proportion; only the leading term of order $\varepsilon^{1/2}$ tends to zero. More precisely, it is replaced by a nontrivial term of order $\varepsilon^{2/3}$ in the limit. Hence, portfolios with soft constraints are in fact less attractive than suggested by the above analysis. This is illustrated in Figure 2.6, which compares the exact and leading-order equivalent safe rates that can be obtained with pairs $(r, \pi_{max})$ satisfying (2.3.2). Whereas the approximation recaptures the qualitative properties of the exact quantity rather well, it severely overestimates the attractiveness of brokers with a soft leverage constraint (i.e., a low margin requirement) and a high lending rate. With the exact quantities, it turns out that the investor typically prefers tighter leverage constraints and lower lending rates, as transaction costs make highly leveraged positions relatively less attractive than they appear to be in frictionless markets.

Illiquid Loans and Deposit Rates

The application in the previous section can also be reinterpreted as follows. Consider a bank, who can borrow from its depositors at a safe rate $r$ to provide illiquid (long-term) loans, whose book values are assumed to follow geometric Brownian motion with constant drift $\mu$ and volatility $\sigma$. To limit excessive risk-taking,
regulating authorities restrict the amount of leverage financial institutions are allowed to use by setting minimal capital requirements, which correspond to the portfolio constraints in our model.

This setting allows us to study how the deposit rates offered by the bank react to harder regulatory constraints. The idea is that the bank will try to achieve the same performance (measured in terms of the equivalent safe rate) with the new constraints. Since tightening the constraints reduces the equivalent safe rate, this means that the bank will decrease its deposit rate. In the absence of transaction costs, i.e., for loans that can be liquidated at their book values, the new deposit rate can be obtained using Formula (2.3.2). However, most long-term loans are not very liquid, incurring substantial transaction costs when liquidated prior to maturity. The corresponding change in the deposit rate can in turn be determined by numerically solving for the safe rate in Theorem 2.2.3 that makes the equivalent safe rate with the new constraint coincide with its counterpart for the old constraint and the old deposit rate.

The results for a concrete example are provided in Table 2.1. In the literature, banks are often modeled as risk-neutral, simply maximizing the present value of future cash flows. Therefore, we also use a low risk aversion ($\gamma = 0.1$) here. As the transaction costs incurred when prematurely liquidating long-term loans are substantial (in particular, compared to the bid-ask spreads observed for equities), we report results for a relative bid-ask spread of up to 10%. It turns out that harder regulatory constraints decrease the deposit rates the most if the long-term loans the bank provides are assumed to be perfectly liquid. If illiquidity is taken into account, then highly leveraged positions are less attractive, and a substantially smaller reduction of the deposit rate is required to compensate the bank for the tighter constraints.

Figure 2.6: Exact (solid) and approximate equivalent safe rate (dotted) with transaction costs (vertical axis, left panel: 1%, right panel: 0.1%) plotted against the constrained weight $\pi_{\text{max}}$ with an interest rate $r$ such that the equivalent safe rate without transaction cost is always 10%. Model parameters are $\bar{\mu} = 8\%$, $\sigma = 16\%$, $\pi_{\text{max}} = 271\%$ and risk aversion is $\gamma = 0.8$. 
Table 2.1: Changes in deposit rates due to a decrease of the upper bound of the risky weight from 220% to 180% (third row) or to 150% (fourth row), in order to retain the initial level of the equivalent safe rate. Model parameters are $\bar{\mu} = 0.08$, $\sigma = 16\%$ and risk aversion is $\gamma = 0.1$.

2.4 Heuristics

In this section, we use methods from stochastic control to heuristically derive a candidate solution.

Fix an upper bound $0 < \pi_{max} = \kappa \pi_* \neq 1$ on the investor's risky weight and consider the problem of maximizing the expected power utility $U(x) = x^{1-\gamma}/(1-\gamma)$ from terminal wealth at time $T$. Denote by $V(t, X^0_t, X_t)$ its value function, which is assumed to depend on time as well as the positions $X^0 = \varphi^0 S^0$ in the safe and $X = \varphi S$ in the risky asset, evaluated in terms of the ask price. Then, by Itô's formula and the self-financing condition (2.2.1):

$$dV(t, X^0_t, X_t) = \left( V_t + r X^0_t V_x + (\mu + r) X_t V_y + \frac{\sigma^2}{2} X^2_t V_{yy} \right) dt + \sigma X_t V_y dW_t$$

$$+ S_t (V_y - V_x) d\varphi_t^\uparrow + S_t (1 - \varepsilon) V_x - V_y) d\varphi_t^\downarrow,$$

where the arguments are omitted for brevity. By the martingale optimality principle, this process has to be a supermartingale for any strategy, and a martingale for the optimizer, leading to the HJB equation:

$$V_t + r X^0_t V_x + (\mu + r) X_t V_y + \frac{\sigma^2}{2} X^2_t V_{yy} = 0, \quad \text{if} \quad 1 < \frac{V_x}{V_y} < \frac{1}{1 - \varepsilon}.$$

The homotheticity of the value function and the observation that – in the long-run – the value function should grow exponentially with the horizon at a constant rate $r + \beta$ suggest the following representation:\(^{11}\)

$$V(t, X^0_t, X_t) = (X^0_t)^{1-\gamma} v\left( X_t/X_t^0 \right) e^{-(1-\gamma)(r+\beta)t}.$$  

\(^{11}\)This representation is valid if the position in the safe asset is non-negative at all times.
Setting $z := y/x$ the HJB equation becomes
\[ \frac{\sigma^2}{2} z^2 v''(z) + \mu z v'(z) - (1 - \gamma) \beta v(z) = 0, \quad \text{if} \quad 1 + z < \frac{(1 - \gamma) v(z)}{v'(z)} < \frac{1}{1 - \varepsilon} + z. \]

The set $\{ z : 1 + z < \frac{(1 - \gamma) v(z)}{v'(z)} < \frac{1}{1 - \varepsilon} + z \}$ corresponds to those values of the stock-cash ratio $X/X^0$ for which the optimal strategy does not move, i.e., the no-trade region. To simplify further, assume that it is given by an interval $l < z < u$, where the lower boundary $l$ is an unknown parameter and the upper boundary $u$ coincides with the constraint, $u = \pi_{\max}/(1 - \pi_{\max})$. Then, one obtains the following system with one free boundary:
\begin{align*}
\frac{\sigma^2}{2} z^2 v''(z) + \mu z v'(z) - (1 - \gamma) \beta v(z) &= 0, \quad \text{if} \quad l < z < u, \quad (2.4.1) \\
(1 + l) v'(l) - (1 - \gamma) v(l) &= 0, \quad (2.4.2) \\
(1/(1 - \varepsilon) + u) v'(u) - (1 - \gamma) v(u) &= 0. \quad (2.4.3)
\end{align*}

These conditions do not suffice to identify the solution, since the ODE (2.4.1) is of order two and the conditions (2.4.2) and (2.4.3) can be matched for any choice of the buying boundary $l$. The optimal buying boundary $l$ is the one that additionally satisfies a smooth pasting condition [36], obtained by formally differentiating (2.4.2):
\[ (1 + l) v''(l) + \gamma v'(l) = 0. \quad (2.4.4) \]

Substituting (2.4.4) and (2.4.2) into (2.4.1) yields
\[ -\frac{\sigma^2 \gamma}{2} \left( \frac{l}{1 + l} \right)^2 + \mu \frac{l}{1 + l} - \beta = 0. \quad (2.4.5) \]

The smaller solution of this quadratic equation determines the lower buying boundary:
\[ \pi_- = \frac{l}{1 + l} = \frac{\mu}{\gamma \sigma^2} - \frac{\sqrt{\mu^2 - 2\beta \gamma \sigma^2}}{\gamma \sigma^2}. \]

Set $\pi_-(1 - \lambda) = (1 - \lambda) \kappa \pi_\ast$ for some $\lambda > 0$. Then, the growth rate $\beta$ can be written as
\[ \beta = \frac{\mu^2}{2 \gamma \sigma^2} (1 - (1 - \kappa (1 - \lambda))^2). \]

$\lambda$ is called the gap, because it describes the deviation of the frictional buying boundary and growth rate from their frictionless counterparts. Since the buying boundary is determined by $\lambda$, the above free-boundary value problem becomes

---

In the leverage case $\pi_{\max} > 1$ one has to factor out $-X^0_t$ instead of $X^0_t$, leading to analogous calculations.

12 Without constraints and transaction costs, the investor would hold an even larger risky weight. Therefore it is natural to assume that the upper selling boundary coincides with the highest value compatible with the constraints.

13 Here, “smooth” means $C^2$ across the boundary. Note that this need not hold at the upper selling boundary, since the latter is fixed by the constraints.
a fixed-boundary value problem with free parameter \( \lambda \) in this notation. The substitution
\[
v(z) = e^{(1-\gamma) \int_0^z u^\gamma(\frac{u}{\lambda}(\lambda)) w(y)dy}, \quad \text{i.e., } \quad w(y) = \frac{l(\lambda)e^y u'(l(\lambda)e^y)}{(1-\gamma)v(l(\lambda)e^y)},
\]
in turn reduces it to a Riccati ODE:
\[
0 = w'(x)+(1-\gamma)w(x)^2+(\frac{2u}{\sigma^2}-1)w(x) - \frac{\mu^2}{\sigma^2}(1-(1-\kappa(1-\lambda))^2), \quad x \in [0, \log (u/l(\lambda))]
\]
with boundary conditions
\[
w(0) = \frac{l(\lambda)}{1+I(\lambda)} = (1-\lambda)\pi_{\text{max}}, \quad (2.4.7)
\]
\[
w \left( \log \left( \frac{u}{l(\lambda)} \right) \right) = \frac{u(1-\varepsilon)}{1+u(1-\varepsilon)} = \frac{\pi_{\text{max}}(1-\varepsilon)}{(1-\pi_{\text{max}}) + \pi_{\text{max}}(1-\varepsilon)}, \quad (2.4.8)
\]
where
\[
\frac{u}{l(\lambda)} = \frac{\pi_{\text{max}}/(1-\pi_{\text{max}})}{\pi_-(\lambda)/(1-\pi_-(\lambda))} = \frac{\pi_{\text{max}}/(1-\pi_{\text{max}})}{(1-\lambda)\pi_{\text{max}}/(1-(1-\lambda)\pi_{\text{max}})}. \quad (2.4.9)
\]
Since the Riccati ODE is of order one, the initial condition (2.4.7) uniquely determines a solution \( w(\lambda, \cdot) \) for any choice of \( \lambda \). The correct one is then identified by the terminal condition (2.4.8). Even though the Riccati ODE (2.4.6)-(2.4.7) can be solved explicitly (cf. Lemma 2.5.1), it is not possible to solve for \( \lambda \) by the terminal condition (2.4.8). Even though the Riccati ODE (2.4.6)-(2.4.7) can be solved explicitly (cf. Lemma 2.5.1), it is not possible to solve for \( \lambda \) in closed form. However, the implicit function theorem readily yields a fractional power series expansion in \( \varepsilon \) (cf. Lemma 2.5.2), which in turn immediately provides the asymptotics for the buying boundary \( \pi_{-} = (1-\lambda)\pi_{\text{max}} \) and the growth rate \( \beta = \frac{\mu^2}{2\gamma\sigma^2}(1-(1-\kappa(1-\lambda))^2) \).

### 2.5 Proofs of Theorem 2.2.3 and Theorem 2.2.4

#### 2.5.1 Proof of Theorem 2.2.3

**2.5.1.1 Explicit Formulae and their Properties**

The first step towards a rigorous verification theorem is to determine an explicit expression for the solution of the Riccati ODE (2.4.6) with initial condition (2.4.7), given a sufficiently small \( \lambda > 0 \).

**Lemma 2.5.1.** Let \( 0 < \pi_{\text{max}} = \kappa \pi_{s} \neq 1 \) and define \( w(\lambda, \cdot) \) by
\[
w(\lambda, x) := \begin{cases}
\frac{a(\lambda) \tan(\text{anh}^{-1}((b(\lambda)/a(\lambda))-a(\lambda)x)+\mu/\sigma^2-1/2)}{\gamma-1}, & \text{if } \gamma \in (0, 1) \text{ and } 0 < \pi_{\text{max}} < 1, \text{ or } \gamma > 1 \text{ and } \pi_{\text{max}} > 1, \\
\frac{a(\lambda) \tan(\text{anh}^{-1}((b(\lambda)/a(\lambda))+a(\lambda)x)+\mu/\sigma^2-1/2)}{\gamma-1}, & \text{if } \gamma > 1 \text{ and } \pi_{\text{max}} \in \left[ \frac{1}{2}(1-\frac{\gamma}{\kappa^{2}-\frac{1}{2}}) \right], \frac{1}{2}(\kappa^{2}-(\gamma-1)(2\kappa-1)) \right], \\
\frac{a(\lambda) \cot(\cot^{-1}((b(\lambda)/a(\lambda))-a(\lambda)x)+\mu/\sigma^2-1/2)}{\gamma-1}, & \text{otherwise},
\end{cases}
\]
with
\[ a(\lambda) := \sqrt{\left(\frac{\lambda}{\gamma \sigma^2} - 1\right) - \left(\frac{1}{2} - \frac{\mu}{\sigma^2}\right)^2}, \]
\[ b(\lambda) := \frac{1}{2} - \frac{\mu}{\sigma^2} + (\gamma - 1)\pi_{\max}(1 - \lambda). \]

Then, for a sufficiently small \( \lambda > 0 \), the mapping \( x \mapsto w(\lambda, x) \) is a local solution of
\[ 0 = w'(x) + (1 - \gamma)w(x)^2 + \left( \frac{\mu}{\sigma^2} - 1 \right)w(x) - \frac{\mu^2}{\gamma \sigma^2}(1 - (1 - \kappa(1 - \lambda))^2), \]
\[ w(0) = (1 - \lambda)\pi_{\max}. \quad (2.5.1) \]

Moreover, \( x \mapsto w(\lambda, x) \) is increasing for \( 0 < \pi_{\max} < 1 \) and decreasing for \( \pi_{\max} > 1 \).

Proof. The first part of the assertion is readily verified by taking derivatives. The second follows by inspection of the explicit formulas.

Next, we establish that the crucial constant \( \lambda \), which determines both the no-trade region and the equivalent safe rate, is well-defined.

Lemma 2.5.2. Let \( 0 < \pi_{\max} = \kappa \pi_* \neq 1 \), define \( w(\lambda, \cdot) \) as in Lemma 2.5.1, and set
\[ l(\lambda) = \frac{(1 - \lambda)\pi_{\max}}{1 - (1 - \lambda)\pi_{\max}} \quad \text{and} \quad u = \frac{\pi_{\max}}{1 - \pi_{\max}}. \quad (2.5.2) \]

Then, for a sufficiently small \( \varepsilon \), there exists a unique solution \( \lambda \) of
\[ w(\lambda, \log (u/l(\lambda))) = \frac{\pi_{\max}(1 - \varepsilon)}{(1 - \pi_{\max}) + \pi_{\max}(1 - \varepsilon)} =: w_. \quad (2.5.3) \]

As \( \varepsilon \downarrow 0 \), it has the asymptotic expansion
\[ \lambda = \left( \frac{1}{\gamma \left( 1 - \kappa \right)} \frac{(1 - \pi_{\max})^2}{\pi_{\max}} \right)^{1/2} \varepsilon^{1/2} + O(\varepsilon). \]

Proof. With minor modifications, this follows as in [43, Lemma B.2]

Henceforth, \( \lambda \) denotes the quantity from Lemma 2.5.2, and we omit the \( \lambda \)-dependence of \( a = a(\lambda), b = b(\lambda), l = l(\lambda), \) and \( w(x) = w(\lambda, x) \).

Corollary 2.5.3. Let \( 0 < \pi_{\max} \neq 1 \) and suppose \( \varepsilon \) is sufficiently small. Then, in all three cases,
\[ w'(0) = \pi_-(1 - \pi_-) \quad (2.5.4) \]
\[ w'(\log (u/l)) \leq w_(1 - w_). \quad (2.5.5) \]

Proof. The assertions follow from the ODE for \( w \) and its boundary conditions in Lemma 2.5.1 and Lemma 2.5.2.
2.5 Proofs of Theorem 2.2.3 and Theorem 2.2.4

2.5.1.2 Shadow Prices and Verification

A key idea for the proof of our verification theorem is to replace the original bid and ask prices by a single fictitious ‘shadow price’ \( \tilde{S} \) evolving within the bid-ask spread, which admits an optimal policy that is feasible (and hence also optimal) in the original market with transaction costs, too. An approach of this kind was first used in [69], and has been utilized in the present setting modulo constraints by [43].

**Definition 2.5.4.** A shadow price is a process \( \tilde{S} \) lying within the bid-ask spread \([1 - \varepsilon)S, S]\), such that there exists a corresponding long-run optimal strategy \((\varphi^0, \varphi)\) of finite variation that satisfies the portfolio constraint (2.2.2) and only entails buying or selling the risky asset when \( \tilde{S} \) equals the buying or selling price, respectively.

The original constraints can be translated as follows:

**Remark 2.5.5.** Let \( \tilde{S} \) be a price process evolving within the bid ask spread \([(1 - \varepsilon)S, S]\). Then if a strategy \((\varphi^0, \varphi)\) satisfies the original portfolio constraint \(\pi_t \leq \pi_{\text{max}}\) on the risky weight computed with the ask price \(S\), it also satisfies the following constraint on the risky weight computed with \(\tilde{S}\):

\[
\tilde{\pi}_t \leq \tilde{\pi}_{\text{max}} := \begin{cases} 
\pi_{\text{max}}, & \text{if } \pi_{\text{max}} \leq 1, \\
\pi_{\text{max}}(1 - \varepsilon) + \pi_{\text{max}}(1 - \varepsilon), & \text{otherwise.}
\end{cases}
\]  

(2.5.6)

The construction successfully used in [88, 9, 60] suggests that the discounted shadow price can be constructed as the marginal rate of substitution of risky for safe assets for the optimal investor, i.e., as the ratio of the partial derivatives of the value function with respect to the numbers of shares in the risky and safe asset, respectively:

\[
\frac{\dot{S}_t}{S_0^\varphi} = \frac{\partial_{\varphi^0} V(t, X_t^0, X_t)}{\partial_{\varphi^0} V(t, X_0^0, X_t)}.
\]

With the candidate value function derived in the above section, this leads to the candidate shadow price

\[
\tilde{S}_t = S_t \frac{w(\log(X_t/X_t^0))}{X_t/X_t^0[1 - w(\log(X_t/(X_t^0)))]} = S_t \frac{w(Y_t)}{le^{Y_t}(1 - w(Y_t))},
\]

(2.5.7)

where \(e^{Y_t} = (X_t/X_t^0)\) is the ratio between the risky and safe positions at the ask price \(S_t\), centered at the buying boundary \(l = \frac{(1-\lambda)\pi_{\text{max}}}{1-(1-\lambda)\pi_{\text{max}}}\). In view of the above heuristics, the stock-cash ratio \(X/X_0\) should remain within the no-trade region \([l, u]\); consequently, \(Y\) should take values in \([0, \log(u/l)]\) (or \([\log(u/l), 0]\), if \(\pi_{\text{max}} > 1\)). In the interior of this interval, the number of risky assets should remain constant, so that the dynamics of \(Y = \log(\varphi/(t\varphi^0)) + \log S/S_0\) coincide with those of the Brownian motion \(\log S/S_0\), which needs to be reflected at the boundaries to remain in \([0, \log(u/l)]\).
These heuristic arguments motivate us to define the process $Y$ as Brownian motion with instantaneous reflection at 0 and $\log(u/l)$:

$$dY_t = (\mu - \sigma^2/2)dt + \sigma dW_t + dL_t - dU_t, \quad (2.5.8)$$

where the local time processes $L$ and $U$ are adapted, continuous, non-decreasing (or non-increasing, if $\pi_{\text{max}} > 1$) and only increase (or decrease, if $\pi_{\text{max}} > 1$) on the sets $\{Y_t = 0\}$ and $\{Y_t = \log(u/l)\}$, respectively. Starting from this process, whose existence dates back to a classical result of [115], the process $\tilde{S}$ can then be defined in accordance with (2.5.7):

**Lemma 2.5.6.** Define

$$y = \begin{cases} 0, & \text{if } l\xi^0 S^0_0 \geq \xi S_0, \\ \log(u/l), & \text{if } u\xi^0 S^0_0 \leq \xi S_0, \\ \log[\xi S_0/(\xi^0 S^0_0 l)], & \text{otherwise}, \end{cases} \quad (2.5.9)$$

and let $Y$ be defined as in (2.5.8), starting at $Y_0 = y$. Then, $\tilde{S} = S_{l \exp(1-\gamma(\log(u/l)))}$, with $w$ as in Lemma 2.5.1, has the dynamics

$$\frac{d\tilde{S}(Y_t)}{\tilde{S}(Y_t)} = (\tilde{\mu}(Y_t) + r)dt + \tilde{\sigma}(Y_t)dW_t + \left(1 - \frac{w'(y)}{(1-w)w}\left(\log\left(\frac{u}{l}\right)\right)\right) dU_t,$$

where $\tilde{\mu}(\cdot)$ and $\tilde{\sigma}(\cdot)$ are given by

$$\tilde{\mu}(y) = \frac{\sigma^2 w'(y)}{w(y)(1-w(y))}\left(\frac{w'(y)}{1-w(y)} - (1-\gamma)w(y)\right), \quad \tilde{\sigma}(y) = \frac{\sigma w'(y)}{w(y)(1-w(y))}.$$

Moreover, the process $\tilde{S}$ takes values within the bid-ask spread $[(1-\varepsilon)S, S]$.

Note that the first two cases in (2.5.9) arise if the initial stock-cash ratio $\xi S_0/(\xi^0 S^0_0)$ lies outside of the interval $[l, u]$. Then, a jump from the initial position $(\varphi^{-}, \varphi^{-}) = (\xi^0, \xi)$ to the nearest boundary value of $[l, u]$ is required. This transfer necessitates the purchase or sale of the risky asset, and hence the initial price $\tilde{S}_0$ is defined to match the buying or selling price, respectively, of the risky asset.

**Proof.** The dynamics of $\tilde{S}$ result from Itô’s formula, the dynamics of $Y$, and the identity

$$w''(y) = -2(1-\gamma)w'(y)w(y) - (2\mu/\sigma^2 - 1)w'(y), \quad (2.5.10)$$

which is a direct consequence of Lemma 2.5.1. In addition, the boundary conditions for $w$ and $w'$ imply that

$$w''(0) - w'(0) + 2w(0)w'(0) = 2w'(0)(\gamma w(0) - \frac{\mu}{\sigma^2}) = 2\pi_{\text{max}}(1-\lambda)(1-\pi_{\text{max}}(1-\lambda))\frac{\mu}{\sigma^2}(\kappa(1-\lambda) - 1)$$

is negative (or positive, if $\pi_{\text{max}} > 1$). Thus, a comparison argument yields that the derivative of the function $\eta : y \mapsto \frac{w(y)}{l \exp(1-\gamma(\log(u/l)))}$ is negative (or positive, if $\pi_{\text{max}} > 1$). Taking into account $\frac{w(0)}{l(1-w(0))} = 1$ and $\frac{w(0)}{u(1-(\log(u/l)))} = 1 - \varepsilon$ completes the proof. \qed
2.5 Proofs of Theorem 2.2.3 and Theorem 2.2.4 29

Unlike in the absence of constraints, the dynamics of \( \tilde{S} \) involve a singular part, such that the shadow market is no longer arbitrage-free. Indeed, whenever \( \tilde{S} \) hits the lower bid price \((1 - \varepsilon)S\) it is reflected upwards, such that one can make a riskless profit by buying risky assets and immediately selling them after the reflection has taken place.

Due to the constraints (2.2.2), however, such arbitrage opportunities cannot be scaled arbitrarily. Consequently, there exists a discount factor that turns all admissible wealth processes into supermartingales (compare [23]):

**Lemma 2.5.7.** For a fixed time horizon \( T \), denote by \( \tilde{X}_T^\psi \) the shadow payoff of an admissible strategy \((\psi^0, \psi)\) in the frictionless shadow market \((\tilde{S}^0, \tilde{S})\), satisfying the constraint (2.2.2) and hence also the shadow constraint \( \tilde{\pi}^\psi \leq \tilde{\pi}_{\text{max}} \) from Remark 2.5.5. Define the process \( \tilde{M} \) by

\[
\tilde{M}_t := e^{-\tilde{\rho} t} \mathcal{E} \left( - \int_0^t \tilde{\mu}(Y_u) dW_u \right) e^{-\tilde{\rho} \max \left( 1 - \frac{\tilde{\omega}(\log(u/l))}{(1-w)\omega(u/l)} \right) U_t},
\]

with the local time process \( U \) from (2.5.8). Then \( \tilde{M} \) is a discount factor:

\[
\mathbb{E}[\tilde{X}_T^\psi \tilde{N}_T] \leq \tilde{X}_0^\psi.
\]

**Proof.** First, notice that \( \tilde{\mu}, \tilde{\sigma} \) and \( w \) are functions of \( Y \), but the argument is omitted throughout to ease notation. Inserting the dynamics of \( \tilde{S} \) yields

\[
\tilde{X}_T^\psi = \tilde{X}_0^\psi \mathcal{E} \left( \int_0^T \left( r + \tilde{\pi}^\psi_T \tilde{\mu} \right) dt + \int_0^T \tilde{\pi}^\psi_T \tilde{\sigma} dW_t \right) e^{\int_0^T \tilde{\pi}^\psi_T \left( 1 - \frac{\tilde{\omega}(\log(u/l))}{(1-w)\omega(u/l)} \right) dU_t},
\]

where \( \tilde{\pi}^\psi \) denotes the risky weight in terms of \( \tilde{S} \). In view of Remark 2.5.5, we have \( \tilde{\pi}^\psi \leq \tilde{\pi}_{\text{max}} \). Furthermore, the identity (2.5.5) implies \( 1 - \frac{\tilde{\omega}(\log(u/l))}{(1-w)\omega(u/l)} \geq 0 \) (or \( \leq 0 \), if \( \pi_{\text{max}} \) > 1) on the set \( \{ y = \log(u/l) \} \) where \( U \) increases (or decreases, if \( \pi_{\text{max}} > 1 \)). Hence,

\[
\mathbb{E}[\tilde{X}_T^\psi \tilde{M}_T] = \mathbb{E} \left[ \tilde{X}_0^\psi \mathcal{E} \left( \int_0^T \left( r + \tilde{\pi}^\psi_T \tilde{\mu} \right) dt + \int_0^T \tilde{\pi}^\psi_T \tilde{\sigma} dW_t \right) e^{\int_0^T \tilde{\pi}^\psi_T \left( 1 - \frac{\tilde{\omega}(\log(u/l))}{(1-w)\omega(u/l)} \right) dU_t} \right]
\]

\[
\leq \mathbb{E} \left[ \tilde{X}_0^\psi \mathcal{E} \left( \int_0^T \left( \tilde{\pi}^\psi_T \tilde{\sigma} - \frac{\tilde{\mu}}{\tilde{\sigma}} \right) dW_t \right) \right] \leq \tilde{X}_0^\psi,
\]

where we have used for the last inequality that the positive local martingale \( \mathcal{E}(\int_0^T (\tilde{\pi}^\psi_T \tilde{\sigma} - \frac{\tilde{\mu}}{\tilde{\sigma}}) dW_t) \) is a supermartingale. \( \square \)

If the candidate process \( \tilde{S} \) from Lemma 2.5.6 is indeed a shadow price, then the optimal numbers of safe and risky assets should be the same as in the original market with transaction costs. By definition of \( \tilde{S} \), the candidate optimal strategy derived heuristically in Section 2.4 therefore leads to the following candidate for the long-run optimal risky weight in the shadow market:

\[
\tilde{\pi}(Y_t) = \frac{\bar{\varphi}_i \tilde{S}_t}{\bar{\varphi}_i^{0} S_t^0 + \bar{\varphi}_i \tilde{S}_t} = \frac{\varphi_i S_t}{\bar{\varphi}_i^{0} S_t^0 + \varphi_i S_t} = \frac{w(Y_t)}{1 - w(Y_t)},
\]

where

\[
\bar{\varphi}_i^{0} S_t^0 + \bar{\varphi}_i \tilde{S}_t = \frac{w(Y_t)}{1 - w(Y_t)}.
\]
To show that this risky weight is indeed long-run optimal for \( \tilde{S} \), we first establish the following finite-horizon identities in analogy to the frictionless case [56]:

**Lemma 2.5.8.** For a fixed time horizon \( T > 0 \), let \( \beta = \frac{\kappa^2}{2\sigma^2}(1 - (1 - \kappa(1 - \lambda))^2) \) and let the function \( w \) be defined as in Lemma 2.5.1. Then, for the shadow payoff \( \tilde{X}_T \) corresponding to the policy \( \tilde{\pi}(Y) = w(Y) \) and the shadow discount factor \( \tilde{M}_T \) introduced in Lemma 2.5.7, the following identities hold true:

\[
\begin{align*}
\mathbb{E}[\tilde{X}_T^{-1}] &= \tilde{X}_0^{-1} \mathbb{E}[e^{(1-\gamma)(r+\beta)\tau}] \mathbb{E}[e^{(1-\gamma)\tilde{\pi}(Y_T)-\tilde{\pi}(Y_0)}], \\
\mathbb{E}[\tilde{M}_T^{-1}] &= e^{(1-\gamma)(r+\beta)\tau} \mathbb{E}[e^{(1-\gamma)\tilde{\pi}(Y_T)-\tilde{\pi}(Y_0)}],
\end{align*}
\]

(2.5.11)

(2.5.12)

where \( \tilde{\pi}(y) := \int_0^y \left( \frac{w'(z)}{1-w(z)} - w(z) \right) dz \) and \( \tilde{\mathbb{E}}[\cdot] \) denotes the expectation with respect to the myopic probability \( \tilde{\mathbb{P}} \), defined by

\[
\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} = \mathbb{E} \left( \int_0^T \left( -\tilde{\mu}(Y_t) + \tilde{\sigma}(Y_t) \tilde{\pi}(Y_t) \right) dW_t \right).
\]

**Proof.** To ease the notation, we again omit the argument \( Y \) of the functions \( \tilde{\mu}, \tilde{\sigma}, \tilde{\pi} \) and \( w \). To obtain (2.5.11), notice that

\[
\tilde{X}_T^{-1} = \tilde{X}_0^{-1} e^{(1-\gamma)\int_0^T (r+\tilde{\mu}w - \frac{\kappa}{2} w^2) dt + (1-\gamma) \int_0^T \tilde{\sigma}wdW_t + \int_0^T (1-\gamma)(w - \frac{w'}{1-w}) \tilde{U}_t}
\]

Hence,

\[
\tilde{X}_T^{-1} = \tilde{X}_0^{-1} \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} \int_0^T \left[ e^{(1-\gamma)(r+\tilde{\mu}w - \frac{\kappa}{2} w^2) t + \frac{1}{2} (-\tilde{\sigma} + \tilde{\pi}w)^2} \right] dt
\]

\[
\times \left[ e^{T} \left[ (1-\gamma)\tilde{\sigma}w - (\frac{\kappa}{2} + \tilde{\pi}w) \right] dW_t + \int_0^T (1-\gamma)(w - \frac{w'}{1-w}) d\tilde{U}_t \right].
\]

Inserting the definitions of \( \tilde{\mu} \) and \( \tilde{\sigma} \), we simplify the second integrand to \((1-\gamma)\sigma((\frac{w'}{1-w}) - w)\). Similarly, the first integrand reduces to \((1-\gamma)(r + \frac{\kappa}{2} \left( \frac{w'}{1-w} \right)^2 - (1-\gamma)\sigma^2 w (\frac{w'}{1-w}) + (1-\gamma)\frac{\kappa}{2} w^2 \) \). In summary:

\[
\tilde{X}_T^{-1} = \tilde{X}_0^{-1} \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} \int_0^T \left[ e^{(1-\gamma)\int_0^T (r + \frac{\kappa}{2} \left( \frac{w'}{1-w} \right)^2 - (1-\gamma)\sigma^2 w (\frac{w'}{1-w}) + (1-\gamma)\frac{\kappa}{2} w^2 dt + \int_0^T (1-\gamma)(w - \frac{w'}{1-w}) d\tilde{U}_t \right]
\]

(2.5.13)

Itô’s formula and the boundary conditions for \( w \) imply that

\[
\tilde{\pi}(Y_T) - \tilde{\pi}(Y_0) = \int_0^T \left( \frac{w'}{1-w} - w \right) \sigma dW_t - \int_0^T \left( \frac{w'}{1-w} - w \right) d\tilde{U}_t
\]

\[
+ \int_0^T \left( \mu - \frac{\sigma^2}{2} \right) \left( \frac{w'}{1-w} - w \right) dt
\]

\[
+ \int_0^T \frac{\sigma^2}{2} \left( \frac{w''(1-w) + w^2}{(1-w)^2} - w' \right) dt.
\]

(2.5.14)
Substituting the second derivative $w''$ according to the equation (2.5.10) and using this identity to replace the stochastic integral in (2.5.13) yields

$$\tilde{X}_{T}^{1-\gamma} = \tilde{X}_{0}^{1-\gamma} \frac{d\tilde{P}}{dP} e^{(1-\gamma)\int_{0}^{T} (r+\frac{1}{2}w'\gamma+(1-\gamma)\frac{\sigma^2}{2}w^2+(\mu-\frac{\sigma^2}{2})w)dt} e^{(1-\gamma)(\tilde{q}(Y_T)-\tilde{q}(Y_0))}.$$

Thus, the first equality results from taking expectations on both sides and using the ODE for $w$ (cf. Lemma 2.5.1).

Similarly, plugging in the definitions of $\tilde{\mu}$ and $\tilde{\sigma}$, the (shadow) discount factor $\tilde{M}_T$ and the myopic probability measure $\tilde{P}$ satisfy

$$\tilde{M}_{T}^{1-\frac{1}{\gamma}} = e^{\frac{1-\gamma}{\gamma} \int_{0}^{T} \frac{\tilde{\mu}}{\gamma} dw_t + \frac{1-\gamma}{\gamma} \int_{0}^{T} \left(r+\frac{\tilde{\sigma}^2}{2}\right) dt + \frac{1-\gamma}{\gamma} \int_{0}^{T} \left(w-\frac{w'}{\gamma}w\right) dt}.$$

Substituting again the stochastic integral with equation (2.5.14) and (2.5.10) we obtain

$$\tilde{M}_{T}^{1-\frac{1}{\gamma}} = \frac{d\tilde{P}}{dP} e^{\frac{1-\gamma}{\gamma} \int_{0}^{T} \left(r+\frac{\sigma^2}{2}\gamma+(1-\gamma)\frac{\sigma^2}{2}w^2+(\mu-\frac{\sigma^2}{2})w\right) dt} e^{\frac{1-\gamma}{\gamma}(\tilde{q}(Y_T)-\tilde{q}(Y_0))}.$$

Finally, the second equality results from taking the expectations on both sides, raising it to the power of $\gamma$, and using the ODE for $w$ from Lemma 2.5.1. \(\square\)

With the finite-horizon identities at hand, it is now straightforward to establish the long-run optimality of the policy $\tilde{\pi}(Y) = w(Y)$ in the shadow market. Note that this strategy does not take advantage of the available arbitrage opportunities: It already attains the portfolio constraint whenever the singular component of $\tilde{S}$ acts, hence the constraints prevent it from further increasing its risky position to profit from this.\(^{14}\)

**Lemma 2.5.9.** Let $0 < \pi_{\text{max}} \neq 1$ and let $w$ be defined as in Lemma 2.5.1. Then, the policy $\tilde{\pi}(Y) = w(Y)$ is long-run optimal in the shadow market $(S^0, \tilde{S})$ with equivalent safe rate $r + \beta$, where $\beta$ is specified in Lemma 2.5.8. For $t \geq 0$, the corresponding wealth process and the numbers of safe and risky assets are given

\(^{14}\)A similar example of a market with arbitrage opportunities for which the optimal policy does not exploit these can be found in [47, Example 5.1]. Whereas the latter is somewhat pathological, this situation arises naturally in the present context when looking for an equivalent frictionless shadow market.
Itô’s formula, the boundary conditions for $w$, and Equation (2.5.10) imply
\[
dw(Y_t) = -(1 - \gamma) \sigma^2 w'(Y_t)w(Y_t)dt + \sigma w'(Y_t) dW_t + w'(Y_t)(dL_t - dU_t).
\]
Integrating $\varphi = w(Y)\tilde{X}/\tilde{S}$ by parts twice, using the dynamics of $w(Y), \tilde{X}$ and $\tilde{S}$, and simplifying yields
\[
d\varphi_t/\varphi_t = w'(Y_t)/w(Y_t)dL_t - (1 - w(Y_t))dU_t.
\]
Since $L_t$ and $U_t$ only increase (or decrease, if $\pi_{\text{max}} > 1$) on $\{Y_t = 0\}$ and $\{Y_t = \log(u/l)\}$, respectively, the assertion then follows from the boundary conditions for $w$ and identity (2.5.4).

Since the shadow price takes values in the bid-ask spread, it allows the investor to trade at more favorable prices than in the original market with transaction costs. As the optimal strategy $(\varphi^0, \varphi)$ only entails buying or selling the risky asset when $\tilde{S}$ coincides with the ask or bid price, respectively, it is also feasible with transaction costs, and therefore optimal with the same growth rate.

**Lemma 2.5.11.** For sufficiently small $\varepsilon$, the policy $(\varphi^0, \varphi)$ from Lemma 2.5.9 is also long-run optimal in the original market with transaction costs, with the same equivalent safe rate $r + \beta$.

**Proof.** This follows along the lines of [43, Proposition C.5].

### 2.5.1.3 Trading Volume

The formulas for share and wealth turnover follow from [43, Lemma D.2] along the lines of [43, Corollary D.3]. Plugging the asymptotics for the gap $\lambda$ into these explicit formulas yields the corresponding expansions for share and wealth turnover.

### 2.5.2 Proof of Theorem 2.2.4

Finally we show that, like in the absence of constraints, the stationary long-run optimal policy is also approximately optimal for any finite horizon $T > 0$.

**Proof of Theorem 2.2.4.** For a fixed time horizon $T$ let $(\psi^0, \psi)$ be any admissible strategy. The liquidation value $\Xi^\psi_T$ of this strategy is smaller than the corresponding shadow payoff $\tilde{X}_T^\psi$, i.e., $\Xi^\psi_T \leq \tilde{X}_T^\psi = \psi^0_0 + \psi_0\tilde{S}_0 + \int_0^T \psi_s d\tilde{S}_s$. Thus, the standard duality bound for power utility as in [56, Lemma 5] and the second equality (2.5.12) in Lemma 2.5.8 imply that
\[
\frac{1}{(1 - \gamma)T} \log \mathbb{E}[(\Xi^\psi_T)^{1-\gamma}] \leq r + \beta + \frac{1}{T} \log (\psi^0_0 + \psi_0 - S_0) + \frac{\gamma}{(1 - \gamma)T} \log \mathbb{E}[e^{(\frac{1}{\gamma} - 1)(\tilde{q}(Y_T) - \tilde{q}(Y_0))}]. \tag{2.5.15}
\]
For the long-run optimal strategy $(\varphi^0, \varphi)$ we have $\Xi^\varphi_T \geq (1 - \pi_{\text{max}}\varepsilon/(1 - \varepsilon))\tilde{X}_T$. 


Hence, the first equality (2.5.11) in Lemma 2.5.8 yields
\[
\frac{1}{(1-\gamma)T} \log \mathbb{E}[(\Xi_T^\gamma)^{1-\gamma}] \geq r + \beta + \frac{1}{T} \log (\varphi_0^{\gamma} + \varphi_0 - \hat{S}_0) \\
+ \frac{1}{(1-\gamma)T} \log \mathbb{E}[e^{(1-\gamma)(\tilde{q}(Y_T) - \tilde{q}(Y_0))}] \\
+ \frac{1}{T} \log \left( 1 - \frac{\epsilon}{1 - \epsilon} \pi_{\text{max}} \right). \tag{2.5.16}
\]

For \( \epsilon \downarrow 0 \), we have \( \log (1 - \pi_{\text{max}}\epsilon/(1-\epsilon)) = \mathcal{O}(\epsilon) \) and \( \log (\varphi_0^{\gamma} + \varphi_0 - \hat{S}_0) \geq \log (\varphi_0^{\gamma} + \varphi_0 - S_0) + \mathcal{O}(\epsilon) \) because \( \hat{S} \in [(1-\epsilon)S, S] \). Moreover, since \( \lambda = \mathcal{O}(\epsilon^{1/2}) \) and \( \log (u/l) = \mathcal{O}(\epsilon^{1/2}) \), the identities (2.5.4), (2.5.10) and Taylor expansion yield
\[
\frac{w'(z)}{1 - w(z)} - w(z) = \frac{\kappa^2 \mu^2}{\gamma \sigma^2} \left( 1 - \frac{1}{\kappa} \right) \epsilon^{1/2} + \mathcal{O}(\epsilon).
\]

Therefore, \( \tilde{q}(y) = \int_y^0 \frac{w'(z)}{1 - w(z)} - w(z)dz \) is also of order \( \mathcal{O}(\epsilon) \) for \( y \leq |\log(u/l)| \), which completes the proof. \( \square \)
Chapter 3

Transaction Costs and Linear Price Impact\(^1\)

3.1 Introduction

Proportional transaction costs are ubiquitous even in the most liquid financial markets in the form of bid-spreads. For large institutional investors, the price impact of their trades also is a key concern.\(^2\) As a result, both frictions have generated voluminous literatures that analyze how to balance the gains and costs of portfolio rebalancing optimally. Bid-ask spreads lead to trading costs linear in the amounts transacted. Then, it is optimal to refrain from trading while the uncontrolled portfolio lies inside some “no-trade region” around the frictionless optimum; once its boundaries are reached, one performs the minimal amount of rebalancing necessary to remain inside.\(^3\) In contrast, linear price impact leads to quadratic trading costs, which are less severe for small trades but become prohibitively expensive for larger orders. As a result, optimal policies typically prescribe rebalancing at all times but at a finite absolutely continuous rate, in contrast to the singular controls used with proportional transaction costs.\(^4\)

All of the extant literature studies either linear or quadratic trading costs. This chapter fills this gap by analyzing the joint impact of proportional transaction costs and linear price impact on portfolio rebalancing.

To make the model tractable, we focus (as in Gerhold et al. [43] and Guasoni and Weber [57] for proportional and quadratic costs, respectively, treated separately) on a single risky asset with constant investment opportunities, and a

\(^1\)This chapter is based on Liu, Muhle-Karbe and Weber [87].

\(^2\)For small private investors, fixed commissions, levied on each trade regardless of its size, are also crucial (cf., e.g., [79] and the references therein). Yet, for large portfolios, their influence becomes negligible [5], and we disregard them in the present chapter.

\(^3\)See [91, 22, 31, 37, 113, 29] for models with constant investment opportunities, [113, 62, 14, 43] for corresponding explicit asymptotic formulas, and [94, 117, 107, 71, 68] for recent extensions to more general settings.

\(^4\)Price impact has been studied extensively in the optimal execution literature, see, e.g., [12, 3, 111]. Recently, increasing attention has also been devoted to its influence on dynamic portfolio choice [42, 41, 4, 57, 21, 101].
representative agent with constant relative risk aversion and a long horizon. The representative investor’s wealth serves as a proxy for a multiple of market capitalization, and price impact is assumed to be inversely proportional to latter. That is, a trade of a given size has a smaller impact as markets grow, compare the discussion in Guasoni and Weber [57]. The present study builds on the results of Guasoni and Weber [57] by adding a bid-ask spread. This price impact model leads to stable long-term behavior. The most popular alternative – constant price impact as in [12, 90, 3] – is well-suited to the short-horizons optimal execution problems these authors have in mind. However, it leads to degenerate results in the long run, as growing fund values eventually make rebalancing prohibitively expensive. To bridge the gap between the different models, we argue informally in Section 3.6 that, asymptotically for small costs, our results are robust with respect to the specification of the price impact function.

The optimal policy in the presence of both frictions turns out to be of the following form. As with proportional transaction costs, there is a no-trade region, where it is optimal to simply hold the current portfolio. Once its boundaries are breached, price impact rules out singular controls; hence, one instead starts trading at some finite rate so as to steer the portfolio back to the no-trade region. Since this policy does not allow us to keep the portfolio uniformly close to the frictionless target, trading starts earlier than in a model with only proportional costs, i.e., the width of the no-trade region is decreased by the additional price impact. On the other hand, since there are now two frictions contributing to the total rebalancing cost, the total trading rate is always lower than in a model that only takes into account price impact. We prove a rigorous verification theorem that identifies the trading boundaries, the trading rate, and the associated welfare through the solution of a nonlinear free-boundary value problem that can be solved numerically. To ease implementation, we also present formal asymptotics for small linear and quadratic costs, which reduce the computation of the optimal policy and welfare to finding the root of a scalar function. With two competing frictions, it is particularly important to assess the quality of the asymptotic approximations, since it is not clear a priori whether the matched rescaling leads to accurate results. Our exact formulas allow us to do this, and show that the small-cost approximations perform very well.

Similarly as for proportional costs [71] and price impact [101] considered separately, our formal asymptotics show that the optimal policy is robust to the form of the trading cost. To wit, in Section 3.6, we consider general price impact functions depending on the large investor's wealth and an additional exogenous state variable (say, the wealth of other traders) in an arbitrary manner. Then, the same asymptotic trading rate as in our baseline model obtains, substituting the current value of the trading cost at each point in time.

The remainder of the chapter is organized as follows. Section 3.2 describes the model. Our main results are presented in Section 3.3. Subsequently, we illustrate them with some numerical examples. Section 3.5 contains a heuristic derivation

\[5\] This is similar in complexity to the asymptotics for small fixed and proportional costs studied by [79].
of our main results, that is complemented by formal asymptotics developed in Section 3.6. The rigorous proofs of our main results are collected in Section 3.7.

3.2 Model

Consider a market consisting of one safe asset normalized to one,⁶ and one risky asset, whose mid price $S_t$ follows geometric Brownian motion:

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t.$$  \hfill (3.2.1)

Here, $(W_t)_{t \geq 0}$ is a standard one-dimensional Brownian motion, $\mu > 0$ is the expected excess return, and $\sigma > 0$ is the volatility. Trades are not settled at the idealized best quote $S_t$. Instead, sales only earn lower bid prices, whereas purchases are charged higher ask prices. Moreover, trading large positions quickly moves prices further in an adverse direction. To wit, the average execution price for trading $\Delta \theta$ shares over a time interval $\Delta t$ is

$$S_t \left( 1 + \varepsilon \text{sgn}(\Delta \theta) + \lambda \frac{S_t \Delta \theta}{X_t \Delta t} \right).$$  \hfill (3.2.2)

Here, the first term corresponds to a relative bid-ask spread $\varepsilon$, i.e., a higher ask price $(1 + \varepsilon)S_t$ for purchases and a lower bid price $(1 - \varepsilon)S_t$ for sales, respectively. The second term describes the additional (relative) price impact of large trades executed quickly. This price impact is proportional to the monetary trading rate $S_t \Delta \theta / \Delta t$, and inversely proportional to market capitalization, which is proxied by the representative investor’s wealth $X_t$.⁷ The constant of proportionality $\lambda$ in turn quantifies the market’s limited liquidity; put differently, $1/\lambda$ measures market “depth”. For $\varepsilon, \lambda \to 0$, one recovers the classical frictionless case, where arbitrary amounts $\Delta \theta$ can be purchased or sold at the mid price for $S_t \Delta \theta$. Nontrivial bid-ask spreads ($\varepsilon > 0$) and finite market depth ($\lambda > 0$) lead to additional linear and quadratic trading costs, respectively. Specifically, with both frictions, the execution cost of trading $\Delta \theta$ shares over a time interval $\Delta t$ is given by⁸

$$S_t \Delta \theta + \varepsilon S_t \left| \frac{\Delta \theta}{\Delta t} \right| \Delta t + \lambda \frac{S_t^2 \Delta \theta^2}{X_t \Delta t^2} \Delta t.$$  \hfill (3.2.3)

⁶That is, the interest rate $r$ is set equal to zero.

⁷See [57] for more details on this price impact model and the related literature.

⁸Note the price impact is purely temporary in our model, in that no trade influences the subsequent ones. There is a large literature on optimal execution with persistent price impact, that only wears off gradually after the completion of each trade (cf. [106] as well as many more recent studies). Since we are working on a much longer time scale than in this literature, we abstract from this issue and instead suppose that the temporary and persistent impact costs generated by various “sub-trades” on a finer “execution time-grid” are all aggregated into our price impact cost. Indeed, suppose each infinitesimal sub-trade is executed in the setting of Obizhaeva and Wang [106]. Then, the expected execution costs are of the same linear-quadratic form as in our model.
For tractability, we now pass to the continuous-time limit. Denote by \( \theta_t \) the number of risky shares the investor holds at time \( t \), and replace \( \Delta \theta/\Delta t \) in (3.2.3) with \( \dot{\theta}_t := \lim_{h \downarrow 0} \frac{\theta_{t+h} - \theta_t}{h} \). Then, the investor’s cash position \( C_t = X_t - S_t \theta_t \) evolves as

\[
dC_t = -S_t d\theta_t - \varepsilon S_t |\dot{\theta}_t| dt - \frac{\sigma^2 \dot{\theta}_t^2}{X_t} dt.
\]

Write \( u_t := \dot{\theta}_t S_t / X_t \) for the wealth turnover at time \( t \). With this notation, a straightforward application of Itô’s formula (cf. [57, Lemma 8]) shows that the corresponding wealth process \( X_t := \theta_t S_t + C_t \) and risky weight \( Y_t := \theta_t S_t / X_t \) have the following dynamics:

\[
dX_t = Y_t(\mu dt + \sigma dW_t) - \varepsilon |u_t| dt - \lambda u_t^2 dt,
\]

\[
dY_t = (Y_t(1 - Y_t)(\mu - Y_t \sigma^2) + u_t + \varepsilon |u_t| Y_t + \lambda Y_t u_t^2) dt + Y_t(1 - Y_t) \sigma dW_t.
\]

As without proportional transaction costs [57], linear price impact implies that the risky weight is no longer a control variable that can be specified freely by the investor. Instead, it becomes a state variable, for which only the drift rate can be influenced by applying the control \( u \). To make this precise, fix a filtered probability space \((\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})\) supporting a Brownian motion \((W_t)_{t \geq 0}\), where \( \mathcal{F}_t \) is the augmentation of the filtration generated by \( W \). We then define strategies in terms of the control variable \((u_t)_{t \geq 0}\). To rule out doubling strategies, we focus on \emph{admissible} strategies with positive wealth process \( X^u \):

**Definition 3.2.1.** An admissible strategy is an adapted process \((u_t)_{t \geq 0}\), which is square-integrable (i.e., \( \int_0^T u_t^2 dt < \infty \) a.s. for all \( T > 0 \)) and such that (3.2.4) has a unique strong solution on \([0, \infty)\) for any \( Y_0 \in [0, 1] \). For any such admissible strategy, the corresponding wealth process is\(^9\)

\[
X^u_t = X_0 \exp \left( \int_0^T \left( \mu Y_t - \frac{\sigma^2}{2} Y_t^2 - \varepsilon |u_t| - \lambda u_t^2 \right) dt + \int_0^T \sigma Y_t dW_t \right).
\]

As in [37, 49, 50], the representative investor has constant relative risk aversion and maximizes the growth rate of her expected utility from terminal wealth over a long horizon. Put differently, she maximizes the “equivalent safe rate”, for which a full safe investment yields the same utility as investing optimally in the original market:

**Definition 3.2.2.** An admissible strategy \((u_t)_{t \geq 0}\) is called long-run optimal, if it maximizes the equivalent safe rate

\[
ESR_\gamma(u) := \lim_{T \to \infty} \frac{1}{T} \log \mathbb{E} \left[ \left( X^u_T / Y_0 \right)^{\lambda} \right]^{1/\lambda}
\]

over all admissible strategies, where \( 0 < \gamma \neq 1 \) denotes the investor’s relative risk aversion.

\(^9\)If \( y_* := \mu / (\gamma \sigma^2) \in (0, 1) \), i.e., the frictionless target portfolio prescribes neither leverage nor short selling, then Lemma 3.7.6 shows that \( Y_t \) takes values in \([0, 1]\) almost surely for all \( t \). In particular, \( \int_0^T \dot{Y}_t^2 dt < \infty \) so that the process \( X^u \) is well defined in this case.
3.3 Main Results

Our main results can be summarized as follows:

**Theorem 3.3.1.** Assume an investor with constant relative risk aversion $0 < \gamma < 1$ trades to maximize the equivalent safe rate (3.2.6), in the presence of a nontrivial bid-ask spread $\varepsilon$ and finite market depth $1/\lambda$. Then, if $y_* := \mu/(\gamma \sigma^2) \in (0, 1)$ and $\varepsilon, \lambda$ are sufficiently small:

i) There exist constants $\beta \in \max\{0, \mu - \gamma \sigma^2/2, \mu^2/(2\gamma \sigma^2)\}$ and $0 \leq y_- \leq y_+ \leq 1$, as well as a $C^1$-function $q : [0, 1] \to \mathbb{R}$ which solve the ODE

$$0 = -\beta + \mu y - \frac{\gamma \sigma^2}{2} y^2 + y(1-y)(\mu - \gamma \sigma^2 y)q$$

$$+ \frac{\sigma^2}{2} y^2 (1-y)^2 (q' + (1-\gamma)q^2)$$

$$+ \begin{cases} \frac{1}{4\lambda} \frac{(q-\varepsilon(1-yq))^2}{1-yq}, & \text{if } y \in [0, y_-], \\ 0, & \text{if } y \in [y_-, y_+], \\ \frac{1}{4\lambda} \frac{(q+\varepsilon(1-yq))^2}{1-yq}, & \text{if } y \in [y_+, 1], \end{cases}$$

with boundary conditions

$$q(0^+) = \varepsilon + 2\sqrt{\lambda\beta},$$

$$q(1^-) = \frac{\lambda d - \varepsilon(1-\varepsilon) - \sqrt{\lambda d(\lambda d - 2 + 2\varepsilon)}}{(1-\varepsilon)^2},$$

$$\text{where } d := -\gamma \sigma^2 - 2\beta + 2\mu,$$

$$q(y_-) = \frac{\varepsilon}{1+\varepsilon y_-},$$

$$q(y_+) = \frac{-\varepsilon}{1-\varepsilon y_+}.$$

ii) A long-run optimal strategy $\hat{u}$ is to remain inactive while the corresponding risky weight lies in the no-trade region $[y_-, y_+]$, and to rebalance at the following wealth turnover rate if it does not:

$$\hat{u}(y) = \begin{cases} \frac{1}{2\lambda} \left( \frac{q(y)}{1-yq(y)} - \varepsilon \right) \geq 0, & \text{if } y \in [0, y_-], \\ \frac{1}{2\lambda} \left( \frac{q(y)}{1-yq(y)} + \varepsilon \right) \leq 0, & \text{if } y \in [y_+, 1]. \end{cases}$$

iii) The maximal equivalent safe rate is $\beta$.

The constant $y_* = \mu/(\gamma \sigma^2)$ is the risky weight without frictions [98]. Thus, $y_* \in (0, 1)$ means that the frictionless optimal strategy neither shorts nor leverages the risky asset. As shown by Guasoni and Weber [57, Theorem 3], levered or short positions cannot be admissible with linear price impact, because they cannot be liquidated quickly enough to offset unfavorable diffusive price moves. This is only exacerbated by the additional linear trading cost. Hence, buy-and-hold strategies are optimal for $y_* \notin (0, 1)$, as in [57, Theorem 3]:
Proposition 3.3.2. Under the assumptions of Theorem 3.3.1:

(i) If \( \mu / (\gamma \sigma^2) \leq 0 \), then \( Y_t = 0 \) and \( \hat{u}_t = 0 \) for all \( t \) is long-run optimal, and
\[ \text{ESR}_\gamma(\hat{u}) = 0. \]

(ii) If \( \mu / (\gamma \sigma^2) \geq 1 \), then \( Y_t = 1 \) and \( \hat{u}_t = 0 \) for all \( t \) is long-run optimal, and
\[ \text{ESR}_\gamma(\hat{u}) = \mu - \gamma \sigma^2/2. \]

The objective function (3.2.6) uses paper wealth rather than the liquidation value of the portfolio. This seems more reasonable for long-run investments meant to run indefinitely, such as trust funds and university endowments. However, similarly as in [57, Lemma 4] for purely quadratic costs, it can be shown that this choice is of little consequence.\(^{10}\) To wit, assuming a constant best quote (which is justified if the portfolio is sold quickly), a policy of selling at a constant turnover rate completes the liquidation within a short period of time and at a small fraction of portfolio value. Formally:

Lemma 3.3.3. Let \( S_t \equiv S \) be constant for \( t \geq T \). Then, the liquidation time \( L(u) = \inf \{ t \geq 0 : \theta_{T+t} = 0 \} \) of the constant selling policy \( u_t \equiv u < 0 \), equals
\[ L(u) = -\frac{\log (1 + (\varepsilon |u| + \lambda u^2)Y_T)}{\varepsilon |u| + \lambda u^2} \sim -\frac{Y_T}{u}. \]

The corresponding relative liquidation cost is
\[ \frac{X_T - X_{T+L(u)}}{X_T} = \varepsilon Y_T - \lambda u Y_T. \]

For instance, if\(^ {11}\) \( u = -\lambda^{-1/2} \) the liquidation time is less than \( \lambda^{1/2} \) years, since \( Y_T \in [0, 1] \). Furthermore, the corresponding liquidation cost is less than \( (\varepsilon + \lambda^{1/2}) \) times the terminal wealth. Using the estimation interval of \( [10^{-3}, 10^{-7}] \) for \( \lambda \) (cf. [57, Section 3.1]) and assuming a liquid stock with a bid-ask spread of 10 basis points yields a liquidation time between 0.08 and 7.91 days and a relative liquidation cost between 0.13% and 3.26%. As the horizon increases, the impact of these small costs on the equivalent safe rate vanishes. Moreover, the short liquidation time supports the constant best quote assumption.

3.4 Numerical Examples

In this section, we investigate the properties of the optimal rebalancing policy from Theorem 3.3.1 in some numerical examples. This also allows us to assess the quality of the formal asymptotics derived in Section 3.6, which turns out to be excellent.

\(^{10}\) In the proof of [57, Lemma 4], use (3.2.4-3.2.5) instead of [57, Equation (4) and (5)]. The rest of the proof carries through unchanged.

\(^{11}\) As shown in [57, Equation (12)] the optimal turnover \( u^{0,\lambda} \) without linear costs admits the following asymptotic expansion:
\[ u^{0,\lambda}(y) = \sigma (\gamma/2)^{1/2} (y_* - y)\lambda^{-1/2} + o(\lambda^{-1/2}). \]
Figure 3.1: left panel: optimal wealth turnover $\hat{u}$ (Dotted: $\lambda = 0.001\%$, Solid: $\lambda = 0.01\%$, Dashed: $\lambda = 0.1\%$) against the risky weight $Y$ with $\varepsilon = 0.1\%$ fixed. Right panel: wealth turnover $\hat{u}$ (Dotted: $\varepsilon = 0.01\%$, Solid $\varepsilon = 0.1\%$, Dashed: $\varepsilon = 1\%$) against the risky weight $Y$ with $\lambda = 0.01\%$ fixed. Model parameters are $\mu = 8\%$, $\sigma = 16\%$, and $\gamma = 5$.

Figure 3.1 displays the optimal policy for various trading costs. The left panel shows how the turnover rate and the corresponding no-trade region depend on the price impact parameter $\lambda$. As the latter decreases, turnover quickly increases near the boundary of the no-trade region, converging to the singular controls ("pushing at an infinite rate") applied there with only proportional costs. For higher price impact costs, the optimal trading rate is almost linear in the deviation from the trading boundaries. Moreover, the width of the no-trade region decreases in this case, as investors start trading earlier to compensate for the slower trading rate at the boundary. However, the size of this effect is quite small, i.e., the width of the optimal no-trade region is relatively insensitive to the quadratic costs.

The right panel in Figure 3.1 plots the trading rate for different widths $\varepsilon$ of the bid-ask spread. As the latter decreases, the no-trade region shrinks to zero and the optimal policy converges to the one with price impact only, i.e., rebalancing at a rate essentially proportional to the deviation from the frictionless Merton portfolio [57]. For larger spreads, the no-trade regions widens quickly, and the optimal rebalancing rate increases much faster near the trading boundaries than further away from these.

In summary, the optimal policy prescribes to i) start trading earlier than with only proportional costs, and ii) rebalance slower than with only price impact. The turnover rate increases faster near the trading boundaries; further away from these, it approaches its counterpart for only quadratic costs.

As a complement, the quality of the small-cost asymptotics derived in Section 3.6 is assessed in Figure 3.2. There, we compare the optimal turnover rate to its asymptotic expansion (3.6.14) for two combinations of trading costs. Even for rather large frictions, the approximations provide an excellent fit. Hence, the computational load can be eased to finding the root of a single scalar function with little loss in accuracy.

---

12There, it grows according to the asymptotic formula (3.6.15) corresponding to a model with only quadratic costs.
3.5 Heuristics

In this section, we use arguments from stochastic control to heuristically derive a candidate solution for the long-run problem (3.2.6). To this end, consider the maximization of expected power utility $U(x) = x^{1-\gamma}/(1-\gamma)$ from terminal wealth at time $T > 0$. Denote by $V(t, X_t, Y_t)$ the corresponding value function, which is assumed to depend on the current wealth $X_t$, the current risky weight $Y_t$, and time $t$. For any given strategy $u$, Itô’s formula yields:

$$dV(t, X_t, Y_t) = V_t dt + V_x dX_t + V_y dY_t + \frac{V_{xx}}{2} d\langle X \rangle_t + \frac{V_{yy}}{2} d\langle Y \rangle_t + V_{xy} d\langle X,Y \rangle_t$$

$$= V_t dt + V_x (\mu X_t Y_t - \varepsilon X_t|u_t| - \lambda X_t u_t^2) dt + V_y Y_t (1 - Y_t)(\mu - Y_t \sigma^2) + u_t + \varepsilon Y_t|u_t| + \lambda Y_t u_t^2) dt$$

$$+ V_y Y_t (1 - Y_t) \sigma dW_t$$

$$+ \left( \frac{\sigma^2}{2} V_{xx} X_t^2 Y_t^2 + \frac{\sigma^2}{2} V_{yy} Y_t^2 (1 - Y_t)^2 + \sigma^2 V_{xy} X_t Y_t^2 (1 - Y_t) \right) dt.$$  

By the martingale optimality principle of stochastic control, the value function $V(t, X_t, Y_t)$ must be a supermartingale for any admissible strategy, and a martingale for the optimal one. That is, the drift of $V(t, X_t, Y_t)$ cannot be positive and must become zero for the optimizer. This leads to the Hamilton-Jacobi-Bellman (henceforth HJB) equation:

$$0 = V_t + y(1-y) (\mu - \sigma^2 y) V_y + \mu x y V_x$$

$$+ \frac{\sigma^2 y^2}{2} (x^2 V_{xx} + (1-y)^2 V_{yy} + 2 x (1-y) V_{xy})$$

$$+ \max_u \left( -\lambda x u^2 V_x - \varepsilon x |u| V_x + V_y (u + \varepsilon y |u| + \lambda y u^2) \right).$$

The homotheticity $U(x) = x^{1-\gamma}U(1)$ of the power utility function and the conjecture that in the long run utility should grow at a constant exponential rate lead to the following ansatz for the long-run value function:

$$V(t, x, y) = \frac{x^{1-\gamma}}{1-\gamma} e^{-(1-\gamma)(\beta(T-t) + \int_0^t q(z)dz)}.$$  \hfill (3.5.1)
Note that the function $q$ is defined up to some arbitrary $p$. This definition of $V$ leads to the long-run version of the HJB-equation:

$$0 = -\beta + \mu y - \frac{\gamma \sigma^2}{2} y^2 + y(1 - y)(\mu - \gamma \sigma^2 y)q + \frac{\sigma^2}{2} y^2(1 - y)^2 (q' + (1 - \gamma)q^2) + \max_u (-\lambda u^2 - \varepsilon |u| + (u + \varepsilon |u| y + \lambda y u^2)q).$$

(3.5.2)

Decomposing wealth turnover $u$ into purchase and sale turnover, i.e., $u = u^+ - u^-$, the HJB equation reduces to

$$0 = -\beta + \mu y - \frac{\gamma \sigma^2}{2} y^2 + y(1 - y)(\mu - \gamma \sigma^2 y)q + \frac{\sigma^2}{2} y^2(1 - y)^2 (q' + (1 - \gamma)q^2) + \max_{u^+ \geq 0} (-\lambda (u^+)^2 - \varepsilon u^+ + (u^+ + \varepsilon u^+ y + \lambda y (u^+)^2)q) + \max_{u^- \geq 0} (-\lambda (u^-)^2 - \varepsilon u^- + (-u^- + \varepsilon u^- y + \lambda y (u^-)^2)q).$$

(3.5.3)

Suppose the “second-order condition” $q(y)y < 1$ is satisfied (this holds for the function $q$ constructed in Lemma 3.7.2). Then, the maxima are attained at

$$u^+(y) = \max \left( \frac{1}{2\lambda} \left( \frac{q(y)}{(1 - yq(y))} - \varepsilon \right), 0 \right),$$

(3.5.4)

$$u^-(y) = \max \left( -\frac{1}{2\lambda} \left( \frac{q(y)}{(1 - yq(y))} + \varepsilon \right), 0 \right).$$

(3.5.5)

The optimizer with proportional transaction costs but without price impact is characterized by a no-trade interval around the frictionless optimum $y^*_\ast = \mu/(\gamma \sigma^2)$. Hence, we conjecture that the no-trade region in the present setting,

$$\left\{ y : u^+(y) = u^-(y) = 0 \right\} = \left\{ y : -\varepsilon < \frac{q(y)}{1 - yq(y)} < \varepsilon \right\},$$

is also given by some interval $[y_-, y_+]$. Substituting the optimal turnover rates (3.5.4-3.5.5) back into (3.5.3), we see that the HJB equation in turn simplifies to the ODE (3.3.1). Imposing continuity across the boundaries $y_-, y_+$ of the no-trade region in turn yields (3.3.4-3.3.5). Since the differential equation (3.3.1) is of order one and there are four unknowns to be determined ($\beta, y_-, y_+, \text{ and } q$), the value matching conditions (3.3.4-3.3.5) are not sufficient to characterize the solution. As a way out, we add two additional boundary conditions that become active when the investor’s portfolio approaches full safe ($Y_t = 0$) or full risky investment ($Y_t = 1$). The idea is that the trading rate (3.5.4-3.5.5) should remain finite in each case; moreover, it should be positive at $Y_t = 0$ and negative at $Y_t = 1$ so as to keep the risky weight in $[0, 1]$. Solving the ODE (3.3.1) at $y \in \{0, 1\}$ leads to a quadratic equation for the boundary value of $q$; choosing the

\[\text{Recall that this is needed to ensure solvency, since portfolios involving short or levered positions lead to bankruptcy with positive probability.}\]
solution with the correct sign in turn gives

\[ q(0^+) = \varepsilon + 2\sqrt{\lambda\beta}, \quad (3.5.6) \]

\[ q(1^-) = \frac{\lambda d - \varepsilon(1 - \varepsilon) - \sqrt{\lambda d(\lambda d - 2 + 2\varepsilon)}}{(1 - \varepsilon)^2}, \quad d = -\gamma \sigma^2 - 2\beta + 2\mu. \quad (3.5.7) \]

Together with the value matching conditions (3.3.4-3.3.5), this yields the representation from Theorem 3.3.1. For \( y_* \in (0,1) \), this informal derivation indeed leads to the correct answer (cf. the rigorous verification theorem in Section 3.7). For \( y_* \not\in (0,1) \), however, the candidate risky weight explodes with positive probability, and the massive rebalancing that comes along with this reduces the corresponding wealth to zero (cf. Lemma 3.7.6). Hence, our candidate strategy is not even admissible in this case, and a simple buy-and-hold strategy turns out to be optimal instead (cf. Proposition 3.3.2). This stresses the need for rigorous verification theorems to complement heuristic considerations, which might otherwise lead to wrong results.

### 3.6 Formal Asymptotics

The differential equation (3.3.1) in Theorem 3.3.1 is of Abel type; no explicit solution is known. However, it is possible to formally obtain asymptotic expansions for the no-trade region \([y_-, y_+]\), the trading rate \(\hat{u}\), and the corresponding equivalent safe rate \(\beta\) as the market frictions tend to zero. Then, the computation of the optimal policy and welfare is simplified from the solution of a nonlinear free-boundary value problem to finding the root of a nonlinear function, similarly as in Korn [79] for proportional and fixed costs. Making the asymptotics rigorous is a purely analytical problem, that would require substantial extra effort beyond the scope of this thesis.

Consider the limiting regime where both the proportional transaction cost \(\varepsilon\) and the price impact parameter \(\lambda\) tend to zero. If these frictions are considered separately, their leading-order impact on the equivalent safe rate is of order \(\varepsilon^{2/3}\) and \(\lambda^{1/2}\), respectively (cf. [43, Formula (2.7)] or [57, Formula (13)]). To obtain an expansion in which neither friction vanishes, we rescale them appropriately to put their asymptotic contributions on the same scale:

\[ \lambda = K \varepsilon^{4/3}, \quad \text{where } K > 0. \]

With this substitution the problem of finding the asymptotic expansion splits into two regions: close to the no-trade region\(^{14}\) and far away from it. To study the first regime, one can argue using homogenization techniques as in Soner and Touzi [117] (see also [107, 5, 101]). To this end, first derive the HJB equation

\(^{14}\)For small transaction costs - without market price impact - the no-trade region \([y_-, y_+]\) contains the Merton proportion \(y_*\), its width is of order \(\varepsilon^{1/3}\), and the corresponding welfare effect is of order \(\varepsilon^{2/3}\), see [62]. With price impact, the no-trade region is again an interval, which contains the Merton proportion provided that \(\varepsilon\) is not too small compared to \(\lambda\) (compare Remark 3.7.5).
for the value function $V$ and postulate an appropriate asymptotic expansion. Then, substitute the expansion back into HJB equation, and collect the leading order terms. This in turn leads to the so-called corrector equation. In the pure transaction cost case, the appropriate ansatz for $V$ is

$$V(t, x, y) = V_0(t, x) - \varepsilon^{2/3}u(t, x) - \varepsilon^{4/3}w(t, x, z) + O(\varepsilon),$$

(3.6.1)

with the fast variable $z = \frac{y-y_s}{\varepsilon^{1/3}}$. Here, $V_0$ denotes the frictionless value function, $x$ the total wealth, and $y$ the wealth invested in the risky asset. Note that the HJB equation for $V$ is of second order. Hence, taking the second derivative of the function $w$ and multiplying it with $\varepsilon^{4/3}$ produces again a term which is of order $\varepsilon^{2/3}$ justifying the power $4/3$ in the expansion.

Our long-run objective forces us to work with the reduced value function $q\ast$.

$$q\ast(y) := \varepsilon \times \begin{cases} 
  r_B(z), & y \in [0, y_-], \\
  r(z), & y \in [y_-, y_+], \\
  r_S(z), & y \in [y_+, 1],
\end{cases} \quad (3.6.3)$$

for constants $0 \leq y_- \leq y_+ \leq 1$, $l > 0$, and functions $r_B, r, r_S$ to be determined. Inserting (3.6.3) back into (3.6.2) yields that, for each $h \in \{r_B, r, r_S\}$:

$$V(t, x, y) = V_0(t, x) \cdot e^{(1-\gamma)(\beta(T-t)+\int_0^y h(z)dz)}$$

$$= V_0(t, x) - V_0(t, x)(1 - \gamma)(T - t)l \cdot \varepsilon^{2/3}$$

$$+ V_0(t, x)(1 - \gamma) \int_y^y h(z)dz \cdot \varepsilon + O(\varepsilon).$$

(3.6.4)

Note that we multiply the third term in Ansatz (3.6.4) with $\varepsilon$, since the exponential transformation (3.6.2) implies a first-order HJB equation for $q$ (cf. (3.3.1)).

To determine the boundaries of the no-trade region $y_-$ and $y_+$ we need two more conditions. To this end, we consider the limit behavior of $q$ as the risky weight $y$ approaches the full investment levels 0 and 1. There, we ignore the fast variable $z$ in (3.6.4), because we are far away from the no-trade region. Again matching the terms at the leading order $\varepsilon^{2/3}$ we define $q(y) = \varepsilon^{2/3}q^\ast(y)$. As $\varepsilon \downarrow 0$, the ODE (3.3.1) then simplifies to

$$q^\ast(y) = (2K\gamma\sigma^2)^{1/2} (y_s - y).$$

(3.6.5)

This equation describes the behavior of $q^\ast$ (and in turn $q$) away from the no trade region and therefore yields the required conditions to determine $y_-$ and $y_+$.\footnote{A similar “pasting” of two different expansions is used in [19] to deal with a small capital gains tax.}
As $\varepsilon \downarrow 0$, the transformation (3.6.3) reduces the ODE (3.3.1) to an inhomogeneous Riccati equation:

\[
0 = -\frac{\gamma \sigma^2}{2} z^2 + l + \frac{\sigma^2}{2} y^2(1 - y^*)^2 r_B + \frac{1}{4K} (r_B - 1)^2, \quad z \in (-\infty, z_-], \quad (3.6.6)
\]

\[
0 = -\frac{\gamma \sigma^2}{2} z^2 + l + \frac{\sigma^2}{2} y^2(1 - y^*)^2 r_s', \quad z \in [z_-, z_+], \quad (3.6.7)
\]

\[
0 = -\frac{\gamma \sigma^2}{2} z^2 + l + \frac{\sigma^2}{2} y^2(1 - y^*)^2 r_s' + \frac{1}{4K} (r_s + 1)^2, \quad z \in [z_+, +\infty), \quad (3.6.8)
\]

where the rescaled buying and selling boundaries $z_-$ and $z_+$ are determined by the following value matching conditions:

\[
r_B(z_-) = r(z_-) = 1, \quad r_s(z_+) = r(z_+) = -1. \quad (3.6.9)
\]

As $z \to -\infty$ (or $z \to +\infty$), the function $r_B$ (or $r_S$) must diverge with the same rate as in our first expansion in (3.6.5):

\[
\lim_{z \to -\infty} \frac{r_B(z)}{-\sqrt{2K\gamma\sigma^2z}} = 1 \quad \text{and} \quad \lim_{z \to +\infty} \frac{r_S(z)}{-\sqrt{2K\gamma\sigma^2z}} = 1. \quad (3.6.10)
\]

In summary, this leads to the ODE (3.6.6-3.6.8) with value matching and growth conditions (3.6.9-3.6.10) – the corrector equation in the present setting.

The next proposition (proved in Section 3.7.2) shows that the ODE (3.6.6-3.6.8) with boundary conditions (3.6.9-3.6.10) admits a unique solution. As a side product, we show that additional price impact decreases the width of the no-trade region compared to the pure linear cost case.

**Proposition 3.6.1.** There are unique $z_-$, $z_+$ and $l_*>0$ such that equation (3.6.6-3.6.8) has a solution that satisfies the value matching conditions (3.6.9) and the growth conditions (3.6.10). In particular,

\[
l_* > \max \left( \sqrt{\frac{\gamma K}{2}} \frac{\sigma^3 y^2(1 - y^*)^2}{2}, \left( \frac{3}{4} \sqrt{2} \frac{\sigma^3 y^2(1 - y^*)^2}{2} \right)^{2/3} \right).
\]

Also, if $z^0_-$ and $z^0_+$ are the boundaries of no-trade region without price impact (see [43, Formula (2.9)]), then $z_- > z^0_-$ and $z_+ < z^0_+$.

**Proof.** See Section 3.7.2. \hfill \square

To find the explicit solution of the ODE (3.6.6-3.6.8) with boundary conditions (3.6.9-3.6.10) we proceed as follows. A simple calculation shows that the linear, inhomogeneous ODE (3.6.7) has the explicit solution

\[
r(z) = \frac{2}{\sigma^2 y^2(1 - y^*)^2} \left( \frac{\gamma \sigma^2}{6} z^3 - l z \right).
\]
Since \( r \) is an odd function, the boundary conditions \( r(z_-) = 1 \) and \( r(z_+) = -1 \) can be replaced by \( r(z_-) = 1 \) and \( r(0) = 0 \). Then, \( z_- = -z_+ \) and it suffices to determine \( z_- \) because the constant \( l \) is linked to \( z_- \) via the condition \( r(z_-) = 1 \):

\[
l(z_-) = \frac{\gamma \sigma^2}{6} z_-^2 - \frac{\sigma^2 y_+^2 (1 - y_+)^2}{2z_-}.
\]

(3.6.11)

To find the rescaled trading boundary \( z_- \) determined by \( r_B(z_-) = 1 \), the Riccati equation (3.6.6) with boundary condition (3.6.10) needs to be solved first.

Equation (3.6.6) is – up to transformations [72, Transformations 1.25, 1.105, 2.220, and 2.273, Formula (9)] – equivalent to a Whittaker equation for which explicit solutions (in terms of the Whittaker functions) are known (cf. [116, Section 1]). For brevity, we omit the derivation of the solution and simply state the final result:

\[
r_B(z, l) := -\frac{1}{z} \left( \frac{1}{2a} + \frac{c}{2a} \sqrt{\frac{1}{2K \gamma \sigma^2}} \right) + 1 + \sqrt{2K \gamma \sigma^2 z} \tag{3.6.12}
\]

where

\[
a = \frac{1}{2K \sigma^2 y_+^2 (1 - y_+)^2}, \quad c = \frac{2l}{\sigma^2 y_+^2 (1 - y_+)^2}, \quad k = \frac{c}{4} \sqrt{\frac{1}{2K \gamma \sigma^2}},
\]

and the Whittaker function \( W \) is defined via the Kummer function\(^{16}\) \( _1F_1(a, b, x) \) and the Tricomi function \( U(\xi, \eta, x) \) cf. [116, Chapter 1]:

\[
U(\xi, \eta, x) := \frac{\Gamma(1 - \eta)}{\Gamma(1 + \xi - \eta)} _1F_1(\xi, \eta, x) + \frac{\Gamma(\eta - 1)}{\Gamma(\xi)} x^{1-\eta} _1F_1(1 + \xi - \eta, 2 - \eta, x),
\]

\[
W(k, m, x) := x^{\frac{1}{2} + m} e^{-\frac{1}{2} x} U(1/2 + m - k, 1 + 2m, x).
\]

We note that the general solution of the Whittaker equation is given by a linear combination of the Whittaker functions\(^{17}\) \( M(k, m, x) \) and \( M(k, -m, x) \) (cf. [123, Section 16]) with

\[
M(k, m, x) := x^{\frac{1}{2} + m} e^{-\frac{1}{2} x} _1F_1(1/2 + m - k, 1 + 2m, x).
\]

\(^{16}\)For \( b \neq 0, -1, -2, \ldots \), the Kummer function \( _1F_1(a, b, x) \) is defined through the following absolute convergent series [116, Chapter 1]:

\[
_1F_1(a, b, x) := \sum_{n=0}^{\infty} \frac{a^{(n)} x^n}{b^{(n)} n!}.
\]

Here, the Pochhammer symbol \( a^{(n)} \) is given by \( a^{(n)} := a(a + 1)(a + 2) \cdots (a + n - 1) \). For \( a \neq 0, -1, -2, \ldots \), we can write \( a^{(n)} = \Gamma(a + n)/\Gamma(a) \), where \( \Gamma(x) \) denotes the Gamma function.

\(^{17}\)Note that the Whittaker function \( W \) (cf. [116, Formula 1.7.1]) can be written as

\[
W(k, m, x) = \frac{\pi}{\sin(2m\pi)} \left( \frac{-M(k, m, x)}{\Gamma(1/2 - m - k)\Gamma(1 + 2m)} + \frac{M(k, -m, x)}{\Gamma(1/2 + m - k)\Gamma(1 - 2m)} \right).
\]
The Riccati equation needs to be solved with the initial condition (3.6.10). Therefore, the Whittaker function \( W(k, m, x) \) is the only candidate, since it has the correct asymptotic growth [123, Section 16.31]:

\[
W(k, m, x) \sim x^k e^{-\frac{1}{2}x}, \quad \text{as} \quad x \to \infty.
\]

Indeed, for any \( l \in \mathbb{R}, \) (3.6.12) satisfies the boundary condition (3.6.10):

\[
\lim_{z \to -\infty} \frac{r_B(z, l)}{-\sqrt{2K\gamma\sigma^2}z} = \lim_{z \to -\infty} \frac{1}{-\sqrt{2K\gamma\sigma^2}z} \left( -\frac{1}{z} \left( \frac{1}{2a} + \frac{c}{2a} \sqrt{\frac{1}{2K\gamma\sigma^2}} \right) + 1 \right)
\]

\[
+ \frac{\sqrt{2K\gamma\sigma^2}}{-\sqrt{2K\gamma\sigma^2}z} \left[ z - \frac{2z}{a\sqrt{2K\gamma\sigma^2}z^2} \right] \times \frac{W(k + 1, -1/4, a\sqrt{2K\gamma\sigma^2}z^2)}{W(k, -1/4, a\sqrt{2K\gamma\sigma^2}z^2)}
\]

\[
= 1,
\]

where we have used (3.6.13) in the last step.

Now, we can put everything together. \( z_- \) is determined by \( r_B(z_-, l(z_-)) = 1 \) with \( l(z_-) \) from (3.6.11). Hence, the asymptotic expansions for the growth rate \( \beta \) and the boundaries \( y_-, y_+ \) of the no-trade region in (3.6.3) are given by:

\[
\beta = \frac{\mu^2}{2\gamma\sigma^2} - \left( \gamma\sigma^2 \frac{\gamma\sigma^2}{6} z_-^2 - \frac{\gamma\sigma^2}{2z_-} \right) \varepsilon^2 + O(\varepsilon),
\]

\[
y_+ = y_* \pm z_- \varepsilon^\frac{1}{3} + O(\varepsilon^\frac{2}{3}),
\]

where \( z_- \) is the root of the equation \( r_B(z, l(z)) = 1 \) with \( l(z) \) from (3.6.11). The asymptotics for the turnover rate \( \hat{u} \) can be derived similarly. Recall from Theorem 3.3.1 that

\[
\hat{u}(y) = \begin{cases} 
\frac{1}{2\lambda} \left( \frac{q(y)}{1 - yq(y)} - \varepsilon \right) \geq 0, & \text{if} \ y \in [0, y_-], \\
0, & \text{if} \ y \in [y_-, y_+], \\
\frac{1}{2\lambda} \left( \frac{q(y)}{1 - yq(y)} + \varepsilon \right) \leq 0, & \text{if} \ y \in [y_+, 1],
\end{cases}
\]

Using the same transformations as in (3.6.3) for \( y \) and \( q \), a straightforward calculation shows that

\[
\hat{u}(y) = \begin{cases} 
\frac{1}{2\lambda} (r_B(z) - 1) \varepsilon^{-\frac{1}{3}} + O(1), & \text{if} \ z = (y - y_*)\varepsilon^{-1/3} \in (-\infty, z_-], \\
0, & \text{if} \ z = (y - y_*)\varepsilon^{-1/3} \in [z_-, z_+], \\
\frac{1}{2\lambda} (r_S(z) + 1) \varepsilon^{-\frac{1}{3}} + O(1), & \text{if} \ z = (y - y_*)\varepsilon^{-1/3} \in [z_+, +\infty),
\end{cases}
\]

(3.6.14)

where we abbreviate \( r_B(z) := r_B(z, l(z_-)) \) and \( r_S(z) := r_B(z) - 2 \).

In summary, asymptotically for small trading costs, the solution of the nonlinear free-boundary problem (3.3.1) can be reduced to finding the root of a scalar
function. The approximation (3.6.14) performs very well, even for relatively large values of the asymptotic parameters $\varepsilon, \lambda$, cf. Figure 3.2. Moreover, it also allows us to say more about the structure of the optimal turnover rate near the trading boundaries and far away from these.

Far away from the no-trade region these are boundary conditions. In particular, as $z \to -\infty$ (or, equivalently, $y \downarrow 0$) we have the limit

$$\lim_{z \to -\infty} \frac{r_B(z)}{\sqrt{-2K\gamma\sigma^2z}} = 1.$$ 

As the other extreme, as $z \to \infty$ (or, equivalently, $y \uparrow 1$)

$$\lim_{z \to \infty} \frac{r_S(z)}{\sqrt{-2K\gamma\sigma^2z}} = 1.$$ 

Together, these two boundary conditions imply

$$\hat{u}(y) = \frac{1}{2K} \left( -\sqrt{2K\gamma\sigma^2z} - 1 \right) \varepsilon^{-1/3} + O(1)$$

$$= \frac{1}{2K} \sqrt{2K\gamma\sigma^2(y_* - y)} \varepsilon^{-2/3} + O(\varepsilon^{-1/3})$$

$$= \sigma \sqrt{\frac{\gamma}{2}} (y_* - y) \lambda^{-1/2} + O(\lambda^{-1/4}).$$

(3.6.15)

Hence, for large deviations from the trading boundaries we recover the leading-order expansion of the wealth turnover without proportional transaction costs [57, Formula (12)].

Close to the trading boundaries, we can apply Taylor’s Theorem to get a first-order approximation as well. To this end, we first compute the derivative of the corresponding Whittaker functions. The differential property [116, Formula (2.4.24)] and the recurrence relation [116, Formula (2.5.11)] of $W(k, m, x)$ show that

$$\left( \frac{W(k + 1, -1/4, x)}{xW(k, -1/4, x)} \right)' = \frac{W(k + 1, -1/4, x)}{xW(k, -1/4, x)} \times \left( \frac{W'(k + 1, -1/4, x)}{W(k + 1, -1/4, x)} - \frac{W'(k, -1/4, x)}{W(k, -1/4, x)} - \frac{1}{x} \right)$$

$$= \frac{W(k + 1, -1/4, x)}{xW(k, -1/4, x)} \left[ \frac{W(k + 1, -1/4, x)}{xW(k, -1/4, x)} - \frac{2}{x} \right.$$

$$- \frac{1}{x^2} \left( -\frac{1}{4} - (k + 1) + \frac{1}{2} \right) \left( -\frac{1}{4} + (k + 1) - \frac{1}{2} \right)$$

$$\times \frac{xW(k, -1/4, x)}{W(k + 1, -1/4, x)} - (2(k + 1)x - x^2) \left. \right],$$

where $x = a \sqrt{2K\gamma\sigma^2z^2}$. Taking into account the value matching condition $r_B(z_-) = 1$, a straightforward computation yields that the wealth turnover close
Figure 3.3: optimal wealth turnover $\hat{u}$ from Theorem 3.3.1 (solid), and its piecewise linear approximation (dotted) by (3.6.16) close to the trading boundaries and (3.6.15) further away from these. Model parameters are $\mu = 8\%$, $\sigma = 16\%$, $\gamma = 5$, $\lambda = 0.01\%$, and $\varepsilon = 0.1\%$ (left panel) or $\varepsilon = 0.5\%$ (right panel).

To the trading boundaries $y_{\pm}$ is given by

$$\hat{u}(y) = \begin{cases} \varepsilon^{1/3} \frac{e^{-1/3}}{2K} \times F \times (z - z_{-}) + \mathcal{O}(1), & \text{if } z = (y - y_{*}) \varepsilon^{-1/3} < z_{-}, \\ \varepsilon^{1/3} \frac{e^{-1/3}}{2K} \times F \times (z - z_{+}) + \mathcal{O}(1), & \text{if } z = (y - y_{*}) \varepsilon^{-1/3} > z_{+}, \end{cases} \tag{3.6.16}$$

where

$$x_{-} := a \sqrt{2K\gamma\sigma^{2}z_{-}^{2}},$$

$$D := \frac{1}{2} \left( 1 - \frac{1}{\sqrt{2K\gamma\sigma^{2}}z_{-}^{2}} \left( \frac{1}{2a} + \frac{c}{2a} \sqrt{\frac{1}{2K\gamma\sigma^{2}}} \right) \right),$$

$$E := D \left[ (D - \frac{2}{x_{-}} - \frac{1}{x_{-}^{2}} \right) \times \left( \frac{1}{D} \left( -\frac{1}{4} - (k + 1) + \frac{1}{2} \right) \left( -\frac{1}{4} + (k + 1) \right) - (2(k + 1)x_{-} - x_{-}^{2}) \right],$$

$$F := \left( \frac{1}{2a} + \frac{c}{2a} \sqrt{\frac{1}{2K\gamma\sigma^{2}}} \right) \frac{1}{z_{-}^{2}} + \sqrt{2K\gamma\sigma^{2} - 2\sqrt{2K\gamma\sigma^{2}}(D + 2a\sqrt{2K\gamma\sigma^{2}Ez_{-}^{2}}) \right).$$

In particular, for small deviations from the trading boundaries the wealth turnover – at the first order – is again linear, however with a different slope. This is illustrated in Figure 3.3. If the proportional costs are small compared to the price impact, then the two slopes are very similar. For larger spreads, however, turnover grows significantly faster near the trading boundaries, compare the right panel in Figure 3.3.

**Robustness of the Asymptotics** In this section we argue informally that – in the small-cost limit – the model (3.2.2) is robust to the specification of the price
impact parameter, in parallel with results for linear and quadratic costs treated separately [71, 101].

To this end, consider an exogenously given diffusion process $\xi$ with dynamics
\[ d\xi_t = \mu_\xi(\xi_t)dt + \sigma_\xi(\xi_t)dW_t, \] (3.6.17)
where $\mu_\xi, \sigma_\xi$ are smooth functions such that the SDE (3.6.17) is well-defined. This additional state variable can model the wealth of other agents trading in the market, for example. The average execution price for the large investor is in turn given by
\[ S_t \left( 1 + \varepsilon \cdot sgn(\Delta \theta) + \lambda(\xi_t, X_t) \frac{S_t \Delta \theta}{\Delta t} \right). \]

To wit, the price impact is a general function of the large investor’s wealth $X$ and the exogenous process $\xi$. If $\lambda(\xi, x) = \lambda/x$ we recover the model (3.2.2).

An easy application of Itô’s Lemma shows that in this setting, the wealth process $X$ and the risky weight $Y$ satisfy:
\[ \frac{dX_t}{X_t} = \left( Y_t (\mu - Y_t \sigma^2) + u_t + \varepsilon |u_t|Y_t + \lambda(\xi_t, X_t) X_t Y_t u_t^2 \right) dt \]
+ $Y_t (1 - Y_t) \sigma dW_t$. (3.6.18)

Let $V(t, X_t, Y_t, \xi_t)$ denote the corresponding finite horizon value function, which is assumed to depend on the current wealth $X_t$, the current risky weight $Y_t$, the exogenous state $\xi_t$, and time $t$. Arguing as in Section 3.5, it follows that
\[ V(t, x, y, \xi) = \frac{x^{1-\gamma}}{1-\gamma} e^{(1-\gamma)(\beta(T-t)+\int_{y_*}^y q(\xi, x, u)du)}, \] (3.6.20)
where the function $q(\xi, x, y)$ satisfies
\[ 0 = -\beta + \mu y - \frac{\gamma \sigma^2}{2} y^2 + y(1 - y)(\mu - \gamma \sigma^2 y)q + \frac{\sigma^2}{2} y^2 (1 - y)^2 (q_y + (1 - \gamma)q^2) \]
+ $\mu x \int_{y_*}^y q_x(\xi, x, u)du + \sigma^2 y^2 (1 - y) \\
\times \left( (1 - \gamma)x q(\xi, x, y) \int_{y_*}^y q_x(\xi, x, u)du + x q_x(\xi, x, y) \right) \\
+ \frac{\sigma^2 y^2}{2} \left[ 2(1 - \gamma)x \int_{y_*}^y q_x(\xi, x, u)du + (1 - \gamma) \left( x \int_{y_*}^y q_x(\xi, x, u)du \right)^2 \right] \\
+ x^2 \int_{y_*}^y q_{xx}(\xi, x, u)du \right] + \mu_x(\xi) \int_{y_*}^y q_x(\xi, x, u)du \\
+ \frac{\sigma^2_x(\xi)}{2} \left[ \left( \int_{y_*}^y q_x(\xi, x, u)du \right)^2 (1 - \gamma) + \int_{y_*}^y q_{xx}(\xi, x, u)du \right] \\
+ \ldots \] (3.6.21)
\[
+ \sigma \sigma (\xi) y (1 - y) \left( 1 - \gamma \right) q (\xi, x, y) \int_{y_*}^y q_x (\xi, x, u) du + q_x (\xi, x, y) \right) \\
+ \sigma \sigma (\xi) y \left[ 1 - \gamma \right] \int_{y_*}^y q_x (\xi, x, u) du + (1 - \gamma) x \int_{y_*}^y q_x (\xi, x, u) du \\
\times \int_{y_*}^y q_x (\xi, x, u) du + x \int_{y_*}^y q_x (\xi, x, u) du \\
\times \left\{ \begin{array}{ll}
\frac{1}{4 \lambda (\xi, x)} \left[ q_\varepsilon (1 - y - y^*) + q_\varepsilon (\xi, x, u) du \right] & \text{if } y \in [0, y_-(\xi, x)], \\
0 & \text{if } y \in [y_-(\xi, x), y_+(\xi, x)], \\
\frac{1}{4 \lambda (\xi, x)} \left[ q_\varepsilon (1 - y - y^*) + q_\varepsilon (\xi, x, u) du \right] & \text{if } y \in [y_+(\xi, x), 1], 
\end{array} \right. (3.6.22)
\]

and
\[
q (\xi, x, y_-(\xi, x)) = \frac{\varepsilon + \varepsilon x \int_{y_*}^{y_-(\xi, x)} q_x (\xi, x, u) du}{1 + \varepsilon y_-(\xi, x)}, \\
q (\xi, x, y_+(\xi, x)) = \frac{-\varepsilon - \varepsilon x \int_{y_*}^{y_+(\xi, x)} q_x (\xi, x, u) du}{1 - \varepsilon y_+(\xi, x)},
\]

for functions \(0 \leq y_-(\xi, x) \leq y_+(\xi, x) \leq 1\) to be determined. To derive the corresponding small-cost asymptotics, rescale transaction costs and price impact accordingly:
\[
\lambda (\xi, x) x = K (\xi, x) \varepsilon^{4/3},
\]
for some function \(K\). Close to the no-trade region we again apply the homogenization approach and define
\[
z := \frac{y - y_*}{\varepsilon^{1/3}}, (3.6.23)
\]
\[
q (\xi, x, y) := \varepsilon \times \left\{ \begin{array}{ll}
r_B (\xi, x, z), & y \in [0, y_-(\xi, x)], \\
r (\xi, x, z), & y \in [y_-(\xi, x), y_+(\xi, x)], \\
r_S (\xi, x, z), & y \in [y_+(\xi, x), 1], 
\end{array} \right. (3.6.24)
\]
\[
\beta := \frac{\mu^2}{2 \gamma \sigma^2} - l \varepsilon^{2/3}, (3.6.25)
\]
for a constant\(^{18}\) \(l > 0\) and functions \(0 \leq y_-(\xi, x) \leq y_+(\xi, x) \leq 1\), as well as \(r_B (\xi, x, z), r (\xi, x, z), r_S (\xi, x, z)\) to be determined. To identify the boundaries of the no-trade region we need additional conditions. To this end, we again consider \(q (\xi, x, y) = \varepsilon^{2/3} q^* (\xi, x, y)\). As \(\varepsilon \downarrow 0\), a direct calculation shows that the ODE \(3.6.22\) reduces to
\[
q^* (\xi, x, y) = (2 K (\xi, x) \gamma \sigma^2)^{1/2} (y_* - y). (3.6.26)
\]
\(^{18}\)Note that the optimal long-run growth rate is typically constant even in factor models with an additional state variable, compare [56].
If the functions $r_B, r, r_S$ are sufficiently smooth, then as $\varepsilon \downarrow 0$ transformation (3.6.23-3.6.25) reduces the ODE (3.6.22) to

$$0 = -\frac{\gamma}{2} z^2 + l + \frac{\sigma^2}{2} y_*^2 (1 - y_*)^2 r_B + \frac{1}{4K(\xi, x)} (r_B - 1)^2, \quad z \in (-\infty, z_-(\xi, x)],$$

(3.6.27)

$$0 = -\frac{\gamma}{2} z^2 + l + \frac{\sigma^2}{2} y_*^2 (1 - y_*)^2 r', \quad z \in [z_-(\xi, x), z_+(\xi, x)],$$

(3.6.28)

$$0 = -\frac{\gamma}{2} z^2 + l + \frac{\sigma^2}{2} y_*^2 (1 - y_*)^2 r_S' + \frac{1}{4K(\xi, x)} (r_S + 1)^2, \quad z \in [z_+(\xi, x), +\infty),$$

(3.6.29)

where the rescaled buying and selling boundaries $z_-(\xi, x)$ and $z_+(\xi, x)$ satisfy:

$$r_B(\xi, x, z_-(\xi, x)) = r(\xi, x, z_-(\xi, x)) = 1,$$

(3.6.30)

$$r_S(\xi, x, z_+(\xi, x)) = r(\xi, x, z_+(\xi, x)) = -1.$$  

(3.6.31)

As $z \to -\infty$ (or $z \to \infty$) the function $r_B(\xi, x, \cdot)$ (or $r_S(\xi, x, \cdot)$) must diverge with the same rate as in (3.6.26):

$$\lim_{z \to -\infty} \frac{r_B(\xi, x, z)}{-\sqrt{2K(\xi, x)}} = 1 \quad \text{and} \quad \lim_{z \to \infty} \frac{r_S(\xi, x, z)}{-\sqrt{2K(\xi, x)}} = 1.$$  

(3.6.32)

As a result, the asymptotic trading boundaries and trading rate are still determined by an inhomogenous Riccati equation, but by a different one for each value of the additional state variables. To wit, the asymptotic solutions for more general cost structures are obtained by plugging in the current value of the cost into the asymptotic expansions derived in the previous section.

### 3.7 Proofs of Theorem 3.3.1 and Theorem 3.6.1

#### 3.7.1 Proof of Theorem 3.3.1

Assume throughout that $y_* \in (0, 1)$. The first step towards a rigorous verification theorem is to show that the differential equation (3.3.1) indeed admits a solution with the required properties (3.3.2-3.3.5). To this end, we first rewrite (3.3.1) in slope field notation:

$$q' = f(y, q) = \begin{cases} f_B(y, q), & q \geq \frac{\varepsilon}{1 + \varepsilon}, \\ f_{NT}(y, q), & -\frac{\varepsilon}{1 - \varepsilon} \leq q \leq \frac{\varepsilon}{1 + \varepsilon}, \\ f_S(y, q), & q \leq -\frac{\varepsilon}{1 - \varepsilon}, \end{cases}$$

(3.7.1)

For each $h \in \{r_B, r, r_S\}$, note that if $h(\xi, \cdot)$ is continuous then

$$\int_{y_*}^y h(\xi, u) du = \varepsilon \int_{y_*}^y h(\xi, u - y_*) u^{-1/3} du = \varepsilon^{4/3} \int_0^y h(\xi, u) du = o(\varepsilon^{4/3}).$$

This will be justified a posteriori by the monotonicity properties of $q$ established in Lemma 3.7.2.
Notice that \( f(y, q) \) is well defined because \( f_B(y, q) = f_N(y, q) \) on \( q = \frac{\gamma}{1 + \gamma y} \) and \( f_N(y, q) = f_S(y, q) \) on \( q = -\frac{\gamma}{1 - \gamma y} \).

**Remark 3.7.1.** Allocating the entire wealth into the riskless asset or into the risky asset is an admissible strategy that does not require trading. The corresponding equivalent safe rate is 0 or \( \mu - \frac{\gamma^2}{2} \), respectively. This provides a natural lower bound for the optimal equivalent safe rate, namely \( \beta \geq \max\{\mu - \frac{\gamma^2}{2}, 0\} \).

Conversely, an upper bound is given by the frictionless equivalent safe rate \( \mu \frac{\gamma^2}{2} \).

**Lemma 3.7.2.** Suppose \( \lambda \) and \( \varepsilon \) are sufficiently small. Then, for a suitable \( \beta \in \left[ \max\{\mu - \frac{\gamma^2}{2}, 0\}, \frac{\mu^2}{2 \gamma^2} \right] \), there is a solution of \( q' = f(y, q) \) such that

\[
q(0^+) = b_0(\varepsilon, \lambda) := \varepsilon + 2\sqrt{\lambda \beta},
\]

\[
q(1^-) = b_1(\varepsilon, \lambda) := \frac{\lambda d - \varepsilon (1 - \varepsilon) - \sqrt{\lambda d (\lambda d - 2 + 2\varepsilon)}}{(1 - \varepsilon)^2}, d = -\gamma \sigma^2 - 2\beta + 2\mu.
\]

In particular, there exist \( y_-, y_+ \in [0, 1] \) satisfying (3.3.1) and

\[
\begin{cases}
q(y) > \frac{\varepsilon}{1 + \gamma y}, & \text{if } y \in [0, y_-], \\
\frac{\varepsilon}{1 + \gamma y} \leq q(y) \leq \frac{\varepsilon}{1 - \gamma y}, & \text{if } y \in [y_-, y_+], \\
q(y) \leq \frac{\varepsilon}{1 - \gamma y}, & \text{if } y \in (y_+, 1].
\end{cases}
\]

Moreover, the solution \( q(y) \) fulfills \( q(y) < 1 \) for all \( y \in [0, 1] \).

**Proof.** First, notice that for every \( y \in (0, 1) \) we have \( \frac{1}{y} > \frac{\varepsilon}{1 + \gamma y} \), and thus

\[
\lim_{q \to (\frac{1}{y})^-} f(y, q) = \lim_{q \to (\frac{1}{y})^-} f_B(y, q) = -\infty.
\]

Hence, every solution to \( q' = f(y, q) \) starting below the curve \( \frac{1}{y} \) must remain below this curve. The rest of the proof proceeds as follows:

(i) For every \( \beta > \max\{\mu - \frac{\gamma^2}{2}, 0\} \), there is a unique solution \( q_0(y) \) defined on \( (0, y_0) \subset (0, 1) \) that satisfies the boundary condition (3.7.2) at \( y = 0 \) and a unique solution \( q_1(y) \) defined on \( (y_1, 1) \subset (0, 1) \) that satisfies the boundary condition (3.7.3) at \( y = 1 \).

(ii) If \( \beta > \frac{\mu^2}{2 \gamma^2} \), then \( q_0(y) > 0 \) and \( 0 > q_1(y) \) on their respective definition intervals.

(iii) Set \( \beta = \frac{\mu^2}{2 \gamma^2} - c \), with \( c > 0 \). If \( \lambda \) and \( \varepsilon \) are sufficiently small, we have \( q_0(y) < q_1(y) \) on \( (0, y_0) \cap (y_1, 1) \). Additionally, if \( y_0 < 1 \) then \( \lim_{y \to y_0^-} q_0(y) = -\infty \), and if \( y_1 > 0 \) then \( \lim_{y \to y_1^+} q_1(y) = +\infty \).

The solution to \( q' = f(y, q) \) depends continuously on \( \beta \). Hence, if \( \lambda \) and \( \varepsilon \) are sufficiently small, we thus have \( q_0 \equiv q_1 \) for some \( \beta \in [\max\{\mu - \gamma \sigma^2/2, 0\}, \mu^2/(2 \gamma^2)] \).
Proof of (i): This follows similarly as in [57, Lemma 15(i)]. Indeed, replace \(q^2/4\lambda\) by \((q-\varepsilon)^2/4\lambda\) in the first line of the proof of [57, Lemma 15(i)]. Then, the proof proceeds analogously, leading to

\[
h(0) = \varepsilon + 2\sqrt{\lambda \beta}, \quad h'(0) = -\left(\frac{\lambda}{\beta}\right)^{1/2} (\mu + (\mu + \beta)h(0)) - \varepsilon h(0) < 0.
\]

This term is negative like the corresponding expression in the first displayed equation in [57, Lemma 15(i)]. Hence, the remaining steps can be carried through unchanged.

Proof of (ii): This follows verbatim as in [57, Lemma 15(iii)].

Proof of (iii): Here, additional work is required compared to [57]. The proof is based on the following observations:

**Remark 3.7.3.** On \(\{ (y, q) : y \in (0, 1), q < \frac{1}{y} \}\), we have \(f(y, q) \leq f_{NT}(y, q)\).

**Remark 3.7.4.** There are \(\alpha, \zeta > 0\), and \(\eta < 0\) independent of \(\lambda\) and \(\varepsilon\) such that

\[
f(y, q) \leq f_{NT}(y, q) \leq \eta < 0, \quad \text{on } [y_* - \alpha, y_* + \alpha] \times [-\zeta, \zeta].
\]

**Proof of Remark 3.7.4.** On \([y_-, y_+]\), Equation (3.3.1) can be rewritten as

\[
y^2(1 - y)^2 f_{NT}(y, q) = -c + k_1(y, q) + k_2(y, q),
\]

where \(\lim_{y \to y_*} k_1(y, q) = 0\) and \(\lim_{q \to 0} k_2(y, q) = 0\) uniformly. In particular, there is a negative constant \(\eta\) such that, if \(y\) is sufficiently close to \(y_*\) and \(q\) to 0, we have \(f_{NT}(y, q) \leq \eta\).

A straightforward computation shows that one can choose \(\lambda_1\) and \(\varepsilon_1\) small enough such that, for any \(\lambda \leq \lambda_1\) and \(\varepsilon \leq \varepsilon_1\):

- \(\frac{d}{dy}[y^2(1 - y)^2 f_{NT}(y, b_0(\lambda, \varepsilon))] < 0\) on \((0, y_* - \alpha)\);
- \(\frac{d}{dy}[y^2(1 - y)^2 f_{NT}(y, b_1(\lambda, \varepsilon))] > 0\) on \((y_* + \alpha, 1)\);
- \(\max\{b_0, -b_1\} < \zeta\) and \(\eta < \frac{b_1(\lambda, \varepsilon) - b_0(\lambda, \varepsilon)}{2\alpha} < 0\).

From the first point, and since \(y^2(1 - y)^2 f_{NT}(y, b_0(\lambda, \varepsilon)) = 0\) for \(y = 0\), we get \(f_{NT}(y, b_0(\lambda, \varepsilon)) < 0\) on \((0, y_* - \alpha)\). Remark 3.7.3 implies that \(f(y, b_0(\lambda, \varepsilon)) < 0\) on \((0, y_* - \alpha)\).

The second point, with the same arguments, yields \(f(y, b_1(\lambda, \varepsilon)) < 0\) on \((y_* + \alpha, 1)\).

Consider now the line \(q = \frac{b_1 - b_0}{2\alpha} (y - y_* - \alpha) + b_1\). Given a solution \(q(y)\), the third point and Remark 3.7.4 imply that if \(q(y_* - \alpha) < b_0\) then \(q(y) < \frac{b_1 - b_0}{2\alpha} (y - y_* - \alpha) + b_1\) on \([y_* - \alpha, y_* + \alpha]\). Define the following function:

\[
g(y) = \begin{cases} b_0, & y \in (0, y_* - \alpha), \\ \frac{b_1 - b_0}{2\alpha} (y - y_* - \alpha) + b_1, & y \in [y_* - \alpha, y_* + \alpha], \\ b_1, & y \in (y_* + \alpha, 1). \end{cases}
\]  

(3.7.5)
We have just shown that $f(y, g(y)) < g'(y)$ on $(0, 1) \setminus \{y_0 - \alpha, y_0 + \alpha\}$, thus $q_0(y) < g(y)$ on the definition interval of $q_0$. Analogously, $q_1(y) > g(y)$ and so $q_0(y) < q_1(y)$ on their common definition interval.

This proves that there exists $\beta = \frac{\mu^2}{2\gamma^2} - c(\varepsilon, \lambda)$ and a solution $q(y)$ satisfying conditions (3.7.2) and (3.7.3). Finally, we show that the set \(\{y : -\varepsilon < \frac{q(y)}{1+y} < \varepsilon\}\) is an interval \([y_-, y_+]\). In particular, it is enough to show that the solution $q(y)$ crosses the curves $q = \frac{\varepsilon}{1+\varepsilon y}$ and $q = -\frac{\varepsilon}{1-\varepsilon y}$ just once, in $y_-$ and $y_+$, respectively.

The equation $f(y, \frac{\varepsilon}{1+\varepsilon y}) = \frac{d}{dy}(\frac{\varepsilon}{1+\varepsilon y})$ has exactly two solutions $y_2 < y_3$ in $(0, 1)$. In particular, $f(y, \frac{\varepsilon}{1+\varepsilon y}) > \frac{d}{dy}(\frac{\varepsilon}{1+\varepsilon y})$ on $(0, y_2) \cup (y_3, 1)$ and $f(y, \frac{\varepsilon}{1+\varepsilon y}) < \frac{d}{dy}(\frac{\varepsilon}{1+\varepsilon y})$ on $(y_2, y_3)$. Since $q(0^+) \geq \varepsilon$ and $q(1^-) < \frac{\varepsilon}{1+\varepsilon}$, the solution $q(y)$ crosses $q = \frac{\varepsilon}{1+\varepsilon y}$ just once in $y_- \in (y_2, y_3)$. By the same arguments, $q(y)$ crosses $q = -\frac{\varepsilon}{1-\varepsilon y}$ just once in $y_+$.

**Remark 3.7.5.** Unlike for small proportional transaction costs [62, 43], the frictionless Merton proportion $y_*$ does not generally lie in the no-trade region $[y_-, y_+]$ in the present setting. To see this, recall from Guasoni and Weber [57, Remark 20] that in their model with price impact – but without proportional transaction costs – turnover is zero at exactly one point $y_1$, which is $\mathcal{O}(\lambda^{1/2})$-close but not identical to $y_*$ for small $\lambda$. For a given small price impact $\lambda$, the trading boundaries $y_-, y_+$ converge to $y_1$ as $\varepsilon \downarrow 0$. Hence, $y_* \notin [y_-, y_+]$ if the transaction cost is sufficiently small compared to the price impact. However, numerical evidence indicates that this effect only appears if the ratio $\varepsilon/\lambda$ is extremely small. Otherwise, the Merton proportion is contained in the no-trade region, compare Figures 3.1 and 3.2.

As shown in [57, Theorem 3], levered or short positions in the risky asset cannot be admissible with linear price impact. This remains true in the present setting with additional proportional costs:

**Lemma 3.7.6.** Let $u$ be an admissible strategy. Then, for sufficiently small $\varepsilon$ and $\lambda$, the corresponding risky weight

\[
\begin{aligned}
dY_t &= (Y_t(1 - Y_t)(\mu - Y_t\sigma^2) + (u(Y_t) + \varepsilon Y_t|u(Y_t)| + \lambda Y_t u(Y_t)^2))dt \\
&+ Y_t(1 - Y_t)\sigma dW_t, \\
Y_0 &= y \in (0, 1),
\end{aligned}
\]  

(3.7.6)

(3.7.7)

takes values in $[0, 1]$ a.s. for all $t$.

**Proof.** With minor modifications, the assertion follows along the lines of [57, Lemma 9, Lemma 10, Lemma 11, and Theorem 3]. For the sake of completeness, we briefly recall the main ideas here. Let $u$ be any admissible strategy. First,

\footnote{In the proof of [57, Lemma 9], use $-(1 - \varepsilon y)/4\lambda$ instead of $-1/4\lambda$ in the definition of $\bar{\mu}$; then the proof can be carried through along the same lines. In the proof of Lemma 3.7.6, our analogue of [57, Theorem 3], proportional transaction costs lead to an additional term $\int_0^T |\theta_t|dt$ in the numerator of the expression analyzed in [57, Lemma 11]. However, the latter still converges to zero by the same arguments as in the proof of [57, Lemma 11].}
verify that a stochastic process with the dynamics (3.7.6) and initial value \( y \in (1, \infty) \) or \( y \in (-\infty, 0) \), has a finite exploding time \( \tau \) with positive probability, i.e., \( \mathbb{P}[\tau < \infty] > 0 \). In a second step, show that the corresponding wealth process \( X^y \) satisfies, \( X^y(\omega) = 0 \) a.s. on \( \{ \tau < \infty \} \) and \( \theta_x > 0 \) a.s. This in turn implies that any admissible strategy must fulfill \( Y_t \in [0, 1] \) a.s. for all \( t \). \( \square \)

Next, verify that the candidate strategy \( \hat{u} \) from Theorem 3.3.1 is admissible.

**Lemma 3.7.7.** Define \( \beta, q, y_-, y_+ \) as in Lemma 3.7.2 and set

\[
\hat{u}(y) = \begin{cases} 
\frac{1}{2\lambda} \left( \frac{q(y)}{1-q(y)} - \varepsilon \right), & \text{if } y \in [0, y_-), \\
0, & \text{if } y \in [y_-, y_+], \\
\frac{1}{2\lambda} \left( \frac{q(y)}{1-q(y)} + \varepsilon \right), & \text{if } y \in (y_+, 1].
\end{cases}
\]

Then, for sufficiently small \( \varepsilon \) and \( \lambda \), the SDE

\[
dY^\hat{u}_t = \left( Y^\hat{u}_t(1 - Y^\hat{u}_t)(\mu - Y^\hat{u}_t \sigma^2) + (\hat{u}(Y^\hat{u}_t) + \varepsilon Y^\hat{u}_t \sigma)\hat{u}(Y^\hat{u}_t) \right) dt + \lambda Y^\hat{u}_t(1 - Y^\hat{u}_t) \sigma dW_t,
\]

\[
Y^\hat{u}_0 = y \in (0, 1),
\]

has a unique strong solution which takes values in \([0, 1]\) a.s. for all \( t \). In particular, the strategy \( \hat{u} \) is admissible.

**Proof.** Lemma 3.7.2 shows that \( \hat{u}(y) \) is a bounded, continuous function on \([0, 1]\) which satisfies \( \hat{u}(0) > 0 \) and \( \hat{u}(1) < 0 \), i.e., the strategy buys at full investment and sells at zero investment. Furthermore, notice that the scale function of the process \( Y^\hat{u} \) is given by

\[
s(x) = \int_c^x \exp \left[ -2 \int_c^y \frac{z(1-z)(\mu - z\sigma^2) + \hat{u}(z) + \varepsilon |\hat{u}(z)| + \lambda z \hat{u}(z)^2}{z^2(1-z)^2 \sigma^2} dz \right] dy,
\]

for \( c \in (0, 1) \). For sufficiently small \( \varepsilon \) and \( \lambda \), \( \hat{u}(1) + \varepsilon |\hat{u}(1)| + \lambda \hat{u}(1)^2 < 0 \). A straightforward computation shows that \( s(0^+) = -\infty \) and \( s(1^-) = \infty \), so that [76, Proposition 5.5.22] yields the first assertion. Finally, since \( \hat{u} \) and \( Y^\hat{u} \) are both bounded, the admissibility of the strategy follows. \( \square \)

With the function \( q \), the constant \( \beta \), and the boundaries \( y_-, y_+ \) of the no-trade region at hand, we can now use a variant of the verification argument\(^{22}\) of Guasoni and Robertson [56, Theorem 7] to compute an upper bound for the equivalent safe rate of any admissible strategy:

\(^{22}\)The verification argument used in the proof was first used by Guasoni and Robertson [56, Theorem 7] in a general Markovian frictionless setting. It was in turn adapted by Guasoni and Weber [57, Lemma 14] to a Black-Scholes model with quadratic trading costs. Here, we extend it to a setting with quadratic and linear trading costs, which leads to an additional nonlinearity in the HJB equation (3.7.14).
Lemma 3.7.8. Let $y \in (0,1)$ be the initial risky weight. Further, define $\beta, q$ as in Lemma 3.7.2, and set $Q(\xi) = \int_0^\xi q(z)dz$. Then, the terminal wealth $X_u^T$ of any given admissible strategy $u$ satisfies:

$$
\mathbb{E}[(X_u^T)^{1-\gamma}]^{\frac{1}{1-\gamma}} \leq X_0 e^{\beta T + Q(y)} \mathbb{E}_{\hat{\mathbb{P}}^u} \left[e^{-(1-\gamma)Q(Y_u^T)}\right]^{\frac{1}{1-\gamma}},
$$

(3.7.8)

where

$$
\frac{d\hat{\mathbb{P}}^u}{d\mathbb{P}}|_{\mathcal{F}_T} = \mathcal{E} \left( \int_0^T (1 - \gamma) Y_u^u(1 + q(Y_u^u)(1 - Y_u^u))\sigma dW_s \right). 
$$

(3.7.9)

Moreover, equality holds in (3.7.8) for the strategy $\hat{u}$ from Lemma 3.7.7.

Proof. Fix an admissible strategy $u$ and omit the $u-$dependence of $X, Y,$ and $\hat{\mathbb{P}}$ for the sake of clarity in the rest of the proof. Lemma 3.7.2, Lemma 3.7.6, and Novikov’s Condition imply that the stochastic exponential on the right-hand side of (3.7.9) is a true martingale and therefore is the density process of $\hat{\mathbb{P}}^u$ with respect to $\mathbb{P}$.

Now, one readily checks that the assertion follows from

$$
\log X_T - \log X_0 - \frac{1}{1-\gamma} \log \left( \frac{d\hat{\mathbb{P}}}{d\mathbb{P}} \right) \leq \beta T - Q(Y_T) + Q(y).
$$

(3.7.10)

To verify (3.7.10), recall the dynamics of the wealth process $X$ and the risky weight $Y$ from (3.2.4-3.2.5) and apply Itô’s formula to $Q(Y_T)$ and $\log(X_T)$, obtaining:

$$
Q(Y_T) - Q(y) = \int_0^T q(Y_t)(Y_t(1 - Y_t)(\mu - Y_t\sigma^2)) + u_t + \varepsilon|u_t|Y_t + \lambda Y_t u_t^2)dt
+ \int_0^T \frac{1}{2} q'(Y_t)Y_t^2(1 - Y_t)^2\sigma^2 dt + \int_0^T q(Y_t)\sigma dW_t,
$$

(3.7.11)

$$
\log X_T - \log X_0 = \int_0^T \left( Y_t\mu - \varepsilon|u_t| - \lambda u_t^2 - \frac{1}{2}\sigma^2 Y_t^2 \right) dt + \int_0^T \sigma Y_t dW_t.
$$

(3.7.12)

After substituting (3.7.11-3.7.12) into (3.7.10), this inequality reads as

$$
\int_0^T \mu Y_t - \varepsilon|u_t| - \lambda u_t^2 - \frac{1}{2}\sigma^2 Y_t^2 + \frac{1}{2}\sigma^2(1 - \gamma)Y_t^2(1 + q(Y_t)(1 - Y_t))^2 dt
\leq \int_0^T \left( \beta - q(Y_t)(Y_t(1 - Y_t)(\mu - Y_t\sigma^2)) + u_t
+ \varepsilon|u_t|Y_t + \lambda Y_t u_t^2 - \frac{1}{2}\sigma^2 Y_t^2(1 - Y_t)^2q'(Y_t) \right) dt.
$$
Hence, it remains to verify that, for all \( u \in \mathbb{R} \) and \( y \in [0, 1] \):

\[
\mu y - \varepsilon |u| - \lambda u^2 - \frac{\sigma^2}{2} y^2 + \frac{1 - \gamma}{2} \sigma^2 y^2 (1 + q(y)(1 - y))^2
\]

\[
\leq \beta - q(y)(y(1 - y)(\mu - y\sigma^2) + u + \varepsilon |u|y + \lambda yu^2) - \frac{\sigma^2}{2} y^2 (1 - y)^2 q'(y).
\]

(3.7.13)

Rearranging (3.7.13), it suffices to check that, for all \( u \in \mathbb{R} \) and \( y \in [0, 1] \):

\[
0 \geq -\beta + \mu y - \frac{\gamma \sigma^2}{2} y^2 + y(1 - y)(\mu - \gamma \sigma^2 y)q + \frac{\sigma^2}{2} y^2 (1 - y)^2 (q' + (1 - \gamma)q^2)
\]

\[
- \lambda u^2 - \varepsilon |u| + (u + \varepsilon |u|y + \lambda yu^2)q.
\]

(3.7.14)

Maximizing \(-\lambda u^2 - \varepsilon |u| + (u + \varepsilon |u|y + \lambda yu^2)q\) over \( u \) shows that the maximum\(^{23}\) is attained at

\[
\hat{u}(y) = \begin{cases} 
\frac{1}{2\lambda} \left( \frac{q(y)}{1 - yq(y)} - \varepsilon \right), & \text{if } \frac{q(y)}{1 - yq(y)} \geq \varepsilon, \\
0, & \text{if } -\varepsilon \leq \frac{q(y)}{1 - yq(y)} \leq \varepsilon, \\
\frac{1}{2\lambda} \left( \frac{q(y)}{1 - yq(y)} + \varepsilon \right), & \text{if } \frac{q(y)}{1 - yq(y)} \leq -\varepsilon. 
\end{cases}
\]

In view of (3.7.4),

\[
\hat{u}(y) = \begin{cases} 
\frac{1}{2\lambda} \left( \frac{q(y)}{1 - yq(y)} - \varepsilon \right), & \text{if } y \in [0, y_-), \\
0, & \text{if } y \in [y_-, y_+], \\
\frac{1}{2\lambda} \left( \frac{q(y)}{1 - yq(y)} + \varepsilon \right), & \text{if } y \in (y_+, 1]. 
\end{cases}
\]

Now, the inequality (3.7.14) follows after substituting the ODE (3.3.1) for \( q \) and using the maximality of \( \hat{u} \). Evidently, this inequality becomes an equality for the strategy \( \hat{u} \) from Lemma 3.7.7.

To complete the proof of Theorem 3.3.1 we now verify that, as \( T \to \infty \), the upper bound in Lemma 3.7.8 converges to \( \beta \) for any admissible strategy, and is attained for \( \hat{u} \) from Lemma 3.7.7.

**Proof of Theorem 3.3.1.** Let \( \beta \) and \( q \) be defined as in Lemma 3.7.2 and let \( u \) be an arbitrary admissible strategy. By Lemma 3.7.6, we have \( Y^u_t \in [0, 1] \) for all \( t \). As \( q \) is bounded on \([0, 1]\) due to Lemma 3.7.2, the function \( Q(\xi) = \int_0^\xi q(z)dz \) is also bounded on \([0, 1]\). Thus, for every admissible strategy, we have

\[
\lim_{T \to \infty} \frac{1}{(1 - \gamma)T} \log \mathbb{E}^{\hat{u}_T} [e^{-(1 - \gamma)Q(Y^T - Q(y))}] = 0.
\]

\(^{23}\)The condition \( q(y)y < 1 \) on \([0, 1]\) guarantees that the critical point is indeed a maximum.
As \( T \to \infty \), Lemma 3.7.8 therefore provides a strategy-independent upper bound for the equivalent safe rate:

\[
\operatorname{ESR}_\gamma(u) = \lim_{T \to \infty} \frac{1}{(1 - \gamma) T} \log \mathbb{E}[(X_T^n)^{1 - \gamma}] \leq \beta.
\]

This upper bound is attained for the admissible strategy \( \hat{u} \) from Lemma 3.7.7. Hence, the latter is long-run optimal with equivalent safe rate \( \beta \) as claimed. \( \square \)

### 3.7.2 Proof of Proposition 3.6.1

In this section, we establish existence and uniqueness for the inhomogeneous Riccati equation that determines the small-cost asymptotics in Section 3.6. To this end, we first prove an auxiliary result about Riccati ODEs:

**Lemma 3.7.9.** The Riccati equation

\[
y'(x) = f(x, y(x)) := -ax^2 + b + cy^2(x), \tag{3.7.15}
\]

with \( a, b, c > 0 \) has a unique solution such that

\[
\lim_{x \to \infty} \frac{y(x)}{\sqrt{a/cx}} = 1. \tag{3.7.16}
\]

Furthermore, in equation (3.7.15) replace the parameter \( b \) with \( b_1 \) or \( b_2 \), and consider the corresponding unique solutions \( y_1(x) \) and \( y_2(x) \) that satisfy (3.7.16). If \( b_1 < b_2 \), then \( y_1(x) > y_2(x) \).

**Proof.** On \((\sqrt{b/a}, +\infty)\), define the function \( h(x) := \sqrt{ax^2/c - b/c} \). Notice that by definition of \( h(x) \) we have \( f(x, h(x)) = 0 \). For each \( \bar{x} \in (\sqrt{b/a}, +\infty) \) consider the solution \( y(x; \bar{x}, h(\bar{x})) \) with initial condition \((\bar{x}, h(\bar{x}))\) and define \( y_*(x) := \sup\{y(x; \bar{x}, h(\bar{x})) : \bar{x} \in (\sqrt{b/a}, +\infty)\} \).

For any \( x_1 \) there is a large \( y_1 \) such that the linear function \( \tilde{y}(x) = y_1 + \sqrt{a/c}(x - x_1) \) is a subsolution to (3.7.15) whose graph does not intersect the graph of \( h(x) \). In particular, the solution \( y(x; x_1, y_1) \) to (3.7.15) with initial condition \((x_1, y_1)\) is (on its definition interval) strictly larger than \( h(x) \). Since for any \( \bar{x} \) in its definition interval \( y(\bar{x}; x_1, y_1) > h(\bar{x}) = y(\bar{x}; \bar{x}, h(\bar{x})) \), we have also that \( y_1 = y_1(x_1; x_1, y_1) > y_1(x_1; \bar{x}, h(\bar{x})) \) and thus \(+\infty > y_1 \geq y_*(x_1)\). This argument can be repeated for any \( x_1 \in (\sqrt{b/a}, +\infty) \), hence \( y_*(x) < +\infty \) on \((\sqrt{b/a}, +\infty)\).

We want to prove that \( y_*(x) \) is the unique solution that satisfies (3.7.16). By construction, \( y_*(x) \) has the following properties:

i) \( y_*(x) \geq h(x) \);

ii) \( (\sqrt{b/a}, +\infty) \subset D \), where \( D \) is the domain of \( y_*(x) \).

From property (i) it follows that \( \liminf_{x \to \infty} \frac{y_*(x)}{\sqrt{a/cx}} \geq 1 \). Next we show that \( L := \lim_{x \to \infty} \frac{y_*(x)}{\sqrt{a/cx}} \) exists. Assume \( \limsup_{x \to \infty} \frac{y_*(x)}{\sqrt{a/cx}} =: M \in (1, +\infty) \) (the case \( M = \)
+∞ is analogous). Then, there is a sequence \((x_n)_{n \geq 0}\) such that \(\lim_{n \to \infty} \frac{y_*(x_n)}{\sqrt{a/cx_n}} = M\). In particular, for any \(\delta \in (0, M - 1)\) there exists \(N_\delta \in \mathbb{N}\) such that for all \(n \geq N_\delta\) we have \(y_*(x_n) \geq (M - \delta)\sqrt{a/cx_n}\). For large \(x\), the function \(s(x) = (M - \delta)\sqrt{a/cx}\) is a subsolution to (3.7.15), because

\[
(M - \delta)\sqrt{\frac{a}{c}} = s'(x) \leq -ax^2 + b + cs^2(x) = ax^2((M - \delta)^2 - 1) + b.
\]

Thus, for every \(\delta \in (0, M - 1)\) and some \(\bar{x}\), we have \(y_*(x) \geq (M - \delta)\sqrt{a/cx}\) for \(x \geq \bar{x}\). In particular, \(\liminf_{x \to \infty} \frac{y_*(x)}{\sqrt{a/cx}} \geq M - \delta\) for any small \(\delta\), and \(\liminf_{x \to \infty} \frac{y_*(x)}{\sqrt{a/cx}} = M = \limsup_{x \to \infty} \frac{y_*(x)}{\sqrt{a/cx}}\). In other terms, the limit \(L\) exists.

We prove next that \(L = 1\). First, assume by contradiction that \(1 < L < +\infty\). Since \(\lim_{x \to \infty} \frac{y_*(x)}{\sqrt{a/cx}} = L < +\infty\), the function \(y_*(x)\) grows linearly. On the other hand from (3.7.15) one gets

\[
\lim_{x \to \infty} \frac{y_*'(x)}{ax^2} = L - 1 > 0,
\]

which implies that \(y_*'(x)\) grows quadratically, leading to a contradiction.

Assume now that \(L = +\infty\). From (3.7.15) it follows that

\[
\lim_{x \to \infty} \frac{y_*'(x)}{cy_*^2(x)} = 1.
\]

For small \(\delta\) and sufficiently large \(x\), we have \((1 - \delta)cy_*^2(x) \leq y_*'(x)\). This implies that \(y_*(x)\) is bounded from below by a positive function of the form \(\frac{1}{k - (1 - \delta)x}\) for some \(k > 0\). In particular, \(y_*(x)\) would have a vertical asymptote, contradicting property (ii). This proves that \(L = 1\).

The next step is to prove uniqueness. Consider a sufficiently small \(\delta > 0\) and \(\bar{x}\) such that for any \(x \geq \bar{x}\):

\[
\frac{1}{\sqrt{acx^2}} \leq \delta \quad \text{and} \quad \frac{y_*(x)}{\sqrt{a/cx}} \geq 1 - \delta.
\]

For any \(d > 0\), consider the function \(w(x; d) = y_*(x) + dx\). We now show that for \(x \geq \bar{x}\), \(w(x; d)\) is a subsolution to (3.7.15), i.e., \(w'(x; d) \leq f(x, w(x; d))\). Since \(y_*(x)\) is a solution to (3.7.15), this inequality is equivalent to \(d \leq cd^2x^2 + 2cdy_*(x)x\), which can be rearranged to

\[
\frac{1}{\sqrt{acx^2}} \leq \sqrt{\frac{c}{a}}d + 2\sqrt{\frac{c}{a}}\frac{y_*(x)}{x}.
\]

Since \(x \geq \bar{x}\), this inequality follows from \(\delta \leq \sqrt{\frac{c}{a}}d + 2\sqrt{\frac{c}{a}}\), given that \(\delta\) was chosen appropriately. Thus, \(w(x; d)\) is a subsolution for any \(d > 0\). In particular, let \(y_2(x) > y_*(x)\) be a solution to (3.7.15) and choose \(d_\star\) such that \(y_2(\bar{x}) = y_*(\bar{x}) + d_\star \bar{x}\). Then \(y_2(x) \geq w(x; d_\star) = y_*(x) + d_\star x\) for \(x \geq \bar{x}\) and \(y_2(x)\) cannot satisfy (3.7.16).
Since any solution smaller than \( y_*(x) \) is also for large \( x \) smaller than \( h(x) \) and thus eventually decreasing, this is enough to prove uniqueness.

Finally, define as before \( h_1(x) = \sqrt{ax^2/c - b_1/c} \) and \( h_2(x) = \sqrt{ax^2/c - b_2/c} \). Since \( h_1(x) > h_2(x) \) and since any solution to equation (3.7.15) with coefficient \( b_1 \) is a subsolution for equation (3.7.15) with coefficient \( b_2 \), the solution \( y_1(x, \bar{x}) \) to the first equation with initial condition \( (\bar{x}, h_1(\bar{x})) \) is above the solution \( y_2(x, \bar{x}) \) to the second equation with initial condition \( (\bar{x}, h_2(\bar{x})) \) on \( (-\infty, \bar{x}) \). It follows that \( y_1(x) > y_2(x) \).

**Proof of Proposition 3.6.1.** For any \( l > 0 \), define \( r_B(z; l) \) as the unique solution of (3.6.6) that satisfies

\[
\lim_{z \to -\infty} \frac{r_B(z)}{\sqrt{2K\gamma \sigma^2 z}} = 1. \tag{3.7.17}
\]

Here, existence and uniqueness of this solution follow from Lemma 3.7.9. Let \( r(z; l) \) be the unique solution to (3.6.7) with initial condition \( r(0; l) = 0 \). It is enough to prove that there exists a unique \( l \) such that for some \( z_- < 0 \) we get \( r_B(z_-; l) = r(z_-; l) = 1 \), \( r_B(z_-; l) > 1 \) on \( (-\infty, z_-) \) and \( r(z_-; l) < 1 \) on \( (z_-, 0) \).

Define \( l_\lambda := \sqrt{\frac{\gamma K}{2}} \sigma^3 y_*^2 (1 - y_*)^2 \) and \( l_e := \left( \frac{3}{4} \sqrt{\frac{\gamma}{2}} \sigma^3 y_*^2 (1 - y_*)^2 \right)^{2/3} \). If the following values exist, define \( z_B(l) \) the (only) value such that \( r_B(z_B(l); l) = 1 \) and \( r_B(z_B(l); l) < 0 \), and \( z_{NT}(l) \) the (only) value such that \( r(z_{NT}(l); l) = 1 \) and \( r'(z_{NT}(l); l) < 0 \).

If \( l = l_\lambda \), the solution to ODE (3.6.6) with boundary condition (3.7.17) is

\[
r_B(z; l_\lambda) = -\sqrt{2K\gamma \sigma z} + 1.
\]

Thus, \( z_B(l_\lambda) = 0 \). From Lemma 3.7.9 it follows that if \( l_1 < l_2 \), then \( r_B(z; l_1) > r_B(z; l_2) \) and thus \( z_B(l_1) > z_B(l_2) \). Hence, the function \( l \to z_B(l) \) is well-defined and decreasing on \( [l_\lambda, +\infty) \). Furthermore, \( \lim_{l \to +\infty} z_B(l) = -\infty \).

Since \( \max_{z \in (-\infty, 0)} r(z; l) < 1 \) if and only if \( l < l_e \), the function \( l \to z_{NT}(l) \) is well-defined only on \( [l_e, +\infty) \). Furthermore, it is increasing and \( \lim_{l \to +\infty} z_{NT}(l) = 0 \).

Let \( l_M := \max\{l_\lambda, l_e\} \). If \( l_M \geq l_e \), then \( z_{NT}(l_\lambda) < z_B(l_\lambda) = 0 \). A brief calculation shows that \( z_{NT}(l_\lambda) = -\sqrt{\frac{2l}{\gamma \sigma}} \). Thus, for \( (z, r) \in (-\infty, z_{NT}(l_\lambda)) \times \{1\} \) any solution to (3.6.6) has strictly positive derivative, while it is strictly negative for \( (z, r) \in (z_{NT}(l_\lambda), 0) \times \{1\} \). In particular, since \( r_B'(z_B(l); l) < 0 \), if \( l \geq l_\lambda \), then \( z_B(l) \in (z_{NT}(l), 0) \). In both cases, \( z_{NT}(l_M) < z_B(l_M) \). Given the monotonicity of the functions \( z_{NT}(l) \) and \( z_B(l) \), there exists a unique \( l_* \in (l_M, +\infty) \) such that \( z_{NT}(l_*) = z_B(l_*) \).

Without price impact the no-trade region is given by \( (z_{NT}(l_e), -z_{NT}(l_\lambda)) \) (see [43, Formula (2.9)]). Since \( z_{NT}(l) \) is an increasing function, the no-trade region with price impact is strictly smaller: \( (z_{NT}(l_*), -z_{NT}(l_\lambda)) \subset (z_{NT}(l_\lambda), -z_{NT}(l_\lambda)) \).
Chapter 4

Transaction Costs and Continuous Dividends

4.1 Introduction

Buying and holding a portfolio mostly invested in stocks is typical advice for long-term investors. Prominent academic voices, like [114] and [92], argue at length in support of buy-and-hold strategies. Warren Buffett [17] famously quipped that “our favorite holding period is forever”.

Yet, a closer look at theoretical work reveals rather different conclusions. Dynamic portfolio choice models, since Merton’s [98, 99] and its numerous extensions, imply optimal policies that frequently buy and sell securities, as to keep portfolio weights close to a desired target. Proportional transaction costs\(^1\) entail large no-trade regions, but still dictate selling stocks when their portfolio weight is too high. Overall, in the portfolio choice literature, with or without frictions, the only buy-and-hold investor has only risky assets and no interest in safe assets — due to very particular preferences.

This chapter proves that buying and holding stocks, not selling them is optimal for a realistic range of market and preference parameters, if investment opportunities are constant, transaction costs are proportional, and — crucially — stocks have a positive constant dividend yield. For example, for an investor who would hold 90\% of wealth in stocks in the absence of frictions,\(^3\) transaction costs of 1\% and dividends of 3\% make it optimal to buy stocks whenever their weight falls below 90\%, and to always refrain from selling. In general, if both transaction costs and dividends are present, the optimal policy for a long-term investor can be two of types: if the dividend yield is very low, then the no-trade region lies within two boundaries, one for buying and one for selling, as in [37]. Increasing the dividend payout while holding the total return constant makes the

---

\(^1\)This chapter is based on Guasoni, Liu and Muhle-Karbe [51].
\(^2\)For portfolio choice with transaction costs without dividends, see [91, 22, 31, 37, 113].
\(^3\)The example is not artificial. Writing about his will, Buffett [18] plainly states “My advice to the trustee could not be more simple: Put 10\% of the cash in short-term government bonds and 90\% in a very low-cost S&P 500 index fund.”
sells the boundary rise quickly. When the dividend yield reaches a critical level, it becomes optimal to buy stocks when their weight is too low, otherwise hold them, and never sell.

The main intuition is that dividends have an automatic rebalancing effect that reduces the need for selling. The reason is not that dividends provide a cash flow, because even the interest rate generates a cash flow, without altering the conclusions of the Merton model. Instead, the point is that a constant dividend yield produces a cash flow proportional to stock holdings, which makes the weight of cash grow quickly when it is low, and slowly when it is high – portfolio weights are mean-reverting. Such mean-reversion can make selling redundant through two channels: first, the stock portfolio weight spends more time near its frictionless target, reducing the need to sell. Second, even when the portfolio weight departs significantly from the target, waiting for it to revert towards the target can be wiser than selling stocks.

Even then, the question of whether selling is ever optimal remains a quantitative matter, and its answer critically depends on parameter values. Indeed, never-sell strategies are optimal if transaction costs and dividends are large, and the frictionless target weight is close to one. But even when these conditions are not met, never selling is close to optimal in that, even if the optimal policy entails selling, ignoring its selling signals leads to a negligible loss compared to the loss of ignoring buying signals.

Our main result has three broad implications. First, it shows that portfolio rebalancing is a weak motive for selling stocks, since even small dividends can make selling suboptimal when combined with trading costs. The conventional wisdom is that dividends are irrelevant, since a dividend-paying asset is equivalent to another asset, which pays no dividends, but has a higher expected return. In the context of valuation, this is the Miller-Modigliani Theorem [100] for frictionless markets. The portfolio choice literature has largely followed this tradition, focusing on total returns, and neglecting dividends – even in models with frictions.4 We argue that frictions make dividends relevant for portfolio choice, and that optimal trading policies are very sensitive to dividend yields in addition to total expected returns.

Second, we show that a more complex model can actually lead to a simpler trading policy, as the combined effect of two small deviations from the benchmark model deviates from the overlap of their separate impacts. To wit, dividends alone are irrelevant for frictionless models, and transaction costs alone imply a no-trade region between a buy and a sell boundary. When taken together, however, dividends and transaction costs can cause the sell boundary to disappear, making dynamic buy-and-hold optimal. This result casts in a cautionary light the established practice of studying market frictions as arbitrarily small deviations from the classical benchmark, to be studied asymptotically.5 While asymptotics

4The only exception is [64], who solves numerically a model similar to ours for some combinations of parameter values that yield solutions similar to the case without dividends.
5Early references include [113, 122, 62]. For more recent developments, cf. [117, 71, 34] and the references therein.
are a crucial analytical tool, it is equally important to understand the range of parameters for which such approximations are accurate, and the potential regime shifts that may occur when they are not.

Third, we present the first closed-form solution to a portfolio choice problem with capital gain taxes. With hindsight, this result is a simple or even obvious observation, because capital gains cannot arise without selling. Hence, if a buy-and-hold strategy is optimal without accounting for capital gains taxes – which may only decrease the performance of other competing strategies – it is a fortiori also optimal when such taxes are taken into account. More importantly, although a model that neglects taxes on capital gains may suggest that selling is optimal in some circumstances, such a policy may perform more poorly than a similar alternative that never sells. We investigate this issue through Monte-Carlo simulations, and find that capital-gains taxes expand the range of parameter values for which dynamic buy-and-hold is almost optimal. Although the additional welfare is modest because transaction costs already imply wide no-trade regions, and hence infrequent realizations of capital gains, the novelty is that a strategy that foregoes selling altogether is close to optimal, and its performance is robust to any capital gain tax structure, unlike the performance of strategies that involve selling.

We qualify the scope of the optimality of dynamic buy-and-hold strategies by investigating the conditions under which selling becomes part of the optimal trading policy. Two main selling motives stand out: high risk aversion and intertemporal consumption. Highly risk-averse long-term investors should sell stocks, as the rebalancing provided by dividends is not enough to keep the portfolio weight near its desired target – a result that we prove analytically. When capital gains taxes are not considered, selling is optimal for investors who would ideally (i.e., if trading were costless) hold approximately 50% of their wealth in stocks, or less.

Intermediate consumption can be a valid reason to sell stocks even for more risk-tolerant investors, as we find with Monte-Carlo simulations. The main intuition is straightforward: while dividends over time tilt a portfolio towards cash, consumption partially offsets this effect by draining cash at an approximately constant rate. The question is whether dividends can support both consumption and rebalancing needs, and the answer is rather sensitive to both the dividend yield and the desired consumption rate. For example, consider an investor who aims at an allocation in stocks of 50% and requires a consumption stream of about 3% a year.\(^7\) Then, assuming typical market parameters, it turns out that a never sell policy is inferior to a policy that buys and sells for a 3% dividend yield, but is close to optimal – and superior to alternative tractable policies – for a 4% dividend yield.

These observations are relevant for retirement planning. They suggest that

---

\(^6\) For numerical studies cf., e.g., [38, 30, 35, 118, 28]; also compare [19] for an asymptotic analysis of a particular model.

\(^7\) To be precise, the figures in the first row in Table 4.3 below correspond to a consumption rate of 2.84% per year, obtained from equation (4.2.6) with \(\gamma = 6.25\) and \(\rho = 2\%\).
Transaction Costs and Continuous Dividends

A broad range of investors with moderate risk aversion should avoid selling securities for rebalancing purposes before retirement, as buy-and-hold policies are optimal when the horizon is long and intermediate consumption is absent. On the other hand, at the onset of retirement it may become optimal to sell stocks to finance required payouts, unless dividend yields are sufficiently high to support both rebalancing and consumption. This result supports in part the preference of income-seeking investors for high-dividend funds, which is unfounded in frictionless models.

The broad optimality of strategies that never sell also sheds new light on the value of tax-sheltered or tax-deferred retirement plans, such as 401(k)s and Roth accounts. Assets held inside these vehicles are shielded from taxes on the dividends that they generate and on gains realized from trades, as long as funds remain in the account. For investors who divide their investments into a bond and a stock fund, our results indicate that in these accounts the tax shelter on dividends is more important before retirement than the one on capital gains, because selling is not optimal.

Our analysis assumes for simplicity that investors are infinitely lived, but accounting for finite lives would strengthen our conclusions. In the United States and other jurisdictions, the cost basis of stocks is reset at their current value upon an investor's death, thereby exempting heirs from taxes on capital gains accumulated during the investor's life, estate taxes notwithstanding. For investors with a bequest motive, this exemption is an additional advantage of policies that never sell.

The remainder of this chapter is organized as follows: Section 4.2 describes the model, discusses the main results and their economic implications, and then provides their formal statement. Section 4.3 derives the optimal trading strategy using heuristic control arguments and all proofs are contained in Section 4.4.

4.2 Model and Main Results

4.2.1 Model

The market includes a safe asset earning a constant interest rate \( r \) and a risky asset that trades at price \( S_t \), subject to a proportional transaction cost \( \varepsilon S_t \) per share, for some \( \varepsilon \in (0, 1) \). Thus, the effective bid and ask prices are \((1-\varepsilon)S_t\) and \((1+\varepsilon)S_t\), respectively. The risky asset pays a continuous dividend stream at a rate \( \delta S_t \) (paid out in cash), hence \( \delta > 0 \) represents its dividend yield, also assumed constant.

The instantaneous return of the risky asset \( dS_t/S_t \) is decomposed into an expected ex-dividend return \( \mu - \delta + r \) and a normally distributed noise component with constant volatility \( \sigma > 0 \):

\[
\frac{dS_t}{S_t} = (\mu - \delta + r)dt + \sigma dW_t, \quad (4.2.1)
\]

where \( W_t \) is a standard Brownian motion, and \( \mu \) represents the expected total return in excess of the safe rate. Without transaction costs \( (\varepsilon = 0) \), the value of
the dividend yield $\delta$ is irrelevant: a higher $\delta$ implies higher dividend payments that are compensated by lower expected returns, and a dynamic strategy of share repurchases can replicate the return on an asset with the same total return $\mu + r$ but $\delta = 0$. With transaction costs, a share repurchase strategy would only deliver an expected return of $\mu - \delta \frac{\varepsilon}{1+\varepsilon}$. Hence, dividends become a relevant aspect of investment opportunities.

As in [37, 49, 50], we consider an investor with constant relative risk aversion $\gamma > 0$, who maximizes expected utility from terminal wealth over a long horizon. Performance is measured in terms of a hypothetical equivalent safe rate at which a full safe long-run investment yields the same expected utility as trading optimally in the original market.

Without transaction costs ($\varepsilon = 0$), the dynamics of wealth does not depend on dividends, but only on the total return, so that the optimal policy is to hold the risky weight equal to the constant Merton proportion $\pi^* = \mu/(\gamma \sigma^2)$, as without dividends [98, 99]. In view of this observation, the standard assumption in portfolio choice models is to omit dividends, even though such an approach is suspicious when trading costs are present — and dividends relevant.

With transaction costs — but without dividends — it is optimal to keep the risky weight in a no-trade region around the frictionless target weight $\pi^*$ [91, 31, 37, 113]. No trading takes place inside this interval, while the investor buys or sells at each boundary as little as necessary to keep the risky weight within the interval.

### 4.2.2 Main Results

The main result clarifies the interplay of dividends and transaction costs in determining optimal policies and their corresponding welfare. For the sake of clarity we first explain these results informally, discussing their robustness to taxes and consumption, and then provide their precise statements in Section 4.2.3. Throughout, we assume that the optimal frictionless portfolio is neither short nor leveraged:

$$0 < \pi^* < 1.$$  

This condition essentially amounts to assuming that the equity premium is positive, but not so high that the investor wants risky investments in excess of her own capital. When this condition is not satisfied, selling the risky asset becomes inevitable in order to prevent bankruptcy — buy and hold is not an option.

**Optimal Policy**

Our main observation is that — with proportional transaction costs $\varepsilon$ and dividends $\delta$ — it is optimal to *never sell* the risky asset for a broad range of realistic parameter values. Then, the optimal policy is fully described by a minimum risky weight $\pi_-$, at which the investor buys stocks to keep their portfolio weight above this level, and otherwise holds them forever.

This phenomenon crucially requires sufficiently large dividend yields. With
transaction costs but without dividends, the optimal policy is determined by buying and selling boundaries \( \pi_-, \pi_+ \), symmetric around the frictionless target weight \( \pi^* \) [43]. The left panel in Figure 4.1 shows the effect of a positive dividend yield \( \delta \). For small \( \delta \), both trading boundaries increase by a similar amount, counteracting the dividend cash flow to keep the average portfolio weight near the frictionless level. However, the buying boundary quickly stabilizes near the frictionless target weight, whereas the selling boundary continues to increase, reaching the full investment level at around \( \delta = 2\% \). At that point, the optimal strategy does not involve selling any more, because a long-only portfolio cannot reach full investment without trading.

These comparative statics crucially depend on the Merton proportion \( \pi^* \), which equals 90.6% and therefore is quite close to the full investment level in the above example. To shed more light on the joint dependence on the dividend yield and the frictionless target, the right panel in Figure 4.1 shows the combinations of both variables for which it is optimal never to sell the risky asset. Lower Merton proportions require larger dividend yields to maintain the optimality of a never-sell strategy, as a larger shift in the selling boundary becomes necessary to reach full investment.

The main intuition behind this result stems from the qualitative difference in the dynamics of static strategies when dividends are present. To see this, consider a portfolio with \( X_t \) dollars in cash and \( Y_t \) dollars in stock. Without trades, the joint dynamics of these positions is

\[
\begin{align*}
\frac{dX_t}{rX_t} &= dt + \delta Y_t dt \\
\frac{dY_t}{(\mu - \delta + r)Y_t} &= dt + \sigma Y_t dW_t
\end{align*}
\]

Thus, the stock/cash ratio \( Z_t = Y_t / X_t \), which equals the ratio of portfolio weights \( \frac{Y_t}{X_t} / \frac{X_t}{Y_t} \), by Itô’s formula follows the dynamics

\[
dZ_t = (\mu - \delta - \delta Z_t) Z_t dt + \sigma Z_t dW_t.
\]

Figure 4.1: Left panel: optimal no-trade region (vertical axis) against the dividend yield \( \delta \) (horizontal axis) for \( \pi^* = 90.6\% \). Right panel: never-sell region (shaded) for combinations of dividend yield \( \delta \) (horizontal axis) and frictionless portfolio weight \( \pi^* = \mu / (\gamma \sigma^2) \) (vertical axis). Parameters are \( \mu = 8\% \), \( \sigma = 16\% \) and \( \varepsilon = 1\% \).
4.2 Model and Main Results

\[
\begin{array}{cccc}
\pi_* & \text{optimal} & \text{never sell} & \text{buy & hold} \\
[\pi_-, \pi_+] & [\pi_-, 1] & [0, 1] \\
50\% & 1.67\% & 2.00\% & 4.67\% \\
60\% & 1.76\% & 1.76\% & 4.41\% \\
70\% & 1.58\% & 1.58\% & 4.21\% \\
80\% & 1.43\% & 1.43\% & 3.81\% \\
90\% & 1.52\% & 1.52\% & 3.70\%
\end{array}
\]

Table 4.1: Relative equivalent safe rate loss \((\text{ESR}_0 - \text{ESR})/\text{ESR}_0\) of the optimal \(([\pi_-, \pi_+])\), never sell \(([\pi_-, 1])\) and buy-and-hold \(([0, 1])\) strategies with \(\pi_\pm\) as in (4.2.9-4.2.10), computed using Monte-Carlo simulation with \(T = 20\), time step \(dt = 1/250\), and sample size \(N = 2 \times 10^7\). Parameters are \(\mu = 8\%, \sigma = 16\%, r = 1\%, \delta = 2\%, \) and \(\varepsilon = 1\%\).

This equation clearly displays the mean-reversion effect of dividends on portfolio weights. Without dividends \((\delta = 0)\), the stock/cash ratio has a positive growth rate (as the equity premium \(\mu\) is positive). By contrast, a positive dividend rate makes the stock/cash ratio oscillate around the long-term mean \(Z = Y/X = \mu/\delta - 1\), which corresponds to the long-term average portfolio weight

\[
\pi = \frac{Y}{X + Y} = \frac{Z}{1 + Z} = 1 - \delta/\mu.
\]

(4.2.5)

This formula implies that, if dividends are absent \((\delta = 0)\), a portfolio initially divided equally between stocks and cash, without rebalancing will shift over time most of its weight on stocks \((\pi = 1)\). But with a dividend yield of \(\delta = 2\%\) and an equity premium of \(\mu = 8\%\), the same portfolio will oscillate over time around a weight of \(\pi = 75\%\), as growth in stocks spills over to the cash position.

Thus, the stabilizing effect of dividends reduces the need to sell for two related reasons: first, the portfolio weight has a lower long-term average, hence it is less likely to reach excessively high levels. Second, the portfolio weight automatically reverts to its mean, offering a cheaper alternative to selling – waiting.

**Robustness**

For a given transaction cost \(\varepsilon\) and dividend yield \(\delta\), a never-sell strategy is only optimal if the Merton proportion \(\pi_*\) is sufficiently close to the full-investment level (i.e., one). But even when it is not optimal, we argue that a never-sell policy is almost optimal under rather broad conditions. We use Monte-Carlo simulations to compare the performance of i) the optimal policy with buying and selling boundaries \(\pi_-, \pi_+\), ii) a never-sell strategy with the same buying boundary \(\pi_-\), and iii) a static buy-and-hold strategy that starts from the Merton proportion \(\pi_*\) and never trades. The performance of these strategies is expressed in terms of equivalent safe rates: for each policy, we find the hypothetical safe rate ESR for which a full safe investment would yield the same utility as the policy under
consideration. Table 4.1 reports the relative underperformance of each policy compared to the frictionless equivalent safe rate $\text{ESR}_0 = r + \mu^2/(2\gamma\sigma^2)$.

Even for a relatively low Merton proportion of 50%, the never-sell strategy achieves almost the same performance as the optimal policy. In particular, the loss compared to the optimum is substantially smaller than the gain compared to a buy-and-hold strategy that neither buys nor sells the risky asset.

Taxes

Next, we assess the robustness of never-sell policies to taxes on dividends and capital gains.

The model easily accommodates proportional taxes on dividends. If dividends are taxed at a rate $\tau \in (0, 1)$, then the after-tax dividend rate is $(1 - \tau)\delta$, while the expected, ex-dividend return on the risky asset remains $\mu - \delta$ in excess of the safe rate $r$. In fact, such a model is equivalent to another one without dividend taxes, but with a dividend yield of $\tilde{\delta} = (1 - \tau)\delta$ and expected total return of $\tilde{\mu} = \mu - \tau$, which leads to the same asset dynamics.

The effect of capital gains taxes on investment policies depends on the convention used for the cost basis. We focus on two common methods: with the Weighted Average Cost Basis (adopted in Canada and Italy, and one of the options in the United States), the cost (and hence the tax basis) of each share sold is the average cost of all shares previously acquired (compare [30, 118]). With the Specific Share Identification, also an option in the United States, an investor specifies which shares are being sold, freely choosing among lots with different cost. With this method, any rational investor sells the shares with the highest cost first. In both cases, we assume that capital gains taxes are a constant proportion $\alpha \in (0, 1)$ of the difference between the sale price and the cost basis.

For any specification, capital gains taxes introduce the additional state variables of cost bases, thereby making the computation of optimal strategies analytically intractable in general. Therefore we compare through Monte Carlo simulations, similar to the preceding section, the performance of three alternative strategies. The first one is the optimal strategy with dividends and transaction costs, but without taxes. The second one uses the same buying boundary as the first one, but never sells; the third one is a static buy-and-hold strategy. Table 4.2 reports the relative loss, when the equivalent safe rate of each strategy is compared to the equivalent safe rate in a frictionless setting.\(^8\)

Of course, capital gains taxes are irrelevant for strategies that never sell and thereby never realize capital gains. For the policy that is optimal with dividends and transaction costs (but without taxes), the differences between the two cost-basis specifications turns out to be negligible.

As in the absence of taxes, the never-sell (i.e., dynamic buy and hold) strategy significantly outperforms static buy-and-hold. But accounting for taxes, the never-sell strategy also dominates the policy that is optimal with dividends and transaction costs, but without capital gains taxes, even for Merton weights as

\(^8\)Here, the frictionless benchmark adjusted for the dividend tax $\tau$ is $r + (\mu - \delta\tau)^2/2\gamma\sigma^2$. 


4.2 Model and Main Results

**Table 4.2: Relative equivalent safe rate loss** (ESR\(_{0,\tau} - \text{ESR}\)) with capital gains taxes, for optimal ([\(\pi_-\), \(\pi_+\)]), never sell ([\(\pi_-\), 1]), and buy-and-hold ([0, 1]) strategies with \(\pi_{\pm}\) as in (4.2.9-4.2.10). Rates are computed through Monte-Carlo simulation with \(T = 20\), time step \(dt = 1/250\), and sample size \(N = 2 \times 10^7\). Parameters are \(\mu = 8\%\), \(\sigma = 16\%\), \(\alpha = 20\%\), \(\tau = 20\%\), \(r = 1\%\), \(\delta = 2\%\), and \(\varepsilon = 1\%\). The static buy-and-hold policy starts with a risky portfolio weight equal to \(\pi_* = \mu/\left(\gamma\sigma^2\right)\).

<table>
<thead>
<tr>
<th>(\pi_*) (average)</th>
<th>(\pi_-), (\pi_+) (specific)</th>
<th>never sell</th>
<th>buy &amp; hold</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>2.41%</td>
<td>2.41%</td>
<td>2.07%</td>
</tr>
<tr>
<td>60%</td>
<td>2.13%</td>
<td>2.13%</td>
<td>1.83%</td>
</tr>
<tr>
<td>70%</td>
<td>1.91%</td>
<td>1.91%</td>
<td>1.64%</td>
</tr>
<tr>
<td>80%</td>
<td>1.49%</td>
<td>1.49%</td>
<td>1.49%</td>
</tr>
<tr>
<td>90%</td>
<td>1.36%</td>
<td>1.36%</td>
<td>1.36%</td>
</tr>
</tbody>
</table>

In other words, the ostensible advantage of buying and selling towards the frictionless weight is lost once capital gains are acknowledged. The quantitative welfare gains are small, because the wide no-trade regions implied by transaction costs alone already lead to infrequent sales, and hence modest capital gains. Yet, the qualitative insight is that capital gains taxes significantly expand the range of superiority of the never-sell policy.

**Consumption**

The previous discussion indicates that in many realistic situations it is better to let dividends automatically rebalance a portfolio rather than selling stocks. We now heuristically show that this result is sensitive to intertemporal consumption. Relatively high dividends can support both rebalancing and consumption for investors with certain preferences. However, highly impatient or risk-averse individuals will typically require more consumption than can be supported by typical dividend yields alone. In these cases, selling therefore becomes inevitable to avoid an excessively leveraged portfolio. The broad message for retirement planning is that selling is typically unnecessary until retirement. During retirement, an investor may either use sales to finance consumption while keeping an appropriate amount of risk, or can avoid selling by investing in securities with high dividend yield.

To justify these statements, consider the maximization of lifetime consumption over an infinite horizon,

\[
E \left[ \int_0^\infty e^{-\rho t} \frac{C_t^{1-\gamma}}{1-\gamma} dt \right],
\]

where \(C_t > 0\) is the investor’s consumption rate and \(\rho > 0\) is the rate of time preference. Without frictions, the optimal consumption/wealth ratio is constant
Table 4.3: Relative equivalent safe rate loss \((ESR_0 - ESR)/ESR_0\) of the asymptotically optimal \(([\pi^{JS}_-, \pi^{JS}_+]), never-sell \(([\pi_-, 1])\) and static buy-and-hold \(([0, 1])\) strategies with \(\pi^{JS}_\pm\) as in [62, Theorem 2] and \(\pi_-\) as in (4.2.9), computed using Monte Carlo simulation with \(T = 20\), time step \(dt = 1/250\), and sample size \(N = 2 \times 10^7\). Parameters are \(\mu = 8\%\), \(\sigma = 16\%\), \(\rho = 2\%\), \(r = 1\%\), \(\tau = 0\%\), \(\varepsilon = 1\%\) and \(\delta = 3\%\) (left panel) or \(\delta = 4\%\) (right panel).

\[
\frac{C^*_t}{X_t + Y_t} = \frac{\rho}{\gamma} + \left(1 - \frac{1}{\gamma}\right) \left(r + \frac{\mu^2}{2\gamma\sigma^2}\right), \tag{4.2.6}
\]

where \(X_t\) and \(Y_t\) denote the investor’s safe and risky positions, respectively. Since the same consumption policy is approximately optimal even with small proportional transaction costs [22, 71], we adopt it in combination with each of the investment strategies considered (the exact optimal policies are unknown, and the corresponding problems analytically intractable).

We compare the performances of three investment strategies. The first one is the optimal strategy for consumption with small transaction costs, but without dividends [62, Theorem 2], which entails buying and selling boundaries \(\pi^{JS}_-, \pi^{JS}_+\) symmetric around the frictionless Merton proportion \(\pi_*\). The second one is the never-sell strategy with a buying boundary that takes into account dividends but neglects consumption. The third policy is a simple static buy-and-hold strategy that starts from the frictionless Merton proportion. Performances are reported as equivalent safe rates: for each policy, we identify the hypothetical safe rate which would make an investor indifferent between either investing all wealth at this rate and consuming optimally at rate (4.2.6), or investing in the original market, with the true safe rate, but also risky assets, transaction costs and dividends, and again consuming according to (4.2.6). (Risk-aversion and time-preference are held constant in this comparison, and this equivalent safe rate readily follows by setting \(\mu = 0\) in Formula (2.4) of [62].)

The left panel in Table 4.3 reports the relative losses with respect to the frictionless benchmark, when the dividend yield is 3%. As before, the differences between the three policies are higher when the benchmark portfolio weight is lower. Yet, with intermediate consumption, the never-sell strategy performs only slightly better than static buy-and-hold and is significantly worse than the symmetric rebalancing rule of [62], even though the latter neglects dividends. The
reason is that much of the dividend flow – and possibly more – is expended for consumption, thereby offsetting its rebalancing effect.

However, this conclusion is reversed for a higher dividend yield of 4%, reported in the right panel of Table 4.3, with which the never-sell policy is again preferable to a static buy-and-hold, and essentially equivalent to the asymptotically optimal policy. Overall, the tension between consumption rates and dividend yields is a critical determinant of the relative performance of the never-sell policy – a relatively high yield makes it nearly optimal. This observation supports the preference of income-seeking investors for high-dividend funds, which offer an effective vehicle to achieve automatic rebalancing with passive strategies, while also supporting a consumption stream.

4.2.3 Statement of the Results

This section contains the formal statement of the results discussed above. With transaction costs, a self-financing trading strategy is described by a predictable process of finite variation $\varphi$, where $\varphi_0 = \xi \in \mathbb{R}$ denotes the initial number of shares in the risky asset, and $\varphi_t$ denotes the number of shares at time $t$. Writing $\varphi_t = \varphi_{-t} + \varphi_{t}^+ - \varphi_{t}^-$ as the difference between the cumulative number of shares bought $\varphi_{t}^+$ and sold $\varphi_{t}^-$ by time $t$, the self-financing condition states that the safe position $X_t$ changes only through the cash flow generated by interest and dividends, as well as trades in the risky asset:

$$dX_t = rX_t dt + \delta\varphi_t S_t dt - (1 + \varepsilon)S_t d\varphi_t^+ + (1 - \varepsilon)S_t d\varphi_t^-, \quad \forall t \geq 0. \quad (4.2.7)$$

**Definition 4.2.1.** A self-financing trading strategy $\varphi$ is admissible if its liquidation value is positive at all times:

$$\Xi_t^\varphi := X_t + (1 - \varepsilon)S_t \varphi_t^+ - (1 + \varepsilon)S_t \varphi_t^- \geq 0, \quad \forall t \geq 0.$$ 

As in [37, 49, 50], the investor has constant relative risk aversion $0 < \gamma \neq 1$ and an infinite planning horizon:

**Definition 4.2.2.** An admissible strategy $\varphi$ is long-run optimal if it maximizes the equivalent safe rate

$$\liminf_{T \to \infty} \frac{1}{T} \log \mathbb{E} \left[ (\Xi_T^\varphi)^{1-\gamma} \right]^{\frac{1}{1-\gamma}} \quad (4.2.8)$$

over all admissible strategies, where $0 < \gamma \neq 1$ denotes the investor’s relative risk aversion.

To wit, a full safe long-run investment at the equivalent safe rate yields the same utility as trading optimally in the original market. We now state our main verification theorem, which is valid under the following assumption. Its first case (for which buying and selling are optimal) applies in particular, if the bid-ask spread $\varepsilon$ is small enough. Sufficient conditions for the validity of the second case (in which it is optimal to never sell), are provided in Proposition 4.4.4.

---

*The corresponding results for logarithmic utility are readily obtained by passing to the limit $\gamma \to 1$.**
Assumption 4.2.3. Set

\[ \pi^{-}(\lambda) = \frac{\mu - \varepsilon \delta / (1 + \varepsilon) - \sqrt{\lambda^2 - 2\mu \varepsilon \delta / (1 + \varepsilon) + (\varepsilon \delta / (1 + \varepsilon))^2}}{\gamma \sigma^2}, \]
\[ \pi^{+}(\lambda) = \min \left( \frac{\mu + \varepsilon \delta / (1 - \varepsilon) + \sqrt{\lambda^2 + 2\mu \varepsilon / (1 - \varepsilon) + (\varepsilon \delta / (1 - \varepsilon))^2}}{\gamma \sigma^2}, 1 \right), \]

(4.2.9)

and assume that one of the following two conditions is satisfied:

(a) There exists \( \lambda > 0 \) such that \( \pi^{+}(\lambda) < 1 \) and the (unique) solution \( w(x, \lambda) \) of

\[
0 = w'(x) + (1 - \gamma)w(x)^2 + \left( 2\gamma - 1 - \frac{2(\mu - \delta)}{\sigma^2} + \frac{2\delta}{\sigma^2 e^{\lambda} u(\lambda)} \right)w(x) \\
- \left( \gamma + \frac{\mu^2 - \lambda^2}{\gamma \sigma^4} - \frac{2(\mu - \delta)}{\sigma^2} \right),
\]

\[
w \left( \log \left( \frac{l(\lambda)}{u(\lambda)} \right) \right) = \frac{l(\lambda)}{1 + \varepsilon + l(\lambda)},
\]

with

\[ l(\lambda) = (1 + \varepsilon) \frac{1 - \pi^{-}(\lambda)}{\pi^{-}(\lambda)}, \quad u(\lambda) = (1 - \varepsilon) \frac{1 - \pi^{+}(\lambda)}{\pi^{+}(\lambda)}, \]

also satisfies the following initial condition:

\[ w(0, \lambda) = \frac{u(\lambda)}{1 - \varepsilon + u(\lambda)}. \]

(b) There exists \( \lambda > 0 \) such that \( \pi^{+}(\lambda) = 1 \) and the corresponding (unique) solution \( w(x, \lambda) \) of

\[
0 = w'(x) + (1 - \gamma)w(x)^2 + \left( 1 - 2\gamma + \frac{2(\mu - \delta)}{\sigma^2} - \frac{2\delta e^{\lambda}}{\sigma^2 l(\lambda)} \right)w(x) \\
- \left( \gamma + \frac{\mu^2 - \lambda^2}{\gamma \sigma^4} - \frac{2(\mu - \delta)}{\sigma^2} \right),
\]

\[ 0 = \lim_{x \to \infty} w(x), \]

also satisfies the following initial condition:

\[ w(0, \lambda) = \frac{-l(\lambda)}{1 + \varepsilon + l(\lambda)}. \]

Note that if Condition 4.2.3 is satisfied, then the corresponding \( \lambda \) is unique (cf. Lemma 4.4.12). Our main theorem now reads as follows:
Theorem 4.2.4. Let Assumption 4.2.3 hold. Assume an investor with constant relative risk aversion \(0 < \gamma \neq 1\) maximizes the equivalent safe rate (4.2.8) when the risky asset pays a constant dividend yield \(\delta\) and trades with proportional transaction cost \(\varepsilon\). Then:

(i) Optimal Strategy:
It is optimal to refrain from trading if the risky weight lies between the buying and selling boundaries \(\pi_-\) and \(\pi_+\). At the boundaries, the minimal amount of trading is performed to keep the risky weight inside \([\pi_-, \pi_+]\). Here, the buying and selling boundaries are evaluated at the ask price \((1 + \varepsilon)S_t\) and the bid price \((1 - \varepsilon)S_t\), respectively.

(ii) Equivalent Safe Rate:
For the investor, trading the dividend-paying risky asset with transaction costs is equivalent to leaving all wealth in a hypothetical safe asset, which pays the higher equivalent safe rate

\[
ESR = r + \frac{\mu^2 - \lambda^2}{2\gamma\sigma^2}.
\]

We now describe the practical use of Theorem 4.2.4 for computing the optimal policy and its corresponding welfare. First, compute the upper boundary \(\pi_+(\lambda)\) for \(\lambda = 0\). If \(\pi_+(0) = 1\), then only Assumption 4.2.3(b) may hold, because \(\pi_+(\lambda)\) increases in \(\lambda\). In this case, a solution of the ODE in Assumption 4.2.3(b) is given by the following ratio of Whittaker functions (cf. Proposition 4.4.2 below):

\[
w(x, \lambda) = \left(\left(\frac{1}{4} + N\right)^2 - m^2\right) \frac{W_{-N-1,m}(\frac{2\delta}{f(\lambda)\sigma_x^2e^{-\lambda}})}{W_{-N,m}(\frac{2\delta}{f(\lambda)\sigma_x^2e^{-\lambda}})},
\]

where

\[
N = \gamma - \frac{\mu - \delta}{\sigma^2} - 1, \quad m = \sqrt{\frac{1}{4} + N(N + 1) + (1 - \gamma)} \left(\gamma + \frac{\mu^2 - \lambda^2}{\gamma\sigma^4} - \frac{2(\mu - \delta)}{\sigma^2}\right).
\]

With this solution at hand, now apply a numerical one-dimensional search for a value of \(\lambda\) that satisfies the initial condition \(w(0, \lambda) = -l(\lambda)/(1 + \varepsilon + l(\lambda))\) in Assumption 4.2.3(b). If such a value is found, then Assumption 4.2.3(b) holds and the selling boundary \(\pi_+ = 1\) is never reached — a dynamic buy-and-hold strategy is optimal.

If the numerical search does not yield a solution, try instead to verify Assumption 4.2.3(a) as follows. Again, perform a one-dimensional numerical search for a value of \(\lambda\) for which the initial condition in Assumption 4.2.3(a) holds. The general form of \(w(\lambda, x)\) again involves Whittaker functions (cf. Section 4.3) or,

\[\footnote{Here we assume that the Whittaker functions (cf. Section 4.3) are well defined. Sufficient conditions are provided in Proposition 4.4.4 (iii). The Whittaker function \(W_{k,m}(x)\) is implemented in Mathematica 9 as WhittakerW[k, m, x].} \]
alternatively, a standard numerical ODE solver can be employed, as the terminal condition is at a finite boundary rather than at infinity. If this search for a value of \( \lambda \) is successful, the sell boundary \( \pi_+ < 1 \) is reached, and the optimal strategy does buy and sell.

The asymptotic expansion for small transaction costs \( \varepsilon \sim 0 \) offers some intuition on the conditions under which each of the two regimes applies. Since the target weight without transaction costs is \( \pi_* < 1 \), for sufficiently small transaction costs the selling boundary also lies below full investment \( (\pi_+ < 1) \), as is readily verified through the implicit function theorem (as in [43]). That is, Assumption 4.2.3(a) always holds for sufficiently small transaction costs \( \varepsilon \), and the trading boundaries \( \pi_\pm \) have the asymptotic expansions

\[
\pi_\pm = \pi_* \pm \left( \frac{3}{2\gamma \pi_*^2(1-\pi_*)^2} \right)^{1/3} \varepsilon^{1/3} + \frac{\delta}{\gamma \sigma^2} \left( \frac{2\gamma \pi_*}{3(1-\pi_*)^2} \right)^{1/3} \varepsilon^{2/3} + O(\varepsilon). \tag{4.2.11}
\]

The first two terms in this formula do not contain the dividend yield parameter \( \delta \), which means that for very small costs the trading boundaries are unaffected by dividends and, in particular, are symmetric around the frictionless target \( \pi_* \) as in [43]. For somewhat larger costs, the third term shifts both boundaries upwards by the same amount: this correction is important if either transaction costs are large, dividends are substantial, or the frictionless target weight is close to one. Consistently with the numerical results in Figure 4.1 (left panel), small dividends shift both trading boundaries upwards by a similar amount. Moreover (right panel), the never-sell regime applies for parameter constellations where dividends are sufficiently large and the frictionless weight is close enough to full investment, as the asymptotic expansion (4.2.11) suggests a large shift upward.

### 4.3 Heuristics

This section describes how to derive the candidate solution to the optimization problem through heuristic control arguments.

#### 4.3.1 Hamilton-Jacobi-Bellman Equation

For a self-financing trading strategy \((\varphi^0_t, \varphi^*_t)\), write the number of risky shares \( \varphi_t = \varphi^+_t - \varphi^-_t \) as the difference of the cumulated units purchased and sold, and denote by

\[
X_t = \varphi^0_t S_t, \quad Y_t = \varphi^*_t S_t,
\]

the values of the safe and risky positions in terms of the mid price \( S_t \). Then, the self-financing condition (4.2.7), and the dynamics of \( S^b_t \) and \( S_t \) imply

\[
\begin{align*}
    dX_t &= rX_t dt + \delta Y_t dt - (1 + \varepsilon) S_t d\varphi^+_t + (1 - \varepsilon) S_t d\varphi^-_t, \\
    dY_t &= (\mu - \delta + r) Y_t dt + \sigma Y_t dW_t + S_t d\varphi^+_t - S_t d\varphi^-_t. \tag{4.3.1}
\end{align*}
\]

Consider the maximization of expected power utility \( U(x) = x^{1-\gamma}/(1-\gamma) \) from terminal wealth at time \( T \). Denote by \( V(t, x, y) \) the corresponding value function,
which is assumed to depend on time and the value of the safe and risky positions. Itô’s formula yields

\[ dV(t, X_t, Y_t) = V_t dt + V_x dX_t + V_y dY_t + \frac{1}{2} V_{yy} d\langle Y, Y \rangle_t \]

\[ = \left( V_t + r X_t V_x + \delta Y_t V_x + (\mu - \delta + r) Y_t V_y + \frac{\sigma^2}{2} V_{yy} \right) dt \]

\[ + S_t (V_y - (1 + \varepsilon) V_x) d\varphi^\dagger_t + S_t ((1 - \varepsilon) V_x - V_y) d\varphi^\downarrow_t + \sigma Y_t dW_t, \]

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

\[ \frac{1}{1 + \varepsilon} \leq \frac{V_x}{V_y} \leq \frac{1}{1 - \varepsilon}. \]

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

\[ V_t + r X_t V_x + \delta Y_t V_x + (\mu - \delta + r) Y_t V_y + \frac{\sigma^2}{2} V_{yy} = 0, \text{ if } \frac{1}{1 + \varepsilon} < \frac{V_x}{V_y} < \frac{1}{1 - \varepsilon}. \]

(4.3.3)

To simplify further, note that the value function must be homogeneous with respect to wealth, and that in the long run it should grow exponentially with the horizon. These arguments lead to guess that

\[ V(t, X_t, Y_t) = (Y_t)^{1-\gamma} v(X_t/Y_t) e^{-(1-\gamma)(r+\beta)t} \]

for some unknown function \( v \) and some unknown rate \( \beta \). Setting \( z = x/y \), we reduce the Hamilton-Jacobi-Bellman (henceforth HJB) equation (4.3.3) to

\[ 0 = \frac{\sigma^2}{2} (-\gamma (1 - \gamma) v(z) + 2 \gamma z v'(z) + z^2 v''(z)) + (\mu - \delta)((1 - \gamma) v(z) - z v'(z)), \]

\[ + \delta v'(z) - \beta (1 - \gamma) v(z), \quad \text{if } 1 - \varepsilon + z \leq \frac{(1 - \gamma) v(z)}{v'(z)} \leq 1 + \varepsilon + z. \]

Assuming that the no-trade region \( \{ z : 1 - \varepsilon + z \leq \frac{(1 - \gamma) v(z)}{v'(z)} \leq 1 + \varepsilon + z \} \) coincides with some interval \( \varepsilon < z \leq l \) to be determined, we obtain the following free boundary problem:\textsuperscript{11}

\[ 0 = \frac{\sigma^2}{2} (-\gamma (1 - \gamma) v(z) + 2 \gamma z v'(z) + z^2 v''(z)) + (\mu - \delta)((1 - \gamma) v(z) - z v'(z)) \]

\[ + \delta v'(z) - \beta (1 - \gamma) v(z), \quad (4.3.4) \]

\[ 0 = (1 - \varepsilon + u) v'(u) - (1 - \gamma) v(u), \]

\[ 0 = (1 + \varepsilon + l) v'(l) - (1 - \gamma) v(l). \]

\[ (4.3.5) \quad (4.3.6) \]

\textsuperscript{11}As the dividend vanishes (\( \delta \to 0 \)), this free boundary problem reduces to the one in [43, Equations (4.3-4.5)] after applying the transformation \( V(t, X_t, Y_t) = (Y_t)^{1-\gamma} v(X_t/Y_t) e^{-(1-\gamma)(r+\beta)t} \) to [43, Equation (4.1)].
In addition to the reduced value function \( v \), this system requires us to solve for the excess certainty equivalent rate \( \beta \) and the trading boundaries \( u \) and \( l \). The conditions (4.3.5) and (4.3.6) are therefore not enough to identify the solution, because they can be matched for any choice of the trading boundaries \( u \) and \( l \).

The optimal boundaries are the ones that additionally satisfy the smooth-pasting conditions [10, 36], formally obtained by differentiating (4.3.5) and (4.3.6) with respect to \( u \) and \( l \), respectively:

\[
(1 - \varepsilon + u)v''(u) + \gamma v'(u) = 0, \tag{4.3.7}
\]

\[
(1 + \varepsilon + l)v''(l) + \gamma v'(l) = 0. \tag{4.3.8}
\]

Substituting (4.3.8) and (4.3.6) into (4.3.4) we have

\[
0 = -\frac{\sigma^2 \gamma}{2} \left(1 - \frac{2}{1 + \varepsilon + l} + \left(\frac{l}{1 + \varepsilon + l}\right)^2\right) + (\mu - \delta) \left(1 - \frac{l}{1 + \varepsilon + l}\right)
\]

\[
+ \delta \left(\frac{1}{1 + \varepsilon + l}\right) - \beta.
\]

Setting \( \pi_- = (1 + \varepsilon)/(1 + \varepsilon + l) \), an easy computation shows that

\[
-\frac{\gamma \sigma^2}{2} \pi_-^2 + \left(\mu - \frac{\varepsilon \delta}{1 + \varepsilon}\right) \pi_- - \beta = 0,
\]

that is,

\[
\pi_- = \frac{\mu - \varepsilon \delta/1 + \varepsilon}{\gamma \sigma^2} \pm \sqrt{\left(\frac{\mu - \varepsilon \delta/1 + \varepsilon}{\gamma \sigma^2}\right)^2 - 2\beta \gamma \sigma^2}.
\]

The smaller solution is the natural candidate, since it is the correct solution in the limit \( \delta \to 0 \) (cf. [43, Section 4.1]). Analogously, setting \( \pi_+ = (1 - \varepsilon)/(1 - \varepsilon + u) \) leads to

\[
-\frac{\gamma \sigma^2}{2} \pi_+^2 + \left(\mu + \frac{\varepsilon \delta}{1 - \varepsilon}\right) \pi_+ - \beta = 0.
\]

Hence:

\[
\pi_+ = \frac{\mu + \varepsilon \delta/(1 - \varepsilon) + \sqrt{(\mu + \varepsilon \delta/(1 - \varepsilon))^2 - 2\beta \gamma \sigma^2}}{\gamma \sigma^2}.
\]

Note that, for a sufficiently large Merton proportion \( \mu/(\gamma \sigma^2) \) and dividend yield \( \delta \), the selling boundary exceeds one. However, since the paths of the risky asset \( S_t \) are continuous, the risky weight cannot exceed unity without additional transactions, so that we should have \( \pi_+ \leq 1 \). Setting

\[
\beta = \frac{\mu^2 - \lambda^2}{2 \gamma \sigma^2}, \tag{4.3.9}
\]
we see that the candidate trading boundaries become:

\[ \pi_- = \frac{\mu - \varepsilon \delta / (1 + \varepsilon) - \sqrt{\lambda^2 - 2 \mu \varepsilon \delta / (1 + \varepsilon) + (\varepsilon \delta / (1 + \varepsilon))^2}}{\gamma \sigma^2}, \quad (4.3.10) \]

\[ \pi_+ = \min \left\{ \frac{\mu + \varepsilon \delta / (1 - \varepsilon) + \sqrt{\lambda^2 + 2 \mu \varepsilon \delta / (1 - \varepsilon) + (\varepsilon \delta / (1 - \varepsilon))^2}}{\gamma \sigma^2}, 1 \right\}. \quad (4.3.11) \]

After deriving the trading boundaries \( u \) and \( l \), we observe that the free boundary problem is no longer free but fixed. Setting \( B = \frac{2 \delta}{\sigma^2} \) and applying the substitution, we get

\[ v(z) =: \left( \frac{B}{z} \right)^N f \left( \frac{B}{z} \right) \]

\[ =: \left( \frac{B}{z} \right)^N \exp \left( \frac{B}{2z} \right) h \left( \frac{B}{z} \right) \]

the differential equation (4.3.4) simplifies to Whittaker’s equation, cf. e.g., [116, Chapter 1] for more details:

\[ 0 = \left( \frac{B}{z} \right)^2 f'' \left( \frac{B}{z} \right) + \left( \frac{2(N + 1 - A)}{z} - \frac{B}{z} \right) f' \left( \frac{B}{z} \right) \]

\[ + \left( N(N + 1) - 2AN - C - N \frac{B}{z} \right) f \left( \frac{B}{z} \right) \]

\[ = h'' \left( \frac{B}{z} \right) + \left( -\frac{1}{4} + \frac{N}{B/z} + \frac{1/4 - m^2}{(B/z)^2} \right) h \left( \frac{B}{z} \right), \quad (4.3.14) \]

where

\[ A = \gamma - \frac{\mu - \delta}{\sigma^2}, \quad C = (1 - \gamma) \left( \frac{\mu^2 - \lambda^2}{\gamma \sigma^4} - \frac{2(\mu - \delta)}{\sigma^2} \right), \quad (4.3.15) \]

\[ N = A - 1, \quad m = \sqrt{1/4 + N(N + 1) + C}. \quad (4.3.16) \]

The solution to (4.3.14) is given by a linear combination of the Whittaker functions \( M_{-N, \pm m} \) [123, Chapter 16]:

\[ h \left( \frac{B}{z} \right) = C_1 \left( M_{-N, m} \left( \frac{B}{z} \right) + M_{-N, -m} \left( \frac{B}{z} \right) \right) + C_2 W_{-N, m} \left( \frac{B}{z} \right), \quad (4.3.17) \]
where $C_1, C_2$ are two real constants and the Whittaker functions $M_{-N, m}(x)$ and $W_{-N, m}(x)$ are defined via the Kummer function$^{13} \ _1 F_1(a, b, x)$ and the Tricomi function $U(\xi, \eta, x)$: (cf. [116, Chapter 1])

$$M_{-N, m}(x) := x^{1/2 + m} e^{-1/2x} \ _1 F_1(1/2 + m + N, 1 + 2m, x), \tag{4.3.18}$$

$$U(\xi, \eta, x) := \frac{\Gamma(1-\eta)}{\Gamma(1 + \xi - \eta)} F_1(1 - \eta, 1 + \xi - \eta, x), \tag{4.3.19}$$

$$W_{-N, m}(x) := x^{1/2 + m} e^{-1/2x} \ _1 F_1(1/2 + m + N, 1 + 2m, x). \tag{4.3.20}$$

In general, the Whittaker functions $M_{-N, \pm m}$ are complex valued, while the value function of the problem needs to be real valued. To achieve this, we consider the special alignment of the functions $M_{-N, \pm m}$ and $W_{-N, m}$ in (4.3.17), which ensures that the function $h$ and hence the value function $v$ remains real for all model parameters (cf. Lemma 4.4.1 (ii) below).

To determine the real constants $C_1$ and $C_2$ in (4.3.17), observe that the differential equation (4.3.4) and its boundary conditions (4.3.5), (4.3.6) are invariant under scalar multiplications, allowing us to set $C_2 = 1$. To identify $C_1$, first assume $\pi_+ < 1$ and hence $u > 0$, so that $B/z$ admits a finite upper bound. In this case, $C_1$ is determined by the boundary condition (4.3.5). In the limiting case $z \to u = 0$, i.e., $\pi_+ = 1$, the argument of the function $h$ diverges, so that there is no boundary condition from which to determine $C_1$. Instead, recall that the utility should remain finite as the selling boundary approaches unity. As a result, $C_1$ has to be chosen so that the value function $v$ remains finite as its argument tends to zero. In view of (4.3.13) this condition is equivalent to

$$h \left( \frac{B}{z} \right) \sim \left( \frac{B}{z} \right)^{-N} \exp \left( -\frac{B}{2z} \right), \quad \text{for } z \downarrow 0,$n

which is satisfied for $h(x) = W_{-N, m}(x)$ (i.e., $C_1 = 0$) in view of the asymptotics of the Whittaker function $W_{-N, m}(x)$ [123, Section 16.31].

Finally, the parameter $\lambda$ is determined by (4.3.6).

### 4.4 Proof of Theorem 4.2.4

#### 4.4.1 Preliminaries

The starting point for the proof is the free boundary problem (4.3.4-4.3.6). In Section 4.3 we transformed the differential equation (4.3.4) into a Whittaker equa-

\[ _1 F_1(a, b, x) := \sum_{n=0}^{\infty} \frac{a^{(n)} x^n}{b^{(n)} n!}, \]

Here the Pochhammer symbol $a^{(n)}$ is given by $a^{(n)} := (a + 1)(a + 2) \cdots (a + n - 1)$. For $a \neq 0, -1, -2, \ldots$, we can write $a^{(n)} = \Gamma(a + n)/\Gamma(a)$, where $\Gamma(x)$ denotes the Gamma function.
4.4 Proof of Theorem 4.2.4

In this subsection, we first establish some important properties of Whittaker functions and their relation to the boundary value problem (4.3.4-4.3.6).

**Lemma 4.4.1.** Let $C_1$ and $C_2$ be real constants and define $N, m$ as in (4.3.16). If $2m \not\in \mathbb{Z}$, then:

(i) The function

$$v(z) := \left(\frac{2\delta}{\sigma^2 z}\right)^{N} \exp\left(\frac{\delta}{\sigma^2 z}\right) \left( C_1 \left( M_{N,m} \left( \frac{2\delta}{\sigma^2 z} \right) + M_{N,-m} \left( \frac{2\delta}{\sigma^2 z} \right) \right) + C_2 W_{N,m} \left( \frac{2\delta}{\sigma^2 z} \right) \right)$$

with $M_{N,\pm m}$, $W_{N,m}$ as in (4.3.20), solves the differential equation (4.4.4).

(ii) For $x \in \mathbb{R}_{>0}$ the Whittaker functions

$$M_{N,m}(x) + M_{N,-m}(x) \quad \text{and} \quad W_{N,m}(x)$$

are real valued.

(iii) For each $m \in i\mathbb{R}_{>0}$ there is a constant $R > 0$ such that $v$ has no zeros on $(R, \infty)$.

(iv) The function

$$v_1(z) := \left(\frac{2\delta}{\sigma^2 z}\right)^{N} \exp\left(\frac{\delta}{\sigma^2 z}\right) W_{N,m} \left( \frac{2\delta}{\sigma^2 z} \right) \quad (4.4.1)$$

satisfies $\lim_{z \to 0} v_1(z) < \infty$. If $m \in \mathbb{R}_{>0}$ then $v_1$ has no positive real zeros provided that $\xi := 1/2 + N + m > 0$.

**Proof.** First, note that for $2m \not\in \mathbb{Z}$ the Whittaker functions $M_{N,\pm m}$ and $W_{N,m}$ are well defined.

(i) A straightforward computation shows that $v$ solves the differential equation (4.3.4) if and only if the Whittaker functions $M_{N,\pm m}$ and $W_{N,m}$ are solutions of the differential equation (4.3.14). This is readily verified by insertion.

(ii) Without loss of generality, assume that $m \in i\mathbb{R}_{>0}$. Setting $\eta = 1 + 2m$ yields

$$x^N (M_{N,m}(x) + M_{N,-m}(x)) = e^{-\frac{1}{2}x} \left( x^\xi \mathcal{F}_1(\xi, \eta, x) + x^\xi \mathcal{F}_1(\bar{\xi}, \bar{\eta}, x) \right), \quad (4.4.2)$$

where $\mathcal{F}_1(a, b, x)$ denotes the Kummer function and $\bar{a}$ the complex conjugate of $a \in \mathbb{C}$. The power series expansion of the Kummer function
shows that \( \xi_1 F_1(\xi, \eta, x) \) and \( \bar{\xi}_1 F_1(\bar{\xi}, \bar{\eta}, x) \) are complex conjugate, so that the sum in (4.4.2) is real valued. By definition of the Whittaker function \( W_{-N,m} \) (cf. (4.3.20)),

\[
x^N W_{-N,m}(x) = x^N e^{-\frac{1}{2}x} x^{\frac{1}{2}+m} \left( \frac{\Gamma(1-\eta)}{\Gamma(1+\xi-\eta)} F_1(\xi, \eta, x) \right.
+ \frac{\Gamma(\eta-1)}{\Gamma(\xi)} x^{1-\eta} F_1(1+\xi-\eta, 2-\eta, x) \left. \right) \tag{4.4.3}
\]

where we have used in the second step that

\[
1 - \eta + \xi = \bar{\xi}, \quad 2 - \eta = \bar{\eta}. \tag{4.4.4}
\]

As \( \Gamma(\bar{\xi}) = \Gamma(\xi) \), this completes the proof.

(iii) Taking into account the following alternative representation of the Whittaker function \( W_{-N,m} \) [116, Formula 1.7.1],

\[
W_{-N,m}(x) = \frac{\pi}{\sin(2m\pi)} \left( \frac{-M_{-N,m}(x)}{\Gamma(1/2 - m + N)\Gamma(1 + 2m)} \right.
+ \frac{M_{-N,-m}(x)}{\Gamma(1/2 + m + N)\Gamma(1 - 2m)} \left. \right),
\]

we obtain

\[
v(z) = x^N e^{\frac{1}{2}z} \left[ \left( C_1 - \frac{C_2}{\sin(2m\pi)\Gamma(\xi)\Gamma(\eta)} \right) M_{-N,m}(x) \right.
+ \left. \left( C_1 + \frac{C_2}{\sin(2m\pi)\Gamma(\xi)\Gamma(\bar{\eta})} \right) M_{-N,-m}(x) \right], \tag{4.4.5}
\]

where \( x = 2\delta/\sigma^2 z \). As a result, any positive real root of \( v \) is automatically a positive real root of the Whittaker function \( M_{-N,m} \), because the terms in front of the two Whittaker functions in (4.4.5) are complex conjugate (recall that \( \Gamma(\bar{\xi}) = \Gamma(\xi) \)). However, for \( m \in i\mathbb{R}_{>0} \), the set of positive real zeros of the Whittaker function \( W_{-N,m} \) and hence \( M_{-N,m} \) is bounded [39, Theorem 3]. That is, \( v \) does not vanish on \( (R, \infty) \) for some constant \( R > 0 \).

(iv) The first part of this assertion follows directly from the asymptotics of the Whittaker function \( W_{-N,m} \) [123, Chapter 16]:

\[
W_{-N,m}(x) \sim x^{-N} e^{-\frac{1}{2}x}, \quad \text{as} \quad x \to \infty. \tag{4.4.6}
\]

Moreover, the definition (4.3.20) shows that the function \( v_1 \) has the same zeros as the Tricomi function \( U(\xi, \eta, x) \), which does not admit any positive real zeros under the assumption \( \xi > 0 \) [116, Section 6.1.2].
Aiming for simpler formulas, we exploit another transformation of the boundary value problem problem (4.3.4-4.3.6):

Case \( \pi_+ < 1 \): The substitution

\[
v(z) = e^{(1-\gamma) \int_0^z (\log(z/u(\lambda))) w(y) \, dy}, \quad \text{i.e.,} \quad w(x) = \frac{u(\lambda)e^x v'(u(\lambda)e^x)}{(1-\gamma)v(u(\lambda)e^x)}, \tag{4.4.7}
\]

reduces the boundary value problem (4.3.4-4.3.6) to an (inhomogeneous) Riccati equation:

\[
0 = w'(x) + (1-\gamma)w(x)^2 + \left(2\gamma - 1 - \frac{2(\mu - \delta)}{\sigma^2} + \frac{2\delta}{\sigma^2 e^x u}\right)w(x) - \left(\gamma + \frac{\mu^2 - \lambda^2}{\gamma \sigma^2} - \frac{2(\mu - \delta)}{\sigma^2}\right), \tag{4.4.8}
\]

\[
w(0, \lambda) = \frac{u}{1-\varepsilon + u}, \tag{4.4.9}
\]

\[
w \left( \log \left( \frac{l(\lambda)}{u(\lambda)} \right), \lambda \right) = \frac{l}{1 + \varepsilon + \lambda}, \tag{4.4.10}
\]

where

\[
u = (1-\varepsilon) \frac{1 - \pi_+}{\pi_+},
\]

\[
\pi_+ = \frac{\mu + \varepsilon\delta/(1-\varepsilon) + \sqrt{\lambda^2 + 2\mu\varepsilon\delta/(1-\varepsilon) + (\varepsilon\delta/(1-\varepsilon))^2}}{\gamma \sigma^2},
\]

\[
l = (1+\varepsilon) \frac{1 - \pi_-}{\pi_-},
\]

\[
\pi_- = \frac{\mu - \varepsilon\delta/(1+\varepsilon) - \sqrt{\lambda^2 - 2\mu\varepsilon\delta/(1+\varepsilon) + (\varepsilon\delta/(1+\varepsilon))^2}}{\gamma \sigma^2}.
\]

In Section 4.3 we have identified the solutions of the original problem (4.3.4-4.3.5) by means of the Whittaker functions \( M_{-N, m} \) and \( W_{-N, m} \). Thus, the transformation (4.4.7) relates the solutions of the boundary value problem (4.4.8-4.4.9) to the Whittaker functions. Notice that in the limiting case \( \delta \to 0 \) we recover the transformed Hamilton-Jacobi-Bellman equation in [43, Equations (4.8-4.10)].\(^{14}\)

\(^{14}\)To see this, apply the transformation

\[V(t, X_t, Y_t) = (Y_t)^{1-\gamma} e^{(X_t/Y_t) e^{-(1-\gamma)(r+\beta)t}}\]

to [43, Equation 4.1] and follow the calculations in [43, Section 4].
Case $\pi_+ = 1$: In this case, the lower boundary $u$ equals 0. Hence we use a different substitution, i.e.,

$$v(z) = e^{(1-\gamma) \int_0^z l(y) \, dy},$$

i.e.,

$$w(x) = \frac{-l(l\lambda)e^{-x}v'(l\lambda)e^{-x}}{(1-\gamma)u(l\lambda)e^{-x)},}
$$

(4.4.11)

to also reduce the boundary value problem (4.3.4-4.3.6) to an (inhomogeneous) Riccati equation:

$$0 = w'(x) + (1-\gamma)w(x)^2 + \left(1 - 2\gamma + \frac{2(\mu - \delta)}{\sigma^2} - \frac{2\delta}{\sigma^2 e^{-x}}\right)w(x)$$

$$- \left(\gamma + \frac{\mu^2 - \lambda^2}{\gamma \sigma^4} - \frac{2(\mu - \delta)}{\sigma^2}\right),$$

(4.4.12)

$$\lim_{x \to \infty} w(x, \lambda) = \lim_{x \to \infty} \frac{-l(l\lambda)e^{-x}}{1 - \varepsilon + l(l\lambda)e^{-x}} = 0,$$

(4.4.13)

$$w(0, \lambda) = \frac{-l(l\lambda)}{1 + \varepsilon + l(l\lambda)}.$$

(4.4.14)

Riccati equations with non-constant coefficients typically do not admit closed-form solutions. However, with the explicit solution of the HJB equation at hand, it is possible to derive a solution of the corresponding Riccati equation (4.4.12) in terms of Whittaker functions:

**Proposition 4.4.2.** Suppose the positive real zeros of the Whittaker function $W_{-N,m}(\frac{25}{12\sigma^2 e^{-x}})$ are bounded, so that the function

$$w(x, \lambda) = \left(\frac{N}{2} + \frac{1}{2} \right)^2 - \frac{m^2}{1 - \gamma} \frac{W_{-N-1,m}(\frac{25}{12\sigma^2 e^{-x}})}{W_{-N,m}(\frac{25}{12\sigma^2 e^{-x}})},$$

(4.4.15)

with $N$ and $m$ as in (4.3.16) is well defined for sufficiently large $x > 0$. Then, this function is the unique solution of the boundary value problem (4.4.12-4.4.13).

**Proof.** In view of Lemma 4.4.1(i), the function $v_1$ in (4.4.1) solves the HJB equation (4.3.4). Using the differentiation formula [116, Formula 2.4.15]

$$\frac{d}{dx} \left(x^N e^{\frac{1}{2}x} W_{N,m}(x)\right) = \left(\frac{1}{2} + m + N\right) \left(\frac{1}{2} - m + N\right) x^{N-1} e^{\frac{1}{2}x} W_{N-1,m}(x),$$

(4.4.16)

a tedious but straightforward calculation shows that

$$w(x) = -le^{-x} \frac{v_1'(l e^{-x})}{(1-\gamma)v_1(l e^{-x})} = \left(\frac{1}{2} + N\right)^2 - \frac{m^2}{1 - \gamma} \frac{W_{-N-1,m}(\frac{25}{12\sigma^2 e^{-x}})}{W_{-N,m}(\frac{25}{12\sigma^2 e^{-x}})}.$$

Here, the upper boundary condition (4.4.13) is obtained from (4.3.5) by applying the transformation (4.4.11) and passing to the limit $u \to 0$. 

---

15 Here, the upper boundary condition (4.4.13) is obtained from (4.3.5) by applying the transformation (4.4.11) and passing to the limit $u \to 0$. 

solves the corresponding Riccati equation (4.4.12). Moreover, the boundary condition (4.4.13) is satisfied due to the asymptotics of the Whittaker function [123, Chapter 16]:

$$w(x) \sim \left( \frac{(\frac{1}{2} + N)^2 - m^2}{1 - \gamma} \right) \frac{\sigma^2}{2\delta} le^{-x}, \quad \text{as} \quad x \to \infty. \quad (4.4.17)$$

Note that the Whittaker functions $W_{-N,m}(x)$ and $W_{N,m}(-x)$ form a fundamental system of solutions of the Whittaker equation (4.3.14) (cf. [123, Section 16.31]). Since the boundary condition (4.4.13) rules out the solution $W_{N,m}(-x)$, the transformation (4.4.11) shows that the solution (4.4.15) is unique.

A similar result can be derived along the same lines for the Riccati equation (4.4.8-4.4.9) arising in the case $\pi_+ < 1$.

Henceforth, we assume that there exists a suitable $\lambda > 0$ for which the additional boundary condition (4.4.14) is also satisfied unless $\pi_+ < 1$, in which case we assume instead that the analogous condition (4.4.10) holds. That is, we assume that the following equivalent form of Assumption 4.2.3 is fulfilled.

**Assumption 4.4.3.** If $\pi_+ = 1$, there exists a solution of the Riccati equation (4.4.12) satisfying the boundary conditions (4.4.13-4.4.14). If $\pi_+ < 1$, there exists a solution of the Riccati equation (4.4.8) with boundary conditions (4.4.9-4.4.10). That is, there exists $\lambda \in [\lambda_{\min}, \min(\lambda_{\max}, \mu)]$ such that $\pi_+ = 1$ and the conditions (4.4.12-4.4.14) are satisfied or $\pi_+ < 1$ and the conditions (4.4.8-4.4.10) are satisfied, respectively. Here, the lower and upper bounds\(^\text{16}\) are defined as

$$\lambda_{\min} := \sqrt{\frac{\delta \varepsilon}{1 + \varepsilon} \left( 2\mu - \frac{\delta \varepsilon}{1 + \varepsilon} \right)}, \quad (4.4.18)$$

$$\lambda_{\max} := \sqrt{\frac{1}{1 - \gamma} \left( (\mu - \delta \gamma)^2 + \delta^2 \gamma(1 - \gamma) + \gamma^2 \sigma^4 \left( \frac{1}{4\gamma} - \frac{\mu - \delta}{\gamma \sigma^2} \right) \right)}. \quad (4.4.19)$$

As in [43], the implicit function theorem\(^\text{17}\) guarantees the existence and uniqueness of such a parameter $\lambda$ and a corresponding $\pi_+ < 1$ if the transaction costs $\varepsilon$ are sufficiently small. The following lemma provides a sufficient condition on the model parameters for the existence of the gap parameter $\lambda$ for $\pi_+ = 1$:

\(^{16}\)The lower bound $\lambda_{\min}$ is chosen so that the corresponding buying boundary $\pi_-$ in (4.3.10) is real valued for any $\lambda \geq 0$. The upper bound $\lambda_{\max}$ guarantees that the parameter $m$ from (4.3.16) is purely imaginary or real valued if $\gamma > 1$ or $0 < \gamma < 1$, respectively.

\(^{17}\)For the transformation (4.4.7) to be well defined, the function $\nu$ from Lemma 4.4.1(i) cannot admit any zeros on $[u, l]$. In the economically more relevant case $\gamma > 1$, the condition in Proposition 4.4.4(iii) is sufficient to guarantee this condition. For $0 < \gamma < 1$, finding an analogous condition is tantamount to identifying the real zeros of the sum of the Whittaker functions $M_{-N, \pm m}$. This task, which appears to have limited interest for applications, is beyond the scope of this thesis, where we simply assume that the parameters are chosen in such a way that the transformation (4.4.7) is always applicable.
Proposition 4.4.4. Assumption 4.4.3 holds if the following parameter restrictions are satisfied:

(i) The selling boundary is sufficiently large, in that:

\[ \pi_+ := \frac{\mu + \varepsilon \delta / (1 - \varepsilon) + \sqrt{2 \mu \varepsilon \delta / (1 - \varepsilon) + (\varepsilon \delta / (1 - \varepsilon))^2}}{\gamma \sigma^2} \geq 1. \]

(ii) \( \lambda_{\min} \) and \( \lambda_{\max} \) defined in (4.4.18-4.4.19) are real numbers satisfying \( \lambda_{\min} \leq \min(\lambda_{\max}, \mu) \), and the function \( g : [\lambda_{\min}, \min(\lambda_{\max}, \mu)] \mapsto \mathbb{R} \) given by

\[ g(\lambda) = -\frac{(\frac{1}{2} + N)^2 - m(\lambda)^2}{l(\lambda)} W_{-N-1,m}(\lambda) \left( \frac{2\delta}{l(\lambda)\sigma^2} \right) - \frac{1 - \gamma}{1 + \varepsilon + l(\lambda)}, \]

changes sign on \([\lambda_{\min}, \min(\lambda_{\max}, \mu)]\), where the constants \( N, m \) and the Whittaker function \( W_{-N,m} \) are as defined in (4.3.15-4.3.16) and (4.3.20), respectively. In particular, there exists \( \lambda_\ast \in [\lambda_{\min}, \min(\lambda_{\max}, \mu)] \) such that

\[ g(\lambda_\ast) = 0. \]

(iii) The function \( w \) specified in (4.4.15) is well defined. For example, this holds if

\begin{align*}
\{ & \xi = 1/2 + N + m(\lambda_\ast) > 0 \quad \text{and} \quad 2m(\lambda_\ast) \notin \mathbb{Z}, \quad \text{for} \quad 0 < \gamma < 1, \\
& x_{\min} := \frac{2\delta}{(2\pi l(\lambda_\ast))} > R := \max(r_{N+1}, r_N), \quad \text{for} \quad \gamma > 1,
\end{align*}

where \( r_n := \sup\{x > 0 | W_{-n,m(\lambda_\ast)}(x) = 0\} < \infty \) (cf. [39, Theorem 3]).

Proof. Under condition (i), the selling boundary (4.3.11) is given by \( \pi_+ = 1 \). Since \( g \) is a continuous function of \( \lambda \), the intermediate value theorem yields the existence of \( \lambda_\ast \in [\lambda_{\min}, \min(\lambda_{\max}, \mu)] \) such that

\[ 0 = g(\lambda_\ast) = \frac{1 - \gamma}{-l(\lambda_\ast)} w(0, \lambda_\ast) - \frac{1 - \gamma}{1 + \varepsilon + l(\lambda_\ast)}, \]

which is equivalent to (4.4.14). For \( \lambda_\ast \in [\lambda_{\min}, \min(\lambda_{\max}, \mu)] \) a simple calculation shows that \( m \) is purely imaginary or real valued if \( \gamma > 1 \) or \( 0 < \gamma < 1 \), respectively. By Lemma 4.4.1(iii) and (iv) the function \( v \) does not admit any zeros on \((0, l(\lambda_\ast)]\) and hence the function \( w \) is well defined.

For given model parameters, the parameter conditions in Proposition 4.4.4 can be readily verified by insertion, compare Table 4.4. Henceforth, we always assume that Assumption 4.4.3 holds. Moreover, note that if a suitable \( \lambda \) as in Assumption 4.4.3 exists then it is also unique (cf. Lemma 4.4.12). Thus, \( \lambda \) henceforth denotes this candidate value, and we omit the \( \lambda \)-dependence of \( l = l(\lambda) \), \( \pi_- = \pi_-(\lambda) \) and \( w(x) = w(\lambda, x) \) in the sequel. The proof techniques used below

\[ \text{This condition ensures that the maximum in (4.3.11) equals 1 for any } \lambda \geq 0, \text{ so that any solution necessarily corresponds to } \pi_+ = 1. \]
4.4 Proof of Theorem 4.2.4

Table 4.4: Verification of the conditions in Proposition 4.4.4. For $\gamma = 3.45$ we fix $\mu = 8\%$, $\sigma = 16\%$, $\delta = 4\%$, and $\varepsilon = 1\%$. For $\gamma = 0.6$ we use $\mu = 5\%$, $\sigma = 30\%$, $\delta = 3\%$, and $\varepsilon = 1\%$.

can be applied both for $\pi_+ = 1$ and $\pi_+ < 1$ by using the transformations (4.4.11) and (4.4.7), respectively. For the sake of brevity we only report the case $\pi_+ = 1$.

For later use, we make the following simple observation, which translates the smooth pasting condition for the candidate value function $v$ to its “reduced” counterpart $w$:

**Corollary 4.4.5.** The function $w$ from (4.4.15) satisfies

$$w'(0) = -w(0)(1 + w(0)) = \frac{l(1 + \varepsilon)}{(1 + \varepsilon + l)^2} > 0. \tag{4.4.20}$$

Moreover, it is strictly increasing and takes values in $[-l/(1 + \varepsilon + l), 0)$.

**Proof.** The first assertion is verified directly by inserting the boundary condition (4.4.14) into the differential equation (4.4.12). Under Assumption 4.4.3, both Whittaker functions $W_{-N-1,m}$ and $W_{-N,m}$ do not admit any zeros on $(0, l]$. As $w(0) = -l/(1 + \varepsilon + l) < 0$, we infer that $w < 0$ on $[0, \infty)$. It remains to show that $w$ is strictly increasing, for which it suffices to establish $w' > 0$. Since $w'(0) > 0$, let us by contradiction assume that there exists a $x_0 > 0$ such that $w'(x_0) = 0$. Differentiating (4.4.12) yields

$$w''(x) = -2(1 - \gamma)w'(x)w(x) - \left(1 - 2\gamma + \frac{2(\mu - \delta)}{\sigma^2} - \frac{2\delta e^x}{\sigma^2 l}\right)w'(x)$$
$$+ \frac{2\delta}{\sigma^2 l} e^x w(x). \tag{4.4.21}$$

In particular,

$$w''(x_0) = \frac{2\delta}{\sigma^2 l} e^{x_0} w(x_0) < 0.$$ 

Therefore, $w' \leq 0$ on $[x_0, \infty)$ and hence $w$ is decreasing on $[x_0, \infty)$, contradicting the boundary condition (4.4.13).
4.4.2 Shadow Prices and Long-Run Verification

A key idea for the proof of our verification theorem is to replace the original bid and ask prices by a single, fictitious “shadow price” \( \tilde{S} \). Any frictionless price process that evolves in the bid-ask spread and pays the same dividends leads to at least as favorable terms of trade as the original market with transaction costs. The shadow price \( \tilde{S} \) is distinguished by an optimal policy that is also feasible with transaction costs, and thereby optimal in the original market as well (compare [43, Proposition C.5]):

**Definition 4.4.6.** A shadow price is a frictionless price process \( \tilde{S} \) lying within the bid-ask spread \([(1 - \varepsilon)S, (1 + \varepsilon)S]\) that pays the same dividends as the risky asset with frictions, and which admits a long-run optimal strategy \((\varphi^0, \varphi)\) of finite variation that only entails buying or selling the risky asset when \( \tilde{S} \) equals the buying or selling price, respectively.

The argument in [43, Section B.1] suggests that the discounted shadow price can be constructed as the marginal rate of substitution of risky for safe assets for the optimal investor, i.e., as the ratio of the partial derivatives of the value function with respect to the numbers of shares in the risky and safe asset:

\[
\frac{\tilde{S}_t}{S^0_t} = \frac{\partial_{\varphi_t} V(t, X_t, Y_t)}{\partial_{\varphi^0_t} V(t, X_t, Y_t)}.
\]

With the candidate value function derived in Section 4.3, this leads to the candidate shadow price

\[
\tilde{S}_t = -S_t \frac{e^{-\Upsilon_t}(1 + w(\Upsilon_t))}{w(\Upsilon_t)},
\]

(4.4.22)

where \( e^{\Upsilon_t} = lY_t/X_t \) is the ratio between the risky and safe positions at the mid price \( S_t \), centered at the buying boundary \( l = (1 + \varepsilon)(1 - \pi)/\pi \). In view of the above heuristics, the cash-stock ratio \( X/Y \) should remain within the no-trade region \([0, l]\); consequently, \( \Upsilon \) should take values in \([0, \infty)\).

These heuristic arguments and the dynamics (4.3.1) and (4.3.2) motivates us to define the process \( \Upsilon \) as a diffusion process with instantaneous reflection at 0:

\[
d\Upsilon_t = (\mu - \delta - \sigma^2/2 - \delta e^{\Upsilon_t}/l)dt + \sigma dW_t + dL_t,
\]

(4.4.23)

where \( L \) is the minimal adapted, continuous, non-decreasing process that only increases on \( \{\Upsilon_t = 0\} \) and keeps \( \Upsilon_t \) positive. Starting from this process, whose existence dates back to a classical result of [115], the process \( \tilde{S} \) can then be defined in accordance with (4.4.22):

\(^{19}\)An approach of this kind was first used in [69], and has been utilized in the present setting without dividends by [43].

\(^{20}\)Also compare [88, 9, 60].
Lemma 4.4.7. Define

\[
\Upsilon_0 = \begin{cases} 
\log [\pi_*/(1 - \pi_*)], & \text{if } \xi^0 S^0_0 > l_\xi S_0, \\
\log [\pi_*/(1 - \pi_*)], & \text{if } \xi^0 S^0_0 \leq 0, \\
\log [l_\xi S_0/(\xi^0 S^0_0)], & \text{otherwise}.
\end{cases}
\] (4.4.24)

Let \( \Upsilon \) be defined as in (4.4.23), starting at \( \Upsilon_0 \). Then,

\[
\tilde{S} = -S le^{-\Upsilon(1 + w(\Upsilon))}/w(\Upsilon),
\]

with \( w \) from (4.4.12-4.4.14), has the dynamics

\[
d\tilde{S}/\tilde{S} = (r + \tilde{\mu}(\Upsilon_t) - \tilde{\delta}(\Upsilon_t))dt + \tilde{\sigma}(\Upsilon_t)dW_t,
\]

where \( \tilde{\mu}(\cdot), \tilde{\sigma}(\cdot) \) and \( \tilde{\delta}(\cdot) \) are given by

\[
\tilde{\mu}(y) = \frac{\sigma^2w'(y)}{w(y)(1 + w(y))}\left(\frac{w'(y)}{w(y)} + (1 - \gamma)w(y) + (1 - \gamma)\right),
\] (4.4.25)

\[
\tilde{\sigma}(y) = -\frac{\sigma w'(y)}{w(y)(1 + w(y))},
\] (4.4.26)

\[
\tilde{\delta}(y) = \frac{-\delta e^y w(y)}{l(1 + w(y))}.
\] (4.4.27)

Moreover, the process \( \tilde{S} \) takes values within the bid-ask spread \([ (1 - \varepsilon)S, (1 + \varepsilon)S ] \)
and \( \delta S = \tilde{\delta} \tilde{S} \).

Note that the first two cases in (4.4.24) arise if the initial cash-stock ratio \( \xi^0 S^0_0/\xi S_0 \) lies outside of the interval \((0, l] \). Then, a jump from the initial position \( (\varphi^0_0, \varphi^-_0) = (\xi^0, \xi) \) to the Merton proportion \( \pi_* \) moves the cash-stock ratio \( \xi^0 S^0_0/\xi S_0 \) into the interval \((0, l] \). This transfer necessitates the purchase or sale of the risky asset and hence the initial price \( \tilde{S}_0 \) is defined to match the buying or selling price, respectively, of the risky asset.

Proof. The dynamics of \( \tilde{S} \) result from Itô’s formula, the dynamics of \( \Upsilon \), the boundary condition (4.4.14), and the identity (4.4.21).\footnote{The calculations and simplifications provided in this proof have been carried out using Mathematica 9; the code is available from the author upon request.} Furthermore, a direct computation shows that the dividend rate (4.4.27) satisfies \( \delta S = \tilde{\delta} \tilde{S} \). For the range of the shadow price first note that the boundary condition (4.4.14) implies

\[
-\frac{le^0(1 + w(0))}{w(0)} = \frac{-l}{1 + \varepsilon} = 1 + \varepsilon.
\]
Moreover, the asymptotic formula (4.4.17) shows

\[
G := \lim_{y \to \infty} -\frac{ye^{-y}(1 + w(y))}{w(y)} = \lim_{y \to \infty} -\frac{e^{-y}}{w(y)} = -\frac{2\delta}{\sigma^2} \left( \frac{\left( \frac{1}{2} + N \right)^2 - m^2}{1 - \gamma} \right) = -\frac{2\delta \gamma \sigma^2}{2\gamma \sigma^2 + (\mu - \gamma \sigma^2)^2 - \lambda^2}.
\]

In particular, by definition of the selling boundary (4.3.11), we have \( G \geq (1 - \varepsilon) \) and equality holds if and only if

\[
\pi_+ = \frac{\mu + \varepsilon \delta / (1 - \varepsilon) + \sqrt{\lambda^2 + 2 \mu \varepsilon \delta / (1 - \varepsilon) + (\varepsilon \delta / (1 - \varepsilon))^2}}{\gamma \sigma^2} = 1.
\]

To complete the proof it remains to check that the function \( y \mapsto -\frac{ye^{-y}(1 + w(y))}{w(y)} \) is decreasing, or equivalently that \( g(y) := w'(y) + w(y) + w'(y) \leq 0 \) for all \( y \geq 0 \). To this end, we first note that the identity (4.4.20) yields \( g(0) = 0 \). Moreover, the boundary condition (4.4.13) and the asymptotic formula (4.4.17) for \( w \) show

\[
\lim_{y \to \infty} g(y) = \lim_{y \to \infty} \gamma w(y)^2 + \left( 2\gamma - \frac{2(\mu - \delta)}{\sigma^2} + \frac{2\delta}{\sigma^2} e^{-y} \right) w(y)
\]

\[
+ \left( \gamma + \frac{\mu^2 - \lambda^2}{\gamma \sigma^4} - \frac{2(\mu - \delta)}{\sigma^2} \right)
\]

\[
= \left( \frac{\left( \frac{1}{2} + N \right)^2 - m^2}{1 - \gamma} \right) + \left( \gamma + \frac{\mu^2 - \lambda^2}{\gamma \sigma^4} - \frac{2(\mu - \delta)}{\sigma^2} \right) = 0.
\]

The differential equation (4.4.12), the boundary condition (4.4.14), and the identity (4.4.20) therefore give

\[
g'(0) = 2\gamma w'(0) \left( (1 + w(0)) - \frac{\mu - \delta}{\gamma \sigma^2} \right) + \frac{2\delta}{\sigma^2} e^y (w'(0) + w(0))
\]

\[
= -\frac{2(1 - \pi_+) \pi_+ ((1 + \varepsilon)(\mu - \gamma \sigma^2 \pi_+) - \delta \varepsilon)}{(1 + \varepsilon) \sigma^2} < 0.
\]

By contradiction let us now assume that there exists a \( y_1 > 0 \) such that \( g(y_1) = \nu > 0 \). Then, by Rolle’s Theorem there exists a \( y_2 > 0 \) with \( g(y_2) > 0 \) and \( g'(y_2) = 0 \), since \( g(0) = 0 \), \( g'(0) < 0 \) and \( \lim_{y \to \infty} g(y) = 0 \). Moreover, the convergence of \( g \) shows that \( y_2 \) can be chosen to be a (local) maximum. Taking the second derivative of \( g \) in (4.4.28) and substituting, we obtain

\[
0 = g'(y_2) = w''(y_2) + w'(y_2) + 2w'(y_2)w(y_2)
\]

\[
= 2\gamma w'(y) \left( (1 + w(y)) - \frac{\mu - \delta}{\gamma \sigma^2} \right) + \frac{2\delta}{\sigma^2} e^{-y} (w'(y) + w(y))
\]
which yields
\[ g''(y_2) = 2\gamma(-w'(y_2) - 2w(y_2)w(y_2)) \left( 1 + w(y_2) - \frac{\mu - \delta}{\gamma \sigma^2} \right) + 2\gamma(w')^2(y_2) + \frac{2\delta}{\sigma^2}e^{-y_2}(w'(y_2) + w(y_2) - 2w(y_2)w'(y_2)) \]
\[ = \frac{2\delta}{\sigma^2}e^{-y_2}g(y_2) + 2\gamma(w')^2(y_2) > 0, \]
which contradicts the (local) maximality of \( y_2 \).

If the candidate process \( \tilde{S} \) from Lemma 4.4.7 is indeed a shadow price, then the optimal numbers of safe and risky assets must be the same as in the original market with transaction costs, since the dividend payments in the shadow market match their counterparts in the original market with transaction costs by (4.4.27). By definition of \( \tilde{S} \), the candidate optimal strategy derived heuristically in Section 4.3 therefore leads to the following candidate for the long-run optimal risky weight in the shadow market:

\[ \tilde{\pi}(Y_t) = \frac{\varphi_t \tilde{S}_t}{\varphi_0 \tilde{S}_0 + \varphi_t \tilde{S}_t} = \frac{\varphi_t S_t s_0 - \frac{e^{-Y_t(1+w(Y_t))}}{w(Y_t)}}{\varphi_0 S_0 s_0 + \varphi_t S_t - \frac{e^{-Y_t(1+w(Y_t))}}{w(Y_t)}} = \frac{-1 + w(Y_t)}{1 - \frac{1 + w(Y_t)}{w(Y_t)}} = 1 + w(Y_t). \]

(4.4.29)

To show that this risky weight is indeed long-run optimal for \( \tilde{S} \), we first establish the following finite-horizon bounds in analogy to the frictionless case [56]:

**Lemma 4.4.8.** Let \( T > 0 \) be a fixed time horizon and set \( \beta := \frac{\mu^2 - \lambda^2}{2\gamma \sigma^2} \). For the shadow payoff \( \tilde{X}_T \) corresponding to the policy \( \tilde{\pi}(T) = 1 + w(Y) \) and the shadow discount factor \( \tilde{M}_T := e^{-rT} \mathcal{E}(\int_0^T \frac{\tilde{\pi}}{\sigma} dW_t)_T \), the following bounds hold:

\[ \mathbb{E}[\tilde{X}_T] = X_0^{1-\gamma} e^{(1-\gamma)(r+\beta)T} \mathbb{E}[e^{(1-\gamma)(\tilde{\pi}(Y_T) - \tilde{\pi}(Y_0))}], \]

(4.4.30)

\[ \mathbb{E}[\tilde{M}_T] = e^{(1-\gamma)(r+\beta)T} \mathbb{E}[e^{(1-\gamma)(\tilde{\pi}(Y_T) - \tilde{\pi}(Y_0))}], \]

(4.4.31)

where the function\(^{22}\) \( \tilde{\pi} \) is given by

\[ \tilde{\pi}(y) = \int_0^y - \left( \frac{w'(x)}{w(x)} + w(x) + 1 \right) dx, \]

\(^{22}\)Note that the dividend yield vanishes from the frictionless wealth dynamics.

\(^{23}\)Since the support of \( Y \) is unbounded, we have to verify that \( \int_0^\infty \left( \frac{w'(x)}{w(x)} + w(x) + 1 \right) dx \) exists.

To this end, note that the transformation (4.4.11), the differentiation formula (4.4.16), and the asymptotics of the Whittaker function (4.4.6) imply

\[ \frac{w'(x)}{w(x)} + w(x) + 1 = e^{-x} \left( -\frac{w''(e^{-x})}{w'(e^{-x})} - \gamma \frac{w'(e^{-x})}{(1-\gamma)w(e^{-x})} \right) = O(e^{-x}), \quad \text{as} \quad x \to \infty, \]

so that the integral up to infinity is indeed well defined.
and \( \hat{E}[] \) denotes the expectation with respect to the myopic probability \( \hat{P} \) with density

\[
\frac{d\hat{P}}{dP} = \mathcal{E} \left( \int_0^T \left( -\frac{\hat{\mu}(Y_t)}{\hat{\sigma}(Y_t)} + \hat{\sigma}(Y_t) \hat{\pi}(Y_t) \right) dW_t \right).
\]

**Proof.** First, we verify that the measure \( \hat{P} \) is indeed a probability measure. Inserting the definition of \( \hat{\mu}(Y), \hat{\sigma}(Y) \) and \( \hat{\pi}(Y) \) we have

\[
\frac{d\hat{P}}{dP} = \exp \left( \int_0^T \sigma(1-\gamma)(1+w(Y_t))dW_t - \frac{1}{2} \int_0^T \sigma^2(1-\gamma)^2(1+w(Y_t))^2 dt \right).
\]

Since the function \( w \) can only take values between \([-l/(1+\varepsilon+l), 0]\) by Corollary 4.4.5, Novikov's Condition shows that the density process is a true martingale.

To ease notation in the rest of the proof, we omit the argument \( Y \) of the functions \( \hat{\mu}, \hat{\sigma}, \hat{\delta}, \hat{\pi} \) and \( w \). To obtain (4.4.30), notice that

\[
\tilde{X}_{1-\gamma}^T = \tilde{X}_0^{1-\gamma} e^{(1-\gamma)\int_0^T \left( r + \hat{\mu}(1+w) - \frac{\sigma^2}{2} (1+w)^2 \right) dt + (1-\gamma)\int_0^T \hat{\sigma}(1+w)dW_t}.
\]

Hence:

\[
\tilde{X}_{1-\gamma}^T = \tilde{X}_0^{1-\gamma} \frac{d\hat{P}}{dP} e^{\int_0^T \left[ (1-\gamma)(r+\hat{\mu}(1+w) - \frac{\sigma^2}{2} (1+w)^2) + \frac{1}{2} (-\frac{\sigma^2}{2} + \hat{\sigma}(1+w))^2 \right] dt}
\times e^{\int_0^T \left[ (1-\gamma)\hat{\sigma}(1+w) - (\frac{\sigma^2}{2} + \hat{\sigma}(1+w)) \right] dW_t}.
\]

Inserting the definitions of \( \hat{\mu} \) and \( \hat{\sigma} \), we see that the second integrand simplifies to \(- (1-\gamma) (w'/w + w + 1)\). Similarly, the first integrand reduces\(^{24}\) to \((1-\gamma) H(Y_t)\) with

\[
H(y) := \frac{1}{2w^2} \left[ 2(1-\gamma)\sigma^2 w^3 + (1-\gamma)\sigma^2 w^4 + 2(1-\gamma)\sigma^2 w w' + \sigma^2 (w')^2 \right. \\
\left. + w^2 (2r + (1-\gamma)\sigma^2 + 2(1-\gamma)\sigma^2 w') \right].
\]

In summary,

\[
\tilde{X}_{1-\gamma}^T = \tilde{X}_0^{1-\gamma} \frac{d\hat{P}}{dP} e^{(1-\gamma)\int_0^T H(Y_t) dt} e^{-(1-\gamma)\int_0^T \left( \frac{w'}{w} + w + 1 \right) \sigma dW_t}.
\]

\[\text{(4.4.32)}\]

Itô's formula and the boundary condition (4.4.14) yield

\[
\tilde{q}(Y_T) - \tilde{q}(Y_0) = -\int_0^T \left[ \left( \mu - \frac{\sigma^2}{2} - \delta - \delta e^y \right) \left( \frac{w'}{w} + w + 1 \right) \\
+ \frac{\sigma^2}{2} \left( \frac{w'}{w} + w + 1 \right) \right] dt - \int_0^T \left( \frac{w'}{w} + w + 1 \right) \sigma dW_t.
\]

\[\text{(4.4.33)}\]

\(^{24}\)The calculations and simplifications provided in this proof have been carried out using Mathematica 9; the code is available from the author upon request.
We substitute the first or second derivative of $w$ in accordance with (4.4.12) or (4.4.21), respectively, use this identity to replace the stochastic integral in (4.4.32), and simplify. This yields

$$\tilde{X}_t^{1-\gamma} = \tilde{X}_0^{1-\gamma} \left. \frac{d\hat{P}}{d\tilde{P}} \middle|_{\hat{P}} \right|_{\tilde{P}} e^{(1-\gamma) \int_0^T (r+\beta) dt} e^{(1-\gamma)(\hat{q}(\Upsilon_T)-\hat{q}(\Upsilon_0))}.$$

Thus, the first bound results from taking expectations on both sides. Similarly, inserting the definitions of $\tilde{\mu}$ and $\tilde{\sigma}$, the (shadow) discount factor $\tilde{M}_T$ and the myopic probability measure $\hat{P}$ satisfy

$$\tilde{M}_T^{1-\frac{1}{\gamma}} = e^{\frac{1-\gamma}{\gamma} \int_0^T \frac{\hat{\mu} - \gamma}{\gamma} \left( \hat{\sigma}^2 \right) ds + \frac{1-\gamma}{\gamma} \int_0^T (r + \hat{\mu}^2 + \gamma \hat{\sigma}^2 \hat{\sigma}^2) dt} = \left. \frac{d\hat{P}}{d\tilde{P}} \middle|_{\hat{P}} \right|_{\tilde{P}} e^{\frac{1-\gamma}{\gamma} \int_0^T \frac{\hat{\mu} + \gamma}{\gamma} \left( \hat{\sigma} \right) d\tilde{W}_s}.$$ 

Substituting again the stochastic integral with equation (4.4.33), and replacing the first or second derivative of $w$ in accordance with (4.4.12) or (4.4.21), respectively, shows that

$$\tilde{M}_T^{1-\frac{1}{\gamma}} = \left. \frac{d\hat{P}}{d\tilde{P}} \middle|_{\hat{P}} \right|_{\tilde{P}} e^{\frac{1-\gamma}{\gamma} \int_0^T (r + \beta) dt} e^{\frac{1-\gamma}{\gamma} (\hat{q}(\Upsilon_T)-\hat{q}(\Upsilon_0))}.$$ 

Finally, the second bound results after taking expectations on both sides and raising to the power of $\gamma$.

With the finite-horizon bounds at hand, we can now establish the long-run optimality of the policy $\hat{\pi}(\Upsilon) = 1 + w(\Upsilon)$ in the frictionless shadow market. In contrast to [43, Lemma C.3] the convergence of the finite-horizon bounds requires a more detailed analysis here because the support of the state variable $\Upsilon$ is no longer compact.

**Lemma 4.4.9.** The policy $\hat{\pi}(\Upsilon) = 1 + w(\Upsilon)$ is long-run optimal in the shadow market $(S^0, \hat{S})$ with equivalent safe rate $r + \beta$, where $\beta$ is specified in Lemma 4.4.8. For $t \geq 0$, the corresponding wealth process and the numbers of safe and risky assets are given by

$$\tilde{X}_t = (\xi^0 S^0_t + \xi \hat{S}_t) \times \mathcal{E} \left( \int_0^t r + (1 + w(\Upsilon_s)) \hat{\mu}(\Upsilon_s) ds + \int_0^t (1 + w(\Upsilon_s)) \hat{\sigma}(\Upsilon_s) dW_s \right),$$

$$\varphi_0 = \xi, \quad \varphi_t = (1 + w(\Upsilon_t)) \tilde{X}_t / \hat{S}_t, \quad \varphi_0^0 = \xi^0, \quad \varphi_t^0 = -w(\Upsilon_t) \tilde{X}_t / S^0_t.$$

**Proof.** The formulas for the wealth process and the number of safe and risky units follow directly from the respective definitions.
To verify the optimality of this policy, we use the standard duality bound for power utility [56, Lemma 5], valid for the shadow payoff $\tilde{X}_T^\psi$ of any admissible strategy $(\psi^0, \psi)$ in the shadow market:

$$\mathbb{E}[(\tilde{X}_T^\psi)^{1-\gamma}]^{1/(1-\gamma)} \leq \tilde{X}_0 \mathbb{E}[(\tilde{M}_T)^{2/(1-\gamma)}].$$

This inequality and the second bound (4.4.31) in Lemma 4.4.8 yield an upper bound for the equivalent safe rate:

$$\liminf_{T \to \infty} \frac{1}{(1-\gamma)T \log \mathbb{E}[(\tilde{X}_T^\psi)^{1-\gamma}]} \leq \liminf_{T \to \infty} \frac{\gamma}{(1-\gamma)T \log \mathbb{E}[(\tilde{M}_T)^{2/(1-\gamma)}]}.$$ 

In view of (4.4.31), the upper bound equals $r + \beta$ provided that

$$\liminf_{T \to \infty} \hat{\mathbb{E}}[e^{(1-\gamma)(1+\hat{\psi}(\Upsilon_T))}] < \infty.$$ 

To this end, we first recall the dynamics of the process $\Upsilon$ under the myopic probability $\hat{\mathbb{P}}$. Inserting the definitions of $\tilde{\mu}$, $\tilde{\sigma}$ and $\tilde{\pi}$ yields

$$d\Upsilon_t = \left(\mu - \frac{\sigma^2}{2} - \frac{\delta \Upsilon_t}{l} - \sigma \frac{\tilde{\mu}(\Upsilon_t)}{\tilde{\sigma}(\Upsilon_t)}(1 + w(\Upsilon_t))\right) dt + \sigma d\hat{W}_t$$

$$= \left(\mu - \frac{\sigma^2}{2} - \frac{\delta \Upsilon_t}{l} - \sigma^2 (1 - \gamma)(1 + w(\Upsilon_t))\right) dt + \sigma d\hat{W}_t,$$

where $\hat{W}_t = W_t + \frac{\tilde{\mu}(\Upsilon_t)}{\tilde{\sigma}(\Upsilon_t)} - \frac{\tilde{\sigma}(\Upsilon_t)}{\tilde{\pi}(\Upsilon_t)}$ is a $\hat{\mathbb{P}}$-Brownian motion by Girsanov’s Theorem. Therefore, the scale function and speed measure of the diffusion $\Upsilon$ are given by

$$s(x) = \int_0^x \exp\left(1 - \frac{2\mu}{\sigma^2} + \frac{2\delta}{\sigma^2}\right) \eta + \frac{2\delta}{\sigma^2 l} (e^\eta - 1)$$

$$+ 2(1 - \gamma) \int_0^\eta (1 + w(z))dz\right) d\eta,$$

$$m(dx) = 1_{(0,\infty)}(x) \frac{2dx}{\sigma^2 s'(x)} = 1_{(0,\infty)}(x) \frac{2}{\sigma^2} \exp(\kappa(x))dx,$$

with

$$\kappa(x) := \frac{2\mu}{\sigma^2} - 1 - \frac{2\delta}{\sigma^2} x - \frac{2\delta}{\sigma^2 l} (e^x - 1) - 2(1 - \gamma) \int_0^x (1 + w(y))dy.$$ 

The boundary condition (4.4.13) shows that $\kappa(x)$ is dominated by $-2\delta e^x/(\sigma^2 l)$ for large values of $x$. Therefore, the invariant distribution $\nu(dx) := m(dx)/m([0, \infty))$ is well defined, since exponential decay prevails. By [13, Proposition V.10.3] the

---

\[25\] The reasoning in the case $\pi_+ < 1$ is straightforward (cf. [43, Lemma C.3]), since the corresponding function $\tilde{q}$ is then bounded on the compact support of its argument $\Upsilon$. 

4.4 Proof of Theorem 4.2.4

Process $\Upsilon$ is positively recurrent, so that the ergodic theorem \cite{16, Formula II.35} gives

$$
\lim_{T \to \infty} \hat{E} \left[ e^{(\frac{1}{\gamma} - 1)\hat{q}(\Upsilon_T)} \right] = \int_0^{\infty} e^{(\frac{1}{\gamma} - 1)\hat{q}(y)} \nu(dy)
$$

$$
= \frac{2}{\sigma^2 m([0, \infty))} e^{\frac{2\hat{t}}{\sigma^2}}
\times \int_0^{\infty} \exp \left[ \int_0^y -\left( \frac{1}{\gamma} - 1 \right) \left( \frac{w'(z)}{w(z)} + w(z) + 1 \right) dz \right]
+ \left( \frac{2\mu}{\sigma^2} - 1 - \frac{2\delta}{\sigma^2} \right) y - \frac{2\delta}{\sigma^2} e^y - 2(1 - \gamma) \int_0^y \tilde{\pi}(z)dz \right] dy
< \infty.
$$

In the last step we have again used that the expression in the square bracket is dominated by $-2\delta e^y/\sigma^2$ for large values of $y$. The same idea applied to the first bound (4.4.30) in Lemma 4.4.8 implies that the upper bound $r + \beta$ is attained by the policy $\tilde{\pi}$, which corresponds to the strategy $(\varphi^0, \varphi)$. \hfill $\square$

To conclude that $\tilde{S}$ is indeed a shadow price, it remains to check that the optimal strategy $(\varphi^0, \varphi)$ only acts on $\{\Upsilon_t = 0\}$. \cite{27}

**Lemma 4.4.10.** The number of shares $\varphi = (1 + w(\Upsilon))\tilde{X}/\tilde{S}$ for the optimal policy from Lemma 4.4.9 has the dynamics

$$
d\varphi_t/\varphi_t = \frac{l}{1 + \varepsilon + l} dL_t.
$$

Thus, $\varphi_t$ increases only when $\{\Upsilon_t = 0\}$, that is, when $\tilde{S}_t$ equals the ask price. In particular, it is of finite variation.

**Proof.** Itô’s formula shows

$$
d(1 + w(\Upsilon_t)) = \left( \mu - \delta - \frac{\sigma^2}{2} - \frac{\delta e^{\Upsilon_t}}{l} \right) w'(\Upsilon_t) + \frac{\sigma^2 w''(\Upsilon_t)}{2} dt
$$

$$
+ w'(\Upsilon_t)\sigma dw_t + w'(\Upsilon_t)dL_t.
$$

Integrating $\varphi = (1 + w(\Upsilon))\tilde{X}/\tilde{S}$ by parts twice, using the dynamics of $(1 + w(\Upsilon))\tilde{X}$ and $\tilde{S}$, substituting the first or second derivative of $w$ in accordance with (4.4.12) or (4.4.21), respectively, and simplifying yields \cite{28}

$$
d\varphi_t/\varphi_t = w'(\Upsilon_t)/(1 + w(\Upsilon_t))dL_t.
$$

Since $L_t$ only increases on $\{\Upsilon_t = 0\}$, the assertion therefore follows from the boundary conditions (4.4.14) and (4.4.20). \hfill $\square$

---

26 Note that the function $y \mapsto e^{(\frac{1}{\gamma} - 1)\hat{q}(y)}$ is positive and bounded in $y$ on $[0, \infty)$, cf. Footnote 23.

27 In the case of $\tilde{\pi} < 1$, one analogously verifies that the optimal strategy $(\varphi^0, \varphi)$ only acts on $\{\Upsilon_t = 0\}$ and $\{\Upsilon_t = \log(l/u)\}$.

28 The calculations in this proof have been carried out using Mathematica 9; the code is available from the author upon request.
As the shadow price takes values in the bid-ask spread, it allows the investor to trade at more favorable prices than in the original market with transaction costs. Moreover, dividend payments coincide in both markets by (4.4.27). As the optimal strategy \((\varphi^0, \varphi)\) only entails buying the risky asset when \(\tilde{S}\) coincides with the ask price, it is also feasible with transaction costs, and therefore optimal with the same growth rate.

**Lemma 4.4.11.** The policy \((\varphi^0, \varphi)\) from Lemma 4.4.9 is also long-run optimal in the original market with transaction costs, with the same equivalent safe rate \(r + \beta\).

**Proof.** The formula (4.4.27) shows that the dividend payments in the shadow market equal to its counterparts in the original bid-ask spread market: \(\tilde{\delta} \tilde{S} = \delta S\). The proof now follows along the lines of [43, Proposition C.5].

To complete the proof, we show that the equivalent safe rate of the original market is unique.

**Lemma 4.4.12.** If Assumption 4.4.3 holds, then the corresponding \(\lambda\) is unique.

**Proof.** Suppose there are \(\lambda_1 \neq \lambda_2\) both satisfying Assumption 4.4.3. Following the same lines of reasoning as above there exist two long-run optimal investment strategies in the original market with different long-run growth rates. This is a contradiction.

\(\square\)
Chapter 5

Transaction Costs and Multiple Risky Assets$^1$

5.1 Introduction

A basic problem in asset management is when and how to rebalance portfolios. In doing so, traders need to strike a balance between closely tracking a frictionless target portfolio that implements the optimal risk-return tradeoff, and limiting transaction costs, which make all too frequent trades prohibitively expensive.

In practice, this issue is often addressed by specifying a “suitable” trading frequency, daily, weekly, or monthly rebalancing say, in a more or less ad hoc manner. Theory, however, points out that this approach is not optimal. Indeed, for proportional transaction costs – a reasonable assumption for mid-sized investors – there is a large body of literature documenting that optimal strategies should be “move-” rather than “time-based” [91, 22, 31, 113, 122]. To wit, rebalancing times should not be chosen exogenously, but determined endogenously by the excursions of the investors’ actual portfolios from their frictionless target. In recent years, there has been substantial progress on problems of this type, leading to a rather complete analysis and explicit rebalancing rules in the limiting regime of small transaction costs for the case of a single risky asset [117, 109, 94, 71].

In contrast, much less is known about the practically very important case of many risky assets. There are some numerical results for the bivariate Black-Scholes model [103], but despite some recent progress in asymptotic analysis [15, 107], problems of this kind have remained rather intractable.$^2$

In this chapter, we therefore revisit the simpler time-based approach. In a general multidimensional diffusion setting, we consider an investor with constant relative risk aversion maximizing expected profits penalized for risk and transaction costs.$^3$ In the limit for small transaction costs, we explicitly determine

---

$^1$This chapter is based on Ekren, Liu and Muhle-Karbe [40].

$^2$Models with fixed or quadratic trading costs are more tractable, see [5] or [42, 21, 101, 58], but the issue is only exacerbated for nonlinear transaction costs such as the ones derived from the square-root price impact advocated by many practitioners [2, 120].

$^3$Related “local” criteria are used in [109, 94, 42, 52]. Under suitable integrability conditions
the optimal “time-based rebalancing rule”, for which the next trading time is already specified when the current trade is implemented.\footnote{A special case are the constant trading frequencies considered in the previous literature, e.g., \cite{11, 59}; more generally, the waiting times can depend on current market characteristics.} We also provide an explicit asymptotic formula quantifying the performance loss compared to the frictionless case. These results allow us to shed new light on the rebalancing of (multidimensional) portfolios in a number of ways.

First, for a single risky asset, they allow us to explicitly quantify the suboptimality of time-based relative to move-based rebalancing rules \cite{62, 94, 117, 71}. It turns out that the ratio of the corresponding utility losses compared to the frictionless case is a universal constant, independent of market or preference parameters: \((12/\pi)^{1/3} \approx 1.56\). This number should be compared to the constant \(2^{1/3} \approx 1.26\) differentiating the optimal move-based strategy \cite{52} (which is implemented by reflection off its trading boundaries) from the optimal strategy within the class of strategies that trade back immediately to the frictionless target once these boundaries are breached \cite{97}. Whence, trading back immediately to the frictionless target accounts for approximately 26\% of the losses, whereas the remaining 30\% are due to the pre-specified rebalancing times.

Second, we provide an explicit formula for the performance of a simple benchmark strategy in a general multidimensional setting. In the one-dimensional case, the crucial statistic for the welfare losses due to transaction costs is the diffusion coefficient of the frictionless target weight \cite{122, 62, 71}. In our setting, this role is played by the diffusion matrix of the difference between the frictionless target and its buy-and-hold approximation. This quantity determines both the optimal waiting times and the performance of the corresponding policy through its \(L_{2,1}\)-norm and, weighted with the risky assets’ diffusion matrix, through its trace norm.

Third, we perform a number of numerical tests that compare the performance of our time-based trading rule with various alternatives. For Black-Scholes models with constant investment opportunities, we find that the performance of the optimal time-based rebalancing rule virtually coincides with the optimal move-based portfolio, both for one and for two risky assets. In these settings, the optimal trading frequency for a 1\% transaction cost is less than every two years, so that the particular method for rebalancing does not play a crucial role. To illustrate how this changes in models with stochastic investment opportunities, we consider a truncated version of the model of Kim and Omberg \cite{77}, where expected returns are mean-reverting and optimal strategies are of trend-following type. Here, the optimal trading frequency increases to around once every half year, and optimal move-based strategies offer a marked improvement when they can be computed explicitly for a single risky asset. With more than one risky asset, the no-trade regions characterizing optimal move-based strategies are no longer known explicitly, and numerical algorithms for their computation are not available either. However, we find that even for correlated risky assets, a concatenation of univariate no-trade regions (which is asymptotically optimal for

\footnote{compare, e.g., \cite{68}}, our asymptotic results could be extended to global criteria.
uncorrelated risky assets in the high risk aversion limit [53]) still outperforms a
time-based rebalancing rule.

The message of these results is mixed. With constant investment opportuni-
ties, time-based rebalancing delivers virtually optimal results, with explicit formu-
las both for optimal trading times and their performance in arbitrary dimensions.
Whence, they are an appealingly simple alternative to the less tractable optimal
policies in these settings. In contrast, with stochastic investment opportunities,
e.g., if optimal strategies are of trend following type, our results suggest that even
ad-hoc multidimensional no-trade regions appear to be better suited than time-

based rebalancing plans. Further developments in this direction – maybe using
a more refined discretization scheme\(^5\) as in Gobet and Landon [46] – therefore
remain a challenging direction for future research.

The remainder of this chapter is organized as follows: Section 5.2 introduces
the model and the optimization problem. Our main results are collected in Sec-
tion 5.3. Subsequently, we numerically analyze the performance of our optimal
strategy in different baseline models and compare it to a number of alternatives
by means of Monte Carlo simulations. All proofs are collected in Section 5.5.

5.2 Model

5.2.1 Market

On a filtered probability space \((\Omega, \mathcal{F}, \mathbb{P})\), we consider a financial market which
consists of \(m + 1\) assets: one riskless asset, normalized to one, and \(m\) risky assets
driven by a \(d\)-dimensional Brownian motion \(B = (B_1, \ldots, B_d)\):

\[
\frac{dS^i_t}{S^i_t} = \mu^i(Y_t)dt + \sum_{j=1}^d \sigma^{ij}(Y_t)dB^j_t, \quad i = 1, \ldots, m.
\]

Here, the state variable \(Y = (Y^1, \ldots, Y^p)^	op \in \mathbb{R}^p\) is an autonomous diffusion
process,

\[
dY^i_t = b^i(Y_t)dt + \sum_{j=1}^d g^{ij}(Y_t)dB^j_t, \quad i = 1, \ldots, p.
\]

The expected excess returns of the risky assets are collected in the vector-valued
process \(\mu(Y) = (\mu^1(Y), \ldots, \mu^m(Y))\); the entries \(\sigma^{ij}(Y)\) of the matrix-valued
process \(\sigma(Y)\) describe the exposures of risky asset \(i\) with respect to the shocks
induced by the \(j\)-th component of the Brownian motion \(B\). Likewise, \(b(Y) =
(b^1(Y), \ldots, b^p(Y))\) \(\in \mathbb{R}^m\) and \(g(Y) = (g^{ij}(Y))_{1 \leq i \leq p, 1 \leq j \leq d} \in \mathbb{R}^{p \times d}\) are the drift
coefficient and diffusion matrix of the state variable. Throughout, we impose the
following regularity conditions:

\(^5\)Gobet and Landon [46] study the discretization error for strategies that rebalance when
price increments exit an ellipsoid. Jointly minimizing transaction costs and this discretization
error is more involved, requiring numerical methods to optimize functions from \(\mathbb{R}^d\) to \(\mathbb{R}\) derived
from the expectations of hitting times.
Assumption 5.2.1. Let $E \subset \mathbb{R}^p$ be the support of the state variable $Y$.

(a) We assume that $\mu \in C^2_b(E; \mathbb{R}^m)$, $b \in C^2_b(E; \mathbb{R}^p)$, $\sigma \in C^2_b(E, \mathcal{M}_{m \times d}(\mathbb{R}))$, and $g \in C^2_b(E, \mathcal{M}_{p \times d}(\mathbb{R}))$. That is, the drift and diffusion coefficients are bounded and twice continuously differentiable with bounded derivatives. In particular, $\sup_{y \in E} \|\sigma(y)\|_{2,1} < K_\sigma$ for some constant $K_\sigma > 0$.

(b) The covariance matrix of the risky assets, $\Sigma(Y_t) := \sigma(Y_t)\sigma(Y_t)^T$ is uniformly elliptic, so that its inverse $\Sigma^{-1} = (\xi_{ij})_{1 \leq i,j \leq m}$ exists for all $t \in [0, T]$ and is uniformly bounded.

5.2.2 Frictionless Optimization

To set the stage, we first briefly recapitulate the frictionless case, where portfolio rebalancing is costless. To this end, consider an investor with initial wealth $v > 0$. If she holds fractions $w_t = (w^1_t, \ldots, w^m_t)^T$ in the risky assets at time $t$, her wealth has the following dynamics:

$$\frac{dV_t}{V_t} = \sum_{i=1}^m w^i_t (\mu^i(Y_t) dt + \sigma^i(Y_t) dB_t), \quad V_0 = v.$$  

As in [21, 52], the investor maximizes her expected relative returns, penalized for the corresponding variances. In the continuous-time limit, this leads to the following local mean-variance criterion:

$$F(w) := \frac{1}{T} \mathbb{E} \left[ \int_0^T \frac{dV_t}{V_t} - \frac{\gamma}{2} \int_0^T \frac{d[V]^2}{V^2_t} \right] = \frac{1}{T} \int_0^T \mathbb{E} \left[ (w(Y_t)^T \mu(Y_t) - \frac{\gamma}{2} w(Y_t)^T \Sigma(Y_t) w(Y_t)) dt \right] \rightarrow \max! \quad (5.2.1)$$

Here, the risk aversion parameter $\gamma > 0$ trades off the relative importance of expected returns and variances. Pointwise maximization readily yields that this objective function is maximized by the Merton portfolio:

$$w^*_t(Y_t) := (w^{1,*}_t(Y_t), \ldots, w^{m,*}_t(Y_t))^T = \frac{1}{\gamma} \Sigma^{-1}(Y_t) \mu(Y_t). \quad (5.2.2)$$

The optimal frictionless performance is then obtained by plugging (5.2.2) back into (5.2.1):

$$F(w^*) = \frac{1}{T} \mathbb{E} \left[ \int_0^T \mu(Y_t)^T \Sigma^{-1}(Y_t) \mu(Y_t) \right] = \frac{1}{2\gamma} \mathbb{E} \left[ \int_0^T \mu(Y_t)^T \Sigma^{-1}(Y_t) \mu(Y_t) \right].$$

---

6. Here, the $L_{2,1}$-norm of matrix $A \in \mathbb{R}^{m \times d}$ is defined as follows

$$\|A\|_{2,1} := \sum_{i=1}^m \sqrt{\sum_{j=1}^d |A_{ij}|^2}.$$  

Moreover, $A^i$ denotes the i-th row of the matrix $A \in \mathbb{R}^{m \times d}$.

7. The second equality follows from Assumption 5.2.1(ii). Similar criteria formulated in terms of the absolute increments of the wealth process have been studied by [67] in the frictionless case and by [96, 42, 41, 94] with trading costs.
Throughout, we assume that the Merton portfolio short sells neither the safe nor the risky assets, and never invests all funds into one of these:

Assumption 5.2.2. \(0 \leq w^* i(Y_t) < 1, \ i = 1, \ldots, m,\) and \(0 < \sum_{i=1}^{m} w^* i(Y_t) \leq 1.\)

5.2.3 Introducing transaction costs

Now, we add proportional transaction costs \(\varepsilon > 0\) to the optimization problem (5.2.1). Even with arbitrarily small transaction costs, any strategy of infinite variation leads to immediate bankruptcy. In particular, the Merton portfolio (5.2.2) cannot be implemented, because it is generically a diffusion process.\(^8\)

In the general multidimensional setting considered here, it is infeasible to compute the no-trade region characterizing the optimal policy, even in the limit for small transaction costs. Therefore, we focus here on the less ambitious goal of determining the optimal trading frequency within the class of “time-based” rebalancing strategies, which can be described as follows. At time \(\tau_0 = 0,\) set up the Merton portfolio \(w^*(Y_{\tau_0})\) from (5.2.2), and choose the next rebalancing date \(\tau_1\) based on the current market characteristics. Until time \(\tau_1,\) let the portfolio evolve uncontrolled, before rebalancing it back to the Merton portfolio \(w^*(Y_{\tau_1})\) at \(\tau_1.\) Then, choose the next trading time \(\tau_2\) based on the information available at time \(\tau_1,\) and repeat until the terminal time \(T\) is reached. Since we are eventually interested in the limit \(\varepsilon \downarrow 0\) of small transaction costs, we parametrize the waiting times between successive trades as follows:

Definition 5.2.3. Let \(\alpha \in (0, 2).\) A discretization rule is an adapted, continuous, and positive process \(A\) such that

\[
E \left[ \int_0^T A_t dt \right] < \infty \quad \text{and} \quad E \left[ \frac{1}{\inf_{t \in [0,T]} A_t} \right] < \infty.
\]

The trading times associated with the discretization rule \(A\) and the transaction cost \(\varepsilon > 0\) are given by the following increasing sequence of stopping times:

\[
\tau_0 = 0, \quad \tau_j = \tau_{j-1} + \varepsilon^\alpha A_{\tau_{j-1}}, \ j = 1, 2, \ldots
\]

To wit, the parameter \(\alpha\) governs how trading is sped up with smaller transaction costs, whereas the process \((A_t)_{t \in [0,T]}\) incorporates the current market characteristics. Notice that the second requirement in (5.2.3) implies that the number of trades until maturity is a.s. finite:

\[
N := \inf \{ j \geq 0 : \tau_{j+1} \geq T \} \leq \frac{\varepsilon^{-\alpha} T}{\inf_{t \in [0,T]} A_t} < \infty \quad \mathbb{P} \text{-a.s.}
\]

Moreover, using the continuity of \(A,\) the first requirement in (5.2.3) yields that

\[
\lim_{\varepsilon \to 0} \sup_{j \leq N} \{ \tau_{j+1} - \tau_j \} \leq \lim_{\varepsilon \to 0} \varepsilon^\alpha \sup_{t \in [0,T]} A_t = 0 \quad \mathbb{P} \text{-a.s.}
\]
That is, the mesh width of the discretization is indeed governed by \( \varepsilon^\alpha \). Given a discretization rule and the initial wealth, the corresponding wealth process can be described recursively as follows:

**Definition 5.2.4.** Fix an initial wealth \( v > 0 \) and a discretization rule \( A \). Then, the evolution of the dollar amounts \( V^{\varepsilon,i}(A), i = 0, \ldots, m \) invested in each of the assets and in turn the total wealth \( V^\varepsilon(A) := \sum_{i=0}^m V^{\varepsilon,i}(A) \) evolve as follows. At the initial time \( t = \tau_0 = 0 \), we have

\[
V^{\varepsilon,i}_0(A) := vw^{*,i}(Y_0), \quad i = 1, \ldots, m, \quad V^{\varepsilon,0}_0(A) := v - \sum_{i=1}^m V^{\varepsilon,i}_0.
\]

On \((\tau_{j-1}, \tau_j)\) no transaction is made, so that the wealth process evolves uncontrollably as

\[
V^{\varepsilon,0}_t(A) = V^{\varepsilon,0}_{\tau_{j-1}}(A),
\]

\[
V^{\varepsilon,i}_t(A) = V^{\varepsilon,i}_{\tau_{j-1}}(A) \exp \left( \int_{\tau_{j-1}}^t \left( \mu^i(Y_s) - \frac{1}{2} ||\sigma^i(Y_s)||^2_{\mathbb{R}^d} \right) ds + \int_{\tau_{j-1}}^t \sigma^i(Y_s) dB_s \right),
\]

with associated risky weights \( w^{\varepsilon,i}_t(A) := V^{\varepsilon,i}_t(A)/V^\varepsilon_t(A), i = 0, \ldots, m \). At time \( t = \tau_j \), the portfolio is rebalanced back to the Merton portfolio \( w^*(Y_{\tau_j}) \) from (5.2.2). That is, the dollar amount \( V^{\varepsilon}_{\tau_j-} \Delta L^i_{\tau_j} \) traded in asset \( i \) is determined via

\[
w^{\varepsilon,i}_{\tau_j} = \frac{V^{\varepsilon}_{\tau_j-}(A) \left( w^{\varepsilon,i}_{\tau_j-} + \Delta L^i_{\tau_j} \right)}{V^{\varepsilon}_{\tau_j-}(A) \left( 1 - \varepsilon \sum_{i=1}^m |\Delta L^i_{\tau_j}| \right)} = \frac{w^{\varepsilon,i}_{\tau_j-} + \Delta L^i_{\tau_j}}{1 - \varepsilon \sum_{i=1}^m |\Delta L^i_{\tau_j}|} w^{*,i}_{\tau_j}, \quad i = 1, \ldots, m.
\]

Then, the risky weights \( w^\varepsilon_{\tau_j}(A) \) match the Merton weights \( w^*(Y_{\tau_j}) \) after subtracting the transaction costs from the safe account and in turn the total wealth:

\[
V^\varepsilon_{\tau_j}(A) = V^\varepsilon_{\tau_j-}(A) \left( 1 - \varepsilon \sum_{i=1}^m |\Delta L^i_{\tau_j}| \right).
\]

With the wealth dynamics at hand, we now formulate the investor’s local mean-variance criterion with transaction costs in direct analogy to its frictionless counterpart (5.2.1):

\[
F^\varepsilon(A) := \frac{1}{T} \mathbb{E} \left[ \int_0^T \frac{dV^\varepsilon_t(A)}{V^\varepsilon_t(A)} - \frac{\gamma}{2} \int_0^T \frac{d[V^\varepsilon_t(A)]_t}{V^\varepsilon_t(A)^2} \right] \rightarrow \max!
\]  

(5.2.6)

### 5.3 Main Results

Since the optimization problem (5.2.6) cannot be solved in closed form, we study its limit as the transaction cost \( \varepsilon \) tends to zero and the solution approaches its frictionless counterpart (5.2.2). Asymptotically, the objective function can then be decomposed into its frictionless counterpart, as well as losses caused directly by the transaction costs and displacement from the frictionless target, respectively (compare [108, 62, 63, 71]):
Proposition 5.3.1. Under Assumptions 5.2.1 and 5.2.2, for \( 0 < \alpha < 2 \) and a discretization rule \( A \), the objective function has the following expansion as \( \varepsilon \downarrow 0 \):

\[
F^\varepsilon(A) = \frac{1}{T} \mathbb{E} \left[ \int_0^T \left( \mu(Y_t)^\top \Sigma^{-1}(Y_t) \mu(Y_t) \right) \frac{dt}{2\gamma} \right]
- \frac{1}{T} \mathbb{E} \left[ \text{TAC}(A) + \frac{\gamma}{2} \text{DE}(A) \right] + O(\varepsilon^{2-\alpha}).
\]

(5.3.1)

Here, the transaction costs \( \text{TAC}(A) \) and the discretization error \( \text{DE}(A) \) are given by

\[
\text{TAC}(A) := \varepsilon \sum_{i=1}^m \sum_{j=1}^N |\Delta L_{r_j}^i|,
\]

(5.3.3)

\[
\text{DE}(A) := \int_0^T (w^*(Y_t) - w^\varepsilon(Y_t))^\top \Sigma(Y_t)(w^*(Y_t) - w^\varepsilon(Y_t)) dt.
\]

(5.3.4)

Proof. See Section 5.5.2.

In the above decomposition, the term \( \text{TAC}(A) \) tracks the transaction costs accumulated by applying the discretization rule \( A \). The term \( \text{DE}(A) \) in turn measures the remaining utility loss, which is accrued due to displacement from the frictionless target portfolio. As the transaction cost \( \varepsilon \) tends to zero, these terms tend to zero at different asymptotic rates determined by the parameter \( \alpha \) from Definition 5.2.3:

Lemma 5.3.2. Define \( \beta(Y_t) = (\beta^1(Y_t), \ldots, \beta^m(Y_t))^\top \) with

\[
\beta^i(Y_t) := \tilde{\sigma}^i(Y_t) - w^*(Y_t) \left( \sigma^i(Y_t) - \sum_{k=1}^m w^*k(Y_t) \sigma^k(Y_t) \right) \in \mathbb{R}^d,
\]

where \( \tilde{\sigma}(Y_t) \) is the diffusion coefficient (5.5.2) of the Merton portfolio (5.2.2), and suppose that

\[
\mathbb{E} \left[ \frac{1}{\inf_{t \in [0,T]} \| \beta(Y_t) \|_{2,1}^{2/3}} \right] < \infty.
\]

(5.3.6)

Then, for \( 0 < \alpha < 2 \), the following expansions hold in the limit \( \varepsilon \downarrow 0 \):

\[
\mathbb{E}[\text{TAC}(A)] = \mathbb{E} \left[ \varepsilon \sum_{i=1}^m \sum_{j=1}^N |\Delta L_{r_j}^i| \right]
= \varepsilon^{1-\alpha/2} \mathbb{E} \left[ \sqrt{\frac{2}{\pi}} \int_0^T \left\| \beta_t \right\|_{2,1} \sqrt{A_t} dt \right] + o(\varepsilon^{1-\alpha/2}),
\]

(5.3.7)

\[
\mathbb{E}[\text{DE}(A)] = \mathbb{E} \left[ \int_0^T (w^*_t - w^\varepsilon_t)^\top \Sigma_t(w^*_t - w^\varepsilon_t) dt \right]
= \frac{\varepsilon^\alpha}{2} \mathbb{E} \left[ \int_0^T \text{tr} \left( \beta_t^\top \Sigma_t \beta_t \right) A_t dt \right] + o(\varepsilon^\alpha),
\]

(5.3.8)

where all the expectations are positive and finite.
Proof. See Section 5.5.3.

Rebalancing more frequently evidently reduces the discretization error but also increases the incurred transaction costs. Therefore, to maximize the objective function \( F^\varepsilon \) over discretization rules we have to choose \( \alpha > 0 \) so that the leading orders of both error terms are of the same magnitude. Lemma 5.3.2 shows that the leading orders match at \( \varepsilon^{2/3} \) for \( \alpha = 2/3 \). With this choice, maximizing the local mean variance criterion with transaction costs is at the leading order tantamount to minimizing the sum of i) transaction costs and ii) discretization error, weighted by risk aversion.

**Definition 5.3.3.** A discretization rule \( A \) is called asymptotically optimal if it minimizes the leading-order total cost,

\[
TC(A) := \lim_{\varepsilon \to 0} \frac{E[TAC(A) + \frac{\gamma^2}{2} DE(A)]}{\varepsilon^{2/3}}.
\]

The optimal discretization rule and performance for this asymptotic criterion can be computed explicitly:

**Theorem 5.3.4.** Suppose Assumptions 5.2.1, 5.2.2 and Condition (5.3.6) are satisfied. Then,

\[
TC(A) = E \left[ \int_0^T \frac{\gamma^2}{4} \text{tr} \left( \beta(Y_t) \Sigma(Y_t) \beta(Y_t) \right) A_t dt + \frac{\sqrt{2}}{\pi} \int_0^T \frac{\|\beta(Y_t)\|_{2,1}}{\sqrt{A_t}} dt \right]. \tag{5.3.9}
\]

An asymptotically optimal discretization rule is given by

\[
A^\ast(Y_t) = \left( \frac{\sqrt{2} \|\beta(Y_t)\|_{2,1}}{\frac{\gamma}{2} \text{tr} \left( \beta(Y_t) \Sigma(Y_t) \beta(Y_t) \right)} \right)^{2/3}, \tag{5.3.10}
\]

with associated trading times \( \tau_0^\ast = 0 \) and \( \tau_j^\ast = \tau_{j-1}^\ast + \varepsilon^{2/3} A^\ast(Y_{\tau_{j-1}}) \), \( j = 1, 2, \ldots \). The corresponding minimal leading-order total cost is given by

\[
TC(A^\ast) = \frac{3}{2} E \left[ \int_0^T \left( \frac{\sqrt{2}}{\pi} \|\beta(Y_t)\|_{2,1} \right)^{2/3} \left( \frac{\gamma}{2} \text{tr} \left( \beta(Y_t) \Sigma(Y_t) \beta(Y_t) \right) \right)^{1/3} dt \right]. \tag{5.3.11}
\]

The crucial quantity in the above formulas is the vector \( \beta \). Its entries measure the difference between the diffusion coefficients of the frictionless target weight and the discretely rebalanced version, cf. Lemma 5.5.1 and Lemma 5.5.3. The asymptotically optimal trading frequency and the corresponding welfare loss relative to the frictionless case are completely determined by \( \beta \) and the covariance.

---

9 A similar criterion in terms of absolute quantities is directly used by [109].

10 Note that even though the numbers of shares in the discretely rebalanced portfolio are of finite variation, this is not the case for the corresponding risky weights, which fluctuate with diffusive price shocks.
matrix $\Sigma$ of the risky assets – like in other models with small trading costs, the expected returns do not contribute directly at the leading order.

The derivation of Formula (5.3.11) reveals that direct transaction costs contribute two thirds of the total cost $TC(A^*)$, while the remaining one third is due to discretization error. Maybe surprisingly, these universal relative contributions do not depend on the number of risky assets, and also agree with the corresponding result for univariate asymptotically optimal move-based strategies [71].

Mutatis mutandis, the same arguments can also be used to assess the performance of a constant rebalancing frequency, i.e., trading times $\tau_j = \tau_{j-1} + \varepsilon^{2/3} A$ for some constant $A > 0$. These policies are even easier to interpret and implement, but lead to somewhat more cumbersome formulas. Indeed, the optimal constant discretization rule turns out to be

$$A^* = \left( \frac{E \left[ \sqrt{\frac{2}{\pi}} \int_0^T \|\beta(Y_t)\|_{2,1} dt \right]}{\frac{3}{2} \int_0^T \text{tr} (\beta(Y_t)\Sigma(Y_t)\beta(Y_t)) dt} \right)^{2/3},$$

(5.3.12)

which leads to a total cost of

$$TC(A^*) = \frac{3}{2} \left( \frac{E \left[ \sqrt{\frac{2}{\pi}} \int_0^T \|\beta(Y_t)\|_{2,1} dt \right]}{\frac{3}{2} \int_0^T \text{tr} (\beta(Y_t)\Sigma(Y_t)\beta(Y_t)) dt} \right)^{2/3} \times \left( \frac{E \left[ \frac{\gamma}{2} \int_0^T \text{tr} (\beta(Y_t)\Sigma(Y_t)\beta(Y_t)) dt \right]}{\frac{3}{2} \int_0^T \text{tr} (\beta(Y_t)\Sigma(Y_t)\beta(Y_t)) dt} \right)^{1/3}.$$

5.4 Examples and Implications

In this section, we illustrate our results with a number of examples. First, we consider the case of a single risky asset, and compare the performance of our time-based rebalancing rule with the asymptotically optimal move-based policies [52, 97]. Afterwards, we turn to models with two risky assets, where we use Monte Carlo simulations to benchmark our policies against i) a simple buy-and-hold strategy, ii) univariate move-based strategies pasted together component-wise, and iii) the optimal bivariate move-based strategy, computed numerically [6].

5.4.1 Single Asset

**Black-Scholes Model** We first consider the univariate Black-Scholes model, where the expected excess return $\mu$, the volatility $\sigma$ of the risky asset, and in turn the frictionless Merton portfolio $w^* = \mu/(\gamma \sigma^2)$ are positive constants. Then, the volatility $\tilde{\sigma}$ of $w^*$ vanishes, and we tacitly assume that $\gamma > \mu/\sigma^2$ so that Assumption 5.2.2 is satisfied and (5.3.6) holds, too, because

$$\beta = \sigma(1 - w^*)w^* \in (0, 1).$$
Table 5.1: Simulated expected profits (5.2.6) for different strategies. Parameters are: sample size \(N = 10^6\), \(dt = 1/250\), \(\mu = 8\%\), \(\sigma = 16\%\), \(\gamma = 5\), \(T = 20\), and \(\varepsilon = 1\%\).

<table>
<thead>
<tr>
<th>strategy</th>
<th>simulated profits</th>
</tr>
</thead>
<tbody>
<tr>
<td>frictionless move based</td>
<td>2.50%</td>
</tr>
<tr>
<td>time based</td>
<td>2.47%</td>
</tr>
<tr>
<td>buy &amp; hold</td>
<td>2.46%</td>
</tr>
<tr>
<td>buy &amp; hold</td>
<td>2.32%</td>
</tr>
</tbody>
</table>

Whence, Theorem 5.3.4 is applicable and yields the following asymptotically optimal discretization rule:

\[
A^* = \left( \frac{\sqrt{2} \sigma |(1 - w^*)w^*|}{2 \sigma^4 (1 - w^*)^2 (w^*)^2} \right)^{2/3} = \left( \frac{8/\pi}{\gamma \sigma^3 |w^*(1 - w^*)|} \right)^{2/3}.
\]

Thus, the waiting periods between the trading times \(\tau_{j+1}^* = \tau_j^* + \varepsilon^{2/3} A^*\) are long if i) transaction costs \(\varepsilon\) are substantial, ii) the target portfolio is close to a buy-and-hold strategy, iii) risk aversion \(\gamma\) is low, or iv) the market volatility \(\sigma\) is low. The corresponding leading-order performance loss is given by

\[
\sigma^2 T \left( \frac{27}{8\pi} \gamma \varepsilon^2 |w^*(1 - w^*)|^4 \right)^{1/3}.
\]

It is equivalent to an annuity accrued in business time \(\sigma^2 T = d\langle S \rangle_T / S_T^2\), and determined by risk aversion \(\gamma\), transaction costs \(\varepsilon\), and the term \(|w^*(1 - w^*)|\) which measures the frictionless optimizer’s distance from a buy-and-hold strategy.\(^{11}\) The same ingredients also appear in the asymptotically optimal move-based performance studied in [52]:

\[
\sigma^2 T \left( \frac{9}{32} \gamma \varepsilon^2 |w^*(1 - w^*)|^4 \right)^{1/3}.
\]

Whence, the asymptotic welfare loss that can be achieved by a move-based strategy differs from the optimal time-based performance by a universal factor of \((\frac{12}{\pi})^{1/3} \approx 1.56\), independent of market and preference parameters. This complements a result of [97], who find that, for unit relative risk aversion,\(^{12}\) a similar relationship holds for move-based strategies that rebalance by means of reflection off the trading boundaries, and move-based strategies that rebalance back directly to the frictionless target. Then, the universal constant linking the two respective optimizers is \(2^{1/3} \approx 1.26\). In summary, about 26% of the performance are lost due to trading back immediately to the Merton portfolio, whereas the remaining 30% stem from the predetermined rebalancing times.

To get a feel for the magnitude of these effects, let us consider a concrete example. For \(\mu = 8\%\), \(\sigma = 16\%\), and \(\gamma = 5\), the frictionless target portfolio is

\(^{11}\)This quantity can also be interpreted as the sensitivity of the frictionless target weight with respect to relative changes of the stock price, compare [62].

\(^{12}\)Heuristic arguments as in [71, 97] suggest that the same universal constant also applies for more general preferences.
given by \( w^* = 62.5\% \). For a 1% transaction cost, our time-based rebalancing rule in turn prescribes to trade once every 2.23 years. Even for transaction costs of only \( \varepsilon = 0.1\% \), the waiting times are still almost 6 months. Accordingly, the welfare losses of all rebalancing strategies are relatively small. Indeed, the difference between the optimal time- and move-based rebalancing rules turns out to be negligibly small here, cf. Table 5.1.

**Kim-Omberg Model** Next, we pass to a setting where stochastic investment opportunities provide additional motives to trade, which makes transaction costs more important, cf. [89]. To this end, we consider a variant of the model of Kim and Omberg [77], where expected returns are mean reverting. In order to satisfy our regularity assumptions, we consider a version where the expected returns are truncated for excessively large values of the state variable:

\[
\begin{align*}
\sigma \tilde{Y} & = \tilde{Y} - \bar{y} \eta \sigma - \tilde{Y} \sigma (y)
\end{align*}
\]

where \( \tilde{Y} \) is a \((0,1)\)-valued function. The state variable \( Y \) follows a modified Ornstein-Uhlenbeck process with long-run mean \( 0 < \tilde{Y} \in (y_{\text{min}}, y_{\text{max}}) \), mean reversion speed \( \lambda_Y > 0 \). The diffusion vector of \( Y \) is then given by \((\alpha_Y \eta, \alpha_Y \sqrt{1 - \eta^2})\), where \( \alpha_Y > 0 \) and \( \eta \in (-1, 1) \).

The additional state variable implies that the frictionless Merton solution is no longer constant but stochastic \( w^*(Y_t) = \mu(Y_t) / \gamma \sigma^2 \). We assume that the risk aversion \( \gamma \) and the cut-off levels \( y_{\text{min}}, y_{\text{max}} \) for \( \mu^1 \) are chosen so that \( \delta_1 \leq w^* \leq \delta_2 \) for \( 0 < \delta_1 < \delta_2 < 1 \). Then, the Merton proportion \( w^*(Y) \) is again an Itô process (cf. Lemma 5.5.1):

\[
dw^*(Y_t) = \hat{\mu}(Y_t) dt + \hat{\sigma}(Y_t) dB_t,
\]

with

\[
\hat{\sigma}(Y_t) = \begin{cases} 
\frac{1}{\gamma \sigma^2} (\alpha_Y \eta, \alpha_Y \sqrt{1 - \eta^2})^T, & \text{for } Y_t \in [y_{\text{min}} + \xi, y_{\text{max}} - \xi], \\
\frac{1}{\gamma \sigma^2} (\alpha_Y \eta, \alpha_Y \sqrt{1 - \eta^2})^T, & \text{for } Y_t \in (y_{\text{min}}, y_{\text{min}} + \xi) \cup (y_{\text{max}} - \xi, y_{\text{max}}), \\
0, & \text{otherwise}.
\end{cases}
\]
In particular, there exists a constant trading frequency (5.3.12), which is a straightforward computation shows that the matrix \( \beta(Y) \) from (5.3.5) is given by

\[
\beta(Y) = \begin{cases}
\frac{1}{\sqrt{\pi}} (\alpha Y \eta, \alpha Y \sqrt{1 - \eta^2})^\top - w^*(Y_t)(1 - w^*(Y_t)) (\sigma, 0)^\top, \\
\text{for } Y_t \in [y_{\min} + \xi, y_{\max} - \xi], \\
\frac{s(Y_t)}{\sqrt{\pi}} (\alpha Y \eta, \alpha Y \sqrt{1 - \eta^2})^\top - w^*(Y_t)(1 - w^*(Y_t)) (\sigma, 0)^\top, \\
\text{for } Y_t \in (y_{\min}, y_{\min} + \xi) \cup (y_{\max} - \xi, y_{\max}), \\
-w^*(Y_t)(1 - w^*(Y_t)) (\sigma, 0)^\top, \\
\end{cases}
\]

In particular, there exists a \( C > 0 \) such that

\[
\|\beta(Y_t)\|_{2,1} \geq \begin{cases}
|\beta(Y_t)|^{12} = \frac{\alpha Y \sqrt{1 - \eta^2}}{\sqrt{\pi}} > 0, \text{ for } Y_t \in [y_{\min} + \xi, y_{\max} - \xi], \\
\geq C\|\beta(Y_t)\|_1 = C \left( \frac{s(Y_t)}{\sqrt{\pi}} \alpha Y \eta - \sigma w^*(Y_t)(1 - w^*(Y_t)) \right) \\
+ \left| \frac{s(Y_t)}{\sqrt{\pi}} \alpha Y \sqrt{1 - \eta^2} \right| > 0, \\
\text{for } Y_t \in (y_{\min}, y_{\min} + \xi) \cup (y_{\max} - \xi, y_{\max}), \\
\end{cases}
\]

which in turn implies that Assumption 5.2.2 and Condition (5.3.6) are satisfied. As a result, Theorem 5.3.4 shows that the optimal trading frequency now depends on the state variable and is given by

\[
(\tau^*_j - \tau^*_{j-1})(Y_{t-1}) = \left( \frac{8}{\pi \gamma \sigma^2 \|\beta(Y_{t-1})\|_{2,1}} \right)^{2/3}, \quad \forall j \geq 1,
\]

where \( \|\beta(y)\|_{2,1} = 2\sigma^2 \left( \frac{1}{2} (w^*(y)(1 - w^*(y)))^2 - \frac{\alpha Y \eta}{\gamma \sigma^2} w^*(y)(1 - w^*(y)) + \frac{\alpha^2}{2 \gamma^2 \sigma^2} \right) \) on \([y_{\min} + \xi, y_{\max} - \xi]\). The corresponding leading-order performance loss (5.3.11) reads as follows:

\[
\sigma^2 E \left[ \int_0^T \left( \frac{27}{2 \pi \gamma \varepsilon^2 \|\beta(Y_t)\|_{2,1}^2} \right)^{1/3} dt \right].
\]

For unit relative risk aversion, the leading-order loss of the optimal move-based strategy is again smaller by a universal factor of \((12/\pi)^{1/3}\), just like in the Black-Scholes model, [97, Theorem 4.1]. Heuristic arguments as in [71, 97] again suggest that this relationship remains true for more general preferences.

To illustrate these results, we consider parameters estimated from a long time series of US equity market data [7]:

\[
\bar{Y} = 5.60\%, \quad \alpha_Y = 3.68\%, \quad \lambda_Y = 0.2712, \quad \sigma = 14.28\%, \quad \eta = -0.9351.
\]

Table 5.2 collects Monte-Carlo estimates for the performances of the optimal time-based rebalancing rule, its move-based counterpart, and a simple buy-and-hold strategy. In addition, we also consider the strategy associated to the optimal constant trading frequency (5.3.12), which is 6.7 months for a 1% transaction.
5.4 Examples and Implications

<table>
<thead>
<tr>
<th></th>
<th>frictionless move based</th>
<th>time based</th>
<th>constant frequency</th>
<th>buy &amp; hold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.65%</td>
<td>2.28%</td>
<td>2.09%</td>
<td>2.09%</td>
</tr>
</tbody>
</table>

Table 5.2: Simulated expected profits (5.2.6) for different strategies. Parameters are taken from [7]: $Y = 5.60\%$, $\alpha_Y = 3.68\%$, $\lambda_Y = 0.2712$, $\sigma = 14.28\%$, $\gamma = 5$, $\eta = -0.9351$, $T = 20$, $\varepsilon = 1\%$, $dt = 1/250$ and $N = 10^6$.

cost and an investor with risk aversion $\gamma = 5$. We observe that the differences between the various strategies are much more pronounced than for the Black-Scholes model, in line with the results of [89]. However, the relative magnitudes of these differences are virtually the same, and in excellent agreement with our asymptotic results. Finally, note that adapting the time-based rule to changing market characteristics only has a very small effect here; the performance of the simple constant discretization rule is virtually the same.

5.4.2 Two Risky Assets

**Black-Scholes Model** Now, we turn to a Black-Scholes model with two risky assets, with expected excess returns $\mu = (\mu^1, \mu^2) \in \mathbb{R}^2_{>0}$ and diffusion matrix

$$\sigma = \begin{pmatrix} \alpha & 0 \\ v \cdot \rho & v \sqrt{1 - \rho^2} \end{pmatrix},$$

i.e., the risky assets have volatilities $\alpha, v > 0$, and correlation $\rho \in [-1,1]$. As before, we assume that the risk aversion is sufficiently large so that the Merton portfolio

$$w^* := (w^{*,1}, w^{*,2})^\top = \frac{1}{\gamma} \Sigma^{-1} \mu, \quad \text{with} \quad \Sigma = \sigma \sigma^\top,$$

satisfies Assumption 5.2.2. The diffusion coefficient of this constant portfolio is zero; using (5.3.5) a straightforward calculation shows

$$\beta = \begin{pmatrix} (w^{*,1} - 1)w^{*,1}\alpha + w^{*,1}w^{*,2}v\rho & w^{*,1}w^{*,2}v\sqrt{1 - \rho^2} \\ (w^{*,1}w^{*,2}\alpha + w^{*,2}(w^{*,2} - 1)v\rho & w^{*,2}(w^{*,2} - 1)v\sqrt{1 - \rho^2} \end{pmatrix}.$$  

Therefore, we have $\|\beta\|_{2,1} \geq |\beta^{22}| > 0$ if $w^{*,2} > 0$ and $\|\beta\|_{2,1} = |(w^{*,1} - 1)w^{*,1}\alpha| > 0$ if $w^{*,2} = 0$, which shows that Condition (5.3.6) is satisfied. Theorem 5.3.4 in turn yields that the optimal trading times are given by

$$\tau_{j+1}^* - \tau_j^* = \left( \frac{\varepsilon \sqrt{\frac{2}{\sigma}} \|\beta\|_{2,1}}{\frac{3}{2} \cdot \text{tr} (\beta^\top \Sigma \beta)} \right)^{2/3}.$$  

For risky assets with identical expected excess returns and volatilities, the effect of a nonzero correlation is illustrated in Figure 5.1 (left panel). For very
Figure 5.1: Left Panel: The optimal waiting times $\tau_j^* - \tau_{j-1}^*$ plotted for various correlations $\rho$ in the two-dimensional (solid) or one-dimensional (dotted) Black-Scholes model. Right Panel: Expected profits (5.2.6) plotted for various correlations $\rho$ in the two-dimensional (solid) or one-dimensional (dotted) Black-Scholes model. Parameters are $\mu^1 = \mu^2 = 8\%$, $\alpha = v = 16\%$, $\gamma = 5$ and $\varepsilon = 1\%$.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>frictionless</th>
<th>move based</th>
<th>time based</th>
<th>buy &amp; hold</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>3.84%</td>
<td>3.80%</td>
<td>3.77%</td>
<td>3.67%</td>
</tr>
<tr>
<td>0.6</td>
<td>3.12%</td>
<td>3.09%</td>
<td>3.08%</td>
<td>3.01%</td>
</tr>
<tr>
<td>0.9</td>
<td>2.62%</td>
<td>2.60%</td>
<td>2.59%</td>
<td>2.48%</td>
</tr>
</tbody>
</table>

Table 5.3: Simulated expected profits (5.2.6) for different strategies. Parameters are: sample size $N = 10^6$, $dt = 1/250$, $\mu^1 = \mu^2 = 8\%$, $\alpha = v = 16\%$, $\gamma = 5$, $T = 20$, and $\varepsilon = 1\%$.

large correlation ($\rho \approx 1$), the market is essentially equivalent to one with only a single risky asset. Accordingly, the optimal trading frequency converges to its univariate counterpart as $\rho \uparrow 1$. For intermediate correlations $\rho \in (0,1)$, the relationship is surprisingly non-monotonic even though the associated optimal welfare (5.2.6) is decreasing in $\rho$ (Figure 5.1, right panel).

**Numerical Results** We again compare the performance of our time-based policy to a number of alternatives. The first benchmark is the asymptotically optimal move-based strategy. For more than one risky asset, explicit formulas are no longer available because one still needs to solve a free boundary problem even after passing to the small cost limit [107]. For our illustration, we use the policy iteration algorithm proposed in [6] to carry out these computations. The second competitor is again a simple buy-and-hold portfolio. The simulated performances are collected in Table 5.3. As in the one-dimensional case, the difference between the optimal time- and move-based strategies is very small. Whence, with constant investment opportunities, our simple, explicit trading rule is an appealing alternative to the much less tractable optimizer.
Kim-Omberg Model Finally, we consider a Kim and Omberg-type model with two risky assets:

\[
\begin{align*}
    dS_1^i &= S_1^i \left( \mu^1(Y_i)dt + \alpha dB_1^i \right), \\
    dS_2^i &= S_2^i \left( \mu^2(Y_i)dt + v\rho dB_1^i + v\sqrt{1-\rho^2} dB_2^i \right), \\
    dY_i &= \lambda Y_i (Y_i - \mu(Y_i)) dt + \alpha\eta dB_1^i + \alpha\gamma \sqrt{1-\eta^2} dB_2^i;
\end{align*}
\]

Similarly as in the one-dimensional case, \( B = (B_1^1, B_2^1)^\top \) is two-dimensional Brownian motion and for each \( i \in \{1, 2\} \) the function \( \mu^i \) is a smooth truncation of the identity function with cut-off levels \( y_{i,\text{min}} \) and \( y_{i,\text{max}} \) and the associated small positive constant \( \xi_i \), chosen so that \( \mu^i \) is \( C^2 \) and has bounded derivative. Again, the state variable \( Y \) follows a modified Ornstein-Uhlenbeck process with long-run mean \( 0 < \bar{Y} \in (y_{1,\text{min}}, y_{1,\text{max}}) \), mean reversion speed \( \lambda_Y > 0 \) and volatility vector \( (\alpha\eta, \alpha\gamma \sqrt{1-\eta^2})^\top \) with \( \eta \in (-1, 1) \).

The volatility matrix of the risky assets \( S = (S_1^1, S_2^1)^\top \) is given by

\[
\sigma = \begin{pmatrix}
    \alpha & 0 \\
    v \cdot \rho & v\sqrt{1-\rho^2}
\end{pmatrix},
\]

with \( \rho \in (-1, 1) \) and \( 0 < \alpha < 1/\rho \) in case of \( \rho > 0 \), and the frictionless Merton portfolio is

\[
w^*(Y_i) := (w^{*1}(Y_i), w^{*2}(Y_i))^\top = \frac{1}{\gamma} \Sigma^{-1}(\mu^1(Y_i), \mu^2(Y_i))^\top,
\]

with \( \Sigma = \sigma \sigma^\top \) and

\[
\Sigma^{-1} = (\xi^{ij})_{1 \leq i, j \leq 2} = \frac{1}{\alpha^2 v^2 (1-\rho^2)} \begin{pmatrix}
v^2 & -\rho \alpha v \\
-\rho \alpha v & \alpha^2
\end{pmatrix}.
\]

We assume that the risk aversion \( \gamma \) and the cut-off levels for \( \mu^1, \mu^2 \) are chosen so that

\[
\delta_1 \leq w^{*,1} < 1, \quad \delta_2 \leq w^{*,2} < 1, \quad 0 < w^{*,1} + w^{*,2} \leq 1,
\]

for some \( \delta_1, \delta_2 > 0 \). Similar calculations as before yield that for \( i \in \{1, 2\} \) and \( y \in [y_{1,\text{min}} + \xi_1, y_{1,\text{max}} - \xi_1] \) the diffusion coefficients of the Merton proportion \( \hat{\sigma}^i \) and the matrix \( \beta(Y_i) \) are given by

\[
\begin{align*}
    \hat{\sigma}^i(Y_i) &= \frac{1}{\gamma} (\xi^{i1} + \xi^{i2})(\alpha\gamma \eta, \alpha\gamma \sqrt{1-\eta^2})^\top, \\
    \beta^1(Y_i) &= \hat{\sigma}^1(Y_i) - w^{*,1}(Y_i)((\alpha,0)^T - w^{*,1}(Y_i)(\alpha,0)^T - w^{*,2}(Y_i)(v\rho, v\sqrt{1-\rho^2})^T) \\
    \beta^2(Y_i) &= \hat{\sigma}^2(Y_i) - w^{*,2}(Y_i)((v\rho, v\sqrt{1-\rho^2})^T - w^{*,1}(Y_i)(\alpha,0)^T \\
    &- w^{*,2}(Y_i)(v\rho, v\sqrt{1-\rho^2})^T).
\end{align*}
\]

\(^{13}\)Note that we use the same state variable for both risky assets here, but allow for different degrees of correlations with return shocks.
<table>
<thead>
<tr>
<th>$\rho$</th>
<th>frictionless</th>
<th>pasted</th>
<th>time based</th>
<th>buy &amp; hold</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>4.07%</td>
<td>3.53%</td>
<td>3.26%</td>
<td>2.25%</td>
</tr>
<tr>
<td>0.6</td>
<td>3.31%</td>
<td>2.80%</td>
<td>2.64%</td>
<td>1.79%</td>
</tr>
<tr>
<td>0.9</td>
<td>2.79%</td>
<td>2.27%</td>
<td>2.21%</td>
<td>1.40%</td>
</tr>
</tbody>
</table>

Table 5.4: Simulated expected profits (5.2.6) for different strategies. Parameters are taken from [7]: $\bar{Y}^1 = \bar{Y}^2 = 5.60\%$, $\alpha_Y = \nu_Y = 3.68\%$, $\lambda_1^Y = \lambda_2^Y = 0.2712$, $\alpha = v = 14.28\%$, $\gamma = 5$, $\eta = -0.9351$, $T = 20$, $\varepsilon = 1\%$, $dt = 1/250$ and $N = 10^6$.

On $\mathbb{R}\setminus[y_{1,\text{min}} + \xi_1, y_{1,\text{max}} - \xi_1]$ similar formulas can be derived. In particular, we have for all $y \in \mathbb{R}$

$$\|\beta(y)\|_{2,1} \geq |\beta^{12}(y)| \geq w^{*1}(y)w^{*2}(y)v\sqrt{1-\rho^2} > \delta_1\delta_2v\sqrt{1-\rho^2} > 0,$$

which in turn shows that Assumption 5.2.2 and Condition (5.3.6) are satisfied.

Numerical Results  In Table 5.4, we again compare the performance of our time-based rebalancing rules to two alternative strategies. With mean-reverting returns, time-based rebalancing offers marked gains compared to buy-and-hold. However, the performance loss compared to the frictionless benchmark is also not negligible, similarly as in the one-dimensional case. The asymptotically optimal move-based strategy in this setting is not known explicitly, and even computing it numerically would be quite involved. As a partial remedy, we therefore consider a “pasted” move-based strategy, where the univariate no-trade regions are simply concatenated. This is motivated by results that link portfolio choice with high constant relative risk aversion to constant absolute risk aversion [105, 53], for which the multivariate problem factorizes for uncorrelated assets [84, 53]. Whence, these policies are expected to be useful proxies if relative risk aversion is sufficiently high and correlation between assets is sufficiently low. We find that for relative risk aversion $\gamma = 5$, the pasted strategy outperforms the time-based rule even for high correlations, though the difference declines as correlation is increased.

5.5 Proofs of Proposition 5.3.1, Lemma 5.3.2 and Theorem 5.3.4

Throughout this section $C$ stands for a generic positive constant that may vary from expression to expression. Moreover, we often suppress the dependence on the state variable $Y$ to ease notation. For example, $\mu'(Y_t)$ is frequently abbreviated to $\mu'_t$.

5.5.1 Portfolio Dynamics

First, we compute the dynamics of the Merton portfolio (5.2.2) and its discretized version $w^\varepsilon$ (cf. Definition 5.2.4):
Lemma 5.5.1. Under Assumptions 5.2.1 and 5.2.2, the Merton portfolio $w^*(Y_t)$ is an Itô process with the following dynamics:

$$dw^*{}_{i}^{}(Y_t) = \bar{\mu}^i(Y_t)dt + \bar{\sigma}^i(Y_t)dB_t, \quad i = 1, \ldots, m, \quad (5.5.1)$$

with

$$\mathbb{R}^d \ni \bar{\sigma}^i(Y_t) = \left( \sum_{l=1}^{m} \sum_{k=1}^{m} (w^{}_{i}^{*l}(Y_t))_{\xi}^{lk}(Y_t)(\nabla \Sigma^{kl})(G(Y_t)) \right) + \frac{1}{\gamma} \xi^i(Y_t)(\nabla \mu^l)(G(Y_t))^\top, \quad (5.5.2)$$

where

$$\nabla \Sigma^{kl} = \left( \frac{\partial \Sigma^{kl}}{\partial y_1}, \ldots, \frac{\partial \Sigma^{kl}}{\partial y_p} \right) (Y_t), \quad \nabla \mu^l = \left( \frac{\partial \mu^l}{\partial y_1}, \ldots, \frac{\partial \mu^l}{\partial y_p} \right) (Y_t).$$

Moreover, its drift and diffusion coefficients are bounded,

$$\sup_{y \in \mathcal{E}} \|\bar{\mu}(y)\|_{\mathbb{R}^m} < \infty, \quad \sup_{y \in \mathcal{E}} \|\bar{\sigma}(y)\|_{2,1} < \infty.$$

Proof. Under Assumption 5.2.1, the Itô representation and the formulas for the corresponding drift and diffusion coefficients follow from Itô’s formula. Moreover, as the Merton portfolio shorts none of the assets by Assumption 5.2.2 ($0 \leq w^{}_{i}^{*l}(Y_t) < 1$ for $i = 1, \ldots, m$ and $t \in [0, T]$), all functions appearing in (5.5.2) are bounded. An analogous argument applies for the drift coefficients $\bar{\mu}^i$. \hfill \Box

Corollary 5.5.2. Under Assumptions 5.2.1 and 5.2.2, $\beta^j_t = (\beta^1_t, \ldots, \beta^m_t)^\top$ with $\beta^i$ as in (5.3.5) has bounded $L_{2,1}$ norm, i.e., there exists a constant $K_{\beta} < \infty$ such that $\sup_{y \in \mathcal{E}} \|\beta(y)\|_{2,1} < K_{\beta}$.

We now turn to the dynamics of the discretized Merton portfolio:

Lemma 5.5.3. Define the rebalancing times $(\tau_j)_{j \in \mathbb{N}}$ as in Definition 5.2.3. Under Assumptions 5.2.1 and 5.2.2, the risky weights $w^\varepsilon = (w^\varepsilon,1, \ldots, w^\varepsilon,m)^\top$ from Definition 5.2.4 satisfy:

$$dw^\varepsilon{}_{i}^{}(t) = w^\varepsilon{}_{i}^{}(t) \left( \mu^w_{i} - m \sum_{k=1}^{m} w^\varepsilon{}_{i}^{k} \sigma^{k}_{i} \right) dt + \left( \sum_{k=1}^{m} w^\varepsilon{}_{i}^{k} \sigma^{k}_{i} \right) dB_t, \quad \text{on } [\tau_{j-1}, \tau_j), \quad j \geq 1, \quad (5.5.3)$$

with

$$\mu^w_{i} := \mu^i_t - w^\varepsilon{}_{i}^{}(t) \mu_t - \sigma^{i}_{i} \sum_{k=1}^{m} w^\varepsilon{}_{i}^{k} \sigma^{k}_{i} + \left( \sum_{k=1}^{m} w^\varepsilon{}_{i}^{k} \sigma^{k}_{i} \right) \left( \sum_{k=1}^{m} w^\varepsilon{}_{i}^{k} \sigma^{k}_{i} \right), \quad i = 1, \ldots, m.$$
and
\[ w_{\tau_{j-1}}^i = w_{\tau_{j}}^i, \quad j \geq 1. \quad (5.5.4) \]

In particular, the processes \( w_t^i, i = 1, \ldots, m \), are well defined and take values in \([0, 1)\).

**Proof.** The definition of the portfolio \( w^\varepsilon \) yields (5.5.4) and, together with Itô’s formula and the dynamics of the risky assets \( S^i \) and the wealth process \( V^\varepsilon \) (cf. Definition 5.2.4), the dynamics (5.5.3). Since both the drift and diffusion coefficients in (5.5.3) are locally Lipschitz continuous, for the given initial value \( w_{\tau_{j-1}}^* \) there exists a unique local solution \((w_t^\varepsilon)_{\tau_{j-1} \leq t \leq \tau_j}\) up to an explosion time \( \tau \). By definition, \( w_{\tau_{j-1}}^\varepsilon = w_{\tau_{j}}^* \in [0, 1) \) for \( i = 0, \ldots, m \) and \( j \geq 1 \). Since the price process \( S \) is continuous, Definition 5.2.4 implies \( w_t^\varepsilon \in [0, 1) \) for \( i \in \{0, \ldots, m\}, t \in [\tau_{j-1}, \tau_j) \), and \( j \geq 1 \). In summary, the process \( w^\varepsilon \) therefore remains \([0, 1)\)-valued on \([0, T]\). □

### 5.5.2 Proof of Proposition 5.3.1

**Proof of Proposition 5.3.1. Step 1: Estimation of the Trade Sizes.** Define the rebalancing times \((\tau_j)_{j \in \mathbb{N}}\) as in Definition 5.2.3. Then, by Definition 5.2.4, the wealth process \( V^\varepsilon \) of the corresponding strategy \( w^\varepsilon \) satisfies
\[
\frac{dV_t^\varepsilon}{V_t^\varepsilon} = \sum_{i=1}^{m} w_t^i \mu_t^i dt + \sigma_t^i dB_t
\]
on \((\tau_{j-1}, \tau_j)\), and at time \( t = \tau_j \) the risky weights \((w_{\tau_{j-1}}^1, \ldots, w_{\tau_{j}}^m)\) are rebalanced back to the frictionless targets \((w_{\tau_{j-1}}^*, \ldots, w_{\tau_{j}}^*)\). Whence, for each \( i = 1, \ldots, m \) the respective dollar amounts \( V_{\tau_{j}}^\varepsilon - \Delta L_{\tau_j}^i \) transferred satisfy the following rebalancing condition:
\[
V_{\tau_{j}}^\varepsilon (w_{\tau_{j-1}}^i + \Delta L_{\tau_j}^i) = w_{\tau_{j}}^* V_{\tau_j}^\varepsilon = w_{\tau_{j}}^* V_{\tau_{j}}^\varepsilon \left( 1 - \varepsilon \sum_{k=1}^{m} |\Delta L_{\tau_j}^k| \right).
\]

Put differently, the changes \( \Delta L_{\tau_j}^i \) in the risky weights are given by
\[
\Delta L_{\tau_j}^i + \varepsilon w_{\tau_{j}}^* \sum_{k=1}^{m} |\Delta L_{\tau_j}^k| = w_{\tau_{j}}^* - w_{\tau_{j-1}}^*.
\]

As a consequence, the implicit function theorem yields that, for small \( \varepsilon > 0 \):
\[
\Delta L_{\tau_j}^i = w_{\tau_{j}}^* - w_{\tau_{j-1}}^* + P_{TAC}^i(\tau_j), \quad (5.5.5)
\]
where the remainder term \( P_{TAC}^i(\tau_j) \) satisfies under Condition (5.2.2) (no shortselling)
\[
|P_{TAC}^i(\tau_j)| \leq C \varepsilon \|w_{\tau_{j}}^* - w_{\tau_{j-1}}^*\|_{\mathbb{R}^m}. \quad (5.5.6)
\]
In addition,
\[
\frac{V_{\tau_j} - V_{\tau_j^-}}{V_{\tau_j^-}} = -\varepsilon \sum_{k=1}^{m} |\Delta L_{kj}^i|.
\]
(5.5.7)

Putting everything together we obtain
\[
\int_0^T \frac{dV_t^\varepsilon}{V_t^\varepsilon} = \int_0^T w_t^\varepsilon (\mu_t dt + \sigma_t dB_t) - \varepsilon \sum_{i=1}^{m} \sum_{j=1}^{N} \left[ w_{\tau_j}^{*,i} - w_{\tau_j^-}^{*,i} + P_{TAC}^i(\tau_j) \right].
\]
(5.5.8)

**Step 2: Estimation of the Quadratic Variation of the Jumps.** Transaction costs are only paid at the trading times \(\tau_j\). Hence, the cumulative amount of transaction costs paid up to terminal time is a pure jump process and its quadratic variation is negligible at the leading order. Indeed, (5.5.7) gives
\[
E \left[ \sum_{j=1}^{N} \left( \frac{V_{\tau_j} - V_{\tau_j^-}}{V_{\tau_j^-}} \right)^2 \right] = \varepsilon^2 E \left[ \sum_{j=1}^{N} \left( \sum_{i=1}^{m} |\Delta L_{i\tau_j}^i| \right)^2 \right].
\]

By (5.5.5), (5.5.6), and Assumption 5.2.2 (no shortselling), the changes \(\Delta L_{i\tau_j}^i\) of the risky weights are bounded. Therefore, \(E[N] = O(\varepsilon^{-\alpha})\) (cf. (5.2.4)) implies that
\[
E \left[ \sum_{j=1}^{m} \left( \frac{V_{\tau_j} - V_{\tau_j^-}}{V_{\tau_j^-}} \right)^2 \right] = O(\varepsilon^{2-\alpha}).
\]

In addition, on \((\tau_{j-1}, \tau_j)\) we have \(d[V_t^\varepsilon]/(V_t^\varepsilon)^2 = (w_t^\varepsilon)^\top \Sigma_t w_t^\varepsilon dt\). Thus,
\[
E \left[ \left| \int_0^T \frac{d[V_t^\varepsilon]}{(V_t^\varepsilon)^2} - \int_0^T (w_t^\varepsilon)^\top \Sigma_t w_t^\varepsilon dt \right| \right] = O(\varepsilon^{2-\alpha}).
\]
(5.5.9)

**Step 3: Expansion of the local mean-variance criterion \(F^\varepsilon\).** Inserting (5.5.8) and (5.5.9) into the definition of \(F^\varepsilon\) (cf. (5.2.6)), we see that the objective function simplifies to
\[
F^\varepsilon(A) = \frac{1}{T} E \left[ \int_0^T ((w_t^\varepsilon)^\top \mu_t - \frac{\gamma}{2} (w_t^\varepsilon)^\top \Sigma_t w_t^\varepsilon) dt \right] - \varepsilon \sum_{i=1}^{m} \sum_{j=1}^{N} \left[ w_{\tau_j}^{*,i} - w_{\tau_j^-}^{*,i} + P_{TAC}^i(\tau_j) \right] + O(\varepsilon^{2-\alpha}).
\]
Furthermore, the definition of the Merton proportion \( \mathbf{w}_t^* = \Sigma_t^{-1} \mathbf{u}_t / \gamma \) yields

\[
F^\varepsilon(A) = \frac{1}{T} \mathbb{E} \left[ \int_0^T ((\mathbf{w}_t^*)^\top \mathbf{u}_t - \frac{\gamma}{2} (\mathbf{w}_t^*)^\top \Sigma_t \mathbf{w}_t^*) dt \right] - \frac{\gamma}{2T} \mathbb{E} \left[ \int_0^T (\mathbf{w}_t^* - \mathbf{w}_t^\varepsilon)^\top \Sigma_t (\mathbf{w}_t^* - \mathbf{w}_t^\varepsilon) dt \right] + \frac{1}{T} \mathbb{E} \left[ \int_0^T (\mathbf{w}_t^* - \mathbf{w}_t^\varepsilon)^\top (\mathbf{u}_t - \gamma \Sigma_t \mathbf{w}_t^\varepsilon) dt \right] - \frac{\varepsilon}{T} \mathbb{E} \left[ \sum_{i=1}^m \sum_{j=1}^N |w_{\tau_j,i}^\varepsilon - w_{\tau_j,i}^\varepsilon + P_{\text{TAC}}(\tau_j)| \right] + O(\varepsilon^{2-\alpha})
\]

\[
= \frac{1}{T} \mathbb{E} \left[ \int_0^T (\mathbf{w}_t^*)^\top \mathbf{u}_t - \frac{\gamma}{2} (\mathbf{w}_t^*)^\top \Sigma_t \mathbf{w}_t^* \right] - \frac{1}{T} \mathbb{E} \left[ \text{TAC}(A) + \frac{\gamma}{2} \text{DE}(A) \right] + O(\varepsilon^{2-\alpha})
\]

\[
= \frac{1}{T} \mathbb{E} \left[ \int_0^T \mu_t^\top \Sigma_t^{-1} \mu_t dt \right] - \frac{1}{T} \mathbb{E} \left[ \text{TAC}(A) + \frac{\gamma}{2} \text{DE}(A) \right] + O(\varepsilon^{2-\alpha}).
\]

This completes the proof of Proposition 5.3.1. \qed

### 5.5.3 Proof of Lemma 5.3.2

To establish Lemma 5.3.2, we first derive the following useful asymptotic result:

**Lemma 5.5.4.** Define the rebalancing times \((\tau_j)_{j=0,1,...}\) as in Definition 5.2.3 and let \((X^\varepsilon)_{\varepsilon>0}\) be a family of right-continuous, \(m\)-dimensional processes such that:

(i) The process \(X^\varepsilon\) satisfies \(dX_t^\varepsilon = \Theta_t^\varepsilon dt + \Gamma_t^\varepsilon dB_t\) for \(t \in (\tau_{j-1}, \tau_j)\).

(ii) \(X_t^\varepsilon = 0\) for all \(j\).

(iii) There exists a constant \(C > 0\) such that

\[
\sup_{t \in [0,T]} (\|\Theta_t^\varepsilon\|_{\mathbb{R}^m} + \|\Gamma_t^\varepsilon\|_{2,1}) \leq C, \quad \lim_{\varepsilon \downarrow 0} \mathbb{E} \left[ \sup_{t \in [0,T]} \|\Gamma_t^\varepsilon\|_{2,1}^2 \right] = 0.
\]

Then:

\[
\mathbb{E} \left[ \sum_{j=1}^N \|X_{\tau_j}^\varepsilon - X_{\tau_{j-1}}^\varepsilon\|_{\mathbb{R}^m} \right] = o(\varepsilon^{-\alpha/2}).
\]

Proof. The Burkholder-Davis-Gundy inequality yields
\[
E \left[ \sum_{j=1}^{N} \left\| X_{\tau_{j}}^{\varepsilon} - X_{\tau_{j-1}}^{\varepsilon} \right\|_{\mathbb{R}^m} \right] = E \left[ \sum_{j=1}^{N} \left\| \int_{\tau_{j-1}}^{\tau_{j}} \Theta_{t}^{\varepsilon} dt + \Gamma_{t}^{\varepsilon} dB_{t} \right\|_{\mathbb{R}^m} \right] \\
\leq E \left[ \sum_{j=1}^{N} \int_{\tau_{j-1}}^{\tau_{j}} \left\| \Theta_{t}^{\varepsilon} \right\|_{\mathbb{R}^m} dt \right] + E \left[ \sum_{j=1}^{N} \left\| \Gamma_{t}^{\varepsilon} dB_{t} \right\|_{\mathbb{R}^m} \right] \\
\leq CE \left[ \sum_{j=1}^{N} (\tau_{j} - \tau_{j-1}) \right] \\
+ CE \left[ \sup_{t \in [0,T]} \left\| \Gamma_{t}^{\varepsilon} \right\|_{2,1} \sum_{j=1}^{N} (\tau_{j} - \tau_{j-1})^{1/2} \right] \\
\leq C + C\varepsilon^{-\alpha/2} E \left[ \sup_{t \in [0,T]} \left\| \Gamma_{t}^{\varepsilon} \right\|_{2,1} \int_{0}^{T} A_{t}^{-1/2} dt \right] \\
+ C\varepsilon^{-\alpha/2} E \left[ \sup_{t \in [0,T]} \left\| \Gamma_{t}^{\varepsilon} \right\|_{2,1} \right. \\
\times \left( \sum_{j=1}^{N} \frac{\tau_{j} - \tau_{j-1}}{A_{\tau_{j-1}}^{1/2}} - \int_{0}^{T} A_{t}^{-1/2} dt \right) \right].
\]

Here, we have used in the last step that \((\tau_{j} - \tau_{j-1})^{1/2} = \frac{\tau_{j} - \tau_{j-1}}{\varepsilon^{\alpha/2} A_{\tau_{j-1}}^{1/2}}\), and passed to the integral in the limit, taking into account the continuity and integrability of \(A\). In view of the integrability assumptions on \(\Gamma^{\varepsilon}\) and \(A\), Hölder’s inequality yields
\[
E \left[ \sum_{j=1}^{N} \left\| X_{\tau_{j}}^{\varepsilon} - X_{\tau_{j-1}}^{\varepsilon} \right\|_{\mathbb{R}^m} \right] = o(\varepsilon^{-\alpha/2}),
\]
as claimed.

By combining Lemma 5.5.4 with elementary estimates for the normal distribution, we can now establish Lemma 5.3.2:

Proof of Lemma 5.3.2. The finiteness of the expectations follows from the integrability conditions on \(A\) and the boundedness of \(\Sigma\) and \(\beta\).

Step 1: Expansion of the Transaction Cost Loss. We now perform a more detailed analysis of the leading order term \(w_{\tau_{j}}^{*} - w_{\tau_{j}}^{\varepsilon, i}\) in (5.5.5). Using the dynamics of \(w^{*}\) and \(w^{\varepsilon}\) in (5.5.1) and (5.5.3), respectively, we find
\[
w_{\tau_{j}}^{*} - w_{\tau_{j}}^{\varepsilon, i} = \int_{\tau_{j-1}}^{\tau_{j}} \left( \hat{\mu}_{t}^{i} - \mu_{t}^{\varepsilon, i} \right) dt + \int_{\tau_{j-1}}^{\tau_{j}} \left[ \sigma_{t}^{i} - w_{t}^{\varepsilon, i} \left( \sigma_{t}^{i} - \sum_{k=1}^{m} w_{t}^{\varepsilon, k} \sigma_{t}^{k} \right) \right] dB_{t}.
\] (5.5.10)
Adding and subtracting \(\int_{\tau_{j-1}}^{\tau_{j}} \beta_{\tau_{j}}^{i} dB_{t}\) in (5.5.10) yields
\[
w_{\tau_{j}}^{*} - w_{\tau_{j}}^{\varepsilon, i} = \beta_{\tau_{j}}^{i} (B_{\tau_{j}} - B_{\tau_{j-1}}) + R_{TAC}(\tau_{j}),
\]

for some process \(R_{TAC}(\tau_{j})\). For \(\beta_{\tau_{j}}^{i}\), we can apply Lemma 5.3.2 to obtain the desired bounds.
where \( R_i^{TAC}(\tau_j) \) is defined as follows:

\[
R_i^{TAC}(\tau_j) := \int_{\tau_{j-1}}^{\tau_j} J_t^{FV,i} dt + \int_{\tau_{j-1}}^{\tau_j} J_t^{S,i} dB_t, \tag{5.5.11}
\]

with

\[
J_t^{FV,i} = \tilde{\mu}_t - w^\varepsilon,t \mu_t^w,
\]

\[
J_t^{S,i} = \tilde{\sigma}_t - w^\varepsilon,t \left( \sigma_t^i - \sum_{k=1}^m w^\varepsilon,k \sigma_t^k \right) - \beta_{\tau_j-1}, \quad \text{for } t \in (\tau_{j-1}, \tau_j).
\]

In vector or matrix notation we have \( J^{FV} = (J^{FV,1}, \ldots, J^{FV,m})^\top \in \mathbb{R}^m \) and \( J^{S} = (J^{S,1}, \ldots, J^{S,m})^\top \in \mathcal{M}_{m \times d}(\mathbb{R}) \), respectively. Lemma 5.5.3 and Lemma 5.5.1 imply that \( J^{FV} \) is bounded. Using that the mapping

\[
w \mapsto \left[ \tilde{\sigma}_t - w^i \left( \sigma_t^i - \sum_{k=1}^m w^k \sigma_t^k \right) \right]
\]

is locally Lipschitz continuous with Lipschitz constant \( L \) depending on \( m \) and \( K_\sigma \), we obtain that

\[
\|J_t^{S,i}\|_{\mathbb{R}^d} \leq L \|w^i - w^i_*\|_{\mathbb{R}^m} \quad \text{and} \quad |J_t^{FV,i}| \leq C \quad \text{on } (\tau_{j-1}, \tau_j),
\]

where \( C \) denotes a real constant. In view of (5.5.5) we also have

\[
\varepsilon \mathbb{E} \left[ \sum_{i=1}^m \sum_{j=1}^N |\Delta L_{\tau_j}^i| \right] = \varepsilon \mathbb{E} \left[ \sum_{i=1}^m \sum_{j=1}^N |\beta_{\tau_j-1}^i (B_{\tau_j} - B_{\tau_j-1}) + R_i^{TAC}(\tau_j) + P_i^{TAC}(\tau_j)| \right],
\]

where \( R^i \) and \( P^i \) are defined as in (5.5.11) and (5.5.5), respectively.

In order to prove the expansion (5.3.7) it suffices to show that:

\[
\varepsilon \mathbb{E} \left[ \sum_{i=1}^m \sum_{j=1}^N |\beta_{\tau_j-1}^i (B_{\tau_j} - B_{\tau_j-1})| \right] = \varepsilon^{1-\alpha/2} \mathbb{E} \left[ \sqrt{2 \pi} \int_0^T \frac{\|\beta_t\|^2}{\sqrt{A_t}} dt \right] + o(\varepsilon^{1-\alpha/2}), \tag{5.5.12}
\]

\[
\varepsilon \mathbb{E} \left[ \sum_{i=1}^m \sum_{j=1}^N |R_i^{TAC}(\tau_j)| + |P_i^{TAC}(\tau_j)| \right] = o(\varepsilon^{1-\alpha/2}). \tag{5.5.13}
\]

For (5.5.12) we use the independent increments property of Brownian motion and
the scaling property of the normal distribution to obtain
\[
\varepsilon E\left[\sum_{i=1}^{m} \sum_{j=1}^{N} |\beta_{t_{j-1}}^i (B_{t_j} - B_{t_{j-1}})| \right] \\
= \varepsilon E\left[ \sum_{j=1}^{\infty} \sum_{i=1}^{m} 1_{\{t_j < T\}} |\beta_{t_{j-1}}^i (B_{t_j} - B_{t_{j-1}})| \right] \\
= \varepsilon E\left[ \sum_{j=1}^{\infty} 1_{\{t_j < T\}} \sum_{i=1}^{m} \sum_{0 < t_j < T} \frac{2}{\pi} \| \beta_{t_{j-1}}^i \|_{\mathbb{R}^d} |\sqrt{t_j - t_{j-1}}| \right] \\
= \varepsilon E\left[ \sum_{0 < t_j < T} \sum_{i=1}^{m} \sqrt{\frac{2}{\pi}} \| \beta_{t_{j-1}}^i \|_{\mathbb{R}^d} \sqrt{e^\alpha A_{t_{j-1}}} \right].
\]

Here, \( Z \) denotes an independent univariate standard normal random variable, for which \( E[|Z|] = \sqrt{2/\pi} \). To approximate the random sum over \( t_j \) we rewrite the expression inside the sum:
\[
E[TAC(A)] = \varepsilon E \left[ \sum_{0 < t_j < T} \sum_{i=1}^{m} \sqrt{\frac{2}{\pi}} \| \beta_{t_{j-1}}^i \|_{\mathbb{R}^d} \frac{e^\alpha A_{t_{j-1}}}{\sqrt{e^\alpha A_{t_{j-1}}}} \right] \\
= E \left[ \sum_{0 < t_j < T} \sum_{i=1}^{m} \sqrt{\frac{2}{\pi}} e^{1-\alpha/2} \| \beta_{t_{j-1}}^i \|_{\mathbb{R}^d} \frac{A_{t_{j-1}}}{\sqrt{A_{t_{j-1}}}} \right].
\]

Since \( \text{sup}_{\varepsilon \in \mathbb{E}} \| \beta(y) \|_{2,1} \) is bounded by Corollary 5.5.2, the lower bound on \( A \) from Definition 5.2.3 and the estimate (5.2.5) imply
\[
E[TAC(A)] = E \left[ \sqrt{\frac{2}{\pi}} e^{1-\alpha/2} \int_0^T \| \beta_t \|_{2,1} dt \right] \\
+ \sqrt{\frac{2}{\pi}} e^{1-\alpha/2} \left( \sum_{0 < t_j < T} \frac{\| \beta_{t_{j-1}} \|_{2,1}}{\sqrt{A_{t_{j-1}}}} (t_j - t_{j-1}) - \int_0^T \frac{\| \beta_t \|_{2,1}}{\sqrt{A_t}} dt \right) \\
= E \left[ \sqrt{\frac{2}{\pi}} e^{1-\alpha/2} \int_0^T \frac{\| \beta_t \|_{2,1}}{\sqrt{A_t}} dt \right] + O(e^{1-\alpha/2}). \quad (5.5.14)
\]

Here, dominated convergence is applicable in the last step due to the boundedness of \( \beta \) and the integrability of \( A \).

We now turn to (5.5.13) and consider the first term in the corresponding expectation. To estimate it, we apply Lemma 5.5.4 with
\[
X_t^\varepsilon := \int_{t_{j-1}}^{t} J_{t}^{FV} ds + \int_{t_{j-1}}^{t} J_{t}^{S} dB_s \quad \text{on} \quad [t_{j-1}, t_j].
\]
In view of the uniform boundedness of $J^S$ and $J^{FV}$ and because
\[
E \left[ \sup_{s \in [0,T]} \| J^S_s \|_{1,1}^2 \right] \leq 2^{m-1} L^2 E \left[ \sup_{s \in [0,T]} \| \hat{w}_s^\varepsilon - w_s^* \|_{R^m}^2 \right],
\]
it then remains to show $E[\sup_{s \in [0,T]} \| w_s^\varepsilon - w_s^* \|_{R^m}^2] \to 0$ as $\varepsilon \downarrow 0$.

To this end, recall that
\[
\tilde{X}_t^\varepsilon := w_t^\varepsilon - w_t^* = \int_{\tau_{j-1}}^t J^{FV}_s \, ds + \int_{\tau_{j-1}}^t (J^S_s + \beta_{t_{j-1}}) \, dB_s, \quad \text{for } t \in [\tau_{j-1}, \tau_j).
\]

For all $\omega \in \Omega$, denote by $\tilde{K}^\varepsilon(\omega)$ the $1/16$--Hölder constant of this process on $[0,T]$. Using the Burkholder-Davis-Gundy inequality and the uniform boundedness of $J^{FV}$ and $J^S$, we see that
\[
E \left[ \| \tilde{X}_t^\varepsilon - \tilde{X}_{s}^\varepsilon \|_{R^m} \right] \leq C(t - s)^2. \tag{5.5.15}
\]

Kolmogorov’s Hölder continuity criterion in turn shows that $E \left[ (\tilde{K}^\varepsilon)^4 \right] < C$ for a constant $C$ that does not depend on $\varepsilon$. To prove $E \left[ \sup_{s \in [0,T]} \| w_s^\varepsilon - w_s^* \|_{R^m}^2 \right] \to 0$ as $\varepsilon \to 0$, we fix an arbitrary small $r > 0$ and find a sufficiently large $M_r > 0$ such that the set $\Omega_r := \{ \sup_s A_s^{1/8} \leq M_r \}$ satisfies
\[
\mathbb{P}(\Omega_r) \geq 1 - \frac{r}{2m}.
\]

Then, since for each $i = 1, \ldots, m$ the processes $w^{*,i}$ and $w^{\varepsilon,i}$ are $[0,1)$-valued, using the definition of the trading times we obtain
\[
E \left[ \sup_{s \in [0,T]} \| w_s^\varepsilon - w_s^* \|_{R^m}^2 \right] = E \left[ \sup_{s \in [0,T]} \| w_s^\varepsilon - w_s^* \|_{R^m}^2 \mathbf{1}_{\Omega_r} \right] + E \left[ \sup_{s \in [0,T]} \| w_s^\varepsilon - w_s^* \|_{R^m}^2 \mathbf{1}_{\in \Omega_r} \right] \leq E \left[ (\tilde{K}^\varepsilon)^2 \right] \varepsilon^{\alpha/8} \sup_{s \in [0,T]} A_s^{1/8} \mathbf{1}_{\Omega_r} + E \left[ \sup_{s \in [0,T]} \| w_s^\varepsilon - w_s^* \|_{R^m}^2 \mathbf{1}_{\in \Omega_r} \right] \leq M_r \varepsilon^{\alpha/8} E[(\tilde{K}^\varepsilon)^2] + r/2.
\]

Due to the uniform bound on $E[(\tilde{K}^\varepsilon)^4]$, we can pick $\varepsilon > 0$ small enough to ensure $M_r \varepsilon^{\alpha/8} E[(\tilde{K}^\varepsilon)^2] \leq r/2$ and in turn
\[
E \left[ \sup_{s \in [0,T]} \| w_s^\varepsilon - w_s^* \|_{R^m}^2 \right] \leq r.
\]
As \( r \) was arbitrary, this shows that Lemma 5.5.4 is applicable for \( X^\varepsilon \) and yields

\[
\varepsilon E \left[ \sum_{i=1}^{m} \sum_{j=1}^{N} |R_{TAC}^i(\tau_j)| \right] = o(\varepsilon^{1-\alpha/2}).
\]

Therefore, the first term in (5.5.13) is indeed of the claimed asymptotic order. As for the second, notice that for \( \xi, \eta \in \mathbb{R} \) the triangle inequalities \( |\xi| - |\eta| \leq |\xi + \eta| \leq |\xi| + |\eta| \) and (5.5.12) show that

\[
\lim_{\varepsilon \to 0} \varepsilon E \left[ \sum_{i=1}^{m} \sum_{j=1}^{N} |w_{ij} - u_{ij} - \beta_{ij} (B_{ij} - B_{ij-1})| \right] = 1.
\]

Moreover, the inequality (5.5.6) gives

\[
\varepsilon E \left[ \sum_{i=1}^{m} \sum_{j=1}^{N} |\beta_{ij} (B_{ij} - B_{ij-1}) + R_{TAC}^i(\tau_j)| \right] \leq C m \varepsilon^2 E \left[ \sum_{i=1}^{m} \sum_{j=1}^{N} |w_{ij} - u_{ij}| \right].
\]

Therefore,

\[
\lim_{\varepsilon \to 0} \varepsilon E \left[ \sum_{i=1}^{m} \sum_{j=1}^{N} |P_{TAC}^i(\tau_j)| \right] \leq C m \lim_{\varepsilon \to 0} \varepsilon E \left[ \sum_{i=1}^{m} \sum_{j=1}^{N} |w_{ij} - u_{ij}| \right] = 0,
\]

where we have used (5.5.16) and (5.5.12) in the last step. This completes the proof of (5.5.13) and in turn (5.3.7).

**Step 2: Expansion of the Discretization Error.** For the discretization error we proceed similarly. For \( j \in \mathbb{N} \) and \( r \in [\tau_{j-1}, \tau_j) \) recall that

\[
w_{ij}^r - w_{ij}^\varepsilon = \int_{\tau_{j-1}}^{r} (\bar{\mu}_t - \bar{\mu}_t^\varepsilon \mu_t^w) dt + \left[ \bar{\sigma}_t - \sigma_t \left( \sum_{k=1}^{m} w_{ijk}^\varepsilon \sigma_t^k \right) \right] dB_t.
\]

Arguing similarly as in Step 1, we can replace \( w_{ij}^r - w_{ij}^\varepsilon \) and \( \Sigma_t \) at the leading order with \( \int_{\tau_{j-1}}^{r} \beta_{ij} dB_t \in \mathbb{R}^m \) and \( \Sigma_{\tau_{j-1}} \), respectively. Therefore, the discretization error can be rewritten as

\[
E[DE(A)] = E \left[ \int_{0}^{T} (w_{ij}^r - w_{ij}^\varepsilon) \Sigma_t (w_{ij}^r - w_{ij}^\varepsilon) dt \right]
\]

\[
= E \left[ \sum_{0 < \tau_t < T} \int_{\tau_{j-1}}^{\tau_j} \left( \int_{\tau_{j-1}}^{t} \beta_{ij-1} dB_t \right) \Sigma_{\tau_{j-1}} \left( \int_{\tau_{j-1}}^{t} \beta_{ij-1} dB_t \right) dt \right] + o(\varepsilon^\alpha).
\]
Conditioning on $\mathcal{F}_{\tau_{j-1}}$ and integrating over $t$ in turn yields

$$
E[DE(A)] = E \left[ \sum_{0 < \tau_j < T} \int_{\tau_{j-1}}^{\tau_j} \text{tr} \left( \beta_{\tau_{j-1}}^T \Sigma_{\tau_{j-1}} \beta_{\tau_{j-1}} \right) (t - \tau_{j-1}) dt \right] + o(\varepsilon^\alpha)
$$

$$
= \frac{1}{2} E \left[ \sum_{0 < \tau_j < T} \text{tr} \left( \beta_{\tau_{j-1}}^T \Sigma_{\tau_{j-1}} \beta_{\tau_{j-1}} \right) (\tau_j - \tau_{j-1})^2 \right] + o(\varepsilon^\alpha)
$$

$$
= \frac{\varepsilon^\alpha}{2} E \left[ \sum_{0 < \tau_j < T} \text{tr} \left( \beta_{\tau_{j-1}}^T \Sigma_{\tau_{j-1}} \beta_{\tau_{j-1}} \right) A_{\tau_{j-1}} (\tau_j - \tau_{j-1}) \right] + o(\varepsilon^\alpha).
$$

Since the processes $\Sigma$ and $\beta$ are assumed to be bounded and continuous and $A$ satisfies (5.2.3), applying the same dominated convergence argument as before implies that the sum can again be approximated by the corresponding integral:

$$
E[DE(A)] = \frac{\varepsilon^\alpha}{2} E \left[ \int_0^T \text{tr} \left( \beta_t^T \Sigma_t \beta_t \right) A_t dt \right] + o(\varepsilon^\alpha).
$$

This completes the proof of (5.3.8) and in turn Lemma 5.3.2.

5.5.4 Proof of Theorem 5.3.4

Proof of Theorem 5.3.4. To minimize the total loss, we have to match the two leading orders in the asymptotic formulas (5.3.7-5.3.8) for TAC and DE:

$$
1 - \frac{\alpha}{2} = \alpha \Leftrightarrow \alpha = \frac{2}{3}.
$$

With this choice of $\alpha$, the asymptotic expansion of $F^\varepsilon$ simplifies to

$$
F^\varepsilon(A) = \frac{1}{T} E \left[ \int_0^T \frac{\mu_t^T \Sigma_t^{-1} \mu_t}{2\gamma} dt \right]
$$

$$
- \frac{\varepsilon^{2/3}}{T} E \left[ \int_0^T \frac{\gamma}{4} \text{tr} (\beta_t^T \Sigma_t \beta_t) A_t dt + \sqrt{\frac{2}{\pi}} \int_0^T \frac{\|\beta_t\|_{2,1}}{\sqrt{A_t}} dt \right] + o(\varepsilon^{2/3}).
$$

To obtain the optimal discretization rule $A$, it therefore remains to solve the following optimization problem:

$$
\min_{(A_t)_{t \in [0, T]}} \varepsilon^{2/3} E \left[ \int_0^T \frac{\gamma}{4} \text{tr} (\beta_t^T \Sigma_t \beta_t) A_t dt + \sqrt{\frac{2}{\pi}} \frac{\|\beta_t\|_{2,1}}{\sqrt{A_t}} dt \right].
$$

(5.5.17)

Pointwise minimization of the integrand yields the optimizer from (5.3.10):

$$
A_t^* = \left( \frac{\sqrt{\frac{2}{\pi}} \|\beta_t\|_{2,1}}{\frac{\gamma}{4} \text{tr} (\beta_t^T \Sigma_t \beta_t)} \right)^{2/3}.
$$
Plugging $A^*$ back into the cost functional (5.5.17) in turn yields that the corresponding minimal leading-order total cost is given by the formula reported in (5.3.11).

To complete the proof, it remains to show that the process $A^*$ is admissible in the sense of Definition 5.2.3. To this end, notice that the sub-multiplicativity of the Frobenius norm yields

$$\text{tr} (\beta_t^T \Sigma_t \beta_t) = \text{tr} (\beta_t^T \sigma_t \sigma_t^T \beta_t) = \|\beta_t \sigma_t\|_F^2 \leq \|\beta_t\|_F^2 \|\sigma_t\|_F^2 \leq \|\beta_t\|_{2,1} \|\sigma_t\|_{2,1}^2.$$ 

In view of Corollary 5.5.2, it follows that

$$A_t^* \geq C \left( \frac{\|\beta_t\|_{2,1}}{\|\beta_t\|_{2,1} \|\sigma_t\|_{2,1}^2} \right)^{2/3} \geq C \left( \frac{1}{K_2^2 K_\beta} \right)^{2/3} > 0, \quad \text{for all } t \in [0, T].$$

Therefore, $E[(\inf_{t \in [0,T]} A_t^*)^{-1}] < \infty$. Moreover, the uniform ellipticity of the covariance matrix $\Sigma$ implies that

$$\text{tr} (\beta_t^T \Sigma_t \beta_t) \geq C \text{ tr} (\beta_t^T \beta_t) = C \|\beta_t\|_F^2 \geq C \|\beta_t\|_{2,1}^2.$$ 

As a result:

$$A_t^* \leq C \left( \frac{\|\beta_t\|_{2,1}}{\|\beta_t\|_{2,1}^2} \right)^{2/3} \leq \frac{C}{\|\beta_t\|_{2,1}^{2/3}},$$

so that $\int_0^T A_t^* dt$ has finite expectation by Assumption (5.3.6). In summary, the discretization rule $A^*$ is admissible and therefore indeed asymptotically optimal. $\square$
Appendix A

General Notation

\[ \mathbb{Z} \quad \{ \ldots, -2, -1, 0, 1, 2, \ldots \} \]
\[ \mathbb{R}, \mathbb{R}_{\geq 0}, \mathbb{R}_{> 0} \quad (-\infty, \infty), [0, \infty), (0, \infty) \]
\[ i\mathbb{R}_{> 0} \quad \text{the set of purely imaginary numbers with a strictly positive imaginary part, i.e., } \{i \cdot x | x \in \mathbb{R}_{> 0}\} \]
\[ \mathbb{R}^m, \mathbb{C}^m \quad \text{the } m\text{-dimensional Euclidean or unitary space} \]
\[ \bar{\xi} \quad \text{the complex conjugate of } \xi \in \mathbb{C} \]
\[ x^-, x^+ \quad \max\{-x, 0\}, \max\{x, 0\} \]
\[ |x| \quad \text{the absolute value of } x \in \mathbb{R} \]
\[ \mathcal{O}(\cdot), \mathcal{o}(\cdot) \quad \text{the Landau symbols} \]
\[ f(x^-), f(x^+) \quad \text{the left (or right) limit } \lim_{s\uparrow x} f(s) \]
\[ (or \lim_{s\downarrow x} f(s)) \text{ of the real function } f \text{ at } x \in \mathbb{R} \]
\[ \dot{\theta}, \theta' \quad \text{the time derivative (or space derivative) of the function } \theta : \mathbb{R}_{\geq 0} \mapsto \mathbb{R} \text{ (or } \theta : \mathbb{R} \mapsto \mathbb{R}) \]
\[ f_{x_i} \quad \text{the partial derivative of the function } f : \mathbb{R}^m \mapsto \mathbb{R} \]
\[ \text{with respect to } x_i \]
\[ \nabla f \quad \text{the gradient of the function } f : \mathbb{R}^m \mapsto \mathbb{R} \]
\[ \text{sgn}(\cdot) \quad \text{the sign function} \]
\[ a^{(n)} \quad \text{the Pochhammer symbol} \]
\[ _1F_1(a, b, \cdot) \quad \text{the Kummer function} \]
\[ \Gamma(\cdot) \quad \text{the Gamma function} \]
\[ U(k, m, \cdot) \quad \text{the Tricomi function} \]
\[ M(k, m, \cdot), M_{k, m}(\cdot) \quad \text{the Whittaker function of the first kind} \]
\[ W(k, m, \cdot), W_{k, m}(\cdot) \quad \text{the Whittaker function of the second kind} \]
\[ \|b\|_{\mathbb{R}^m} \quad \text{the Euclidean norm of the vector } b = (b^1, \ldots, b^m) \]
\[ bc \quad \text{the inner product of } b \in \mathbb{R}^m \text{ and } c \in \mathbb{R}^m \]
\[ \mathbb{R}^{m \times d} \quad \text{the space of } m \times d \text{ matrices with real entries} \]
\[ A^\top \quad \text{the transpose of the matrix } A \in \mathbb{R}^{m \times d} \]
\[ \|A\|_F, \|A\|_1 \quad \text{the Frobenius (or sum) norm of the matrix } A \in \mathbb{R}^{m \times d} \]
\[ \|A\|_{2,1} \quad \text{i.e., } \|A\|_{2,1} := \sum_{i=1}^m \sqrt{\sum_{j=1}^d |A_{ij}|^2} \]
\( \mathcal{M}_{m \times d}(\mathbb{R}) \) the space of \( m \times d \) matrices with real entries equipped with the \( L_{2,1} \) norm

\( A^i \) the \( i \)-th row of the matrix \( A \in \mathbb{R}^{m \times d} \)

\( \text{tr}(A) \) the trace of the matrix \( A \in \mathbb{R}^{m \times m} \)

\( A^{-1} \) the inverse of the matrix \( A \in \mathbb{R}^{m \times m} \)

\( C^j_b(\mathbb{R}^m) \) the class of \( \mathbb{R}^m \)-valued, bounded, \( j \) times differentiable functions with bounded derivatives up to order \( j \)

\( C^j_b(\mathbb{R}^{m \times d}) \) the class of \( \mathbb{R}^{m \times d} \)-valued, bounded, \( j \) times differentiable functions with bounded derivatives up to order \( j \)

\( \mathbf{1}_D \) the indicator function of the set \( D \)

\( \mathcal{F}, \mathcal{G} \) \( \sigma \)-fields

\( \mathbb{P}, \tilde{\mathbb{P}}, \hat{\mathbb{P}}, \hat{\mathbb{P}} \) probability measures

\( \mathbb{Q} \ll \mathbb{P} \) absolute continuity of \( \mathbb{Q} \) with respect to \( \mathbb{P} \)

\( \mathcal{F} = (\mathcal{F}_t)_{t \geq 0} \) filtration

\( (\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P}) \) filtered probability space

\( \mathbb{E}, \tilde{\mathbb{E}}, \hat{\mathbb{E}} \) expectation under \( \mathbb{P} \) (or \( \tilde{\mathbb{P}}, \hat{\mathbb{P}} \))

\( \mathbb{P} \text{- a.s.} \) almost surely with respect to \( \mathbb{P} \)

\( \mathbb{E}[X|\mathcal{F}] \) the conditional expectation of \( X \) given \( \mathcal{F} \) under \( \mathbb{P} \)

\( \frac{d\mathbb{Q}}{d\mathbb{P}} \) Radon-Nikodým derivative of \( \mathbb{Q} \ll \mathbb{P} \)

\( \mathbb{P}|_{\mathcal{G}} \) the restriction of the measure \( \mathbb{P} \) to the \( \sigma \)-field \( \mathcal{G} \)

\( X_{t-} \) the left limit \( \lim_{s \uparrow t} X_s \) of the semimartingale \( X \)

\( \Delta X_t \) the jump \( X_t - X_{t-} \) of the semimartingale \( X \)

\( \mathcal{E}(X) \) the stochastic exponential of the semimartingale \( X \)

\( \langle X, Y \rangle \) the predictable quadratic covariation of the semimartingales \( X \) and \( Y \)

\( [X, Y] \) the quadratic covariation of the semimartingales \( X \) and \( Y \)

\( \langle X \rangle \) \( \langle X, X \rangle \)

\( [X] \) \( [X, X] \)
Bibliography


Curriculum Vitae

Ren Liu
born February 03, 1987
Swiss citizen

Education

2012/09-present  PhD in Mathematics, ETH Zurich, Switzerland.
PhD Thesis: “Portfolio Selection with Frictions”
Supervisor: Prof. Dr. J. Muhle-Karbe

2010/09-2012/07  Master of Science in Quantitative Finance,
ETH Zurich & University of Zurich, Switzerland.
Master Thesis: “Portfolio Selection under Transaction Costs
and Leverage Constraints”
Supervisor: Prof. Dr. J. Muhle-Karbe

2007/09-2010/09  Bachelor of Mathematics, ETH Zurich, Switzerland.
Bachelor Thesis: “Duality Mapping and Geometry of Banach
Spaces”
Supervisor: Prof. Dr. M. Struwe

2003/08-2007/07  Matura, Alte Kantonsschule Aarau, Switzerland.

Academic Employment

2012/06-present  Scientific Researcher, ETH Zurich.

2009/09-2011/12  Supporting Assistant, ETH Zurich.