Limit order book models and optimal trading strategies

vorgelegt von
Marcel Höschler
Master of Science
aus Frechen

Von der Fakultät II – Mathematik und Naturwissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
– Dr. rer. nat. –

genehmigte Dissertation

Vorsitzender: Prof. Dr. Harry Yserentant
Berichter: Prof. Dr. Peter Bank
Prof. Dr. Bruno Bouchard

Tag der wissenschaftlichen Aussprache: 7.7.2011
Berlin 2011
D 83

Zusammenfassung

In den letzten Jahren haben fast alle großen Börsen elektronische Orderbücher eingeführt. Diese sammeln eingehende Limitorder und führen Marktorder automatisch zum bestmöglichen Preis aus. Durch die Einführung von Orderbüchern haben sich Handelsstrategien stark verändert; dies liegt zum einen an der sehr viel höheren Handelsgeschwindigkeit sowie an den verschiedenen Ordertypen, aus denen die Händler nun auswählen können. Es stellt sich daher die Frage, welcher Ordertyp unter welchen Umständen verwendet werden sollte, oder - allgemeiner - ob und wie optimale Handelsstrategien gefunden werden können. Während einige dieser Fragen in der Wirtschafts- und ökonometrischen Literatur betrachtet wurden, fehlt oft noch eine strenge mathematische Behandlung. In dieser Arbeit entwickeln wir geeignete mathematische Modelle und finden die angemessenen mathematischen Werkzeuge, um diese Fragen zu beantworten.

Im ersten Teil entwickeln wir ein mathematisches Modell für ein dynamisches, zeitstetiges Orderbuch. Innerhalb dieses Modells untersuchen wir, wie der aktuelle Zustand des Orderbuchs seine kurzfristige zukünftige Entwicklung bestimmt. Insbesondere analysieren wir die Verteilung des Ausführungszeitpunkts einer Limitorder. Da automatisierte Micro-Trader die Order innerhalb von Millisekunden platzieren müssen, bestimmen wir eine Näherungsformel für die Laplace-Transformation und die Momente des Ausführungszeitpunkts, die sehr effizient berechnet werden kann. Anschließend testen wir das Modell mit realen Hochfrequenz-Orderbuchdaten und zeigen, dass wichtige Eigenschaften sehr gut durch das Modell wiedergegeben werden.

Im zweiten Teil dieser Arbeit analysieren wir optimale Handelsstrategien in Orderbüchern. Zunächst bleiben im Rahmen des Modells aus Teil I und berechnen die optimale Handelsstrategie, wenn der Händler nur Marktorder benutzt. Danach berechnen wir optimale Handelsstrategien mit sowohl Markt- als auch Limitordern in einem vereinfachten Orderbuchmodell. Schließlich betrachten wir das Problem des Kaufs einer einzelnen Aktie. Der Händler platziert eine Limitorder zu Beginn der Handelsperiode. Nun gilt es, den optimalen Zeitpunkt zu finden, wenn die Limitorder in eine Marktorder umgewandelt werden sollte, falls sie noch nicht ausgeführt wurde. Wir zeigen, wie dieser Zeitpunkt vom *spread* abhängig ist, d.h. dem zusätzlichen Preis, den man bei der Umwandlung der Limit- zur Marktorder zahlen muss.

Summary

In the last few years, almost all major stock exchanges have introduced electronic limit order books, which collect incoming limit orders and automatically match market orders against the best available limit order. The introduction of limit order books has significantly changed trading strategies as the speed of trading increased dramatically and traders have the choice between different order types. This automatically raises the question which order type should be used under which circumstances, and more generally, if and how optimal trading strategies can be found. While some of these questions have been considered in the economic and econometric literature, a rigorous mathematical treatment of is often still lacking. In this thesis we develop suitable mathematical frameworks and find appropriate mathematical tools to address these questions.

In the first part, we propose a mathematical model for a dynamic, continuous time limit order book. Within this model, we study how the current state of the order book determines its short-time evolution. In particular, we analyse the distribution of the time-to-fill of a limit order. Since automated microtraders have to place orders within milliseconds, we also propose approximate formulae for the Laplace transform and the moments of the time-to-fill that can be computed very efficiently. Finally, we test the model with real-world high-frequency order book data and show that important properties are well reproduced by the model.

In the second part of this thesis, we analyse optimal trading strategies in limit order books. We first remain in the setting of the model of part I, and compute optimal liquidation strategies when the trader is restricted to use only market orders. Next, we compute optimal liquidation strategies with both market and limit orders in a simplified order book model. Finally, we turn to the problem of buying a single share. The trader places a limit order at the beginning of the trading period. The question is to find the optimal time when the limit order should be converted to a market order if it has not been filled yet. We show how this time depends on the spread, i.e. the additional price that is charged when converting the limit to a market order.

Contents

Chapter 1

Introduction

1.1 The limit order book

Stock market trading has changed dramatically in the last two decades. Next to globalization and deregulation of the financial market, technological innovation has been one of the main drivers of these changes: While traditionally a market maker collected buy and sell orders and provided liquidity by setting bid and ask quotes, nowadays most exchanges work with *order-driven* systems. These fully automated electronic trading platforms collect and match orders. *Electronic Communication Networks* (ECN) aggregate incoming limit orders at each price level. They constitute the overall liquidity and are made available to all market participants in the *limit order book* (LOB) by financial market data providers, see for example figure 1.1. Market orders are automatically executed against the best available limit orders. The automated electronic matching of orders significantly increases the speed of trading. It now often only takes a few milli-seconds from sending an order to its execution.

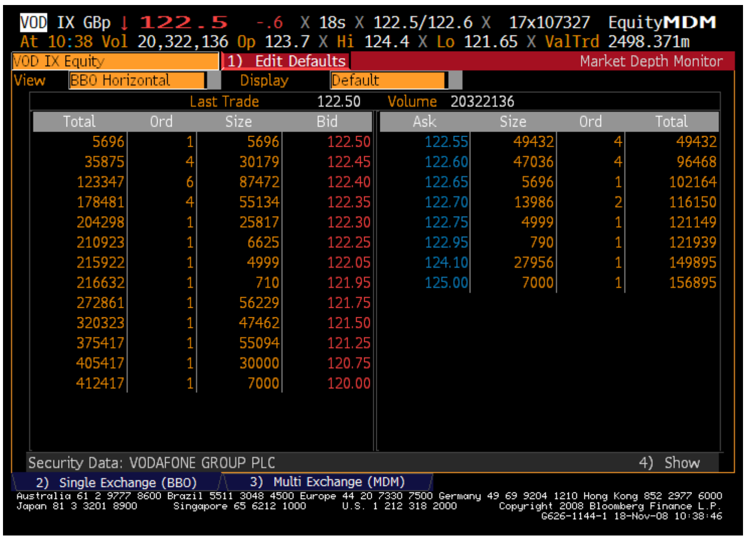


Figure 1.1: Bloomberg screenshot of Vodafone limit order book, with sizes of limit orders stored on different price ticks. The bid side (red) displays buy offers, and the ask side (blue) displays sell offers.

Most major exchanges such as Deutsche Börse, London Stock Exchange, Nasdaq, NYSE, Paris

Bourse and Tokyo Stock Exchange rely on either pure electronic limit order trading, or on hybrid trading systems combining order- and quote-driven markets. Electronic platforms have been playing a more and more important role over the last few years. For example, ECNs captured over 30 percent of Nasdaq trades within 3 years of existence, as was noted by Jain (2003), and more than 90 percent[1] of all share trading at German exchanges in 2010 was handled through Xetra, the electronic trading platform of Deutsche Börse. According to Jain (2003), over 80 percent of the world's stock exchanges have introduced some form of electronic trading mechanism with automatic execution.

Electronic trading platforms and automated trading have led to less execution costs (due to lower transaction costs and less need for human intervention) as well as higher execution efficiency (due to quicker reaction to incoming orders and market changes). However, decisions in the trading strategy have to be taken much faster, as the speed of trading significantly increased. Moreover, trading strategies are more complex, as traders typically have to choose between at least three types of operations[2] (for both buying and selling):

- *Limit orders* of a given size can be placed at any price tick, where they enter the queue of outstanding limit orders.
- *Cancellations* of (parts of) limit orders that are still in the queue can be sent.
- *Market orders* of a given size can be sent; they are immediately executed against the limit orders at the best available price tick, where limit orders that were placed earlier on the same price tick are considered first (first-in first-out policy).

Different risk categories are associated to these operations: firstly, risk associated to *execution time* (or *time-to-fill*), i.e. the time between sending the order and its execution, and secondly, risk associated to *execution price*. The main risk of a limit order arises from the time-to-fill, whereas its execution price is fixed by the price tick where the limit order was placed. A market order, on the contrary, is executed immediately, so its time-to-fill can be assumed to be zero. Its execution price, however, depends on the size of the market order and the liquidity provided. Moreover, large orders have an impact on the subsequent behaviour of the order book: market (buy) orders drive up prices and widen the spread (i.e. the gap between best limit buy and sell order). Large limit orders, which are visible to all market participants, change the supply and demand balance of the asset and therefore also have an impact on the price, even without being executed.

Traders are therefore facing a wide range of questions:

- How does the current state of the order book influence the short-time evolution of the order book?
- What is the distribution of the time-to-fill of a limit order, given current market conditions?
- What type of order should be used (market or limit)?

[1] Data taken from Xetra webpage on 09.12.2010

[2] We just list the three basic operations common to most ECNs. Special operations such as iceberg orders, or special trading platforms such as dark pools are deliberately not considered.

- Should an order be split into pieces and if so, how large should the pieces be, and when should they be placed?
- If a limit order is used, at which price tick should it be placed?
- If a limit order is *not* filled, when should it be cancelled and converted to a market order?

While some of these questions have been considered in the economic and econometric literature, a rigourous mathematical treatment is often still lacking. The main contribution of this thesis is to develop suitable mathematical frameworks and find appropriate mathematical tools to address these questions. We propose a new way of modelling order books on the basis of stochastic analysis. In particular, we include both sides of the order book and take account of the interaction of bid and ask sides. We show how to reduce the modelling dimensions significantly, and can thus identify the *key parameters* of the order book model. The dynamics of the order book are then easily expressed in terms of these key parameters. Our model can be used both on the *descriptive* level, e.g. to analyse midquote-price, spread and time-to-fill, and on the *normative* level, e.g. to compute optimal trading strategies.

Another focus of this work is applicability: the author strongly believes that applied mathematics should not only be practiced with regard to mathematical beauty, but great importance should be attached to its capability to be applied in the real world. In many occasions, we therefore

- show how model parameters can be estimated,
- perform this estimation with real market data, and
- compare the outcome of our theoretical findings to the results stemming from real market data.

The *existence* of suitable algorithms to perform these tasks is one issue. For limit order book trading, where decisions often have to be taken within milli-seconds, the issue of *speed* is equally important, and will be addressed in this work. Thus, another contribution is to partly bridge the inevitable gap between the world of mathematical models and real markets.

Before turning to the discussion of the mathematical results of this thesis, and their connection to the typical questions traders are facing, let us first outline the economic background.

1.2 Economic background

1.2.1 Empirical studies on limit order book

Numerous empiricial studies on order books have been carried out in the last few years, as more and more order book data was available. See for example Biais, Hillion, and Spatt (1995), Omura, Tanigawa, and Uno (2000), Maslov and Mills (2001), Zovko and Farmer (2002), Bouchaud, Mézard, and Potters (2002), Potters and Bouchaud (2003), Ranaldo (2004), Hall and Hautsch (2006), Cont, Kukanov, and Stoikov (2010) and Malo and Pennanen (2010).

The objective of these studies is mainly on the descriptive level. They investigate various statistical properties such as shape of the average order book, price impact, order flow, life time and execution probabilility of limit orders and order aggressiveness. In section 3.1, we will give an overview of some of these properties, and analyse if and how they are reproduced by our model.

1.2.2 Dynamic limit order book models

Various dynamic models of limit order books have recently been introduced. One line of studies focuses on equilibrium models, such as Parlour (1998), Bovier, Cerny, and Hryniv (2006), Foucault, Kadan, and Kandel (2005) and Rosu (2009). They analyse properties of the price formation, trade-off between market and limit order, and the form of the order book. Inevitably, all of these models depend on the agents' preferences which makes it difficult to incorporate these in a real trading engine.

Another line of research proposes stochastical models of the order book dynamics similar to the one we introduce in chapters 2 and 3. Luckock (2003), Osterrieder (2007), Cont, Stoikov, and Talreja (2010) and Smith, Farmer, Gillemot, and Krishnamurthy (2003) use queueing techniques to model the order book explicitly, whereas Malo and Pennanen (2010) model the order book using continuous stochastic differential equations (SDEs). Luckock computes the steady-state distribution of the order book, under the assumption that order arrival rates are independent of the order book state. Osterrieder assumes that the transaction price is given by a continuous stochastic process, independent of limit orders which are modeled by Poisson random measures, and stored in the order book. When the transaction price hits a price level where a limit order is stored, it is executed and removed from the book. Osterrieder analyses several properties such as the behaviour of the spread, and the time to execution/cancellation of a limit order. Cont, Stoikov, and Talreja model the arrival rate of limit orders conditional on their position in the book. They use Laplace transform techniques to compute the probabilities of various events such as execution of a limit order before the midquote price moves *conditional* on the current order book state. Malo and Pennanen use SDEs to model the midquote price and certain liquidity factors. They show how their model can be calibrated to market data and that it reproduces several empirically observed properties of order book dynamics such as liquidity mean-reversion.

1.2.3 Optimal trading in limit order books

Most studies focus on the problem of optimal portfolio liquidation. Early work such as Bertsimas and Lo (1998) and Huberman and Stanzl (2005) look at optimal trading strategies in a market impact model, however, but do not model the order book explicitly.

Almgren and Chriss (2001) look at the problem of portfolio liquidation with market orders. Their model captures both temporary and permanent price impact, but assumes that the temporary price impact vanishes immediately. Although not stated explicitly, their market impact model can be considered as a limit order book model. Using a mean-variance criterion, they can compute explicit optimal strategies.

Thanks to its ability to include different price impacts, the realistic structure of its optimal strategies and its mathematical tractability, this model has become very popular among practitioners and academics. It has been extended in many ways, see for example Almgren (2003), Almgren and Lorenz (2007), Almgren (2009), Schied and Schöneborn (2009a), Schied and Schöneborn (2009b), Schied, Schöneborn, and Tehranchi (2010) and Gatheral and Schied (2010).

Obizhaeva and Wang (2005) propose an order book model in which the temporary price impact does not vanish at once. Instead, they assume resilience, i.e. the price recovers gradually after a large market order. They also compute optimal liquidation strategies using market orders. This model was extended by Alfonsi, Fruth, and Schied (2008), Alfonsi, Schied, and Slynko (2009) Alfonsi, Fruth, and Schied (2010), Alfonsi and Schied (2010), Gatheral, Schied, and Slynko (2010b), Gatheral, Schied, and Slynko (2010a) and Predoiu, Shaikhet, and Shreve (2010).

1.3 Mathematical and economic results

1.3.1 Part I: Limit order book models

In part I, we set up our limit order book model and analyse various quantities such as the best bid, best ask and time-to-fill of a limit order. We then calibrate and test the model with real high-frequency order book data. Note that up to date, only few order book models exist and most models either focus on particular properties of the limit order book, or are too complex to be analysed and to formulate optimal control problems. One contribution of this work is to build an order book model which is mathematically tractable and reproduces most important properties observed empirically.

We start in chapter 2 by setting up a general framework for order book models. This framework will be the building stone for our studies in part I. We first formalize the bookkeeping of a static order book by introducing two equivalent order book representations. We then turn to the question how a dynamic order book evolves in time. Similar to Cont, Kukanov, and Stoikov (2010) (but slightly more general, because we do not restrict to the first level of the book), we look at the different components of the order flow (in particular market order, limit orders and cancellations) and their effect on the order book. Under the crucial modelling assumption 2.6 that limit orders are placed and cancelled proportionally to the existing limit order, the model equations greatly simplify to

$$\mathrm{d}\mathfrak{A}(t,p) = \mathrm{d}X(t) + \mathfrak{A}(t-,p)\mathrm{d}Y(t),$$

where $\mathfrak{A}(t,p)$ denotes the maximal quantity of shares a trader can buy at time t when she is willing to pay at most p for a single share, $\mathrm{d}X(t)$ denotes the cumulative order flow of market orders and limit orders placed inside the spread, and $\mathrm{d}Y(t)$ denotes the rate of limit order placements and cancellations deeper in the book. Note that our modelling framework also includes jumps. We then answer the question of existence and uniqueness of the order book. Moreover, we show how to reduce the number of dimensions of the order book from infinity (number of limit orders stored at any price level $p \in \mathbb{R}$) to four *key parameters* best bid, best ask, volume on best bid and volume on best ask. Empirical studies from Pascual Gascó and Veredas (2008) and Cao, Hansch, and Wang (2009) confirm this approach, as they show that most of the informational content of the orderbook is represented by these four key parameters.

This dimension reduction not only greatly reduces the model complexity, but also helps us in identifying the most important parameters of the model under minimal assumptions.

In chapter 3 we give an overview of theoretical and empricial results on the order flow. In particular we consider conditional order flow which depends on the current state of the order book. We then propose a system of continuous SDEs for the order flow of market and limit orders, which depends on the order book state represented by the four key parameters. In particular, we include the inter-dependence of *both* sides of the order book. We verify that the proposed dynamics reproduce the behaviour that was found in empirical studies.

This model is analysed in detail in chapter 4. We first prove consistency of the model and derive closed form solutions for mean and variance of the key parameters at arbitrary times t. We also investigate how the different parts of the dynamics can be decomposed in microstructure drift and volatility, and exogenous (market) drift and volatility. We then turn to the important issue of analysing the time-to-fill τ_p of a limit order placed at price p: for a trader it is crucial to know the full distribution of the time-to-fill in order to decide if and when a limit order should be converted to a market order, because each time this operation is performed, she 'loses the spread'. As always, there is the trade-off between immediate execution at a bad price, and uncertain execution at a good price. We propose three methods to compute either the distribution of the time-to-fill τ_p or its Laplace transfrom $u(x)$ conditional on the initial order book state x. Firstly, we show that $u(x)$ can be computed as the solution of a Dirichlet problem. Secondly, we can determine the cumulative distribution function of the time-to-fill using Monte Carlo simulations. Although exact in principle, both methods take a long time to complete, as they either require complex numerical computations or converge very slowy. Since the decision about market vs. limit order often has to be taken within milli-seconds, these approaches are not applicable in practice. Therefore we propose a third method, inspired by the work of Fouque, Papanicolaou, and Sircar (2000). We introduce an extra parameter ϵ in the model equations and make an asymptotic approximation of the Laplace transform of τ_p:

$$u(x) \approx u^{(0)}(x) + \epsilon u^{(1)}(x)$$

where $u^{(0)}$ and $u^{(1)}$ are given in closed form. Economically, it turns out that the zero-order term $u^{(0)}$ corresponds to a simple one-dimensional Bachelier model for the midquote price with fixed spread. The first order term $u^{(1)}$ accounts for the first order correction of introducing a stochastic spread. Mathematically, the asymptotic analysis allows us to reduce a computationally complex PDE to two simple ODEs that can be solved explicitly. We then give an estimate of the error and test the method by comparing it to the Monte Carlo method. It turns out that the asymptotic method is fairly accurate, and extremely fast, and therefore well-suited for real-world applications.

In chapter 5, we show how the order book model can easily be calibrated to high-frequency order book data from Nasdaq, and test it by comparing real time-to-fill from the Nasdaq data with the theoretical time-to-fill of the order book model, conditional on the order book state at the time when the limit order was placed. It turns out that we get a good fit in general, with better results if we include more parameters in the order book model.

1.3.2 Part II: Optimal trading strategies in limit order books

While we focused in the first part of this thesis on descriptive statements on the orderbook and its dynamics, our objective in the second part is to make normative assertions. For this end, we take the position of one particular trader. We will consider two typical optimal control problems that traders (or automated microtraders) are faced with: portfolio liquidation and 'peg-cross' strategies.

We first remain in the setting of a simplified version of the model developed in part I. In chapter 6 our objective is to compute optimal static strategies for the portfolio liquidation problem when only the use of market orders is allowed. Using the Euler-Lagrange method on a heuristic level, we can divide the state space in 3 different regions, in which the optimal strategy is either to trade by a large single block trade, to trade continuously, or to wait and not trade at all. It turns out that the parameter which decides in which region we are is the ratio

$$\frac{s-\mu}{x}$$

where s is the current expected spread, μ is the average spread size, and x is the current number of shares that remains to be traded. By a verification argument, the candidate optimal strategy obtained by the Euler-Lagrange method is then shown to be the true optimal strategy. It also turns out that the base case considered in Obizhaeva and Wang (2005) and Alfonsi, Fruth, and Schied (2008) corresponds to a special case of the model developed in part I. As a by-product, we obtain a method to estimate the resilience parameter that appears in these models.

In chapter 7, we again treat the problem of optimal portfolio liquidation, this time, however, we want to compute optimal adaptive strategies that allow the use of market and limit orders, and cancellation of limit orders. In order to make the analysis tractable, we work in an orderbook model with instantaneous resilience, which can be considered as an extension of the market impact model in Almgren and Chriss (2001). Execution of limit orders are modelled by a compound Poisson process which allows us to include partial execution of limit orders. In spite of all simplifications, the model captures the main characteristics and trade-offs of market and limit orders:

(i) Market orders are executed immediately, while limit orders are (partially) filled at a random time.

(ii) Limit orders are executed at a better price than market orders.

(iii) Both market and limit orders have a market impact: orders of higher size increase the execution costs.

In section 7.3, we first consider the liquidation problem on an infinite time interval. While this problem formulation does not seem to be of interest for a practitioner, it has two important advantages. Mathematically, we reduce the number of dimensions by one and therefore simplify the problem. Economically, we can reduce the optimal strategies to simple functions of four *key market parameters*, which are simple ratios of our model parameters. We are able to compute a closed-form expression for the optimal strategy, which consists of two components:

(i) A market order trading rate which is a fraction of the outstanding assets. This corresponds to liquidating at (constant) exponential speed.

(ii) A limit order which is placed right at the beginning of the trading program. Its size is also a certain fraction of the outstanding assets. As long as it is not filled, a certain part is constantly cancelled, due to market order trading. As soon as the limit order is executed, a new limit order is placed.

We then investigate different kind of market models, and show how the key market parameters influence the trading strategies.

In section 7.5 we consider the liquidation problem on a finite time interval. When the remaining time for the liquidation programm is large, the optimal trading strategy is similar to the infinite time interval case, and can therefore easily be expressed in terms of the key market parameters. When the 'time-to-go' is short, the remaining shares are just liquidated linearly using market orders. A similiar problem has been considered in Kratz (2011) and Naujokat and Westray (2010), but to the knowledge of the author, this is the first time that market impact and partial execution of limit orders has been included. Another contribution in this chapter is the detailed analysis of the optimal strategy and the reduction to the key market parameters.

Finally, in chapter 8, we turn to the important issue of 'peg-cross' strategies. To the knowledge of the author, this is the first time that this problem has been treated mathematically. It deals with the task of buying (or selling) a single share package, which is not split up into smaller pieces anymore. This job is often carried out by automated microtraders. The order can either be placed as a single 'block' market or a limit order. When a market order is sent, the trading program is finished. When a limit order is placed, there are two possibilites: the trader can either *peg*, that is maintain the limit order (and wait for execution), or *cross*, that is cancel the limit order, cross the spread and convert it to a market order. The trading program is finished when either the limit order is executed or converted to a market order. The main drivers for the peg-cross decision is the size of the spread (i.e. the extra charge of a market order compared to a limit order) and the probability that the limit order is filled. We model the spread by a positive diffusion process

$$\mathrm{d}\mathfrak{s}(t) = r(\mathfrak{s}(t))\mathrm{d}t + \sigma(\mathfrak{s}(t))\mathrm{d}W(t)$$

and the execution of the limit order by a Poisson process, as in chapter 7. We can then formulate the peg-cross problem as an optimal stopping problem, and focus on the case of infinite time horizon. In particular, we compute closed form solutions when the spread is (i) a geometric Brownian motion (GBM), and (ii) a Cox-Ingersoll-Ross (CIR) process. Using a heuristic approach, we guess that there is an optimal crossing level: When the spread is smaller than this level, it is optimal to cross immediately and pay the extra charge. When the spread is above the optimal level, pegging is optimal. We then verify that this corresponds to the optimal strategy. It turns out that the expected costs are given by two components:

- The cost of simply waiting for the limit order.
- A discount that takes account of the fact that the spread might fall to the optimal level at a later stage.

We also perform a test of our optimal strategy with real-world order book data, which performs surprisingly well, given the simplicity of the model. We find that the CIR model performs slightly better than the GBM model.

1.4 Acknowledgements

First of all, I would like to thank my advisor Prof. Peter Bank for his support and ideas over the last years. He was always open for discussions and helped me with both technical details and fundamental questions.

Moreover, I am particularly grateful to my colleagues at Deutsche Bank Quantitative Products Laboratory, especially Antje Fruth, Mikhail Urusov and Almut Birsner. They have always been a great source of support and motivation. In particular, the discussions with Antje and Mikhail were always a great inspiration. As an employee of the the Quantitative Products Laboratory, I was able to benefit from its outstanding working atmosphere, infrastructure and funding over several years. Therefore, I would like to sincerely thank the people from Humboldt Universität, Technische Universität Berlin and Deutsche Bank, in particular Prof. Alexander Schied and Marcus Overhaus, for making this joint project possible. Moreover, the support by Technische Universität Berlin is acknowledged.

Last but not least, I wish to express my deepest gratitude to my family and friends, in particular my parents Renate and Andreas and my sister Mira. Their love and support has been invaluable.

Part I

Limit order book models

Chapter 2

A general framework for limit orderbook models

Our objective is to set up a framework for limit order book models that allows at the same time for high flexibility of statistical properties, dynamics and shape of the order book, and sufficient mathematical tractability. Currently, there is no standard limit order book microstructure model. Therefore we have to carry out some basic work until we can finally give a precise definition of an order book in definition 2.1. It is clear that the dynamics of an orderbook depend on random events such as future submissions of market orders by other market participants. We will therefore place ourselves for the rest of this thesis in a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathrm{P})$ satisfying the usual conditions.

2.1 Modelling framework

We start by setting up a general framework that allows to represent orderbooks and describe their evolution over time. For this we need to introduce a number of modelling assumptions:

Modelling assumption 2.1 (Logarithmic prices)**.** We consider logarithmic prices. Thus the price range is $\mathbb{R}$.

The order book describes the stock price at its very micro-level. Typical questions we try to answer within these models (e.g. time-to-fill, pegging vs. crossing) have a time horizon of a few minutes, and sometimes only milli-seconds. Thus, at this level, the difference between a linear and a logarithmic price scale is negligeable.

Moreover, we introduce

Modelling assumption 2.2 (Infinitesimal tick size)**.** Orders can be placed and cancelled at any (real-valued) price. Thus the tick-size is infinitesimal.

and

Modelling assumption 2.3 (Continuous order size)**.** Any positive real-valued number of limit or market orders can be placed/executed. Any positive real-valued amount can be cancelled, if it has been placed as a limit order and has not been filled yet.

In real markets, we usually have discrete and equi-spaced price ticks, i.e. orders can only be placed at fixed discrete price levels, for example at every full cent. Fractional order sizes are not possible and often there is even a minimum order size (e.g. 100 shares). The continuity assumptions 2.2- 2.3 are idealizations which we make in order to be able to describe and study orderbook dynamics via SDEs.

We shall also require

Modelling assumption 2.4 (Limit order placement in spread)**.** Given the current order book state and an incoming limit order of known size x that is placed inside the spread, the limit order is placed right left to the best ask (for a sell limit order) and the spread is filled up with the limit order according to a deterministic rule specified by the model.

If we want to model the evolution of the order book in time, and a new limit order is placed inside the spread, we need a rule where exactly in the spread it is placed. If the spread comprises more than one price tick, the order can be placed on any of these ticks, or be split up and placed on several ticks simultaneously. We require that the limit order is always placed right next to the corresponding best quote, such that there is no gap (empty tick) in the orderbook. Moreover, we assume that there is a deterministic rule which tells us, how much of the tick should be filled, and when the limit order should be placed on the next tick.

2.1.1 Representation of limit order books

We will now introduce two equivalent order book representations that satisfy the restrictions imposed by modelling assumptions 2.1-2.4.

Volume-density representation

Notation 2.1 (Volume-density represenation)**.** The *volume-density representation* of a stochastic order book is a quadruple $(\mathfrak{a}, \mathfrak{b}, \alpha, \beta)$ of functions where $\mathfrak{a}, \mathfrak{b} : \mathbb{R}^+ \times \mathbb{R} \times \Omega \to (0, \infty)$ and $\alpha, \beta : \mathbb{R}^+ \times \Omega \to \mathbb{R}$, such that $\mathbb{P}$-a.s. for all $t \geq 0$, $p \to \mathfrak{a}(t, p, \omega)$ and $p \to \mathfrak{b}(t, p, \omega)$ are bounded on compacts and Lebesgue measureable.

The function $\mathfrak{a}$ corresponds to the ask side and $\mathfrak{b}$ to the bid side of the book. $\alpha(t)$ and $\beta(t)$ denote the best ask and bid at time t, so we expect that the inequality $\beta(t) \leq \alpha(t)$ always holds. $\mathfrak{a}(t, p)$ is the volume-density at time t at price p for $p \geq \alpha(t)$, i.e. the density of limit orders stored at price p: at time t there are $\int_p^{p+\Delta p} \mathfrak{a}(t, r) \mathrm{d}r$ limit orders stored in the price range between p and $p + \Delta p$. The part of the curve for $p < \alpha(t)$ ensures that modelling assumption 2.4 is satisfied: if a limit order of size x is placed inside the spread, the area of size x under the curve $p \to \mathfrak{a}(t, p)$ is filled left from the best ask, and the best ask is updated accordingly. As $\mathfrak{a}(t, p) > 0$, there cannot be gaps in the orderbook. Figure 2.1 shows a limit order book in volume-density representation with best ask $\alpha = 27.11$, best bid $\beta = 27.09$,

The advantage of this representation is that we can easily fit the model at time $t = 0$ to market data: The discrete tick size of real order books will result in $\mathfrak{a}(0, p), p \geq \alpha(0)$ being piece-wise constant. By approximation of the step function with a smooth function $p \to \mathfrak{a}(0, p)$ we can reach any degree of smoothness. Additionally, we need to extend $\mathfrak{a}(0, p)$ beyond the best ask,

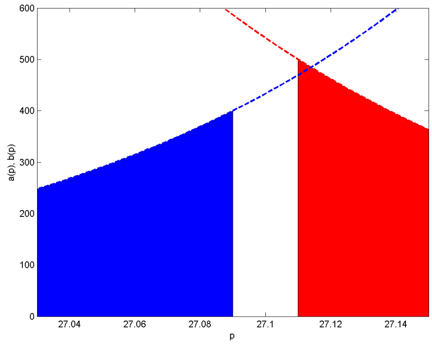


Figure 2.1: Volume-density representation of order book: The bid side $\mathfrak{b}$ in blue is filled up to best bid at price 27.09, ask side $\mathfrak{a}$ in red is filled up to best ask at price 27.11, the spread equals 0.02. The part of the $\mathfrak{a}$ and $\mathfrak{b}$ beyond the best quotes is represented as a dashed line.

i.e. for $p < \alpha(0)$. When a new limit order of a certain size is placed inside the spread, one needs to know *how* it is placed inside the spread. The shape of the curve beyond the best ask defines how the limit order is placed: the curve will just be filled up until the complete size of the limit order has been attained. Modelling assumption 2.4 guarantees that we know this part of the function $\mathfrak{a}$. In general it is not clear how $\mathfrak{a}(0, p)$ for $p < \alpha(0)$ can be fitted to market data. We will see later, that for the models of interest, this will not pose a problem.

Price-cumulated representation

By integrating the volume-density representation with respect to price p and choosing the integration constant in a way such that the resulting curve intersects the price-axis at the best ask, we obtain an equivalent representation:

Notation 2.2 (Price-cumulated represenation)**.** Given an order book in volume-density representation $(\mathfrak{a}, \mathfrak{b}, \alpha, \beta)$, we define the corresponding *price-cumulated representation* as the pair $(\mathfrak{A}, \mathfrak{B})$ of functions where $\mathfrak{A}, \mathfrak{B} : \mathbb{R}^+ \times \mathbb{R} \times \Omega \to \mathbb{R}$ are defined for each t by

$$\mathfrak{A}(t, p, \omega) = \int_{\alpha(t)}^{p} \mathfrak{a}(t, s, \omega) \mathrm{d}s \tag{2.1}$$

$$\mathfrak{B}(t, p, \omega) = \int_{p}^{\beta(t)} \mathfrak{b}(t, s, \omega) \mathrm{d}s \tag{2.2}$$

Remark 2.1. Note that by definition $\mathfrak{A}(t, \alpha(t)) = \mathfrak{B}(t, \beta(t)) = 0$, and that for all $t \geq 0$ $p \to \mathfrak{A}(t, p)$ is strictly increasing. The price-cumulated representation can be interpreted as follows: for $p \geq \alpha(t)$, $\mathfrak{A}(t, p)$ is the maximal quantity of shares you can buy when you are willing to pay at most price p for a unit of shares at time t. For $p < \alpha(t)$, the quantity $|\mathfrak{A}(t, p)|$ denotes minimal size of a new limit order that needs to be placed inside the spread such that price level p is attained by the limit order. Similarly for $p \leq \beta(t)$, $\mathfrak{B}(t, p)$ is the maximal quantity of shares you can sell when you are only willing to sell at prices $\geq p$ per share unit at time t. Figure 2.2 shows the the price-cumulated representation corresponding to the volume density representation in figure2.1.

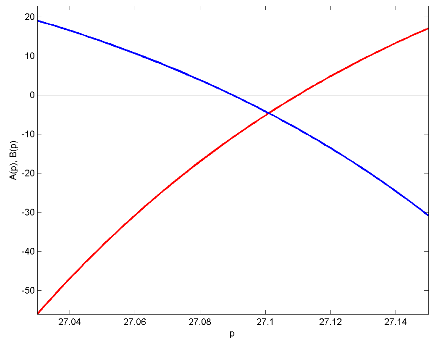


Figure 2.2: Price-cumulated representation of the same order book as in figure 2.1. Again the bid side is represented in blue, and the ask side in red.

Both representations being equivalent, we will freely switch between them. The volume-density representation has the advantage that it is closer to real-world orderbook and thus more intuitive. From a modelling point of view, however, the price-cumulated representation also has a number of advantages, e.g.:

- We only need to keep track of $\mathfrak{A}$ and $\mathfrak{B}$ instead of $\mathfrak{a}, \mathfrak{b}, \alpha, \beta$.
- For each t, $p \to \mathfrak{A}(t,p)$ is differentiable.
- The impact of a market order is easier to describe: if at time t a buy market order of size $x > 0$ is executed, the price-cumulated ask curve is shifted downwards by x: $\mathfrak{A}(t_+, p) = \mathfrak{A}(t,p) - x$ for all $p \in \mathbb{R}$. In volume-density representation this operation is harder to describe: one first has to compute how far the market order of size x 'eats' into the order book, and then update the best ask parameter accordingly. The volume-density curve is a good choice if you know how $\alpha(t)$ and $\beta(t)$ behave. In our approach, however, the behaviour of best ask/bid are an output and not on input, as will be made clear in the next part.

Before we proceed, let us introduce some more order book terminology:

Notation 2.3 (Spread)**.** The *spread* $\mathfrak{s}(t)$ is defined to be the distance between best bid an best ask

$$\mathfrak{s}(t) := \alpha(t) - \beta(t). \tag{2.3}$$

Notation 2.4 (Midquote price)**.** The *midquote price* $\mathfrak{m}(t)$ is defined to be the midpoint between best bid and best ask

$$\mathfrak{m}(t) := \frac{\alpha(t) + \beta(t)}{2}. \tag{2.4}$$

Notation 2.5 (Volume)**.** The *(best) ask volume* $V^{\mathfrak{A}}(t)$ of the ask side denotes the quantity of sell orders stored on the best ask $\alpha(t)$ at time t and defined by

$$V^{\mathfrak{A}}(t) := \mathfrak{a}(t, \alpha(t)). \tag{2.5}$$

Similary we define the *(best) bid volume* $V^{\mathfrak{B}}(t)$ by

$$V^{\mathfrak{B}}(t) := \mathfrak{b}(t, \beta(t)). \tag{2.6}$$

Notation 2.6 (Volume imbalance)**.** The *volume imbalance* $\mathfrak{z}(t)$ indicates the (im-)balance of volume on bid vs. ask side. It is defined by

$$\mathfrak{z}(t) := \log \frac{V^{\mathfrak{B}}(t)}{V^{\mathfrak{A}}(t)}.$$

Note that we can easily express these quantities in terms of the price-cumulated representation, e.g.,

$$V^{\mathfrak{A}}(t) = \frac{\partial \mathfrak{A}}{\partial p}(t, \alpha(t)). \tag{2.7}$$

2.1.2 Dynamics of limit order book models

We now turn to the question how the order book evolves in time. From now on, we will only consider the ask side, the bid side being symmetric. When both sides are considered, we add the superscript $^{\mathfrak{A}}$ (respectively $^{\mathfrak{B}}$) to all processes, parameters, etc. referring to the ask (bid) side. Note that at this stage, the derivation of SDEs is purely formal. No considerations of existence and uniqueness (in any sense) will be made. The reason is that we will not (yet) make any further assumptions on the stochastic processes that govern the SDEs. At the moment we are only interested in the 'bookkeeping'.

We make the following assumptions on the evolution of the order book

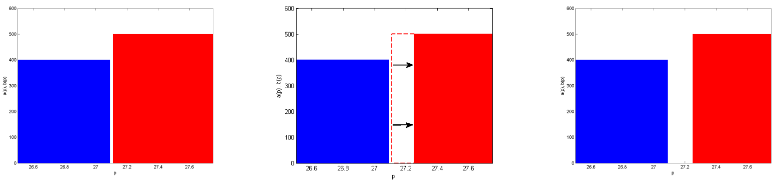
Modelling assumption 2.5 (Dynamics of order book)**.** The state of the order book at time $T \geq 0$ is fully determined by

- its initial state at time 0
- market and limit orders placed between time $[0, T]$
- cancellation of limit orders between time $[0, T]$

We will distinguish between limit orders placed on exisiting order (which do not influence the best ask) and those placed inside the spread (and thus decrease the best ask). Hence the evolution of the ask side of an orderbook is determined by the following processes:

Notation 2.7 (Order flow processes)**.** For $t \geq 0$, we denote

- $M(t, \omega)$ the number of buy market orders submitted up to time t,
- $S^+(t, \omega)$ the number of sell limit orders placed inside the spread up to time t,
- $S^-(t, \omega)$ the number of sell limit orders cancelled inside the spread up to time t,
- $\mathrm{d}Y^+(t, \omega)$ the intensity/speed with which limit orders are placed at time t,
- $l^+(t, p, \omega)$ the number of limit orders placed on the order book at price level p,
- $\mathrm{d}Y^-(t, \omega)$ the intensity/speed with which limit orders are cancelled at time t,



(a) Orderbook before buy market order execution **(b)** Execution of market buy order 'eats' into order book **(c)** Orderbook after buy market order execution

Figure 2.3: Effect of a buy market order in a block-shaped order book in volume-density representation. A market order of size $x = 70$ is executed at the best ask. The price level of the best ask is moved from 27.11 to 27.25. The bid side is unaffected from the buy market order.

- $l^-(t, p, \omega)$ the number of limit orders cancelled in the order book at price level p.

Remark 2.2. Note that M, S^+, S^-, Y^+, Y^- are all non-decreasing, and l^+, l^- are positive.

In the next sections we will explain how assumption 2.5 will be translated in an infinite dimensional system of SDEs.

We will sometimes explain the effect of market/ limit orders in volume-density represenation, whenever this seems more intuitive and then translate into price-cumulated representation. The final equations will all be in terms of price-cumulated representation. Of course, we can switch between both representations at any time.

Initial state

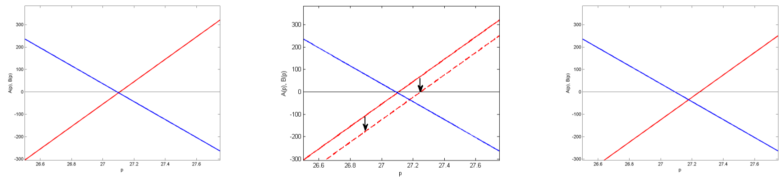
We assume that the initial state $(\mathfrak{A}(0,\cdot), \mathfrak{B}(0,\cdot))$ is given by the order book at time 0. The initial state can be fitted to empirical order book data and will be deterministic in most applications.

Market orders

When a buy market order of size x is executed, then the limit orders on the ask side will be traded against it, in increasing price order, starting with the best ask, until all x shares have been traded. In the volume-density representation this corresponds to 'eating' into the order book from the left, starting at the best ask, as illustrated in figure 2.3. Thus the best ask will be shifted to a higher price level: $\alpha(t_+) = \alpha(t) + p(x)$ where $p(x)$ satisfies $\int_{\alpha(t)}^{\alpha(t)+p(x)} \mathfrak{a}(t, r)\mathrm{d}r = x$. In the price-cumulated representation this corresponds to shifting $\mathfrak{A}$ downwards by x, as shown in figure 2.4. Thus, when considering the isolated effect of a buy market order on the orderbook

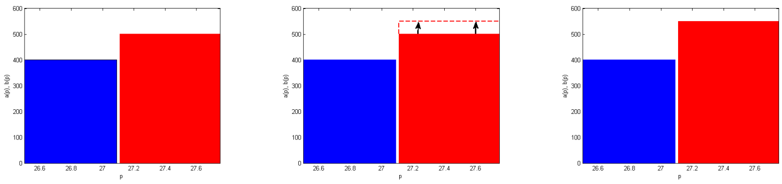
$$\mathrm{d}\mathfrak{A}(t, p) = -\mathrm{d}M(t) \tag{2.8}$$

is the differential description of the input of a market order.



(a) Orderbook before buy market order execution **(b)** Execution of market buy order shifts $\mathfrak{A}$ down **(c)** Orderbook after buy market order execution

Figure 2.4: Effect of a buy market order in a block-shaped order book in price-cumulated representation. A market order of size $x = 70$ is executed at the best ask.



(a) Orderbook before limit order placement on ask side **(b)** Placing limit orders increases ask side **(c)** Orderbook after limit order placement on ask side

Figure 2.5: Effect of a placing a sell limit order in a block-shaped order book in volume-density representation. The ask volume increases from 500 to 550. The bid side is unaffected by the sell limit order.

Spread limit orders

Suppose a sell limit order of size x is placed inside the spread. By assumption 2.4 limit orders are placed to the left of the best ask, according to the rule given by the model. In volume-density represenation the part of the order book to the left of the best ask, $\mathfrak{a}(t,p), p < \alpha(t)$ prescribes how the spread is decreased. More specifically, it is filled from $\alpha(t)$ up to $p(x) < \alpha(t)$ such that $\int_{p(x)}^{\alpha(t)} \mathfrak{a}(t,s)\mathrm{d}s = x$ holds.

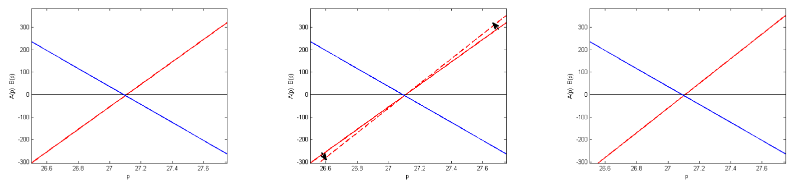
In price-cumulated representation this corresponds to vertically shifting the $\mathfrak{A}(t,\cdot)$-curve upwards by x. The isolated effect is described by the SDE

$$\mathrm{d}\mathfrak{A}(t,p) = \mathrm{d}S^{+}(t). \tag{2.9}$$

Cancelled spread limit orders

When a sell limit order of size x is cancelled at the spread, this has exactly the same effect as executing a buy market order of size x; i.e. 'eating' into the order book from the left, starting at the best ask (in volume-density representation) and shifting $\mathfrak{A}$ downwards by x in price-cumulated representation. Hence the isolated effect of cancelling a spread limit order is given by the SDE

$$\mathrm{d}\mathfrak{A}(t,p) = -\mathrm{d}S^{-}(t). \tag{2.10}$$



(a) Orderbook before limit order placement on ask side **(b)** Placing limit orders changes gradient of $p \mapsto \mathfrak{A}(t,p)$ **(c)** Orderbook after limit order placement on ask side

Figure 2.6: Effect of a placing a sell limit order in a block-shaped order book in price-cumulated representation.

Placing limit orders

Placing limit orders on existing orders does not change the best ask, and increases the density of limit orders stored in the order book. In volume-density representation this corresponds to

$$\mathrm{d}\mathfrak{a}(t,p) = l^+(t-,p)\mathrm{d}Y^+(t).$$

In price-cumulated representation we get the corresponding SDE

$$\mathrm{d}\mathfrak{A}(t,p) = L^+(t-,p)\mathrm{d}Y^+(t), \tag{2.11}$$

where $L^+(t,p) = \int_{\alpha(t)}^{p} l^+(t,q)\mathrm{d}q$. Figures 2.5 and 2.6 illustrate this effect for a block-shaped order book with $l^+(t,p) = \mathfrak{a}(t,p)$ in volume-density representation and price-cumulated representation, respectively.

Cancelled limit orders

Cancelling a limit order corresponds to the opposite effect of placing a limit order. Hence, in price-cumulated representation we have

$$\mathrm{d}\mathfrak{A}(t,p) = -L^-(t-,p)\mathrm{d}Y^-(t), \tag{2.12}$$

where $L^-(t,p) = \int_{\alpha(t)}^{p} l^-(t,q)\mathrm{d}q$.

Full equation

Now group market and spread limit orders together by setting $X(t) := S^+(t) - S^-(t) - M(t)$. Then equations (2.8), (2.9), (2.10), (2.11), (2.12) can be summarized at each price level p by the following orderbook dynamics:

$$\mathrm{d}\mathfrak{A}(t,p) = \underbrace{\mathrm{d}X(t)}_{\text{MO and (cancelled) SLO}} + \underbrace{L^+(t-,p)\mathrm{d}Y^+(t) - L^-(t-,p)\mathrm{d}Y^-(t)}_{\text{LO and cancelled LO}} \tag{2.13}$$

2.1.3 Proportional placing/cancelling of limit orders

The orderbook model can be made significantly more tractable by making the additional

Modelling assumption 2.6 (Proportional limit order placement/cancellation)**.** The number of new limit orders placed on the order book at price level p is proportional to the current number of limit orders stored at price p. The number of limit orders cancelled at price level p is proportional to the current number of limit orders stored at price p.

The assumption of proportional limit order placement and cancellation is very strong and needs some comments. It turns out (see theorem 2.2) that this assumption is crucial to reduce the dimensionality of the order book from infinity to a finite number. The entire state of the order book can be reduced to a set of key parameters. This seems to be very restrictive at first sight, however, it was found empirically that those key parameters carry most of the explanatory power of the order book in two respects: the true value of the stock and traders behaviour such as order aggressiveness and timing of order submissions. We can therefore justify modelling assumption 2.6 by the significant advantage of dimension reduction and the empirical fact that we still retain the most relevant parts orderbook information. In addition, modelling assumption 2.6 can be justified by the empirical results we obtain when testing the model with real orderbook data in chapter 5. For a more detailed discussion we refer to remark 2.3 following theorem 2.2 on page 31.

In mathematical terms, modelling assumption 2.6 simply means that

$$\begin{aligned} l^+(t,p) &= \mathfrak{a}(t,p) \\ l^-(t,p) &= \mathfrak{b}(t,p). \end{aligned}$$

Now write

$$Y(t) := Y^+(t) - Y^-(t)$$

and (2.13) simplifies to

$$\mathrm{d}\mathfrak{A}(t,p) = \underbrace{\mathrm{d}X^{\mathfrak{A}}(t)}_{\text{MO and (cancelled) SLO}} + \underbrace{\mathfrak{A}(t-,p)\mathrm{d}Y^{\mathfrak{A}}(t)}_{\text{LO and cancelled LO}}, \tag{2.14}$$

and for the bid side

$$\mathrm{d}\mathfrak{B}(t,p) = \underbrace{\mathrm{d}X^{\mathfrak{B}}(t)}_{\text{MO and (cancelled) SLO}} + \underbrace{\mathfrak{B}(t-,p)\mathrm{d}Y^{\mathfrak{B}}(t)}_{\text{LO and cancelled LO}}. \tag{2.15}$$

We can now give a precise definition of an order book model with proportional placing and cancellation of limit orders.

Definition 2.1 (Orderbook model)**.** A *proportional orderbook model* in price-cumulated representation is given by a tupel $\left(X^{\mathfrak{A}}, Y^{\mathfrak{A}}, X^{\mathfrak{B}}, Y^{\mathfrak{B}}, \mathfrak{A}(0,\cdot), \mathfrak{B}(0,\cdot)\right)$ where $X^{\mathfrak{A}}, Y^{\mathfrak{A}}, X^{\mathfrak{B}}, Y^{\mathfrak{B}}$ are $\mathbb{R}$-valued stochastic processes and $\mathfrak{A}(0,p), \mathfrak{B}(0,p) : \mathbb{R} \to \mathbb{R}$ are continuous deterministic functions such that

(i) $X^{\mathfrak{A}}(0) = Y^{\mathfrak{A}}(0) = X^{\mathfrak{B}}(0) = Y^{\mathfrak{B}}(0) = 0$

(ii) for all $p \in \mathbb{R}$ there exist unique strong solutions to (2.14) and (2.15) with initial conditions $\mathfrak{A}(0,p), \mathfrak{B}(0,p)$

If a.s. for all $t \geq 0$ there exist unique $\alpha(t)$, $\beta(t)$ satisfying $\mathfrak{A}(t,\alpha(t)) = \mathfrak{B}(t,\beta(t)) = 0$, the model is called *valid*. If we have a.s. $\alpha(t) \geq \beta(t)$ for all $t \geq 0$, the model is called *consistent*. $\left(X^{\mathfrak{A}}, Y^{\mathfrak{A}}, X^{\mathfrak{B}}, Y^{\mathfrak{B}}\right)$ are called the *order flow processes*, and $\mathfrak{A}(0,\cdot), \mathfrak{B}(0,\cdot)$ the *initial state* of the ask/bid side of the order book.

2.2 Dimension reduction

Henceforth we will assume proportional limit order placement 2.6 and work with the simplified equation (2.14). Note that in general this is an infinite-dimensional coupled system of SDEs:

- We have infinitely many equations for each price p.
- The equations are coupled, because $X(t) = X(t, (\mathfrak{A}(t,\cdot,\omega), \mathfrak{B}(t,\cdot,\omega)))$ and $Y(t) = Y(t, (\mathfrak{A}(t,\cdot,\omega), \mathfrak{B}(t,\cdot,\omega)))$ depend on the entire curve.

However, the next theorem will show that modelling assumption 2.6 does greatly simplify the analysis, and that we can reduce the complexity of the system. We focus on the case of continuous semimartigales because the resulting formula is much nicer, and we will only be using this case in our following analysis.

Theorem 2.1 (Existence of order book). *1. Let $(X, Y, \mathfrak{A}(t,\cdot))$ be the ask side of an order book model. Suppose that X and Y are continuous semimartingales satisfying $X(0) = Y(0) = 0$. Then for each p, the order book equation*

$$\mathrm{d}\mathfrak{A}(t,p) = \mathrm{d}X(t) + \mathfrak{A}(t,p)\mathrm{d}Y(t) \tag{2.16}$$

admits a unique solution in $\mathbb{D}$ (the space of adapted càdlàg processes), which is again a continuous semimartingale.

2. Moreover, for each p, the order book $\mathfrak{A}$ is given explicitly by

$$\mathfrak{A}(t,p) = \mathcal{E}(Y)(t)\left[\mathfrak{A}(0,p) + \int_0^t \frac{\mathrm{d}X(s) - \mathrm{d}[X,Y](s)}{\mathcal{E}(Y)(s)}\right], \tag{2.17}$$

where

$$\mathcal{E}(Y) = \exp\left\{Y(t) - \frac{1}{2}[Y,Y](t)\right\}$$

The proof of this and all following results are given at the end of the corresponding chapter.

We immediately have the following

Corollary 2.1 (Preservation of properties of initial order book). *Under the assumptions of Theorem 2.1, the order book preserves its initial properties: if $p \mapsto \mathfrak{A}(0,p)$ is continuous (differentiable, strictly increasing) then a.s. the same is true for the order book $p \mapsto \mathfrak{A}(t,p)$ at any time $t > 0$.*

In the next theorem, we will see that it is enough to know the initial order book form at time 0 and keep track of the best ask (bid) and the volume on both sides, to have all information about the limit order book. Moreover, we give an explicit expression of the dynamics of best ask and volume in terms of incoming and cancelled market and limit orders given by (X, Y). We will write $p \mapsto \mathfrak{A}_0(p)$ for the function $p \mapsto \mathfrak{A}(0, p)$.

Theorem 2.2 (Reduction to key parameters). *Let $(X, Y, \mathfrak{A}(t, \cdot))$ be the ask side of an order book model. Suppose that the assumptions of Theorem 2.1 hold, that $p \to \mathfrak{A}_0(p)$ is strictly increasing and three times continuously differentiable, and that $p \to \mathfrak{A}_0^{-1}(p)$ is twice continuously differentiable. Suppose moreover that the model is valid, i.e. X and Y are such that a.s. for all $t \geq 0$ there exists $\alpha(t) \in \mathbb{R}$ satisfying $\mathfrak{A}(t, \alpha(t)) = 0$. Define $V(t) = \frac{\partial \mathfrak{A}}{\partial p}(t, \alpha(t))$. Then*

(i) α and V are continuous semimartingales.

(ii) V is a.s. strictly positive.

(iii) The order book can be represented in parameterized form

$$\mathfrak{A}(t, p) = \frac{V(t)}{\mathfrak{A}_0'(\alpha(t))} \left[\mathfrak{A}_0(p) - \mathfrak{A}_0(\alpha(t))\right]. \tag{2.18}$$

(iv) The dynamics of (α, V) are given by the system of SDEs

$$\mathrm{d}\alpha(t) = -\frac{\mathrm{d}X(t)}{V(t)} + \frac{\mathrm{d}[X, Y](t)}{V(t)} - \frac{\mathfrak{A}_0''(\alpha(t))}{2\mathfrak{A}_0'(\alpha(t))} \frac{\mathrm{d}[X](t)}{V(t)^2} \tag{2.19}$$

$$\begin{aligned} \mathrm{d}V(t) &= V(t)\mathrm{d}Y(t) - \frac{\mathfrak{A}_0''(\alpha(t))}{\mathfrak{A}_0'(\alpha(t))}\mathrm{d}X(t) \\ &\quad + \frac{1}{2V(t)} \left\{ \frac{\mathfrak{A}_0'''(\alpha(t))}{\mathfrak{A}_0'(\alpha(t))} - \left(\frac{\mathfrak{A}_0''(\alpha(t))}{\mathfrak{A}_0'(\alpha(t))} \right)^2 \right\} \mathrm{d}[X](t). \end{aligned} \tag{2.20}$$

Remark 2.3. The main implication of the above theorem is that, for a fixed time t, the infinite-dimensional order book

$$(\mathfrak{A}(t, p), \mathfrak{B}(t, p)), \quad p \in \mathbb{R}$$

can be reduced to the four-dimensional set of key parameters best bid/ask and volume on best bid/ask

$$(\alpha(t), \beta(t), V^{\mathfrak{A}}(t), V^{\mathfrak{B}}(t))$$

with dynamics given by by (2.19) and (2.20) (and the corresponding SDEs for the bid side).

The dimension reduction and focus on the best quotes is due to the proportional placement and cancellation of limit orders, as described in modelling assumption 2.6. To justify this strong assumption, we will discuss why it makes sense to focus on the four-dimensional model $(\alpha(t), \beta(t), V^{\mathfrak{A}}(t), V^{\mathfrak{B}}(t))$. For this we will consider two empirical articles, in which the informational value of different parts of the order book are evaluated.

Pascual Gascó and Veredas (2008) study empirically which particular pieces of the order book do characterize the upcoming order flow, based on 6-months high frequency data from the Spanish Stock Exchange in 2000. They analyse three different probit models: a baseline model, which does not know anything about the order book state, a best quotes model which basically knows the four key parameters $(\alpha, \beta, V^{\mathfrak{A}}, V^{\mathfrak{B}})$, and the complete model which in addition has explanatory variables with information about the 5 ticks beyond best bid/ask. First they look at order aggressiveness, ranging from very aggressive large market orders to little aggressive limit orders and even less aggressive cancellations. It turns out that while the complete model performs best, most of the explanatory power is already concentrated on the key parameters, and the inclusion of the additional 5 best ticks does not provide much extra information. Interestingly, the authors find a difference between active and passive traders: The four key parameters are by far the most informative piece of the order book for active traders (liquidity consumers who trade with market orders). However, passive traders (liquidity providers who trade with limit orders) seem to base their order submissions on the complete order book. This suggests that the four-dimensional model might not be sufficient for very sophisticated traders. The authors also analyse which pieces of the order book are important in explaining the time between two order submission/ cancellations. It turns out that only the spread (which is a function of the four key parameters) shows an effect on the timing of orders.

Cao, Hansch, and Wang (2009) analyse the incremental information content of the order book, i.e. the additional information of order book data beyond the best quotes compared to the information of the four key parameters from the perspective of the true value of the security, using March 2000 data from the Australian Stock Exchange. The authors assess the incremental informational content of the order book on the true price. They find that the key parameters together with the last transaction price has an information share of approximately 78%, whereas the average information share of the 9 ticks beyond the best quotes (on both sides) is approximately 22%. When considering the ticks $2-4$ and $5-10$ separately, both are found to have an information share of more than 10%.

The above empirical results have the following implications for our model

- Most of the informational content on order flow, order timing and true price is already contained in the key parameters $(\alpha, \beta, V^{\mathfrak{A}}, V^{\mathfrak{B}})$. Thus it is a good tradeoff between model tractability and model realism to introduce the proportional limit order placing/cancelling in modelling assumption 2.6: the number of dimensions is reduced from infinity to only four key parameters, but the most important parts of the limit order book are still accounted for.

- *If* we want to include more information than just the 4 key parameters, but not keep track of the full (infinite-dimensional) order book, the empirical results do not help in finding *which* additional parts of the order book should be included; e.g. it does not make sense to include ticks $2-4$, but exclude ticks $5-10$.

Hence there are good mathematical and empirical reasons why we should work with the *proportional* orderbook model.

2.3 Linear orderbook models and fundamental examples

We will now look at two particular examples for the order book form. Before introducing these models, let us motivate them. It turns out to be particularly useful if the orderbook took the *linear* form given by

$$\mathfrak{A}(t,p) = \frac{V(t)}{V(0)}\mathfrak{A}_0(p-\alpha(t)+\alpha(0)), \tag{2.21}$$

where the volume $V(t)$ scales the height of the curve and best ask $\alpha(t)$ is shifted along the curve, as t varies.

To understand the implications of (2.21), assume for simplicity that X and Y are given by continuous finite variation processes. Then (2.19) and (2.20) simplify to

$$\begin{aligned} \mathrm{d}\alpha(t) &= -\frac{\mathrm{d}X(t)}{V(t)} \\ \mathrm{d}V(t) &= V(t)\mathrm{d}Y(t) - \frac{\mathfrak{A}_0''(\alpha(t))}{\mathfrak{A}_0'(\alpha(t))}\mathrm{d}X(t) \end{aligned}$$

We equate (2.18) and (2.21) to obtain

$$\frac{\mathfrak{A}_0(p-\alpha(t)+\alpha(0))}{V(0)} = \frac{\mathfrak{A}_0(p)-\mathfrak{A}_0(\alpha(t))}{\mathfrak{A}_0'(\alpha(t))} \tag{2.22}$$

$$\frac{\mathfrak{A}_0'(p-\alpha(t)+\alpha(0))}{V(0)} = \frac{\mathfrak{A}_0'(p)}{\mathfrak{A}_0'(\alpha(t))} \tag{2.23}$$

Now from (2.21), we get

$$\begin{aligned} \mathrm{d}\mathfrak{A}(t,p) &= \frac{1}{V(0)}\left[\mathfrak{A}_0(p-\alpha(t)+\alpha(0))\mathrm{d}V(t) - \mathfrak{A}_0'(p-\alpha(t)+\alpha(0))V(t)\mathrm{d}\alpha(t)\right] \\ &= \frac{1}{V(0)}\mathfrak{A}_0(p-\alpha(t)+\alpha(0))V(t)\mathrm{d}Y(t) - \frac{1}{V(0)}\frac{\mathfrak{A}_0''(\alpha(t))}{\mathfrak{A}_0'(\alpha(t))}\mathfrak{A}_0(p-\alpha(t)+\alpha(0))\mathrm{d}X(t) \\ &\quad + \frac{1}{V(0)}\mathfrak{A}_0'(p-\alpha(t)+\alpha(0))\mathrm{d}X(t) \\ &= \frac{1}{V(0)}\left[\mathfrak{A}_0'(p-\alpha(t)+\alpha(0)) - \frac{\mathfrak{A}_0''(\alpha(t))}{\mathfrak{A}_0'(\alpha(t))}\mathfrak{A}_0(p-\alpha(t)+\alpha(0))\right]\mathrm{d}X(t) + \mathfrak{A}(t,p)\mathrm{d}Y(t) \end{aligned}$$

Comparing with (2.14) we obtain

$$\frac{1}{V(0)}\left[\mathfrak{A}_0'(p-\alpha(t)+\alpha(0)) - \frac{\mathfrak{A}_0''(\alpha(t))}{\mathfrak{A}_0'(\alpha(t))}\mathfrak{A}_0(p-\alpha(t)+\alpha(0))\right] = 1$$

Now plug in (2.22) and (2.23), note that the equation should hold true for any $\alpha(t)$, so in particular for $\alpha(t)=\alpha(0)$ and write the constant $b = -\frac{\mathfrak{A}_0''(\alpha(0))}{\mathfrak{A}_0'(\alpha(0))}$. We get a differential equation for $p \to \mathfrak{A}_0(p)$

$$\mathfrak{A}_0'(p) + b\mathfrak{A}_0(p) = V(0), \qquad \mathfrak{A}_0(\alpha(0)) = 0$$

which has the unique solution

$$\mathfrak{A}_0(p) = \begin{cases} V(0)(p-\alpha(0)) & \text{if } b = 0 \\ \frac{V(0)}{b}\left(1-\exp^{-b(p-\alpha(0))}\right) & \text{if } b \neq 0 \end{cases}$$

Combining these solution with (2.21), we obtain the two fundamental examples for orderbooks in scalable form:

Example 2.1 (Block-shape model). For $b = 0$, we have the block-shape model

$$\mathfrak{A}(t,p) = V(t)(p - \alpha(t)) \tag{2.24}$$

where the dynamics of best ask and ask volume are given by

$$\mathrm{d}\alpha(t) = -\frac{\mathrm{d}X(t)}{V(t)} + \frac{\mathrm{d}[X,Y](t)}{V(t)}, \tag{2.25}$$

$$\mathrm{d}V(t) = V(t)\mathrm{d}Y(t), \tag{2.26}$$

for general market and limit order dynamics X, Y.

Note that the dynamics of the volume only depend on Y, i.e. placing and cancellation of limit orders on existing limit orders. This model corresponds to an order book in block form, see figure 2.7a. The total amount of limit orders stored on the ask side is infinite.

Example 2.2 (Exponential model). For $b \neq 0$, we have the exponential model

$$\mathfrak{A}(t,p) = \frac{V(t)}{b}\left(1 - \exp^{-b(p-\alpha(t))}\right) \tag{2.27}$$

where the dynamics of best ask and ask volume are given by

$$\mathrm{d}\alpha(t) = -\frac{\mathrm{d}X(t)}{V(t)} + \frac{\mathrm{d}[X,Y](t)}{V(t)} - b\frac{\mathrm{d}[X](t)}{V(t)^2},$$

$$\mathrm{d}V(t) = V(t)\mathrm{d}Y(t) + b\mathrm{d}X(t),$$

for general market and limit order dynamics X, Y.

This time the dynamics of the volume depend on both Y and X. This model corresponds to an order book where the density of order decreases (positive b) or increases (negative b) exponentially as you move from the best ask deeper inside the order book as illustrated in figure 2.7b. In real markets we typically observe that the liquidity decreases, as we move deeper into the order book, hence we will usually make the assumption $b > 0$. Note that in this case, the total amount of limit orders stored on the ask side is finite, it equals $\frac{V(t)}{b}$ at time t. Since we assume that X and Y are such that for all t there exists $\alpha(t) \in \mathbb{R}$ such that $\mathfrak{A}(t, \alpha(t)) = 0$, no limit order of size $\geq \frac{V(t)}{b}$ can be submitted at time t which ensures that $V(t)$ stays positive.

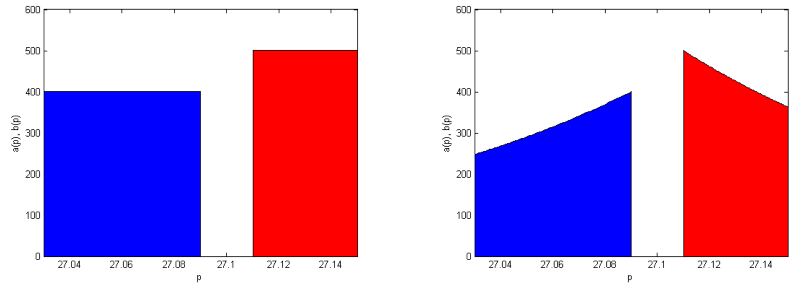
2.4 Proofs

We start by recalling two following well-known results (see, e.g., Protter, 2004)

Proposition 2.1 (Doléans-Dade exponential). *Let X be a semimartingale, with $X(0) = 0$ and satisfying a.s. for all $t \geq 0$*

$$\sum_{0 \leq s \leq t} |\Delta X(s)| < \infty.$$

Then there exists a unique semimartigale Z satisfying the equation $Z(t) = 1 + \int_0^t Z(s-)\mathrm{d}X(s)$. Z is called the Doléans-Dade (or stochastic) exponential of X, sometimes written as $Z = \mathcal{E}(X)$,



(a) Block-shape model **(b)** Exponential model

Figure 2.7: Block-shape and exponential (with $b = 8$) orderbook model in volume-density representation. The part of $\mathfrak{a}$ and $\mathfrak{b}$ beyond the best quotes is left out. The key parameters of both order books are $(\alpha, \beta, V^{\mathfrak{A}}, V^{\mathfrak{B}}) = (27.11, 27.09, 400, 500)$.

and is given explicitly by

$$\begin{aligned} Z(t) &= \exp\left\{X(t) - \frac{1}{2}[X,X](t)\right\} \prod_{0\le s\le t} (1+\Delta X(s)) \exp\left\{-\Delta X(s) + \frac{1}{2}(\Delta X(s))^2\right\} \\ &= \exp\left\{X^c(t) - \frac{1}{2}[X^c,X^c](t)\right\} \prod_{0\le s\le t} (1+\Delta X(s)), \end{aligned}$$

where X^c is the continuous part of X.

Proposition 2.2 (Existence and uniqueness)**.** *Given a vector of semimartingales $\mathbf{Z} = (Z^1, ..., Z^d)$, $\mathbf{Z}_0 = 0$ processes $J^i \in \mathbb{D}, 1 \le i \le n$, and operators F^i_j which are functional Lipschitz ($1 \le i \le n, 1 \le j \le d$), the system of equations*

$$X^i_t = J^i_t + \sum_{j=1}^{d} \int_0^t F^i_j(\mathbf{X})_{s-} dZ^j_s$$

($1 \le i \le n$) has a solution in $\mathbb{D}^n$, and it is unique. Moreover if $(J^i)_{i\le n}$ is a vector of semimartingales, then so is $(X^i)_{i\le n}$.

Now we proceed with the proofs.

Proof of Theorem 2.1. 1. This follows directly from Proposition 2.2

2. Fix $p \in \mathbb{R}$, and write $Z(t) = \mathcal{E}(Y)(t)$. We guess a solution of the form $\mathfrak{A}(t,p) = Z(t)(\mathfrak{A}(0,p) + U(t))$, for some $U(t)$. Use the product rule to obtain

$$\begin{aligned} \mathrm{d}\mathfrak{A}(t,p) &= (\mathfrak{A}(0,p) + U(t))\mathrm{d}Z(t) + Z(t)\mathrm{d}U(t) + \mathrm{d}[U,Z](t) \\ &= (\mathfrak{A}(0,p) + U(t))Z(t)\mathrm{d}Y(t) + Z(t)\mathrm{d}U(t) + \mathrm{d}[U,Z](t) \\ &= \mathfrak{A}(t,p)\mathrm{d}Y(t) + Z(t)\mathrm{d}U(t) + \mathrm{d}[U,Z](t) \end{aligned}$$

and equate with (2.16) to get

$$\mathrm{d}X(t) = Z(t)\mathrm{d}U(t) + \mathrm{d}[U,Z](t)$$

Now write $X(t) = M(t) + A(t)$ and $U(t) = N(t) + B(t)$ where M, N are continuous martingales and A, B are continuous finite variation processes. Thus we get

$$\mathrm{d}(M(t) + A(t)) = Z(t)\mathrm{d}(N(t) + B(t)) + \mathrm{d}[U, Z](t)$$

Note that $Z = \mathcal{E}(Y)$ is strictly positive and $1/Z$ is locally bounded. Equating the continuous martingale part we obtain

$$\mathrm{d}N(t) = \frac{\mathrm{d}M(t)}{Z(t)}.$$

Use the fact that

$$\mathrm{d}[U, Z](t) = \mathrm{d}[X, Y](t)$$

and equate the continuous finite variation part to obtain

$$\mathrm{d}B(t) = \frac{\mathrm{d}A(t) - \mathrm{d}[X, Y](t)}{Z(t)}.$$

Thus we obtain

$$\mathrm{d}U(t) = \frac{\mathrm{d}X(t) - \mathrm{d}[X, Y](t)}{Z(t)},$$

and finally we have equation (2.17).

□

Proof of Corollary 2.1. This follows immediately from the representation (2.17). □

Proof of Theorem 2.2. Write $Z(t) = \mathcal{E}(Y)(t)$. From Theorem 2.1 we have

$$\mathfrak{A}(t, p) = Z(t)(\mathfrak{A}_0(p) + U(t)), \tag{2.28}$$

where

$$U(t) = \int_0^t \frac{\mathrm{d}X(s) - \mathrm{d}[X, Y](s)}{Z(s)}.$$

and thus

$$0 = \mathfrak{A}(t, \alpha(t)) = Z(t)(\mathfrak{A}_0(\alpha(t)) + U(t))$$

and since $Z(t)$ is strictly positive, we get

$$0 = \mathfrak{A}_0(\alpha(t)) + U(t) \tag{2.29}$$

Moreover

$$V(t) = V^{\mathfrak{A}}(t) = \mathfrak{a}(t, \alpha(t)) = \frac{\partial \mathfrak{A}}{\partial p}(t, \alpha(t)) = Z(t)\mathfrak{A}_0'(\alpha(t)) \tag{2.30}$$

First note that since $\mathfrak{A}_0$ and $\mathfrak{A}_0^{-1}$ are two times continuous differentiable, we have that $\alpha(t) = \mathfrak{A}_0^{-1}(-U(t))$ and $V(t) = Z(t)\mathfrak{A}_0'(\alpha(t))$ are semimartingales, which shows (i).

By (2.30), we also have $V(t) > 0$ a.s. and thus (ii). Putting formulae (2.28), (2.29) and (2.30) together, we obtain

$$\mathfrak{A}(t, p) = \frac{V(t)}{\mathfrak{A}_0'(\alpha(t))}\left[\mathfrak{A}_0(p) - \mathfrak{A}_0(\alpha(t))\right]$$

which is (iii).

Finally we will show property (iv): Apply Itô's formula to (2.29) to obtain

$$0 = \mathfrak{A}_0'(\alpha(t))\mathrm{d}\alpha(t) + \frac{1}{2}\mathfrak{A}_0''(\alpha(t))\mathrm{d}[\alpha](t) + \frac{\mathrm{d}X(t) - \mathrm{d}[X,Y](t)}{Z(t)}.$$

From this we infer

$$\mathrm{d}[\alpha](t) = \frac{\mathrm{d}[X](t)}{V^2(t)}.$$

Hence we obtain (2.19).

Now apply Itô's formula to $V(t) = Z(t)\mathfrak{A}_0'(\alpha(t))$. We obtain

$$\begin{aligned}
\mathrm{d}V(t) &= \mathfrak{A}_0'(\alpha(t))\mathrm{d}Z(t) + Z(t)\mathfrak{A}_0''(\alpha(t))\mathrm{d}\alpha(t) \\
&\quad + \frac{1}{2}Z(t)\mathfrak{A}_0'''(\alpha(t))\mathrm{d}[\alpha](t) + \mathrm{d}[Z,\mathfrak{A}_0'(\alpha(\cdot))](t) \\
&= V(t)\mathrm{d}Y(t) \\
&\quad + V(t)\frac{\mathfrak{A}_0''(\alpha(t))}{\mathfrak{A}_0'(\alpha(t))}\left\{-\frac{\mathrm{d}X(t)}{V(t)} + \frac{\mathrm{d}[X,Y](t)}{V(t)} - \frac{\mathfrak{A}_0''(\alpha(t))}{2\mathfrak{A}_0'(\alpha(t))}\frac{\mathrm{d}[X](t)}{V^2(t)}\right\} \\
&\quad + \frac{V(t)\mathfrak{A}_0'''(\alpha(t))}{2\mathfrak{A}_0'(\alpha(t))V^2(t)}\mathrm{d}[X](t) + \mathrm{d}[Z,\mathfrak{A}_0'(\alpha(\cdot))](t) \\
&= V(t)\mathrm{d}Y(t) - \frac{\mathfrak{A}_0''(\alpha(t))}{\mathfrak{A}_0'(\alpha(t))}\mathrm{d}X(t) + \frac{\mathfrak{A}_0''(\alpha(t))}{\mathfrak{A}_0'(\alpha(t))}\mathrm{d}[X,Y](t) \\
&\quad + \frac{1}{2V(t)}\left\{\frac{\mathfrak{A}_0'''(\alpha(t))}{\mathfrak{A}_0'(\alpha(t))} - \left(\frac{\mathfrak{A}_0''(\alpha(t))}{\mathfrak{A}_0'(\alpha(t))}\right)^2\right\}\mathrm{d}[X](t) \\
&\quad + \mathrm{d}[Z,\mathfrak{A}_0'(\alpha(\cdot))] \\
&= V(t)\mathrm{d}Y(t) - \frac{\mathfrak{A}_0''(\alpha(t))}{\mathfrak{A}_0'(\alpha(t))}\mathrm{d}X(t) \\
&\quad + \frac{1}{2V(t)}\left\{\frac{\mathfrak{A}_0'''(\alpha(t))}{\mathfrak{A}_0'(\alpha(t))} - \left(\frac{\mathfrak{A}_0''(\alpha(t))}{\mathfrak{A}_0'(\alpha(t))}\right)^2\right\}\mathrm{d}[X](t),
\end{aligned}$$

where we used the equation for $\alpha(t)$ above and the identitiy

$$\begin{aligned}
\mathrm{d}[Z,\mathfrak{A}_0'(\alpha(\cdot))](t) &= Z(t)\mathfrak{A}_0''(\alpha(t))\mathrm{d}[Y,\alpha(\cdot)](t) \\
&= -\frac{\mathfrak{A}_0''(\alpha(t))}{\mathfrak{A}_0'(\alpha(t))}\mathrm{d}[X,Y](t).
\end{aligned}$$

This shows (iv). □

Chapter 3

Order flow in limit order books

In the previous chapter we established a framework for limit order book models, under general modelling assumptions 2.1 - 2.6. The dynamics of the model are expressed via the order flow of market and limit orders, on both sides of the order book $\left(X^{\mathfrak{A}}, Y^{\mathfrak{A}}, X^{\mathfrak{B}}, Y^{\mathfrak{B}}\right)$. In particular, we saw that modelling assumption 2.6 implies that the state of the order book can be reduced to the key parameters best quotes and volume on the best quotes. However, this framework only provides the 'bookkeeping' of the order book. In this chapter we will therefore analyse the dynamics of the order flow and its influence on the order book evolution. We start in section 3.1 with a review of the literature on order flow dynamics in limit order books and a synthesis of implications on the order flow processes $\left(X^{\mathfrak{A}}, Y^{\mathfrak{A}}, X^{\mathfrak{B}}, Y^{\mathfrak{B}}\right)$ in our modelling framework. Next, we propose a particular set of order flow processes in section 3.2 motivated by the order flow dynamics observed in limit order markets.

3.1 State-dependent behaviour of the order flow

In the following, we will give an overview over theoretical and empirical results on order flow dynamics in a limit order book. As we saw in theorem 2.2, the state of the order book (in our modelling framework) can be reduced to four key parameters best ask, best bid, volume on best ask and volume on best bid. We therefore focus on properties of the order flow that depend on the four key parameters and derived quantities such as spread, volume-imbalance and midquote price.

On the theoretical side, one of the first models was proposed by Cohen, Maier, Schwartz, and Whitcomb (1981) who find that a limit order placed in the order book has the following property: suppose an investor wants to buy shares and has the choice between market and limit orders. As the best bid increases to the best ask, it becomes relatively more attractive to place a market order and have certainty of execution.

Foucault (1999) proposes a trading game in which traders arrive sequentially and can choose to submit either a market or a limit order, and he computes the equilibrium. He finds that the proportion of limit orders is positively related to the average size of the spread.

Rosu (2009) studies the equilibrium of a trading game where players with different degrees of patience arrive to the market, submit limit orders or market orders, and can respond to new information by changing their strategy dynamically. He finds that a market sell order leads to a decrease in both the bid and the ask, the decrease in the bid being larger.

On the empirical side, Biais, Hillion, and Spatt (1995) analyse the order flow of the CAC 40 stocks at the Paris Bourse on 19 trading days in 1991. The data set includes the first five bid and ask quotes. In particular, they analyse the order flow conditional on the state of the book (spread and number of shares offered at the quotes). They find that market order trades are relatively more frequent when the spread is tight, whereas new limit orders inside the spread are relatively more frequent when the spread is large. They also notice that when the volume at the quotes is large, more new limit orders are placed inside the spread. Conversely, when the volume at the quotes is low, more new limit order are placed at the quotes. This is explained by a tradeoff between undercutting the best quote to obtain time priority and queuing up the current quote. They also analyse the average time interval between two orders, conditional on the bid-ask spread. It turns out that the average time interval is lowest when the spread is very large.

Omura, Tanigawa, and Uno (2000) analyse order book data from 50 stocks from the Tokyo Stock Exchange in December 1998. The data set only contains the best ask/bid prices and volumes, but no quotes deeper inside the order book. Contrary to the Paris Bourse data analysed in Biais, Hillion, and Spatt (1995), Tokyo Stock Exchange has a different microstructure setting: the placement of hidden limit orders is not allowed, and hence it is closer to our modelling framework which does not include hidden limit orders either. The authors analyse the probability that a limit order placed at a certain time is executed by the end of the trading day. They formulate several hypotheses and test them using a probit model. Amongst others, it turns out that firstly, execution probabilities of sell limit orders are lower when the volume of the ask side of the book is higher. Secondly, they find that execution probabilities of sell limit orders are higher, when the volume of the bid side of the book is higher. Thirdly, execution probabilities are lower, when there are open ticks between the bid-ask spread (at least for those stocks which have a high trading volume).

Maslov and Mills (2001) study statistical properties of NASDAQ data. They observe that an imbalance of the volume on the best ask versus best bid causes a predictable change of the midquote price in the near future. When there is more volume on the best bid than on the best ask, the midquote price will increase in the near future, and vice-versa.

Ranaldo (2004) examines the relationship between the state of the limit order book and the aggressiveness of traders. Aggressiveness is defined by the trader's choice of placing orders: it ranges from very aggressive (placing a large market order) to little aggressive (placing a limit order at the best quote or cancelling an order). The empirical analysis is based on order book data of 15 stocks from the Swiss Stock Exchange in 1997, which is a pure, order-driven electronic market without market makers. The author comes to the conclusion that a large volume on the best bid increases the aggressiveness of a buyer. Moreover he finds that when the spread size widens, the aggressiveness of an order submission decreases.

Hall and Hautsch (2006) analyse order book data from the five most liquid stocks at the Australian Stock Exchange in 2002. Contrary to previous studies where order aggressiveness was just coded as a univariate variable, they use a multivariate model that allows for the possibility that market orders, limit orders and cancellations behave differently in their dependence on the order book state. They find that a high volume on the best ask decreases the intensity of aggressive buys and increases it for aggressive sells, as a high volume on the best ask reflects a higher proportion of volume to be sold at a comparatively low price. This negative price signal increases the market participants' preference to aggressively sell their positions by posting market orders. Moreover, a high volume on the best ask increases the probability of cancellations on the ask side. They also find that aggressive market trading (on both sides of the book) significantly decreases when the spread rises. Conversely, the aggressiveness of limit order traders increases. Moreover, they find weak evidence that a large spread implies fewer cancellations of limit orders on both sides of the book.

Summarizing and synthesizing the above theoretical and empirical findings on the order flow in limit orderbooks, we are led to formulate a number of assumptions concerning the impact of the state of the order book on the order flow rates: We first consider the effect on $\mathrm{d}X^{\mathfrak{A}}$, which we defined by :

$$X^{\mathfrak{A}}(t) := \underbrace{S^{\mathfrak{A}+}(t)}_{\text{sell LO placed inside spread}} - \underbrace{S^{\mathfrak{A}-}(t)}_{\text{sell LO inside spread cancelled}} - \underbrace{M^{\mathfrak{A}}(t)}_{\text{buy MO executed}}$$

X1 When the volume on the best ask is large, the amount of sell limit orders on the best ask is high, and buy traders can place buy market orders with only little market impact of market orders.

X2 When the volume on the best ask is large, sell traders want to undercut the current best ask and place sell limit orders inside the spread.

X3 When the spread is large, market orders are expensive compared to limit orders, and traders act less aggressively; they place fewer market orders and more limit orders inside the spread.

X4 When the spread is small, traders are more aggressive, as the price difference between market and limit orders is small. However, market orders are executed immediately, without risk of non-execution and therefore preferable. Hence, more market order will be submitted and fewer limit orders will be placed inside the spread.

X5 When the spread is equal to zero, no limit orders can be placed inside the spread, and traders can only cancel limit orders at the best quotes or submit market orders.

X6 When there is more volume on the best bid than on the best ask, there is an excess of supply of buy limit orders. Hence traders place more buy market orders and fewer sell limit orders inside the spread, or even cancel limit orders at the best ask.

X7 When there is more volume on the best ask than on the best bid, there is an excess of supply of ask limit orders. Hence traders place fewer buy market orders and more sell limit orders inside the spread.

Similarly, we consider the effect on $\mathrm{d}Y^{\mathfrak{A}}$, the rate of (limit orders placed) $-$ (limit orders cancelled) on the order book:

Y1 When the spread is large, executing market orders is expensive compared to limit orders. Hence traders place more limit orders on the order book.

Y2 When the spread is small, the price advantage of a limit order compared to a market order is small. However, market orders are executed immediately, without risk of non-execution. Hence traders cancel limit orders stored on the order book (and convert them to market orders).

Y3 When there is more volume on the best bid than on the best ask, there is an excess of supply of buy limit orders. Patient traders place more limit orders on the ask side of the order book.

Y4 When there is more volume on the best ask than on the best bid, there is an excess of supply of ask limit orders. Impatient traders cancel limit orders on the ask side of the order book.

Note that effects induced by the spread are the same for the bid side $X^{\mathfrak{B}}, Y^{\mathfrak{B}}$, whereas the effects induced by volume-imbalance are exactly opposite.

3.2 Choice of order flow processes

We will work in the block-shape model given by equation (2.24). For this model, we shall propose order flow processes $\left(X^{\mathfrak{A}}, Y^{\mathfrak{A}}, X^{\mathfrak{B}}, Y^{\mathfrak{B}}\right)$ which satisfy assumptions (X1)-(X7) and (Y1)-(Y4). As we work in the framework developed in chapter 2, the dynamics depend on the state of the order book which reduces to the key parameters best quotes α, β, volume $V^{\mathfrak{A}}, V^{\mathfrak{B}}$, spread $\mathfrak{s} = \alpha - \beta$ and volume-imbalance $\mathfrak{z} = \log \frac{V^{\mathfrak{B}}}{V^{\mathfrak{A}}}$.

Now define the order flow processes $\left(X^{\mathfrak{A}}, Y^{\mathfrak{A}}, X^{\mathfrak{B}}, Y^{\mathfrak{B}}\right)$ by the SDEs

$$\mathrm{d}X^{\mathfrak{A}}(t) = -\frac{V^{\mathfrak{A}}(t)}{2}\left(\kappa(\mu - \mathfrak{s}(t))\mathrm{d}t + (d_1 + d_2\mathfrak{z}(t))\mathrm{d}t + \sigma_1\sqrt{\mathfrak{s}(t)}\mathrm{d}W_1(t) + \sigma_0\mathrm{d}W_0(t)\right) \tag{3.1}$$

$$\mathrm{d}X^{\mathfrak{B}}(t) = -\frac{V^{\mathfrak{B}}(t)}{2}\left(\kappa(\mu - \mathfrak{s}(t))\mathrm{d}t - (d_1 + d_2\mathfrak{z}(t))\mathrm{d}t + \sigma_1\sqrt{\mathfrak{s}(t)}\mathrm{d}W_2(t) - \sigma_0\mathrm{d}W_0(t)\right) \tag{3.2}$$

$$\mathrm{d}Y^{\mathfrak{A}}(t) = -d_3(\nu - \mathfrak{z}(t))\mathrm{d}t - \delta(\mu - \mathfrak{s}(t))\mathrm{d}t + \sigma_2\mathrm{d}W_3(t) \tag{3.3}$$

$$\mathrm{d}Y^{\mathfrak{B}}(t) = d_4(\nu - \mathfrak{z}(t))\mathrm{d}t - \delta(\mu - \mathfrak{s}(t))\mathrm{d}t + \sigma_2\mathrm{d}W_4(t) \tag{3.4}$$

where W_i, $i = 0, 1, 2, 3, 4$ are independent standard Brownian motions, $\kappa, \delta, \mu, \sigma_0, \sigma_1, \sigma_2, d_i, \in \mathbb{R}_{\geq 0}$ for $i = 2, 3, 4$ and $d_1 \in \mathbb{R}$. Moreover we assume initial values $X^{\mathfrak{A}}(0) = Y^{\mathfrak{A}}(0) = X^{\mathfrak{B}}(0) = Y^{\mathfrak{B}}(0) = 0$.

To justify this choice, we first check that $\mathrm{d}X^{\mathfrak{A}}$ defined by equation (3.1) satisfies $(X1) - (X7)$:

X1: The rate of buy market order increases (linearly) in $V^{\mathfrak{A}}$; when more sell limit orders are available, more buy market orders will be matched against them.

X2: The rate of spread market order increases (linearly) in $V^{\mathfrak{A}}$; when more sell limit orders are available on the best ask, more limit orders are placed inside the spread to undercut the current best ask.

X3: When the spread is large (compared to the reference value μ), the term $-\kappa(\mu - \mathfrak{s}(t))\mathrm{d}t$ induces a positive drift. As we will see later, μ can be interpreted as the long-time mean of the spread, and therefore it is reasonable to say that the spread $\mathfrak{s}$ is large when it exceeds μ.

X4: Similarly, when the spread is small (compared to the reference value μ), the term $-\kappa(\mu - \mathfrak{s}(t))\mathrm{d}t$ adds a negative drift.

X5: Suppose we have a zero spread ($\mathfrak{s} = 0$) and the external drift $(d_1 + d_2\mathfrak{z}) = 0$, then $\mathrm{d}X^{\mathfrak{A}}(t) = -\frac{V^{\mathfrak{A}}(t)\kappa\mu}{2}\mathrm{d}t < 0$, and no limit orders are placed inside the spread.

X6: When the volume-imbalance $\mathfrak{z}$ is positive, $-d_2\mathfrak{z}(t)\mathrm{d}t$ induces a negative drift term.

X7: When the volume-imbalance $\mathfrak{z}$ is negative, $-d_2\mathfrak{z}(t)\mathrm{d}t$ induces a positive drift term.

In addition, there might be an extra drift induced by external market events and which does not depend on order book parameters. This external drift is represented by d_1.

Next, we check that $\mathrm{d}Y^{\mathfrak{A}}$ as defined in equation (3.3) satisfies (Y1)-(Y4):

Y1: When the spread is large (i.e. above μ), $-\delta(\mu - \mathfrak{s}(t))\mathrm{d}t$ adds a positive drift-term.

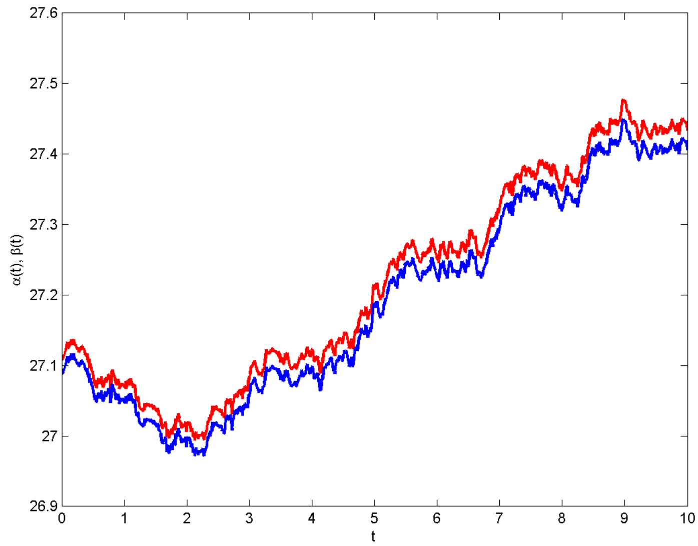
Y2: When the spread is small (i.e. below μ), $-\delta(\mu - \mathfrak{s}(t))\mathrm{d}t$ adds a negative drift-term.

Y3: When the volume imbalance is high (compared to the reference value ν), there is an excess of buy limit orders and the positive drift term $-d_3(\nu - \mathfrak{z}(t))\mathrm{d}t$ increases the volume of ask limit orders. As we will see, ν can be interpreted as the long-time mean of the volume imbalance.

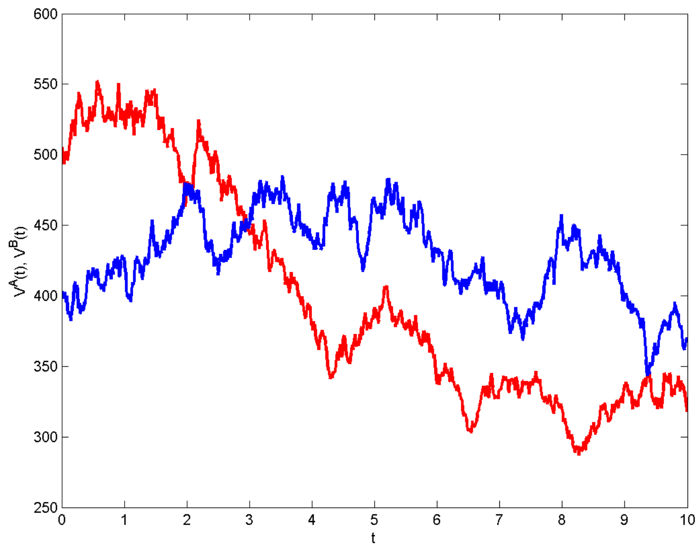
Y4: When the volume imbalance is small (compared to the reference value ν), there is an excess of ask limit orders (compared to average) and the negative drift term $-d_3(\nu - \mathfrak{z}(t))\mathrm{d}t$ decreases the volume of ask limit orders.

Thus the state-dependent behaviour of the order flow which can expressed in terms of the model key parameters that was noticed in theoretical models as well as empirical studies, and which we synthesized in (X1)-(X7) and (Y1)-(Y4) is reflected in our choice of order flow processes defined above.

Figure 3.1 shows a sample paths of (α, β) and $(V^{\mathfrak{A}}, V^{\mathfrak{B}})$ for $t \in [0, 10]$ with parameters $d_1 = 0.01, d_2 = 0.2, d_3 = d_4 = 0.1, \kappa = 0.1, \delta = 0.1, \mu = 0.03, \sigma_0 = 0.1, \sigma_1 = 0.01, \sigma_2 = 0.1$ and initial values $(\alpha(0), \beta(0), V^{\mathfrak{A}}(0), V^{\mathfrak{B}}(0)) = (27.11, 27.09, 500, 400)$.



(a) Sample path of best ask α (red) and best bid β (blue)



(b) Sample path of ask volume $V^{\mathfrak{A}}$ (red) and bid volume $V^{\mathfrak{B}}$ (blue)

Figure 3.1: Sample path of order book with dynamics given as in (3.1)- (3.4). Note that (i) the best ask is always above the best bid, so the spread is always positive, (ii) one clearly notes the drift induced by volume-imbalance; when there is more volume on the best bid (and thus more demand for the stock), the price is driven up.

Chapter 4

Analysis of order book

We will now analyse the order book constructed in chapter 3. We start with the basic properties, such as existence and consistency of the order book and moments of best ask/bid and volume imbalance. We will also give economic interpretations of each term appearing in the model equations.

Then we will move on to the analysis of the time-to-fill, i.e. the time it takes for a limit order to be executed. It will be analysed using a Dirichlet problem formulation, asymptotic analysis and via Monte Carlo methods.

4.1 Basic order book properties

Let us start by translating the order flow processes (3.1) - (3.4) in terms of dynamics for best ask/bid and volume. As we are working in the block-shape orderbook model, we can plug (3.1) - (3.4) into (2.25) and (2.26), and obtain the following SDEs for the key parameters $(\alpha, \beta, V^{\mathfrak{A}}, V^{\mathfrak{B}})$:

$$\begin{aligned}
\mathrm{d}\alpha(t) &= \frac{\kappa}{2}(\mu - \mathfrak{s}(t))\mathrm{d}t + \frac{1}{2}(d_1 + d_2\mathfrak{z}(t))\mathrm{d}t + \frac{\sigma_1}{2}\sqrt{\mathfrak{s}(t)}\mathrm{d}W_1(t) + \frac{\sigma_0}{2}\mathrm{d}W_0(t), && (4.1)\\
\mathrm{d}\beta(t) &= -\frac{\kappa}{2}(\mu - \mathfrak{s}(t))\mathrm{d}t + \frac{1}{2}(d_1 + d_2\mathfrak{z}(t))\mathrm{d}t - \frac{\sigma_1}{2}\sqrt{\mathfrak{s}(t)}\mathrm{d}W_2(t) + \frac{\sigma_0}{2}\mathrm{d}W_0(t), && (4.2)\\
\mathrm{d}V^{\mathfrak{A}}(t) &= V^{\mathfrak{A}}(t)\left(d_3\mathfrak{z}(t)\mathrm{d}t - \delta(\mu - \mathfrak{s}(t))\mathrm{d}t + \sigma_2\mathrm{d}W_3(t)\right), && (4.3)\\
\mathrm{d}V^{\mathfrak{B}}(t) &= V^{\mathfrak{B}}(t)\left(-d_4\mathfrak{z}(t)\mathrm{d}t - \delta(\mu - \mathfrak{s}(t))\mathrm{d}t + \sigma_2\mathrm{d}W_4(t)\right). && (4.4)
\end{aligned}$$

By theorem 2.1, we know that the orderbook exists. The block-shape order book model is then given by

$$\mathfrak{A}(t,p) = V^{\mathfrak{A}}(t)(p - \alpha(t)),$$
$$\mathfrak{B}(t,p) = V^{\mathfrak{B}}(t)(\beta(t) - p).$$

Moreover, the proposition below summarizes the basic properties of the orderbook model.

Proposition 4.1 (Properties of order book)**.** *Let $\alpha, \beta, V^{\mathfrak{A}}, V^{\mathfrak{B}}$ be given by (4.1) - (4.4) and initial values $\beta(0) \le \alpha(0)$, $V^{\mathfrak{A}}(0) > 0, V^{\mathfrak{B}}(0) > 0$. Then there exist four independent standard Brownian motions B_0, B_1, B_2, B_3 such that* *(i) The midquote price satisfies the SDE*

$$\mathrm{d}\mathfrak{m}(t) = \frac{1}{2}(d_1 + d_2\mathfrak{z}(t))\mathrm{d}t + \frac{\sigma_1}{2\sqrt{2}}\sqrt{\mathfrak{s}(t)}\mathrm{d}B_2(t) + \frac{\sigma_0}{2}\mathrm{d}B_0(t). \tag{4.5}$$

(ii) The spread is a mean-reverting Cox-Ingersoll-Ross (CIR) process satisfying the SDE

$$\mathrm{d}\mathfrak{s}(t) = \kappa(\mu - \mathfrak{s}(t))\mathrm{d}t + \frac{\sigma_1}{\sqrt{2}}\sqrt{\mathfrak{s}(t)}\mathrm{d}B_1(t). \tag{4.6}$$

(iii) The volume-imbalance is a mean-reverting Ornstein-Uhlenbeck (OU) process satisfying the SDE

$$\mathrm{d}\mathfrak{z}(t) = \rho(\nu - \mathfrak{z}(t))\mathrm{d}t + \frac{\sigma_2}{\sqrt{2}}\mathrm{d}B_3(t), \tag{4.7}$$

where $\rho = d_3 + d_4 \geq 0$.

(iv) There exist unique strong solutions for $\alpha, \beta, V^{\mathfrak{A}}, V^{\mathfrak{B}}$.

(v) The model is consistent, that is a.s. for all $t \geq 0$, $\alpha(t) \geq \beta(t)$. *Moreover, if* $4\kappa\mu \geq \sigma_1^2$, *then a.s. for all* $t \geq 0$, $\alpha(t) > \beta(t)$.

(vi) We have closed-form expressions for mean and variance of best ask/bid, spread, midquote-price and volume imbalance:

$$\begin{aligned}
\mathrm{E}[\alpha(t)] &= \frac{\alpha(0)+\beta(0)}{2} + \frac{1}{2}d_1 t + \frac{d_2}{2\rho}\left((\mathfrak{z}(0)-\nu)\left(1-e^{-\rho t}\right) + \nu\rho t\right) \\
&\quad + e^{-\kappa t}\frac{\alpha(0)-\beta(0)}{2} + \frac{\mu}{2}\left(1-e^{-\kappa t}\right) \\
\mathrm{E}[\beta(t)] &= \frac{\alpha(0)+\beta(0)}{2} + \frac{1}{2}d_1 t + \frac{d_2}{2\rho}\left((\mathfrak{z}(0)-\nu)\left(1-e^{-\rho t}\right) + \nu\rho t\right) \\
&\quad - e^{-\kappa t}\frac{\alpha(0)-\beta(0)}{2} - \frac{\mu}{2}\left(1-e^{-\kappa t}\right) \\
\mathrm{Var}(\alpha(t)) &= \mathrm{Var}(\beta(t)) \\
&= \frac{d_2^2\sigma_2^2}{16\rho^2}\left(2\rho t - 3 + 4e^{-\rho t} - e^{-2\rho t}\right) \\
&\quad + \frac{\sigma_1^2}{8\kappa}\left(t\kappa\mu + (\mathfrak{s}(0)-\mu)(1-e^{-t\kappa})\right) \\
&\quad + \frac{\sigma_1^2}{8\kappa}\mathfrak{s}(0)\left(e^{-\kappa t} - e^{-2\kappa t}\right) + \frac{\mu\sigma_1^2}{16\kappa}\left(1 - 2e^{-\kappa t} + e^{-2\kappa t}\right) + \frac{\sigma_0^2}{4}t
\end{aligned}$$

$$
\begin{aligned}
\mathrm{E}[\mathfrak{s}(t)] &= e^{-\kappa t}\mathfrak{s}(0) + \mu\left(1 - e^{-\kappa t}\right) \\
\mathrm{Var}(\mathfrak{s}(t)) &= \frac{\sigma_1^2}{2\kappa}\mathfrak{s}(0)\left(e^{-\kappa t} - e^{-2\kappa t}\right) + \frac{\mu\sigma_1^2}{4\kappa}\left(1 - 2e^{-\kappa t} + e^{-2\kappa t}\right) \\
\mathrm{E}[\mathfrak{m}(t)] &= \mathfrak{m}(0) + \frac{1}{2}d_1 t + \frac{d_2}{2\rho}\left((\mathfrak{z}(0) - \nu)\left(1 - e^{-\rho t}\right) + \nu\rho t\right) \\
\mathrm{Var}(\mathfrak{m}(t)) &= \frac{d_2^2\sigma_2^2}{16\rho^2}\left(2\rho t - 3 + 4e^{-\rho t} - e^{-2\rho t}\right) + \frac{\sigma_0^2}{4}t \\
&\quad + \frac{\sigma_1^2}{8\kappa}\left(t\kappa\mu + (\mathfrak{s}(0) - \mu)(1 - e^{-t\kappa})\right) \\
\mathrm{E}[\mathfrak{z}(t)] &= e^{-\rho t}\mathfrak{z}(0) + \nu(1 - e^{-\rho t}) \\
\mathrm{Var}(\mathfrak{z}(t)) &= \frac{\sigma_2^2}{4\rho}\left(1 - e^{-2\rho t}\right)
\end{aligned}
$$

Let us consider the different components that drive the dynamics of the best ask:

$$
\mathrm{d}\alpha(t) = \underbrace{\frac{\kappa}{2}(\mu - \mathfrak{s}(t))\mathrm{d}t}_{\text{(D1)}} + \underbrace{\frac{1}{2}d_2\mathfrak{z}(t)\mathrm{d}t}_{\text{(D2)}} + \underbrace{\frac{1}{2}d_1\mathrm{d}t}_{\text{(D3)}} + \underbrace{\frac{\sigma_1}{2}\sqrt{\mathfrak{s}(t)}\mathrm{d}W_1(t)}_{\text{(V1)}} + \underbrace{\frac{\sigma_0}{2}\mathrm{d}W_0(t)}_{\text{(V2)}}
$$

The five components can be interpreted in the following way

(D1) is a microstructure drift term, that is endogenously given by the spread of the stock. A large spread pulls the best ask down, whereas a small spread pushes it up. The mean-reversion property of the spread ensures that the sign of this drift term alternates constantly. It depends on the current state of the order book, but not on external events.

(D2) is another microstructure drift term that results from volume-imbalance. A positive volume imbalance pushes the best ask up, since there is an excess of supply of buy offers. The mean-reversion property of the volume imbalance ensures that the sign of this drift term alternates constantly. As (D1), it depends on the current state of the order book, but not on external events.

(D3) is an exogenously given drift term which describes the general market trend. It is influenced by external events and news, and does not depend on the current microstructure of the order book.

(V1) is a microstructure volatility term, that is endogenously given by the spread of the stock. The larger the spread, the larger the volatility. A zero spread corresponds to no volatility contribution from the microstructure of the order book.

(V2) is an exogenously given volatility term. It captures the volatility which does not depend on the current state of the order book, but comes from external risk sources.

The effect of (D1) and (D2) were already noted in the discussion at the beginning of section 3.1. It is clear that the price drift cannot depend solely on endogenous terms (spread, volume imbalance) of the order book microstructure, but also depends on external events and market trends. This motives the existence of (D3).

It has long been noticed that a high (total) volatility of an asset results in higher spreads: Harris (2003) notes that since volatility increases limit order option values, traders widen their spreads, when trading volatile instruments to minimize the value of the timing option.

Bollerslev and Melvin (1994) examine the relationship between bid-ask spreads for exchange-rate quotes and the volatility of the underlying exchange-rate process. They find a positive relationship between latent volatility and observed spreads in the Deutschemark/dollar foreign exchange market. This is captured by the microstructure volatility term (V1): when a higher (total) volatility is observed, this effect must stem from a higher spread, as the exogenous volatility (V2) is assumed to be constant.

As in the drift case, the exogenous volatility (V2) is explained by external risks that are independent of the risk associated to the order book microstructure (e.g. the risk of crossing a large spread, instead of waiting for the limit order).

Note that the midquote price

$$\mathrm{d}\mathfrak{m}(t) = \underbrace{\frac{1}{2}d_{2\mathfrak{z}}(t)\mathrm{d}t}_{(D2)} + \underbrace{\frac{1}{2}d_1\mathrm{d}t}_{(D3)} + \underbrace{\frac{\sigma_1}{2\sqrt{2}}\sqrt{\mathfrak{s}(t)}\mathrm{d}B_2(t)}_{(V1)} + \underbrace{\frac{\sigma_0}{2}\mathrm{d}B_0(t)}_{(V2)}$$

has the same components as the best ask, except for the microstructure drift term stemming from the spread. This is due to the assumption that best bid and ask have exactly opposite spread behaviors which neutralize each other and thus vanish in the midquote price. Moreover the volatility components (V1) and (V2) are now driven by other Browian motions, which are correlated with the Brownian motions driving the best ask.

4.2 Time-to-fill

4.2.1 Problem definition

A trader who wants to buy or sell assets in a limit order market can choose between two main types of orders as part of her trading strategy: market or limit orders.

The market order is executed immediately, however its price is usually worse than that of a limit order, especially for large orders and in volatile markets. In chapter 6 we will analyse optimal trading strategies with market orders, in the framework of the model described in this chapter.

The other important order type is a limit order: when such an order is submitted the trader offers to buy or sell a prespecified number of shares at a prespecified price. In particular, this imlies that there is no price risk associated to the limit order. On the downside, there is a risk of non-execution, and even if the limit order is executed at some time, this time is random and may be too long for an impatient trader.

To balance the price risk of a market order against the risk of non-execution within a certain time horizon of a limit order, it is crucial to analyse the *time-to-fill* of a limit order; when a limit order is placed at time 0 at a certain price level $p \geq \alpha(0)$, this is the time it takes until a market order is executed against it. Define $\hat{\tau}_p$ to be the time-to-fill of a limit order placed at price level p. As an approximation to the time-to-fill we shall consider the first-passage time

$$\tau_p = \inf\{t \geq 0 : \alpha(t) \geq p\}$$

The first-passage time provides lower and upper bounds for the time-to-fill:

$$\tau_p \leq \hat{\tau}_p \leq \tau_{p+\Delta p} \tag{4.8}$$

where $\Delta p > 0$ denotes the tick size of the order book. In the following, we will call τ_p the time-to-fill, as this is the best approximation we can make within our modelling framework.

In order to simplify the analyis, we consider the two-dimensional model given by

$$\begin{aligned} \mathrm{d}\alpha(t) &= \frac{\kappa}{2}(\mu - \mathfrak{s}(t))\mathrm{d}t + \frac{d_1}{2}\mathrm{d}t + \frac{\sigma_1}{2}\sqrt{\mathfrak{s}(t)}\mathrm{d}W_1(t) + \frac{\sigma_0}{2}\mathrm{d}W_0(t), \\ \mathrm{d}\beta(t) &= -\frac{\kappa}{2}(\mu - \mathfrak{s}(t))\mathrm{d}t + \frac{d_1}{2}\mathrm{d}t - \frac{\sigma_1}{2}\sqrt{\mathfrak{s}(t)}\mathrm{d}W_2(t) + \frac{\sigma_0}{2}\mathrm{d}W_0(t), \end{aligned}$$

.

It is obtained by setting the volume-imbalance parameter $d_2 = 0$ in the general model equations (3.1) - (3.4).

For the ease of reference, we also write down the dynamics of the equivalent model in terms of midquote-price and spread:

$$\begin{aligned} \mathrm{d}\mathfrak{m}(t) &= \frac{d_1}{2}\mathrm{d}t + \frac{\sigma_1}{2\sqrt{2}}\sqrt{\mathfrak{s}(t)}\mathrm{d}B_2(t) + \frac{\sigma_0}{2}\mathrm{d}B_0(t), \\ \mathrm{d}\mathfrak{s}(t) &= \kappa(\mu - \mathfrak{s}(t))\mathrm{d}t + \frac{\sigma_1}{\sqrt{2}}\sqrt{\mathfrak{s}(t)}\mathrm{d}B_1(t), \end{aligned}$$

where B_0, B_1, B_2 are independent standard Brownian motions.

4.2.2 Formulation as Dirichlet problem and viscosity solutions

As we shall see below, the distribution of τ_p can be computed by solving a Dirichlet problem. Since the process $X = (\mathfrak{m}, \mathfrak{s})$ is not uniformly elliptic, we will use the notion of viscosity solutions.

Let

$$(\mathscr{L}u)(m,s) = \frac{d_1}{2}u_m(m,s) + \frac{\sigma_0^2}{8}u_{mm}(m,s) + \frac{\sigma_1^2}{16}su_{mm}(m,s) + \kappa(\mu - s)u_s(m,s) + \frac{\sigma_1^2}{4}su_{ss}(m,s).$$

be the infinitesimal generator of X, and consider the *generalised Dirichlet problem*

$$\mathscr{L}u - \lambda u + h = 0 \quad \text{in } D, \tag{4.9}$$

$$u = c \quad \text{in } \Gamma, \tag{4.10}$$

where $D = \left\{(m,s) : m + \frac{s}{2} < p, s \geq 0\right\} \subseteq \mathbb{R}^2$ and $\Gamma = \left\{(m,s) : m + \frac{s}{2} = p, s \geq 0\right\}$, c is a constant, and $h : \bar{D} \mapsto \mathbb{R}$ is a continuous function. We can now define

Definition 4.1 (Viscosity solution)**.** (i) A function $u : \bar{D} \mapsto \mathbb{R}$ is a *viscosity subsolution* of (4.9)-(4.10) if

$$\begin{aligned} -\left(\mathscr{L}\phi(x_0) - \lambda u^*(x_0) + h(x_0)\right) \leq 0 &\quad \text{holds for } x_0 \in D, \\ \min\left\{-\left(\mathscr{L}\phi(x_0) - \lambda u^*(x_0) + h(x_0)\right), u^*(x_0) - c\right\} \leq 0 &\quad \text{holds for } x_0 \in \partial D, \end{aligned}$$

and all $\phi \in C^2(\mathbb{R}^2)$ such that $u^* - \phi$ attains a global maximum on $\bar{D}$ at x_0, and $u^*(x) = \limsup_{y \to x} u(y)$.

(ii) A function $u : D \mapsto \mathbb{R}$ is a *viscosity supersolution* of (4.9)- (4.10) if

$$
\begin{aligned}
-\left(\mathscr{L}\phi(x_0) - \lambda u_*(x_0) + h(x_0)\right) \geq 0 \quad & \text{holds for } x_0 \in D, \\
\max\left\{-\left(\mathscr{L}\phi(x_0) - \lambda u_*(x_0) + h(x_0)\right), u_*(x_0) - c\right\} \geq 0 \quad & \text{holds for } x_0 \in \partial D,
\end{aligned}
$$

and all $\phi \in C^2(\mathbb{R}^2)$ such that $u_* - \phi$ attains a global minimum on $\bar{D}$ at x_0, and $u_*(x) = \liminf_{y \to x} u(y)$.

(iii) A function $u : D \mapsto \mathbb{R}$ is a *viscosity solution* of (4.9)- (4.10) if it is both a viscosity sub- and supersolution of (4.9)- (4.10).

For a general introduction to viscosity solutions and a motivation of the above definition, we refer to Crandall, Ishii, and Lions (1992).

We now have

Proposition 4.2 (Time-to-fill). *Suppose the order book model is given by* (4.5) - (4.6) *with* $d_2 = 0$ *and* $\sigma_0 > 0$. *For* $\lambda > 0$, *define the Laplace transform of the time-to-fill*

$$u(m, s) = \mathrm{E}^{(m,s)}\left[e^{-\lambda \tau_p}\right]$$

where $\mathrm{E}^{(m,s)}$ *denotes the expectation conditional on* $(\mathfrak{m}(0), \mathfrak{s}(0)) = (m, s)$. *Then* u *is a viscosity solution of the system of equations*

$$
\begin{aligned}
\mathscr{L}u - \lambda u = 0 \quad & \text{in } D, \\
u = 1 \quad & \text{in } \Gamma.
\end{aligned}
$$

4.3 Asymptotic analysis of time-to-fill

4.3.1 Motivation

In practice, the trader (or an algorithm within a microtrader) has to decide within milli-seconds if a limit order or a market order should be placed. This requires fast estimation of the risk associated to a limit order, and thus efficient assessment of the time-to-fill. One approach to compute the time-to-fill is via the Dirichlet problem described in propositon 4.2 or, alternatively, Monte Carlo simulations can be used to estimate τ_p.

Both methods, however, require numerical schemes which are potentially computationally intense. Due to the extreme time constraint in which a decision about market vs. limit order has to be taken, both methods might be too slow to be put in practice.

In this section we will therefore perform an asymptotic analysis of the time-to-fill with the objective of finding a computationally efficient (asymptotic) approximation of the time-to-fill, which is suitable for real-world applications. The idea of this approximation is mainly inspired by the work of Fouque, Papanicolaou, and Sircar (2000).

We will work with a modified version of equations (4.5)- (4.6), obtained by multiplying the mean-reversion term (D1) by factor ϵ and the microstructure volatility term (V1) by factor $\sqrt{\epsilon}$ (see discussion at the end of section 4.1):

$$\mathrm{d}\mathfrak{m}^{(\epsilon)}(t) = \frac{d_1}{2}\mathrm{d}t + \sqrt{\epsilon}\frac{\sigma_1}{2\sqrt{2}}\sqrt{\mathfrak{s}^{(\epsilon)}(t)}\mathrm{d}B_2(t) + \frac{\sigma_0}{2}\mathrm{d}B_0(t), \tag{4.11}$$

$$\mathrm{d}\mathfrak{s}^{(\epsilon)}(t) = \epsilon\kappa(\mu - \mathfrak{s}^{(\epsilon)}(t))\mathrm{d}t + \sqrt{\epsilon}\frac{\sigma_1}{\sqrt{2}}\sqrt{\mathfrak{s}^{(\epsilon)}(t)}\mathrm{d}B_1(t), \tag{4.12}$$

where B_0, B_1, B_2 are independent standard Brownian motions.

For $\epsilon = 0$, we have $\mathrm{d}\mathfrak{s}^{(\epsilon)}(t) = 0$ and thus the spread is constant at all times. Therefore $\mathfrak{m}^{(\epsilon)}$ is just a Brownian motion with drift. Hence we are in the setting of the one-dimensional *Bachelier model*, given by an arithmetic Brownian motion with constant drift and constant volatility, and a spread that stays constant at all times.

For $\epsilon = 1$, we retrieve the original two-dimensional model from subsection 4.2.1

We will analyse the model for small values of ϵ which corresponds to slow spread dynamics: While the midquote-price corresponds to the 'fair' value of the stock and changes frequently with incoming market information, the spread reflects some part of the liquidity risk (more precisely, the costs of a 'round-trip', i.e. buying and selling a share at the same time). For many stocks, the spread stays constant most of the time which corresponds to small values of ϵ.

As before, we consider the time-to-fill

$$\tau_p^{(\epsilon)} = \inf\left\{t \geq 0 : \alpha^{(\epsilon)}(t) \geq p\right\} = \inf\left\{t \geq 0 : \mathfrak{m}^{(\epsilon)}(t) + \mathfrak{s}^{(\epsilon)}(t)/2 \geq p\right\},$$

and its Laplace transform $\mathrm{E}^{(m,s)}\left[e^{-\lambda\tau_p^{(\epsilon)}}\right]$.

Note that for $\epsilon = 0$, $\tau_p^{(\epsilon)}$ corresponds to the first hitting time of a Brownian motion with constant volatility and constant drift. There exists a well-known explicit formula for the Laplace transform of $\tau_p^{(\epsilon)}$.

For $\epsilon = 1$, however, no explicit formula is known for the Laplace transform.

The idea is to make an asymptotic expansion of $\mathrm{E}^{(m,s)}\left[e^{-\lambda\tau_p^{(\epsilon)}}\right]$ for small values of ϵ and hope for an explicit formula in this expansion.

4.3.2 Modified Dirichlet problem

Just as in propositon 4.2, we start with the Dirichlet problem

$$\mathscr{L}^{(\epsilon)}u^{(\epsilon)} = \lambda u^{(\epsilon)} \qquad \text{in } D, \tag{4.13}$$

$$u^{(\epsilon)} = 1 \qquad \text{on } \partial D, \tag{4.14}$$

where $\lambda > 0$,

$$D = \{(m, s) : m \in \mathbb{R}, s \in \mathbb{R}_+, m + s/2 < p\},$$

and $\mathscr{L}^{(\epsilon)}$ is the infinitesimal generator of the process $(\mathfrak{m}^{(\epsilon)}(t), \mathfrak{s}^{(\epsilon)}(t))$

We are looking for solutions u bounded in m. By lemma 4.1, we know that the Laplace transform of the time-to-fill solves the Dirichlet problem (4.13)- (4.14) in the viscosity sense:

$$u^{(\epsilon)}(m,s) = \mathrm{E}^{(m,s)}\left[e^{-\lambda \tau_p^{(\epsilon)}}\right]$$

Next, we note that the operator $\mathscr{L}^{(\epsilon)}$ can be written as

$$\mathscr{L}^{(\epsilon)} = \mathscr{L}^{(0)} + \epsilon \mathscr{L}^{(1)} \tag{4.15}$$

where

$$(\mathscr{L}^{(0)}u)(m,s) = \frac{d_1}{2}u_m(m,s) + \frac{\sigma_0^2}{8}u_{mm}(m,s), \tag{4.16}$$

$$(\mathscr{L}^{(1)}u)(m,s) = \frac{\sigma_1^2}{16}su_{mm}(m,s) + \kappa(\mu - s)u_s(m,s) + \frac{\sigma_1^2}{4}su_{ss}(m,s) \tag{4.17}$$

are two infinitesimal generators independent of ϵ.

4.3.3 Asymptotic approximation

Asymptotic expansion

We now formally expand the Lapace transform $u^{(\epsilon)}(m,s)$ in powers of ϵ:

$$u^{(\epsilon)}(m,s) = u^{(0)}(m,s) + \epsilon u^{(1)}(m,s) + \epsilon^2 u^{(2)}(m,s) + \cdots . \tag{4.18}$$

We are interested in the approximation up to order one, i.e.

$$u^{(\epsilon)}(m,s) \approx u^{(0)}(m,s) + \epsilon u^{(1)}(m,s).$$

Plugging (4.18) and (4.15) into (4.13), we formally obtain the asymptotic expansion (up to order 1) of the Dirichlet problem

$$\mathscr{L}^{(0)}u^{(0)} + \epsilon\mathscr{L}^{(1)}u^{(0)} + \epsilon\mathscr{L}^{(0)}u^{(1)} + \cdots = \lambda\left(u^{(0)} + \epsilon u^{(1)} + \cdots\right) \tag{4.19}$$

By comparing coefficients of different powers of ϵ, we deduce differential equations for the terms up to order one:

$$\begin{aligned} \mathscr{L}^{(0)}u^{(0)} &= \lambda u^{(0)} \quad \text{(order zero)} \\ \mathscr{L}^{(0)}u^{(1)} + \mathscr{L}^{(1)}u^{(0)} &= \lambda u^{(1)} \quad \text{(order one)} \end{aligned}$$

We also need to specify the boundary conditions, which need to be in line with (4.14) and (4.18). Since we would like the zero order term $u^{(0)}$ to correspond to the Laplace transform of the time-to-fill in the one-dimensional Bachelier model ($\epsilon = 0$), we choose $u^{(0)} = 1$ on Γ. Given this choice for $u^{(0)}$, the boundary conditions for the first order term are now imposed: $u^{(1)} = 0$ on Γ.

Zero order term

We want to solve the Dirichlet problem

$$\mathscr{L}^{(0)} u^{(0)} = \lambda u^{(0)} \qquad \text{in } D,$$
$$u^{(0)} = 1 \qquad \text{on } \Gamma.$$

This second-order ODE is known to have a unique solution bounded in m given explicitly by

$$u^{(0)}(m,s) = \exp\left\{-\left(p - m - \frac{s}{2}\right)\frac{2}{\sigma_0^2}\left(\sqrt{d_1^2 + 2\lambda\sigma_0^2} - d_1\right)\right\} \tag{4.20}$$

Cleary, the zero order term corresponds to the Laplace transform of the hitting time of a Brownian motion with drift $d_1/2$ and volatility $\sigma_0/2$, i.e.

$$u^{(0)}(m,s) = \mathrm{E}^{(m,s)}\left[e^{-\lambda\tau_p^{(0)}}\right],$$

where

$$\tau_p^{(0)} = \inf\left\{t \geq 0 : \mathfrak{m}^{(0)}(t) \geq p - \mathfrak{s}^{(0)}(t)/2\right\}.$$

First order term

The first order term solves the Dirichlet problem

$$\mathscr{L}^{(0)} u^{(1)} + \mathscr{L}^{(1)} u^{(0)} = \lambda u^{(1)} \qquad \text{in } D,$$
$$u^{(1)} = 0 \qquad \text{on } \Gamma.$$

Note that this is again a second order ODE in $u^{(1)}$ which can be solved explicitly, and has the unique solution bounded in m

$$u^{(1)}(m,s) = \left(p - m - \frac{s}{2}\right)\frac{2}{\sqrt{d_1^2 + 2\lambda\sigma_0^2}}\left(c_0 + c_1 s\right) u^{(0)}(m,s), \tag{4.21}$$

where

$$c_0 = \frac{\kappa\mu}{\sigma_0^2}\left(\sqrt{d_1^2 + 2\lambda\sigma_0^2} - d_1\right)$$
$$c_1 = \frac{1}{\sigma_0^4}\left\{d_1^2\sigma_1^2 + \sigma_0^2\left(\lambda\sigma_1^2 - \kappa\sqrt{d_1^2 + 2\lambda\sigma_0^2}\right) + d_1\left(\kappa\sigma_0^2 - \sigma_1^2\sqrt{d_1^2 + 2\lambda\sigma_0^2}\right)\right\}$$

are two constants.

4.3.4 Error estimation

We want to estimate the accuracy of the approximation

$$u^{(\epsilon)}(m,s) \approx u^{(0)}(m,s) + \epsilon u^{(1)}(m,s).$$

We therefore write

$$u^{(\epsilon)}(m,s) = u^{(0)}(m,s) + \epsilon u^{(1)}(m,s) - R^{(\epsilon)}(m,s), \tag{4.22}$$

where $R^{(\epsilon)}$ denotes the remainder or error term.

Our objective is to bound the error term and show that it is of order ϵ^2.

Instead of giving a bound for the error term directly, we will derive a set of equations that $R^{(\epsilon)}$ satisfies and compute a viscosity solution. We will then give bounds on the viscosity solution. In particular, this implies that, whenever there is a unique solution for the viscosity solution, the bounds will also be true for $R^{(\epsilon)}$.

Proposition 4.3 (Error estimate)**.** *Let $\epsilon \in [0,1]$. Then $R^{(\epsilon)}$ satisfies a system of equations which admits a viscosity solution $\tilde{R}^{(\epsilon)}$ such that there exists a constant $c(s) > 0$ (independent of ϵ and m) with*

$$|\tilde{R}^{(\epsilon)}(m,s)| \leq \epsilon^2 c(s). \tag{4.23}$$

More precisely, we have

$$\begin{aligned}|\tilde{R}^{(\epsilon)}(m,s)| \leq & \frac{\epsilon^2}{4\sqrt{d_1^2 + 2\lambda\sigma_0^2}} \times \\ & \left\{ a_0 \frac{1}{\lambda} + a_1 \frac{s\lambda + \epsilon\kappa\mu}{\lambda^2 + \epsilon\kappa\lambda} + a_2 \frac{2s^2\lambda^2 + s\epsilon\lambda(2s\kappa + 4\kappa\mu + \sigma_1^2) + \epsilon^2\kappa\mu(4\kappa\mu + \sigma_1^2)}{2\lambda^3 + 6\epsilon\kappa\lambda^2 + 4\epsilon^2\kappa^2} \right\},\end{aligned} \tag{4.24}$$

where

$$a_i = \left| k_i + l_i e^{-1} \frac{\sigma_0^2}{\sqrt{d_1^2 + 2\lambda\sigma_0^2} - d_1} \right|, \quad i = 0,1,2 \tag{4.25}$$

are positive constants and the k_i, l_i are given by

$$k_0 = \frac{4\kappa^2\mu^2}{\sigma_0^2}\left(\sqrt{d_1^2+2\lambda\sigma_0^2}-d_1\right) \tag{4.26}$$

$$\begin{aligned} k_1 =& \frac{2}{\sigma_0^4}\left\{-d_1^2\left(6\kappa\mu\sigma_1^2+\sigma_1^4\right)+\right.\\ & \sigma_0^2\left(4\kappa^2\mu\sqrt{d_1^2+2\lambda\sigma_0^2}+\kappa\sigma_1^2\left(\sqrt{d_1^2+2\lambda\sigma_0^2}-6\lambda\mu\right)-\lambda\sigma_1^4\right)-\\ & \left.d_1\left(4\kappa^2\mu\sigma_0^2+\kappa\sigma_1^2\left(\sigma_0^2-6\mu\sqrt{d_1^2+2\lambda\sigma_0^2}\right)-\sigma_1^4\sqrt{d_1^2+2\lambda\sigma_0^2}\right)\right\} \end{aligned} \tag{4.27}$$

$$\begin{aligned} k_2 =& \left(\frac{4\kappa}{\sigma_0^4}-\frac{4\sigma_1^2\left(\sqrt{d_1^2+2\lambda\sigma_0^2}-d_1\right)}{\sigma_0^6}\right)\times\\ & \left(d_1^2\sigma_1^2+\sigma_0^2\left(\lambda\sigma_1^2-\kappa\sqrt{d_1^2+2\lambda\sigma_0^2}\right)+d_1\left(\kappa\sigma_0^2-\sqrt{d_1^2+2\lambda\sigma_0^2}\right)\right) \end{aligned} \tag{4.28}$$

$$\begin{aligned} l_0 =& \frac{4\kappa\mu}{\sigma_0^4}\left\{d_1^2(2\kappa\mu+\sigma_1^2)+\sigma_0^2\left(2\kappa\lambda\mu-\kappa\sqrt{d_1^2+2\lambda\sigma_0^2}+\lambda\sigma_1^2\right)+\right.\\ & \left.d_1\left(\kappa\left(\sigma_0^2-2\mu\sqrt{d_1^2+2\lambda\sigma_0^2}\right)-\sigma_1^2\sqrt{d_1^2+2\lambda\sigma_0^2}\right)\right\} \end{aligned} \tag{4.29}$$

$$\begin{aligned} l_1 =& \frac{2}{\sigma_0^6}\left(-2\kappa\sigma_0^2+(4\kappa\mu+\sigma_1^2)\left(\sqrt{d_1^2+2\lambda\sigma_0^2}-d_1\right)\right)\times\\ & \left(-d_1^2\sigma_1^2+\sigma_0^2\left(\kappa\sqrt{d_1^2+2\lambda\sigma_0^2}-\lambda\sigma_1^2\right)+d_1\left(\sigma_1^2\sqrt{d_1^2+2\lambda\sigma_0^2}-\kappa\sigma_0^2\right)\right) \end{aligned} \tag{4.30}$$

$$l_2 = -\frac{4}{\sigma_0^8}\left(d_1^2\sigma_1^2+\sigma_0^2\left(\lambda\sigma_1^2-\kappa\sqrt{d_1^2+2\lambda\sigma_0^2}\right)+d_1\left(\kappa\sigma_0^2-\sigma_1^2\sqrt{d_1^2+2\lambda\sigma_0^2}\right)\right)^2 \tag{4.31}$$

Remark 4.1. If we could show uniqueness in the class of functions bounded in m and with polynomial growth in s, we could conclude that $\tilde{R}^{(\epsilon)} = R^{(\epsilon)}$. However, the unboundedness of the domain makes it hard to establish such a result. For bounded domains, there are many uniqueness results available, see e.g. Barles and Burdeau (1995, Thm. 4.1) and Díaz (2004, Thm. 4.1). When it comes to practical applications, we can of course always restrict to bounded domains.

4.3.5 Discussion of asymptotic approximation

Interpretation of asymptotic formula

We will now give an interpretation of the asymptotic approximation of the Laplace transform of the time-to-fill

$$\begin{aligned} \mathrm{E}^{(m,s)}\left[e^{-\lambda\tau_p^{(\epsilon)}}\right] =& u^{(\epsilon)}(m,s)\\ \approx& \underbrace{u^{(0)}(m,s)}_{\text{zero order with constant spread}} + \underbrace{\epsilon\times u^{(1)}(m,s)}_{\text{first order correction of stochastic spread}} \end{aligned} \tag{4.32}$$

As already noted in the discussion following equation (4.20) of the zero order term, $u^{(0)}$ corresponds to an orderbook model with fixed spread. In this one-dimensional model (which corresponds to the Bachelier model with a constant spread), the distribution of the time-to-fill is known explicitly. It only depends on the drift d_1 and the exogeneously given volatility term σ_0 of the midquote-price.

The more interesting part of the approximation (4.32) is the first order correction term, which appears when the dynamics of the spread act on the orderbook model. We can rewrite the first order term as

$$u^{(1)}(m,s) = \underbrace{c_1\left(p - m - \frac{s}{2}\right)u^{(0)}(m,s)}_{\text{tick factor}}\left\{\underbrace{\kappa(\mu - s)c_2}_{\text{spread mean-reversion}} + \underbrace{sc_3}_{\text{spread volatility}}\right\}, \tag{4.33}$$

where

$$c_1 = \frac{2}{\sigma_0^2}\sqrt{d_1^2 + 2\lambda\sigma_0^2} \geq 0, \tag{4.34}$$

$$c_2 = \sqrt{d_1^2 + 2\lambda\sigma_0^2} - d_1 \geq 0, \tag{4.35}$$

$$c_3 = \frac{\sigma_1^2}{\sigma_0^2}\left(\lambda\sigma_0^2 - d_1\left(\sqrt{d_1^2 + 2\lambda\sigma_0^2} - d_1\right)\right) \geq 0. \tag{4.36}$$

Let us now analyse the three parts of the zero order correction term:

The *tick factor* is always positive, and decreases to 0 as the initial best ask $\alpha(0) = m + \frac{s}{2}$ increases to the price level p where the limit order has been placed. Thus is depends on the distance (i.e. number of ticks) between the initial best ask and the limit order:

- When the distance is small, the influence of the spread is small, as the best ask is likely to immediately hit the price level p, irrespective of the dynamics of the spread.
- The larger the distance, the more weight is assigned to effects induced by the initial spread.

Hence this weight is called *tick factor*, and multiplies the sum of the other two terms which describe how the spread affects the time-to-fill.

The *spread mean-reversion* captures the effect of the mean-reversion dynamics of $\mathfrak{s}(t)$ on the time-to-fill:

- When $s < \mu$, the spread is pulled up, and the best ask is likely to reach price level p faster. Since $c_2 \geq 0$ this corresponds to a larger $u^{(1)}$ and thus a smaller $\tau_p^{(\epsilon)}$, as expected.
- When $s > \mu$, the spread is pulled down, and the best ask is likely to reach price level p slower. Since $c_2 \geq 0$ this corresponds to a smaller $u^{(1)}$ and thus a larger $\tau_p^{(\epsilon)}$.
- A larger mean-reversion speed κ corresponds to a higher mean-reversion effect on the time-to-fill.

The *spread volatility* measures the effect of the micro-structure volatility on the time-to-fill:

- The higher s the more micro-structure volatility is in the orderbook. Thus the best ask is likely to reach price level p faster. Since $c_3 \geq 0$ this corresponds to a larger $u^{(1)}$ and thus a smaller $\tau_p^{(\epsilon)}$, as expected.
- The effect of the micro-structure volatility is scaled by the factor c_3. This is the only term in $u^{(1)}$ which depends (quadratically) on the micro-structure volatility parameter σ_1.

Remark 4.2 (Trade-off between spread effects)*.* As we can see in the approximation formula (4.32), the initial spread s can have opposed effects. On the one hand, a large value of s ($s > \mu$) increases the time-to-fill, because the mean-reversion pulls down the spread. On the other hand, a large value of s increases the (initial) micro-structure volatility, and thus decreases the time-to-fill. The model parameters $d_1, \kappa, \mu, \sigma_0$ and σ_1 then determine which of the two opposed effects is decisive.

Remark 4.3 (Effects not taken account of in first order correction)*.* As we only look at the first two terms of the approximation, we expect that there are behaviours of the orderbook model that are not taken account of in formula (4.32). One effect that is seen to be missing is the *long-term micro-structure volatility*: when the initial spread is zero, and $\mu \neq 0$, it is pushed up by the mean-reversion, and hence the micro-structure volatility starts acting in the model dynamics. In the approximation formula, however, this effect is not represented: when the initital spread s is zero, c_3 - the only term containing σ_1 - vanishes.

Pseudo-mean and variance

We will now see how the asymptotic formula can be used in financial engineering applications: In general, the Laplace transform of some random variable can be used to calculate its moments, when they are finite: we have

$$(-1)^n \lim_{\lambda \to 0} \frac{\partial^n}{(\partial\lambda)^n} \left(\mathrm{E}^{(m,s)} \left[e^{-\lambda \tau_p^{(\epsilon)}} \right] \right) = \mathrm{E}^{(m,s)} \left[\left(\tau_p^{(\epsilon)} \right)^n \right]$$

when this limit exists.

Let us now replace the exact Laplace transform by the approximation in (4.32).

Definition 4.2 (Pseudo-moments)**.** We call

$$(-1)^n \lim_{\lambda \to 0} \frac{\partial^n}{(\partial\lambda)^n} \left(u^{(0)}(m,s) + \epsilon u^{(1)}(m,s) \right)$$

the n$^{\mathrm{th}}$ *pseudo-moment* of $\tau_p^{(\epsilon)}$ whenever the limit exists, and denote it by

$$\mathrm{E_{pseudo}}^{(m,s)} \left[\left(\tau_p^{(\epsilon)} \right)^n \right].$$

A similar notation applies to the variance.

A simple computation then gives explicit formulae for the pseudo-mean and -variance of $\tau_p^{(\epsilon)}$ when $d_1 > 0$. We have

$$\mathrm{E_{pseudo}}^{(m,s)} \left[\tau_p^{(\epsilon)} \right] = 2 \left(p - m - \frac{s}{2} \right) \left(\frac{1}{d_1} + \epsilon \frac{\kappa(s-\mu)}{d_1^2} \right), \tag{4.37}$$

$$\begin{aligned} \mathrm{Var_{pseudo}}^{(m,s)} \left[\tau_p^{(\epsilon)} \right] =& 2 \frac{p - m - \frac{s}{2}}{d_1^4} \left\{ d_1 \sigma_0^2 + \right. \\ & \left. \epsilon \left(3\sigma_0^2 \kappa (s-\mu) + d_1 s \sigma_1^2 \right) - \epsilon^2 2 \left(p - m - \frac{s}{2} \right) \kappa^2 (s-\mu)^2 \right\}. \end{aligned} \tag{4.38}$$

ϵ	s	$\mathrm{E_{pseudo}}^{(m,s)}\left[\tau_p^{(\epsilon)}\right]$	$\mathrm{E_{MC}}^{(m,s)}\left[\tau_p^{(\epsilon)}\right]$	$\mathrm{Var_{pseudo}}^{(m,s)}\left[\tau_p^{(\epsilon)}\right]$	$\mathrm{Var_{MC}}^{(m,s)}\left[\tau_p^{(\epsilon)}\right]$
0	0.1	5.00	5.01	7.81	7.58
0.1	0.1	4.50	4.69	5.22	6.30
0.1	0.3	4.75	4.82	6.59	7.10
0.1	0.5	5.00	5.05	7.83	7.79
0.1	0.7	5.25	5.20	8.94	8.76
0.1	0.9	5.50	5.45	9.93	9.42

Table 4.1: Comparison of mean and variance of time-to-fill computed via asymptotic formula and Monte Carlo simulation for parameters $d_1 = 0.1, \kappa = 1.0, \mu = 0.5, \sigma_0 = 0.5, \sigma_1 = 0.1$ and distance $p - m - s/2 = 1$

ϵ	s	$\mathrm{E_{pseudo}}^{(m,s)}\left[\tau_p^{(\epsilon)}\right]$	$\mathrm{E_{MC}}^{(m,s)}\left[\tau_p^{(\epsilon)}\right]$	$\mathrm{Var_{pseudo}}^{(m,s)}\left[\tau_p^{(\epsilon)}\right]$	$\mathrm{Var_{MC}}^{(m,s)}\left[\tau_p^{(\epsilon)}\right]$
0	0.1	5.00	5.06	31.25	31.61
0.1	0.1	4.50	4.83	21.70	30.93
0.1	0.3	4.75	4.90	26.73	30.17
0.1	0.5	5.00	5.00	31.64	31.27
0.1	0.7	5.25	5.15	36.42	32.73
0.1	0.9	5.50	5.40	41.08	36.81

Table 4.2: Comparison of mean and variance of time-to-fill computed via asymptotic formula and Monte Carlo simulation for parameters $d_1 = 0.1, \kappa = 1.0, \mu = 0.5, \sigma_0 = 1.0, \sigma_1 = 0.5$ and distance $p - m - s/2 = 1$

Remark 4.4 (Is the approximation a Laplace transform?). Since we are using the approximation $u^{(0)} + \epsilon u^{(1)}$ to the Laplace transform $u^{(\epsilon)}$ of the stopping time $\tau_p^{(\epsilon)}$, one obvious question is, if there exists some random variable $\hat{\tau}_p^{(\epsilon)}$ of which $u^{(0)} + \epsilon u^{(1)}$ is the Laplace transform. If that was the case, the pseudo-variance given in equation (4.38) would be the true variance. However, for some parameters, we see that $\mathrm{Var_{pseudo}}^{(m,s)}\left[\tau_p^{(\epsilon)}\right]$ can be strictly negative. Thus, in general, $u^{(0)} + \epsilon u^{(1)}$ cannot be a Laplace transform of a random variable.

ϵ	s	$\mathrm{E_{pseudo}}^{(m,s)}\left[\tau_p^{(\epsilon)}\right]$	$\mathrm{E_{MC}}^{(m,s)}\left[\tau_p^{(\epsilon)}\right]$	$\mathrm{Var_{pseudo}}^{(m,s)}\left[\tau_p^{(\epsilon)}\right]$	$\mathrm{Var_{MC}}^{(m,s)}\left[\tau_p^{(\epsilon)}\right]$
0	0.1	2.50	2.52	3.91	3.88
0.1	0.1	2.44	2.47	3.62	3.82
0.1	0.3	2.47	2.48	3.79	3.98
0.1	0.5	2.50	2.52	3.96	3.96
0.1	0.7	2.53	2.55	4.12	4.10
0.1	0.9	2.56	2.57	4.28	4.15

Table 4.3: Comparison of mean and variance of time-to-fill computed via asymptotic formula and Monte Carlo simulation for parameters $d_1 = 0.2, \kappa = 0.5, \mu = 0.5, \sigma_0 = 1.0, \sigma_1 = 0.5$ and distance $p - m - s/2 = 1$

ϵ	s	$\mathrm{E_{pseudo}}^{(m,s)}\left[\tau_p^{(\epsilon)}\right]$	$\mathrm{E_{MC}}^{(m,s)}\left[\tau_p^{(\epsilon)}\right]$	$\mathrm{Var_{pseudo}}^{(m,s)}\left[\tau_p^{(\epsilon)}\right]$	$\mathrm{Var_{MC}}^{(m,s)}\left[\tau_p^{(\epsilon)}\right]$
0	0.1	1.00	1.01	0.25	0.26
0.5	0.1	0.95	0.98	0.21	0.24
0.5	0.3	0.98	0.99	0.24	0.25
0.5	0.5	1.00	1.02	0.27	0.28
0.5	0.7	1.02	1.03	0.29	0.28
0.5	0.9	1.05	1.06	0.31	0.30

Table 4.4: Comparison of mean and variance of time-to-fill computed via asymptotic formula and Monte Carlo simulation for parameters $d_1 = 0.5, \kappa = 0.5, \mu = 0.5, \sigma_0 = 1.0, \sigma_1 = 0.5$ and distance $p - m - s/2 = 1$

Numerical performance

We will now analyse how well the asymptotic approximation of the time-to-fill performs in practice. While theorem 4.3 gives an explicit error bound, it is not very intuitive. Therefore we will compare the approximation (4.32) with results obtained from Monte Carlo simulations.

We choose the pseudo-mean and -variance computed in equations (4.37)- (4.38), and compare it with the corresponding values obtained via Monte Carlo simulations. Note that we could equally well have chosen the Laplace transform instead of mean/variance. However, we opted for the latter, because of their intuitive meaning, and because it is not clear which values of λ should be considered in the Laplace transform.

Let $\mathrm{E_{MC}}^{(m,s)}$ and $\mathrm{Var_{MC}}^{(m,s)}$ denote the mean and variance obtained via Monte Carlo simulations, conditional on $(\mathfrak{m}(0), \mathfrak{s}(0)) = (m, s)$. We use a simple Euler scheme for the simulation of the SDEs. We compute one instance of $\tau_p^{(\epsilon)}$, using N Euler steps, and perform this simulation M times. For the base-case $\epsilon = 0$, we know that our asymptotic formula (4.32) is exact. It turns out that, in order to achieve a rough equality of pseudo- and Monte-Carlo mean, we need to choose $N = 500,000$ and $M = 10,000$.

For a certain set of parameters, we first compute the four quantities (pseudo- and Monte Carlo mean, pseudo- and Monte Carlo variance) for the base-case $\epsilon = 0$ and then for $\epsilon > 0$ and different values of s. Note that m and p are always chosen such that the distance of the initial best ask to the limit order placed at price tick $p - m - s/2$ equals 1.

The results are given in tables 4.1 - 4.4.

Let us start with some comments on the performance. The Monte Carlo simulation converges extremely slowly. To compute mean and variance for one parameter set, $N \times M = 5,000,000,000$ loops are run through, and within each loop one Euler step with a few basic operations (addition, multiplication) is carried out[1]. Even with this extreme computational effort, the Monte Carlo method fails to achieve accuracy up to 2 digits after the decimal point, as can be seen in the first line of each table, where $\epsilon = 0$, and the asymptotic formula is exact.

Next, we investigate the accuracy of the approximation. All comments have to be considered with caution, as we compare our asymptotic approximation with results from Monte Carlo simulations, which are themselves approximations. Nevertheless, a few key points can be readily

[1]With a 3.6 Ghz CPU and 2 GB of RAM, the computation for one parameter set takes more than 30 minutes.

observed:

(1) In general, the formula for the pseudo-mean appears to be a very good approximation to the Monte Carlo mean.

(2) Both pseudo- and Monte Carlo variance are increasing in s.

(3) For $s < \mu$, we have $\mathrm{Var}_{\mathrm{pseudo}}{}^{(m,s)} < \mathrm{Var}_{\mathrm{MC}}{}^{(m,s)}$.

(4) For s close to μ, we have $\mathrm{Var}_{\mathrm{pseudo}}{}^{(m,s)} \approx \mathrm{Var}_{\mathrm{MC}}{}^{(m,s)}$.

(5) For $s > \mu$, we have $\mathrm{Var}_{\mathrm{pseudo}}{}^{(m,s)} > \mathrm{Var}_{\mathrm{MC}}{}^{(m,s)}$.

Point (2) is easily explained by the fact, that a larger s leads to a larger initial microstructure volatility and - due to the mean-reversion - a longer time-to-fill. Both effects increase the variance.

Points (3)-(5) can be explained by remark 4.3: the missing effect of long-term microstructure volatility leads to a too small (large) pseudo-variance for small (large) s. For $s \approx \mu$, there is no under-/over-estimation of the variance.

Conclusion and extensions

While the approximation (4.32) diverge from the true value, especially for initial spread values that are far away from the long-time average spread, it captures most of the main effects of introducing the second dimension *spread* to the one-dimensional Bachelier model, such as mean-reversion and initial microstructure volatility. Given that the decision if (and where) a limit order order should be placed has to be made within milli-seconds, Monte Carlo methods can certainly not be used, as they converge far too slowly. The fast computation of formula (4.32), together with its good approximation results of the time-to-fill, make it a very promising candidate as part of an automated microtrader. Moreover, we briefly discuss two extensions which could be useful for real-world applications.

As a first possible extension, we could include the second order term in our approximation. The inclusion of this term will remove some of the flaws (see remark 4.3) of the first order approximation. Thus we look for an approximation formula of the form

$$\mathrm{E}^{(m,s)}\left[e^{-\lambda\tau_p^{(\epsilon)}}\right] \approx u^{(0)}(m,s) + \epsilon u^{(1)}(m,s) + \epsilon^2 u^{(2)}(m,s)$$

Using the same reasoning as above, we find that $u^{(2)}$ must solve a Poisson problem

$$\begin{aligned} \mathscr{L}^{(0)}u^{(2)} + g = \lambda u^{(2)} \qquad & \text{in } D, \\ u^{(2)} = 0 \qquad & \text{on } \Gamma \end{aligned}$$

where

$$g = \mathscr{L}^{(1)}u^{(1)}.$$

Note that, since $\mathscr{L}^{(0)}$ is a linear second order operator in m, and, as a function of m alone, we have

$$g(m) = e^{r_0 m}(r_1 + r_2 m)$$

for some constants r_0, r_1 and r_2. Thus we only need to solve a linear second order ODE with the above boundary conditions, to compute the second order term. $u^{(2)}$ can be computed explicitly, but the expression is too complicated to admit an intuitive interpretation as was carried out in section 4.3.5.

Another extension is the inclusion of the third dimension volume-imbalance. We consider the three dimensional system

$$
\begin{aligned}
\mathrm{d}\mathfrak{m}^{(\epsilon,\delta)}(t) &= \frac{d_1 + d_2\mathfrak{z}(t)}{2}\mathrm{d}t + \sqrt{\epsilon}\frac{\sigma_1}{2\sqrt{2}}\sqrt{\mathfrak{s}^{(\epsilon,\delta)}(t)}\mathrm{d}B_2(t) + \frac{\sigma_0}{2}\mathrm{d}B_0(t),\\
\mathrm{d}\mathfrak{s}^{(\epsilon,\delta)}(t) &= \epsilon\kappa(\mu - \mathfrak{s}^{(\epsilon,\delta)}(t))\mathrm{d}t + \sqrt{\epsilon}\frac{\sigma_1}{\sqrt{2}}\sqrt{\mathfrak{s}^{(\epsilon,\delta)}(t)}\mathrm{d}B_1(t),\\
\mathrm{d}\mathfrak{z}^{(\epsilon,\delta)}(t) &= \delta\rho(\nu - \mathfrak{z}^{(\epsilon,\delta)}(t))\mathrm{d}t + \sqrt{\delta}\frac{\sigma_2}{\sqrt{2}}\mathrm{d}B_3(t),
\end{aligned}
$$

where B_0, B_1, B_2 and B_3 are independent standard Brownian motions.

It has infinitesimal generator

$$\mathscr{L}^{(\epsilon)} = \mathscr{L}^{(0)} + \epsilon\mathscr{L}^{(1)} + \delta\mathscr{L}^{(2)}$$

with

$$
\begin{aligned}
(\mathscr{L}^{(0)}u)(m,s,z) &= \frac{d_1 + d_2 z}{2}u_m(m,s,z) + \frac{\sigma_0^2}{8}u_{mm}(m,s,z)\\
(\mathscr{L}^{(1)}u)(m,s,z) &= \frac{\sigma_1^2}{16}su_{mm}(m,s,z) + \kappa(\mu - s)u_s(m,s,z) + \frac{\sigma_1^2}{4}su_{ss}(m,s,z)\\
(\mathscr{L}^{(2)}u)(m,s,z) &= \rho(\nu - z)u_z(m,s,z) + \frac{\sigma_2^2}{4}u_{zz}(m,s,z).
\end{aligned}
$$

We are looking for a first order approximation in both ϵ and δ:

$$u^{(\epsilon,\delta)}(m,s,z) \approx u^{(0,0)}(m,s,z) + \epsilon u^{(1,0)}(m,s,z) + \delta u^{(0,1)}(m,s,z)$$

Exactly as in the two-dimensional case, we can compute $u^{(0,0)}$ and $u^{(1,0)}$, and we can derive a linear second order ODE for $u^{(0,1)}$ which can be solved explicitly (but again leads to a rather complicated, unintuitive formula).

4.4 Proofs

Proof of Proposition 4.1. The proof of the below standard results are provided here for the ease of reference.

(i), (ii), (iii): Using Itô's formula and the definition of spread $\mathfrak{s} = \alpha - \beta$, midquote-price $\mathfrak{m} = \frac{1}{2}(\alpha + \beta)$ and volume-imbalace $\mathfrak{z} = \log\frac{V^{\mathfrak{B}}}{V^{\mathfrak{A}}}$, we immediately obtain (4.5), (4.6), and (4.7) with $B_0 = W_0, B_1 = (W_1 - W_2)/\sqrt{2}\,, B_2 = (W_1 + W_2)/\sqrt{2}\,, B_3 = (W_3 + W_4)/\sqrt{2}$ four independent standard brownian motions.

(iv) By standard results on CIR and OU processes, there exist unique strong solutions for $\mathfrak{s}$ and $\mathfrak{z}$. Given these solutions, it is clear that unique solutions for $\mathfrak{m}, V^{\mathfrak{A}}$ and $V^{\mathfrak{B}}$ exist, and hence also for $\alpha(t) = \mathfrak{m}(t) + \frac{\mathfrak{s}(t)}{2}$ and $\beta(t) = \mathfrak{m}(t) - \frac{\mathfrak{s}(t)}{2}$

(v) This follows by standard results for the CIR process.

(vi) Again by standard results for CIR and OU processes, we obtain mean and variance of $\mathfrak{s}$ and $\mathfrak{z}$. Moreover

$$\begin{aligned}\mathfrak{m}(t) &= \mathfrak{m}(0) + \frac{1}{2}d_1 t + \frac{d_2}{2}\int_0^t \mathfrak{z}(u)\mathrm{d}u + \frac{\sigma_1}{2\sqrt{2}}\int_0^t \sqrt{\mathfrak{s}(u)}\mathrm{d}B_2(u) + \frac{\sigma_0}{2}B_0(t) \\ &= \mathfrak{m}(0) + \frac{1}{2}d_1 t + I(t) + M(t) + \frac{\sigma_0}{2}B_0(t)\end{aligned}$$

where all four terms are independent and M is a local martingale. By Fubini's theorem (recall that $\mathfrak{s} \geq 0$)

$$\begin{aligned}\mathrm{E}\left[[M](t)\right] &= \frac{\sigma_1^2}{8}\mathrm{E}\left[\int_0^t \mathfrak{s}(u)\mathrm{d}u\right] = \frac{\sigma_1^2}{8}\int_0^t \mathrm{E}[\mathfrak{s}(u)]\mathrm{d}u \\ &= \frac{\sigma_1^2}{8\kappa}\left(t\kappa\mu + (\mathfrak{s}(0)-\mu)(1-e^{-t\kappa})\right) < \infty\end{aligned}$$

so M is a true square-integrable martingale and $\mathrm{E}[M](t) = 0$. Now note that $|\mathfrak{z}(u)|$ has a folded normal distribution, and thus

$$\mathrm{E}\left[|\mathfrak{z}(u)|\right] = \sigma_u\sqrt{2/\pi}e^{-\mu_u^2/(2\sigma_u^2)} + \mu_u(1-2\Phi(-\mu_u/\sigma_u))$$

where $\mu_u = \mathrm{E}[\mathfrak{z}(u)], \sigma_u = \mathrm{Var}(\mathfrak{z}(u))$ and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. We deduce that $\int_0^t |\mathfrak{z}(u)|\mathrm{d}u < \infty$, so we can apply Fubini's theorem and obtain

$$\begin{aligned}\mathrm{E}\left[I(t)\right] &= \frac{d_2}{2}\int_0^t \mathrm{E}[\mathfrak{z}(u)]\mathrm{d}u \\ &= \frac{d_2}{2\rho}\left((\mathfrak{z}(0)-\nu)\left(1-e^{-\rho t}\right) + \nu\rho t\right)\end{aligned}$$

Note that by independence $\mathrm{Var}(\mathfrak{m}(t)) = \mathrm{Var}(I(t)) + \mathrm{Var}(M(t)) + \frac{\sigma_0^2}{4}t$, and since M is square-integrable, we have $\mathrm{Var}(M(t)) = \mathrm{E}[M^2(t)] = \mathrm{E}\left[[M](t)\right]$. Using Fubini's theorem, we obtain

$$\begin{aligned}\mathrm{Var}\left(\int_0^t \mathfrak{z}(u)\mathrm{d}u\right) &= \int_0^t\int_0^t \left(\mathrm{E}[\mathfrak{z}(u)\mathfrak{z}(s)] - \mathrm{E}[\mathfrak{z}(u)]\mathrm{E}[\mathfrak{z}(s)]\right)\mathrm{d}s\mathrm{d}u \\ &= \int_0^t\int_0^t \mathrm{Cov}\left(\mathfrak{z}(u),\mathfrak{z}(s)\right)\mathrm{d}s\mathrm{d}u \\ &= 2\iint_{0\leq s\leq t,\, 0\leq u\leq s} \mathrm{Cov}\left(\mathfrak{z}(u),\mathfrak{z}(s)\right)\mathrm{d}s\mathrm{d}u \\ &= 2\iint_{0\leq s\leq t,\, 0\leq u\leq s} \frac{\sigma_2^2}{4\rho}\left(e^{-\rho(s-u)} - e^{-\rho(s+u)}\right) \\ &= \frac{\sigma_2^2}{4\rho^2}\left(2\rho t - 3 + 4e^{-\rho t} - e^{-2\rho t}\right)\end{aligned}$$

Finally use the decomposition of best ask/bid in terms of midquote price and spread to compute mean and variance of best ask and bid. □

Proof of Proposition 4.2. This follows directly from lemma 4.1 below with $h = 0$ and $c = 1$. □

Lemma 4.1 (Viscosity solution). *Consider the two-dimensional diffusion $X(t) = (m(t), s(t))$ with dynamics*

$$\mathrm{d}m(t) = d\mathrm{d}t + \sigma_0 \mathrm{d}B_0(t) + \sigma_1\sqrt{s(t)}\mathrm{d}B_1(t), \tag{4.39}$$

$$\mathrm{d}s(t) = \kappa(\mu - s(t))\mathrm{d}t + \sigma_2\sqrt{s(t)}\mathrm{d}B_2(t), \tag{4.40}$$

where B_0, B_1, B_2 are independent standard Brownian motions, and $\kappa > 0, \mu \geq 0, \sigma_0 > 0$. It has infinitesimal generator

$$(\mathscr{L}u)(m,s) = du_m(m,s) + \frac{\sigma_0^2}{2}u_{mm}(m,s) + \frac{\sigma_1^2}{2}su_{mm}(m,s) + \kappa(\mu - s)u_s(m,s) + \frac{\sigma_2^2}{2}su_{ss}(m,s).$$

Let $\lambda > 0$, $c, p \in \mathbb{R}$, and define the hitting time

$$\tau = \inf\left\{t \geq 0 : m(t) + s(t)/2 \geq p\right\},$$

Also define the regions

$$D = \left\{(m,s) : m + \frac{s}{2} < p, s \geq 0\right\} \subseteq \mathbb{R}^2$$

$$\Gamma = \left\{(m,s) : m + \frac{s}{2} = p, s \geq 0\right\} \subseteq \mathbb{R}^2$$

and let $h : \bar{D} \mapsto \mathbb{R}$ be a continuous function such that $\mathrm{E}^x\left[\int_0^\tau e^{-\lambda s}|h\left(X(s)\right)|\mathrm{d}s\right] < \infty$. Then

$$u(x) = \mathrm{E}^x\left[ce^{-\lambda\tau} + \int_0^\tau e^{-\lambda s}h\left(X(s)\right)\mathrm{d}s\right]$$

is a viscosity solution in the sense of definition 4.1 of the generalised Dirichlet problem (4.9)-(4.10).

Proof of lemma 4.1. We will follow a classical argument, see e.g. Touzi (2004). By the martingale property, for every stopping time T with respect to the filtration of (B_0, B_1, B_2) and for every $x \in \bar{D}$, we have

$$u(x) = \mathrm{E}^x\left[u(X_{T\wedge\tau})e^{-\lambda(T\wedge\tau)} + \int_0^{T\wedge\tau} e^{-\lambda s}h\left(X(s)\right)\mathrm{d}s\right]$$

We first show that u is a subsolution. Let $\phi \in C^2(\mathbb{R}^2)$ and x_0 be a global maximizer of $u^* - \phi$ on $\bar{D}$. Without loss of generality, we may assume that $u^*(x_0) - \phi(x_0) = 0$.

Let us take note that for every $x_0 \in \bar{D}$, we can choose a sequence $x_\epsilon \in D$ such that $x_\epsilon \to x_0$ and $u(x_\epsilon) \to u^*(x_0)$. This is clear for $x_0 \in D$. Now consider the case $x_0 \in \partial D$. If $u^*(x_0) - c \leq 0$, we are done. Thus we may assume $u^*(x_0) > c$. Suppose there does not exist $x_\epsilon \in D$ such that $x_\epsilon \to x_0$ and $u(x_\epsilon) \to u^*(x_0)$. Then there must exist a sequence $x_{\epsilon'} \in \partial D$ such that $x_{\epsilon'} \to x_0$ and $u(x_{\epsilon'}) \to u^*(x_0)$. But then, since $u(x) = c$ for $x \in \partial D$,

$$c = \lim_{\epsilon'} u(x_{\epsilon'}) = u^*(x_0) > c$$

which is a contradiction. Thus for both $x_0 \in D$ and $x_0 \in \partial D$, we may assume that there exists a sequence $x_\epsilon \in D$ such that $x_\epsilon \to x_0$ and $u(x_\epsilon) \to u^*(x_0)$.

Set $\eta_\epsilon = u(x_\epsilon) - \phi(x_\epsilon)$. By continuity of ϕ and choice of x_ϵ, we have $\eta_\epsilon \to 0$. Define the stopping time $T_\epsilon = \inf\{t \geq 0 : X(t) \notin B(x_\epsilon, 1)\} \wedge r_\epsilon$, where

$$r_\epsilon = \sqrt{|\eta_\epsilon|}1_{\{\eta_\epsilon \neq 0\}} + \epsilon 1_{\{\eta_\epsilon = 0\}}.$$

By the martingale property

$$\begin{aligned} 0 &\geq -\mathrm{E}^{x_\epsilon}\left[u(X_{T_\epsilon\wedge\tau})e^{-\lambda(T_\epsilon\wedge\tau)} - u(x_\epsilon) + \int_0^{T_\epsilon\wedge\tau} e^{-\lambda s}h\left(X(s)\right)\mathrm{d}s\right] \\ &\geq \eta_\epsilon - \mathrm{E}^{x_\epsilon}\left[\phi(X_{T_\epsilon\wedge\tau})e^{-\lambda(T_\epsilon\wedge\tau)} - \phi(x_\epsilon) + \int_0^{T_\epsilon\wedge\tau} e^{-\lambda s}h\left(X(s)\right)\mathrm{d}s\right] \end{aligned}$$

Now apply Itô's formula to ϕ and obtain

$$\begin{aligned} 0 \geq& \eta_\epsilon - \mathrm{E}^{x_\epsilon}\left[\int_0^{T_\epsilon\wedge\tau} e^{-\lambda v}\left(\mathscr{L}\phi(X(v)) - \lambda\phi(X(v)) + h(X(v))\right)\mathrm{d}v\right] + \\ &\mathrm{E}^{x_\epsilon}\left[\int_0^{T_\epsilon\wedge\tau} e^{-\lambda v}\left(\phi_m(X(v))\sigma_0\mathrm{d}B_0(v) + \phi_m(X(v))\sigma_1\sqrt{s(v)}\mathrm{d}B_1(v) + \phi_s(X(v))\sigma_2\sqrt{s(v)}\mathrm{d}B_2(v)\right)\right]. \end{aligned}$$

Note that the last term vanishes as any of the integrants is bounded on $[0, T_\epsilon \wedge \tau]$, so

$$0 \geq \frac{\eta_\epsilon}{r_\epsilon} - \mathrm{E}^{x_\epsilon}\left[\frac{1}{r_\epsilon}\int_0^{T_\epsilon\wedge\tau} e^{-\lambda v}\left(\mathscr{L}\phi(X(v)) - \lambda\phi(X(v)) + h(X(v))\right)\mathrm{d}v\right]$$

For ω fixed and $\epsilon > 0$ sufficiently small, we have $T_\epsilon(\omega) \wedge \tau = r_\epsilon$ when X is started at $x_\epsilon \in D$, and by uniform boundedness of $\frac{1}{r_\epsilon}\int_0^{T_\epsilon\wedge\tau} e^{-\lambda v}\left(\mathscr{L}\phi(X(v)) - \lambda\phi(X(v)) + h(X(v))\right)\mathrm{d}v$ in ϵ, we can apply the mean value theorem and dominated convergence to obtain

$$-\left(\mathscr{L}\phi(x_0) - \lambda u^*(x_0) + h(x_0)\right) \leq 0$$

The supersolution property is shown in the same way. □

Proof of proposition 4.3. We start by showing that the error term $R^{(\epsilon)}$ solves a certain Dirichlet problem. Using the definition of $R^{(\epsilon)}$, the Dirichlet problems solved by $u^{(0)}$ and $u^{(1)}$ and lemma 4.1 (by which $\mathscr{L}^{(\epsilon)}u^{(\epsilon)} = \lambda u^{(\epsilon)}$ in the viscosity sense) , we compute

$$\begin{aligned} \mathscr{L}^{(\epsilon)}R^{(\epsilon)} =& \mathscr{L}^{(\epsilon)}\left(u^{(0)} + \epsilon u^{(1)} - u^{(\epsilon)}\right) \\ =& \left(\mathscr{L}^{(0)} + \epsilon\mathscr{L}^{(1)}\right)\left(u^{(0)} + \epsilon u^{(1)}\right) - \mathscr{L}^{(\epsilon)}u^{(\epsilon)} \\ =& \mathscr{L}^{(0)}u^{(0)} + \epsilon\left(\mathscr{L}^{(0)}u^{(1)} + \mathscr{L}^{(1)}u^{(0)}\right) + \epsilon^2\mathscr{L}^{(1)}u^{(1)} - \mathscr{L}^{(\epsilon)}u^{(\epsilon)} \\ =& \lambda\left(u^{(0)} + \epsilon u^{(1)} - u^{(\epsilon)}\right) + \epsilon^2\mathscr{L}^{(1)}u^{(1)} \\ =& \lambda R^{(\epsilon)} + \epsilon^2\mathscr{L}^{(1)}u^{(1)} \end{aligned}$$

in the viscosity sense.

Moreover, using the boundary conditions for $u^{(0)}$, $u^{(1)}$ and $u^{(\epsilon)}$, we deduce the boundary condition for the error term:

$$R^{(\epsilon)} = 0 \qquad \text{on } \Gamma.$$

Hence, by lemma 4.1, with $c = 0$ and $h(m,s) = \left(\mathscr{L}^{(1)}u^{(1)}\right)(m,s)$, the probabilistic interpretation

$$\tilde{R}^{(\epsilon)}(m,s) = -\epsilon^2 \mathrm{E}^{(m,s)}\left[\int_0^{\tau_p^{(\epsilon)}} e^{-\lambda t}\left(\mathscr{L}^{(1)}u^{(1)}\right)(\mathfrak{m}^{(\epsilon)}(t), \mathfrak{s}^{(\epsilon)}(t))\mathrm{d}t\right]$$

is a viscosity solution of the above equations. Using the definition of $\mathscr{L}^{(1)}$ in (4.17) and the formula for $u^{(1)}$ in (4.21), we can compute

$$\left(\mathscr{L}^{(1)}u^{(1)}\right)(m,s) = \frac{u^{(0)}(m,s)}{4\sqrt{d_1^2 + 2\lambda\sigma_0^2}}\left(k_0 + k_1 s + k_2 s^2 + (2m + s - 2p)(l_0 + l_1 s + l_2 s^2)\right)$$

Using the bounds

$$\left|u^{(0)}(m,s)\right| \leq 1 \text{ in } D$$

and

$$\left|u^{(0)}(m,s)(2m + s - 2p)\right| \leq e^{-1}\frac{\sigma_0^2}{\sqrt{d_1^2 + 2\lambda\sigma_0^2} - d_1} \text{ in } D,$$

we obtain the bound on $\mathscr{L}^{(1)}u^{(1)}$

$$\left|\left(\mathscr{L}^{(1)}u^{(1)}\right)(m,s)\right| \leq \frac{1}{4\sqrt{d_1^2 + 2\lambda\sigma_0^2}}\left(a_0 + a_1 s + a_2 s^2\right) \text{ in } D.$$

Thus we have

$$\left|\tilde{R}^{(\epsilon)}(m,s)\right| \leq \frac{\epsilon^2}{4\sqrt{d_1^2 + 2\lambda\sigma_0^2}}\mathrm{E}^{(m,s)}\left[\int_0^{\infty} e^{-\lambda t}\left(a_0 + a_1\mathfrak{s}^{(\epsilon)}(t) + a_2\left(\mathfrak{s}^{(\epsilon)}(t)\right)^2\right)\mathrm{d}t\right] \text{ in } D.$$

Using Fubinis theorem and the explicit formula for the first two moments of $\mathfrak{s}^{(\epsilon)}$

$$\begin{aligned}
\mathrm{E}\left[\mathfrak{s}^{(\epsilon)}(t)\middle|\mathfrak{s}^{(\epsilon)}(0) = s\right] =& \mu + (s - \mu)e^{-\epsilon\kappa t} \\
\mathrm{E}\left[\left(\mathfrak{s}^{(\epsilon)}(t)\right)^2\middle|\mathfrak{s}^{(\epsilon)}(0) = s\right] =& s^2 e^{-2\epsilon\kappa t} + s e^{-2\epsilon\kappa t}\left(e^{\epsilon\kappa t} - 1\right)\frac{4\kappa\mu + \sigma_1^2}{2\kappa} + \\
& \mu e^{-2\epsilon\kappa t}\left(e^{\epsilon\kappa t} - 1\right)^2\frac{4\kappa\mu + \sigma_1^2}{4\kappa},
\end{aligned}$$

we obtain the estimate (4.24) which implies the first bound (4.23) as $\epsilon \in [0,1]$. □

Chapter 5

Model calibration and test

Mathematical models can be of interest on their own on the theoretical side. However, when the model is applied to real-world problems, the estimation of the model parameters are of highest importance. In this chapter, we will investigate how the parameters of the model described in chapter 3 can be estimated using high-frequency orderbook data.

5.1 Method overview

As orderbook models are mainly used in automated high-frequency trading, it is of particular interest that the estimation of model parameters is both *numerically stable* and *fast*, i.e. can be performed in real-time. We will test the model and its parameter estimation by analysing its effectiveness of predicting time-to-fill of limit orders as defined in chapter 4, i.e. $\tau_p = \inf\{t \geq 0 : \alpha(t) \geq p\}$. Thus we are interested in the behaviour of best bid and best ask, and it is sufficient to look at the 3-dimensional model $(\alpha, \beta, \mathfrak{z})$ given by initial values $\alpha(0), \beta(0), \mathfrak{z}(0)$ and the SDEs

$$\begin{aligned}
\mathrm{d}\alpha(t) &= \frac{\kappa}{2}(\mu - \mathfrak{s}(t))\mathrm{d}t + \frac{1}{2}(d_1 + d_2\mathfrak{z}(t))\mathrm{d}t + \frac{\sigma_1}{2}\sqrt{\mathfrak{s}(t)}\mathrm{d}W_1(t) + \frac{\sigma_0}{2}\mathrm{d}W_0(t)\\
\mathrm{d}\beta(t) &= -\frac{\kappa}{2}(\mu - \mathfrak{s}(t))\mathrm{d}t + \frac{1}{2}(d_1 + d_2\mathfrak{z}(t))\mathrm{d}t - \frac{\sigma_1}{2}\sqrt{\mathfrak{s}(t)}\mathrm{d}W_2(t) + \frac{\sigma_0}{2}\mathrm{d}W_0(t)\\
\mathrm{d}\mathfrak{z}(t) &= \rho(\nu - \mathfrak{z}(t))\mathrm{d}t + \frac{\sigma_2}{\sqrt{2}}\mathrm{d}B_3(t),
\end{aligned}$$

or equivalently initial values $\mathfrak{s}(0), \mathfrak{m}(0), \mathfrak{z}(0)$ and

$$\mathrm{d}\mathfrak{s}(t) = \kappa(\mu - \mathfrak{s}(t))\mathrm{d}t + \frac{\sigma_1}{\sqrt{2}}\sqrt{\mathfrak{s}(t)}\mathrm{d}B_1(t) \tag{5.1}$$

$$\mathrm{d}\mathfrak{m}(t) = \frac{1}{2}(d_1 + d_2\mathfrak{z}(t))\mathrm{d}t + \frac{\sigma_1}{2\sqrt{2}}\sqrt{\mathfrak{s}(t)}\mathrm{d}B_2(t) + \frac{\sigma_0}{2}\mathrm{d}B_0(t) \tag{5.2}$$

$$\mathrm{d}\mathfrak{z}(t) = \rho(\nu - \mathfrak{z}(t))\mathrm{d}t + \frac{\sigma_2}{\sqrt{2}}\mathrm{d}B_3(t). \tag{5.3}$$

The model parameters are estimated using different methods which are applied to the relevant processes. Table 5.1 gives an overview how the different parameters are estimated, and which processes are used. To test the model, we compute the full distribution of first passage times

Constants estimated	Process	Method
κ, μ	Spread	Maximum likelihood estimation
ρ, ν	Volume-imbalance	Maximum likelihood estimation
d_1, d_2	Midquote	Least square approximation
σ_1	Spread	Two scales realized volatility
σ_2	Volume-imbalance	Two scales realized volatility
σ_0	Midquote and Spread	Two scales realized volatility

Table 5.1: Calibration method

Stock symbol	Company	Time window	N	Freq.
AAPL	Apple	10.00h-11.00h	17778	4.9
BJS	BJ Services Company	10.00h-15.00h	41807	2.3
MSFT	Microsoft Corporation	10.00h-11.00h	21454	6.0

Table 5.2: Orderbook data

in terms of cumulative distribution functions. Then we compare the three-dimensional model and two reference models to the empirical first passage time computed from orderbook data.

5.2 The data

We use orderbook data from NASDAQ, on February 17, 2010, with different time windows depending on the stock and venue to ensure that the number of data points N is large enough, yet not including any noise from opening/intraday/closing auctions. Table 5.2 gives an overview of the stocks we will analyse. It lists the stock symbol, the full company name, time window, data size N and average trading frequency in trades per second.

The orderbook data we use is four-dimensional: time, best bid, best ask and volume imbalance. After cleaning the raw data and removing duplicate values, we obtain the data in the form of N four-tuples

$$(t_i, \beta(t_i), \alpha(t_i), \mathfrak{z}(t_i))) \qquad i = 1, ..., N,$$

where time is given in milliseconds after midnight, and the best bid/ask prices in US-dollars. Note that this data is sufficient to compute all the paths of the processes given by equations (5.1)-(5.3).

For all stocks the tick size is 0.01 US-dollars. To get an overview over the stock characteristics, figure 5.1 displays for each stock

- the evolution of the midquote price (price given in US-dollars, time given in milliseconds after midnight),
- the number of time points when the spread is equal to $1, 2, 3, \ldots, 10$ ticks,
- the number of time points when the volume-imbalance is equal to $-5, 4, \ldots, 4, 5$, when rounded to the nearest integer.

Note that while the spread for AAPL takes values between 1 and 5 ticks with relatively high frequency, the spread for BJS is equal to 1 in 90.5% of all data points and equals 2 ticks otherwise, and for MSFT, the spread is equal to 1 tick in 96.6% of all data points, otherwise being equal to 2 ticks.

5.3 Calibration

We henceforth assume that we have a cleaned dataset in the form $(t_i, \beta(t_i), \alpha(t_i), \mathfrak{z}(t_i))), \quad i = 1, ..., N$. We want to calibrate a continuous diffusion model to high-frequency financial market data, so we are faced with a number of problems:

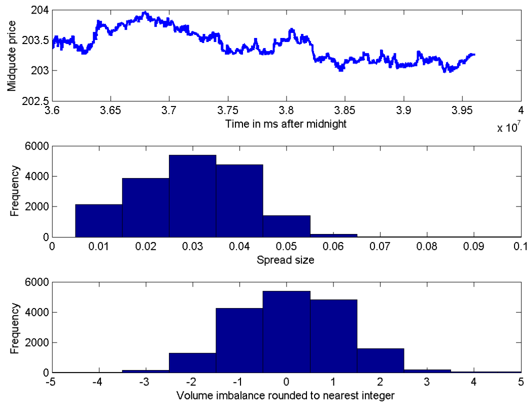
- The data is observed at discrete time points.
- Time points are not evenly spaced.
- Observations are made at extremely high frequency; sometimes there are only a few milliseconds between two trades.
- Observations reflect typical market microstructure noise due to e.g. discreteness of price/volume changes.
- We need to disentangle the different volatility terms described in section 4.1, that is microstructure volatility caused by the spread, microstructure volatility caused by volume-imbalance, and exogeneous volatility.

5.3.1 Estimation of volatility parameters

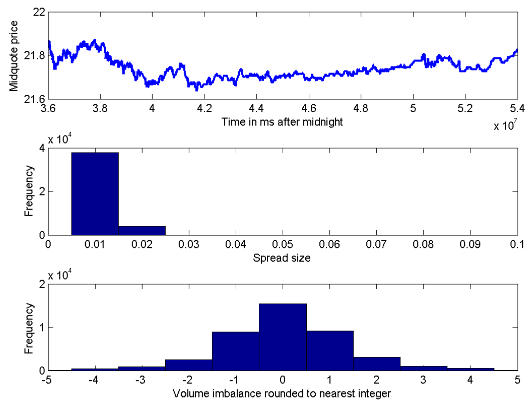
We start by estimating the volatility parameters σ_0, σ_1 and σ_2. If we naively try to estimate them by using every data point, the market microstructure noise introduces a bias. In fact, instead of estimating the volatility of the process we are interested in, we will estimate the volatility of the noise, scaled by the number of observations, as has been noticed in Ait-Sahalia, Zhang, and Mykland (2005, section 2.2). Thus we would obtain estimates for $\sigma_i, i = 0, 1, 2$ which are far too large. The microstructure noise is mainly due to discreteness of price ticks and other microstructure effects which are not explicitly included in our model, stemming, for example, from price ticks beyond the best quotes. To overcome this problem, we use the method of *two scales realized volatility* which was introduced in Ait-Sahalia, Zhang, and Mykland (2005) as *the first-best approach* for volatility estimation of noisy high-frequency data, and which we briefly describe below:

Suppose we want to estimate the quadratic variation $[X]$ of a process X, and we observe the noisy process $Y_{t_i} = X_{t_i} + \epsilon_{t_i}$ at times $t_i, i = 1, \dots, N$, where the ϵ_{t_i}'s are independent noise around the true value of X_{t_i}. We proceed in 3 steps:

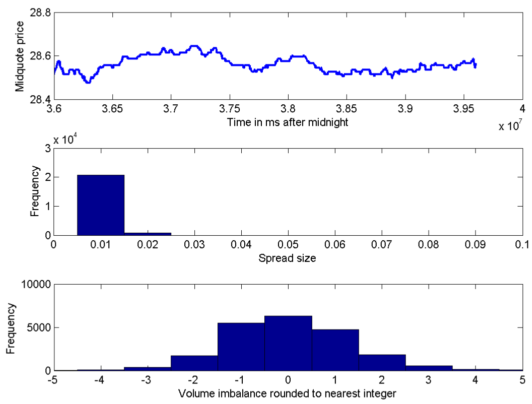
- Partition the original time grid $\mathcal{G} = \{t_1, \dots, t_N\}$ into subsamples $\mathcal{G}^{(k)}, \quad k = 1, \dots, K$, with $\mathcal{G}^{(k)} = \left\{t_i^{(k)} | i = 1, \dots, N^{(k)}\right\}$, $t_i^{(k)} \leq t_{i+1}^{(k)}$, and $N/K \to \infty$ as $N \to \infty$.



(a) AAPL



(b) BJS



(c) MSFT

Figure 5.1: Overview midquote price, spread and volume imbalance for each stock

- For each subsample $\mathcal{G}^{(k)}$, compute $\widehat{[X]}^{(k)} = \sum_{i=1}^{N^{(k)}} \left(X_{t_{i+1}^{(k)}} - X_{t_i^{(k)}} \right)^2$. Then compute $\widehat{[X]}^{\text{all}} = \sum_{i=1}^{N} \left(X_{t_{i+1}} - X_{t_i} \right)^2$ and $\widehat{[X]}^{\text{avg}} = \sum_{i=1}^{K} \widehat{[X]}^{(k)} / K$.
- Finally compute a combination of the estimators obtained over the two time scales 'all' and 'avg', the *bias-adjusted estimator*

$$\widehat{[X]} = \widehat{[X]}^{\text{avg}} - \frac{1}{K} \widehat{[X]}^{\text{all}}$$

The estimator $\widehat{[X]}$ is shown to have desirable properties in Ait-Sahalia, Zhang, and Mykland (2005). In particular one can show that the estimator is correctly centered and converges at rate $n^{1/6}$ ((see Ait-Sahalia, Zhang, and Mykland, 2005, Theorem 4)). Using this method, we obtain estimates for σ_1 and σ_2.

It remains to estimate the parameter σ_0 in (5.2). We need to proceed carefully because the infinite variation part of the midquote price is composed by

- the constant volatility term $\frac{\sigma_0}{2} \mathrm{d}B_0(t)$,
- the stochastic volatiliy term, depending on the current spread $\frac{\sigma_1}{2\sqrt{2}} \sqrt{\mathfrak{s}(t)} \mathrm{d}B_2(t)$, and
- additional volatility coming from market microstructure noise.

Again, using the method of *two scales realized volatility*, we obtain estimators $\widehat{[\mathfrak{m}]}(t)$ for $[\mathfrak{m}](t)$ and $\widehat{[\mathfrak{s}]}(t)$ for $[\mathfrak{s}](t)$. Now note that

$$[\mathfrak{s}](t) = \frac{\sigma_1^2}{2} \int_0^t \mathfrak{s}(u) \mathrm{d}u,$$
$$[\mathfrak{m}](t) = \frac{\sigma_1^2}{8} \int_0^t \mathfrak{s}(u) \mathrm{d}u + \frac{\sigma_0^2}{4} t,$$

so we obtain a good estimator for σ_0 by setting

$$\hat{\sigma}_0 = \sqrt{\frac{4\widehat{[\mathfrak{m}]}(t) - \widehat{[\mathfrak{s}]}(t)}{t}}. \tag{5.4}$$

5.3.2 Estimation of CIR process

Suppose we have data $(t_i, s(t_i)), \quad i = 1, \ldots, N$, and we assume that $s(t)$ is a realization of a CIR process given by the SDE

$$\mathrm{d}s(t) = \kappa(\mu - s(t))\mathrm{d}t + \sigma \sqrt{s(t)} \mathrm{d}W(t),$$

where W is a standard Brownian motion, and σ is known. We will estimate the parameters $\vartheta = (\kappa, \mu) \in \Theta = \mathbb{R}_+ \times \mathbb{R}_+$ using the method of maximum likelihood. Given $s(t)$ at time t, let

$$p(s(t + \Delta t)|s(t); \vartheta, \Delta t)$$

be the conditional density of $s(t+\Delta t)$ at time $t+\Delta t$. A maximum likelihood estimator $\hat{\vartheta}$ is defined as

$$\hat{\vartheta} = \arg\max_{\vartheta\in\Theta} l(\vartheta),$$

where l is log-likelihood function

$$l(\vartheta) = \sum_{i=1}^{N-1} \log p(s(t_{i+1})|s(t_i);\vartheta, t_{i+1}-t_i).$$

The transition density of the CIR process is

$$p(s(t+\Delta t)|s(t);\vartheta,\Delta t) = ce^{-u-v}\,(v/u)^{q/2}\,I_q(2\sqrt{uv}),$$

where I_q is the modified Bessel function of the first kind and order q, and

$$\begin{aligned} c &= \frac{2\kappa}{\sigma^2(1-e^{\kappa\Delta t})}, \\ u &= cs(t)e^{\kappa\Delta t}, \\ v &= cs(t+\Delta t), \\ q &= \frac{2\kappa\mu}{\sigma^2} - 1. \end{aligned}$$

As pointed out in Kladivko (2007), we shall use a scaled version of the Bessel function $I_q^1(x) = e^{-x}I_q(x)$ when implementing the log-likelihood function in MATLAB to ensure stability of the optimisation. For a MATLAB implementation of CIR log-likelihood function see appendix A.1.

Any numerical method used to find the optimal $\hat{\vartheta}$ (e.g., gradient search or derivative-free methods), requires a good initial estimate. As suggested in Kladivko (2007) we use ordinary least square (OLS) approximation to obtain the initial estimate; for this, consider a discretized version of the CIR process:

$$s(t+\Delta t) - s(t) = \kappa(\mu - s(t))\Delta t + \sigma\sqrt{s(t)}\epsilon(\Delta t), \tag{5.5}$$

where $\epsilon(\Delta t)$ is normally distributed with zero mean and variance Δt. Transforming (5.5), we obtain

$$\frac{s(t+\Delta t)-s(t)}{\sqrt{s(t)}} - \frac{\kappa\mu\Delta t}{\sqrt{s(t)}} - \kappa\sqrt{s(t)}\Delta t = \sigma\epsilon(\Delta t)$$

and we take as initial estimate for κ, μ the minimizer of the OLS objective function

$$(\hat{\kappa},\hat{\mu}) = \arg\min_{\kappa,\mu} \sum_{i=1}^{N-1} \left(\frac{s(t_{i+1})-s(t_i)}{\sqrt{s(t_i)}} - \frac{\kappa\mu\Delta t_i}{\sqrt{s(t_i)}} + \kappa\sqrt{s(t_i)}\Delta t_i \right)^2,$$

where $\Delta t_i = t_{i+1} - t_i$. Appendix A.1 contains the corresponding MATLAB code.

5.3.3 Estimation of OU process

Suppose we have data $(t_i, z(t_i))$, $\quad i = 1, \ldots, N$, and it is assumed that $z(t)$ is a realization of an OU process given by the SDE

$$\mathrm{d}z(t) = \rho(\nu - z(t))\mathrm{d}t + \sigma\mathrm{d}W(t),$$

where W is a standard Brownian motion, and σ is known. As in the CIR case we estimate the parameters $\vartheta = (\rho, \nu) \in \Theta = \mathbb{R}_+ \times \mathbb{R}$ using the method of maximum likelihood. Again, the transition density is known in closed form:

$$p(z(t+\Delta t)|z(t); \vartheta, \Delta t) = \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \exp\left\{ -\frac{\left(z(t+\Delta t) - z(t)e^{-\rho\Delta t} - \nu(1-e^{-\rho\Delta t})\right)^2}{2\tilde{\sigma}^2} \right\},$$

with $\tilde{\sigma}^2 = \sigma^2 \frac{1-e^{-2\rho\Delta t}}{2\rho}$. To obtain a good initial estimate, we apply again the OLS method. The discretized version of the OU process reads

$$z(t+\Delta t) - z(t) = \rho(\nu - z(t))\Delta t + \sigma\epsilon(\Delta t), \tag{5.6}$$

where $\epsilon(\Delta t)$ is normally distributed with zero mean and variance Δt. Minimizing the OLS objective function we find initial estimates for the drift term

$$(\hat{\rho}, \hat{\nu}) = \arg\min_{\rho,\nu} \sum_{i=1}^{N-1} \left(z(t_{i+1}) - z(t_i) - \rho\nu\Delta t_i + \rho z(t_i)\Delta t_i\right)^2,$$

with $\Delta t_i = t_{i+1} - t_i$. For a MATLAB implementation see appendix A.1.

5.3.4 Estimation of midquote price

Consider a discretized version of (5.2):

$$\mathfrak{m}(t+\Delta t) - \mathfrak{m}(t) = \frac{1}{2}(d_1 + d_2\mathfrak{z}(t))\Delta t + \epsilon(\Delta t, \mathfrak{s}(t)),$$

where $\epsilon(\Delta t, \mathfrak{s}(t))$ is normally distributed with mean zero and variance $\frac{\sigma_1^2 \mathfrak{s}(t)(\Delta t)^2}{8} + \frac{\sigma_0^2(\Delta t)^2}{4}$. The parameters d_1 and d_2 are estimated by a simple ordinary least square optimization

$$(\hat{d}_1, \hat{d}_2) = \arg\min_{d_1,d_2} \sum_{i=1}^{N-1} \left(\mathfrak{m}(t_{i+1}) - \mathfrak{m}(t_i) - \frac{1}{2}d_1\Delta t_i - \frac{1}{2}d_2 z(t_i)\Delta t_i\right)^2.$$

Again, the MATLAB code can be found in appendix A.1.

5.3.5 Calibration algorithm

The following pseudocode illustrates the calibration algorithm.

Input: Orderbook data $(t_i, \beta_i, \alpha_i, \mathfrak{z}_i)_{i=1}^N$.
Output: Estimators for model parameters $(\hat{\kappa}, \hat{\mu}, \hat{\sigma_1}, \hat{\rho}, \hat{\nu}, \hat{\sigma_2}, \hat{d}_1, \hat{d}_2, \hat{\sigma_0})$.

for $i = 1$ to N
    $\mathfrak{s}_i = \alpha_i - \beta_i$;
    $\mathfrak{m}_i = (\alpha_i + \beta_i)/2$;
end

Parameter	AAPL	BJS	MSFT
κ	7.4×10^{-5}	1.2×10^{-2}	2.6×10^{-4}
μ	3.2×10^{-2}	1.1×10^{-2}	1.0×10^{-2}
σ_1	1.2×10^{-3}	3.7×10^{-4}	3.7×10^{-4}
ρ	8.7×10^{-5}	3.6×10^{-5}	5.6×10^{-5}
ν	-8.9×10^{-2}	-1.1×10^{-1}	-9.7×10^{-3}
σ_2	2.0×10^{-2}	1.8×10^{-2}	2.3×10^{-2}
d_1	-2.3×10^{-7}	2.1×10^{-8}	-7.2×10^{-8}
d_2	1.9×10^{-6}	8.0×10^{-8}	1.4×10^{-7}
σ_0	9.5×10^{-4}	1.2×10^{-4}	1.4×10^{-4}

Table 5.3: Parameters obtained by estimation procedure

Compute $\widehat{[\mathfrak{s}]}^{\text{avg}}, \widehat{[\mathfrak{m}]}^{\text{avg}}, \widehat{[\mathfrak{z}]}^{\text{avg}}$; (section 5.3.1)
Compute $\widehat{[\mathfrak{s}]}^{\text{all}}, \widehat{[\mathfrak{m}]}^{\text{all}}, \widehat{[\mathfrak{z}]}^{\text{all}}$; (section 5.3.1)

$\widehat{[\mathfrak{s}]} = \widehat{[\mathfrak{s}]}^{\text{avg}} - \widehat{[\mathfrak{s}]}^{\text{all}}/K;$
$\widehat{[\mathfrak{m}]} = \widehat{[\mathfrak{m}]}^{\text{avg}} - \widehat{[\mathfrak{m}]}^{\text{all}}/K;$
$\widehat{[\mathfrak{z}]} = \widehat{[\mathfrak{z}]}^{\text{avg}} - \widehat{[\mathfrak{z}]}^{\text{all}}/K;$

$\hat{\sigma_1} = (2\widehat{[\mathfrak{s}]}/\int_{t_1}^{t_N} \mathfrak{s}(u)\mathrm{d}u)^{1/2};$
$\hat{\sigma_2} = (2\widehat{[\mathfrak{z}]}/(t_N - t_1))^{1/2};$
$\hat{\sigma_0} = ((4\widehat{[\mathfrak{m}]} - \widehat{[\mathfrak{s}]})/(t_N - t_1))^{1/2};$

$(\hat{\kappa}, \hat{\mu}) = \arg\min_{\kappa,\mu} \left\{-\text{LogLike}_{\text{CIR}}(\kappa, \mu; \hat{\sigma_1}, (t_i)_{i=1}^N, (\mathfrak{s}_i)_{i=1}^N)\right\};$

$(\hat{\rho}, \hat{\nu}) = \arg\min_{\rho,\nu} \left\{-\text{LogLike}_{\text{OU}}(\rho, \nu; \hat{\sigma_2}, (t_i)_{i=1}^N, (\mathfrak{z}_i)_{i=1}^N)\right\};$

$(\hat{d}_1, \hat{d}_2) = \arg\min_{d_1,d_2} \left\{\sum_{i=1}^{N-1} \left(\mathfrak{m}(t_{i+1}) - \mathfrak{m}(t_i) - \frac{1}{2}d_1\Delta t_i - \frac{1}{2}d_2 z(t_i)\Delta t_i\right)^2\right\};$

The detailed MATLAB code can be found in appendix A.1.

5.3.6 Calibration results

Table 5.3 shows the parameters obtained by the algorithms described above. Note that we have $d_2 > 0$ in all cases, which is consistent with the order flow assumption (X6) regarding volume-imbalance in section 3.1.

Also note that the algorithm achieves our objectives we set at the beginning of the chapter:

- The algorithm is numerically stable, provided that we use stable optimization routines to find the minimum of the log-likelihood functions (such routines certainly exist for 2-dimensional problems we consider here).

- The simplicity of the calibration algorithm makes sure that it is extremely fast, and the model can be re-calibrated in real time, thus allowing to incorporate changes of the microstructural behaviour of the market immediately. Such a situation may arise if e.g. a trader starts selling a large position constantly over a longer period.

5.4 Computation of time-to-fill/first passage time

As discussed in section 4.2, we consider the first passage time τ_p as an approximation of the time-to-fill $\hat{\tau}_p$. Note that neither our model, nor the available data allow us to give an exact definiton of $\hat{\tau}_p$ and compute it. Thus the approximation via first passage time is clearly the best we can achieve, given our model and data constraints. Moreover equation (4.8) gives us bounds of the time-to-fill in terms of first passage times.

Let Δp be the tick size of the orderbook. We are interested in conditional probabilities, namely

$$\mathrm{P}^{(\mathfrak{s}_0,\mathfrak{z}_0)}(\tilde{\tau}_n \leq t) := \mathrm{P}(\tau_{\alpha(0)+n\Delta p} \leq t | \mathfrak{s}(0) = \mathfrak{s}_0, \mathfrak{z}(0) = \mathfrak{z}_0), \quad t \geq 0,$$

the cumulative distribution function (CDF) of the time-to-fill of a sell limit order placed n ticks above the initial best ask (at time 0), conditional on the initial spread to be $\mathfrak{s}_0$ and the initial volume imbalance to be $\mathfrak{z}_0$. Note that this quantity (at least in our modelling framework) does not depend on the initial best ask $\alpha(0)$.

We will compute $\mathrm{P}^{(\mathfrak{s}_0,\mathfrak{z}_0)}(\tilde{\tau}_n \leq t)$ in the range $t \in [0, T_{\text{end}}]$ milli-seconds empirically for the data and for the orderbook model.

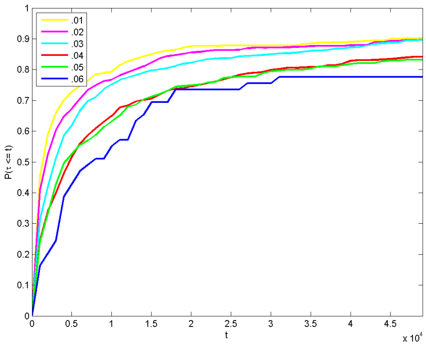
Computing CDF observed in data: Given $\mathfrak{s}_0, \mathfrak{z}_0, n$ and $t \in [0, T_{\text{end}}]$, let I be the subset of $\{1, \ldots, N\}$ such that $i \in I$ if and only if $\alpha(t_i) - \beta(t_i) = \mathfrak{s}_0$, $[\mathfrak{z}(t_i)] = \mathfrak{z}_0$ and $t_N - t_i \geq T_{\text{end}}$. Note that $\mathfrak{z}(t_i)$ is rounded to the nearest integer to ensure that there are enough data points. Then define $\mathrm{P}^{(\mathfrak{s}_0,\mathfrak{z}_0)}_{\text{data}}(\tilde{\tau}_n \leq t)$ to be the number of $i \in I$ where the best asks increases by n or more ticks in time $[t_i, t_i + t]$ divided by the total number of sample points:

$$\mathrm{P}^{(\mathfrak{s}_0,\mathfrak{z}_0)}_{\text{data}}(\tilde{\tau}_n \leq t) = \left(\sum_{i \in I} 1_{\{j | \max\{\alpha(t_k) - \alpha(t_j) \geq n\Delta p | k \geq j, t_k - t_j \leq t\}\}}(i) \right) / |I|$$

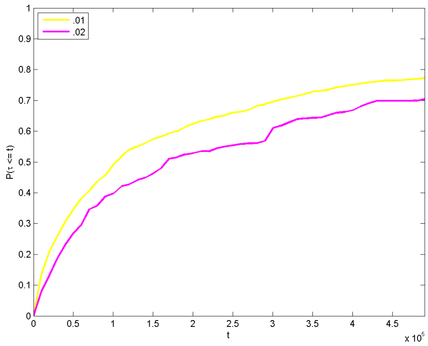
Computing CDF in model: We use a Monte Carlo simulation with initial spread $\mathfrak{s}_0$ and initial volume imbalance $\mathfrak{z}_0$ to compute the hitting times $\mathrm{P}^{(\mathfrak{s}_0,\mathfrak{z}_0)}_{\text{model}}(\tilde{\tau}_n \leq t)$ in the corresponding model.

MATLAB code for these computations can be found in appendix A.2.

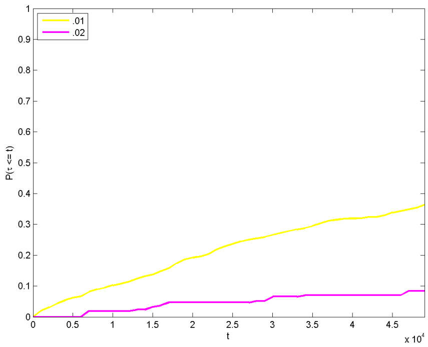
Before comparing model and data, we consider how the first passage time varies in the data, as initial spread and volume-imbalance take different values. Figure 5.2 shows the CDF of the three stocks AAPL, BJS and MSFT for $\mathfrak{z}_0 = 0$, $n = 1$ and varying initial spread from $\mathfrak{s}_0 = 1$ tick (yellow curve) to $\mathfrak{s}_0 = 6$ ticks (blue curve), if applicable. We see that as the initial spread increases, the execution probability of a limit order placed on the best ask decreases. All stocks exhibit a similiar monotone behaviour. The mean-reversion effect of the CIR process that we use to model the spread ensures that we have the same dependence in the model (see also the discussion of the first order term (4.32)). The non-smoothness of some curves (e.g. the blue curve ($\mathfrak{s}_0 = 0.06$) for AAPL) is due to a small sample size $|I|$ in the computation of the CDF.



(a) AAPL



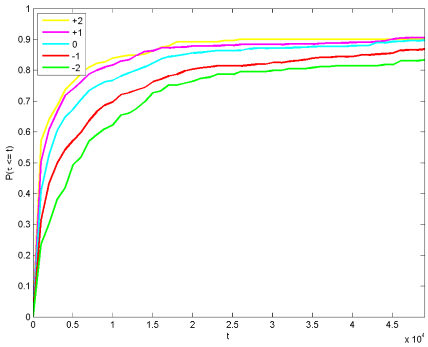
(b) BJS



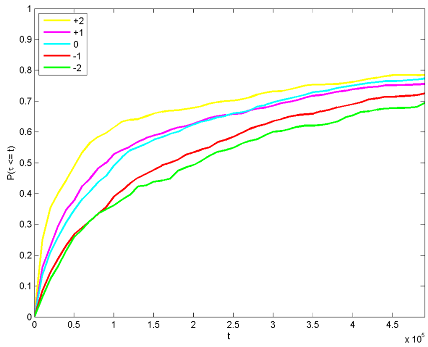
(c) MSFT

Figure 5.2: $\mathrm{P}_{\text{data}}^{(\mathfrak{s}_0,0)}(\tilde{\tau}_1 \leq t)$ with varying $\mathfrak{s}_0$ represented by different colours

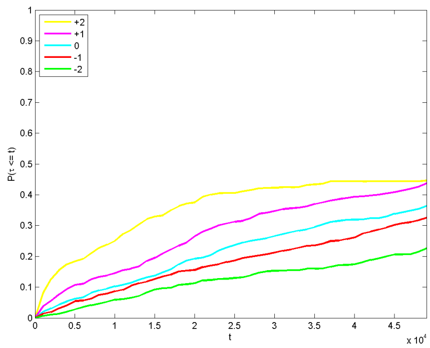
Figure 5.3 shows the CDF of the stocks for a fixed value of $\mathfrak{s}_0$, $n = 1$ and varying initial volume imbalance from $\mathfrak{z}_0 = +2$ (yellow curve) to $\mathfrak{z}_0 = -2$ (blue curve). Again we see a monotone behaviour of the CDFs: the higher the volume imbalance (i.e. more volume on the best bid than on the best ask), the faster a limit order placed on the best ask is executed. In our model, this behaviour is reflected by a positive choice of d_2. We therefore expect that a good calibration will result in a positive estimator of d_2.



(a) AAPL, $\mathfrak{s}_0 = 0.02$



(b) BJS, $\mathfrak{s}_0 = 0.01$



(c) MSFT, $\mathfrak{s}_0 = 0.01$

Figure 5.3: $\mathrm{P}_{\text{data}}^{(\mathfrak{s}_0,\mathfrak{z}_0)}(\tilde{\tau}_1 \le t)$ with varying $\mathfrak{z}_0$ represented by different colours

5.5 Comparison of time-to-fill/first passage times

To test our model, i.e. the general framework in chapter 2 and the choice of (X, Y) in chapter 3 and the calibration methods in section 5.3, we compare the conditional CDFs of first passage times (as an approximation to time-to-fill). We compare empirical CDFs obtained from orderbook data with theoretical CDFs computed in three models:

3-dim model The three-dimensional model $(\mathfrak{s}, \mathfrak{m}, \mathfrak{z})$ given by equations (5.1) - (5.3)

2-dim model A two-dimension model $(\mathfrak{s}, \mathfrak{m})$ given by equations (5.1) and (5.2) where we set $\mathfrak{z} \equiv 0$. In this model, the dynamics of best ask and best bid only depend on the spread.

1-dim model A one-dimensional model α where only the best ask is modelled as a brownian motion with drift: $\mathrm{d}\alpha(t) = \tilde{\mu}\mathrm{d}t + \tilde{\sigma}\mathrm{d}W(t)$.

The parameters of each model are estimated using the methods described in section 5.3. The algorithm for the 3-dim model can easily be adapted to the 2- and 1-dim model. Next, we compute the conditional CDF from the orderbook data $t \mapsto \mathrm{P}^{(\mathfrak{s}_0,\mathfrak{z}_0)}_{\text{data}}(\tau_n \leq t)$. The conditional CDF for the three models, $t \mapsto \mathrm{P}^{(\mathfrak{s}_0,\mathfrak{z}_0)}_{\text{m-dim}}(\tilde{\tau}_n \leq t)$ for $m = 1, 2, 3$ are computed using Monte Carlo simulations. All four CDFs are plotted in the same graph; the solid line shows the CDF from the data, the dashed line corresponds to the 3-, the dotted line to the 2-, and the cross-line to the 1-dim model. For AAPL and MSFT, the CDF is plotted in the range $[0, T_{\text{end}}]$, for $T_{\text{end}} = 50,000$ milliseconds. Due to the lower trading frequency of BJS, we choose $T_{\text{end}} = 500,000$ milliseconds in this case.

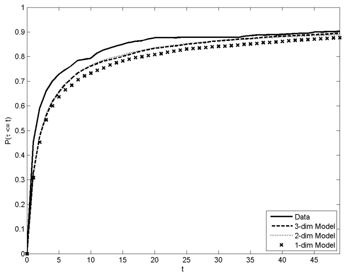
We moreover compute two quantities which serve as a measure of how well the data CDF is approximated by the model CDF. Firstly, we compute the (truncated) L^2-distance between data and model CDF in the range $t \in [0, T_{\text{end}}]$ milli-seconds. Due to the different time range for BJS, we divide the result by 10, so that it can be compared to other stocks. The results are displayed in table 5.4. Secondly, we compute the (truncated) Kolmogorov-Smirnov (K-S) statistic of data and model CDF in the range $t \in [0, T_{\text{end}}]$ milli-seconds. A small K-S value corresponds to a good fit of the CDF. The results are displayed in table 5.5.

Figure 5.4 shows the CDFs for AAPL, with fixed volume-imbalance $\mathfrak{z}_0 = 0$ and initial spread ranging from 1 to 5 ticks. Overall, the CDFs from the three models are very close to the CDF from the data. This suggests that our calibration approach works well. In plot 5.4c, with initial spread $\mathfrak{s}_0 = 0.03$, there is almost no difference between the three models. This is due to the fact that the average spread size is about 0.03 and the average volume-imbalance is zero. The average time-to-fill (which is approximated by the 1-dim model that does not take the spread into account) is close to the time-to-fill conditional on the average spread size (i.e., $\mathfrak{s}_0 = 0.03$) and on the average volume-imbalance (i.e., $\mathfrak{z}_0 = 0$). However, when we consider initial spread values other than 0.03, we see that the 2- and 3-dim model perform much better than the 1-dim model. The smaller the initial spread $\mathfrak{s}_0$, the faster limit orders are filled. This behaviour is reproduced by the 2- and 3-dim model, where the spread is modelled as a mean-reverting CIR process. In all plots, there is no large difference between the 2- and 3-dim model. Again, this is due to the fact that the average volume-imbalance is close to zero, and thus the 2-dim model which is independent of the initial volume-imbalance has a similar behaviour as the 3-dim model.

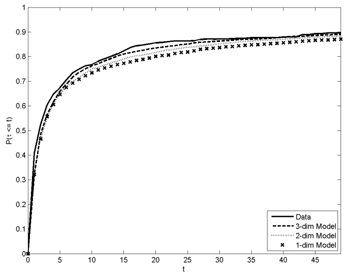
Figure 5.5 shows CDFs for AAPL with fixed initial spread $\mathfrak{s}_0 = 0.02$ and initial volume-imbalance ranging from -2 to 2 (when rounded to the nearest integer). Here it becomes obvious that the additional dimension *volume-imbalance* in the state space of the 3-dim model significantly enhances the model, as compared to the 2-dim model. From the data CDFs, we see that the higher the initial volume-imbalance, the faster limit orders (placed on the best ask) are filled. This behaviour is reproduced by the 3-dim model, where volume-imbalance is modelled as a mean-reverting OU process. The 1- and 2-dim model, which are independent of the initial volume-imbalance, do not display this property.

Figure 5.6 displays the CDFs for BJS. As in the case of AAPL, the model CDFs fits well to the data CDF. Moreover, the characteristic spread dependence of the time to fill is reproduced by the 2- and 3-dim model.

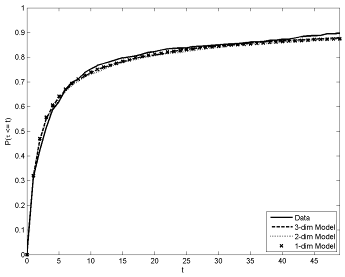
CDFs for MSFT are displayed in figure 5.7. While the model CDFs are not as close to the data CDF when comparing to AAPL and BJS, the 2- and 3-dim model also reproduce the dependence on the initial spread.



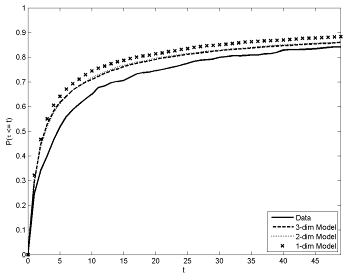
(a) $\mathfrak{s}_0 = 1$ tick



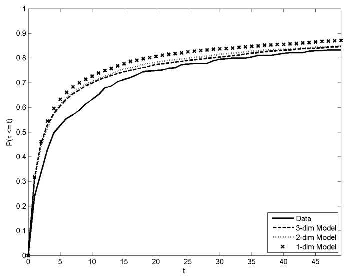
(b) $\mathfrak{s}_0 = 2$ ticks



(c) $\mathfrak{s}_0 = 3$ ticks



(d) $\mathfrak{s}_0 = 4$ ticks

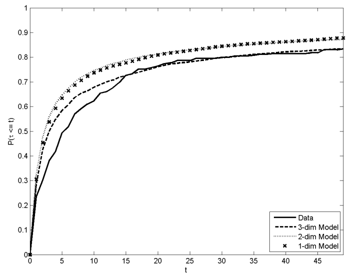


(e) $\mathfrak{s}_0 = 5$ ticks

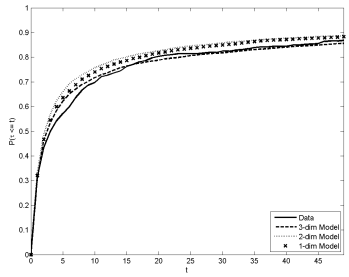
Figure 5.4: $\mathrm{P}^{(\mathfrak{s}_0,0)}(\tilde{\tau}_1 \leq t)$ with varying $\mathfrak{s}_0$ for AAPL data and three models

Stock symbol	$\mathfrak{s}_0$	$\mathfrak{z}_0$	L^2 3-dim	L^2 2-dim	L^2 1-dim
AAPL	1	0	0.31	0.30	0.43
AAPL	2	0	0.13	0.20	0.29
AAPL	3	0	0.09	0.12	0.10
AAPL	4	0	0.36	0.36	0.52
AAPL	5	0	0.30	0.32	0.46
AAPL	2	-2	0.29	0.57	0.53
AAPL	2	-1	0.12	0.30	0.21
AAPL	2	$+1$	0.25	0.43	0.49
AAPL	2	$+2$	0.27	0.53	0.55
BJS	1	0	0.25	0.29	0.17
BJS	2	0	0.20	0.22	0.69
MSFT	1	0	0.54	0.67	0.86
MSFT	2	0	0.86	1.00	2.16

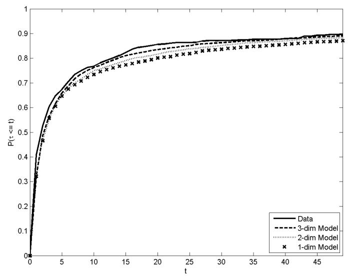
Table 5.4: L^2 comparison of time-to-fill CDF for data and model



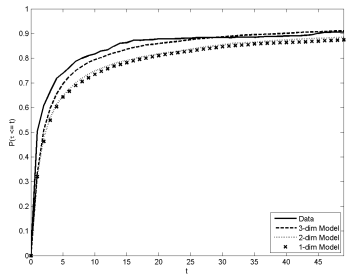
(a) $\mathfrak{z}_0 = -2$



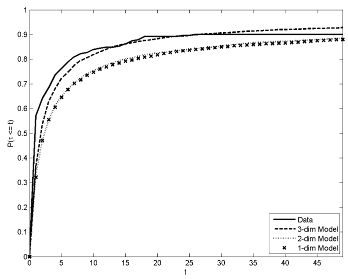
(b) $\mathfrak{z}_0 = -1$



(c) $\mathfrak{z}_0 = 0$

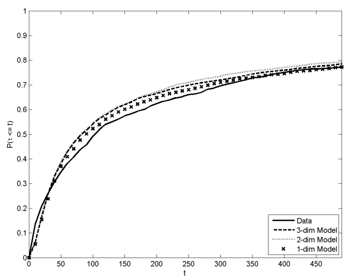


(d) $\mathfrak{z}_0 = +1$

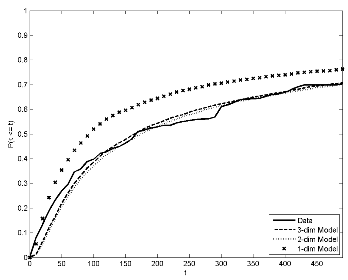


(e) $\mathfrak{z}_0 = +2$

Figure 5.5: $\mathrm{P}^{(0.02,\mathfrak{z}_0)}(\tilde{\tau}_1 \leq t)$ with varying $\mathfrak{z}_0$ for AAPL data and three models

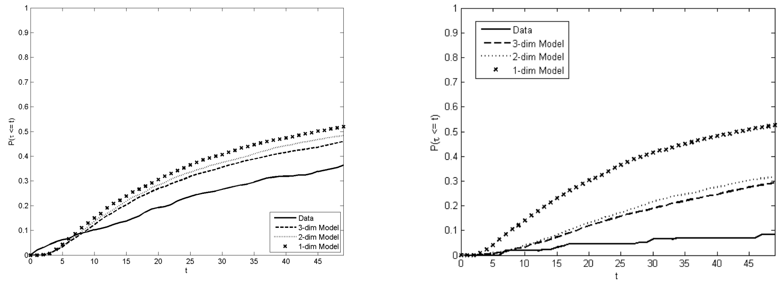


(a) $\mathfrak{s}_0 = 1$ tick



(b) $\mathfrak{s}_0 = 2$ ticks

Figure 5.6: $\mathrm{P}^{(\mathfrak{s}_0,0)}(\tilde{\tau}_1 \leq t)$ with varying $\mathfrak{s}_0$ for BJS data and three models



(a) $\mathfrak{s}_0 = 1$ tick (b) $\mathfrak{s}_0 = 2$ ticks

Figure 5.7: $\mathrm{P}^{(\mathfrak{s}_0,0)}(\tilde{\tau}_1 \leq t)$ with varying $\mathfrak{s}_0$ for MSFT data and three models

Stock symbol	$\mathfrak{s}_0$	$\mathfrak{z}_0$	K-S 3-dim	K-S 2-dim	K-S 1-dim
AAPL	1	0	0.13	0.14	0.14
AAPL	2	0	0.09	0.08	0.09
AAPL	3	0	0.04	0.04	0.05
AAPL	4	0	0.12	0.12	0.15
AAPL	5	0	0.12	0.11	0.13
AAPL	2	−2	0.13	0.19	0.18
AAPL	2	−1	0.05	0.09	0.07
AAPL	2	+1	0.17	0.18	0.18
AAPL	2	+2	0.21	0.24	0.25
BJS	1	0	0.07	0.07	0.08
BJS	2	0	0.07	0.08	0.14
MSFT	1	0	0.10	0.13	0.17
MSFT	2	0	0.21	0.24	0.44

Table 5.5: Kolmogorov-Smirnov statistic of time-to-fill CDF for data and model

Part II

Optimal trading strategies in limit order books

Chapter 6

Optimal trading strategies with market orders

In the previous chapters we focused on an external description of the orderbook and its dynamics. We took the position of an outsider who considers all traders to be equal and has the objective to describe the typical evolution of the orderbook without her interfering with the dynamics.

In this and the following chapters we will take the position of one particular trader. Again, we need to specify the typical evolution of the orderbook *without* intervention of the particular trader. Then we will describe how the trader can interact with the orderbook (e.g. submit market orders, place and cancel limit orders) and how this affects the state of the orderbook. These interactions of the particular trader will be called *control*. We need to specify which controls are admissible, and how to define the trading costs associated to an admissible control. This provides us with a control problem which we shall analyse in detail. In particular we will look at the question of existence and uniqueness of an optimal control (i.e. a control with minimal cost), and we will characterize the optimal control.

6.1 Model assumptions and problem formulation

In this chapter we will consider the problem of optimal execution of a possibly very large order: Suppose a trader wants to buy $x_0 > 0$ shares in a fixed time interval $[0, T]$ at minimum price. She could simply submit a large market order (block trade) of size x_0 at some time between 0 and T, however, due to the price impact of a market order, this drives the best ask price up considerably, unless there is enough liquidity (i.e. sell limit orders awaiting execution) on the best ask. To reduce price impact, it is often a good idea to split the order into smaller pieces. The challenge is to find the optimal *size* and *timing* of the suborders. Similar problems have for example been investigated in Almgren and Chriss (2001), Obizhaeva and Wang (2005) and Predoiu, Shaikhet, and Shreve (2010).

We will work with a version of the order book model developed in chapter 3. To make the analysis tractable, we shall introduce three simplifying assumptions.

Simplifying assumption 6.1. We work in the two-dimensional model given by (5.1) and (5.2) with $\mathfrak{z} \equiv 0$ and $d_1 = 0$. We moreover assume that $V^{\mathfrak{A}} \equiv V^{\mathfrak{B}} \equiv A > 0$.

As we saw in chapter 5, the two-dimensional model, which explicitly models best bid and ask (or, equivalently, spread and midquote-price), but assumes a constant (and equal) volume on both sides of the order book showed good results and captured most of the properties of the more general three-dimensional model.

Before the trader starts trading in the market, the order book dynamics are therefore given by

$$\begin{aligned} \mathrm{d}\mathfrak{s}(t) &= \kappa(\mu - \mathfrak{s}(t))\mathrm{d}t + \frac{\sigma_1}{\sqrt{2}}\sqrt{\mathfrak{s}(t)}\mathrm{d}B_1(t), \\ \mathrm{d}\mathfrak{m}(t) &= \frac{\sigma_1}{2\sqrt{2}}\sqrt{\mathfrak{s}(t)}\mathrm{d}B_2(t) + \frac{\sigma_0}{2}\mathrm{d}B_0(t). \end{aligned}$$

The model parameters $(\kappa, \mu, \sigma_0, \sigma_1)$ can easily be fitted to market data by using the methods of chapter 5.

We also require

Simplifying assumption 6.2. We only admit pure buy strategies that trade exclusively with market orders.

In chapter 7, we will consider the same problem, but allow strategies that include trading with both market orders and limit orders. However, this generalization will not come without a cost: we will need to work in a much simpler model, where the spread is assumed to be zero except for infinitesimal small times, when the impact of a market order widens the spread. In this chapter, however, we model the spread (and therefore the resilience of the order book) explicitly, but have to restrict to trading with market orders.

Moreover we will make the third

Simplifying assumption 6.3. We consider the space of deterministic strategies.

Considering the space of deterministic strategies is clearly a strong assumption. However, some aspects of the problem can only be analysed when we restrict ourselves to the space of deterministic strategies. Note also that in the next chapter we will look at optimal previsible trading strategies for the liquidation problem we consider here (in a different orderbook model).

We thus introduce the set of admissible market order trading strategies:

$$\begin{aligned} \Theta(t,x) = \{\theta : [t,T] \to \mathbb{R}_+,\ &\theta \text{ càglàd} \\ &\theta(t) = 0,\ \theta(T) = x,\ \theta \text{ increasing} \\ &\theta(s) = \sum_{t \le u \le s} \Delta\theta(u) + \int_t^s \dot{\theta}(u)\mathrm{d}u\} \end{aligned} \tag{6.1}$$

Hence any $\theta \in \Theta(t,x)$ is composed of a pure jump part, and a part $\dot{\theta}$ which is absolutely continuous with respect to Lebesgue measure. $\theta \in \Theta(t,x)$ is an admissible strategy for buying x shares in the time interval $[t,T]$. For $t \le s \le T$, $\theta(s)$ denotes the amount of shares already bought up to time s.

Under our simplifying assumptions, what is the effect on the spread and midquote price, when the trader submits a (possibly very large) market order of size $m > 0$ at time t? The spread is increased by x and the midquote price by $\frac{x}{2}$, where x satisfies $m = Ax$. Thus, right after the trade, we have

$$\begin{aligned}\mathfrak{s}(t+) &= \mathfrak{s}(t) + \frac{m}{A},\\ \mathfrak{m}(t+) &= \mathfrak{m}(t) + \frac{m}{2A}.\end{aligned}$$

Given a market order trading strategy $\theta \in \Theta(t, x)$ the *controlled* orderbook model dynamics are therefore a combination of the uncontrolled orderbook model from chapter 3 and the effect of the control on the spread. It is given by the SDEs

$$\mathrm{d}\mathfrak{s}_\theta(u) = \kappa(\mu - \mathfrak{s}_\theta(u))\mathrm{d}u + \frac{\sigma_1}{\sqrt{2}}\sqrt{\mathfrak{s}_\theta(u)}\mathrm{d}B_1(u) + \frac{1}{A}\mathrm{d}\theta(u) \tag{6.2}$$

$$\mathrm{d}\mathfrak{m}_\theta(u) = \frac{\sigma_1}{2\sqrt{2}}\sqrt{\mathfrak{s}_\theta(u)}\mathrm{d}B_2(u) + \frac{\sigma_0}{2}\mathrm{d}B_0(u) + \frac{1}{2A}\mathrm{d}\theta(u) \tag{6.3}$$

with initial conditions $\mathfrak{s}_\theta(t) = \mathfrak{s}$ and $\mathfrak{m}_\theta(t) = \mathfrak{m}$. In the following we will often drop the subscript θ when there is no risk of confusion.

The trader wants to minimize her expected trading costs

$$\mathcal{J}_1(t, \mathfrak{s}, \mathfrak{m}, x; \theta) = \mathrm{E}^{(\mathfrak{s},\mathfrak{m})}\left[\int_t^T \left(\mathfrak{m}_\theta(u) + \frac{\mathfrak{s}_\theta(u)}{2} + \frac{1}{2A}\Delta\theta(u)\right)\mathrm{d}\theta(u)\right] \tag{6.4}$$

where $\mathrm{E}^{(\mathfrak{s},\mathfrak{m})}$ denotes the expectation conditional on $\mathfrak{s}(t) = \mathfrak{s}$ and $\mathfrak{m}(t) = \mathfrak{m}$.

Due to the martingale property of the midquote price, the performance function $\mathcal{J}_1$ may be simplified as follows:

Proposition 6.1 (Transformation of performance function)**.** *Let $\theta \in \Theta(t, x)$. Then the performance function can be written as*

$$\mathcal{J}_1(t, \mathfrak{s}, \mathfrak{m}, x; \theta) = \frac{1}{2}\mathcal{J}_2(t, \mathfrak{s}, x; \theta) + \mathfrak{m}x + \frac{x^2}{4A}$$

where

$$\mathcal{J}_2(t, \mathfrak{s}, x; \theta) = \int_t^T \left(s_\theta(u) + \frac{1}{2A}\Delta\theta(u)\right)\mathrm{d}\theta(u), \tag{6.5}$$

where $s_\theta(t) = \mathfrak{s}$ and the dynamics of s_θ are given by

$$\mathrm{d}s_\theta(u) = \kappa(\mu - s_\theta(u))\mathrm{d}t + \frac{1}{A}\mathrm{d}\theta(u). \tag{6.6}$$

Thus, for fixed $(t, \mathfrak{s}, \mathfrak{m}, x)$, $\mathcal{J}_1$ and $\mathcal{J}_2$ only differ by multiplication with $1/2$ and addition of a constant, and we will henceforth consider the equivalent performance function $\mathcal{J}_2$.

We are interested in finding the optimal trading strategy $\hat{\theta} \in \Theta(t, x)$ satisfying

$$\mathcal{J}_2(t, \mathfrak{s}, x; \hat{\theta}) = \inf_{\theta \in \Theta(t,x)} \mathcal{J}_2(t, \mathfrak{s}, x; \theta),$$

and in calculating the corresponding value function

$$\mathcal{V}(t,\mathfrak{s},x) = \inf_{\theta\in\Theta(t,x)} \mathcal{J}_2(t,\mathfrak{s},x;\theta) \tag{6.7}$$

In (6.5) we see that $\mathcal{J}_2$ only depends on the spread, and not on the midquote price. Intuitively, the optimal strategy $\hat{\theta}$ should thus satisfy the below rules of thumb:

- Buy fast, when the spread is small,
- buy slowly, when the spread is large, and
- wait for a more favourable spread, when it is too large.

In practice, the trader would also place limit orders when the spread is large, instead of slowing down (or stopping) trading. Often traders use the following strategy: they place a limit order and wait a certain time for execution of the limit order. If it is not executed and either time is short, or the spread is relatively small, then the limit order is cancelled and converted into a market order which is immediately executed at the best price available. This is clearly beyond the scope of this chapter. In chapter 8, however, we will analyse this problem in detail, using a modelling framework similar to the one used here. In particular, we will compute the optimal time when a limit order should be converted to a market order.

In the rest of this chapter, we compute the value function $\mathcal{V}$ and the optimal strategy $\hat{\theta} \in \Theta(t,x)$ which minimizes (6.4). We proceed in four steps: first, we use the Euler-Lagrange method to derive candidate optimal strategies. We will see that these strategies depend on the initial value of $\frac{\mathfrak{s}-\mu}{x}$. The candidate optimal strategies obtained by the Euler-Lagrange method also tell us the regions in which the optimal strategies take different forms (e.g. 'wait' instead of 'buy'). In the second step we transform the value function and reduce the number of dimension from two to three. Thirdly, we use the associated HJB equation of the simplified value function from step two and the regions obtained in step one to calculate a candidate value function. Finally, in the fourth step, the candidate value function is shown to be the true value function.

6.2 Derivation of candidate strategy using Euler-Lagrange method

The Euler-Lagrange method is a powerful tool from calculus of variations used for solving optimization problems in which one seeks a function that minimizes a given functional under certain constraints. In our setting the sought function is the trading strategy, the functional is the performance function $\mathcal{J}_2$ given in (6.5) and the constraint is that all shares have to be bought before expiration of the time limit, and the trader is not allowed to sell shares. More detailed information on the calculus of variation, and applications of the Euler-Lagrange method to optimal control problems can be found for example in Gregory and Lin (1996). Note that we use the Euler-Lagrange method in a purely heuristic fashion in this section. In section 6.3 we will give a precise proof of the correctness of our heuristic results obtained below.

To apply the Euler-Lagrange method, we first look at the constrained problem without jumps.

Thus we want to minimize

$$\int_0^T s(t)\dot{\theta}(t)\mathrm{d}t$$

where

$$\dot{s}(t) = \kappa(\mu - s(t)) + \frac{\dot{\theta}(t)}{A}$$

with so called isoperimetric constraint

$$\int_0^T \dot{\theta}(t)\mathrm{d}t = x_0,$$

i.e. all shares have to be bought within the given time horizon.

We now reformulate the problem in terms of $(s, \dot{s})$. Replace $\dot{\theta}(t) = A(\dot{s}(t) - \kappa(\mu - s(t)))$, and rewrite the problem as minimizing

$$\int_0^T f(t, s(t), \dot{s}(t))\mathrm{d}t, \tag{6.8}$$

with isoperimetric contraint

$$\int_0^T g(t, s(t), \dot{s}(t))\mathrm{d}t = x_0, \tag{6.9}$$

(i.e. all shares have to be bought at time T) and positivity constraint

$$g(t, s, \dot{s}) \geq 0, \tag{6.10}$$

(i.e. the trader can only use buy market orders and not sell), where

$$\begin{aligned} f(t, s, \dot{s}) &= As(\dot{s} - \kappa(\mu - s)), \\ g(t, s, \dot{s}) &= A(\dot{s} - \kappa(\mu - s)). \end{aligned}$$

To find the optimal value $s = \hat{s}$, we introduce the Lagrange multipliers $\tilde{\lambda}$ (a constant to be determined) and $\lambda(t)$ (a function to be determined) and the auxiliary function

$$F(t, s, \dot{s}) = f(t, s, \dot{s}) + \tilde{\lambda} g(t, s, \dot{s}) - \lambda(t) g(t, s, \dot{s}).$$

From Gregory and Lin (1996) we have the following necessary conditions that the optimal $\hat{s}$ has to satisfy:
The Euler-Lagrange equation

$$\frac{\mathrm{d}}{\mathrm{d}t} F_{\dot{s}}(t, s, \dot{s}) = F_s(t, s, \dot{s}), \tag{EL1}$$

the Kuhn-Tucker conditions

$$\lambda(t) g(t, s(t), \dot{s}(t)) = 0, \tag{EL2}$$

$$g(t, s(t), \dot{s}(t)) \geq 0, \tag{EL3}$$

and the transversality condition at the right boundary

$$F_{\dot{s}}(T+, s(T+), \dot{s}(T+)) = 0, \tag{EL4}$$

which simplify to

$$\dot{\lambda}(t) = \kappa(\mu - 2s(t) - \tilde{\lambda} + \lambda(t)), \tag{EL1'}$$
$$0 = \lambda(t)A(\dot{s}(t) - \kappa(\mu - s(t))), \tag{EL2'}$$
$$0 \leq A(\dot{s}(t) - \kappa(\mu - s(t))), \tag{EL3'}$$
$$\tilde{\lambda} = \lambda(T+) - s(T+). \tag{EL4'}$$

Now we combine (EL1') and (EL4') to eliminate $\tilde{\lambda}$, and differentiate with respect to time t to obtain

$$s(t) = \frac{\mu + \lambda(t) + s(T+) - \lambda(T+)}{2} - \frac{\dot{\lambda}(t)}{2\kappa}, \tag{6.11}$$
$$\dot{s}(t) = \frac{\dot{\lambda}(t)}{2} - \frac{\ddot{\lambda}(t)}{2\kappa}. \tag{6.12}$$

Next, we plug (6.11) and (6.12) into (EL2'), and we obtain the following ODE for λ

$$\lambda(t)A\left(\frac{\kappa(\mu + \tilde{\lambda})}{2} - \frac{\kappa}{2}\lambda(t) + \frac{1}{2\kappa}\ddot{\lambda}(t)\right) = 0 \tag{6.13}$$

which can be solved explicitly: $\lambda(t) \equiv 0$ or $\lambda(t) = \mu + \tilde{\lambda} + c_1 e^{t\kappa} + c_2 e^{-t\kappa}$, for some constants c_1, c_2 to be determined.

Let us investigate what the two solutions for λ mean in terms of our trading strategy:

- For $\lambda(t) \equiv 0$, we infer from (EL2) and (EL3) that we can have $g(t, s(t), \dot{s}(t)) > 0$ which means that we are buying shares at a constant rate. Hence we are in a *buying region.*
- For $\lambda(t) = \mu + \tilde{\lambda} + c_1 e^{t\kappa} + c_2 e^{-t\kappa} \neq 0$, (EL2) and (EL3) imply that we must have $g(t, s(t), \dot{s}(t)) = 0$ which means that we are not buying shares. Hence we are in a *waiting region.*

Let us now consider both cases in detail and investigate for which initial parameters we are in which of the two regions, and how the buying/waiting is done exactly.

6.2.1 Buying region

We first consider the case $\lambda(t) \equiv 0$. Then (EL1') gives an implicit equation of the optimal spread level $\hat{s}$ in terms of $\tilde{\lambda}$

$$\hat{s}(t) = \frac{\mu - \tilde{\lambda}}{2},\ t \in (0, T)$$

Note that $\hat{s}(t) \equiv \hat{s}$ is constant.

(EL4') gives $\tilde{\lambda} = -s(T+)$. Hence we have $\hat{s} = \frac{\mu + \hat{s}(T+)}{2}$. Now we introduce the jump parts in the candidate optimal strategy. We guess that the optimal strategy has at most two jumps at time $t = 0$ and $t = T$ of size $\Delta\theta(0)$ and $\Delta\theta(T)$ to be determined.

Since $\hat{s}(0+) = s(0) + \Delta\theta(0)/A$ and $s(T+) = \hat{s}(T-) + \Delta\theta(T)/A$, we have

$$\Delta\theta(0) = A\left(\frac{s(T+)+\mu}{2} - s(0)\right) \text{ and } \Delta\theta(T) = A\left(s(T+) - \frac{s(T+)+\mu}{2}\right) \tag{6.14}$$

Using $\dot{\theta}(t) = \frac{A\kappa}{2}(\mu + s(T+))$ and the isoperimetric constraint

$$\Delta\theta(0) + \int_0^T \dot{\theta}(t)\mathrm{d}t + \Delta\theta(T) = x_0$$

we compute $s(T+) = \frac{1}{A(2+\kappa)}(2x_0 + 2A\mathfrak{s}_0 + A\kappa\mu T)$, from which we infer the optimal spread $\hat{s} = \frac{1}{A(2+\kappa)}(x_0 + A(\mathfrak{s}_0 + \mu) + A\kappa\mu T)$ and the candidate optimal strategy

$$\Delta\theta(0) = \frac{1}{2+\kappa T}(x_0 - A(1+T\kappa)(\mathfrak{s}_0 - \mu)) \tag{6.15}$$

$$\Delta\theta(T) = \frac{1}{2+\kappa T}(x_0 + A(\mathfrak{s}_0 - \mu)) \tag{6.16}$$

$$\dot{\theta} = \kappa\Delta\theta(T) = \frac{\kappa}{2+\kappa T}(x_0 + A(\mathfrak{s}_0 - \mu)) \tag{6.17}$$

However, this strategy is only admissible for initial parameters $(\mathfrak{s}_0, x_0)$ which satisfy $-\frac{1}{A} \leq \frac{\mathfrak{s}_0 - \mu}{x_0} \leq \frac{1}{A(1+\kappa T)}$, because we only admit pure buy strategies, i.e. we require $\Delta\theta(0) \geq 0, \dot{\theta} \geq 0$ and $\Delta\theta(T) \geq 0$. If this is satisfied, the candidate optimal strategy consists in trading from time 0 to time T. There are no waiting regions, where it is optimal to wait for a better (that is smaller) spread.

6.2.2 Waiting region

Let us now analyse what happens for other parameter combinations. We consider the non-zero solution $\lambda(t) = \mu + \tilde{\lambda} + c_1 e^{t\kappa} + c_2 e^{-t\kappa}$. Note that for $\lambda(t) \neq 0$ (EL2) implies that $g(t, s(t), \dot{s}(t)) = 0$, hence the optimal strategy consists in not trading and waiting for a more favourable spread. In that case (6.11) and (EL2) give

$$s(t) = \mu + c_2 e^{-t\kappa} \tag{6.18}$$

$$\dot{\theta}(t) = 0 \tag{6.19}$$

For $\dot{\theta}(t) = 0$ the spread $s(t)$ can be calculated explicitly from (6.6) and we obtain $c_2 = \mathfrak{s}_0 - \mu$. Moreover, when not trading is optimal, we have $s(T) = s(T+)$, and we obtain $c_1 = 0$ by combining (EL4'), (6.18) and the solution for $\lambda(t)$.

Then, intuitively, we have the following situation: when the spread is too large, i.e. $\frac{\mathfrak{s}_0 - \mu}{x_0} > \frac{1}{A(1+\kappa T)}$, the solution given by (6.15)- (6.17) violates the positivity constraints. Equations (6.18) and (6.19) say that in this case it is optimal to wait for the spread to decrease. It remains to analyse how long the trader has to wait until she starts trading. Let t_1 be the time when she starts trading. Equation (6.18) implies that $s(t_1) = \mu + (\mathfrak{s}_0 - \mu)e^{-t_1\kappa}$. Moreover, at time t_1 the spread has reached a level such that the strategy (6.15)- (6.17), starting at $(t_1, s(t_1), x_0)$ is optimal, and therefore $s(t_1)$ must be on the boundary of the region where strategy (6.15)- (6.17) is optimal:

$$\frac{s(t_1) - \mu}{x_0} = \frac{1}{A(1+\kappa(T-t_1))}$$

This gives an implicit equation for t_1 in terms of the initial parameters $(\mathfrak{s}_0, x_0)$ and the 'time-to-go' $T - t_1$:

$$(1 + \kappa(T - t_1))e^{-\kappa t_1} = \frac{x_0}{A(\mathfrak{s}_0 - \mu)} \tag{6.20}$$

It is straightforward to check that (6.20) defines a unique $t_1 \in [0, T]$ for $\frac{1}{A(1+\kappa T)} \leq \frac{\mathfrak{s}_0 - \mu}{x_0} \leq \frac{e^{T\kappa}}{A}$.

Hence in this region, the candidate optimal strategy is given by

$$\Delta\theta(0) = 0 \tag{6.21}$$
$$\dot{\theta}(t) = 0, \text{ for } t \in [0, t_1) \tag{6.22}$$
$$\dot{\theta}(t) = \kappa\Delta\theta(T) = \frac{x_0\kappa}{1 + \kappa(T - t_1)}, \text{ for } t \in [t_1, T) \tag{6.23}$$
$$\Delta\theta(T) = \frac{x_0}{1 + \kappa(T - t_1)} \tag{6.24}$$

where t_1 is given implicitely by (6.20).

Finally, for $\frac{\mathfrak{s}_0 - \mu}{x_0} > \frac{e^{T\kappa}}{A}$, we guess the optimal strategy $\Delta\theta(0) = 0$, $\dot{\theta}(t) = 0$ $t \in [0, T]$ and $\Delta\theta(T) = x_0$ since the spread is so large that it is optimal to wait until time T and submit a single buy order. Similarly, for $\frac{\mathfrak{s}_0 - \mu}{x_0} < -\frac{1}{A}$, we guess the optimal strategy $\Delta\theta(0) = x_0$, $\dot{\theta}(t) = 0$ $t \in [0, T]$ and $\Delta\theta(T) = 0$ since the spread is so small (compared to average) that it is optimal to submit a single buy order at time $t = 0$.

6.2.3 Characterization of optimal strategy

Let us summarize our findings. Set $y(t) := \frac{s(t) - \mu}{x(t)}$. Then $(t, y(t))$ lies in the state space $S = [0, T] \times \mathbb{R}$. S can be partitioned in disjoints subsets

$$\begin{aligned}
S_1 &= \left\{(t, y) \in S : y \leq -\frac{1}{A}\right\} \\
S_2 &= \left\{(t, y) \in S : -\frac{1}{A} \leq y < \frac{1}{A(1 + \kappa(T - t))}\right\} \\
S_3 &= \left\{(t, y) \in S : y = \frac{1}{A(1 + \kappa(T - t))}\right\} \\
S_4 &= \left\{(t, y) \in S : \frac{1}{A(1 + \kappa(T - t))} < y < \frac{e^{\kappa(T-t)}}{A}\right\} \\
S_5 &= \left\{(t, y) \in S : \frac{e^{\kappa(T-t)}}{A} \leq y\right\}
\end{aligned} \tag{6.25}$$

The corresponding candidate optimal strategies derived above are

Interpretation	O-W	A-F-S	2-dim
Volume on best ask	q	q	A
Resilience speed	ρ	ρ	κ
Permanent impact	λ	γ	$\frac{1}{2A}$
Temporary impact	$\kappa = \frac{1}{q} - \lambda$	$\kappa = \frac{1}{q} - \gamma$	$\frac{1}{2A}$

Table 6.1: Terminology used in Obizhaeva and Wang, Alfonsi, Fruth, and Schied and in this chapter.

Initial value	Strategy
$(0, y(0)) \in S_1$	Single initial trade of size $\Delta\theta(0) = x_0$
$(0, y(0)) \in S_2$	Initial trade, constant continuous trade intensity $\dot{\theta}$, final trade as given in (6.15)- (6.17)
$(0, y(0)) \in S_3$	No initial trade, constant continuous trade intensity $\dot{\theta}$, final trade as given in (6.16)- (6.17)
$(0, y(0)) \in S_4$	No initial trade, continuous trade intensity $\dot{\theta}1_{[t_1,T]}$, final trade as given in (6.21)- (6.24)
$(0, y(0)) \in S_5$	Single final trade of size $\Delta\theta(T) = x_0$

Interestingly, it turns out that for one special value of the initial spread, these results are identical to a well-known model. When $\mathfrak{s}_0 = \mu$, or equivalently $y(0) = 0$, we are in the 'Initial trade - Continuous trade - Final trade' region S_2 with

$$\Delta\theta(0) = \frac{x_0}{2 + \kappa T} \tag{6.26}$$

$$\Delta\theta(T) = \frac{x_0}{2 + \kappa T} \tag{6.27}$$

$$\dot{\theta} = \kappa\Delta\theta(T) = \frac{\kappa x_0}{2 + \kappa T} \tag{6.28}$$

This is the only case, where the size of inital and final trade are equal. When comparing this strategy to the result in Proposition 3 of Obizhaeva and Wang (2005), we note that both strategies are the same, when identifying the parameter κ used here with the parameter ρ used in Obizhaeva and Wang (2005), and is a continuous version of the discrete-time result obtained in Alfonsi, Fruth, and Schied (2010, Cor. 6.1) and a special case of the result obtained in Predoiu, Shaikhet, and Shreve (2010). We shall henceforth refer to the case $\mathfrak{s}_0 = \mu$ investigated in Obizhaeva and Wang (2005), Alfonsi, Fruth, and Schied (2008) and Alfonsi, Fruth, and Schied (2010) as the 'classical case'. For clarity, we refer to table 6.1 for the different terminologies used in the two articles, and the corresponding terminology we employ here.

This correspondence is not surprising; in fact, admitting only deterministic strategies turns out to be equivalent to removing the volatility of the spread. Thus the non-random spread is given by s_θ defined in (6.6). When the trader does not buy any shares, the non-random spread is given explicitly by

$$s(t) = \mu + e^{-\kappa t}\left(s(0) - \mu\right)$$

which corresponds exactly to the exponential resilience (i.e. recovery of the spread) at speed ρ in the classical case.

In table 6.1, we see that in the classical case, the permanent impact is a model parameter that can be chosen freely (with the constraint $0 \leq \lambda \leq \frac{1}{q}$ whereas in the two-dimensional model considered here, it is determined by the volume on the best ask and equals $\frac{1}{2A}$. This is a direct consequence of the fact that we model both best bid and best ask, and that we assume that both sides behave symmetrically. This equilibrium assumption is rather natural and immediately implies that the permanent impact parameter for both sides must be the same, and thus equals $\frac{1}{2A}$.

Clearly, the two-dimensional model we consider extends the well-known results by admitting initial spread values with $\mathfrak{s}_0 \neq \mu$. This leads to frontloading or backloading of the strategy depending on the size of the spread at the beginning of the trading program. It is also interesting to note that in the classical case, the optimal trading rate does not depend on the volume on the best ask A. Yet, when the initial spread is above its long time average ($\mathfrak{s}_0 > \mu$), the initial trade is decreasing in the best ask volume, whereas the continuous trading rate and the final trade are both increasing in the best ask volume. This is explained by the fact that the initial spread is unfavourable for immediate trading, and trading is postponed towards the end of the programm. The more volume stocked on the best ask, the less market impact does a market order have. Thus more shares can be bought towards the end of the program, when the spread recovered to a more favourable level without driving up the best ask price too much. However, when the initial spread is below its long time average, we have exactly the opposite effect. For $\mathfrak{s}_0 = 0$ both effects balance out, and the trading intensity does not depend on the volume on the best ask.

Even when staying in the classical case, where $\mathfrak{s}_0 = \mu$, the extended framework of the two-sided model (which considers the interaction between best bid and best ask) gives us some additional insight into the classical model: in fact, neither Obizhaeva and Wang (2005) nor Alfonsi, Fruth, and Schied (2010) discuss how to actually estimate the resilience parameter ρ. By considering both sides of the order book, the meaning of ρ becomes clearer: it measures the speed at which the spread returns to its long-time mean when it is above or below it. Using the methods developed in chapter 5, one can easily find good estimators for ρ (i.e. κ in our terminology) from high-frequency order book data.

6.3 Derivation and verification of value function

Note that up to now, we have only heuristically derived the candidate optimal strategies, and have not proved optimality yet. In this section, we will use these heuristic results to compute a candidate value function and verify that it is the true value function. This will imply that the candidate optimal strategies we found by the Euler-Lagrange method are indeed the true optimal strategies.

From the previous section we know that the important parameters for the value function are $\left(t, \frac{\mathfrak{s}-\mu}{x}\right)$. We start by a transformation of the value function $\mathcal{V}$ given in (6.7) in terms of these parameters. First introduce the centered spread process $\bar{s}_\theta(t) = s_\theta(t) - \mu$. It satisfies the dynamics

$$d\bar{s}_\theta(u) = -\kappa \bar{s}_\theta(u)du + \frac{1}{A}d\theta(u) \tag{6.29}$$

and define the value function

$$\mathcal{V}_2(t, \bar{s}_t, x) = \inf_{\theta \in \Theta(t,x)} \int_t^T \left(\bar{s}_\theta(u) + \frac{1}{2A}\Delta\theta(u) \right) d\theta(u)$$

A simple calculation shows that

$$\mathcal{V}(t,s,x) = \mathcal{V}_2(t,s-\mu,x) + \mu x$$

Moreover, for any $\rho > 0$ we have the homotheticity property

$$\mathcal{V}_2(t,\rho s,\rho x) = \rho^2 \mathcal{V}_2(t,s,x),$$

so we can reduce the dimensions from three to two by setting $\mathcal{V}_3(t,y) = \mathcal{V}_2(t,y,1)$ and noting that

$$\mathcal{V}(t,s,x) = x^2 \mathcal{V}_3\left(t, \frac{s-\mu}{x}\right) + \mu x \tag{6.30}$$

The HJB equation corresponding to $\mathcal{V}_3$ is easily derived and equals

$$\min\{\mathcal{L}_1\mathcal{V}_3, \mathcal{L}_2\mathcal{V}_3\} = 0 \tag{6.31}$$

where

$$\begin{aligned}(\mathcal{L}_1\varphi)(t,y) &= \varphi_t(t,y) - \kappa y \varphi_y(t,y)\\ (\mathcal{L}_2\varphi)(t,y) &= Ay - 2A\varphi(t,y) + (1+Ay)\varphi_y(t,y)\end{aligned}$$

Suppose $\mathcal{V}_3$ is in $C^1(S)$ and satisfies (6.31). Heuristically, the different regions of S correspond to different optimal trading strategies:

Region	Optimal strategy
$\mathcal{L}_1\mathcal{V}_3 > 0$ and $\mathcal{L}_2\mathcal{V}_3 = 0$	discrete trade at time 0 that brings spread up to its optimal level
$\mathcal{L}_1\mathcal{V}_3 = 0$ and $\mathcal{L}_2\mathcal{V}_3 = 0$	trade continously and hold optimal spread level
$\mathcal{L}_1\mathcal{V}_3 = 0$ and $\mathcal{L}_2\mathcal{V}_3 > 0$	do not trade and wait for spread to decrease to its optimal level

Using the candidate optimal strategies computed in section 6.2, it is straightforward to compute a candidate value function $v : S \to \mathbb{R}$ of the HJB equation (6.31): $v(t,y) = v_i(t,y)$ for $(t,y) \in S_i$, $i = 1,\ldots,5$, where

$$\begin{aligned}v_1(t,y) &= y + \frac{1}{2A}\\ v_2(t,y) = v_3(t,y) &= y + \frac{1}{2A} - \frac{\kappa(T-t)(1+Ay)^2}{2A\kappa(T-t)+4A}\\ v_4(t,y) &= \frac{3+2\kappa(T-\phi^{-1}(ye^{\kappa t})}{2A(1+\kappa(T-\phi^{-1}(ye^{\kappa t})))^2}\\ v_5(t,y) &= ye^{-\kappa(T-t)} + \frac{1}{2A}\end{aligned} \tag{6.32}$$

and

$$\phi(t) = \frac{e^{\kappa t}}{A(1+\kappa(T-t))} \tag{6.33}$$

It is tedious, but elementary to check that on $[0,T)\times\mathbb{R}$

$$\begin{aligned}\mathcal{L}_1 v_i > 0 \text{ and } \mathcal{L}_2 v_i = 0 \text{ in } S_i,\ i=1,2\\ \mathcal{L}_1 v_3 = 0 \text{ and } \mathcal{L}_2 v_3 = 0 \text{ in } S_3\\ \mathcal{L}_1 v_i = 0 \text{ and } \mathcal{L}_2 v_i > 0 \text{ in } S_i,\ i=4,5\end{aligned} \tag{6.34}$$

and also holds for $[0,T]\times\mathbb{R}$ when replacing the 'strictly greater than' by 'greater or equal' signs.

Now we can show that

Theorem 6.1. *The value function defined in* (6.7) *is given by*

$$\mathcal{V}(t,s,x)=\begin{cases} xs+\frac{x^2}{2A} & \textit{if } (t,(s-\mu)/x)\in S_1\\ \frac{2x^2-A^2(T-t)\kappa(s-\mu)^2+2Ax(2s+\kappa\mu(T-t)}{2A(2+\kappa(T-t))} & \textit{if } (t,(s-\mu)/x)\in S_2\cup S_3\\ x^2\frac{3+2\kappa(T-\phi^{-1}(\frac{s-\mu}{x}e^{\kappa t})}{2A(1+\kappa(T-\phi^{-1}(\frac{s-\mu}{x}e^{\kappa t})))^2}+\mu x & \textit{if } (t,(s-\mu)/x)\in S_4\\ x(\mu+(s-\mu)e^{-\kappa(T-t)})+\frac{x^2}{2A} & \textit{if } (t,(s-\mu)/x)\in S_5\end{cases} \tag{6.35}$$

The optimal strategy $\hat{\theta}$ equals the candidate optimal strategy derived in section 6.2.

6.4 Proofs

Proof of Proposition 6.1. We have

$$\mathfrak{m}_\theta(t)=\mathfrak{m}(t)+\frac{1}{2A}\theta(t).$$

Also, for $u\in[t,T]$ we have $0\le \mathrm{E}[\mathfrak{s}_\theta(u)]\le \mathfrak{s}e^{-\kappa(u-t)}+\mu(1-e^{-\kappa(u-t)})+x$, and thus, by computing the quadratic variation of the local martingale $u\to\mathfrak{m}(u)$ we see that it is in fact a true square integrable martingale. Now we use integration by parts and obtain

$$\int_t^T \mathfrak{m}(u)\mathrm{d}\theta(u)=\mathfrak{m}(T)\theta(T)-\mathfrak{m}(0)\theta(0)-\int_t^T\theta(u)\mathrm{d}\mathfrak{m}(u).$$

We take expectation and find that

$$\mathrm{E}\left[\int_t^T \mathfrak{m}(u)\mathrm{d}\theta(u)\right]=\mathfrak{m}(0)x \tag{6.36}$$

Again by integration by parts we have

$$\theta^2(T)-\theta^2(t)=2\int_t^T\theta(u)\mathrm{d}\theta(u)+\sum_{t\le u\le T}(\Delta\theta(u))^2$$

and thus

$$\frac{1}{2A}\int_t^T\theta(u)\mathrm{d}\theta(u)=\frac{1}{4A}x^2-\frac{1}{4A}\sum_{t\le u\le T}(\Delta\theta(u))^2 \tag{6.37}$$

Plug (6.36) and (6.37) into (6.4) to obtain

$$\mathcal{J}_1(t,\mathfrak{s},\mathfrak{m},x;\theta)=\frac{1}{2}\mathrm{E}^{\mathfrak{s}}\left[\int_t^T\left(\mathfrak{s}_\theta(u)+\frac{1}{2A}\Delta\theta(u)\right)\mathrm{d}\theta(u)\right]+\mathfrak{m}x+\frac{x^2}{4A}$$

Now note that $\mathrm{d}s_\theta(t) = \mathrm{d}\mathfrak{s}_\theta(t) + \mathrm{d}M(t)$ where M is a true martingale, so we can compute the expectation, using the fact that θ is deterministic and we obtain the required form for $\mathcal{J}_2$ and the dynamics for s. □

Proof of Theorem 6.1. Suppose we can show that $\mathcal{V}_3 = v$ as defined in equation (6.32). The formula (6.35) is then obtained by using the identity $\mathcal{V}(t, s, x) = x^2\mathcal{V}_3\left(t, \frac{s-\mu}{x}\right) + \mu x$. Hence it suffices to prove that $\mathcal{V}_3 = v$. Since v corresponds to the value of the candidate optimal strategy obtained by the Euler-Lagrange method (which is admissible by construction), it is enough to show that for any $\theta \in \Theta(t, x)$, we have

$$Y(T) \geq x^2 v\left(t, \frac{\bar{s}(t)}{x}\right), \tag{6.38}$$

where for $\tau \in [t, T]$

$$Y(\tau) = \int_t^\tau \left(\bar{s}_\theta(u) + \frac{1}{2A}\Delta\theta(u)\right)\mathrm{d}\theta(u) + x^2(\tau)v\left(\tau, \frac{\bar{s}_\theta(\tau)}{x(\tau)}\right).$$

It is tedious, but elementary to show that v is continuously differentiable *everywhere* in S. Thus we can apply integration by parts and obtain

$$\begin{aligned} Y(\tau) =& x^2 v\left(t, \frac{\bar{s}_t}{x}\right) + \int_t^\tau x^2(u)(\mathcal{L}_1 v)(u, y(u)))\mathrm{d}u + \frac{1}{A}\int_t^\tau x(u)\dot{\theta}(u)(\mathcal{L}_2 v)(u, y(u)))\mathrm{d}u + \\ & \sum_{t\leq u\leq\tau}\left\{\bar{s}_\theta(u-)\Delta\theta(u) + \frac{1}{2A}\left(\Delta\theta(u)\right)^2 + \Delta\left(x^2(u)v(u, y(u))\right)\right\} \\ =& x^2 v\left(t, \frac{\bar{s}_t}{x}\right) + \int_t^\tau I_1(u)\mathrm{d}u + \frac{1}{A}\int_t^\tau I_2(u)\mathrm{d}u + \sum_{t\leq u\leq\tau} I_3(u) \end{aligned}$$

where $y(u) = \frac{\bar{s}_\theta(u)}{x(u)}$. From (6.34) we immediately have that $I_1(u) \geq 0$, $I_2(u) \geq 0$. Next we show that also $I_3(u) \geq 0$. Suppose that there is a jump of size $z \in [0, x]$ at time u and $s(u-) = s$, $x(u-) = x$. Then $I_3(u) \geq 0$ if and only if $g(z) \geq 0$, where

$$g(z) := sz + \frac{z^2}{2A} + (x-z)^2 v\left(u, \frac{s + z/A}{x - z}\right) - x^2 v\left(u, \frac{s}{x}\right)$$

Since $g(0) = 0$, it is enough to show that $g^{'}(z) \geq 0$ in $[0, x]$. Note that

$$g^{'}(z) = s + \frac{z}{A} - 2(x-z)v(u, y_0) + \left(s + \frac{x}{A}\right)v_y(u, y_0)$$

where $y_0 = \frac{s+z/A}{x-z}$.

Multiply by $\frac{A}{x-z}$ to see that this is equivalent to showing that

$$Ay_0 - 2Av(u, y_0) + (1 + Ay_0)\, v_y(u, y_0) \geq 0.$$

But this last inequality holds by equation (6.34), and so we have that $I_3(u) \geq 0$.

We obtain for any admissible θ the inequality

$$\int_t^\tau \left(\bar{s}_\theta(u) + \frac{1}{2A}\Delta\theta(u)\right)\mathrm{d}\theta(u) + x_\tau^2 v\left(\tau, \frac{\bar{s}_\theta(\tau)}{x(\tau)}\right) \geq x^2 v\left(t, \frac{\bar{s}_t}{x}\right)$$

and thus (6.38).

□

Chapter 7

Optimal trading using market and limit orders with partial filling

Similarly as in chapter 6, we analyse optimal trading strategies in a limit order book. However, in this chapter, the trader can use both market and limit orders to buy and sell asset, and we further simplify our model. Furthermore, in constrast to the last chapter, where the trader was risk-neutral, we will allow a risk-averse trader.

7.1 Model assumptions and trading costs

In order to make the analysis tractable, we will work in a simplified model. As in the last chapter we assume a block-shaped model as given by equation (2.24). The first main difference to chapter 6 is that we assume that the spread $\mathfrak{s}$ is constant and equal to zero (so $\alpha \equiv \beta \equiv \mathfrak{m}$) and that the midquote price is a Brownian motion with zero drift: $\mathfrak{m}(t) = \sigma B(t)$ for some $\sigma > 0$ and a standard Brownian motion B. Moreover, the volume on both sides is assumed to be constant $V^{\mathfrak{A}}(t) \equiv V^{\mathfrak{B}}(t) \equiv \frac{1}{2\mu}$, for some constant μ. We will always assume $\mu > 0$, since in the case $\mu = 0$ there is infinite volume stored on the best quotes and any arbitrarily large amount of assets can be bought at the midquote price, without increasing it.

Let us first look at trading costs associated to market orders. Suppose a buy market order of size $m > 0$ is submitted at time t. As we are working in the block-shaped model, the best ask is driven up to $\mathfrak{m}(t) + 2\mu m$, i.e. the part of the order book between $\mathfrak{m}(t)$ and $\mathfrak{m}(t) + 2\mu m$ is 'consumed' by the market order. The average cost per share of the order equals $\mathfrak{m}(t) + \mu m$, and the total cost of the trade equals $\mathfrak{m}(t)m + \mu m^2$. The second main difference to chapter 6 is that we replace the exponential resilience by instantaneous resilience: we assume that the best ask immediately reverts back to its original level: $\alpha(t+) = \mathfrak{m}(t)$ and therefore also $\mathfrak{m}(t+) = \mathfrak{m}(t)$. Note that this corresponds to the linear temporary impact model for market order trading that was introduced in Almgren and Chriss (1999) and Almgren and Chriss (2001).

The third main difference to chapter 6 is, of course, the inclusion of limit order trading. We shall now discuss how this is achieved in our model, and what the associated trading costs are. Suppose the trader decides to place a buy limit order of size $l > 0$ at time $t-$. Its execution is modelled by a compound Poisson process

$$Q(t) = \sum_{i=1}^{\pi(t)} Z_i$$

where Z_i, $i \in \mathbb{N}$, are independent and identically distributed random variables and $\pi(t)$ is a Poisson process of intensity λ, independent of all the Z_i's. A jump of π at time t corresponds to the partial execution of the limit order of size l that the trader had stored in the order book at time $t-$, and the size of the jump Z_i corresponds to the proportion executed of the total amount l that was present in the orderbook at execution time of the limit order. At most the entire limit order l can be executed, therefore we assume $Z_i \in [0, 1]$. We introduce the notation $\beta := \mathrm{E}[Z_i]$ and $\gamma := \mathrm{E}[Z_i^2]$ for the first two moments of the jump size. For a Borel set $U \subseteq [0, 1]$, we define the *Poisson random measure* of Q to be

$$N(t, U) = \sum_{0 \leq s \leq t} 1_U(\Delta Q(s)) = \sum_{n=0}^{\pi(t)} 1_U(Z_n),$$

so $N(t, U)$ equals the number of jumps of size contained in U, occuring before time t. Denote the *compensator* by $\nu(U) = \mathrm{E}[N(1, U)]$, and the *compensated Poisson random measure* by $\tilde{N}(\mathrm{d}t, \mathrm{d}z) = N(\mathrm{d}t, \mathrm{d}z) - \nu(\mathrm{d}z)\mathrm{d}t$.

Note that we made an important choice here: we model the *proportion* executed of the limit order placed in the book. Another, and arguably more realistic approach would be to model the *number* of limit orders executed. This approach, however, will greatly complexify the control problem and we will therefore not pursue it here.

Placing a buy limit order of size l implicitly affects the midquote price. Hautsch and Huang (2009) empirically analysed the price impact of a limit order depending on its position in the order book, its size and the state of the order book, using high frequency data from Euronext Amsterdam. They find that large limit orders that are placed aggressively in the book induce a price movement. This is explained by signalling effects: when a large limit order is placed on the best bid (or inside the spread) this signals to the market an increase in the demand of the stock. Hence prices are driven up. This is in line with the observation by Maslov and Mills (2001) which we already noted in chapter 3 when studying the order flow. They found that a volume-imbalance (due to an excess of limit orders on one side of the book compared to the other side) causes a predictable change of the midquote price. When we assume that the volumes on both sides are balanced without the intervention of the trader, this implies that placing a limit order has an immediate (temporary) price impact. We considerably simplify this effect by assuming a linear temporary impact on the average execution price when a limit order is placed in the order book, i.e. we assume that at execution at time t, the average price per share equals $\mathfrak{m}(t) + \kappa l$, where $\kappa \geq 0$ captures the market impact of the limit order. Let $\hat{l} \in [0, l]$ be the part of the limit order that was executed. Then the total cost of a limit order execution at time t equals $\hat{l}(\mathfrak{m}(t) + \kappa l)$. Note that the market impact of the limit order depends on the total amount l placed in the order book, and not on the amount $\hat{l}$ that is actually executed, since other market participants observe the entire limit order of size l and the price is shifted accordingly. Usually one would assume $\mu > \kappa$, since limit orders are cheaper than market orders. Note that we do not distinguish between types of limit orders, that is we implicitly assume that all limit orders are placed on the best quotes, and not deeper inside the book.

The model we introduced above extends the discrete-time model proposed in Almgren and Chriss (2001) to continuous time and adds the possibility of trading by limit orders.

Despite all simplifications, the model captures the main characteristics of an order book and execution of market/limit orders:

- Market orders are executed immediately.
- Limit orders are filled at a random time that cannot be influenced.
- Limit orders are not necessarily completely filled. The proportion that is actually executed is random.
- Limit orders are cheaper than market orders (assuming parameters $\mu > \kappa$).
- Both market and limit orders have a market impact: orders of higher size increase the execution costs.

7.2 Problem formulation

We consider the problem of optimal trading, i.e. buying or selling a fixed number of assets within a time horizon T. The buy and sell problems are symmetric, so we will focus on the buy problem. Let $x_t \geq 0$ be the number of assets the trader wants to buy in time $[t, T]$. We assume that the trading strategies with market and limit orders are absolutely continuous, so in particular we do not allow market order block trades. In fact, due to the quadratic nature of the control problem, block trades would not be optimal anyway. Let $m(s)$ denote the rate of trading with market orders at time $s \in [t, T]$ and $l(s)$ the size of the limit order currently placed in the order book. Positive (resp. negative) $m(s)$ and $l(s)$ correspond to buy (resp. sell) orders.

Let $X_{(m,l)}(s)$, $t \leq s \leq T$ denote the number of outstanding assets to be bought at time s. Then

$$\begin{aligned} \mathrm{d}X_{(m,l)}(s) &= -m(s)\mathrm{d}s - \int_0^1 z l(s) N(\mathrm{d}s, \mathrm{d}z) \\ X_{(m,l)}(t) &= x_t. \end{aligned} \tag{7.1}$$

Whenever there is no risk of confusion, we will drop the subindex (m, l) in $X_{(m,l)}$ and just write X.

The trader aims to minimize her expected trading costs and at the same time her trading risks, which is another difference to the previous chapter, where the trader was risk-neutral. For market orders, trading risks arise from movements of the midquote-price. For limit orders, there are two sources of risk: firstly, there is the risk of non-execution (or late execution) and secondly, there is the risk of an unfavorable market movement. All three types of risks are assumed to be summarized in the risk term $\alpha \int_t^T X^2(s)\mathrm{d}s$. The traders performance criterion would thus be given by

$$\mathrm{E}\left[\underbrace{\int_t^T m(s)(\mathfrak{m}(s) + \mu m(s))\mathrm{d}s}_{\text{market order trading costs}} + \underbrace{\int_t^T \int_0^1 z l(s)\left(\mathfrak{m}(s) + \kappa l(s)\right) N(\mathrm{d}s, \mathrm{d}z)}_{\text{limit order trading costs}} + \underbrace{\alpha \int_t^T X^2(s)\mathrm{d}s}_{\text{risk terms}}\right] \tag{7.2}$$

where $\alpha \geq 0$ denotes the risk aversion of the trader: the higher α the faster the trader aims to buy the remaining assets.

Before starting with the analysis of the control problem, we need to define the set of strategies which we allow in our optimisation problem: Introduce $\Theta(t, x_t)$, the set of admissible strategies $(m, l) : [t, T] \times \Omega \to \mathbb{R}^2$ satisfying

(i) m, l predictable,

(ii) $\mathrm{E}\left[\int_t^T |m(s)| + |l(s)| \mathrm{d}s\right] < \infty$,

(iii) $\mathrm{E}\left[\int_t^T m^2(s) + l^2(s) + X^2_{(m,l)}(s) \mathrm{d}s\right] < \infty$,

(iv) $\lim_{S \to T} \int_t^S m(s)\mathrm{d}s + \int_t^S \int_0^1 z l(s) N(\mathrm{d}s, \mathrm{d}z) = x_t$ a.s.

(v) $\mathrm{E}\left[\int_t^T l^4(s) \mathrm{d}s\right] < \infty$.

To avoid tedious case distinctions we assume henceforth (unless stated otherwise) that $\mu, \kappa, \alpha, \lambda$ are *strictly* positive and that $\mathrm{P}(Z_i > 0) > 0$.

Why do we impose conditions (i)-(v) on our admissible strategies?

(i) is needed, as we do not want to allow the strategy to look ahead, since otherwise we could increase l when the cheap execution of a limit order is immediate,

(ii) is necessary to ensure that $X_{(m,l)}$ exists and $\mathrm{E}[X_{(m,l)}(s)] < \infty$,

(iii) is imposed, since without loss of generality we can restrict to those strategies which give rise to a finite performance function (otherwise they cannot be optimal),

(iv) is needed to make sure that the desired portfolio is indeed bought at final time T.

(v) is necessary to simplify the performance function, so that is becomes independent of the midquote price

By integration by parts we obtain

$$\int_t^T \mathfrak{m}(s) m(s) \mathrm{d}s + \int_t^T \mathfrak{m}(s) \int_0^1 z l(s) N(\mathrm{d}s, \mathrm{d}z) = -\int_t^T \mathfrak{m}(s) \mathrm{d}X(s)$$
$$= \int_t^T X(s) \mathrm{d}\mathfrak{m}(s) - X(T)\mathfrak{m}(T) + X(t)\mathfrak{m}(t).$$

Since $s \to \int_t^s X(u) \mathrm{d}\mathfrak{m}(u)$ is a true martingale by admissibility assumption (iii), we can take expectation and obtain

$$\mathrm{E}\left[\int_t^T m(s)\mathfrak{m}(s)\mathrm{d}s + \int_t^T \int_0^1 \mathfrak{m}(s) z l(s) N(\mathrm{d}s, \mathrm{d}z)\right] = x_t \mathfrak{m}(t) = \text{constant}.$$

As the local martingale $s \mapsto \int_t^s \int_0^1 \kappa z l^2(u) \tilde{N}(\mathrm{d}s, \mathrm{d}z)$ is a true martingale by admissibility assumption (v), we have

$$\mathrm{E}\left[\int_t^T \int_0^1 \kappa z l^2(s) N(\mathrm{d}s, \mathrm{d}z)\right] = \mathrm{E}\left[\int_t^T \kappa\lambda\beta l^2(s) \mathrm{d}s\right].$$

Thus the performance criterion (7.2) is equivalent to (up to a constant)

$$\mathcal{J}(t,x,(m,l)) = \mathrm{E}^x\left[\int_t^T \mu m^2(s) + \kappa\lambda\beta l^2(s) + \alpha X^2(s)\mathrm{d}s\right] \tag{7.3}$$

where E^x is the conditional expectation under $X(t) = x$.

The control problem is to find an optimal strategy $(\hat{m},\hat{l}) \in \Theta(t,x)$ such that $\mathcal{J}(t,x,(\hat{m},\hat{l})) = \inf_{(m,l)\in\Theta(t,x)} \mathcal{J}(t,x,(m,l))$.

The above calculations greatly reduce the complexity of the control problem: due to the martingale property of the midquote price $\mathfrak{m}$, it disappears in the performance criterion (7.3). The only source of randomness that is left is given by the execution of the limit orders which is driven by the compound Poisson process Q.

Let us investigate the performance criterion (7.3) more closely:

- The first term in (7.3) denotes the additional costs stemming from the temporary price impact of market orders, given in terms of μ.
- The second term aggregates the cost incurred by the execution of limit orders placed by the trader; κ denotes the temporary price of limit orders, λ the probability that a limit order occurs in a unit time interval and β the average proportion of the limit order that is actually exectued.
- The third term in (7.3) describes the non-execution risk; the (squared) amount of outstanding orders is penalized with the risk-aversion factor α.

7.3 Optimal strategies in infinite-time horizon

We first will first consider the case $T = \infty$. Although it seems unreasonable to consider the portfolio liquidation problem on an infinite time interval, there are several good reasons to do so:

Mathematically, the control problem is now independent of time, since at any time point t, the remaining 'time-to-go' equals infinity. Thus, the number of dimensions of the problem is reduced by one, which makes the problem more tractable. By setting $T = \infty$, we also introduce new mathematical problems though, e.g. it is not clear at first sight that the candidate optimal strategy liquidates the entire portfolio (see Lemma 7.2).

Economically, it turns out that all main characteristics of the finite-time problem carry over to the infinite-time problem. In particular, the optimal strategies have a similar structure. This structure, however, becomes clearer in the infinite-time case and it is easier to analyse the relevant parameters. We will see that the optimal solution in the finite time horizon case behaves similar to the infinite-time horizon case, when the time-to-go is large, and then gradually shifts to a simple risk-free liquidation program, where the remaining shares are liquidated linearly with market orders.

We introduce the value function

$$\mathcal{V}_1(x) = \inf_{(m,l)\in\Theta_1(x)} \mathcal{J}(0,x,(m,l)) \tag{7.4}$$

where $\Theta_1(x_0)$ is the set of admissible strategies $(m, l) : [0, \infty] \times \Omega \to \mathbb{R}^2$ satisfying

(i) m, l predictable,

(ii) $\mathrm{E}\left[\int_0^\infty |m(s)| + |l(s)| \mathrm{d}s\right] < \infty$,

(iii) $\mathrm{E}\left[\int_0^\infty m^2(s) + l^2(s) + X^2_{(m,l)}(s)\mathrm{d}s\right] < \infty$,

(iv) $\lim_{t\to\infty} X_{(m,l)}(t) = 0$ a.s.,

(v) $\mathrm{E}\left[\int_0^\infty l^4(s)\mathrm{d}s\right] < \infty$,

(vi) $\lim_{t\to\infty} \mathrm{E}[X^2_{(m,l)}(t)] = 0$,

(vii) $\mathrm{E}\left[\int_0^t l^2(s) X^2_{(m,l)}(s-)\mathrm{d}s\right] < \infty$, $\forall t \geq 0$.

Heuristic arguments suggest that $\mathcal{V}_1$ should satisfy the Hamilton-Jacobi-Bellmann (HJB) equation

$$\inf_{(m,l)\in\mathbb{R}^2} \left\{ -m v_x(x) + \int_0^1 (v(x - zl) - v(x))\nu(\mathrm{d}z) + \mu m^2 + \kappa\lambda\beta l^2 + \alpha x^2 \right\} = 0 \tag{7.5}$$

Now make the quadratic ansatz $v(x) = Ax^2$ for some $A > 0$ and equation (7.5) simplifies to

$$\inf_{(m,l)\in\mathbb{R}^2} \left\{ -2Amx + A\lambda(l^2\gamma - 2lx\beta) + \mu m^2 + \kappa\lambda\beta l^2 + \alpha x^2 \right\} = 0 \tag{7.6}$$

The infimum is attained at $\hat{m} = x\frac{A}{\mu}$ and $\hat{l} = x\frac{A\beta}{A\gamma+\beta\kappa}$. Plugging $\hat{m}, \hat{l}$ back into (7.6) yields

$$x^2 H(A) = 0 \tag{7.7}$$

where

$$H(A) = -\frac{A^2}{\mu} - \frac{A^2\beta^2\lambda}{A\gamma + \beta\kappa} + \alpha. \tag{7.8}$$

It is easy to see that H admits a unique real positive root $\hat{A}$ satisfying

$$0 \leq \frac{1}{2\gamma}\left(\sqrt{(\beta\lambda\mu)^2 + 4\alpha\mu\gamma^2} - \beta\lambda\mu\right) \leq \hat{A} \leq \sqrt{\alpha\mu}. \tag{7.9}$$

Hence we obtain our candidate value function $v(x) := \hat{A}x^2$.

Let us now introduce a condition on the proportion Z_i ($i \in \mathbb{N}$) of the limit orders executed:

$$\lambda \mathrm{E}\left[\log\left|1 - Z_i \frac{\hat{A}\beta}{\hat{A}\gamma + \beta\kappa}\right|\right] < \frac{\hat{A}}{\mu} \tag{C}$$

Condition (C) is a condition on the distribution on Z_i that is needed to make sure that a.s. the entire portfolio is liquidated when using the optimal strategy. However, condition (C) can only fail under aberrent market conditions, as described in subsubsection 7.3.4, with $\Pi \in (2, \infty)$.

In fact, assuming that (C) holds, the candidate value function is the true value function of the problem, as is made precise in the below theorem.

Theorem 7.1. *Suppose that condition (C) holds. Then the value function for the infinite-time problem in (7.4) is given by $\mathcal{V}_1(x) = \hat{A}x^2$, where $\hat{A}$ is the unique positive root of (7.8). The associated trading problem admits a unique solution given in feedback form by the optimal trading strategies*

$$\hat{m}(t, X(t-)) = X(t-)\frac{\hat{A}}{\mu}$$
$$\hat{l}(t, X(t-)) = X(t-)\frac{\hat{A}\beta}{\hat{A}\gamma + \beta\kappa}$$

The optimal strategy depends on the market which is characterized by the parameter set $(\mu, \kappa, \lambda, \beta, \gamma)$, and the risk-aversion of the trader α. Note that all market parameters $(\mu, \kappa, \lambda, \beta, \gamma)$ describe the liquidity of the market:

- μ characterises the temporary price impact of a market order
- $(\kappa, \lambda, \beta, \gamma)$ characterise price impact, frequency and degree of filling of limit order

In the following we will investigate how the model parameters influence the optimal strategy in different markets. For ease of exposition, we will assume that whenever a limit order is executed, it is completely filled, so there is no partial execution and $\beta = \gamma = 1$.

It turns out that the form of the optimal strategies can be explained by a set of simple ratios involving the model parameters.

We first define the *risk/market order impact ratio*

$$\mathfrak{R}_m := \frac{\alpha}{\mu} = \frac{\text{risk-aversion}}{\text{price impact of market orders}}$$

The risk/market order impact ratio describes the trade-off between risk-aversion and market order impact and intuitively determines how fast the trader should liquidate the portfolio using market orders: a high value of $\mathfrak{R}_m$ suggests that it is favourable to use large market orders, whereas a low value of $\mathfrak{R}_m$ indicates that the trader should decrease the speed of market order trading to save trading costs.

The corresponding definition for limit orders is given by the *risk/limit order impact ratio*

$$\mathfrak{R}_l := \frac{\alpha}{\kappa} = \frac{\text{risk-aversion}}{\text{price impact of limit orders}}$$

The risk/limit order impact ratio describes the trade-off between risk-aversion and limit order impact. It influences the size of limit orders that the trader should use: if $\mathfrak{R}_l$ is large, this means that it is better to use large limit orders to minimize the expected trading costs. Conversely, a low value of $\mathfrak{R}_l$ indicates that the trader should place smaller limit orders.

Also define the *limit order liquidity ratio*

$$\mathfrak{L} := \frac{\lambda}{\kappa} = \frac{\text{intensity of limit order execution}}{\text{price impact of limit orders}}$$

Market type	parameters infl. market order trading	parameters infl. limit order trading
Pure market order	$\mathfrak{R}_m$	-
Pure limit order	-	$\mathfrak{R}_l, \mathfrak{L}$
Mixed, no limit order impact	$\mathfrak{R}_m, \lambda$	indep. of key market param.
Mixed, first order approx.	$\mathfrak{R}_m, \lambda, \mathfrak{I}$	$\mathfrak{R}_l, \mathfrak{R}_m, \lambda, \mathfrak{I}$

Table 7.1: Dependence of trading rates on key market parameters $(\mathfrak{R}_m, \mathfrak{R}_l, \mathfrak{L}, \mathfrak{I}, \lambda)$.

The limit order liquidity ratio is a measure for the limit order liquidity of the market, i.e. the price impact and non-execution risk associated to limit orders. The higher $\mathfrak{L}$, the more liquid is the market for limit orders, since the frequency of limit orders executions is large compared to the price impact of limit orders. Note that a frequent limit order execution is equivalent to a low non-execution risk. Conversely, low values of $\mathfrak{L}$ indicate that the limit order price impact is large compared to the frequency of limit order execution.

Finally we define the *limit/market order impact ratio*

$$\mathfrak{I} := \frac{\kappa}{\mu} = \frac{\text{price impact of limit orders}}{\text{price impact of market orders}}$$

The limit/market order impact ratio denotes the trade-off between price impact of limit orders and price impact of market orders. A high value of $\mathfrak{I}$ means that limit orders are relatively expensive (compared to market orders), whereas a low value of $\mathfrak{I}$ indicates that limit orders are cheap, compared to market orders.

In the next sections, we will analyse different types of markets and investigate how the optimal trading rates are determined by the ratios above. We start by looking at pure market order market, then a pure limit order market, followed by the analysis of a mixed market/limit order market *without* temporary impact of limit orders, and finally the general market/limit order market. As the optimal strategies can all be expressed in term of $(\mathfrak{R}_m, \mathfrak{R}_l, \mathfrak{L}, \mathfrak{I}, \lambda)$, we call them the *key market parameters*. Table 7.1 shows which parameters influence the trading rates in different markets.

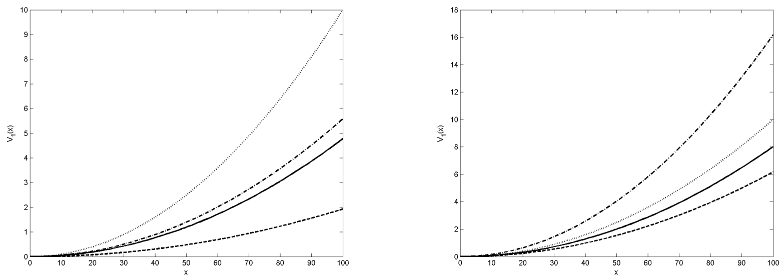
7.3.1 Trading in a pure market order market

We first consider the simplest case, where we only trade with market orders. This is either the case when limit orders are never executed, i.e. $\lambda = \kappa = 0$, or when the market impact costs of limit orders are infinite ($\kappa = \infty$), and it is therefore never optimal to place a limit order. In both cases, we find ourselves in the classical pure market order framework described in Almgren and Chriss (2001). Contrary to Almgren and Chriss (2001) we consider continuous time trading and an infinite time horizon. Then cleary for all $t \geq 0$, $\hat{l}(t, X(t-)) = 0$, so no limit orders are placed at any time. We readily compute that $\hat{A} = \sqrt{\alpha\mu}$ which gives us the optimal trading intensity with market orders

$$\hat{m}(t, X(t-)) = X(t-)\sqrt{\mathfrak{R}_m} \tag{7.10}$$

and thus the optimal trading trajectory

$$\hat{X}(t) = x_0 e^{-t\sqrt{\mathfrak{R}_m}}$$



(a) Value function $\mathcal{V}_1(x)$ for different markets with limit order frequency $\lambda = 0.5$

(b) Value function $\mathcal{V}_1(x)$ for different markets with limit order frequency $\lambda = 0.1$

Figure 7.1: The solid line represents $\mathcal{V}_1(x)$ for general limit order market with $\mu = 0.01, \kappa = 0.001, \alpha = 0.0001$ and complete limit order execution ($\beta = \gamma = 1$). The dotted line represents $\mathcal{V}_1(x)$ for pure market order market with $\mu = 0.01, \kappa = \infty, \alpha = 0.0001$. The dashed line shows $\mathcal{V}_1(x)$ for a dark pool market (i.e. a limit order market without price impact of limit orders) with $\mu = 0.01, \kappa = 0, \alpha = 0.0001$ and complete limit order execution ($\beta = \gamma = 1$). The dashed-dotted line represents $\mathcal{V}_1(x)$ for a pure limit order market with $\mu = \infty, \kappa = 0.001, \alpha = 0.0001$ and complete limit order execution ($\beta = \gamma = 1$).

When comparing our solution with the optimal strategy found in Almgren and Chriss (2001), we see that the basic properties of the optimal strategy are preserved:

- For a sell programm, the share holdings decrease monotonically, when using the optimal liquidation strategy.
- The number of share holdings decreases from its initial position to 0 at a constant rate.
- The number of share holdings is a continuous function of time.
- The optimal strategy is deterministic.

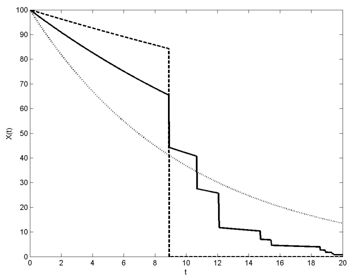
The dotted line in figures 7.1 and 7.2 show the value function $\mathcal{V}_1(x)$ and optimal liquidation trajectories $X(t)$ for the pure market order market.

Dependence of the optimal strategy on key market parameters

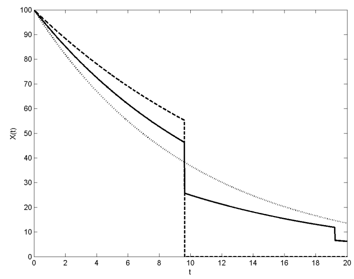
By inspection of the optimal strategy given by (7.10), it turns out that the key constant that determines the strategy is the risk/market order impact ratio $\mathfrak{R}_m$. It determines the constant rate at which market orders are sent to the order book. A higher risk-aversion will increase the speed liquidation speed, whereas a higher temporary market impact of market orders decreases the liquidation speed. For any $\rho > 0$, a scaling of the model parameters $(\mu, \alpha) \mapsto (\rho\mu, \rho\alpha)$ does not change the optimal strategy and the associated trading costs. This homotheticity property can also be directly derived from the performance criterion (7.3), when setting $\lambda = 0$.

7.3.2 Trading in a pure limit order market

Now consider a market, where the cost of executing a market order is infinite, i.e. $\mu = \infty$. This situation might occur in a pure 'dark pool-like' market, where it is technically impossible to



(a) Optimal trajectory $X(t)$ for different markets with limit order frequency $\lambda = 0.5$



(b) Optimal trajectory $X(t)$ for different markets with limit order frequency $\lambda = 0.1$

Figure 7.2: The solid line represents $X(t)$ for general limit order market with $\mu = 0.01, \kappa = 0.001, \alpha = 0.0001$ and complete limit order execution ($\beta = \gamma = 1$). The dotted line represents $X(t)$ for pure market order market with $\mu = 0.01, \kappa = \infty, \alpha = 0.0001$. The dashed line shows $X(t)$ for a dark pool market (i.e. a limit order market without price impact of limit orders) with $\mu = 0.01, \kappa = 0, \alpha = 0.0001$ and complete limit order execution ($\beta = \gamma = 1$).

submit market orders and only limit orders can be executed. We obviously do not trade with market orders, so $\hat{m}(t, X(t-)) = 0$. Remember that for clarity of exposition, we supposed that a limit order is always fully executed, i.e. $\beta = \gamma = 1$. We compute

$$\hat{A} = \frac{1}{2\lambda}\left(\sqrt{\alpha^2 + 4\alpha\kappa\lambda} - \alpha\right)$$

and thus

$$\begin{aligned}\hat{l}(t, X(t-)) &= X(t-)\frac{2\alpha}{\alpha + \sqrt{\alpha^2 + 4\kappa\lambda}} \\ &= X(t-)\frac{2\mathfrak{R}_l}{\mathfrak{R}_l + \sqrt{\mathfrak{R}_l^2 + 4\mathfrak{L}}} \qquad (7.11)\end{aligned}$$

Moreover, we can give the optimal trajectory in explicit form (see proposition 7.1 for the general formula):

$$\begin{aligned}\hat{X}(t) &= x_0\left(1 - \frac{2\alpha}{\alpha + \sqrt{\alpha^2 + 4\kappa\lambda}}\right)^{\pi(t)} \\ &= x_0\left(\frac{\sqrt{\alpha^2 + 4\kappa\lambda} - \alpha}{\sqrt{\alpha^2 + 4\kappa\lambda} + \alpha}\right)^{\pi(t)} \\ &= x_0\left(\frac{\sqrt{\mathfrak{R}_l^2 + 4\mathfrak{L}} - \mathfrak{R}_l}{\sqrt{\mathfrak{R}_l^2 + 4\mathfrak{L}} + \mathfrak{R}_l}\right)^{\pi(t)}\end{aligned}$$

We see that the basic properties of the optimal strategy differs somewhat from the pure market order market:

- For a sell programm, the share holdings decrease monotonically, when using the optimal liquidation strategy.
- The number of share holdings decreases from its initial position to 0 in a sequence of jumps. The jump size is a constant fraction of the current number of share holdings.
- The number of share holdings is a pure jump process.
- The optimal strategy is stochastic. The optimal trajectory at time t depends only on the total number of limit order executions up to time t, $\pi(t)$.

The dashed-dotted line in figures 7.1 and 7.3 show the value function $\mathcal{V}_1(x)$ and optimal liquidation trajectories $X(t)$ for the pure limit order market. In figure 7.1, we see that, depending on the model parameters (in this case limit order frequency), trading in a pure limit order market can be cheaper or more costly than trading in a pure market order market.

A straightforward computation gives

$$\mathrm{E}\left[\hat{X}(t)\right] = x_0 \exp\left\{-t\frac{2\lambda\alpha}{\sqrt{\alpha^2+4\kappa\lambda}+\alpha}\right\}$$

Thus, as the intensity of limit order executions increases to infinity, the entire share holdings will be liquidated right after the start of the programm:

$$\lim_{\lambda\to\infty} \mathrm{E}\left[\hat{X}(t)\right] = 0 \quad \text{, for all } t > 0.$$

Let us also consider what happens in the limiting cases $\alpha \to 0$ and $\alpha \to \infty$:

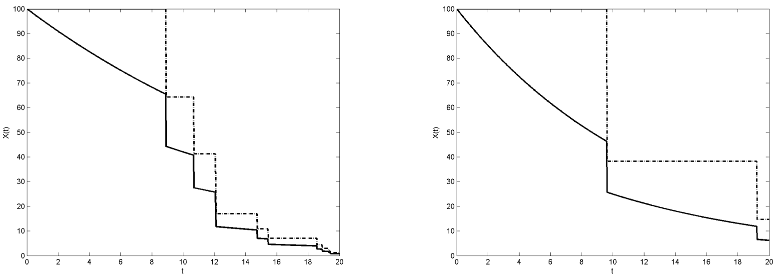
- As $\alpha \to 0$, we have $\frac{2\lambda\alpha}{\sqrt{\alpha^2+4\kappa\lambda}+\alpha} \to 0$: the trader has no incentive to place any limit order (unless $\kappa = 0$). Thus the asymptotic optimal strategy in the limiting case is $\hat{l}(t, X(t-)) = 0$, and therefore $\mathrm{E}\left[\hat{X}(t)\right] = x_0$.
- As $\alpha \to \infty$, we have $\frac{2\lambda\alpha}{\sqrt{\alpha^2+4\kappa\lambda}+\alpha} \to \lambda$: the trader wants to liquidate her position as fast as possible. The optimal strategy consists in placing all remaining shares as a limit order: $\hat{l}(t, X(t-)) = X(t-)$. However, the trader has to wait for the uncertain execution. The average execution rate equals λ, hence we have $\mathrm{E}\left[\hat{X}(t)\right] = x_0 e^{-t\lambda}$.

Dependence of the optimal strategy on key market parameters

Let $l := \frac{2\mathfrak{R}_l}{\mathfrak{R}_l+\sqrt{\mathfrak{R}_l^2+4\mathfrak{L}}}$ be the constant proportion of the total share holdings that is placed as a limit order, when following the optimal strategy.

Then it is easily verified that

- l is increasing in $\mathfrak{R}_l$: when the risk/limit order impact ratio increases, the proportion placed as a limit order increases, as either the risk of large share holdings increase, or the price impact of limit orders decrease, or both,



(a) Optimal trajectory $X(t)$ for different markets with limit order frequency $\lambda = 0.5$

(b) Optimal trajectory $X(t)$ for different markets with limit order frequency $\lambda = 0.1$

Figure 7.3: The solid line represents $X(t)$ for general limit order market with $\mu = 0.01, \kappa = 0.001, \alpha = 0.0001$ and complete limit order execution ($\beta = \gamma = 1$). The dashed-dotted line represents $X(t)$ for a pure limit order market with $\mu = \infty, \kappa = 0.001, \alpha = 0.0001$ and complete limit order execution ($\beta = \gamma = 1$).

- l is decreasing in $\mathfrak{L}$: a rise of limit order liquidity leads to fewer limit orders placed. This might seem unintuitive at first sight, however note that a rise of the limit order liquidity $\mathfrak{L}$, while the risk/limit order impact trade-off $\mathfrak{R}_l$ stays constant, means that the frequency of limit orders increases. Thus even though the trader places smaller limit orders (and thereby decreases her market impact), the higher frequency of execution balances this effect.

7.3.3 Trading in a market without limit order impact

For clarity of exposition, we assume again that a limit order is always completely filled, when a limit order execution occurs, that is $\beta = \gamma = 1$.

The absence of market impact of limit orders needs of course justification: For example, if the asset position that needs to be liquidated is small, we can assume that placing a limit order of this size is not observed by the market and we can set $\kappa = 0$. Another situation where the zero-impact model is justified is dark-pool trading. In this case, the $l(t)$ denotes the order placed in dark pool, whereas $m(t)$ denotes the trading intensity in the primary venue by market orders. As the dark pool order is not observed by other market participants, it is natural to assume that it has no market impact and we can set $\kappa = 0$. Kratz and Schöneborn (2010) introduced this model to analyse optimal trading strategies in dark pools. They analyse in great detail discrete-time version of the model, considering the finite and infinite horizon, as well as the single and multi-asset problem. In Kratz (2011) their model is extended to a continuous time setting.

Assuming that the absence of limit order impact can be justified, we compute

$$\hat{A} = \frac{1}{2}\left(\sqrt{(\lambda\mu)^2 + 4\alpha\mu} - \lambda\mu\right)$$

and obtain the optimal strategy, which is given by

$$\hat{m}(t, X(t-)) = X(t-)\sqrt{\mathfrak{R}_m + \left(\frac{\lambda}{2}\right)^2} - \frac{\lambda}{2},$$
$$\hat{l}(t, X(t-)) = X(t-). \tag{7.12}$$

Moreover we easily compute the optimal trading trajectory

$$\hat{X}(t) = x_0 e^{-t\left(\sqrt{\mathfrak{R}_m + \left(\frac{\lambda}{2}\right)^2} - \frac{\lambda}{2}\right)} 1_{\{\pi(t)=0\}}.$$

The following properties of the optimal strategy are readily observed:

- For a sell programm, the share holdings decrease monotonically, when using the optimal liquidation strategy.
- Until the first execution of a limit order occurs, the optimal trajectory decreases from its initial point at a constant rate.
- As soon as the first limit order occurs, the liquidation programm is terminated, as all remaining shares have been traded via the limit order.
- The optimal strategy is stochastic. The only stochastic component is the first execution time of a limit order.

Let us quantify the comparative advantage of using limit orders over a trading strategy that uses only market orders: Suppose the trader wants to liquidate a position of x assets. Let $\Delta C(x)$ denote the difference in trading costs when trading only by market orders and when trading by both market and limit orders (with zero limit order impact). Then

$$\Delta C(x) = x^2\mu\left(\sqrt{\mathfrak{R}_m} + \sqrt{\left(\frac{\lambda}{2}\right)^2} - \sqrt{\mathfrak{R}_m + \left(\frac{\lambda}{2}\right)^2}\right) \geq 0.$$

Note that $\Delta C(x)$ is increasing in both $\mathfrak{R}_m$ and λ. An increase in the risk/market order impact ratio $\mathfrak{R}_m$ leads to higher trading costs when using market orders, which increases the difference of trading costs $\Delta C(x)$. Likewise an increase of λ implies that limit orders are executed more frequently and thus leads to a decrease of trading costs when the use of limit orders is permitted, and thus also increases $\Delta C(x)$.

Dependence of the optimal strategy on key market parameters

When the trader uses the optimal strategy (7.12), she places all remaining shares as a limit order in the order book, during the entire trading program. We will therefore focus on the analysis of how the market order trading rate depends on the model parameters. The market order trading rate

$$m = \sqrt{\mathfrak{R}_m + \left(\frac{\lambda}{2}\right)^2} - \frac{\lambda}{2}$$

is

- increasing in $\mathfrak{R}_m$: when the risk/market order impact ratio increases, the trading rate via market orders increases,
- decreasing in λ: when the frequency of limit orders increases, the trading rates via market orders decreases, as the chances of a cheaper limit order execution becomes bigger.

7.3.4 Trading in a market with limit order impact

Finally consider the general case, where the use of market orders is permitted, limit order do occur ($\lambda > 0$) and have strictly positive market impact ($\kappa > 0$). The additional possibility of using limit orders must clearly decrease the value function, as compared to the pure market order case in subsection 7.3.2, and therefore we recover one side of the bound (7.9) above. Similarly, we must incur higher costs than in the market without price impact of limit order from subsection 7.3.3, and so we recover the other side of the bound (7.9).

Figures 7.1 and 7.2 show the value function $\mathcal{V}_1(x)$ and optimal liquidation trajectories $X(t)$ of the limit order market with limit order impact (solid line), as well as the dark pool market (dashed line) and the pure market order market (dotted line). The optimal strategy we calculated for the market with limit order impact is clearly best at capturing the typical trading strategies in limit order markets used in practice: A certain fraction of the outstanding shares is placed as a limit order and wait for execution. As soon as the limit order is executed a new limit order is placed. At the same time market orders are executed (at a less favourable price) to reduce the risk of non-execution.

Dependence of trading costs on model parameters

The trading costs associated to buying x shares using the optimal strategy from Theorem 7.1 are $x^2\hat{A}$. We start by analysing how the trading costs depend on the model parameters $(\alpha, \mu, \lambda, \kappa)$. The trading costs $x^2\hat{A}$ are

(i) increasing in α: a higher risk-aversion increases the average trading costs

(ii) increasing in μ: a higher price impact of market orders increases the average trading costs

(iii) decreasing in λ: a higher frequency of limit orders decreases the average trading costs

(iv) increasing in κ: a higher price impact of limit orders increases the average trading costs

The above results confirm our intuition about trading costs.

Monotonicity of the optimal trajectories

Let us analyse how the optimal trajectories depend on the size of the limit order. Let $\Pi = \frac{\hat{A}\beta}{\hat{A}\gamma+\beta\kappa}$ be the proportion of the remaining share holdings that is placed as a limit order. Depending on Π the trajectories $\hat{X}(t)$ associated with the optimal strategy can take different structures:

- $\Pi \in [0,1]$: In that case the optimal trajectory $\hat{X}(t)$ is a.s. decreasing in t and it satisfies a.s. $0 \le \hat{X}(t) \le x_0$ for all $t \ge 0$.
- $\Pi \in (1,2]$: The absolute value of the optimal trajectory $|\hat{X}(t)|$ is a.s. decreasing in t and it satisfies a.s. $-x_0 \le \hat{X}(t) \le x_0$ for all $t \ge 0$. Note that an optimal limit order can change a positive share holding to a negative one, but the share holdings decrease in absolute value.
- $\Pi \in (2,\infty)$: In this case, an optimal limit order can change a positive share holding to a negative one, and the share holdings might even increase in absolute value.

Figure 7.4 shows optimal trajectories for the three cases, with common model parameters $\mu = 0.01, \kappa = 0.001, \alpha = 0.001, \lambda = 0.2$ and different distributions of the fill-rate Z_i (and thus different values of β and γ). For real markets, one would expect that the optimal liquidation strategy consists of trading in one single direction, i.e. that we are in the case $\Pi \in [0,1]$. Note that in the case of complete filling of limit orders ($\beta = \gamma = 1$), this is always satisfied.

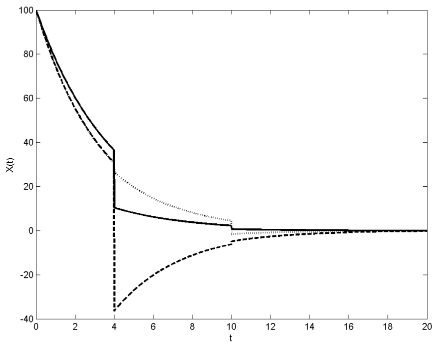


Figure 7.4: Optimal trajectories for different values of Π. The solid line corresponds to $\Pi = 0.716 \in [0,1]$, the dotted line to $\Pi = 1.356 \in (1,2]$, the dashed line to $\Pi = 2.184 \in (2,\infty)$.

First order approximation of limit order impact

Typically the price impact of a limit order κ is much smaller than the impact of a market order μ, as otherwise the trader would have no incentive to place the limit order and run the additional risk of non-execution. However, as noted in subsection 7.3.3, setting the price impact of limit orders to zero leads to unrealistic strategies if the trader wants to buy a large amount of assets. As the impact of κ is very small, compared to other model parameters, yet it cannot be neglected completely, we shall analyse a first order approximation of the price impact of limit orders.

To carry out the first order approximation, we assume that κ is so small, that we can neglect terms of order κ^2 or higher. Note that equation (7.8) reads $H(A) = H_0(A) + \mathcal{O}(\kappa^2)$ where

$$H_0(A) = -\frac{A^2}{\mu} - \frac{A\beta^2\lambda}{\gamma} + \alpha + \lambda\kappa\frac{\beta^3}{\gamma^2}. \tag{7.13}$$

By the first order approximation of the control problem, we mean that we neglect the $\mathcal{O}(\kappa^2)$ terms in (7.8), and instead use (7.13). For ease of exposition, we assume once more complete filling of limit orders, i.e. $\beta = \gamma = 1$. We define A_0 to be the unique positive real root of (7.13), and set m_0, l_0 be the corresponding optimal market order trading rate and optimal limit order size. We then compute $A_0 = \mu\left(\sqrt{\frac{\alpha+\kappa\lambda}{\mu} + \left(\frac{\lambda}{2}^2\right)} - \frac{\lambda}{2}\right)$ and we obtain

$$\begin{aligned} m_0(t, X(t-)) &= X(t-)\left(\sqrt{\mathfrak{R}_m + \lambda\mathfrak{I} + \left(\frac{\lambda}{2}\right)^2} - \frac{\lambda}{2}\right) \\ l_0(t, X(t-)) &= X(t-)\frac{\mathfrak{R}_l + \lambda}{\mathfrak{R}_l + \frac{3}{2}\lambda + \sqrt{\mathfrak{R}_m + \lambda\mathfrak{I} + \left(\frac{\lambda}{2}\right)^2}}. \end{aligned} \tag{7.14}$$

We have the following properties of the optimal trading strategies in the first order approximation:
Let $m_0 = \left(\sqrt{\mathfrak{R}_m + \lambda\mathfrak{I} + \left(\frac{\lambda}{2}\right)^2} - \frac{\lambda}{2}\right)$ be the optimal market order trading rate and $l_0 = \frac{\mathfrak{R}_l+\lambda}{\mathfrak{R}_l+\frac{3}{2}\lambda+\sqrt{\mathfrak{R}_m+\lambda\mathfrak{I}+\left(\frac{\lambda}{2}\right)^2}}$ the optimal fraction of the total share holdings placed as a limit order in the first order approximation of the control problem with complete limit order filling as described above. Then

(i) m_0 is increasing in $\mathfrak{R}_m$

(ii) m_0 is increasing in $\mathfrak{I}$

(iii) l_0 is increasing in $\mathfrak{R}_l$

(iv) l_0 is decreasing in $\mathfrak{I}$

(v) l_0 is decreasing in $\mathfrak{R}_m$

The above dependencies are intuitively clear and can easily be justified economically: When the risk vs. market/limit order impact ratio rises, the trader following the optimal strategy increases the number of market/limit orders. Moreover, it is clear that an increase in the limit/market order impact ratio $\mathfrak{I}$ decreases the trading with limit orders and increases the trading intensity with market orders.

It is interesting to see that the limit order size l_0 is decreasing in $\mathfrak{R}_m$, whereas the market order trading rate m_0 does not depend on $\mathfrak{R}_l$.

Let us now investigate how the limit order intensity λ influences the optimal trading rates. For better comparision, we will not look at the size of the limit order, but at the average limit order execution rate. Intuitively, one would guess that a higher frequency of limit order executions increases the average limit order trading rate and decreases the market order trading rate. The below result shows that, interestingly, this is not always the case:

Let

$$m_0 = \sqrt{\mathfrak{R}_m + \lambda\mathfrak{I} + \left(\frac{\lambda}{2}\right)^2} - \frac{\lambda}{2}$$

be the optimal market order trading rate and

$$r_0 = \frac{\lambda \mathfrak{R}_l + \lambda^2}{\mathfrak{R}_l + \frac{3}{2}\lambda + \sqrt{\mathfrak{R}_m + \lambda \mathfrak{I} + \left(\frac{\lambda}{2}\right)^2}} = \lambda l_0$$

the average limit order trading rate per unit interval in the first order approximation of the control problem with complete limit order filling. Then

(i) m_0 is decreasing in λ if $\mathfrak{R}_m > \mathfrak{I}^2$

(ii) m_0 is a constant function of λ if $\mathfrak{R}_m = \mathfrak{I}^2$

(iii) m_0 is increasing in λ if $\mathfrak{R}_m < \mathfrak{I}^2$

(iv) r_0 is increasing in λ

When the risk/market order impact ratio outweighs the comparative price impact advantage of market order vs. limit order trading (i.e. we have $\mathfrak{R}_m > \mathfrak{I}^2$), then we indeed have that the market order trading rate is a decreasing function of the limit order intensity λ. However, in the case $\mathfrak{R}_m < \mathfrak{I}^2$, the price impact of limit orders is so high that an increase of the limit order intensity leads to an increase of market order trading rate. In real markets, we would not expect this parameter constellation, as the limit order price impact is usually very small.

Optimal trajectory

It is often useful to have information about the optimal trajectory $\hat{X}$. In fact, we can derive an explicit formula for the optimal trajectory:

Proposition 7.1 (Optimal trajectory)**.** *Let $\hat{X} = X_{(\hat{m},\hat{l})}$ be the optimal trajectory corresponding to the optimal strategy $(\hat{m},\hat{l})$ in theorem 7.1. Then*

$$\hat{X}(t) = xe^{-\frac{\hat{A}}{\mu}t} \prod_{i=1}^{\pi(t)} \left(1 - Z_i \frac{\hat{A}\beta}{\hat{A}\gamma + \kappa\beta}\right) \tag{7.15}$$

For the case of a simple Poisson process (that is $Z_i \equiv 1$), the formula simplifies to

$$\hat{X}(t) = xe^{-\frac{\hat{A}}{\mu}t} \left(1 - \frac{\hat{A}}{\hat{A} + \kappa}\right)^{\pi(t)} \tag{7.16}$$

We see that the optimal trajectory is made up of two parts

- A continuous exponential decay, proportional to the current number of share holdings. This continuous decay is to market order trading.
- A pure jump part, with jump size proportional to the current number of share holdings. This jump part is due to the limit order trading.

Time of partial liquidation

In practice, the time until the portfolio has to be liquidated is usually finite. One way to deal with this problem is described in section 7.5 below. Another way is to use the optimal infinite time strategy until the first time τ when all but a certain fraction of the initial share holdings is left for liquidation. This remaining part is then liquidated in a single block trade. If we want to implement such a strategy, we need to analyse τ, in particular it is useful to know the mean and variance of τ to construct a confidence interval for the end of the liquidation program. The following corollary characterises τ and gives an explicit formula for the Laplace transform of its moments, when there is no partial filling of limit orders. Note that Veillette and Taqqu (2010) developed two different approaches to numerically compute the inverse of the Laplace transform which can be used in applications of the below result:

Corollary 7.1 (Time of partial liquidation)**.** *Assume $Q(t) = \pi(t)$ is a simple Poisson process, that is $Z_i \equiv 1$ a.s. Let $a \in (0, \infty)$ and consider*

$$\tau_a = \inf \left\{ t \geq 0 : \hat{X}(t) < e^{-a} x \right\}$$

the first time that at most the e^{-a}-fractional part of x remains to be traded, using the optimal strategy from theorem 7.1. Define the n-th moment of the first passage time

$$e_n(a) = \mathrm{E}[\tau_a^n]$$

Then the Laplace transform of e_n with respect to a is given explicitly by

$$\widehat{e_n}(u) := \int_0^\infty e_n(a) e^{-ua} \mathrm{d}a = \frac{n n!}{u \psi(u)^n}$$

where

$$\psi(u) = u \frac{\hat{A}}{\mu} + \lambda \left(\left(\frac{\kappa}{\hat{A} + \kappa} \right)^u - 1 \right). \tag{7.17}$$

Amount of limit order trading

For higher order layers in the algorithmic trading, one important performance indicator for the microtrader is the amount of shares traded via limit orders versus the amount traded by market orders. The following corollary computes the average amount traded via limit orders when using the optimal strategy, which can be used as a benchmark.

Corollary 7.2 (Average amount traded by limit orders)**.** *Let L denote the number of assets traded via limit orders for initial asset holdings $x = 1$. Let $c_1 = \hat{A}/\mu$ and $c_2 = \frac{\hat{A}\beta}{\hat{A}\gamma + \beta\kappa}$. Assume that*

$$\left| \frac{\lambda}{\lambda + c_1} (1 - c_2 \beta) \right| < 1$$

Then under the optimal strategy from theorem 7.1

$$\mathrm{E}[L] = \frac{\lambda c_2 \beta}{\lambda c_2 \beta + c_1}$$

From the proof of Corollary 7.2, it becomes clear that when $\left| \frac{\lambda}{\lambda + c_1} (1 - c_2 \beta) \right| \geq 1$, the average number of shares traded by limit orders is infinite, so one would not expect such a parameter constellation in real markets.

7.4 Optimal pure buy strategies in infinite-time horizon

In this section, we consider again the case $T = \infty$. This time, however, the trader is assumed to be restricted to *pure buy strategies* (m, l), that is, she is only allowed to trade in one direction. In real markets, such constraints might occur, when there are regulatory restrictions that prevent traders from selling shares when they want to carry out a buy programm. Such restrictions could be imposed by compliance in order to impede market manipulation strategies.

From a mathematical perspective, this is equivalent to saying that the trajectory of asset holdings $X_{(m,l)}$ corresponding to (m, l) is non-increasing. Introduce the value function

$$\mathcal{V}_2(x) = \inf_{(m,l)\in\Theta_2(x)} \mathcal{J}(0, x, (m, l)) \tag{7.18}$$

where $\Theta_2(x_0)$ is the set of admissible strategies $(m, l) : [0, \infty] \times \Omega \to \mathbb{R}^2$ satisfying

(i) m, l predictable,

(ii) $m(t) \geq 0$, $l(t) \in [0, X_{(m,l)}(t-)]$, $t \geq 0$

(iii) $\mathrm{E}\left[\int_0^\infty |m(s)| + |l(s)| \mathrm{d}s\right] < \infty$,

(iv) $\mathrm{E}\left[\int_0^\infty m^2(s) + l^2(s) + X^2_{(m,l)}(s) \mathrm{d}s\right] < \infty$,

(v) $\lim_{t\to\infty} X_{(m,l)}(t) = 0$ a.s.

Now let A_1 be the unique strictly positive real root of

$$H_1(A) = -\frac{A^2}{\mu} - \frac{A^2\beta^2\lambda}{A\gamma + \beta\kappa} + \alpha$$

and A_2 the unique strictly positive real root of

$$H_2(A) = -\frac{A^2}{\mu} - A\lambda(\gamma - 2\beta) + (\alpha + \lambda\beta\kappa).$$

We set

$$\hat{A} = \begin{cases} A_1 & \text{if } A_1 \leq \frac{\beta\kappa}{\beta-\gamma} \\ A_2 & \text{otherwise} \end{cases}$$

and

$$\hat{k} = \begin{cases} \frac{A_1\beta}{A_1\gamma+\beta\kappa} & \text{if } A_1 \leq \frac{\beta\kappa}{\beta-\gamma} \\ 1 & \text{otherwise} \end{cases}$$

We can now formulate

Theorem 7.2. *The value function for the infinite-time problem in* (7.18) *is given by* $\mathcal{V}_2(x) = \hat{A}x^2$. *The associated trading problem admits a unique solution given by the optimal trading strategies*

$$\hat{m}(t, X(t-)) = X(t-)\frac{\hat{A}}{\mu}$$
$$\hat{l}(t, X(t-)) = X(t-)\hat{k}$$

All the analyses we carried out in the unconstrained case, such as dependence on key parameters, optimal trajectories, time of partial liquidation and amount of limit order trading carry over to the constrained case with minor modifications.

7.5 Optimal strategies with finite horizon

In real markets, the trader usually has a time constraint for the trading programme. In this section we will therefore analyse the case $T < \infty$.

7.5.1 Explicit solution in markets without limit order impact

We first assume the special case $\kappa = 0$ (see also Kratz, 2011; Naujokat and Westray, 2010). The value function is defined as

$$\mathcal{V}_3(t,x) = \inf_{(m,l)\in\Theta_3(t,x)} \mathcal{J}(t,x,(m,l)) \tag{7.19}$$

where $\Theta_3(t,x_t)$ is the set of admissible strategies $(m,l) : [t,T] \times \Omega \to \mathbb{R}^2$ satisfying

(i) m, l predictable,

(ii) $\mathrm{E}\left[\int_t^T |m(s)| + |l(s)|\mathrm{d}s\right] < \infty$,

(iii) $\mathrm{E}\left[\int_t^T m^2(s) + l^2(s) + X^2_{(m,l)}(s)\mathrm{d}s\right] < \infty$,

(iv) $\lim_{S\to T} \int_t^S m(s)\mathrm{d}s + \int_t^S \int_0^1 zl(s)N(\mathrm{d}s,\mathrm{d}z) = x_t$ a.s.,

(v) $\mathrm{E}\left[\int_t^T l^4(s)\mathrm{d}s\right] < \infty$,

(vi) $\lim_{t\to T} \mathrm{E}[X^2_{(m,l)}(t)] = 0$,

(vii) $\mathrm{E}\left[\int_t^T l^2(u)X^2_{(m,l)}(u-)\mathrm{d}u\right] < \infty.$,

Then we have

Theorem 7.3. *The value function for the finite-time problem in* (7.19) *is given by* $\mathcal{V}_3(t,x) = \hat{B}(T-t)x^2$ *with*

$$\hat{B}(\tau) = \mu \frac{e^{2\tau\sqrt{\mathfrak{R}_m + \left(\frac{\lambda\beta^2}{2\gamma}\right)^2}}\left(\sqrt{\mathfrak{R}_m + \left(\frac{\lambda\beta^2}{2\gamma}\right)^2} - \frac{\lambda\beta^2}{2\gamma}\right) + \left(\sqrt{\mathfrak{R}_m + \left(\frac{\lambda\beta^2}{2\gamma}\right)^2} + \frac{\lambda\beta^2}{2\gamma}\right)}{e^{2\tau\sqrt{\mathfrak{R}_m + \left(\frac{\lambda\beta^2}{2\gamma}\right)^2}} - 1}$$

The associated trading problem admits a unique solution given by the optimal trading strategies

$$\hat{m}(t,X(t-)) = X(t-)\frac{\hat{B}(T-t)}{\mu}$$
$$\hat{l}(t,X(t-)) = X(t-)\frac{\beta}{\gamma}$$

The proof of the theorem uses that fact that the value function can be computed explicitly, as it reduces to solving a Riccati equation. Hence we obtain a candidate value function which can be shown to be the true value function using a verification argument.

As in subsection 7.3.3, we note the dependence on the key parameters $\mathfrak{R}_m$ and λ. To better understand the structure of the optimal strategy, we will consider different time horizons.

Long time horizon behaviour

When the time-to-go $\tau := T-t$ is large, the optimal strategy is close to the strategy we obtained in the infinite-time case. More precisely,

$$\lim_{\tau\to\infty} \hat{B}(\tau) = \mu \left(\sqrt{\mathfrak{R}_m + \left(\frac{\lambda\beta^2}{2\gamma}\right)^2} - \frac{\lambda\beta^2}{2\gamma} \right)$$

and thus we recover the optimal market order trading strategy from subsection 7.3.3. Note that the optimal limit order trading strategy is independent of time, and thus the same for finite and infinite time horizon.

Short time horizon behaviour

As the time t increases, fewer limit orders are placed (since the total share holdings X decreases), and the market order trading rate increases. When t approaches the final time T, or, equivalently, the time-to-go $\tau := T - t$ tends to zero, the optimal market order trading rate tends to infinity, as $\lim_{\tau\to 0} \hat{B}(\tau) = \infty$. In fact we have

$$\hat{B}(\tau) = \frac{\mu}{\tau} + \mathcal{O}(1)$$

and thus for t close to T approximately

$$\hat{m}(t, X(t-)) \approx \frac{X(t-)}{T-t}$$

and it follows that the remaining shares are liquidated linearly

$$\hat{X}(t) \approx \frac{T-t_0}{T-t} \hat{X}(T-t_0)$$

for $t_0 \le t \le T$ and t_0 close to T.

Thus the short time horizon behaviour of the market order rate does no longer depend on the risk/market order impact trade-off $\mathfrak{R}_m$ and limit order intensity λ: the time-to-go is too short for risk-considerations to have an impact on the strategy and because it is not reasonable to wait for a limit order to be filled in the remaining time.

7.5.2 The general case

We now consider the control problem of section 7.5.1 with market impact of limit orders, i.e. we assume $\kappa > 0$. Take the definition of the value function $\mathcal{V}_3(t, x_t)$ and the set of admissible strategies $\Theta_3(t, x_t)$. Before stating a theorem on optimal strategies, we need an auxiliary result establishing properties of certain differential equations.

Define the differential operators $\mathcal{H}_i$, $i = 1, 2, 3$ by

$$(\mathcal{H}_i B)(s) := B'(s) - h_i(B(s))$$

with

$$\begin{aligned} h_1(b) &= \alpha - \frac{b^2}{\mu} - b\frac{\beta^2\lambda}{\gamma} \\ h_2(b) &= \alpha - \frac{b^2}{\mu} - \frac{b^2\beta^2\lambda}{b\gamma + \beta\kappa} \\ h_3(b) &= \left(\alpha + \frac{\beta^3\lambda\kappa}{\gamma^2}\right) - \frac{b^2}{\mu} - b\frac{\beta^2\lambda}{\gamma}. \end{aligned}$$

Note that the equations $\mathcal{H}_i B = 0$ are quadratic for $i = 1, 3$ and there exists a unique solution in explicit form, as seen in subsection 7.5.1. In lemma 7.1 below, we use the fact that $h_1 \leq h_2 \leq h_3$ on $[0, \infty)$, and the explicit solutions for $i = 1, 3$ to construct a solution $\hat{B}_2(t)$ that satisfies the equation $\mathcal{H}_2\hat{B}_2 = 0$ on $(0, \infty)$ and converges at the 'right' speed to ∞ when t approaches 0.

Lemma 7.1. *Let $T > 0$. For $i = 1, 2, 3$ there exist unique solutions $\hat{B}_i \in C^1((0, T])$ to $(\mathcal{H}_i B) \equiv 0$ satisfying*

(i) $\lim_{t\to 0} \hat{B}_i = \infty$, $i = 1, 2, 3$,

(ii) $\hat{B}_i$ *is decreasing for* $i = 1, 2, 3$,

(iii) $\hat{B}_1(t) \leq \hat{B}_2(t) \leq \hat{B}_3(t)$

Moreover $\hat{B}_1, \hat{B}_3$ *are given explicitly by*

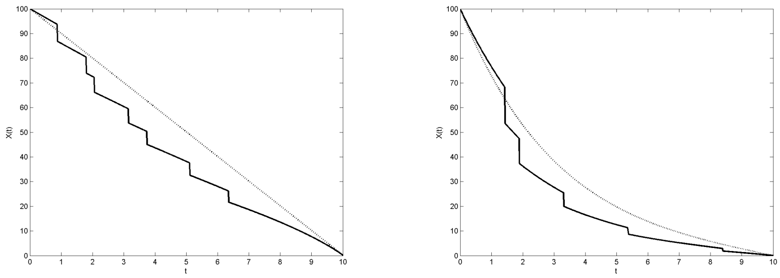
$$\hat{B}_i(t) = \frac{\mu}{2}\frac{\frac{\lambda\beta^2}{\gamma}\left(1 - e^{tc_i}\right) + c_i\left(1 + e^{tc_i}\right)}{e^{tc_i} - 1}, \; i = 1, 3,$$

where $c_1 = \sqrt{\left(\frac{\lambda\beta^2}{\gamma}\right)^2 + 4\alpha/\mu}$ *and* $c_3 = \sqrt{\left(\frac{\lambda\beta^2}{\gamma}\right)^2 + 4/\mu(\alpha + \beta^3\lambda\kappa/\gamma^2)}$.

We can now solve the optimal control problem for $\kappa > 0$.

Theorem 7.4. *The value function for the finite-time problem in* (7.19) *is given by* $\mathcal{V}_3(t, x) = \hat{B}_2(T - t)x^2$ *with* $\hat{B}_2$ *given by lemma 7.1. The associated trading problem admits a unique solution given by the optimal trading strategies*

$$\begin{aligned} \hat{m}(t, X(t)) &= X(t)\frac{\hat{B}_2(T - t)}{\mu} \\ \hat{l}(t, X(t-)) &= X(t-)\frac{\hat{B}_2(T - t)\beta}{\hat{B}_2(T - t)\gamma + \beta\kappa} \end{aligned}$$



(a) Optimal trajectories for risk-neutral ($\alpha = 0$) trader **(b)** Optimal trajectories for risk-averse ($\alpha = 0.000001$) trader

Figure 7.5: The solid line represents $X(t)$ for a general limit order market with $\mu = 0.01, \kappa = 0.001, \lambda = 0.5$ and complete limit order execution ($\beta = \gamma = 1$). The dotted line represents $X(t)$ for a pure market order market with $\mu = 0.01$ and $\kappa = \infty$.

To analyse the structure of the solution in more detail, we proceed as in section 7.3.3, and use a first order approximation of the limit order impact κ. Note that the first order approximation corresponds exactly to the solution $\hat{B}_3$ from lemma 7.1 above. Rewriting the solution, we obtain

$$\hat{B}_3(\tau) = \frac{\mu}{e^{2\tau\sqrt{\mathfrak{R}_m + \frac{\beta^3}{\gamma^2}\lambda\mathfrak{I} + \left(\frac{\lambda\beta^2}{2\gamma}\right)^2}} - 1} \left\{ e^{2\tau\sqrt{\mathfrak{R}_m + \frac{\beta^3}{\gamma^2}\lambda\mathfrak{I} + \left(\frac{\lambda\beta^2}{2\gamma}\right)^2}} \left(\sqrt{\mathfrak{R}_m + \frac{\beta^3}{\gamma^2}\lambda\mathfrak{I} + \left(\frac{\lambda\beta^2}{2\gamma}\right)^2} - \frac{\lambda\beta^2}{2\gamma} \right) + \left(\sqrt{\mathfrak{R}_m + \frac{\beta^3}{\gamma^2}\lambda\mathfrak{I} + \left(\frac{\lambda\beta^2}{2\gamma}\right)^2} + \frac{\lambda\beta^2}{2\gamma} \right) \right\},$$

and thus, as in the case without limit order impact the optimal strategy approaches the infinite-time strategy given in equation (7.14), when the time-to-go $\tau = T - t$ tends to ∞.

The short time horizon behaviour is the same as in the case without limit order impact:

- The market order rate is approximately such that the remaining shares are liquidated linearly.
- A proportion of approximately $\frac{\beta}{\gamma}$ of the remaining shares is placed as a limit order.

Figure 7.5 shows optimal trajectories for the finite-time horizon case, both for a risk neutral and a risk-averse trader. For comparison, we added the dotted line, which denotes the corresponding optimal trajectory in a pure market order market. Towards the end of the trading program (short time horizon), we see that the optimal trajectory follows roughly a linear liquidation with market orders.

7.6 Proofs

Proof of Theorem 7.1. We use a verification argument. Let $x > 0$, $(m, l) \in \Theta_1(x)$ and $X(t)$ be given by (7.1). By Itô's formula

$$v(X(t)) - v(X(0)) = -\int_0^t v_X(X(s))m(s)\mathrm{d}s + \int_0^t \int_0^1 v(X(s-) - zl(s)) - v(X(s-))\nu(\mathrm{d}z)\mathrm{d}s + M(t) \tag{7.20}$$

where

$$M(t) = \int_0^t \int_0^1 v(X(s-) - zl(s)) - v(X(s-))\tilde{N}(\mathrm{d}s, \mathrm{d}z)$$

is a local martingale. Define $Y(t) := \int_0^t \left\{\mu m^2(s)) + \kappa\lambda\beta l^2(s) + \alpha X^2(s)\right\} \mathrm{d}s + v(X(t))$. Plugging in equation (7.20) and using the definition $v(x) = \hat{A}x^2$, we obtain

$$Y(t) = I(t) + M(t) + v(x) \tag{7.21}$$

where

$$\begin{aligned} I(t) = \int_0^t &\left\{\mu m^2(s) + \lambda\beta\kappa l^2(s) + \alpha X^2(s) - \right. \\ &\left. 2\hat{A}X(s-)m(s) + \hat{A}\lambda\gamma l^2(s) - 2\hat{A}l(s)X(s-)\lambda\beta\right\} \mathrm{d}s \end{aligned} \tag{7.22}$$

By choice of $(\hat{m}, \hat{l})$ and $\hat{A}$, we have $I(t) \geq 0$ for all $(m, l) \in \Theta_1(x)$ and $I(t) = 0$ for $(m, l) = (\hat{m}, \hat{l})$. Moreover, note that

$$M(t) = \int_0^t \int_0^1 \left\{\hat{A}z^2 l^2(s) - 2\hat{A}zl(s)X(s-)\right\} \tilde{N}(\mathrm{d}s, \mathrm{d}z)$$

so for $(m, l) \in \Theta_1(x)$ $\mathrm{E}\left[[M](t)\right] < \infty$ for all $t \geq 0$ and $M(t)$ is a true martingale. Hence we can take expectation in (7.21) and obtain

$$\mathrm{E}\left[\int_0^t \mu m^2(s) + \kappa\lambda\beta l^2(s) + \alpha X^2(s)\mathrm{d}s\right] + \mathrm{E}[v(X(t))] = \mathrm{E}[I(t)] + v(x) \geq v(x) \tag{7.23}$$

By lemma 7.2 below, we have that $(\hat{m}, \hat{l})$ is admissible under condition (C) and thus $\lim_{t\to\infty} \mathrm{E}[v(\hat{X}(t))] = 0$. Now take the limit $t \to \infty$ in (7.23) and obtain that $\mathcal{J}(0, x, (m, l)) \geq v(x)$ for all $(m, l) \in \Theta_1(x)$ with equality for $(m, l) = (\hat{m}, \hat{l})$. Thus $\mathcal{V}_1(x) = v(x) = \hat{A}x^2$, and the infimum in (7.4) is attained at $(\hat{m}, \hat{l})$. The uniqueness is a direct consequence of the strict convexity of the mapping $(m, l) \mapsto \mathcal{J}(0, x, (m, l))$. □

Lemma 7.2. *Suppose that condition (C) holds. Then the strategy $(\hat{m}, \hat{l})$ as defined in theorem 7.1 is admissible, that is $(\hat{m}, \hat{l}) \in \Theta_1(x)$.*

Proof. Clearly $(\hat{m}, \hat{l})$ are predictable, so we have (i). Let $\hat{X} = X_{(\hat{m},\hat{l})}$ be the solution of equation (7.1) with $(m, l) = (\hat{m}, \hat{l})$. Putting $(\hat{m}, \hat{l}, \hat{X})$ into equation (7.23), using $I(t) = 0$, and letting $t \to \infty$, we obtain

$$\mathrm{E}\left[\int_0^\infty \mu\hat{m}^2(s)) + \kappa\lambda\beta\hat{l}^2(s) + \alpha\hat{X}^2(s)\mathrm{d}s\right] \leq v(x) < \infty$$

so $\hat{m}, \hat{l}, \hat{X} \in L^2([0, \infty) \times \Omega)$, and we have (iii).

Next, we will show that

$$\lim_{t\to\infty}\int_0^t \hat{m}(s)\mathrm{d}s + \int_0^t\int_0^1 z\hat{l}(s)N(\mathrm{d}s,\mathrm{d}z) = x \text{ a.s.} \tag{7.24}$$

which shows (iv).

By proposition 7.1, we know that

$$\log\left(\left|\hat{X}(t)\right|\right) = \log x + \sum_{i=1}^{\pi(t)} \log\left|1 - Z_i\frac{\hat{A}\beta}{\hat{A}\gamma+\beta\kappa}\right| - \frac{\hat{A}}{\mu}t.$$

Note also that

$$\frac{\pi(t)}{t} \to_{t\to\infty} \lambda \text{ a.s.}$$

Thus by independence of the Z_i and π, and the strong law of large numbers, we have

$$\frac{1}{t}\sum_{i=1}^{\pi(t)} \log\left|1 - Z_i\frac{\hat{A}\beta}{\hat{A}\gamma+\beta\kappa}\right| \to_{t\to\infty} \lambda \mathrm{E}\left[\log\left|1 - Z_i\frac{\hat{A}\beta}{\hat{A}\gamma+\beta\kappa}\right|\right] \text{ a.s.}$$

By condition (C), we obtain

$$\lim_{t\to\infty}\log\left(\left|\hat{X}(t)\right|\right) = -\infty \text{ a.s.}$$

and thus we have (iv).

Let us now show $\mathrm{E}[\hat{X}^2(t)] \to 0$ as $t \to \infty$. Again by proposition 7.1, we have that

$$\hat{X}(t) = xe^{-\frac{\hat{A}}{\mu}t}\prod_{i=1}^{\pi(t)}(1 - Z_i\frac{\hat{A}\beta}{\hat{A}\gamma+\beta\kappa})$$

and, conditioning on $\pi(t)$, we obtain

$$\mathrm{E}[\hat{X}^2(t)] = x^2e^{-(2\frac{\hat{A}}{\mu}+c_2)t} \to 0 \text{ as } t\to\infty,$$

where

$$c_2 = 2\frac{\hat{A}\beta^2}{\hat{A}\gamma+\beta\kappa} - \gamma\left(\frac{\hat{A}\beta}{\hat{A}\gamma+\beta\kappa}\right)^2 > 0.$$

Thus we have (iv). Admissibility properties (ii), (v) and (vii) follow similarly, using the explicit formula for $\hat{X}(t)$ from proposition 7.1. □

Proof of Proposition 7.1. The formula follows directly by the Doléans-Dade exponential from Proposition 2.1. □

Proof of Corollary 7.1. See Kyprianou (2006), Bertoin (1998), Veillette and Taqqu (2010) for a general discussion of first passage times of subordinators. In our particular case, we use the explicit form of $\hat{X}$ from the previous proposition 7.1. After some transformations, we obtain

$$\tau_a = \inf\{t \geq 0 : L(s) > a\}$$

where

$$L(s) = \frac{\hat{A}}{\mu}s + \log\left(\frac{\hat{A}+\kappa}{\kappa}\right)N(s)$$

is a subordinator with characteristic exponent given by (7.17), so by the Lévy-Khintchine formula

$$\mathrm{E}\left[e^{-uL(s)}\right] = e^{-s\psi(u)}.$$

Since L is strictly increasing, we have

$$\{L(s) < t\} = \{\tau_t > s\}.$$

Now use Fubini's theorem and write $F_s(t) = \mathrm{P}(L(s) < t)$ to compute

$$\begin{aligned}
\widehat{e_n}(u) &= \int_0^\infty e_n(t)e^{-ut}\mathrm{d}t \\
&= \int_0^\infty \int_0^\infty ns^{n-1}\mathrm{P}(\tau_t > s)e^{-ut}\mathrm{d}s\mathrm{d}t \\
&= \int_0^\infty \int_0^\infty ns^{n-1}\mathrm{P}(L(s) < t)e^{-ut}\mathrm{d}s\mathrm{d}t \\
&= n\int_0^\infty s^{n-1}\left(\int_0^\infty F_s(t)e^{-ut}\mathrm{d}t\right)\mathrm{d}s
\end{aligned}$$

Using integration by parts, we obtain

$$\begin{aligned}
\widehat{e_n}(u) &= \frac{n}{u}\int_0^\infty s^{n-1}\left(\int_0^\infty e^{-ut}\mathrm{d}F_s(t)\mathrm{d}t\right)\mathrm{d}s \\
&= \frac{n}{u}\int_0^\infty s^{n-1}\mathrm{E}\left[e^{-uL(s)}\right]\mathrm{d}s \\
&= \frac{n}{u}\int_0^\infty s^{n-1}e^{-s\psi(u)}\mathrm{d}s \\
&= \frac{nn!}{u\psi(u)^n}
\end{aligned}$$

□

Proof of Corollary 7.2. From proposition 7.1 we deduce that

$$\begin{aligned}
L &= e^{-c_1\tau_1}c_2Z_1 + e^{-c_1\tau_1}e^{-c_1\tau_2}(1-c_2Z_1)c_2Z_2 + \dots \\
&= \sum_{n=1}^{\infty} e^{-c_1\tau_n}c_2Z_n\prod_{k=1}^{n-1}e^{-c_1\tau_k}(1-c_2Z_k)
\end{aligned}$$

where τ_k is the k-th interarrival time of the Poisson process $\pi(t)$. The τ_k's are i.i.d. following an exponential distribution of rate λ and independent of the Z_k's. By linearity of the expectation and independence, we obtain for $1 - c_2\beta \neq 0$

$$\begin{aligned}
\mathrm{E}[L] &= \sum_{n=1}^{\infty}\frac{\lambda c_2\beta}{\lambda+c_1}\left(\frac{\lambda}{\lambda+c_1}(1-c_2\beta)\right)^{n-1} \\
&= \frac{\frac{\lambda c_2\beta}{\lambda+c_1}}{1-\frac{\lambda}{\lambda+c_1}(1-c_2\beta)} \\
&= \frac{\lambda c_2\beta}{\lambda c_2\beta + c_1}
\end{aligned}$$

and for $1 - c_2\beta = 0$

$$\mathrm{E}[L] = \frac{\lambda}{\lambda + c_1}.$$

□

Proof of Theorem 7.2. We start by showing that $\Theta_2(x) \subset \Theta_1(x)$: Let $(m,l) \in \Theta_2(x)$. Then x is non-increasing on $[0,\infty)$, so $X(t)$ and $l(t)$ are bounded by $x = X(0)$ and we clearly have for all $t \geq 0$

$$\mathrm{E}\left[\int_0^t l^4(s)\mathrm{d}s\right] < \infty, \ \mathrm{E}\left[\int_0^t l^2(s)X^2(s-)\mathrm{d}s\right] < \infty.$$

Moreover $X^2(t) \to 0$ as $t \to \infty$ and $X^2(t)$ is bounded by x^2, so by the dominated convergence theorem we have that $\lim_{t\to\infty} \mathrm{E}[X^2(t)] = 0$. Therefore $(m,l) \in \Theta_1(x)$.

Next, we show that $(\hat{m},\hat{l})$ are admissible: By definition, $\hat{k} \leq 1$, so $\hat{X} = X_{(\hat{m},\hat{l})}$ is non-increasing. The bound $\hat{X}(t) \leq x$ guarantees that the integrability conditions on $(\hat{m},\hat{l})$ and $\hat{x}$ hold. Using similar explicit calculations of the optimal trajectory as in proposition 7.1, we deduce that

$$\hat{X}(t) \leq e^{-\frac{\hat{A}}{\mu}t} \text{ a.s.}$$

so $\lim_{t\to\infty} \hat{X}(t) = 0$ a.s. By definition $\hat{m}$ and $\hat{l}$ are predictable, so $(\hat{m},\hat{l}) \in \Theta_2(x)$.

Since $\Theta_2(x) \subset \Theta_1(x)$, we can proceed analogously to the proof of Theorem 7.1. It only remains to show that (cf. equation (7.22)) for

$$I((m,l);t) = \int_0^t \left\{\mu m^2(s) + \lambda\beta\kappa l^2(s) + \alpha X^2(s) - 2\hat{A}X(s-)m(s) + \hat{A}\lambda\gamma l^2(s) - 2\hat{A}l(s)X(s-)\lambda\beta\right\} \mathrm{d}s$$

we have

$$\begin{aligned} &I((m,l);t) \geq 0, \ \forall (m,l) \in \Theta_2(x) \\ &I((m,l);t) = 0, \ (m,l) = (\hat{m},\hat{l}) \end{aligned} \tag{7.25}$$

Thus to prove (7.25) is suffices to show that

$$G(m,l,\hat{A}) \geq G(\hat{m},\hat{l},\hat{A}) = 0, \ \forall m \geq 0, \ l \in [0,x] \tag{7.26}$$

where

$$G(m,l,A) := \mu m^2 + \lambda\beta\kappa l^2 + \alpha x^2 - 2Axm + A\lambda\gamma l^2 - 2Alx\lambda\beta$$

and $\hat{m} = \hat{A}/\mu x$, $\hat{l} = \hat{k}x$. Considering first the unconstrained problem, we obtain

$$\arg \min_{(m,l)\in\mathbb{R}^2} G(m,l,A) = (x\frac{A}{\mu}, x\frac{A\beta}{A\gamma + \beta\kappa})$$

and $A \mapsto G((x\frac{A}{\mu}, x\frac{A\beta}{A\gamma+\beta\kappa}), A)$ has the unique positive real root A_1. Thus if $\frac{A_1\beta}{A_1\gamma+\beta\kappa} \leq 1$ (or equivalently $A_1 \leq \frac{\beta\kappa}{\beta-\gamma}$), then condition (7.26) is satisfied.

Now assume $\frac{A_1\beta}{A_1\gamma+\beta\kappa} > 1$. Then $(\hat{m},\hat{l},\hat{A}) = (x\frac{A_2}{\mu}, x, A_2)$. The equality in condition (7.26) is checked by direct calculation.

By choice of $\hat{m}$, $G(m,l,\hat{A}) \geq G(\hat{m},l,\hat{A})$. It is thus sufficient to show

$$F(l) := G(\hat{m},l,\hat{A}) \geq 0, \forall l \in [0,x] \tag{7.27}$$

Since $F(\hat{l}) = F(x) = 0$, it is enough to show that F is non-increasing on $[0,x]$. This is equivalent to $F^{'}(l) \leq 0$, $l \in [0,x]$, i.e.

$$2l\lambda(A_2\gamma + \beta\kappa) - 2A_2x\beta\lambda \leq 0,\ l \in [0,x]$$

This inequality is satisfied if we can show that $A_2 \geq \frac{\beta\kappa}{\beta-\gamma}$.

We now argue that $A_1 \leq A_2$: For $(m(t),l(t)) = (X(t-)\frac{A_2}{\mu}, X(t-))$ in equation (7.23), we obtain that A_2x^2 is the value associated to the this strategy. This value must be higher than the one obtained by following the optimal (unconstrained) strategy in $\Theta_1(x)$ which equals A_1x^2.

Therefore $A_2 \geq A_1 > \frac{\beta\kappa}{\beta-\gamma}$, so G_1 is non-increasing on $[0,x]$, and thus condition (7.27) and hence also condition (7.26) is satisfied. □

Proof of Theorem 7.3. Define $v(t,x) := \hat{A}(t)x^2$, with $\hat{A}(t) = \hat{B}(T-t)$. Let $t_0 < T$, $(m,l) \in \Theta_3(t_0, x_{t_0})$ and $X(s)$ be the corresponding process of assets holdings. By Itô's formula for $t \in [t_0,T)$

$$\begin{aligned} v(t,X(t)) - v(t_0,X(t_0)) = &\int_{t_0}^{t} v_t(s,X(s-))\mathrm{d}s - \int_{t_0}^{t} v_X(s,X(s-))m(s)\mathrm{d}s \\ &+ \int_{t_0}^{t}\int_0^1 \{v(s,X(s-) - zl(s)) - v(s,X(s-))\}\,\nu(\mathrm{d}z)\mathrm{d}s + M(t) \end{aligned} \tag{7.28}$$

where

$$M(t) = \int_{t_0}^{t}\int_0^1 \{v(s,X(s-) - zl(s)) - v(s,X(s-))\}\,\tilde{N}(\mathrm{d}s,\mathrm{d}z)$$

is a local martingale. Define

$$Y(t) := \int_{t_0}^{t} \left\{\mu m^2(s)) + \alpha X^2(s)\right\}\mathrm{d}s + v(t,X(t)).$$

Plug in equation (7.28) to obtain

$$Y(t) = I(t) + M(t) + v(t_0,X(t_0)) \tag{7.29}$$

where

$$\begin{aligned} I(t) = &\int_{t_0}^{t}\{\mu m^2(s) + \alpha X^2(s) \\ &+ \hat{A}^{'}(s) - 2\hat{A}(s)X(s-)m(s) + \hat{A}(s)\lambda\gamma l^2(s) - 2\hat{A}(s)l(s)X(s-)\lambda\beta\}\mathrm{d}s. \end{aligned} \tag{7.30}$$

For fixed X the function

$$g(m,l) = \mu m^2 + \alpha X^2 + \hat{A}^{'} - 2\hat{A}Xm + \hat{A}\lambda\gamma l^2 - 2\hat{A}lX\lambda\beta$$

takes its local minimum at $\left(X\frac{\hat{A}}{\mu}, X\frac{\beta}{\gamma}\right)$ which is also the global minimum, as g is quadratic in (m,l). Hence we obtain the candidate optimal strategy given by

$$\hat{m}(t, X(t-)) = X(t-)\frac{\hat{A}(t)}{\mu}$$
$$\hat{l}(t, X(t-)) = X(t-)\frac{\beta}{\gamma}$$

and using $(m,l) = (\hat{m}, \hat{l})$ in (7.30) gives

$$I(t) = \int_{t_0}^{t} X^2(s)(\mathcal{H}\hat{A})(s)\mathrm{d}s$$

where $\mathcal{H}$ is the differential operator

$$(\mathcal{H}A)(s) = A^{'}(s) - \frac{A^2(s)}{\mu} - A(s)\frac{\beta^2\lambda}{\gamma} + \alpha \tag{7.31}$$

Now note that $\hat{A}$ solves $(\mathcal{H}A) = 0$. Hence

$$I(t) \geq 0, \quad \forall (m,l) \in \Theta_3(t_0, x_{t_0}), \tag{7.32}$$
$$I(t) = 0, \quad (m,l) = (\hat{m}, \hat{l}). \tag{7.33}$$

Moreover, note that

$$M(t) = \int_{t_0}^{t}\int_{0}^{1} \left\{\hat{A}(s)z^2l^2(s) - 2\hat{A}(s)zl(s)X(s-)\right\}\tilde{N}(\mathrm{d}s, \mathrm{d}z)$$

so for $(m,l) \in \Theta_3(t_0, x_{t_0})$ $\mathrm{E}\left[[M](t)\right] < \infty$ for all $t \geq 0$ and $M(t)$ is a true martingale. Hence we can take expectation in (7.29) and obtain

$$\mathrm{E}\left[\int_{t_0}^{t}\left\{\mu m^2(s) + \alpha X^2(s)\right\}\mathrm{d}s\right] + \mathrm{E}[v(t, X(t))] = \mathrm{E}[I(t)] + v(t_0, X(t_0)) \tag{7.34}$$

By lemma 7.3 $\lim_{t\to T}\mathrm{E}[v(X(t))] = 0$. Thus as we take the limit $t \to T$ in (7.34), using (7.32) and (7.33), we obtain that $\mathcal{J}(t_0, x_{t_0}, (m,l)) \geq v(t_0, x_{t_0})$ for all $(m,l) \in \Theta_3(t_0, x_{t_0})$ with equality for $(m,l) = (\hat{m}, \hat{l})$. By lemma 7.4 below, $(\hat{m}, \hat{l})$ is admissible, so $\mathcal{V}_3(t,s) = v(t,x) = \hat{A}(t)x^2$, and the infimum in (7.4) is attained at $(\hat{m}, \hat{l})$. Uniqueness follows again from strict convexity of the performance function. □

Lemma 7.3. *Let $(m,l) \in \Theta_3(t_0, x_{t_0})$ and x be the corresponding process of share holdings. Then*

$$\lim_{t\to T}\mathrm{E}\left[\hat{B}(T-t)X^2(t)\right] = 0$$

Proof. Let $\tau = T - t$ be the 'time-to-go'. We want to show that $\lim_{\tau\to 0}\mathrm{E}[\hat{B}(\tau)X^2(T-\tau)] = 0$. Decompose

$$\begin{aligned}\mathrm{E}\left[\hat{B}(\tau)X^2(T-\tau)\right] &= \mathrm{E}\left[\hat{B}(\tau)X^2(T-\tau)1_{F_\tau}\right] + \mathrm{E}\left[\hat{B}(\tau)X^2(T-\tau)1_{F_\tau^c}\right]\\ &= I_1(\tau) + I_2(\tau)\end{aligned}$$

where $F_\tau = \{\pi(T) = \pi(T-\tau)\}$ is the event that no limit order is executed in time $[T-\tau, T]$. By standard results about the Poisson process, we have $\mathrm{P}(F_\tau) = e^{-\lambda\tau}$. We will consider I_1 and I_2 separately

First Term:
On F_τ, no more limit orders are executed, so

$$X(T) - X(T-\tau) = -\int_{T-\tau}^{T} m(s)\mathrm{d}s$$

and thus $X(T-\tau) = \int_{T-\tau}^{T} m(s)\mathrm{d}s$. By Hölder's inequality

$$X^2(T-\tau) \leq \tau \int_{T-\tau}^{T} m^2(s)\mathrm{d}s \tag{7.35}$$

Using e.g. the Taylor expansion of the exponential function around zero, it is easy to see that $\frac{1}{\hat{B}}$ has a pole of order one at $t = 0$, hence there exists a constant $L_1 < \infty$, independent of t such that

$$t\hat{B}(t) \leq L_1,\ t \in [0, T]. \tag{7.36}$$

Hence using first (7.35), then the bound (7.36):

$$\begin{aligned}
I_1(\tau) &= \mathrm{E}\left[\hat{B}(\tau)X^2(T-\tau)1_{F_\tau}\right] \\
&\leq \mathrm{E}\left[\hat{B}(\tau)\tau\int_{T-\tau}^{T} m^2(s)\mathrm{d}s 1_{F_\tau}\right] \\
&\leq \mathrm{E}\left[\hat{B}(\tau)\tau\int_{T-\tau}^{T} m^2(s)\mathrm{d}s\right] \\
&\leq L_1\mathrm{E}\left[\int_{T-\tau}^{T} m^2(s)\mathrm{d}s\right] \to_{\tau\to 0} 0
\end{aligned}$$

Second Term:
Note that F_τ^c only depends on the jumps of $\pi(s)$, $s \in [T-\tau, T]$ and

$$X(T-\tau) = x_{t_0} - \int_{t_0}^{T-\tau} m(s)\mathrm{d}s - \int_{t_0}^{T-\tau}\int_0^1 zl(s)N(\mathrm{d}s, \mathrm{d}z)$$

is $\mathcal{F}_{T-\tau}$-measurable. Hence $X(T-\tau)$ and F_τ^c are independent. Moreover, we have the bound

$$(1 - e^{-\lambda t})\hat{B}(t) \leq L_2,\ t \in [0, T] \tag{7.37}$$

for some constant $L_2 < \infty$, independent of t. Hence we deduce

$$\begin{aligned}
I_2(\tau) &= \mathrm{E}\left[\hat{B}(\tau)1_{F_\tau^c}X^2(T-\tau)\right] \\
&= (1 - e^{-\lambda\tau})\hat{B}(\tau)\mathrm{E}\left[X^2(T-\tau)\right] \\
&\leq L_2\mathrm{E}\left[X^2(T-\tau)\right] \to_{\tau\to 0} 0
\end{aligned}$$

by admissibility property (vi). □

Lemma 7.4. *Let $(\hat{m}, \hat{l})$ be as the strategy as defined in theorem 7.3. Then it is admissible, that is $(\hat{m}, \hat{l}) \in \Theta_3(t_0, x_{t_0})$.*

Proof. By definition of $(\hat{m}, \hat{l})$, it is predictable, so we have (i). Property (iii) follows from equation (7.34) by taking $(m, l) = (\hat{m}, \hat{l})$. Let now $\hat{X} = X_{(\hat{m},\hat{l})}$ denote the process of asset holdings corresponding to the strategy $(\hat{m}, \hat{l})$. We show property (iv), i.e. $\hat{X}(t) \to 0$ a.s. as $t \to T$. By standard results on the Poisson process, we know that a.s. there is *no* jump of π at time T. Thus we can argue ω-wise: there exists $\tau < T$ (depending on ω) such that no jump occurs in $(\tau, T]$. Thus on $(\tau, T]$ we have the dynamics $\mathrm{d}\hat{X}(s) = -\hat{X}(s-)\hat{B}(T-s)/\mu \mathrm{d}s$, and thus for $t \in (\tau, T]$

$$\hat{X}(t) = \hat{X}(\tau) \exp\left\{ -\int_\tau^t \frac{\hat{B}(T-s)}{\mu} \mathrm{d}s \right\} \tag{7.38}$$

There exist strictly positive constants c_1, c_2 independent of $t \in (\tau, T)$ such that

$$\int_\tau^t \hat{B}(T-s)\mathrm{d}s \geq c_1 \int_{T-t}^{T-\tau} \frac{1}{e^{c_2 s} - 1} \mathrm{d}s \tag{7.39}$$

and for $0 < t_1 < t_2$

$$\int_{t_1}^{t_2} \frac{1}{e^{c_2 s} - 1} \mathrm{d}s = \frac{1}{c_2} \left(c_2(t_1 - t_2) + \log\left(e^{c_2 t_2} - 1\right) - \log\left(e^{c_2 t_1} - 1\right)\right) \to_{t_1 \to 0} \infty$$

so $\int_\tau^t \frac{\hat{B}(T-s)}{\mu} \mathrm{d}s \to_{t \to T} \infty$ and thus by equation (7.38) $\hat{X}(t) \to 0$ a.s. as $t \to T$.

Next we show property (vi), i.e. $\mathrm{E}[\hat{X}^2(t)] \to 0$ as $t \to T$. Similar to the proof of proposition 7.1, one can show that

$$\hat{X}(t) = \hat{X}(t_0) \exp\left\{ -\int_{t_0}^t \frac{\hat{B}(T-s)}{\mu} \mathrm{d}s \right\} \prod_{n=1}^{\pi(t)} \left(1 - \frac{\beta}{\gamma} Z_n\right) \tag{7.40}$$

By conditioning on $\pi(t)$ we compute

$$\begin{aligned}
\mathrm{E}[\hat{X}^2(t)] &= X^2(t_0) \exp\left\{ -2\int_{t_0}^t \frac{\hat{B}(T-s)}{\mu} \mathrm{d}s \right\} \mathrm{E}\left[\prod_{n=1}^{\pi(t)} \left(1 - \frac{2\beta}{\gamma} Z_n + \frac{\beta^2}{\gamma^2} Z_n^2\right)\right] \\
&= X^2(t_0) \exp\left\{ -2\int_{t_0}^t \frac{\hat{B}(T-s)}{\mu} \mathrm{d}s \right\} \sum_{N=0}^{\infty} \prod_{n=1}^{N} \mathrm{E}\left[1 - \frac{2\beta}{\gamma} Z_n + \frac{\beta^2}{\gamma^2} Z_n^2\right] \mathrm{P}(\pi(t) = N) \\
&= X^2(t_0) \exp\left\{ -2\int_{t_0}^t \frac{\hat{B}(T-s)}{\mu} \mathrm{d}s \right\} \sum_{N=0}^{\infty} \mathrm{E}\left[1 - \frac{2\beta}{\gamma} Z_1 + \frac{\beta^2}{\gamma^2} Z_1^2\right]^N \frac{(\lambda t)^N e^{-\lambda t}}{N!} \\
&= X^2(t_0) \exp\left\{ -2\int_{t_0}^t \frac{\hat{B}(T-s)}{\mu} \mathrm{d}s - \frac{\beta^2 \lambda t}{\gamma} \right\} \to 0, \text{ as } t \to T
\end{aligned}$$

The other admissibility properties (ii), (v) and (vii) follow similarly from the definition of $(\hat{m}, \hat{l})$ and the explicit formula for the optimal trajectory (7.40). □

Proof of Lemma 7.1. Let $b_i^{(0)}$ be the largest real root of h_i and $b^{(0)} = \max\left\{b_1^{(0)}, b_3^{(0)}, b_3^{(0)}\right\}$ Thus $h_i(b) \leq 0$ for all $b \geq b^{(0)}$, $i = 1, 2, 3$. Let $(b^{(n)})_{n\in\mathbb{N}}$ be a sequence with $b^{(n)} \in (b^{(0)}, \infty)$ such that $b^{(n)} \to \infty$ as $n \to \infty$. We will start by showing for each n the existence of $B_i^{(n)}$ satisfying

$(\mathcal{H}_i B_i^{(n)}) \equiv 0$ with initial condition $B_i^{(n)}(0) = b^{(n)}$, $B_i^{(n)}$ decreasing for $i = 1, 2, 3$ and satisfying $B_1^{(n)}(t) \leq B_2^{(n)}(t) \leq B_3^{(n)}(t)$.

Let n be fixed. Note that h_i is locally Lipschitz on $[0, \infty)$, so by the Picard-Lindelöf theorem, there exists a unique solution on some common open interval of existence I. Call the unique solution $B_i^{(n)}$. Since $h_1 \leq h_2 \leq h_3$ on $[0, \infty)$ and $\frac{\partial B_i^{(n)}(t)}{\partial t} = h_i(B_i^{(n)}(t))$, we have $B_1^{(n)}(t) \leq B_2^{(n)}(t) \leq B_3^{(n)}(t)$. $B_i^{(n)}$ satisfies $\frac{\partial B_i^{(n)}(t)}{\partial t} = h_i(B_i^{(n)}(t)) \leq 0$ for $B_i^{(n)}(t) \geq b_i^{(0)}$, so $B_i^{(n)}$ is decreasing. As $B_i^{(n)}$ is decreasing and bounded below by $b_i^{(0)}$, it cannot explode, and we can assume that the maximal interval of existence is $I = [0, \infty)$. Note also that for $i = 1, 3$ we have the explicit solution

$$B_i^{(n)}(t) = \frac{\mu}{2} \frac{\frac{\lambda\beta^2}{\gamma}\left(1 - e^{tc_i}\right) + c_i\left(1 + e^{tc_i}\right)}{e^{tc_i} - k_i^{(n)}}, \; i = 1, 3$$

where

$$k_i^{(n)} = 1 - \frac{c_i \mu}{b^{(n)}}$$

increases to 1 as $b^{(n)} \to \infty$.

By inspection, we have $B_i^{(n)}(t) \to \hat{B}_i(t)$ as $n \to \infty$ for $i = 1, 3$ and we define the pointwise limit $\hat{B}_2(t) := \lim_{n\to\infty} B_2^{(n)}(t)$. Note that this pointwise limit exists for $t > 0$, because $B_2^{(n)}(t)$ is increasing in n and bounded above by $\hat{B}_3(t)$. Then the properties $(i) - (iii)$ follow immediately from the corresponding properties of $B_i^{(n)}$. It remains to check that $\hat{B}_2(t)$ is continuously differentiable and satisfies $(\mathcal{H}_2 \hat{B}_2) \equiv 0$ on $(0, T]$. For this, it is enough to show that for every $\epsilon > 0$

$$\begin{aligned} B_2^{(n)} &\to \hat{B}_2, \text{ uniformly on } [\epsilon, T], \text{ as } n \to \infty \\ h_2(B_2^{(n)}) &\to h_2(\hat{B}_2), \text{ uniformly on } [\epsilon, T], \text{ as } n \to \infty \end{aligned} \tag{7.41}$$

Note that h_2 is decreasing on $(0, \infty)$ and that $B_2^{(n)} \leq B_2^{(m)}$ for $n \leq m$. For $n \leq m$ define $g_{(m,n)}(t) = B_2^{(m)}(t) - B_2^{(n)}(t)$. Then $g'_{(m,n)}(t) = h_2(B_2^{(m)}(t)) - h_2(B_2^{(n)}(t)) \leq 0$. Thus $g_{(m,n)}$ is decreasing, so for $t_1 \leq t_2$

$$B_2^{(m)}(t_1) - B_2^{(n)}(t_1) \geq B_2^{(m)}(t_2) - B_2^{(n)}(t_2)$$

and letting $m \to \infty$

$$0 \leq \hat{B}_2(t_2) - B_2^{(n)}(t_2) \leq \hat{B}_2(t_1) - B_2^{(n)}(t_1)$$

Let now $\delta > 0$. By pointwise convergence of $B_2^{(n)}(\epsilon) \to \hat{B}_2(\epsilon)$ $\exists N$ such that $\hat{B}_2(\epsilon) - B_2^{(n)}(\epsilon) < \delta$ for $n \geq N$. Thus for all $t \in [\epsilon, T]$

$$0 \leq \hat{B}_2(t) - B_2^{(n)}(t) \leq \hat{B}_2(\epsilon) - B_2^{(n)}(\epsilon) < \delta$$

for $n \geq N$, and thus the convergence is uniform. In particular this implies that $\hat{B}_2$ is continuous on $[\epsilon, T]$ and since $h_2(B_2^{(n)})$ is a monotone decreasing function, the uniform convergence of $h_2(B_2^{(n)}) \to h_2(\hat{B}_2)$ follows by Dini's theorem. □

Proof of Theorem 7.4. The arguments are almost the same as in the proof of Theorem 7.3 with $\hat{B}$ replaced by $\hat{B}_2$, with a few minor changes, relying on the upper and lower bounds for $\hat{B}_2$

$$\hat{B}_1(t) \leq \hat{B}_2(t) \leq \hat{B}_3(t). \tag{7.42}$$

In lemma 7.3, use $\hat{B}_2 \leq \hat{B}_3$, and the explicit form of $\hat{B}_3$ to obtain analogues for the bounds (7.36) and (7.37).

In lemma 7.4, use $\hat{B}_1 \leq \hat{B}_2$, and the explicit form of $\hat{B}_1$ to obtain the bound (7.39) and thus

$$\int_{\tau}^{t} \frac{\hat{B}_2(T-s)}{\mu} \mathrm{d}s \to_{t\to T} \infty. \tag{7.43}$$

It then follows that $\hat{X}(t) \to 0$ a.s. as $t \to T$.

To show that $\mathrm{E}[\hat{X}^2(t)] \to 0$ as $t \to T$, we use the explicit formula

$$\hat{X}(t) = X(t_0) \exp\left\{-\int_{t_0}^{t} \frac{\hat{B}_2(T-s)}{\mu} \mathrm{d}s\right\} \prod_{n=1}^{\pi(t)} (1 - h(t)Z_n),$$

where

$$h(t) = \frac{\hat{B}_2(T-t)\beta}{\hat{B}_2(T-t)\gamma + \kappa\beta}$$

takes values in $[0, \beta/\gamma]$ because $\hat{B}_2$ decreases from $+\infty$ to 0.

Then, as in lemma 7.4, we have

$$\begin{aligned}\mathrm{E}[\hat{X}^2(t)] &= X^2(t_0) \exp\left\{-2\int_{t_0}^{t} \frac{\hat{B}(T-s)}{\mu} \mathrm{d}s\right\} \mathrm{E}\left[\prod_{n=1}^{\pi(t)} \left(1 - 2h(t)Z_n + h^2(t)Z_n^2\right)\right] \\ &= X^2(t_0) \exp\left\{-2\int_{t_0}^{t} \frac{\hat{B}(T-s)}{\mu} \mathrm{d}s\right\} \exp\left\{\lambda t(-2\beta h(t) + \gamma h^2(t))\right\}.\end{aligned}$$

By (7.43) and using

$$-2\beta h(t) + \gamma h^2(t) \leq 0,$$

we obtain that

$$\mathrm{E}[\hat{X}^2(t)] \to_{t\to T} 0$$

The other admissibility properties follow similarly, using the bounds (7.42). □

Chapter 8

Optimal 'Peg-Cross' strategies

8.1 Model assumptions and problem formulation

At the macro-level a high layer in the algorithmic trading strategy transmits a trade assignment to the micro-trader, e.g. the order to buy a certain number of shares within a certain time horizon. The task of the microtrader is optimal execution of the order, i.e. it needs to decide *where* and *when* to place a unit buy order in the order book. First it needs to split the order in smaller packages and determine a time subinterval in which each small package has to be executed. For each package, it has to decide where to place the order. In this chapter we will focus on the problem of optimal placement of a single package. Typically, the micro-trader has a range of different choices of order placement, more precisely

- place a buy market order,
- place a buy limit order inside the spread,
- place a buy limit order on the best bid or,
- place a buy limit order deep in the order book.

If a market order is placed, the order is immediately executed and the micro-trader can turn to the problem of placing the next package. However, if a limit order is placed, the execution time is random and it might even never be executed. However, the price of the limit order is always better than the price of a market order executed at the time when the limit order was placed, because the trader 'earns' the spread. In practice, the micro-trader will 'wake up' at regular time intervals and check whether the order has been (partially) filled. If the limit order has not been filled completely, the micro-trader needs to decide again where and when to place the remaining order, depending on the current state of the order book, the number of outstanding orders and the remaining time until the complete execution of the package. In addition to the 'traditional' four types of order placements mentioned above, the trader might have access to special markets offering even more choices of order placements. These can for example include iceberg orders, hidden orders, or orders placed in alternative trading venues such as dark pools. In this chapter, however, we will focus on the two order types that are available in any order book, namely market and limit orders.

Our objective is to find and analyse optimal placement strategies for the micro-trader. To make our model tractable, we will restrict to two possible order placements: either place a market order, or a limit order on the best bid.

Let x_0 be the size of the (indivisible) package to be executed. At time $t = 0$ a limit order of total package size x_0 will be placed on the best bid: the micro-trader is *pegging* on the best bid. At any time τ, $0 \leq \tau \leq T$, it can decide to *cross* the spread, i.e. cancel the limit order on the best bid and convert it to a market order.

We assume that the best bid is constant in $[0, T]$: If it increases at time t by one tick, the micro-trader will cancel the current limit order, and place it on the new best bid. This is equivalent to the same problem, starting at time t. If the best bid decreases by one tick, this implies that the limit order has already been filled. From the practical point of view, it is therefore no restriction to assume that the best bid β is constant. Without loss of generality, this constant is equal to zero ($\beta \equiv 0$), therefore the 'cost' of buying x_0 assets by a limit order is normalized to zero. Moreover, we assume that the price impact of the limit order is so small that it can be neglected.

Similar as in chapter 7, we assume that the execution of a limit order is modelled by a Poisson process $N(t)$ of rate λ. This time, however, we are dealing with small packages of orders, so we make the assumption that the entire limit order is filled at jump times of $N(t)$ (and thus, contrary to chapter 7, we do not consider compound Poisson processes).

Whenever the micro-trader decides to cancel the limit order and convert it to a market order, it has to cross the spread. The evolution of the spread determines the optimal strategy:

- When the spread is small, a market order is relatively cheap (although still more expensive than the limit order waiting on the best bid), and it can be optimal to cross the spread and convert the limit order into a market order.
- When the spread is large, a market order is expensive compared to the limit order, and it seems optimal to wait for the limit order execution.

The optimal strategy thus depends on the current spread as well as its future evolution. We will model the spread as a positive diffusion process

$$\mathrm{d}\mathfrak{s}(t) = r(\mathfrak{s}(t))\mathrm{d}t + \sigma(\mathfrak{s}(t))\mathrm{d}W(t) \tag{8.1}$$

started at $\mathfrak{s}(0) = \mathfrak{s}_0$, where W is a standard Brownian motion, independent of the Poisson process N, and r and σ are such that a unique strong solution to (8.1) exists and the spread stays positive all time, i.e. $\mathfrak{s}(t) \geq 0$ for all $t \in [0, T]$.

In reality, the microtrader might look at other quantities than the spread, but the spread is clearly the by far most important one. Harris (2003) remarks

> The spread is the most important factor that traders consider when they decide whether to submit limit orders or market orders. When the spread is wide, immediacy is expensive, market order executions are costly, and limit order submission strategies attractive. When the spread is narrow, immediacy is cheap, and market

Limit order		Market order	
+	earn spread	+	immediate execution
+	no price risk	−	pay spread
−	risk of non-execution	−	pay market impact cost

Table 8.1: Typical advantages and disadvantages of using limit and market orders.

> order strategies are attractive. If you are interested in optimizing your order submission strategies, you must understand what determines bid/ask spreads so that you can judge whether they are wide or narrow, given current market conditions.

To write down the optimal stopping problem, we need to know the cost associated to the execution of a market order. We will work with the same market impact model as in chapter 7. Thus our setting will be the block-shape order book model given in (2.24) with $V^{\mathfrak{A}}(t) \equiv V^{\mathfrak{B}}(t) \equiv \frac{1}{2\mu}$, where the cost of a market order of size x_0 equals $x\mathfrak{s} + \mu x^2$ where $\mathfrak{s}$ denotes the current spread.

Note that this model captures the typical effects of the use of limit versus markt orders on the microlevel, as summarized in table 8.1.

We can now formulate the peg-cross problem: let $\mathcal{F}^{\mathfrak{s}} = (\mathcal{F}^{\mathfrak{s}}_t, 0 \leq t \leq T)$ be the filtration generated by the spread $\mathfrak{s}$, and $\mathcal{T}(t,T)$ be the set of all $\mathcal{F}^{\mathfrak{s}}$ stopping times τ with $t \leq \tau \leq T$. Define for $t \in [0,T]$, $\tau \in \mathcal{T}(t,T)$, $x_t \geq 0$, $\mathfrak{s}_t \geq 0$ the performance function

$$\mathcal{J}(\tau; t, x, \mathfrak{s}) := \mathrm{E}^{(x,\mathfrak{s})}\left[\underbrace{X(\tau)\mathfrak{s}(\tau) + \mu X^2(\tau)}_{\text{cost of market order}} + \underbrace{\alpha \int_t^\tau X^2(s)\mathrm{d}s}_{\text{penalization for holding assets}}\right] \tag{8.2}$$

where the dynamics of $\mathfrak{s}$ and X are given by (8.1) and

$$\mathrm{d}X(s) = -X(s-)\mathrm{d}N(s) = x 1_{\{N(s)=0\}}, \tag{8.3}$$

and $\mathrm{E}^{(x,\mathfrak{s})}$ is the conditional expectation, conditioned on $\mathfrak{s}(t) = \mathfrak{s}$ and $X(t) = x$.

Note that X stands for the remaining amount of shares to be bought if no market order has been executed yet.

We are interested in the optimal crossing time, i.e. $\hat{\tau} \in \mathcal{T}(t,T)$ satisfying

$$\mathcal{J}(\hat{\tau}; t, x_t, \mathfrak{s}_t) = \inf_{\tau \in \mathcal{T}(t,T)} \mathcal{J}(\tau; t, x_t, \mathfrak{s}_t) \tag{8.4}$$

if it exists.

Note that as in chapter 7, we introduced the risk-aversion term $\alpha \int_t^\tau X^2(s)\mathrm{d}s$ which penalizes waiting too long until the limit order is converted into a market order. The value function is defined by

$$\mathcal{V}(t, x_t, \mathfrak{s}_t) := \inf_{\tau \in \mathcal{T}(t,T)} \mathcal{J}(\tau; t, x_t, \mathfrak{s}_t) \tag{8.5}$$

The peg-cross problem (8.5) falls into a class a optimal stopping problems analysed in Pham (1998, Sec.2) using viscosity solutions. There it is shown that under some Lipschitz conditions on r and σ, $\mathcal{V}$ is the unique viscosity solution to the associated HJB equation.

Moreover, by a standard result from Peskir and Shiryaev (2006), we have

Theorem 8.1 (Existence of optimal crossing time)**.** *There exists an optimal crossing time* $\hat{\tau}$ *satisfying* (8.4)*.*

Shiryaev (2008) describes iterative numerical procedures which allow the numerical computation of the value function, using the fact that $\mathcal{V}$ can be characterized as the largest subharmonic function that is dominated by G. However, these methods are non-explicit. Another popular technique for finding $\mathcal{V}$ is the solution of a free-boundary problem. In the remainder of the chapter, we will focus on this approach. Our aim is to find explicit solutions in order to perform a detailed analysis of the optimal strategies. Thus we have to introduce simplifying assumptions.

8.2 Optimal strategies without market impact

In this section, we work under the

Simplifying assumption 8.1. The market impact of the market order is zero, i.e. $\mu = 0$.

The parameter μ models the effect that a large market order cannot be filled solely with limit orders stocked on the best quote, and has to 'eat' deeper into the order book. For the peg/cross problem, however, the assumption that $\mu = 0$ is not very restrictive, as we are dealing with small order packages. In this case, the entire market order can often be filled with limit orders sitting on the best ask, and the trader does not have to buy limit orders on price ticks strictly higher than the best ask. In the case where the size of a package equals the smallest tradeable size, we are automatically in the situation where the market order is completely filled with limit orders on the best ask.

Moreover we introduce

Simplifying assumption 8.2. We consider the infinite time horizon problem $T = \infty$.

At first sight, it seems unreasonable to set $T = \infty$, especially from a practical point of view, since the micro-trader only has a finite time interval to buy the package. However, there are several reasons why we do so: Firstly, the introduction of the risk term $\alpha \int_0^\tau X(u)\mathrm{d}u$ penalizes holding the shares package for a long time. In particular this excludes that the trivial solution consisting in waiting forever for the execution of the cheaper limit order is optimal in the infinite time horizon problem. Secondly, as the results of chapter 7 suggest, the difference of the value function for the infinite and finite time horizon problem is small when α is sufficiently large. Thirdly, as seen in section 7.3 and section 7.5 on optimal portfolio liquidation with infinite and finite time horizon, the long-time behaviour of the optimal solution in the finite time horizon problem ressembles the optimal solution in the infinite time horizon problem. We also saw that that a large part of the structure of the optimal solution is preserved in the infinite time case and that it allows us to identify the key parameters of the problem. In the very short time horizon, the only possible solution is clearly to cross the spread immediately. This corresponds to the linear risk-free liquidation with market orders in the short term horizon of the portfolio liquidation problem in section 7.5. Finally, the value function $\mathcal{V}$ is no longer dependent on the 'time-to-go', when we consider the infinite-time problem and we therefore reduce the number of dimensions by one, and make it mathematically much more tractable.

For $\mu = 0$ and $T = \infty$ the performance function can now be written as

$$\begin{aligned}\mathcal{J}(\tau; x_0, s_0) &= \mathrm{E}\left[X(\tau)\mathfrak{s}(\tau) + \alpha \int_0^\tau X^2(u)\mathrm{d}u\right] \\ &= x_0 \mathrm{E}\left[s(\tau \wedge \tau_N) + \alpha x_0 \int_0^{\tau \wedge \tau_N} \mathrm{d}u\right]\end{aligned}$$

where τ_N is the first jump time of N and $s(t)$ defined by

$$\mathrm{d}s(t) = r(s(t))\mathrm{d}t + \sigma(s(t))\mathrm{d}W(t) - s(t-)\mathrm{d}N(t), \quad s(0) = s_0 \tag{8.6}$$

is the 'killed' spread process, which evolves exactly like the spread $\mathfrak{s}(t)$ until the limit order is executed and jumps to 0 at the limit order execution time. Let $\mathcal{L}_s$ be the infinitesimal generator of s defined by

$$(\mathcal{L}_s \varphi)(s) = r(s)\varphi_s(s) + \frac{\sigma^2(s)}{2}\varphi_{ss}(s) + \lambda(\varphi(0) - \varphi(s))$$

Thus, by changing α without loss of generality, we can set $x_0 = 1$ and consider the performance function

$$\mathcal{J}_1(\tau; s_0) := \mathrm{E}\left[\underbrace{s(\tau \wedge \tau_N)}_{\text{cost of market order}} + \underbrace{\alpha \int_0^{\tau \wedge \tau_N} \mathrm{d}u}_{\text{risk of holding assets}}\right] \tag{8.7}$$

and the associated value function

$$\mathcal{V}_1(s) := \inf_{\tau \in \mathcal{T}(0,\infty)} \mathcal{J}_1(\tau; s) \tag{8.8}$$

Simple, but important upper bounds on $\mathcal{V}_1$ are obtained by considering two naive strategies:

Cross immediately Set $\tau = 0$. This corresponds to immediately crossing the spread. Hence we obtain the trivial bound $\mathcal{V}_1(s) \le s$.

Peg forever Set $\tau = \infty$. This corresponds to never crossing the spread and pegging until the limit order is executed. A simple computation gives the bound $\mathcal{V}_1(s) \le \frac{\alpha}{\lambda}$.

The smaller the spread, the smaller the costs to cross it. This suggests that there is an optimal point $\beta \ge 0$ satisfying

- when $\mathfrak{s} \le \beta$, the extra costs of crossing the spread are cheaper than waiting for the execution of the limit order, and being penalized for share holdings while the limit order is not executed yet,
- when $\mathfrak{s} > \beta$, it is too costly to cross the spread, hence the trader pegs on the best ask and waits for the execution of the limit order.

Thus we define our candidate optimal stopping time by

$$\tau_\beta = \inf\{t \ge 0 : s(t) \le \beta\} \tag{8.9}$$

Heuristic arguments (see, e.g. Peskir and Shiryaev (2006) for a deduction) then suggest that our candidate value function $v(s)$ solves the free-boundary problem

$$\mathcal{L}_s v(s) + \alpha = 0, \quad s > \beta \tag{FB1}$$
$$v(s) = s, \quad s = \beta \tag{FB2}$$
$$v'(s) = 1, \quad s = \beta \tag{FB3}$$
$$v(s) < s, \quad s > \beta \tag{FB4}$$
$$v(s) = s, \quad s < \beta \tag{FB5}$$

Remember that we assumed that the spread $\mathfrak{s}(t)$ must necessarily stay positive at all times, since a negative spread cannot be interpreted economically in a meaningful way (it would imply that a market order is cheaper than the price level of a limit order placed at the same time). In the following we consider two important cases of positive processes to model the spread

- the spread evolves as a geometric Brownian motion, treated in section 8.2.1
- the spread evolves as a CIR process, treated in section 8.2.2

8.2.1 Geometric Brownian motion

We start by considering the case where the spread is modelled by a geometric Browian motion, i.e. the 'killed' spread is given by

$$\mathrm{d}s(t) = rs(t)\mathrm{d}t + \sigma s(t)\mathrm{d}W(t) - s(t-)\mathrm{d}N(t) \tag{8.10}$$

where $r \in \mathbb{R}$ is the drift and $\sigma \geq 0$ is the volatility.

Stochastic spread

We first consider the case where $\sigma > 0$. Using the free-boundary problem (FB1) - (FB5), we will derive a candidate value function v. Since $\mathcal{V}_1(0) = 0$, we must also have $v(0) = 0$. (FB1) is easily recognised as an inhomogeneous Cauchy-Euler equation with $\frac{\alpha}{\lambda}$ as a particular non-zero solution. For the homogeneous equation, seek a solution of the form $v(s) = s^p$. We thus obtain the general solution of (FB1)

$$v(s) = \frac{\alpha}{\lambda} + c_1 s^{p_1} + c_2 s^{p_2} \tag{8.11}$$

where c_1, c_2 are arbitrary constants and

$$p_1 = \frac{1}{2\sigma^2}(\sigma^2 - 2r - \sqrt{(\sigma^2 - 2r)^2 + 8\lambda\sigma^2}) < 0$$
$$p_2 = \frac{1}{2\sigma^2}(\sigma^2 - 2r + \sqrt{(\sigma^2 - 2r)^2 + 8\lambda\sigma^2}) > 0$$

From the bound

$$0 \leq \mathcal{V}_1(s) \leq \min\left\{s, \frac{\alpha}{\lambda}\right\},$$

and the fact that $p_2 > 0$, we deduce that we must have $c_2 = 0$.

We can now use (FB2) and (FB3) to determine the free-boundary point β and the unknown constant c_1, and obtain

$$\beta = \frac{\alpha}{4\lambda(\lambda - r)}(\sigma^2 + 4\lambda - 2r - \sqrt{(\sigma^2 - 2r)^2 + 8\lambda\sigma^2}) \tag{8.12}$$

$$c_1 = \frac{1}{p_1}\beta^{1-p_1} \tag{8.13}$$

Rewriting $\beta = \frac{\alpha}{\lambda}\frac{p_1}{p_1-1}$, and noting that $p_1 < 0$, we observe that, (i) $\beta > 0$ and (ii) $\beta < \frac{\alpha}{\lambda}$. This implies also that $c_1 < 0$.

Using a verification argument, we can show that the candidate function equals the true value function:

Theorem 8.2. *For $\sigma > 0$, the value function $\mathcal{V}_1$ is given by*

$$\mathcal{V}_1(s) = \begin{cases} s & \text{if } s < \beta \\ \frac{\alpha}{\lambda} + c_1 s^{p_1} & \text{if } s \geq \beta \end{cases} \tag{8.14}$$

The stopping time τ_β defined in (8.9) *with β from* (8.12) *is optimal.*

This first result takes the basic structure that we will recognize in the following peg-cross problems. It can be summarized as follows:

- While the spread $\mathfrak{s}(t)$ is above the critial level β, it is too expensive to cross the spread, and the micro-trader *pegs* on the best bid and waits (until either the limit order is executed or the spread drops to a level where it is favorable to cross). The costs associated to *pegging* are made up of two parts:

  (i) The cost associated to waiting forever for the limit order which equals $\frac{\alpha}{\lambda}$.

  (ii) A discount (in this case the discount is of size $c_1 s^{p_1} < 0$) which is given, because there is a chance that the spread decreases to the optimal boundary *before* execution of the limit order. Note that the discount is decreasing in s, a fact which we will also see for other problems: clearly the chances of the spread hitting the optimal boundary decrease, when the initial spread increases. It is also interesting to see that in the strict pegging-region (i.e. $\mathfrak{s} > \beta$), the discount is strictly negative. This is due to the fact that the volatility of the spread σ is strictly positive: for any initial spread, there is a strictly positive probability that the optimal boundary gets hit before the limit order is executed.

- As soon as the spread drops below the critial level β, the micro-trader cancels the limit order and *crosses* the spread. The cost associated to *crossing* is just the spread $\mathfrak{s}$ at the time of crossing.

Deterministic spread

For the sake of completeness, we also analyse the deterministic version of the spread, where $\sigma = 0$. Even though the spread $\mathfrak{s}(t)$ evolves deterministically, the problem still involves a stochastic component: the execution of the limit order is determined by the Poisson process N. We have

Theorem 8.3. *Assume the 'killed' spread is given by* (8.10) *with* $\sigma = 0$.

For negative drift $r < 0$*, the optimal point is given by* $\beta = \frac{\alpha}{\lambda - r}$ *and the value function*

$$\mathcal{V}_1(s) = \begin{cases} s & \text{if } s < \beta \\ \frac{\alpha}{\lambda} + cs^{\lambda/r} & \text{if } s \geq \beta \end{cases} \tag{8.15}$$

where $c = \frac{r}{\lambda}\left(\frac{\alpha}{\lambda - r}\right)^{\frac{\lambda - r}{-r}} < 0$.

For positive or zero drift $r \geq 0$*, the optimal point is given by* $\beta = \frac{\alpha}{\lambda}$ *and the value function*

$$\mathcal{V}_1(s) = \begin{cases} s & \text{if } s < \beta \\ \frac{\alpha}{\lambda} & \text{if } s \geq \beta \end{cases} \tag{8.16}$$

The stopping time τ_β *defined in* (8.9) *is optimal.*

We recognize the structure described following theorem 8.2. Note that for the case $r < 0$ the discount is given by $cs^{\lambda/r}$, and thus strictly negative and decreasing in s as in the case of a stochastic spread. However, in the case $r \geq 0$, there is a zero discount; this is due to the fact that the optimal boundary is never reached by the spread, which is is deterministic and non-decreasing.

As expected, the solutions for the determinisic and stochastic spread are consistent: When we take the limit $\sigma \to 0$ in the value function given in (8.14), we recover the results of the deterministic case $\sigma = 0$ of theorem 8.3.

Note also that the problem scales in (α, r, λ): for any $\rho > 0$ the value function doesn't change when we replace (α, r, λ) by $(\rho\alpha, \rho r, \rho\lambda)$.

8.2.2 CIR process

Now we will analyse the optimal peg-cross strategy when the spread is modelled by a mean-reverting CIR process, i.e. the 'killed' spread s is given by

$$\mathrm{d}s(t) = \kappa(\mu - s(t))\mathrm{d}t + \sigma\sqrt{s(t)}\mathrm{d}W(t) - s(t)\mathrm{d}N(t) \tag{8.17}$$

Stochastic spread

Again we start by assuming $\sigma > 0$. As before, we will use the free-boundary problem (FB1)-(FB5) in order to derive a candidate value function v. Since the candidate value function must satisfy $v(0) = 0$, the differential equation (FB1) simplifies to

$$sv''(s) + (b - cs)v'(s) - av(s) + \frac{2\alpha}{\sigma^2} = 0 \tag{8.18}$$

where

$$a = \frac{2\lambda}{\sigma^2}$$
$$b = \frac{2\kappa\mu}{\sigma^2}$$
$$c = \frac{2\kappa}{\sigma^2}$$

Note that the constant function $\frac{\alpha}{\lambda}$ is a non-zero solution of (8.18).

By introducing $w(s) := v(s/c)$, the homogeneous part of (8.18) is recognized as Kummer's equation

$$sw''(s) + (b - s)w'(s) - \frac{a}{c}w(s) = 0$$

which has two independent solutions given by the confluent hypergeometric functions $w_1(s) = M\left(\frac{a}{c}, b, s\right)$ and $w_2(s) = U\left(\frac{a}{c}, b, s\right)$ (see Abramowitz and Stegun, 1964, chap. 13).

Thus (8.18) has the general solution

$$v(s) = \frac{\alpha}{\lambda} + c_1 U\left(\frac{a}{c}, b, cs\right) + c_2 M\left(\frac{a}{c}, b, cs\right)$$

where c_1, c_2 are two arbitrary constants.

As $s \to \infty$, we have the asymptotic behaviour

$$M\left(\frac{a}{c}, b, cs\right) = \frac{\Gamma(b)}{\Gamma(a/c)} e^{cs}(cs)^{a/c-b}(1 + O(s^{-1}))$$
$$U\left(\frac{a}{c}, b, cs\right) = (cs)^{-a/c}(1 + O(s^{-1})).$$

Thus the bounds $0 \le v(s) \le \min\left\{s, \frac{\alpha}{\lambda}\right\}$ imply that we must have $c_2 = 0$.

Now (FB2) and (FB3) can be used to determine the unknown constant c_1 and the free boundary β: β is defined by the equation $f(\beta) = 0$, where

$$f(s) = U\left(\frac{a}{c}, b, cs\right) - a\left(\frac{\alpha}{\lambda} - s\right) U\left(\frac{a}{c} + 1, b + 1, cs\right) \tag{8.19}$$

Note that by lemma 8.2, f admits a unique root in $(0, \alpha/\lambda)$. Moreover, we have

$$c_1 = \left(\beta - \frac{\alpha}{\lambda}\right) \frac{1}{U\left(\frac{a}{c}, b, c\beta\right)} \le 0 \tag{8.20}$$

Thus we obtained the candidate value function

$$v(s) = \begin{cases} s & \text{if } s < \beta \\ \frac{\alpha}{\lambda} + c_1 U\left(\frac{a}{c}, b, cs\right) & \text{if } s \ge \beta. \end{cases} \tag{8.21}$$

Theorem 8.4. *For $\sigma > 0$, the value function $\mathcal{V}_1$ is given by*

$$\mathcal{V}_1(s) = \begin{cases} s & \text{if } s < \beta \\ \frac{\alpha}{\lambda} + c_1 U\left(\frac{\lambda}{\kappa}, \frac{2\kappa\mu}{\sigma^2}, \frac{2\kappa}{\sigma^2}s\right) & \text{if } s \geq \beta \end{cases} \tag{8.22}$$

where β is defined as the unique root of f in $(0, \alpha/\lambda)$ and c_1 is given by (8.20). The stopping time τ_β is optimal.

Again we find the same solution structure as described in the discussion following theorem 8.2.

Deterministic spread

Let us now consider the case when the spread evolves deterministically, that is $\sigma = 0$. We have

Theorem 8.5. *Assume the 'killed' spread is given by (8.17) with $\sigma = 0$.*

For $\frac{\alpha}{\lambda} \geq \mu$, the optimal point is given by $\beta = \frac{\alpha+\kappa\mu}{\kappa+\lambda}$ and the value function

$$\mathcal{V}_1(s) = \begin{cases} s & \text{if } s < \beta \\ \frac{\alpha}{\lambda} + c(\kappa(s-\mu))^{-\lambda/\kappa} & \text{if } s \geq \beta \end{cases} \tag{8.23}$$

where $c = -\frac{1}{\lambda}\left(\frac{\kappa(\alpha-\lambda\mu)}{\kappa+\lambda}\right)^{\lambda/\kappa+1} < 0$.

In the case where $\frac{\alpha}{\lambda} < \mu$, the optimal point is given by $\beta = \frac{\alpha}{\lambda}$ and the value function

$$\mathcal{V}_1(s) = \begin{cases} s & \text{if } s \leq \beta \\ \frac{\alpha}{\lambda} & \text{if } s > \beta \end{cases} \tag{8.24}$$

The stopping time τ_β defined in (8.9) is optimal.

We find the same structure of the solution as in the case of the deterministic geometric Brownian motion. Again there are parameter constellations ($\alpha/\lambda < \mu$) where there is no discount, when the optimal strategy consists in pegging, as the optimal boundary is never reached by the spread.

Note that optimal boundary β is increasing in α and μ and decreasing in λ. This behaviour confirms intuitive ideas about the optimal strategy:

- as α increases, it is optimal to increase the optimal boundary, because the costs for holding the shares increase
- as μ increases, the average level of the spread increases, and thus the optimal boundary increases
- as λ increases, the execution of a limit order is more likely, so the lower bound that the spread has to cross to send a market order decreases

Moreover, when setting $\mu = 0$, we retrieve the results from theorem 8.3 with $r = -\kappa$.

Parameter	Estimate
κ	9.55×10^{-5}
μ	2.99×10^{-2}
σ_{CIR}	9.23×10^{-4}
r	-2.23×10^{-7}
σ_{GBM}	2.51×10^{-4}
λ	6.73×10^{-4}

Table 8.2: Model parameter estimates computed from order book data (AAPL, February 17, 2010, 10am-11am).

8.2.3 Calibration and test

To test our model, check the validity of the analytic results obtained, and to compare different choices for the spread process, we will again consider high-frequency order book data, as in chapter 5. Since we model the spread explicitly, and the optimal crossing boundary is formulated in terms of the spread, we will only consider AAPL data, where the spread varies between 1 and 6 ticks. In our modelling setting, it certainly does not make sense to look at stocks where the spread is constantly equal to one tick. Again we take data from 10am-11am on February 17, 2010, which is given in form of N 4-tuples

$$(t_i, \beta(t_i), \alpha(t_i), \mathfrak{z}(t_i))) \qquad i = 1, ..., N.$$

where $N = 17778$ and time is given in in milliseconds after midnight.

Using the same methods as in chapter 5, the parameters for the CIR process are estimated. The volatility σ of the geometric Brownian motion was also estimated using the method of *subsampling and averaging* explained in section 5.3. Once we obtained an estimate $\hat{\sigma}$ of σ, we estimate the drift r by taking the overall drift of the entire sample:

$$\hat{r} = \frac{\log(\mathfrak{s}(t_N)) - \log(\mathfrak{s}(t_1))}{t_N - t_1} + \frac{1}{2}\hat{\sigma}^2$$

where $\mathfrak{s}$ was of course defined by

$$\mathfrak{s}(t_i) = \alpha(t_i) - \beta(t_i) \qquad i = 1, ..., N.$$

We will again identify execution times of limit orders with first passage times of best bid and best ask. We want to buy shares and thus we are interested in first passage times of the best bid. We therefore define every time point t_i with $\beta(t_i) < \beta(t_{i-1})$ to be a time of limit order execution. Using the maximum likelihood estimator for an exponential distribution, we get an estimate for the limit order rate λ.

Table 8.2 shows the estimates for the model parameters for both the CIR process and geometric Brownian motion. To distinguish the volatility term σ in both models, we add the corresponding subscript. Note that (not surprisingly) the drift of the geometric Brownian motion is almost equal to zero.

Next, we choose values for the parameter α which determines the degree of penalization for holding sharess and we compute the corresponding optimal crossing boundaries for the geometric

α	**Optimal boundary GBM**	**Optimal boundary CIR**
8×10^{-5}	$1,1$ ticks	$1,0$ ticks
16×10^{-5}	$2,2$ ticks	$2,0$ ticks
24×10^{-5}	$3,2$ ticks	$3,1$ ticks
31×10^{-5}	$4,2$ ticks	$4,0$ ticks

Table 8.3: Optimal boundaries for different values of α computed from order book data (AAPL, February 17, 2010, 10am-11am).

Brownian motion (GBM) and CIR model, using the formulae from theorems 8.2 and 8.4. The results are displayed in table 8.3. Note that the tick size is 1 cent.

When the spread $\mathfrak{s}$ is below or equal to the optimal boundary displayed in table 8.3, theorems 8.2 and 8.4 say that it is optimal to cross the spread and cancel the outstanding limit order.

Finally, we will see how the theoretical optimal boundaries perform when tested against real order book data. To carry out this test, we test how the optimal boundaries computed from the 10am-11am data perform, when tested with orderbook data of the same stock, one hour later. For this delayed data, we compute the average costs for six different strategies associated to buying one single share:

'Cross' The costs associated to crossing immediately.

'n ticks' The costs associated to pegging on the best bid until the limit order is executed and crossing as soon as the spread size is $\leq n$ ticks, for $n = 1, ..., 5$.

'Peg' The costs associated to pegging on the best bid until the limit order is executed.

Figure 8.1 displays the costs associated to the six strategies for a risk-neutral trader (i.e. with $\alpha = 0$). Since we normalized the price of the best bid to zero in our model, we will compute the minimum costs of the six strategies from the data and set it to zero. The bars show the average additional cost in US dollar of buying one share using the corresponding strategy, compared to the strategy with minimum costs. For the risk-neutral trader, we see that the minimum cost strategy is pegging on the best bid. This corresponds to our analytical result: when there is no penalization for holding the shares, the best strategy is 'Peg', i.e. the trader waits until the limit order is executed, irrespective of the current spread. The second best strategy is '1 tick', the third-best '2 ticks', and so on, and finally 'Cross' has the highest additional costs compared to 'Peg'.

Figure 8.2 displays the costs associated to the six strategies for different risk-averse traders (i.e. with different levels of risk-aversion α). We see that for $\alpha = 8 \times 10^{-5}$, the optimal strategy is '1 tick', which corresponds to our analytic results, for both GBM and CIR process. Similary, for $\alpha = 16 \times 10^{-5}$, the optimal strategy obtained from analytic and empirical results match and consist in crossing as soon as the spread is lower or equal to 2 ticks. For $\alpha = 24 \times 10^{-5}$, however, the empirical result suggests an optimal crossing boundary of 2 ticks, whereas the analytical result in this section suggests that it should be equal to 3 ticks. Yet, the additional costs of following strategy '3 ticks' instead of the optimal empirical strategy are rather low. Similary, for $\alpha = 24 \times 10^{-5}$ the optimal empirical and analytical strategies do not correspond ('3 ticks' vs. '4 ticks'), but the corresponding additional cost is rather low.

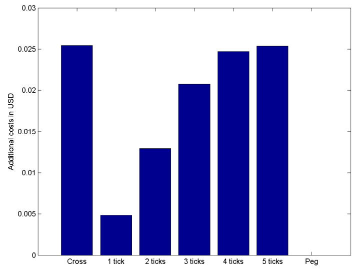


Figure 8.1: Costs associated to different strategies for risk-neutral trader ($\alpha = 0$) computed from order book data (AAPL, February 17, 2010, 11am-12am).

Comparing the optimal boundaries of GBM and CIR model in table 8.3, we note that - for this special case of parameters - there is not much difference between the two models: both models have roughly the same optimal boundaries, with the GBM model giving slightly higher optimal boundaries than the CIR model. Comparing this to the empiricial results in figure 8.2, we see that the CIR model performs slightly better in practice, but the difference is negligible.

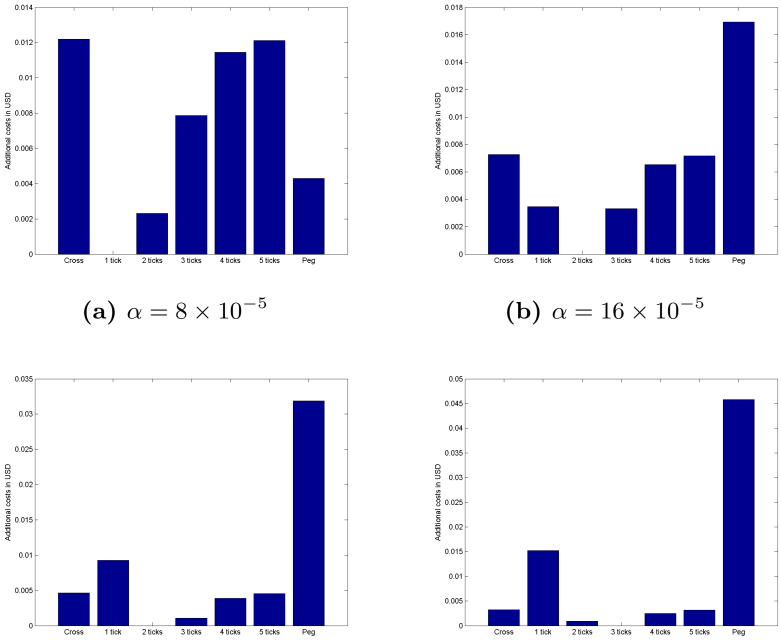
The mean benefit of using the best versus the second best strategy is in the range of 1/10 cent to 5/10 cent. However, one needs to keep in mind that this applies each time a single share is bought. For a large trader engaged in high frequency trading, these fractions of a cent may accumulate to a considerable amount.

Even though the theoretical analysis of the model does not correspond exactly to the optimal strategies found empirically, it is surprising that such a simple model performs so well. Clearly, our model has a lot of shortcomings:

- We model the spread as a diffusion process, however, in reality it is a discrete process, due to the discrete tick size. This leads to results which are hard to interpret, such as an optimal crossing barrier of 3.2 ticks (see table 8.3).
- The time-to-fill is modelled as a simple Poisson process. However, as we saw in chapter 5, it depends heavily on spread and volume-imbalance.
- The best bid is assumed to be constant. A more realistic model would reproduce volatility and drift of the best bid, which will influence the price of the limit order placed on the best bid.

These obvious shortcomings had to be introduced for the sake of mathematical tractability. Despite all this, our the analytical results are very close to the empirical observations.

For practical purposes, I would propose a simple finite-state (or denumerable-state) Markov chain. The state space should at least include best quotes and volume on the best quotes. This model then automatically reflects the discrete tick stucture. Moreover, transition probabilities can easily be estimated from historical data, and will reproduce both movements of best bid/ ask as well as state-dependent time-to-fill. This framework can also easily be extended to include queuing priorities of limit orders. Using a standard dynamic programming algorithm, one can then numerically solve the peg/cross problem with a finite time horizon.



(a) $\alpha = 8 \times 10^{-5}$

(b) $\alpha = 16 \times 10^{-5}$

(c) $\alpha = 24 \times 10^{-5}$

(d) $\alpha = 31 \times 10^{-5}$

Figure 8.2: Costs associated to different strategies for risk-averse trader, with varying level of risk-aversion computed from order book data (AAPL, February 17, 2010, 11am-12am).

8.3 Optimal strategies with market impact

For some markets, assumption 8.1 is not valid. This is the case when the typical package size is larger than the minimal trading size or when markets are very illiquid. In this situation, the price of a market order can considerably exceed the price of the best ask. In this section we will therefore consider $\mu > 0$.

To avoid case distinctions we moreover assume that we work in the non-trivial case of a non-deterministic spread ($\sigma \neq 0$) and strictly positive probability that limit orders are executed ($\lambda > 0$).

As before, we will work under assumption 8.2 of an infinite time horizon. As an immediate consequence we can give a first partial characterization of the optimal crossing time:

Proposition 8.1. *Suppose simplifying assumption 8.2 holds and that $\mu\lambda \geq \alpha$. Then*

$$\mathcal{V}(x, s) = \frac{\alpha}{\lambda} x^2$$

and we have the optimal crossing time $\hat{\tau} = \infty$ a.s.

In markets in which the relation $\mu\lambda \geq \alpha$ holds the optimal strategy consists in always trading by limit orders. There are three reasons why it is favourable not to use market orders in such situations at all:

Market order illiquidity Only few limit orders are stored in the book. Hence buying a large amount of market orders in a single block trade immediately drives the price up. The extra costs incurred by the market order impact are very large compared to the pegging strategy, hence it is better to wait and peg on the best bid.

Limit order liquidity Limit orders are executed frequently. Even when the spread is small, the microtrader will not cross it, as the chances of an imminent limit order execution are high.

Low risk-aversion The microtrader's risk-aversion of holding shares is low. Hence it will always prefer the more risky limit order (the risk of the limit order is the risk of non-execution) to the less risky (immediate execution) but more costly market order.

Note that the condition

$$\lambda \geq \frac{\alpha}{\mu} = \mathfrak{R}_m$$

again involves just the key market parameters we identified in chapter 7. To get an explicit representation of $\mathcal{V}$ for more parameter values, we introduce

Simplifying assumption 8.3. We assume that the spread is a geometric Brownian motion, i.e. $r(t, \mathfrak{s}) \equiv r\mathfrak{s}$ and $\sigma(t, \mathfrak{s}) = \sigma\mathfrak{s}$ in (8.1) for constants $r \in \mathbb{R}$, $\sigma > 0$.

The choice of a geometric Brownian motion for the spread ensures that spread stays positive for all $t \geq 0$. However it does not include the desirable mean-reversion property of the CIR process, and for $r \neq 0$ the mean of the spread converges either to zero or $+\infty$ as $t \to \infty$. However, we need the specific structure of the geometric Brownian motion as it allows us to reduce the number of dimensions in the problem.

In the following analysis, we will compute an explicit expression for $\mathcal{V}$ and identify the optimal crossing time τ. We shall proceed in four steps: first, we heuristically derive properties that the candidate value funtion $v(x, \mathfrak{s})$ should satisfy. Secondly, we transform the problem of determining $v(x, \mathfrak{s})$ into an equivalent problem of determining a function $w(z)$ where $z = x/\mathfrak{s}$. This will reduce the number of dimensions from two to one. Thirdly, we will guess a candidate solution for w and thus (by the inverse transform) for v. Finally, we will use a verification argument to show that the candidate solution v equals the actual value function $\mathcal{V}$.

Intuitively, the micro-trader will submit a market order when x is large (to reduce high costs of holding the assets) and when $\mathfrak{s}$ is small (because in this case the additional cost of crossing the spread are not very high). This suggests that there exists a constant boundary $b \geq 0$ such that the stopping time

$$\tau_b = \inf\left\{t \geq 0 : \frac{X(0)}{\mathfrak{s}(t)} \geq b\right\} \tag{8.25}$$

is optimal.

Heuristic arguments then suggest that our candidate value function $v(x, \mathfrak{s})$ solves the free-boundary problem (V1)-(V5)

$$\mathcal{L}_v v(x,s) + \alpha x^2 = 0, \quad x/s < b \tag{V1}$$
$$v(x,s) = xs + \mu x^2, \quad x/s = b \tag{V2}$$
$$v_X(x,s) = s + 2\mu x, \quad x/s = b \tag{V3a}$$
$$v_s(x,s) = x, \quad x/s = b \tag{V3b}$$
$$v(x,s) < xs + \mu x^2, \quad x/s < b \tag{V4}$$
$$v(x,s) = xs + \mu x^2, \quad x/s > b \tag{V5}$$

where $\mathcal{L}_v$ is the infinitisimal generator of the two-dimensional process $(X(t), \mathfrak{s}(t))$ defined by

$$(\mathcal{L}_v\varphi)(x,s) = rs\varphi_s(x,s) + \frac{\sigma^2 s^2}{2}\varphi_{ss}(x,s) + \lambda\left\{\varphi(0,s) - \varphi(x,s)\right\} \tag{8.26}$$

Now observe that by linearity of the processes $X(t)$ and $\mathfrak{s}(t)$, we have the following homotheticity property: For $\rho > 0$

$$\mathcal{J}(\tau, \rho x_0, \rho\mathfrak{s}_0) = \rho^2 \mathcal{J}(\tau, x_0, \mathfrak{s}_0)$$

and thus

$$\mathcal{V}(\rho x_0, \rho\mathfrak{s}_0) = \rho^2 \mathcal{V}(x_0, \mathfrak{s}_0)$$

Define for $z \geq 0$

$$w(z) = v(z, 1) \tag{8.27}$$

so by the homotheticity property, we have

$$v(x,s) = v(s\frac{x}{s}, s) = s^2 v(\frac{x}{s}, 1) = s^2 w(\frac{x}{s})$$

Now we can rewrite (V1) - (V5) in terms of w and obtain

$$\mathcal{L}_w w(z) + \alpha = 0, \quad z < b \tag{W1}$$
$$w(z) = z + \mu z^2, \quad z = b \tag{W2}$$
$$w'(z) = 1 + 2\mu z^2, \quad z = b \tag{W3}$$
$$w(z) < z + \mu z^2, \quad z < b \tag{W4}$$
$$w(z) = z + \mu z^2, \quad z > b \tag{W5}$$

where $\mathcal{L}_w$ is the differential operator

$$(\mathcal{L}_w\varphi)(z) = \frac{\lambda\varphi(0)}{z^2} + (2r - \lambda + \sigma^2)\frac{\varphi(z)}{z^2} - (r + \sigma^2)\frac{\varphi'(z)}{z} + \frac{\sigma^2}{2}\varphi''(z) \tag{8.28}$$

Since $\mathcal{V}(0, \mathfrak{s}) = 0$, our candidate value function w also needs to satisfy $w(0) = 0$. Then (W1) is an inhomogenenous Cauchy-Euler equation. It can easily be checked that $w_0(z) = \frac{\alpha}{\lambda}z^2$ is a particular non-zero solution of (W1). For the homogeneous ODE (that is, we set $\alpha = 0$ in (W1)), we seek a solution of the form $w(z) = z^p$ which leads to the general solution of (W1)

$$w(z) = \frac{\alpha}{\lambda}z^2 + c_1 z^{p_1} + c_2 z^{p_2} \tag{8.29}$$

where c_1, c_2 are arbitrary constants and

$$p_1 = \frac{3}{2} + \frac{r}{\sigma^2} + \frac{1}{2\sigma^2}\sqrt{(2r-\sigma^2)^2 + 8\lambda\sigma^2}$$
$$p_2 = \frac{3}{2} + \frac{r}{\sigma^2} - \frac{1}{2\sigma^2}\sqrt{(2r-\sigma^2)^2 + 8\lambda\sigma^2}$$

To solve the free boundary problem (W1)-(W5), we need to determine three unknowns c_1, c_2 and b, but we only have two equations (W2) and (W3). However, for $r < \lambda$, we have $p_2 < 1$, thus by taking the limit $z \to 0$ and inequality (W4), we must have $c_2 = 0$. In that case we can solve for c_1 and b and obtain

$$b = \begin{cases} \frac{\lambda}{\alpha-\lambda\mu}\frac{p_1-1}{p_1-2} & \text{if } \alpha > \lambda\mu \\ \infty & \text{if } \alpha \le \lambda\mu \end{cases} \tag{8.30}$$

$$c_1 = \begin{cases} b^{(2-p_1)}\left(\mu - \frac{\alpha}{\lambda}\right) + b^{(1-p_1)} & \text{if } \alpha > \lambda\mu \\ 0 & \text{if } \alpha \le \lambda\mu \end{cases} \tag{8.31}$$

Note that for $\sigma > 0$, $\lambda > 0$, we always have $p_1 > 2$, so b is well-defined in (8.30). We thus have a candidate solution

$$w(z) = \begin{cases} \frac{\alpha}{\lambda}z^2 + c_1 z^{p_1} & \text{if } z < b \\ \mu z^2 + z & \text{if } z \ge b \end{cases} \tag{8.32}$$

which is in line with proprosition 8.1 for $\alpha \le \lambda\mu$ by defintion of b and c_1.

We can now formulate

Theorem 8.6 (Optimal peg-cross strategy)**.** *Under simplifying assumptions 8.2 and 8.3 and for $\lambda > 0$ and $\sigma > 0$, the value function $\mathcal{V}$ is given in explicit form by*

$$\mathcal{V}(x, \mathfrak{s}) = \begin{cases} \frac{\alpha}{\lambda}x^2 + c_1\mathfrak{s}^2\left(\frac{x}{\mathfrak{s}}\right)^{p_1} & \textit{if } \frac{x}{\mathfrak{s}} < b \\ \mu x^2 + x\mathfrak{s} & \textit{if } \frac{x}{\mathfrak{s}} \ge b \end{cases} \tag{8.33}$$

The stopping time τ_b defined in (8.25) *with b from* (8.30) *is optimal.*

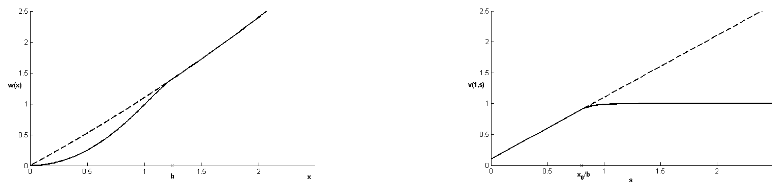
Figure 8.3 shows the value function for parameters $r = 0$, $\sigma = 0.05$, $\mu = 0.1$, $\lambda = 0.1$ and $\alpha = 0.1$.

The optimal peg-cross strategy can be summarized as follows:

- While the spread $\mathfrak{s}(t)$ is above the critial level x_0/b, it is too expensive to cross the spread, and the micro-trader *pegs* on the best bid and waits (until either the limit order is executed or the spread drops to a level where it is favouralbe to cross).
- As soon as the spread drops below the critial level x_0/b, the micro-trader cancels the limit order and *crosses* the spread.

Figure 8.4 illustrates the optimal strategy for two sample paths, using the same parameters as in figure 8.3.

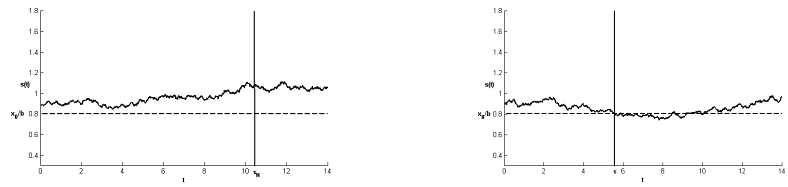
While the micro-trader is pegging (that is for $\mathfrak{s}(t) > x_0/b$), the average cost of the optimal strategy is $\frac{\alpha}{\lambda}x_0^2 + c_1\mathfrak{s}(t)^2\left(\frac{x_0}{\mathfrak{s}(t)}\right)^{p_1}$. The first term $\frac{\alpha}{\lambda}x_0^2$ is the expected cost of using the strategy



(a) Value function as a function of share holdings with constant spread

(b) Value function as a function of spread with constant share holdings

Figure 8.3: Value functions of the peg-cross problem. (a) shows the function $w(x) = v(x, 1)$ (solid line), where x denote the share holdings. Note the kink at $x = b$. For $x < b$, it is advantageous not to cross immediately and either wait for the limit order or a more favouralbe spread. For $x \geq b$ it is optimal to cross the spread immediately. For $x < b$, the dashed line shows the costs associated to the naive strategy of crossing at time 0. (b) shows the function $v(1, s)$ (solid line), where s denotes the size of the spread. For $s \leq x_0/b$ the optimal strategy is crossing immediately and $s \to v(1, s)$ is affine. For $s > x_0/b$ it is optimal to peg. The dashed line shows the costs associated to the naive strategy of crossing immediately.



(a) Peg

(b) Cross

Figure 8.4: Optimal strategies for two different paths of the spread: the evolution of the spread is plotted on the y-axis, time on the x-axis. In (a) the limit order is executed at time τ_N and the spread stays above the critical boundary x_0/b on $[0, \tau_N]$. The optimal strategy consists in pegging on the best ask and waiting for the execution of the limit order. In (b), the spread crosses the critical boundary x_0/b at time τ, and the limit order was not executed before. The optimal strategy consists in cancelling the limit order at time τ and converting it into a market order.

to wait forever for the limit order (and never cross the spread). As the micro-trader has the opportunity to cross the spread at any stopping time τ, it gets a discount of $c_1 \mathfrak{s}(t)^2 \left(\frac{x_0}{\mathfrak{s}(t)}\right)^{p_1}$ (note that $c_1 < 0$).

Note also that by setting $x_0 = 1$ and $\mu = 0$, we recover the optimal boundary β from theorem 8.2:

$$\beta = \frac{1}{b}_{|\mu=0}$$

Proposition 8.2. *Suppose the assumptions of theorem 8.6 are satisfied, and moreover $\alpha > \lambda\mu$ holds. The optimal crossing boundary $\frac{x_0}{b}$ is*

- *increasing in α and r*
- *decreasing in μ, λ and σ.*

The above proposition confirms intuitive ideas about the optimal strategy:

- as the risk-aversion α increases, pegging gets more expensive and one should cross earlier

- as the drift of the spread r increases, the tendency of the spread to widen increases, thus crossing after a certain waiting time gets more expensive, and one should cross earlier
- as the market order impact μ increases, crossing gets more expensive, and one should wait longer for the limit order
- as the intensity of limit orders λ increases, the chances of execution by a limit order increase, and one should wait longer for a limit order
- as the volatility of the spread σ increases, the chances that a small spread is attained (and thus that a cheap crossing in the future is possible) increase, so one should peg and wait (for the limit order or the small spread)

Higher layers in the algorithmic trading strategy determine strategies that the micro-trader uses as input. One important parameter for the higher layer algorithms is the probability that a unit order executed by the micro-trader is executed by a limit order (or equivalently, the probability that it is executed by a market order). The following corollary gives a formula for this probability when the micro-trader uses the optimal peg-cross strategy.

Corollary 8.1. *Suppose the assumptions of theorem 8.6 are satisfied. Let C be the event that the package is bought by crossing the spread when using the optimal strategy. Then*

$$\mathrm{P}(C) = \begin{cases} 1 & \text{if } \alpha > \lambda\mu \text{ and } \mathfrak{s}_0 \leq \frac{x_0}{b} \\ \exp\left\{z(\rho + \sqrt{2\lambda + \rho^2})\right\} & \text{if } \alpha > \lambda\mu \text{ and } \mathfrak{s}_0 > \frac{x_0}{b} \\ 0 & \text{if } \alpha \leq \lambda\mu \end{cases} \tag{8.34}$$

where $z = \frac{1}{\sigma}\log(\frac{x_0}{b\mathfrak{s}_0})$ *and* $\rho = \frac{2r-\sigma^2}{2\sigma}$

Note that for $\alpha > \lambda\mu$ and $\mathfrak{s}_0 > \frac{x_0}{b}$ we have $z < 0$, so $\mathrm{P}(C) \in [0,1]$.

8.4 Proofs

Proof of Theorem 8.1. Enlarge the state space by setting $Z(t) = (t, X(t), \mathfrak{s}(t), I(t))$ where $I(t) := \int_0^t X(u)\mathrm{d}u$. Then $\mathcal{J}(\tau; t, x_t, \mathfrak{s}_t) = \mathrm{E}[G(Z(t))]$ where $G((t,x,s,a)) = xs + \mu x^2 + \alpha a$ is continuous. Therefore by Peskir and Shiryaev (2006, Cor. 2.9 and Rem 2.10), an optimal $\hat{\tau}$ exists. □

Proof of Theorem 8.2. Define the candidate function

$$v(s) = \begin{cases} s & \text{if } s < \beta \\ \frac{\alpha}{\lambda} + c_1 s^{p_1} & \text{if } s \geq \beta \end{cases}$$

We want to show that $v(s) = \mathcal{V}_1(s)$. Note that v is C^2 on $[0,\beta)$ and on (β,∞), but only C^1 at β, so we cannot apply the standard Itô formula. However, we can use the *change-of-variable formula* from Peskir (2007) and obtain

$$v(s(t)) = v(s_0) + \int_0^t \mathcal{L}_s v(s(u))\mathrm{d}u + M_1(t) - M_2(t)$$

where $M_1(t) = \sigma \int_0^t v'(s(u))s(u)\mathrm{d}W(u)$, $M_2(t) = \int_0^t v(s(u))\mathrm{d}\tilde{N}(u)$ and $\tilde{N}(u) = N(u) - \lambda u$.

We have $(\mathcal{L}_s v)(s) + \alpha \geq 0$ for all $s \geq 0$: for $s \geq \beta$ this is trivially true by (FB1), and for $s \in [0, \beta)$ we need to show $s(r - \lambda) + \alpha \geq 0$. In the case $r \geq \lambda$ the inequality is obvious. For $r < \lambda$, it is enough to show $\alpha \geq (\lambda - r)\beta$ which simplifies to $\frac{\lambda(p_1-1)}{(\lambda-r)p_1} \geq 0$, which is true because $p_1 < 0$.

Moreover, using elementary analysis, one can show that $v(s) \leq s$ and $v'(s) \leq 1$ for all $s \geq 0$.

Hence we obtain

$$s(t) + \alpha t \geq v(s_0) + M_1(t) - M_2(t).$$

We note that M_1 and M_2 are true martingales, since

$$\begin{aligned} \mathrm{E}[[M_1](t)] &= \sigma^2 \mathrm{E}[\int_0^t (v'(s(u))s(u))^2 \mathrm{d}u] \\ &\leq \sigma^2 \mathrm{E}[\int_0^t s^2(u)\mathrm{d}u] \\ &= \sigma^2 \int_0^t \mathrm{E}[s^2(u)]\mathrm{d}u < \infty \end{aligned}$$

where we used the bound $v'(s) \leq 1$, Fubinis theorem and the explicit formula for the moments of a geometric Brownian motion. Thus M_1 is a true (square-integrable) martingale. Similary, one can show that M_2 is a true martingale.

We will start by showing that

$$\mathcal{V}_1(s_0) \geq v(s_0). \tag{8.35}$$

To prove (8.35), it suffices to prove that for all stopping times $\tau \in \mathcal{T}(0, \infty)$ we have

$$\mathcal{J}_1(\tau; s_0) \geq v(s_0).$$

By lemma 8.1, the martingales $\tilde{M}_i$ defined by

$$\tilde{M}_i(t) := M_i(t \wedge \tau_N)$$

are uniformly integrable. Let now τ be any stopping time in $\mathcal{T}(0, \infty)$. Then, by the optional sampling theorem

$$\mathrm{E}[s(\tau \wedge \tau_N) + \alpha(\tau \wedge \tau_N)] \geq v(s_0) \tag{8.36}$$

which implies that $v(s_0) \leq \mathcal{V}_1(s_0)$.

To prove equality, consider the candidate optimal stopping time τ_β. Using the fact that v solves (FB1) and similar arguments as above it is straightforward to show that

$$\mathrm{E}[s(\tau_\beta \wedge \tau_N) + \alpha(\tau_\beta \wedge \tau_N)] = v(s_0) \tag{8.37}$$

and thus $v(s_0) = \mathcal{V}_1(s_0)$. □

Lemma 8.1. *Consider the local martingales* $M_1(t) = \sigma \int_0^t v'(s(u))s(u)\mathrm{d}W(u)$ *and* $M_2(t) = \int_0^t v(s(u))\mathrm{d}\tilde{N}(u)$. *Then for* $i = 1, 2$, *the processes* $t \to \tilde{M}_i(t) := M_i(t \wedge \tau_N)$ *are uniformly integrable martingales.*

Proof. To show that $\tilde{M}_1$ is uniformly integrable (u.i.) it is enough to show (e.g. Protter, 2004, Thm. 11 with $G(x) = x^2$) that

$$\sup_{t\geq 0} \mathrm{E}[\tilde{M}_1^2(t)] < \infty$$

Note that

$$sv'(s) = \begin{cases} s & \text{if } s < \beta \\ c_1 p_1 s^{p_1} & \text{if } s \geq \beta \end{cases}$$

As $p_1 < 0$, we have $|sv'(s)| \leq \max\{\beta, |c_1 p_1|\beta^{p_1}\}$, we compute the bound

$$\begin{aligned} \mathrm{E}\left[[\tilde{M}_1](t)\right] &\leq \mathrm{E}\left[[\tilde{M}_1](\infty)\right] \\ &= \mathrm{E}\left[\sigma^2 \int_0^{\tau_N} \left(v'(s(u))s(u)\right)^2 \mathrm{d}u\right] \\ &\leq \sigma^2 \max\{\beta, |c_1 p_1|\beta^{p_1}\}^2\, \mathrm{E}[\tau_N] \\ &\leq \frac{\sigma^2 \max\{\beta, |c_1 p_1|\beta^{p_1}\}^2}{\lambda} =: K \end{aligned}$$

Thus $\sup_{t\geq 0} \mathrm{E}[\tilde{M}_1^2(t)] \leq K < \infty$ and $\tilde{M}_1$ is a u.i. martingale. A similar argument, using $v(s) \leq \frac{\alpha}{\lambda}$ shows that $\tilde{M}_2$ is also a u.i. martingale. □

Proof of Theorem 8.3. As the candidate value function must satisfy $v(0) = 0$, we can solve (FB1), and use equations (FB2) and (FB3) to obtain c, β and the candidate value function

$$v(s) = \begin{cases} s & \text{if } s < \beta \\ \frac{\alpha}{\lambda} + cs^{\lambda/r} & \text{if } s \geq \beta \end{cases}$$

For $r < 0$, it can easily be verified, using the same methods as in theorem 8.2, that the candidate is indeed optimal.

For $r > 0$, however the candidate solution contradicts the bound $\mathcal{V}_1(s) \leq \frac{\alpha}{\lambda}$. Since the spread is deterministic we can apply a direct method to verify $\mathcal{V}_1$ for $r \geq 0$: for any $t \in [0, \infty)$, we calculate

$$\begin{aligned} \mathcal{J}_1(t; s_0) &= \mathrm{E}[(s_0 e^{rt} + \alpha t)1_{\{\tau_N > t\}}] + \mathrm{E}[\alpha \tau_N 1_{\{\tau_N \leq t\}}] && (8.38) \\ &= s_0 e^{(r-\lambda)t} + \frac{\alpha}{\lambda}\left(1 - e^{-\lambda t}\right) && (8.39) \end{aligned}$$

and

$$\mathcal{J}_1(\infty; s_0) = \frac{\alpha}{\lambda}$$

In the case $s_0 < \frac{\alpha}{\lambda}$, we have thus have $\mathcal{J}_1(0; s_0) \leq \mathcal{J}_1(t; s_0)$ for all $t \in [0, \infty]$, so it is optimal to cross at time 0. In the case $s_0 \geq \frac{\alpha}{\lambda}$, we have thus have $\mathcal{J}_1(\infty; s_0) \leq \mathcal{J}_1(t; s_0)$ for all $t \in [0, \infty]$, and hence it is optimal to peg forever. Noting that all stopping times in $\mathcal{T}(0, \infty)$ are deterministic, we obtain the result for $r \geq 0$, by combining the two inequalities. □

Proof of Theorem 8.4. We want to show that $v(s) = \mathcal{V}_1(s)$. Let $\tilde{N}(u) = N(u) - \lambda u$ be the compensated Poisson process. We use again the *change-of-variable formula* from Peskir (2007) and obtain

$$v(s(t)) = v(s_0) + \int_0^t \mathcal{L}_s v(s(u))\mathrm{d}u + M_1(t) - M_2(t)$$

where $M_1(t) = \sigma \int_0^t v'(s(u))\sqrt{s(u)}\mathrm{d}W(u)$, $M_2(t) = \int_0^t v(s(u))\mathrm{d}\tilde{N}(u)$.

Note that we have $(\mathcal{L}v)(s) + \alpha \geq 0$ for all $s \geq 0$: for $s \geq \beta$ it is true by construction of v, and for $s \in [0, \beta)$ it is equivalent to showing

$$\frac{\kappa\mu + \alpha}{\kappa + \lambda} \geq s, \quad \forall s \in [0, \beta)$$

which is implied by

$$\frac{\kappa\mu + \alpha}{\kappa + \lambda} \geq \beta.$$

This is true by lemma 8.2.

Next, we show that $v(s) \leq s$ and $v'(s) \leq 1$ for all $s \geq 0$: for $s < \beta$ this is trivial, for $s \geq \beta$ define $g(s) := s - v(s)$. As $g(\beta) = 0$, it suffices to show that $g'(s) \geq 0$. Using the definition of c_1, we compute

$$g'(s) = 1 - v'(s) = \frac{U\left(\frac{a}{c} + 1, b + 1, cs\right)}{U\left(\frac{a}{c} + 1, b + 1, c\beta\right)} \geq 0$$

for $s \geq \beta$.

Hence we obtain

$$s(t) + \alpha t \geq v(s_0) + M_1(t) - M_2(t)$$

Using the bounds $v(s) \leq s$ and $v'(s) \leq 1$ and Fubini's theorem, it is straightforward to show that M_1 and M_2 are true martingales, using the criterion that $\mathrm{E}[[M_i](t)] < \infty$ for all $t \geq 0$ and $i = 1, 2$. Let now τ be any stopping time in $\mathcal{T}(0, \infty)$. Then, by the optional sampling theorem

$$\mathrm{E}[s(\tau \wedge \tau_N \wedge n) + \alpha(\tau \wedge \tau_N \wedge n)] \geq v(s_0)$$

We want to take the limit $n \to \infty$. We clearly have a.s. convergence: $\lim_{n\to\infty} s(\tau\wedge\tau_N\wedge n)+\alpha(\tau\wedge\tau_N\wedge n) = s(\tau\wedge\tau_N)+\alpha(\tau\wedge\tau_N)$, because $\tau_N < \infty$ a.s. for $\lambda > 0$. By lemma 8.3, $\sup_n s(\tau\wedge\tau_N\wedge n)$ is integrable, so we can use the dominated convergence theorem. For the second term, we simply use the monotone convergence theorem to show that the integral converges. As we take the limit $n \to \infty$, we obtain

$$\mathrm{E}[s(\tau \wedge \tau_N) + \alpha(\tau \wedge \tau_N)] \geq v(s_0) \tag{8.40}$$

which implies that $v(s_0) \leq \mathcal{V}_1(s_0)$.

To prove equality, consider the candidate optimal stopping time τ_β. In this case, using (FB1) and similar arguments as above, it is easy to show that

$$\mathrm{E}[s(\tau_\beta \wedge \tau_N) + \alpha(\tau_\beta \wedge \tau_N)] = v(s_0) \tag{8.41}$$

and thus we have $v(s_0) = \mathcal{V}_1(s_0)$. $\square$

Lemma 8.2. *Let f be defined as in* (8.19)*. Then the equation $f(s) = 0$ has a unique solution $\beta \in (0, \alpha/\lambda)$. Moreover β satisfies*

$$\beta \leq \frac{\kappa\mu + \alpha}{\kappa + \lambda} \tag{8.42}$$

Proof. A straightforward computation gives

$$f'(s) = \frac{4(\kappa+\lambda)(\alpha - s\lambda)}{\sigma^4} U\left(\frac{\lambda}{\kappa}+2, \frac{2\kappa\mu}{\sigma^2}+2, \frac{2\kappa}{\sigma^2}s\right).$$

Hence f is strictly increasing on $(0, \alpha/\lambda)$. Moreover $f(\alpha/\lambda) = U\left(\frac{a}{c}, b, c\alpha/\lambda\right) > 0$, and using the asymptotic behaviour in Abramowitz and Stegun (1964, chap. 13, 13.5.6 and following) we have $\lim_{s\to 0} f(s) = -\infty$. Therefore $f(s)$ has a unique root in $(0, \alpha/\lambda)$, which we denote by β. To show the bound (8.42), it suffices to show that $f\left(\frac{\kappa\mu+\alpha}{\kappa+\lambda}\right) \geq 0$, as $f(s) < 0$ on $(0, \beta)$. We now use the equality

$$U(A, B, z) + (B - z)U(A+1, B+1, z) = -zU'(A+1, B+1, z)$$

from Abramowitz and Stegun (1964, chap. 13, 13.4.27), with $z = \frac{2\kappa(\alpha+\kappa\mu)}{(\lambda+\kappa)\sigma^2}$, $A = a/c$ and $B = b$ to compute

$$\begin{aligned} f\left(\frac{\kappa\mu+\alpha}{\kappa+\lambda}\right) &= -\frac{2\kappa(\alpha+\kappa\mu)}{(\kappa+\lambda)\sigma^2} U'\left(\frac{\lambda}{\kappa}+1, \frac{2\kappa\mu}{\sigma^2}+1, \frac{2\kappa(\alpha+\kappa\mu)}{(\lambda+\kappa)\sigma^2}\right) \\ &= \frac{2\kappa(\alpha+\kappa\mu)}{(\kappa+\lambda)\sigma^2}\left(\frac{\alpha}{\lambda}+1\right) U\left(\frac{\lambda}{\kappa}+2, \frac{2\kappa\mu}{\sigma^2}+2, \frac{2\kappa(\alpha+\kappa\mu)}{(\lambda+\kappa)\sigma^2}\right) \\ &\geq 0 \end{aligned}$$

which completes the proof. □

Lemma 8.3. *The supremum of s until the first jump time of N is integrable:*

$$\mathrm{E}\left[\sup_{0\leq u\leq \tau_N} s(u)\right] < \infty$$

Proof. By definition of the 'killed' spread process s in (8.17) we have the bound

$$s(t) \leq \kappa\mu t + \sigma M(t)$$

where $M(t) = \int_0^t \sqrt{\mathfrak{s}(u)}\mathrm{d}W(u)$. By the Burkholder-Davis-Gundy inequality there exists a constant $C > 0$ such that

$$\begin{aligned} \mathrm{E}\left[\sup_{u\in[0,t]} |M(u)|\right] &\leq C\mathrm{E}\left[[M]^{1/2}(t)\right] \\ &= C\mathrm{E}\left[\left(\int_0^t \mathfrak{s}(u)\mathrm{d}u\right)^{1/2}\right] \\ &\leq C\mathrm{E}\left[1 + \int_0^t \mathfrak{s}(u)\mathrm{d}u\right] \\ &\leq C\left(1 + t\mu + \frac{s_0-\mu}{\kappa}(1-e^{-\kappa t})\right) \end{aligned}$$

where we used Fubini's theorem for the last inequality. Finally by independence of τ_N and M we have

$$\begin{aligned} \mathrm{E}\left[\sup_{0\leq u\leq \tau_N} s(u)\right] &\leq \int_0^\infty \left(\kappa\mu t + \sigma C\left(1 + t\mu + \frac{s_0-\mu}{\kappa}(1-e^{-\kappa t})\right)\right)\lambda e^{-\lambda t}\mathrm{d}t \\ &< \infty. \end{aligned}$$

□

Proof of Theorem 8.5. As the candidate value function must satisfy $v(0) = 0$, we can solve (FB1), and use equations (FB2) and (FB3) to obtain c, β and the candidate value function

$$v(s) = \begin{cases} s & \text{if } s < \beta \\ \frac{\alpha}{\lambda} + c(\kappa(s-\mu))^{-\lambda/\kappa} & \text{if } s \geq \beta \end{cases}$$

In the case $\frac{\alpha}{\lambda} \geq \mu$, we have

$$\frac{\alpha}{\lambda} \geq \beta = \frac{\alpha + \kappa\mu}{\kappa + \lambda} \geq \mu$$

and it can easily be verified that v equals the value function, using the same method as in the proof of theorem 8.4.

For $\frac{\alpha}{\lambda} < \mu$, however, the above candidate solution contradicts the bound $\mathcal{V}_1(s) \leq \frac{\alpha}{\lambda}$. Since the spread is deterministic we will again apply a direct method to verify the form of $\mathcal{V}_1$: for any $t \in [0, \infty)$ we have

$$\begin{aligned} \mathrm{E}[s(t)] &= \mathrm{E}\left[(s_0 e^{-\kappa t} + \mu(1 - e^{-\kappa t}) + \alpha t)1_{\{\tau_N > t\}}\right] + \mathrm{E}\left[\alpha \tau_N 1_{\{\tau_N \leq t\}}\right] \\ &= \frac{\alpha}{\lambda} + (s_0 - \mu)e^{-(\kappa+\lambda)t} + \left(\mu - \frac{\alpha}{\lambda}\right) e^{-\lambda t} \end{aligned}$$

and

$$\mathrm{E}[s(\infty)] = \frac{\alpha}{\lambda}$$

A straightforward computation shows that if $s_0 > \frac{\alpha}{\lambda}$, then for all $t \in [0, \infty]$

$$\mathrm{E}[s(t)] \geq \mathrm{E}[s(\infty)]$$

and if $s_0 \leq \frac{\alpha}{\lambda}$, then for all $t \in [0, \infty]$

$$\mathrm{E}[s(t)] \geq \mathrm{E}[s(0)]$$

As all stopping times in $\mathcal{T}(0, \infty)$ are deterministic, we obtain (8.24) by combining the above inequalities. □

Proof of Proposition 8.1. Suppose the microtrader decides to never cross the spread and wait until the limit order is executed. The associated expected costs are

$$\mathcal{J}(\infty; x_0, \mathfrak{s}_0) = \mathrm{E}\left[\alpha \int_t^\tau X^2(s) \mathrm{d}s\right] = \mathrm{E}\left[\alpha \tau_N X^2(0)\right],$$

where τ_N is the first jump time of the Poisson process N. Hence $\mathcal{J}(\infty; x_0, \mathfrak{s}_0) = \frac{\alpha}{\lambda} x_0^2$.

Now let τ be any stopping time in $\mathcal{T}(0, \infty)$. Then

$$\begin{aligned} \mathcal{J}(\tau; x_0, \mathfrak{s}_0) &\geq \mathrm{E}\left[\frac{\alpha}{\lambda} X^2(\tau) + \alpha \int_{\tau_N}^{\tau} X^2(u) \mathrm{d}u + \alpha \int_0^{\tau_N} X^2(u) \mathrm{d}u\right] \\ &\geq \mathrm{E}\left[\frac{\alpha}{\lambda} X^2(\tau) + \alpha \int_{\tau_N}^{\tau} X^2(u) \mathrm{d}u + \frac{\alpha}{\lambda} x_0^2\right] \end{aligned}$$

Hence it is enough to show that $\mathrm{E}[\frac{1}{\lambda} X^2(\tau) + \int_{\tau_N}^{\tau} X^2(u) \mathrm{d}u] \geq 0$. On the set $\{\tau > \tau_N\}$ this is clear. On $S := \{\tau \leq \tau_N\}$, let μ_τ be the distribution of τ. By independence of τ and τ_N we can calculate explicitly

$$\begin{aligned}
\mathrm{E}\left[1_S\left(\frac{1}{\lambda}X^2(\tau)+\int_{\tau_N}^{\tau}X^2(u)\mathrm{d}u\right)\right] &= x_0^2\mathrm{E}\left[1_S\left(\frac{1}{\lambda}-(\tau_N-\tau)\right)\right]\\
&= \int_0^\infty\int_0^\infty 1_{\{t\le s\}}\left(1/\lambda-(s-t)\right)\lambda e^{-\lambda s}\mathrm{d}s\,\mu_\tau(\mathrm{d}t)\\
&= 0
\end{aligned}$$

Hence $\mathcal{J}(\tau;x_0,\mathfrak{s}_0)\ge\mathcal{J}(\infty;x_0,\mathfrak{s}_0)$ and the result follows. □

Proof of Theorem 8.6. Let $v(x,s)=s^2w(x/s)$. We will show that $v(x,s)=\mathcal{V}(x,s)$. Let $\tilde{N}(t)=N(t)-\lambda t$ be the compensated Poisson process, and $z(t)=X(t)/\mathfrak{s}(t)$. Using Itô's formula we compute $\mathrm{d}z(t)=z(t-)((\sigma^2-r)\mathrm{d}t-\sigma\mathrm{d}W(t)-\mathrm{d}\tilde{N}(t))$ and thus

$$\begin{aligned}
\mathrm{d}v(X(t),\mathfrak{s}(t)) &= (\mathcal{L}_v v)(X(t),\mathfrak{s}(t))\mathrm{d}t+\mathrm{d}M_1(t)-\mathrm{d}M_2(t)\\
&= X^2(t)(\mathcal{L}_w w)(z(t))\mathrm{d}t+\mathrm{d}M_1(t)-\mathrm{d}M_2(t)
\end{aligned}$$

where $M_1(t)=\sigma\int_0^t v_s(X(u),\mathfrak{s}(u))\mathfrak{s}(u)\mathrm{d}W(u)$ and $M_2(t)=\int_0^t v(X(u),\mathfrak{s}(u))\mathrm{d}\tilde{N}(u)$ are local martingales. Hence

$$v(X(t),\mathfrak{s}(t))=v(X(0),\mathfrak{s}(0))+\int_0^t X^2(u)(\mathcal{L}_w w)(z(u))\mathrm{d}u+M_1(t)+M_2(t)$$

By lemma 8.4 $(\mathcal{L}_w w)(z)+\alpha\ge 0$ for all $z\ge 0$. Thus

$$v(X(t),\mathfrak{s}(t))\ge v(X(0),\mathfrak{s}(0))-\int_0^t\alpha X^2(u)\mathrm{d}u+M_1(t)+M_2(t)$$

Let now τ be any stopping time in $\mathcal{T}(0,\infty)$. By lemma 8.5, $\tilde{M}_1$ and $\tilde{M}_2$ defined by $\tilde{M}_i(t)=M_i(t\wedge\tau_N)$ are uniformly integrable martingales, so we can apply the optional sampling theorem and obtain

$$\mathrm{E}[v(X(\tau\wedge\tau_N),\mathfrak{s}(\tau\wedge\tau_N))]\ge v(X(0),\mathfrak{s}(0))-\mathrm{E}\left[\int_0^{\tau\wedge\tau_N}\alpha X^2(u)\mathrm{d}u\right]$$

By lemma 8.6, we have $v(x,s)\le\mu x^2+xs$ for all $x\ge 0$, $s\ge 0$, so

$$\mathrm{E}\left[\mu X^2(\tau\wedge\tau_N)+X(\tau\wedge\tau_N)\mathfrak{s}(\tau\wedge\tau_N)+\int_0^{\tau\wedge\tau_N}\alpha X^2(u)\mathrm{d}u\right]\ge v(X(0),\mathfrak{s}(0))$$

To prove equality for $\tau=\tau_b$, we use (W1) and the same arguments as above to obtain

$$\mathrm{E}\left[\mu X^2(\tau_b)+X(\tau_b)\mathfrak{s}(\tau_b)+\int_0^{\tau_b}\alpha X^2(u)\mathrm{d}u\right]=v(X(0),\mathfrak{s}(0))$$

Thus $\mathcal{V}(x,s)=v(x,s)$. □

Lemma 8.4. *Let w be given as in* (8.32)*. We have*

$$(\mathcal{L}_w w)(z)+\alpha\ge 0\qquad(8.43)$$

for all $z\ge 0$.

Proof. For $z < b$, we have $(\mathcal{L}_w w)(z) + \alpha = 0$. Thus we can discard the case $b = \infty$ and assume henceforth that $\alpha > \lambda\mu$ and $z > b$. In that case $w(z) = \mu z^2 + z$, so, after some algebra, (8.4) is equivalent to

$$(\alpha - \lambda\mu)z + (r - \lambda) \geq 0 \tag{8.44}$$

As $\alpha - \lambda\mu > 0$, it suffices to show (8.44) for $z = b$, which (after some algebra) is equivalent to

$$2\lambda\sigma^2 + r\left(\sqrt{(2r - \sigma^2)^2 + 8\lambda\sigma^2} + (2r - \sigma^2)\right) \geq 0 \tag{8.45}$$

which is clearly true for $r \geq 0$. For $r < 0$, (8.45) is equivalent to

$$2\lambda\sigma^2 + r(2r - \sigma^2) \geq -r\sqrt{(2r - \sigma^2)^2 + 8\lambda\sigma^2}$$

Both sides of the inequality are positive for $r < 0$, so we can square both sides and obtain that (8.45) is equivalent to

$$4\lambda\sigma^4(\lambda - r) \geq 0$$

which is true for $r < 0$. This completes the proof. □

Lemma 8.5. *The local martingales defined by* $\tilde{M}_1(t) = \sigma \int_0^{t\wedge\tau_N} v_s(X(u), \mathfrak{s}(u))\mathfrak{s}(u)\mathrm{d}W(u)$ *and* $\tilde{M}_2(t) = \int_0^{t\wedge\tau_N} v(X(u), \mathfrak{s}(u))\mathrm{d}\tilde{N}(u)$ *are uniformly integrable.*

Proof. Consider first $\tilde{M}_1$. By Protter (2004, Thm. 11 with $G(x) = x^2$) it is enough to show that

$$\sup_{t\geq 0} \mathrm{E}[\tilde{M}_1^2(t)] < \infty$$

Note that

$$sv_s(x,s) = \begin{cases} xs & \text{if } s \leq \frac{x}{b} \\ c_1(2-p_1)x^{p_1}s^{(2-p_1)} & \text{if } s > \frac{x}{b} \end{cases}$$

As $p_1 > 2$, we have $|sv'(s)| \leq \max\left\{\frac{x^2}{b}, c_1(2-p_1)x^{p_1}\left(\frac{x}{b}\right)^{(2-p_1)}\right\} =: L$, we compute the bound

$$\begin{aligned} \mathrm{E}\left[[\tilde{M}_1](t)\right] &\leq \mathrm{E}\left[[\tilde{M}_1](\infty)\right] \\ &= \mathrm{E}\left[\sigma^2 \int_0^{\tau_N} \left(v'(s(u))s(u)\right)^2 \mathrm{d}u\right] \\ &\leq \sigma^2 L^2 \mathrm{E}\left[\tau_N\right] \\ &\leq \frac{\sigma^2 L^2}{\lambda} =: K \end{aligned}$$

Thus $\sup_{t\geq 0} \mathrm{E}[\tilde{M}_1^2(t)] \leq K < \infty$ and $\tilde{M}_1$ is a u.i. martingale. A similar argument, using $v(x,s) \leq \frac{x^2\alpha}{\lambda}$ shows that $\tilde{M}_2$ is also a u.i. martingale. □

Lemma 8.6. *Let w be given as in* (8.32)*. Then we have* $w(z) \leq \mu z^2 + z$ *for all* $z \geq 0$.

Proof. We only need to consider $z < b$.

If $\alpha \leq \lambda\mu$, we have $b = \infty$ and $c_1 = 0$, so

$$w(z) = \frac{\alpha}{\lambda}z^2 \leq \mu z^2 \leq \mu z^2 + z$$

If $\alpha > \lambda\mu$, we need to show that

$$\phi(z) = \left(\mu - \frac{\alpha}{\lambda}\right) z - c_1 z^{(p_1-1)} + 1 \geq 0$$

for $z \in [0, b)$. Since $\phi(0) = 1$ and $\phi(b) = 0$ it is sufficient to show that ϕ is decreasing on $[0, b)$, i.e. $\phi^{'} \leq 0$ on $[0, b)$. Using the definition of c_1 and $p_1 \geq 2$ one shows that $\phi^{'}(b) = \mu - \alpha/\lambda - c_1(p_1 - 1)b^{(p_1-2)} \leq 0$, so it is enough to show that $\phi^{''} \geq 0$ on $[0, b)$. As $\phi^{''}(z) = -c_1(p_1-1)(p_1-2)z^{(p_1-3)}$ the result follows since $c_1 \leq 0$. □

Proof of Proposition 8.2. The result follows by considering the explicit formula

$$b = b(\lambda, r, \sigma, \alpha, \mu) = \frac{\lambda(2r + \sigma^2 + \sqrt{(2r - \sigma^2)^2 + 8\lambda\sigma^2})}{(\alpha - \lambda\mu)(2r - \sigma^2 + \sqrt{(2r - \sigma^2)^2 + 8\lambda\sigma^2})}$$

and taking derivatives with respect to the various parameters. □

Proof of Corollary 8.1. For $\alpha \leq \lambda\mu$, the optimal strategy consists in pegging forever, thus $\mathrm{P}(C) = 0$. For $\alpha > \lambda\mu$ and $\mathfrak{s}_0 \leq \frac{x_0}{b}$, the optimal strategy consists in crossing immediately at $t = 0$, thus $\mathrm{P}(C) = 1$. Now consider the remaining case. Note that $z < 0$ Let τ_N be the first jump time of N. Then by independence of τ_b and τ_N,

$$\begin{aligned}\mathrm{P}(C) &= \mathrm{P}(\tau_b < \tau_N) \\ &= \int_0^\infty \int_t^\infty \lambda e^{-\lambda s} \mathrm{d}s\, \mu_b(\mathrm{d}t)\end{aligned}$$

where μ_b denotes the density of τ_b. Moreover, using the explicit formula for the geometric Brownian motion

$$\begin{aligned}\tau_b &= \inf\left\{t \geq 0 : \mathfrak{s}(t) \leq \frac{x_0}{b}\right\} \\ &= \inf\left\{t \geq 0 : W(t) + \rho t \leq z\right\}\end{aligned}$$

and we deduce, using the density of the first hitting time of a Brownian motion with drift (see, e.g., Borodin and Salminen, 2002) that

$$\mu_b(\mathrm{d}t) = -\frac{z}{\sqrt{2\pi}t^{3/2}} e^{-\frac{(z-\rho t)^2}{2t}} \mathrm{d}t$$

Using this formula, we can compute the double integral and obtain

$$\mathrm{P}(C) = \exp\left\{z(\rho + \sqrt{2\lambda + \rho^2})\right\}$$

□

Appendix A

Matlab code

A.1 Calibration of 3-dimensional model

The main programm `Calibrate.m` estimates the parameters $(\kappa, \mu, \sigma_1, \rho, \nu, \sigma_2, d_1, d_2, \sigma_0)$ of the three-dimensional model defined by equations (5.1) - (5.3).

```
% Estimates coefficients of order book model
% with dynamics given by
% ds(t) =   kappa(mu−s(t)).dt          + sigma1/sqrt(2).dB1(t)
% dz(t) =   rho(nu−z(t)).dt            + sigma2/sqrt(2).dB3(t)
% dm(t) =   0.5(d+d2.z(t)).dt          + simga1/sqrt(8).dB2(t)
%           + sigma0/sqrt(2).dB0(t)
%
% where B0,..., B3 are independent standard brownian motions

close all;

% load orderbook data with filename name.mat
load([name,'.mat'])

% assign order book data to variables:
% T=time, A=best ask, B=best bid
% VA = best ask volume, VB = best bid volume
T = orderbook(:,1);
A = orderbook(:,4);
B = orderbook(:,2);
VA = orderbook(:,5);
VB = orderbook(:,3);
N = length(T);

% Sample sparsely by taking every K−th data point
% choose K s.t. ¬r seconds between two time points in sparse data set
K = 40;

% options for fminsearch
options = optimset('MaxFunEvals',1000);

% calculate spread, midquote, vib
S = A − B;
M= (A+B)/2;
```

```
Z = log(VB./VA);
DeltaT = diff(T);

% setup variables
RV_S_all = 0;
RV_S_avg = zeros(1,K);
RV_M_all = 0;
RV_M_avg = zeros(1,K);
RV_Z_all = 0;
RV_Z_avg = zeros(1,K);

% sample using all points
RV_S_all = sum(diff(S(1:end)).^2);
RV_M_all = sum(diff(M(1:end)).^2);
RV_Z_all = sum(diff(Z(1:end)).^2);


% sample on sparse grid
for k=1:K
    index = k:K:length(S);
    Tsparse = T(index);
    Ssparse = S(index);
    Msparse = M(index);
    Zsparse = Z(index);
    RV_S_avg(k) = sum(diff(Ssparse(1:end)).^2);
    RV_M_avg(k) = sum(diff(Msparse(1:end)).^2);
    RV_Z_avg(k) = sum(diff(Zsparse(1:end)).^2);
end

% use 'First best approach' from Yacine Ait-Sahalia to get rid of
% bias
RV_S = mean(RV_S_avg)-RV_S_all/K;
RV_M = mean(RV_M_avg)-RV_M_all/K;
RV_Z = mean(RV_Z_avg)-RV_Z_all/K;



% compute volatilities
sigma0 = sqrt((4*RV_M - RV_S)/(T(end)-T(1)));
sigma2 = sqrt(2*RV_Z/(T(end)-T(1)));
int_s  = sum(diff(T).*S(1:end-1));
sigma1 = sqrt(RV_S*2/int_s);


% Initial parameters to calibrate CIR spread process via OLS
InitialParamsCIR = InitialPCIR(DeltaT, S);

% Calculate MaxLikelihood estimators of CIR spread
% by minimising log likelihood function
x = fminsearch(@(x) CIRLogLike(x, sigma1, DeltaT, S),InitialParamsCIR(1:2), options);

% Initial parameters to calibrate OU volume imbalance process via OLS
InitialParamsOU = InitialPOU(DeltaT, Z);

% Calculate MaxLikelihood estimators of OU volume imbalance
% by minimising log likelihood function
y = fminsearch(@(y) OULogLike(y, sigma2, DeltaT, Z),InitialParamsOU, options);

% Calculate Midquote Drift parameters using OLS

mq=MidquoteDrift(diff(T), M, Z);
```

```
% assign variables
kappa=x(1);
mu=x(2);

rho=y(1);
nu=y(2);

d1 = mq(1);
d2 = mq(2);
```

The function **InitialPCIR.m** returns initial estimates for the parameters of a CIR process, using the ordinary least square method.

```
function InitialParams = InitialPCIR(DeltaT, S)

% CIR initial parameters estimation
% ds(t) = kappa.(mu-s(t)).dt + sigma.sqrt(s(t)).dW(t)
% input:    (DeltaT, S) = observations of CIR process
% output:   InitialParams = [kappa mu sigma] = initial estimates for CIR
%           parameters

x = S(1:end-1); %
dx = diff(S);
dx = dx./x.^0.5;
regressors = [DeltaT./x.^0.5, DeltaT.*x.^0.5];
drift = regressors\dx; % OLS regressors coefficients estimates
res = regressors*drift - dx;
kappa = -drift(2);
mu = -drift(1)/drift(2);
sigma = sqrt(var(res./sqrt(DeltaT), 1));
InitialParams = [kappa mu sigma];  % return vector of initial parameters
```

The function **InitialPOU.m** returns initial estimates for the parameters of an OU process, using the ordinary least square method.

```
function InitialParams = InitialPOU(DeltaT, Z)

% OU initial parameters estimation
% dz(t)= rho*(nu-z(t))*dt + sigma*dW(t)
% input:    (DeltaT, Z) = observations of OU process
% output:   InitialParams = [rho nu sigma] = initial estimates for OU
%           parameters

x = Z(1:end-1);
dx = diff(Z);
regressors = [Z(1:end-1).*DeltaT, DeltaT];
drift = regressors\dx; % OLS regressors coefficients estimates
res = regressors*drift - dx;
rho = -drift(1);
nu = drift(2)/rho;
sigma = sqrt(var(res./sqrt(DeltaT), 1));
InitialParams = [rho nu sigma]; % return vector of initial parameters
```

The function **CIRLogLike.m** returns the log-likelihood function of a CIR process.

```
function lnL = CIRLogLike(Params, DeltaT, S)

% Log-likelihood function (multiplied by -1) for CIR process
% ds(t) = kappa.(mu-s(t)).dt + sigma.sqrt(s(t)).dW(t)
% input:    (DeltaT, S) = observations of CIR process
%           Params = Model parameters (kappa, mu, sigma)
% output:   Log-likelihood function (multiplied by -1) for CIR process
%           with parameters (kappa, mu, sigma)

kappa = Params(1);
mu = Params(2);
sigma = Params(3);
c = 2*kappa./(sigma^2.*(1-exp(-kappa.*DeltaT)));
q = 2*kappa*mu/sigma^2-1;
u = c.*exp(-kappa.*DeltaT).*S(1:end-1);
v = c.*S(2:end);
z = 2*sqrt(u.*v);
bf = besseli(q,z,1);
lnL=  sum(-log(c) + u + v - 0.5*q*log(v./u) - log(bf) - z);
end
```

The function `OULogLike.m` returns the log-likelihood function of an OU process.

```
function lnL = OULogLike(Params, DeltaT, Z)

% Log-likelihood function (multiplied by -1) for OU process
% dz(t)= rho*(nu-z(t))*dt + sigma*dW(t)
% input:    (DeltaT, Z) = observations of OU process
%           Params = Model parameters (rho, nu, sigma)
% output:   Log-likelihood function (multiplied by -1) for OU process
%           with parameters (rho, nu, sigma)

Data = Z;
DataF = Data(2:end);
DataL = Data(1:end-1);
rho = Params(1);
nu = Params(2);
sigma = sigma2/sqrt(2);
e = exp(-rho.*DeltaT);
c = sigma^2.*(1-e.^2)./(2*rho);
lnL=  sum(   .5.*log(2*pi*c)+((DataF-DataL.*e-nu*(1-e)).^2)./(2*c)    );
end
```

A.2 Computation of first passage time

The program `Data_density.m` computes the CDF of the first passage time from orderbook data and stores it in the variable `CDF`.

```
% Extracts time to fill from data
%
% input:    -NumberTicks, InitialSpread, InitialVib, time_cut and steps
%            need to defined before
%           -Order book data assumed to be stored in 'orderbook'
% output:   CDF = cumulative distribution function of time-to-fill of
```

```
%           sell order placed NumberTicks above initial best ask
%           sample_size = sample size used to compute CDF


% end time in milliseconds
tEnd = time_cut;
% stepsize for density in milliseconds
stepsize = steps;

% time grid
Tgrid = 1:stepsize:tEnd;

% setup cumulative distribution function
CDF = zeros(1,length(Tgrid));

% initialize sample_size to 0
sample_size = 0;

% take orderbook data from workspace
T = orderbook(:,1);
A = orderbook(:,4);
B = orderbook(:,2);
VA = orderbook(:,5);
VB = orderbook(:,3);
N = length(T);


% define spread, midquote, best ask, best bid
% round quantities to .01 precision
S = A - B; S=round(S*100)/100;
M= (A+B)/2; M=round(M*100)/100;
Ask = A; Ask=round(Ask*100)/100;
Bid = B; Bid=round(Bid*100)/100;

% define volume-imbalance rounded to 1.0 precision
VolA = VA;
VolB = VB;
Z = log(VolB./VolA);
Z = round(Z*1)/1;

% tick size of order book
TickSize = 0.01;

% index of states where vib = InitialVib and spread = InitialSpread
% and more than tEnd milliseconds of data left
index = find((S==zeros(size(S))+InitialSpread)
        & (Z==zeros(size(Z))+InitialVib)
        & (T < T(end) - tEnd));

for i=1:length(index)
    %find index of first time when Ask reaches barrier
    k=find( Ask(index(i):end)-Ask(index(i)) ≥ NumberTicks*TickSize, 1,'first');
    if ¬isempty(k)
        % time ellapsed until reaching barrier
        Tellapsed = T(k+index(i)-1) - T(index(i));
        j = floor(Tellapsed/stepsize)+2;
        if j≤length(Tgrid)
            %ind_decile = floor(CDF_MC(j)*10)+1;
            %decile_count(ind_decile) = decile_count(ind_decile)+1;
            CDF(j:end) = CDF(j:end) + 1;
        end
```

```
    end
end

% cumulative distribution function and sample_size
CDF = CDF/length(index);
sample_size = length(index);
```

The program `MonteCarlo_density.m` computes the CDF of the first passage time of the three-dimensional orderbook via Monte-Carlo simulations and stores it in the variable `CDF_MC`. To simulate sample paths, a simple Euler scheme is used.

```
% MonteCarlo simulation of time to fill from model
%
% Underlying model given by
%
% dalpha(t) = +.5*[kappa.(mu-s(t)).dt + (d_1+d_2.z(t)).dt +
% sigma1*sqrt(s(t)).dW1(t) + sigma0.dW0(t)]
% dbeta(t) = -.5*[kappa.(mu-s(t)).dt - (d_1+d_2.z(t)).dt +
% sigma1*sqrt(s(t)).dW2(t) - sigma0.dW0(t)]
% dz(t) = rho(nu-z(t)).dt + sigma2/sqrt(2)*dB3(t)
%
% input:    -NumberTicks, InitialSpread, InitialVib, timecut, steps
%            need to defined before
%           -Model Parameters (kappa, mu, ...) defined in 'Calibrate.m'
% output:   CDF_MC = cumulative distribution function of time-to-fill of
%           sell order placed NumberTicks above initial best ask


% end time in milliseconds
tEnd = time_cut;
% stepsize for density in milliseconds
stepsize = steps;
% Euler-Steps in between two stepsize
eulersteps = 2;
% no. of steps in Euler-Scheme
N= tEnd/stepsize*eulersteps;
% no. of MonteCarlo steps
M = 4000;
% every L-th point corresponds to stepsize
L = round(tEnd/stepsize * N);
ind = 1:L:(N-L);

% tick size of order book
DeltaTick = 0.01;
% initial best ask, can be any arbitrary value
alpha_0 = 12;
% initial spread and best bid
beta_0 = alpha_0-InitialSpread;

% initialize vectors Alpha, Beta, dt and T
Alpha = zeros(1,N);
Alpha(1) = alpha_0;
Beta  = zeros(1,N);
Beta(1)   = beta_0;
Z = zeros(1,N);
Z(1) = InitialVib;
T = linspace(0,tEnd,N);
dt = tEnd/N;

```

```
% count(k) counts no. of sample paths where order is filled after
% k*stepsize milliseconds
count = zeros(1,tEnd/stepsize);

% start MonteCarlo
for j=1:M
    % start Euler
    for n = 2:N
        s = max(Alpha(n-1)- Beta(n-1),0);
        W0 = randn; W1 = randn; W2 = randn; B3 = randn;
        Alpha(n) = Alpha(n-1) + .5*(kappa*(mu-s)*dt + d1*dt
                   + d2*Z(n-1)*dt + sigma1*sqrt(s)*sqrt(dt)*W1 + sigma0*sqrt(dt)*W0);
        Beta(n)  = Beta(n-1)  - .5*(kappa*(mu-s)*dt - d1*dt
                   - d2*Z(n-1)*dt + sigma1*sqrt(s)*sqrt(dt)*W2 - sigma0*sqrt(dt)*W0);
        Z(n) = Z(n-1) + rho*(nu-Z(n-1))*dt + sigma2/sqrt(2)*sqrt(dt)*B3;
    end
    for n=1:tEnd/stepsize
        if (max(Alpha(1:(1+2*(n-1)))) ≥ alpha_0 + NumberTicks*DeltaTick)
            count(n) = count(n) +1;
        end
    end
end
% average over all sample paths gives CDF
CDF_MC = count./M;
```

Bibliography

Abramowitz, M., and I.A. Stegun, 1964, *Handbook of mathematical functions with formulas, graphs, and mathematical tables* (Dover publications).

Ait-Sahalia, Y., L. Zhang, and P.A. Mykland, 2005, A tale of two time scales, *Journal of the American Statistical Association* 100, 1394–1411.

Alfonsi, A., A. Fruth, and A. Schied, 2008, Constrained portfolio liquidation in a limit order book model, *Banach Center Publ* 83, 9–25.

——— , 2010, Optimal execution strategies in limit order books with general shape functions, *Quantitative Finance* 10(2), 143–157.

Alfonsi, A., and A. Schied, 2010, Optimal trade execution and absence of price manipulations in limit order book models, *SSRN eLibrary.*

——— , and A. Slynko, 2009, Order book resilience, price manipulation, and the positive portfolio problem, *Preprint.*

Almgren, R., 2003, Optimal execution with nonlinear impact functions and trading-enhanced risk, *Applied Mathematical Finance* 10, 1–18.

——— , 2009, Optimal trading in a dynamic market, *Preprint.*

——— , and N. Chriss, 1999, Value under liquidation, *Risk* 12, 61–63.

——— , 2001, Optimal execution of portfolio transactions, *Journal of Risk* 3, 5–40.

Almgren, R., and J. Lorenz, 2007, Adaptive arrival price, *Algorithmic Trading III: Precision, Control, Execution.*

Barles, G., and J. Burdeau, 1995, The Dirichlet problem for semilinear second-order degenerate elliptic equations and applications to stochastic exit time control problems, *Communications in Partial Differential Equations* 20, 129–178.

Bertoin, J., 1998, *Lévy processes* (Cambridge University Press).

Bertsimas, D., and A. Lo, 1998, Optimal control of execution costs, *Journal of Financial Markets* 1, 1–50.

Biais, B., P. Hillion, and C. Spatt, 1995, An empirical analysis of the limit order book and the order flow in the Paris Bourse, *Journal of Finance* 50, 1655–1689.

Bollerslev, T., and M. Melvin, 1994, Bid–ask spreads and volatility in the foreign exchange market : An empirical analysis, *Journal of International Economics* 36, 355 – 372.

Borodin, A.N., and P. Salminen, 2002, *Handbook of Brownian motion: facts and formulae* (Birkhauser).

Bouchaud, J.P., M. Mézard, and M. Potters, 2002, Statistical properties of stock order books: empirical results and models, *Quantitative Finance* 2, 251–256.

Bovier, A., J. Cerny, and O. Hryniv, 2006, The opinion game: stock price evolution from microscopic market modelling, *International Journal of Theoretical and Applied Finance* 9, 91–111.

Cao, C., O. Hansch, and X. Wang, 2009, The information content of an open limit-order book, *Journal of Futures Markets* 29, 16–41.

Cohen, K.J., S.F. Maier, R.A. Schwartz, and D.K. Whitcomb, 1981, Transaction costs, order placement strategy, and existence of the bid-ask spread, *The Journal of Political Economy* pp. 287–305.

Cont, R., A. Kukanov, and S. Stoikov, 2010, The price impact of order book events, *ArXiv e-prints.*

Cont, R., S. Stoikov, and R. Talreja, 2010, A stochastic model for order book dynamics, *Operations Research* 58, 676–713.

Crandall, M.G., H. Ishii, and P.L. Lions, 1992, User's guide to viscosity solutions of second order partial differential equations, *Bulletin of the American Mathematical Society* 27, 1–67.

Díaz, G., 2004, On the Dirichlet boundary problem for quasi-degenerate elliptic linear equations, *Contribuciones matemáticas. Homenaje al profesor Enrique Outerelo Domínguez. Universidad Complutense de Madrid* pp. 103–126.

Foucault, T., 1999, Order flow composition and trading costs in a dynamic limit order market, *Journal of Financial markets* 2, 99–134.

——— , O. Kadan, and E. Kandel, 2005, Limit order book as a market for liquidity, *Review of Financial Studies* 18, 1171.

Fouque, J.P., G. Papanicolaou, and K.R. Sircar, 2000, *Derivatives in financial markets with stochastic volatility* (Cambridge University Press).

Gatheral, J., and A. Schied, 2010, Optimal trade execution under geometric brownian motion in the Almgren and Chriss framework, *Preprint.*

——— , and A. Slynko, 2010a, Exponential resilience and decay of market impact, *Preprint.*

——— , 2010b, Transient linear price impact and Fredholm integral equations, *Mathematical Finance, Forthcoming.*

Gregory, J., and C. Lin, 1996, *Constrained optimization in the calculus of variations and optimal control theory* (Springer Verlag).

Hall, A.D., and N. Hautsch, 2006, Order aggressiveness and order book dynamics, *Empirical Economics* 30, 973–1005.

Harris, L., 2003, *Trading and exchanges: Market microstructure for practitioners* (Oxford University Press).

Hautsch, N., and R. Huang, 2009, The market impact of a limit order, *SFB 649 Discussion Papers.*

Huberman, G., and W. Stanzl, 2005, Optimal liquidity trading, *Review of Finance* 9, 165.

Jain, P. K., 2003, Institutional design and liquidity at stock exchanges around the world, *SSRN eLibrary.*

Kladivko, K., 2007, Maximum likelihood estimation of the Cox-Ingersoll-Ross process: The matlab implementation, *Working paper.*

Kratz, P., 2011, Optimal liquidation in dark pools in discrete and continuous time, Ph.D. thesis Humboldt Universität Berlin.

——— , and T. Schöneborn, 2010, Optimal liquidation in dark pools, *Preprint.*

Kyprianou, A.E., 2006, *Introductory lectures on fluctuations of Lévy processes with applications* (Springer Verlag).

Luckock, H., 2003, A steady-state model of the continuous double auction, *Quantitative Finance* 3, 385–404.

Malo, P., and T. Pennanen, 2010, Reduced form modeling of limit order markets, Quantitative Finance Papers 1006.4517 arXiv.org.

Maslov, S., and M. Mills, 2001, Price fluctuations from the order book perspective–empirical facts and a simple model, *Physica A: Statistical Mechanics and its Applications* 299, 234–246.

Naujokat, F., and N. Westray, 2010, Curve following in limit order markets, *Preprint.*

Obizhaeva, A., and J. Wang, 2005, Optimal trading strategy and supply/demand dynamics, *NBER Working Paper.*

Omura, K., Y. Tanigawa, and J. Uno, 2000, Execution probability of limit orders on the Tokyo Stock Exchange, *SSRN eLibrary.*

Osterrieder, J.R., 2007, Arbitrage, the limit order book and market microstructure aspects in financial market models, Ph.D. thesis ETH Zürich.

Parlour, C.A., 1998, Price dynamics in limit order markets, *Review of Financial Studies* 11, 789.

Pascual Gascó, R., and D. Veredas, 2008, What pieces of limit order book information matter in explaining order choice by patient and impatient traders?, *SSRN eLibrary.*

Peskir, Goran, 2007, A change-of-variable formula with local time on surfaces, in *Séminaire de Probabilités XL*, vol. 1899 of *Lecture Notes in Mathematics* . pp. 70–96 (Springer Verlag).

Peskir, G., and A. N. Shiryaev, 2006, *Optimal stopping and free-boundary problems* (Birkhauser).

Pham, H., 1998, Optimal stopping of controlled jump diffusion processes and viscosity solutions, *Journal of Mathematical Systems, Estimation, and Control* 8, 1–27.

Potters, M., and J.P. Bouchaud, 2003, More statistical properties of order books and price impact, *Physica A: Statistical Mechanics and its Applications* 324, 133–140.

Predoiu, S., G. Shaikhet, and S. Shreve, 2010, Optimal execution in a general one-sided limit-order book, *Preprint*.

Protter, P.E., 2004, *Stochastic integration and differential equations* (Springer Verlag).

Ranaldo, A., 2004, Order aggressiveness in limit order book markets, *Journal of Financial Markets* 7, 53–74.

Rosu, I., 2009, A dynamic model of the limit order book, *Review of Financial Studies* 22, 4601.

Schied, A., and T. Schöneborn, 2009a, Liquidation in the face of adversity: Stealth versus sunshine trading, *Preprint*.

——— , 2009b, Risk aversion and the dynamics of optimal liquidation strategies in illiquid markets, *Finance and Stochastics* 13, 181–204.

——— , and M. Tehranchi, 2010, Optimal basket liquidation for CARA investors is deterministic, *Applied Mathematical Finance* 17, 471–489.

Shiryaev, A. N., 2008, *Optimal Stopping Rules* (Springer Verlag).

Smith, E., J.D. Farmer, L. Gillemot, and S. Krishnamurthy, 2003, Statistical theory of the continuous double auction, *Quantitative Finance* 3, 481–514.

Touzi, N., 2004, *Stochastic control problems, viscosity solutions and application to finance* (Lecture notes, Scuola normale superiore di Pisa).

Veillette, M., and M.S. Taqqu, 2010, Numerical computation of first-passage times of increasing Lévy processes, *Methodology and Computing in Applied Probability* 12, 695–729.

Zovko, I., and J.D. Farmer, 2002, The power of patience: a behavioural regularity in limit-order placement, *Quantitative Finance* 2, 387–392.

Eidesstattliche Versicherung

Ich versichere an Eides statt, dass ich die von mir vorgelegte Dissertation selbständig angefertigt habe und alle benutzten Quellen und Hilfsmittel vollständig angegeben habe. Die Zusammenarbeit mit anderen Wissenschaftlern habe ich kenntlich gemacht. Diese Personen haben alle bereits ihr Promotionsverfahren abgeschlossen.

Erklärung

Keine Teile dieser Arbeit sind bereits veröffentlicht.

Eine Anmeldung der Promotionsabsicht habe ich an keiner anderen Fakultät oder Hochschule beantragt.