Optimal Trading with Frictions

Zexin Wang

Supervisor: Prof. Dr. Johannes Muhle-Karbe,

Internal Examiner: Dr. Eyal Neuman,

External Examiner: Prof. Dr. Mathieu Rosenbaum

Submitted for completion of PhD degree

Statement of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Copyright Declaration

Abstract

This thesis studies the optimal trading problem with particular attention to frictions, taking alpha signals as given in several practical settings in modern financial markets.

Chapter 2 provides a reduced-form model for price impact of market orders. As a scaling limit of the econo-physics propagator model, it has both tractability for optimization and good empirical fit. The nonlinearity in propagator model is explained as a effect of intraday stochasticity of the market activity. Optimal trading strategies are given for the case of stochastic alpha signal and volume signals in closed-form solutions. Moreoever, concrete bounds for the absence of price manipulation strategies are provided.

Chapter 3 derives an actionable derivatives hedging strategy with both market and limit orders from the perspective of a central risk book. It is found that limit order is only beneficial for delta-hedging when the gamma of the risky position is negative. Additionally, the interaction between transaction cost, adverse selection and risk aversion can be characterized by a nonlinear PDE that describes the option price. According to empirical analysis, tactical liquidity provision is beneficial for non-competitive market makers for reasonable trading frequencies.

Chapter 4 studies the usage of display and nondisplay limit orders for order execution. A price impact model is postulated and the corresponding scheduling algorithm is derived. In the case where nondisplay limit order (hidden order) is used, there is a time which separates the trading horizon into two regimes: the former only uses hidden order, and the latter uses the mixture of limit and hidden orders. The effectiveness and robustness of the algorithm is shown via numerical testing in both simulated data and NASDAQ 100 Index data.

Acknowledgements

I had a wonderful three and half years of research and writing this thesis, and it is important that I thank the people who contributed.

First and foremost, I would like to show my sincere appreciation towards my supervisor Prof. Dr. Johannes Muhle-Karbe for the support and encouragement. I learned not only about research and career, but also about life in general. I am very fortunate to have such a positive and wise supervisor, who consistently overcame many difficulties to help me.

I would also like to express my gratitude towards my examiners Prof. Dr. Mathieu Rosenbaum and Dr. Eyal Neuman, for their dedication and commitment to serve on my thesis, as well as the very interesting discussions during the defence.

I had the luxury of having a lot of amazing coauthors throughout the course of my research. I have benefited to a great extent from working with my coauthors, and I am very grateful to each and every one of them. Special thanks are due to Dr. Kevin Webster, Ge Zhang, Prof. Dr. Chao Zhou, Prof. Dr. Ying Chen and of course Johannes who contributed to this thesis. I also thank Dr. Weiguan Wang and Prof. Dr. Johannes Ruf, with whom we have another research project which is not included in this thesis.

The financial support was crucial in making all of this possible. For this, the credit goes to the Schrödinger Scholarship provided by the Faculty of Natural Sciences at Imperial, as well as the Roth Scholarship provided by the Department of Mathematics at Imperial.

Support from the Department, in particular the Mathematical Finance section has been very important along the way. For this, special thanks go to Dr. Eyal Neuman and Prof. Dr. Mikko Pakkanen, for acting as my examiners for my Early Stage Assessment and Late Stage Review. I sincerely appreciate their helpful and detailed comments on this thesis, from which I benefited a lot.

Credit also goes to the education I received at the MSc Mathematics and Finance program at Imperial, the Oxford-Imperial CDT in Mathematics of Random System as well as my undergraduate studies in Quantitative Finance at National University of Singapore,

from which I built my foundations of mathematical finance. Special thanks go to Dr. Jonathan Scarlett and Prof. Vincent Y.F. Tan who took me in as a research assistant, and Dr. Vik Gopal, Dr. Wee Seng Ng.

Very special thanks go to my parents Wenhui Wang and Yapin He, as well as my grandparents Chengliang Wang, Ruilan Liang, Shunxin He and especially Eryuan Cai. They are the most caring people I know.

My life had ups and downs, but I could always count on my friends to make the tough times easier and the good times truly memorable. I want to thanks my friends (in chronological, and then alphabetical order): Allan Wang, Tianhan Cheng, Song Kai, Yilin Jia, Song Pan, Yi Qin, Yisi Sun, Hai Wang, Chenmu Wu, Yuxin Wu, Yufei Xie, Penghao Chen, Shaocong Dong, Xiaohang Feng, Xinyi He, Ziyang Mou, Xijin Pu, Bingyi Xu, Muzi Yang, Sirui Yang, Han Zhang, Henry Morco, Mark Ng, Ling Weng, Yifei Zhang, Pengxin You, Naixian Chen, Galvin Ng, Xiaofei Shi, Aitor Muguruza, Claudio Bellani, Alessandro Micheli, Connor Tracy, Joseph Mulligan, Remy Messadene, Mingyu Zhang, Yun Feng and most importantly, *Chang Liu*. This list is by no means exhaustive. I did many internships and I have been lucky enough to stumble upon the most inspiring and helpful people. Here is another non-exhaustive list of the industry professionals who I would like to thank: at Numerical Technologies: Siaw Kian Zhong, Shaoxin Luan, at GIC: Adrian Chiu, Sanjay Yadav, at Citi: Nasir Miam, Anthony Neo, Kumar Gaurav, Xiao Xu, Liyang Sun, at Kaiju Capital Management: Aitor Muguruza, Nicholas Subryan, at DRW: Paolo Mattioli, Charles Morcom, Dean Read, Ben McDonnell, Yun Feng, Yushun Ma, Nicholas Beasley.

Last but by no means the least, I thank my fiancée Zi Li for her loving support and the great discussions we have had.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

'How to trade optimally' is one of the most popular and widely studied research topics in the field of finance. It is generally known that there is no one-size-fit-all optimal trading strategy, due to the complexity and uncertainty in market environment, and the different notions of optimality that are most natural in different contexts. In particular, various *frictions* such as price impact and other trading costs add an extra layer of complexity to the problem. In this dissertation, we explore how to optimally trade in the presence of frictions. In particular, we are interested in the features of the financial market that appear more recently, for example, access to intraday transaction data, central risk books, and non-display orders in lit pool.

Before outlining the agenda of this dissertation in more detail, we must specify the type of trading to be discussed. In practice, trading can be generally separated into the alpha generation process during which the number of units of each asset to be purchased or sold is allocated, and the order execution process during which the allocated number of units is gradually acquired or liquidated with the goal of minimizing costs caused by market frictions. One common practice in the industry is to assign two distinct divisions to handle the two procedures independently, or outsource one of the two entirely. Taking alpha signals as given, this dissertation serves to provide optimal strategies for the order execution process in several scenarios, in which market frictions and trading mechanisms are pertinent.

The primary focus of this thesis, trading frictions manifest itself in many different forms:

- *Price impact:* the adverse effect on asset price in purchasing or selling a large number of units. (Torre & Ferrari, 1997) More specifically, this is the purely mechanical impact caused by the execution. It is the largest at the instant of the execution, and gradually dissipates as time progresses.

- *Transaction costs:* the additional fee charged by the venue as well as market makers in exchange for immediate execution which is proportional to the number of units bought or sold. It contains bid-ask spread, the difference between the best bid and ask prices, and also make-take fee, proportional to the notional used to enter the position with a lower bound.

- *Adverse selection:* the possibility of being picked-off when trading against someone with inside information. On average, with adverse selection, one expects to lose in the zero-sum game as a direct result of asymmetric information. This phenomenon is especially important for market makers or agents posting limit orders, as the execution relies on the counterparty's discretion.

In the following, we present a brief overview of this thesis. For more in-depth discussion of the respective contents as well as detailed literature review, the reader is referred to the respective chapters.

One traditional order execution task is to acquire or liquidate a large number of units, and in turn, price impact is the most important component. How to quickly execute the order while incurring as little impact as possible becomes the main target here. The optimal execution strategy largely depends on the price impact model, which describes the relationship between order flow and asset price. Immediate execution is either extremely expensive or virtually impossible to achieve, and the task boils down to how to split the order into smaller chunks. The econo-physics propagator model (J. Bouchaud, Gefen, Potters, & Wyart, 2004) describes the price as being 'pushed' by the trades, with each push nonlinear in space (to the trade size) and decay in time. Despite its universally

good fit (J. Bouchaud, Farmer, & Lillo, 2009), the propagator model remains intractable due to its nonlinearity in space, and one can only resort to numerical solution under this model (T. Chen, Ludkovski, & Voß, 2022; Curato, Gatheral, & Lillo, 2017). On the other hand, the linear impact models are tractable, but does not fit the empirical data as well as the nonlinear ones (J. Bouchaud et al., 2009). Chapter 2 reconciles the two by showing that a scaling limit of the propagator model emerges as a linear impact model with stochastic liquidity. This limiting model serves as a bridge between the two: it fits the empirical data better than the linear impact model in the empirical analysis, and still remains tractable for optimization. Moreover it answers the practical question: 'how to incorporate intraday transaction data into execution strategy'. The three main results of this chapter are:

- Nonlinear price impact models can be described in reduced form by tractable linear models,in which stochastic price impact parameter links illiquidity and market activity.

- Optimal trading strategies with stochastic alpha and volume signals have *myopic* closed-form solutions for standard price impact models that only depends on one's current alpha, alpha decay, and volume signals.

- We give concrete bounds for volume curves and predictions for the absence of price manipulation strategies. For instance, for fitted square-root price impact models, with all other things being equal, (il)liquidity that doubles within 15 minutes imply price manipulation opportunities.

Accounting for the frictions incurred during derivatives hedging is another popular research question in mathematical finance. In option hedging, as the required position behaves like a diffusion process, transaction cost takes over as the primary minimization objective instead of price impact. With the celebrated frictionless option pricing paper Black and Scholes (1973); Merton (1973) as a basis, Leland (1985) incorporated proportional transaction costs and derived the optimal hedging strategy as well as the corresponding option price. Chapter 3 provides an actionable derivatives hedging strategy

with both market and limit orders from the perspective of a central risk book, where the risks of several trading desks are managed collectively. Limit orders reduce transaction costs, as one earns the half-spread instead of being charged, but bring along the possibility of being picked-off. This chapter has three essential, quantifiable, and reproducible messages:

- Limit order is only beneficial for delta-hedging in the case of negative gamma.
- The nontrivial tradeoff between transaction cost, adverse selection and risk aversion can be characterized by a nonlinear PDE describing the option price.
- According to empirical analysis, tactical liquidity provision is beneficial for non-competitive market makers for reasonable trading frequencies, as per Figure 1.1.

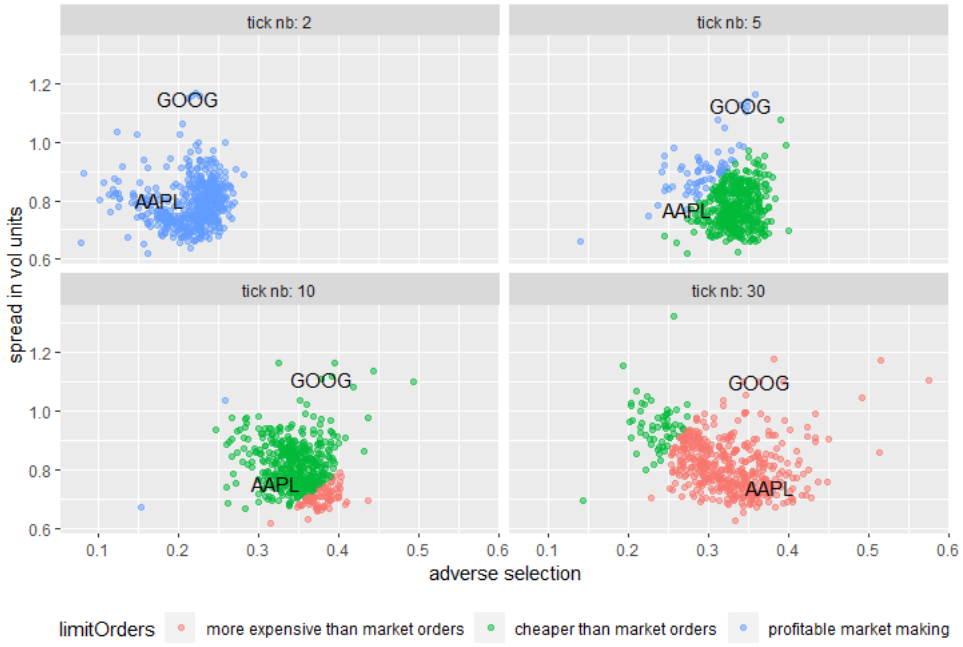

Figure 1.1: Scatter plot of the key variables for limit order profitability: adverse selection and spread. To normalize the spread across the stock universe, one rescales it by the stock's volatility.

The inclusion of other order types in traditional order execution task is also an important research direction of mathematical finance. Different from Chapter 2, Chapter 4 studies the usage of display limit orders and non-display limit orders (hidden orders).

> "The hidden order ratio reveals the extent to which on-exchange trade executions occur in a less than fully transparent fashion due to the participant's preferential use of hidden orders. As market dynamics continue to evolve, changes (or lack thereof) in the hidden order ratio can signal changes in the level of pre-trade transparency on exchanges." (US Securities and Exchange Commission, 2019)

As suggested by the U.S. Securities and Exchange Commission, the nondisplay limit order is a critical type of order that reflects the pre-trade transparency on exchanges. The hidden volume ratio has gradually increased from 10% at 2013 to 28% at 2022, as supported by Figure 1.2. Over all exchanges, it has grown from 9% in 2012 to 15% to 2022(CBOE Insights, 2022). Therefore, it is natural to conclude that hidden order has become an important instrument for on-exchange trading, and should be considered for liquidation tasks.

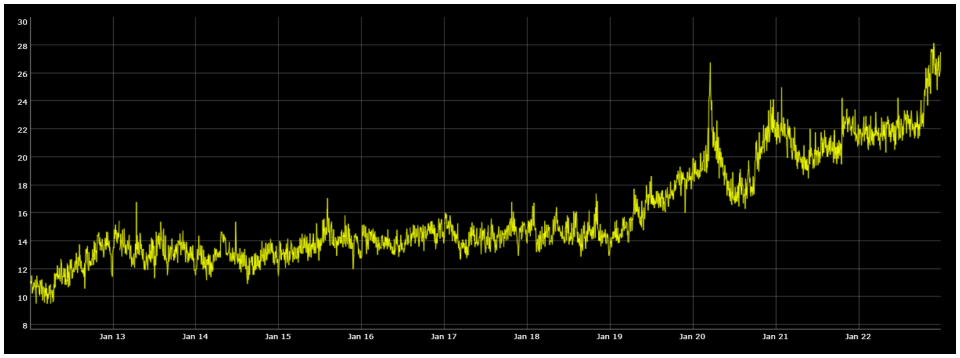

Figure 1.2: Hidden volume ratio - the percentage of the total volume of trades against hidden orders divided by the total volume of all trades. The hidden volume ratio of NASDAQ equities market from 2013 to 2022 is provided by the data visualization tool at `https://www.sec.gov/marketstructure/datavis/ma_exchange_hiddenvolume.html#.ZCKvbHbP2K1`.

Chapter 4 analyzes the passive order perspective of order execution problem. In particular, we postulate and estimate the price impact structure of both limit and nondisplay (hidden) orders, and derive and test the corresponding scheduling algorithm for order execution. Broadly speaking, there are three main takeaways:

- The scheduling algorithm designed for limit and hidden orders is different from the one in R. Almgren and Chriss (2001) due to the structural difference of price

impact for passive orders: The impact scales with the size of the order standing on the book, instead of the size of order executed.

- In the case where hidden order is used, there is a time which separates the trading horizon into two regimes: the former only uses hidden order, and the latter uses the mixture of limit and hidden orders.

- The effectiveness and robustness of the algorithm using combination of hidden and limit orders while incorporating the self-exciting effect of market order is shown via numerical testing in both simulated data and NASDAQ 100 Index data.

Chapter 2

Stochastic Liquidity as a Proxy for Nonlinear Price Impact

This chapter is a joint work with Johannes Muhle-Karbe and Kevin Webster.

2.1 Introduction

Optimal execution and trading algorithms rely on price impact models, like the propagator model, to quantify trading costs. Empirically, price impact is concave in trade sizes, leading to nonlinear models for which optimization problems are intractable and even qualitative properties such as price manipulation are poorly understood. However, we show that in the diffusion limit of small and frequent orders, the nonlinear model converges to a tractable linear model. In this high-frequency limit, a *stochastic* liquidity parameter approximates the original impact function's nonlinearity. We illustrate the approximation's practical performance using limit-order data.

> "Market liquidity – the ability to rapidly execute large financial transactions with a limited price impact – is a key feature of financial market efficiency and functioning." (Committee on the Global Financial System, 2014)

As succinctly summarized by the Committee on the Global Financial System, the price impact of trades is an essential element of financial markets. For instance, price impact

models are used to optimize trade execution (R. Almgren & Chriss, 2001), analyze transaction costs (Kolm & Westray, 2021), and backtest novel trading strategies (Waelbroeck, Federspiel, Marchini, & Gomes, 2013). Therefore, two central challenges in trading are (i) how to measure price impact empirically and (ii) how to incorporate price impact into trading algorithms.

Both challenges have given rise to large and active – but somewhat separate – literatures. Indeed, the empirical literature consistently finds that price impact depends on trade size in a concave, nonlinear manner, both when estimated from proprietary trade data and the public tape (cf. R. Almgren, Thum, Hauptmann, and Li (2005); J. Bouchaud et al. (2004); Frazzini, Israel, and Moskowitz (2018); Gabaix, Gopikrishnan, Plerou, and Stanley (2006); Hasbrouck (1991); Hasbrouck and Seppi (2001); Loeb (1983) and the references therein). In particular, the evolution of the aggregate price impact can be described in a parsimonious yet flexible manner by the "propagator model" of J. Bouchaud, Gefen, Potters, and Wyart (2004): each trade causes an immediate impact, and a decay kernel in turn describes how this effect "propagates" across time.[1]

The empirical prowess of concave propagator models has been reproduced in dozens of studies, as reviewed by J. Bouchaud, Farmer, and Lillo (2009). However, such models are notoriously intractable from an analytical point of view. Consequently, the optimal control literature primarily focuses on linear price impact. An impressive toolbox has emerged for linear models, covering optimal trading (Lehalle & Neuman, 2019; Lorenz & Schied, 2013; Min, Maglaras, & Moallemi, 2022; Neuman & Voß, 2020; Neuman & Voß, 2022; Obizhaeva & Wang, 2013), absence of price manipulation (Fruth, Schöneborn, & Urusov, 2013, 2019), and mean-field games (Fu, Horst, & Xia, 2022; Neuman & Voß, 2021; Schied, Strehle, & Zhang, 2017). Yet, all these results crucially rely on the linearity of the price impact dynamics. In contrast, the analysis of nonlinear propagator models remains limited to numerical studies (T. Chen, Ludkovski, & Voß, 2022; Curato, Gatheral, &

[1] Regressions of price changes on lagged values of the order flow as in Hasbrouck (1991) are an early predecessor of this model class. One mechanism that generates this "impact decay" are the inventory costs of dealers, which can only gradually externalise their positions (Bank, Ekren, & Muhle-Karbe, 2021; Gârleanu & Pedersen, 2016). The same effect also appears when new information arrives continuously over time, but is only gradually incorporated into prices through trading (Bernhardt, Seiler, & Taub, 2010; Vodret, Mastromatteo, Tóth, & Benzaquen, 2021).

Lillo, 2017).

This chapter builds a bridge between these two strands of literature. Indeed, using a functional central limit theorem of Jacod and Protter (2012), we show that linear continuous-time models approximate nonlinear discrete-time propagator models. The approximation applies in the realistic scaling limit of diffusive order flow with a nontrivial quadratic variation. The discrete impact function's *nonlinearity* does not disappear in the high-frequency limit, but instead manifests itself through a *stochastic* price impact parameter, which fluctuates with the volatility of the aggregate order flow. In this way, our limit theorem links the unobservable illiquidity parameter to measured market activity as proxied, e.g., by moving averages of trading volume.

The reduced-form linear limit brings to bear the potent tools developed for such linear impact models and links their results to market activity. For instance, we show that the no price manipulation condition of Fruth et al. (2019) translates into an upper bound for the percentage changes of market activity, which is straightforward to measure empirically. Using public trading data alone, this flags periods to investigate for suspicious trading activity.

As another application, we discuss what market flow dynamics lead to execution prices for which it is difficult to make systematic trading profits. With linear permanent price impact, this is equivalent to order flow being (close to) a martingale as in Kyle (1985) but in stark contrast to the strong persistence observed empirically (Lo & Wang, 2010). When impact decays over time, the execution price's martingale property is instead tantamount to a trend-following order flow, where autocorrelation of the order flow offsets impact decay, compare J. Bouchaud et al. (2004). In this regime, non-trivial flow can create significant price deviations away from fundamental prices as in Black (1986), without leading to statistical arbitrage opportunities for strategic agents.

To demonstrate these analytical results' practical application, we complement them with an empirical study. To wit, we compare the performance of the model with linear but stochastic price impact motivated by our limit theorem to alternative models, including the non-linear propagator model. Our stochastic model does not introduce

additional free parameters: the realized trading volume pins down the model. Despite this parsimony, our stochastic model markedly improves linear price impact's explanatory power over a standard constant parameter model, halving the gap between concave and linear propagator models. The model grounded in our limit theorem also outperforms alternative dynamic models, e.g., based on normalizing with the limit order book depth or by refitting the model on different time bins (Cont, Kukanov, & Stoikov, 2014).

From a practical point of view, this chapter has three broad takeaways:

(a) Nonlinear price impact models can be described in reduced form by linear models, whose stochastic price impact parameter ("Kyle's lambda") mechanically links illiquidity and market activity, measured by moving averages of the order flow or more general volume curves.

(b) Optimal trading strategies with stochastic alpha and volume signals have closed-form solutions for standard price impact models, including the square-root propagator model. The corresponding optimal trading strategy is *myopic:* one derives the optimal trading target considering only one's current alpha, alpha decay, and volume signals.

(c) To rule out price manipulation strategies, volume curves and predictions must satisfy concrete bounds. For instance, for fitted square-root price impact models, volume predictions that – ceteris paribus – double within 15 minutes imply price manipulation opportunities.

The remainder of this chapter is organized as follows. Section 2.2 recalls the discrete nonlinear propagator models of J. Bouchaud et al. (2004). Section 2.3 introduces our continuous-time scaling limit and the corresponding limit theorem. Section 2.4 describes optimal trading strategies in the limiting model, the absence of price manipulation, and order-flow dynamics that lead to martingale execution prices. Finally, Section 2.5 contains our empirical case study. For better readability, all proofs are collected in Appendix 2.7.

Notation

Throughout this chapter, we fix a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{P})$ and denote by $\mathcal{L}^2$ the adapted processes $(X_t)_{t\in[0,T]}$ satisfying $\mathbb{E}[\int_0^T |X_t|^2 dt] < \infty$. We first consider trading on a discrete, equidistant time grid $t_n^N = n\Delta t^N$ with mesh width $\Delta t^N = T/N$. Then, we pass to the continuous-time limit $N \to \infty$ in a second step. The samples of any continuous-time process $(X_t)_{t\in[0,T]}$ along this grid and the corresponding discrete differences are denoted by

$$X_n^N = X_{t_n^N} \quad \text{and} \quad \Delta X_n^N = X_n^N - X_{n-1}^N.$$

2.2 Discrete-Time Microstructure Model

We first recall the *propagator model* of J. Bouchaud et al. (2004), which describes on a discrete microstructure level how each trade's price impact propagates across time.

To ease notation, we focus on a financial market with two assets. The first one is safe, with the interest rate normalized to zero for simplicity. The second asset is risky. A continuous semimartingale S models the asset's *fundamental* price: S describes price changes due to exogenous factors such as news rather than trading. Another continuous semimartingale F models the *(signed) order flow* F. Finally, a price impact model describes how the order flow shifts the actual market price relative to its fundamental level. In the propagator model, each trade has an immediate impact described by a function $g^N(\cdot)$. This initial price shift dissipates over time according to a decay kernel. Here, we focus on the simplest case where impact decays at a constant rate.[2]

Definition 2.2.1. *At the microstructure level,* price impact *has the dynamics*

$$\Delta I_n^N = -\beta I_{n-1}^N \Delta t^N + g^N\left(\Delta F_n^N\right), \tag{2.1}$$

[2]The empirical literature typically finds kernels of power form (e.g., J. Bouchaud, Kockelkoren, and Potters (2006)). However, our fitting results in Section 2.5 suggest that the choice of the decay rate within the exponential class only has a second-order effect compared to the form of the price impact function. Therefore, we do not pursue such more sophisticated but less tractable model extensions here.

for some odd function $g^N \in C^1$ *that is concave on* $[0, \infty)$.

As the function g^N is odd, the price impact of purchases and sales is symmetric. Concavity means that trading twice the amount at most doubles the price impact. The simplest case is *linear* impact, which corresponds to the model of Obizhaeva and Wang (2013). However, for empirical data, one improves the model's fit by passing to a concave parametric class of functions, compare Section 2.5. The literature fits price impact models on proprietary order data (e.g., R. Almgren et al. (2005); Bershova and Rakhlin (2013); Bucci, Benzaquen, Lillo, and Bouchaud (2015); Frazzini et al. (2018)) and public trading data (e.g., J. Bouchaud et al. (2004); Y. Chen, Horst, and Hai Tran (2019); Cont et al. (2014); Loeb (1983)). For a review and comparison of price impact estimations across proprietary data, see Zarinelli, Treccani, Farmer, and Lillo (2015).

The shape of the function g at the microstructure level is a particular focus in three papers:

(a) In the original propagator model of J. Bouchaud et al. (2004), $g^N(x) \propto \log(x)$.

(b) J. Bouchaud et al. (2009) review empirical results around the non-linear behavior of price impact at the fill level. These find impact functions of power form $g^N(x) \propto x^p$ with exponents $p \in [0.2, 0.5]$.

(c) For a database of institutional trades, Bucci, Benzaquen, Lillo, and Bouchaud (2019) document a crossover between a linear market impact regime and a square-root regime as a function of the order's size, i.e., $g^N(x) \propto x$ for small x and $g^N(x) \propto x^{1/2}$ for larger x.

2.3 Scaling Limit

We now consider the continuous-time scaling limit of the impact dynamics (2.1) in the regime where orders are small, frequent, *and* have nontrivial quadratic variation. Figure 2.1 illustrates this diffusive nature of the aggregate order flow for empirical data.

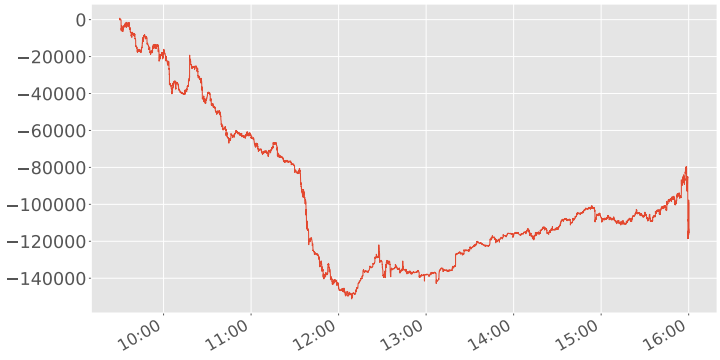

Figure 2.1: Cumulative signed order flow for PayPal on January 7, 2019.

Carmona and Webster (2019); Webster et al. (2015) performed formal statistical tests confirming diffusivity of the flow on Nasdaq's public trading tape. Carmona and Leal (2021) carried out formal tests on the Toronto stock exchange and highlight the diffusivity of individual traders' flow.[3]

We model the diffusive order flow by an Itô process $F \in \mathcal{L}^2$ with dynamics

$$dF_t = \mu_t^F dt + v_t dW_t.$$

Here W is a Brownian motion, the drift rate μ^F is a locally bounded càdlàg process, and the *positive* volatility process v has Itô dynamics

$$dv_t = a_t dt + \eta_t dW_t + b_t dW_t^\perp,$$

for a Brownian motion $W^\perp$ independent of W and locally bounded, càdlàg coefficients a, η, b.

Remark 2.3.1. *The order flow's volatility v corresponds to what practitioners call* volume curves. *For example, a widespread practice is to measure the sum $\sum_n |\Delta F_n^N|$ of unsigned volumes over some time interval. Brokers measure and predict such sums in VWAP (Volume Weighted Average Price) algorithms. Moreover, the daily volume traded on the stock is computed in the same manner.*

[3]The Toronto and Australian stock exchanges are examples where trading is not anonymous, enabling Carmona and Leal (2021) to analyze traders separately rather than in the aggregate. Our central limit theorem 2.3.3 only requires aggregate diffusivity. Our optimal trading strategies from Lemma 2.4.1 imply individual difusivity for statistical arbitrageurs with stochastic signals.

By the functional law of large numbers (Jacod & Protter, 2012, Theorem 7.2.2),

$$\Delta t^N \sum_n \left| \tfrac{1}{\sqrt{\Delta t^N}} \Delta F_n^N \right| \to \sqrt{\tfrac{2}{\pi}} \int v_t dt,$$

uniformly on compacts in probability. Therefore, we refer to v as market activity*, as it is simply a (renormalized) estimate of unsigned market volumes. Practitioners commonly predict the market activity v in volume curves. Such models could be deterministic or include stochastic factors, but are always time-varying: market activity displays a pronounced time-of-day effect.*

Suppose the impact function $g^N(x)$ in (2.1) is linear as in Obizhaeva and Wang (2013). Then, the discrete price impact model has a well-defined continuous-time limit: an Ornstein-Uhlenbeck process driven by the order flow. To obtain a nontrivial, finite limit for diffusive order flow also with nonlinear impact, we consider impact functions that scale as follows:

$$g^N(x) = \sqrt{\Delta t^N} g\left(\tfrac{1}{\sqrt{\Delta t^N}} x\right), \tag{2.2}$$

for a universal odd function $g \in C^1$ that is concave on $[0, \infty)$.[4]

Remark 2.3.2. *Two different perspectives motivate the scaling* (2.2)*:*

(a) As the order flow observed empirically has a nontrivial quadratic variation, any continuous-time limiting model should be compatible with this. Via the functional central limit theorems of Jacod and Protter (2012), this naturally leads to the rescaling (2.2)*.*

Alternatively, one could also model the price impact of rough *trades (say, the overall market) and* smooth *trades (say, the trades of a given agent liquidating a sizeable position) with separate price impact models. However, this conflicts with the "anonymity assumption", which states that the same functional form applies to all trades, regardless of origin. Empirical support for this is provided by Tóth et al. (2011).*

[4]Here we note that the dynamics of order flow F does not depend on N but the impact I does via the propagator model (2.1).

(b) *From a microstructure perspective, one can derive price impact's nonlinearity from a limit order book model. Obizhaeva and Wang (2013) deduce linear price impact from a* block-shaped *order book. Carmona and Webster (2019) derive nonlinear price impact from a general order book shape.If one wants to keep the "distance" between the bid and ask sides of the order book proportional to price volatility (in line with related work on bid-ask spreads such as Wyart, Bouchaud, Kockelkoren, Potters, and Vettorazzo (2008)), then the scaling for a large number of trades again must be of the form* (2.2).

We now state our main result, which shows that the continuous-time limit of the nonlinear impact dynamics (2.1) with the rescaling (2.2) is a *linear* price impact model.[5] Appendix 2.7.1 outlines how Theorem 2.3.3 follows from the functional central limit theorem (Jacod & Protter, 2012, Theorem 10.3.2(a)), and also explains how the linearization can be formally derived from a Taylor expansion in the simplest case of iid order flow.

Theorem 2.3.3. *As the number N of timesteps tends to infinity, the impact I^N in the nonlinear propagator model* (2.1) *converges (stably in law) to the* linear *impact model*

$$dI_t = -\beta I_t dt + \lambda_t dF_t + \eta_t \chi_t dt + \theta_t dW'_t. \tag{2.3}$$

Here,

$$\lambda_t = \int_{-\infty}^{\infty} g'(x)\phi_{v_t}(x)dx, \quad \chi_t = \frac{1}{2}\int_{-\infty}^{\infty} (v_t^{-2}x^2 - 1)g'(x)\phi_{v_t}(x)dx, \quad \theta_t = \sqrt{\lambda_t^2 v_t^2 + \zeta_t^2 - 2\lambda_t h_t},$$

for

$$\zeta_t = \sqrt{\int_{-\infty}^{\infty} g^2(x)\phi_{v_t}(x)dx}, \quad h_t = \int_{-\infty}^{\infty} xg(x)\phi_{v_t}(x)dx.$$

ϕ_v is the density function of a Gaussian law with mean zero and standard deviation $v > 0$,

[5] For i.i.d. order flow, an antecedent of this approximation already appears in J. Bouchaud et al. (2009). Empirical evidence of Patzelt and Bouchaud (2018) supports this for reasonably small trades (in line with our diffusive model for the order flow).

and W' is a Brownian motion whose quadratic covariation with W is

$$d[W, W']_t = \frac{1}{\theta_t}\left(\frac{h_t}{v_t} - \lambda_t v_t\right) dt.$$

Let us briefly discuss the elements of the limiting dynamics (2.3). The first term is the exponential decay kernel directly implied by the linear mean-reversion of the discrete impact process.

The second term is central: the continuous-time limit of the *nonlinear* propagator model (2.1) has *linear* price impact. However, the nonlinearity of the discrete price impact function does *not* disappear in the continuous-time limit. Instead, it is incorporated through a *stochastic* illiquidity parameter λ, which fluctuates with the stochastic volatility v of the order flow F. Collin-Dufresne and Fos (2016) derive a similar link between the volatility of trading volumes and stochastic liquidity in an extension of the Kyle model.

The third term in (2.3) is an additional drift introduced by the concavity of the price impact function. If the volatility of the order flow is smooth (e.g., a deterministic function of time of day only), then this term disappears. However, if the order flow's volatility follows a square-root process, for example, then this drift rate is a constant. We have evaluated this more general version of the limiting model empirically; adding an intercept term to the linear price impact model only marginally improves the model fit, suggesting that this term is only of secondary importance.

Finally, the last term in (2.3) is additional noise typical for functional central limit theorems.

Remark 2.3.4. *Theorem 2.3.3 mechanically links the market activity v with liquidity. Indeed, the price impact parameter λ, sometimes referred to as Kyle's lambda, measures price illiquidity by quantifying how severely prices react to trades and drive up trading costs. Our representation*

$$\lambda_t = \int_{-\infty}^{\infty} g'(x)\phi_{v_t}(x)dx$$

establishes a one-to-one relationship between the illiquidity measure λ and the market

activity v. As a concrete example, if $g(x) \propto x^{1/2}$,[6] then $\lambda_t \propto v_t^{-1/2}$. In this case, doubling the market activity reduces illiquidity by about 30%. In contrast, in the standard Obizhaeva and Wang model, doubling the market activity does not decrease price illiquidity as measured by Kyle's lambda.

2.4 Trading Applications of the Reduced-Form Model

So far, we have analyzed how a *given* diffusive order flow impacts prices. Now, we outline how the limiting dynamics' linearity supports closed-form expressions for optimal trading strategies. This leverages a rich body of work in stochastic control (Ackermann, Kruse, & Urusov, 2021; Alfonsi, Fruth, & Schied, 2010; Becherer, Bilarev, & Frentrup, 2018; Fruth et al., 2019; Fu et al., 2022; Gârleanu & Pedersen, 2016; Graewe & Horst, 2017; Mastromatteo, Benzaquen, Eisler, & Bouchaud, 2017; Min et al., 2022; Neuman & Voß, 2020; Obizhaeva & Wang, 2013; Schied et al., 2017). For the linear limiting price impact model, results of this kind are essentially known. However, our limit theorem shows that the closed-form optimal trading strategies in linear models with suitable "effective" stochastic liquidity are also approximately optimal for nonlinear propagator models. Moreover, this interpretation naturally links optimal trading to the market flow's activity. Our result thereby sheds new light on conditions for the absence of price manipulation. We also derive constraints on *market flow dynamics* that rule out simple trading profits for unconstrained arbitrageurs.

2.4.1 External Inputs and Goal Functional

We now specify the optimization problem of a trader choosing their order flow F. The external inputs are the fundamental price S of the risky asset and the *external* order flow $\bar{F}$ of other market participants. For simplicity, we assume that S and $\bar{F}$ are both Itô

[6] Of course, one cannot directly apply Theorem 2.3.3 to a square-root impact function $g(x) = c\ \mathrm{sgn}(x)\sqrt{|x|}$ due to a lack of differentiability at the origin. However, the limit theorem applies for the smoothed version $g_\varepsilon(x) = \frac{x}{|x|+\varepsilon}\sqrt{|x|} + x\frac{\varepsilon}{|x|+\varepsilon}$, which in fact is consistent with the crossover from linear to square root impact documented in Bucci et al. (2019). Then, the resulting formulas remain well behaved when the regularization parameter ε goes to zero.

processes whose drift rates belong to $\mathcal{L}^2$ and whose volatilities are uniformly bounded and bounded away from zero. Moreover, the external flow's volatility is smooth.

We suppose that the trader we consider has a small influence on the overall order flow's volatility. For the linear dynamics of the aggregate price impact motivated by Theorem 2.3.3,

$$dI_t^\Sigma = -\beta I_t^\Sigma dt + \lambda_t d(F_t + \bar{F}_t) + dM_t, \quad I_0^\Sigma = 0,$$

this means that the external order flow pins down the stochastic liquidity parameter λ and the continuous martingale M. Therefore, the trader treats both as exogenous processes in their optimization problem.[7] This assumption is justified, for example, if the trader's own order flow is smooth, so that it does not contribute to the overall order-flow volatility. Alternatively, this regime also applies in a mean-field setting with many agents where each makes a small contribution to the overall trading volume. Finally, a third example arises in the – realistic – regime when fundamental prices and external impact are close to a martingale, so that the trader's profit opportunities are small relative to the trading activity in the overall market.

Crucially, assuming the impact parameters to be exogenous does *not* imply that the trader's order flow has no price impact. Instead, by linearity, one decomposes the aggregate price impact into the *external impact*

$$d\bar{I}_t = -\beta \bar{I}_t dt + \lambda_t d\bar{F}_t + dM_t, \quad \bar{I}_0 = 0,$$

and the trader's *own impact*

$$dI_t = d(I_t^\Sigma - \bar{I}_t) = -\beta(I_t^\Sigma - \bar{I}_t)dt + \lambda_t dF_t = -\beta I_t dt + \lambda_t dF_t, \quad I_0 = 0.$$

If the trader values the terminal position F_T at the execution price $S_T + \bar{I}_T$ in the absence

[7] Note that the third term in (2.3) is zero here because we have assumed the order flow's volatility to be smooth.

of the trader's own impact,[8] then the corresponding cash position in discrete time is

$$F_N^N(S_N^N + \bar{I}_N^N) - \sum_{n=1}^{N} \left(S_{n-1}^N + \bar{I}_{n-1}^N + I_{n-1}^N + \Delta S_n^N + \Delta \bar{I}_n^N + \tfrac{1}{2}\Delta I_n^N\right) \Delta F_n^N.$$

In this budget equation, trades are settled after the fundamental price has moved, and at the midpoint between the price before and after the trader's own execution. Regarding the contemporaneous execution of external orders, we assume the trade incurs these orders' entire impact. This rules out our trader consistently executing right before or simultaneously to other market participants – a *no-frontrunning* assumption. In continuous time, the trader's P&L converges to

$$F_T(S_T + \bar{I}_T) - \int_0^T (S_{t-} + \bar{I}_{t-} + I_{t-}) dF_t - [S + \bar{I} + \tfrac{1}{2} I, F]_T.$$

Integration by parts applied to $F_T(S_T + \bar{I}_T)$ in turn shows that the trader's expected P&L equals

$$\mathbb{E}\big[\int_0^T F_{t-} d(S_t + \bar{I}_t) - \int_0^T I_{t-} dF_t - [\tfrac{1}{2} I, F]_T\big]. \tag{2.4}$$

2.4.2 Optimal Trading with a Predictive Signal

As the liquidity parameter λ is bounded, bounded away from zero and smooth, one can write it as

$$\lambda_t = e^{\varphi_t},$$

where φ is differentiable with bounded derivative φ'. Risk-neutral optimal trading problems like (2.4) are highly tractable even with stochastic liquidity parameters. This becomes particularly clear after switching the trader's control variable from risky positions to the corresponding impact. Then, the optimization problem becomes *myopic* and can be solved directly by pointwise maximization. This observation first appears in the optimal execution results of Ackermann et al. (2021); Fruth et al. (2019), but also applies if

[8] Alternatively, one could obtain equivalent results if the terminal position is valued at the fundamental value S_T. The crucial point is that the trader's own impact should not enter the terminal valuation to avoid manipulation of the final mark-to-market price.

the trader has a predictive signal about future price changes ("alpha"):

Lemma 2.4.1. *Suppose the* trading signal

$$\alpha_t = \mathbb{E}_t[S_T + \bar{I}_T - S_t - \bar{I}_t]$$

is an Itô process with bounded volatility and denote its drift rate by μ^α. *Then, the goal functional* (2.4) *is equivalent to a* myopic *problem in impact space:*

$$\mathbb{E}\left[\int_0^T e^{-\varphi_t}\left(-I_t\mu_t^\alpha + (\beta+\varphi_t')\alpha_t I_t - \tfrac{2\beta+\varphi_t'}{2} I_t^2\right) dt - \frac{1}{2}e^{-\varphi_T} I_T^2\right]. \tag{2.5}$$

The corresponding holdings can be recovered from $F_t = F_0 + \int_0^t \frac{1}{\lambda}_s dI_s + \int_0^t \frac{\beta}{\lambda_s} I_s ds$. *Under the no price manipulation condition* $2\beta + \varphi_t' > 0$, *the* optimal *impact is*

$$I_t^* = \tfrac{\beta+\varphi_t'}{2\beta+\varphi_t'}\alpha_t - \tfrac{1}{2\beta+\varphi_t'}\mu_t^\alpha, \quad t \in [0,T), \quad I_T^* = 0. \tag{2.6}$$

The myopic relationship between impact, alpha, and liquidity parameters in Equation (2.6) massively simplifies the implementation of optimal trading strategies:

Example 2.4.2 (Fast-decaying impact)**.** *Assume* $\varphi_t' \ll \beta$*: price impact decays much faster than liquidity varies.*[9] *In this regime, the myopic optimality equation* (2.6) *simplifies to*

$$I_t^* = \frac{1}{2}\left(\alpha_t - \beta^{-1}\mu_t^\alpha\right).$$

Therefore, the optimal impact state depends only on (a) the trader's alpha level α_t*, (b) the trader's* alpha decay $-\mu_t^\alpha$*, and (c) price impact's decay* β. *Specifically, the trader simply targets an impact state equal to* half *their alpha level and modulates that impact state based on their alpha's decay.*[10]

Gârleanu and Pedersen (2016) highlight alpha decay's key role for optimal trading strategies:

[9] In practice, the heuristic $\varphi_t' \ll \beta$ is verified for stocks during most of the trading day: the only time it occasionally fails is near the close, when liquidity may sharply rise.

[10] In this regime, the optimal impact state doesn't depend on Kyle's lambda anymore: λ is only used to translate the optimal impact I^* into a trading strategy F^*.

> "The alpha decay is important because it determines how long the investor can enjoy high expected returns and, therefore, affects the trade-off between returns and transactions costs."

Remark 2.4.3 (Non-parametric models)**.** *Today, trading desks typically construct predictive models using non-parametric methods, such as machine learning. The stochastic model variables such as α_t, and λ_t then do not have simple explicit forms. For example, rather than pinning the price dynamics down using an SDE, practitioners compute Bayesian estimators*

$$\alpha_t = \mathbb{E}\left[S_T - S_t \middle| \mathcal{F}_t\right]$$

with a wide array of non-parametric models. This is in stark contrast to traditional control models, where price dynamics and model ingredients usually have explicit forms such as Bachelier or Ornstein-Uhlenbeck dynamics.

By reducing optimization to a myopic problem, Lemma 2.4.1 allows to deal with general non-parametric signals, such as those coming from neural networks, without additional effort. The closed-form solutions for general α_t and λ_t allow one to focus on fitting effective models for alpha, alpha decay and volume, without having to use numerical optimizers to turn these forecasts into an optimal trading strategy.

2.4.3 Absence of Price Manipulation

The wellposedness condition $2\beta + \varphi_t' > 0$ in Lemma 2.4.1 first appears in Fruth et al. (2013, 2019), who point out that:

> "Time-dependent liquidity can potentially lead to price manipulation. In periods of low liquidity, a trader could buy the asset and push market prices up significantly; in a subsequent period of higher liquidity, he might be able to unwind this long position without depressing market prices to their original level, leaving the trader with a profit after such a round trip trade."

To see this, suppose a trader buys Δ_t shares in a bulk trade at time t and sells them in another bulk trade at time $t' > t$. Then, the corresponding trading profits (2.4.1) are

$$\begin{aligned} &- \left(S_t + \tfrac{1}{2}\lambda_t \Delta_t\right) \Delta_t + \left(S_{t'} + e^{-\beta(t'-t)} \lambda_t \Delta_t - \tfrac{1}{2}\lambda_{t'} \Delta_t\right) \Delta_t \\ &= (S_{t'} - S_t)\Delta_t + \tfrac{1}{2}\lambda_t \left(2e^{-\beta(t'-t)} - 1 - e^{\int_t^{t'} \varphi'_s ds}\right) \Delta_t^2. \end{aligned}$$

For sufficiently large trades, the quadratic impact term dominates the linear gains from expected fundamental price changes. For small $t' - t$, the series representation of the exponential shows that this quadratic term is positive if $2\beta + \varphi'_t < 0$. Consequently, if illiquidity decreases too promptly relative to impact decay, then large round trips executed quickly allow the trader to obtain arbitrarily large expected profits. These profits do not exploit any signals about fundamental price changes, but solely rely on the price manipulation through the trader's own impact.

Remark 2.4.4. *Our limit theorem 2.3.3 linking price illiquidity λ_t and market activity v_t makes this trading intuition even more concrete. For example, consider the square-root model $g(x) \propto \sqrt{x}$.*[11] *Under that model, $e^{\varphi_t} = \lambda_t \propto v_t^{-1/2}$ and in turn $2\varphi'_t = -v'_t/v_t$. Therefore, in this model, the no price manipulation condition in Lemma 2.4.1 is an upper bound on percentage changes v'/v in terms of impact decay β: $v'_t/v_t < 4\beta$. Integration of this bound shows that, for a square-root model and ceteris paribus, trading activity cannot double in less than a quarter of the impact model's half-life $\log(2)/\beta$. Concretely, for a price impact with a (realistic) one hour half-life, our no-price manipulation condition states that, under the square-root impact model, trading activity cannot double in less than 15 minutes without creating a price manipulation opportunity.*

Note that price manipulation is asymmetric in liquidity. Negative shocks to market activity, where trading and, in turn, liquidity quickly dries up, do not lead to price manipulation. In contrast, positive liquidity shocks, where trading liquidity quickly floods the market, potentially invite price manipulation.

[11] Note that this discussion focuses on the time scales on which the reduced-form linear limiting model is an effective approximation. At more more granular levels, price manipulation with nonlinear propagator dynamics is more delicate .

Relating Fruth, Schöneborn, and Urusov's no-price manipulation condition to market activity is not merely a theoretical result. Market activity is a straightforward variable to measure, forecast, and monitor: banks provide *volume curves* that model the evolution of v_t over time for execution purposes. Therefore, such models can determine how sensitive a stock is to price manipulation. For example, one could determine, at least ex-post, whether liquidity rose fast enough to make liquidity-based price manipulation possible. Of course, one would still need to prove that a suspicious trader anticipated or caused the liquidity shock ahead of time and traded accordingly, pushing up the price when liquidity was low and exiting their position when liquidity rapidly increased. But this statistic can still flag, e.g., to regulators, what periods should be investigated more closely to identify suspicious trading behavior.

2.4.4 Martingale Flow vs. Martingale Impact

We now consider another novel application of the optimal trading strategies from Lemma 2.4.1 combined with the Limit Theorem 2.3.3. To wit, we discuss how these results identify order-flow dynamics compatible with execution prices that are (almost) martingales, corresponding to (almost) efficient markets where it is difficult to make systematic trading profits.

In models with linear *permanent* price impact such as the Kyle model, martingale order flow directly translates into martingale impact and in turn martingale execution prices. In contrast, if impact mean reverts to zero as in (2.3), then martingale order flow creates a predictive trading signal for observant traders. To wit, even if the fundamental price S is a martingale, the trading signal α in Lemma 2.4.1 is not zero but, instead,

$$\alpha_t = (e^{-\beta(T-t)} - 1)\bar{I}_t, \quad \text{if the external flow is a martingale.}$$

Consequently, $\mu_t^\alpha = \beta \bar{I}_t$, so that the optimal impact from Lemma 2.4.1 is

$$I_t^* = \frac{\beta + \varphi_t'}{2\beta + \varphi_t'}\alpha_t - \frac{1}{2\beta + \varphi_t'}\mu_t^\alpha = -\bar{I}_t + \frac{\beta + \varphi_t'}{2\beta + \varphi_t'}e^{-\beta(T-t)}\bar{I}_t, \quad t \in [0, T), \quad \text{and } I_T^* = 0.$$

Whence, if the time horizon $T - t$ is relatively long compared to impact decay β, then a risk-neutral trader completely offsets any deviation $\bar{I}$ from the fundamental value caused by external noise trading. In particular, there is no aggregate trading activity in this case, as the risk-neutral arbitrageur is happy to provide all liquidity the noise traders demand.

Lemma 2.4.1 suggests that the trading price $S + \bar{I}$ needs to be close to a martingale to rule out such simple profit opportunities. Therefore, with impact decay, external *impact* $\bar{I}$ rather than external *flow* should (approximately) have martingale dynamics. The linear limiting model (2.3.3) reveals that order-flow dynamics compatible with this are feedback functions of the current impact state.[12] To wit, if the external flow has Itô dynamics $d\bar{F}_t = \mu_t^{\bar{F}} dt + v_t dW_t$, then we need $0 = -\beta \bar{I}_t + \lambda_t \mu_t^{\bar{F}}$ and in turn $\mu_t^{\bar{F}} = \frac{\beta}{\lambda_t} \bar{I}_t$. The corresponding external impact then has the form $d\bar{I}_t = \lambda_t v_t dW_t + \theta_t dW_t'$. Thus, the volatility v_t of the order flow can be chosen freely, similar to Heath-Jarrow-Morton models for the term structure of interest rates; the corresponding drift rates are in turn pinned down.

In the linear impact model, martingale dynamics for the external impact therefore require "trend-following" order flow. To wit, after aggregate buying (leading to a positive current impact state $\bar{I}_t$), the order flow continues to drift up at rate $\frac{\beta}{\lambda_t} \bar{I}_t$ to offset the impact decay. Our martingale condition for external price impact is a tractable reduced-form model for the observations made by J. Bouchaud et al. (2004), who discuss the links between positively autocorrelated order flow, mean-reverting price impact dynamics, and martingale prices.

[12] An intriguing mathematical question is whether feedback order flow as above yields the same limiting result when directly applied in the discrete, concave propagator model (2.1). Here, the challenge is that the limit theorems of Jacod and Protter (2012) no longer apply because the order flow is not a fixed process anymore but instead depends on the corresponding impact state.

For realistic parameter values, numerical experiments suggest that the feedback motivated by the limiting model still leads to impact with approximate martingale dynamics. A formal proof of a corresponding functional central limit theorem is an intriguing direction for future research.

2.5 Empirical Analysis

As emphasized by the Committee on the Global Financial System (2014), measuring price impact is crucial, e.g., when executing sizable trades and scaling investment strategies. Such measurements are relatively straightforward once given a parametric model. One can then fit the model parameters on publicly available trading tapes, such as NYSE Trade And Quote (TAQ) or Nasdaq Equities Market Data.[13] This chapter leverages LOBSTER data, which presents all Nasdaq limit order book events in a standard table format. In particular, the data includes all public trades and the limit order book state before and after the trade. Our study covers the S&P 500 constituents over the year 2019.

2.5.1 Methodology

In line with Cont et al. (2014), we bin the data using a bin-size of $\Delta t = 10$ seconds. Therefore, we group and aggregate all discrete events into a set number of bins per day. For instance, there are 390 minutes in a trading day, leading to 2340 bins per stock and date and in turn 1.17 million data points per day. Binning the data reduces its size and standardizes the number of data points across days and stocks. The ten-second bin choice presents a sensible trade-off between accuracy and data reduction but is otherwise arbitrary.

We construct three variables of interest from the events:

- the observed midprice P_t at the start of the bin and the return $(P_{t+\Delta t} - P_t)/P_t$ over the bin. We use geometric rather than arithmetic returns to make them unitless; over the time scales we consider, normalizing by the initial price level instead would lead to virtually the same results;
- the unsigned volume $|\Delta F_t|$ traded over the bin;
- the signed volume ΔF_t traded over the bin.

[13] NYSE and Nasdaq provide customers with public trading data, and the historical datasets are helpful for building price impact models. NYSE describes TAQ data on their website (New York Stock Exchange, n.d.). Nasdaq describes the Nasdaq Equities Market Data on their website (NASDAQ, n.d.).

In the following statistical estimations, we fix the decay parameter (in hourly units) to $\beta = 0.7$ so that the half-life of impact is $\log 2/\beta = 60$ minutes. We have run the empirical tests with different values of the decay parameter: a grid-search over β has negligible impact on the results for our prediction horizons ranging from 1 to 120 minutes.[14]

We compute various quantities from the base features, most notably the impact state I_t at the start of the interval and the market activity v_t. The impact state depends on the impact model. As discussed in Remark 2.3.1, market activity v_t can be estimated by moving averages of absolute trading volumes. In our empirical study, we use an exponential moving average with the same half-life as the impact model.[15]

Price impact behaves differently on different time horizons. For a time horizon h, define:

- the horizon-specific return $\Delta_t^h P = (P_{t+h} - P_t)/P_t$.
- the impact return $\Delta_t^h I = I_{t+h} - I_t$. Note that $\Delta_t^h I$ includes *both* the contemporaneous flow effects ΔF_t and impact decay.

To estimate the parameters of the price impact model, we run the regression

$$\Delta^h P = \Delta^h I(\lambda) + \epsilon.$$

over the parameter choice λ. The resulting regressions for our parametric price impact models are all simple linear regressions: $\Delta^h I(\lambda)$ depends linearly on the parameter λ in each case. We are simply fitting price impact's *magnitude*, as measured by λ, given a parametric model choice. We drop the explicit dependence of I on λ in the following sections.

[14]This lack of sensitivity to β over minute to hour prediction horizons may reflect the representative order length in the public trading tape rather than a fundamental property of price impact. Price impact studies focusing on proprietary multi-day orders and predictions find longer time-scales, cf., e.g., Bershova and Rakhlin (2013); Bucci et al. (2015).

[15]Practitioners often use moving averages as straightforward volume predictors. Our exponential choice is one on many: it performs equally well to regular moving averages on timescales above 15 minutes, but somewhat outperforms these on 1 minute timescales. A more sophisticated high frequency implementation could leverage the results of Aït-Sahalia and Jacod (2014) to estimate a joint model for market activity and the stochastic volatility of prices for shorter timescales.

Cont et al. (2014) present their core results for a timescale of $h = \Delta t$ corresponding to ten seconds. However, the bin size Δt and the prediction horizon h generally need not match: Cont et al. (2014) and Cont, Cucuringu, and Zhang (2021) both briefly mention prediction horizons ranging from fifty milliseconds to one hour (cf., e.g., (Cont et al., 2014, Figure 12) and (Cont et al., 2021, Figure 17)). We study prediction horizons of $h = 1$, 15, and 60 minutes.

2.5.2 Models

In all the models below, let σ be a daily volatility estimate for the stock price and ADV an estimate of the daily traded volume of the stock. For a given day, σ and ADV are based on the average daily realized volatility and volume over the past month. As in R. Almgren et al. (2005) and many subsequent studies, these variables normalize the price impact models across stocks and thereby make the impact coefficient λ comparable in the cross section.

The Original OW Model

Y. Chen et al. (2019) fit an Obizhaeva-Wang model with constant parameters:

$$I_{t+\Delta t} - I_t = -\beta I_t \Delta t + \lambda \sigma \frac{\Delta F_t}{\mathrm{ADV}}.$$

The Concave Propagator Model

J. Bouchaud et al. (2009) review empirical studies of concave propagator models, which suggest $g(x) \propto x^c$ with $c \in [0.2, 0.5]$. We consider the most well-known model specification where impact scales with the square root of trade sizes,

$$I_{t+\Delta t} - I_t = -\beta I_t \Delta t + \lambda \sigma \ \mathrm{sgn}(\Delta F_t) \sqrt{\frac{|\Delta F_t|}{\mathrm{ADV}}}.$$

The Reduced Form Model

Our Limit Theorem 2.3.3 suggests a reduced-form version of the propagator model where the square-root impact is proxied by a stochastic liquidity parameter:

$$\lambda_t \propto 1/\sqrt{v_t}.$$

(Here, v_t is the moving average of trading volume introduced above, which measures market activity.) The corresponding linear price impact model is

$$I_{t+\Delta t} - I_t = -\beta I_t \Delta t + \lambda\sigma \frac{\Delta F_t}{\sqrt{\text{ADV } v_t}}.$$

The Depth Model

Cont et al. (2014) propose an alternative model for stochastic liquidity:[16]

$$\lambda_t \propto 1/D_t.$$

Here, D_t is the current "depth" on the best bid and ask queues. The corresponding linear price impact model is

$$I_{t+\Delta t} - I_t = -\beta I_t \Delta t + \lambda\sigma \frac{\Delta F_t}{D_t}.$$

2.5.3 Core empirical results

Comparing impact models across time-scales involves varying the horizon h. To cover a wide range of frequencies of trading strategies, we consider horizons $h \in \{1, 15, 60\}$, measured in *minutes*.

We perform all statistical estimations in this section on a stock-by-stock basis, using a monthly training sample ("in sample"), no regularization, and the subsequent month as a validation sample ("out of sample"). Therefore, in total, we regress each model 11 times per stock, for a total of 5500 regressions per model and horizon. For a given (model,

[16] In Cont et al. (2014), the main focus is on using this factor to normalize the "order flow imbalance", a variable composed of both trades *and* quotes.

horizon) pair, we provide performance statistics averaged across all 5500 regressions.

We define an estimator's *t-stat* as $\text{mean}(\hat{\lambda})/\text{std}(\hat{\lambda})$. Here, each $\hat{\lambda}$ is a realization of the fitted λ for each model on a given (month, stock) pair. The t-stat assesses the model's robustness across both the universe of stocks and the time period. All else being equal, one prefers a model with a high t-stat, as it is more stable. The ability to compute t-stat across stocks, and therefore, measure model stability across stocks, is made feasible by the model normalizations using σ and ADV. Indeed, without those normalizations, λ would not be unitless and comparable across stocks.

Table 2.1 summarizes and compares the performance of each model on the horizons $h \in \{1, 15, 60\}$.

Price Impact Model	In-sample R^2	Out-of-sample R^2	t-stat
Propagator	19%	18%	6.2
Reduced form	15%	14%	3.7
Depth	13%	11%	1.1
OW	11%	10%	2.7

(a) Performance of various price impact models for $h = 1$ minute.

Price Impact Model	In-sample R^2	Out-of-sample R^2	t-stat
Propagator	14%	13%	3.1
Reduced form	11%	9%	2.1
Depth	10%	6%	0.8
OW	8%	6%	1.9

(b) Performance of various price impact models for $h = 15$ minutes.

Price Impact Model	In-sample R^2	Out-of-sample R^2	t-stat
Propagator	10%	8%	2.0
Reduced form	8%	5%	1.5
Depth	7%	2%	0.7
OW	7%	4%	1.3

(c) Performance of various price impact models for $h = 60$ minutes.

Table 2.1: Performance (averaged across the S&P 500 and the year 2019) of various price impact models across horizons h fitted on training samples of one month over the course of 2019. In each case, the half-life of impact is 60 minutes.

The empirical results provide a clear ordering across the four models. Indeed, the reduced form model notably outperforms the linear model with constant illiquidity parameter, indicating that the theoretically derived stochastic liquidity factor markedly

improves the model fit. Furthermore, the reduced form model also outperforms the depth model, especially out-of-sample and in terms of stability.[17] On the other hand, the reduced form model underperforms the propagator model with square root impact. More specifically, the reduced-form model's metrics fall roughly half-way between the linear model with constant parameters and the square-root model. These statistics indicate that the martingale terms in the convergence theorem, while mean zero and irrelevant for many stochastic control problems, do hold explanatory power from a fitting perspective.

A further observation is that the explanatory power of price impact models is a decreasing function of the horizon h, in line with a broad study by Tomas, Mastromatteo, and Benzaquen (2022).[18]

2.5.4 Sensitivity Analysis

Dynamic models

To incorporate systematic time-of-day effects, Cont et al. (2014) also refit their models with different parameters every thirty minute of the day. Plotting the average model parameter $\hat{\lambda}$ over days as a function of the time of day then provides a visualization of the time of day effect, see (Cont et al., 2014, Figure 10). This approach leads to another natural benchmark, a *nonparametric* fit of λ that changes every thirty minutes. We also implemented this nonparametric approach, using 90 minute intervals for $h = 1$ and 150 minute intervals for $h = 15$. Table 2.2 summarizes the performance of this fitting methodology.

The main takeaways are the following:

(a) Refitting λ to account for time-of-day effects significantly improves the standard

[17] The reduced form model is also more parsimonious and does not require access to limit order book data, only trading volumes. The depth model performs remarkably well when including additional limit order book events, such as in the OFI model of Cont et al. (2014). However, the depth model adds less value when considering trades only. Therefore, we exclude the depth model in the remaining analysis.

[18] Tomas et al. (2022) provide a large number of details for reproducing their analysis. In particular, the authors outline their filtering procedure. The main filter is "removing the beginning and end of the trading period to focus on the intraday behavior of liquidity and volatility and circumvent intraday non-stationary issues." Section 2.5.4 illustrates these time-of-day issues and their effect on price impact's explanatory power for various models. We do not filter trading periods at the start and end of the day as they are important sources of time-varying liquidity.

Price Impact Model	In-sample R^2	Out-of-sample R^2	t-stat
Propagator	25%	24%	2.5
Reduced form	18%	17%	2.5
OW	17%	14%	1.9

(a) Performance of various price impact models for $h = 1$ minute.

Price Impact Model	In-sample R^2	Out-of-sample R^2	t-stat
Propagator	20%	17%	1.5
Reduced form	13%	10%	1.5
OW	13%	8%	1.1

(b) Performance of various price impact models for $h = 15$ minutes.

Table 2.2: Performance across the S&P 500 of various *non-parametric* price impact models across horizons h fitted on monthly samples over the course of 2019. Note that we compute the t-stats by time of day. Therefore, the t-stats only reflect (month, stock) variability. In each model, the half life of price impact is 60 minutes.

Obizhaeva-Wang model. In contrast, the model with stochastic illiquidity parameter does not benefit from the refitting to the same extent, as the market activity that determines the stochastic factor already accounts for most intraday variations.

(b) The refitted version of the Obizhaeva-Wang model does *not* outperform the reduced-form model: the theoretically derived model captures the liquidity dynamics as well as a fully non-parametric model.

(c) The non-parametric models are less stable than their static counterparts across (month, stock) pairs for a given time of day.

(d) The R^2 of the nonlinear propagator model benefits from the refitting, at the expense of model stability.

Universal models

Another important empirical question is whether it is preferable to fit a single "universal" model across stocks, or a collection of bespoke models, one for each stock. The former approach has advantages in terms of stability, while the latter can more dynamically adjust to the data, in this case, stock characteristics.

The first universal price impact models are fitted in Lillo, Farmer, and Mantegna

(2003). Zhou (2012) provides a detailed comparison of universal and specialized models and concludes that universal models serve as important benchmarks:

> "these universal price impact functions are unambiguous targets that any empirical model of order-driven markets must hit." (p. 13)

Table 2.3 compares the in-sample and out-of-sample performance of universal and specialized models. This comparison is possible because we normalize all models using the price volatility σ and the average daily trading volume ADV. Without this normalization, the universal model would not produce a stable λ across stocks, let alone estimates of λ comparable to the specialized models. Three broad observations follow:

(a) All models mechanically perform better in-sample when the fitting model is specialized to each stock.

(b) All universal models are mechanically more stable than their stock-specific counterparts.

(c) The universal models consistently perform better *out-of-sample* than their stock-specific counterparts.

(d) The performance gap increases with the length of the horizon h.

Because universal models are mechanically more stable, they are, all else being equal, preferable to stock-specific models. Furthermore, universal modeling enables the pooling of data across stocks. In turn, one can use this data pooling to fit higher dimensional models, e.g., a neural network for the predicted volume curve v_t or a reinforcement learning algorithm for optimal execution.

2.6 Conclusion

This chapter studies the high-frequency limit of propagator models, where trades cause immediate nonlinear price impact that in turn decays gradually over time. We show that the nonlinearity does not disappear in the continuous-time limit, but can be effectively

Model	univ.	spec.
Propagator	18%	19%
Red. form	13%	15%
OW	10%	11%

(a) In-sample R^2 of various price impact models for $h = 1$ minute.

Model	univ.	spec.
Propagator	18%	18%
Red. form	14%	14%
OW	10%	10%

(b) Out-of-sample R^2 for $h = 1$ minute.

Model	univ.	spec.
Propagator	13%	14%
Red. form	9%	11%
OW	7%	8%

(c) In-sample R^2 for $h = 15$ minute.

Model	univ.	spec.
Propagator	14%	13%
Red. form	10%	9%
OW	8%	6%

(d) Out-of-sample R^2 for $h = 15$ minute.

Model	univ.	spec.
Propagator	8%	10%
Red. form	6%	8%
OW	5%	7%

(e) In-sample R^2 for $h = 60$ minute.

Model	univ.	spec.
Propagator	8%	8%
Red. form	6%	5%
OW	5%	4%

(f) Out-of-sample R^2 for $h = 60$ minute.

Table 2.3: Performance across the S&P 500 of various price impact models across horizons h fitted on training samples of one month over the course of 2019. In each case, the half life of impact is 60 minutes. "Univ." stands for the universal fitting methodology; "spec." for the stock-specific methodology.

approximated by a linear price impact model with a stochastic illiquidity parameter. More specifically, illiquidity in this reduced form model is mechanically linked to market activity.

In the limiting model, optimal trading problems have explicit solutions even for general nonparametric alpha and volume predictions. These results in turn identify bounds on changes in market activity that rule out price manipulation, and relations between autocorrelation of order flow and impact decay that rule out statistical arbitrage.

Our analysis also illustrates the power of two convenient mathematical tools: functional central limit theorems as in Jacod and Protter (2012) and the map to impact space from Fruth et al. (2013). Both apply to a broader range of trading problems.[19] Similarly, our empirical analysis illustrates the availability and applicability of public trading data for price impact models.

2.7 Appendix

2.7.1 Proofs for Section 2.3

The proof of the continuous-time scaling limit from Theorem 2.3.3 is based on the functional central limit theorem as in (Jacod & Protter, 2012, Theorem 10.3.2(a)). For easy reference and to encourage further applications of this powerful tool, we provide a compact statement of this result and the underlying assumptions in our context.

We first summarize the assumptions on the processes involved; these correspond to Assumptions (H) and (K) from (Jacod & Protter, 2012, p. 273 and p. 284) in the absence of jumps.

Assumption 2.7.1. *We consider a real-valued state process X with Itô dynamics,*

$$dX_t = b_t dt + \sigma_t dW_t.$$

[19] The characterization of rough volatility models as scaling limits of discrete order-flow dynamics is one example for the former (Jaisson & Rosenbaum, 2015; Jusselin & Rosenbaum, 2020). Ackermann et al. (2021) cover generalizations of the latter.

Here, the drift rate b is locally bounded; the volatility σ has Itô dynamics,

$$d\sigma_t = \tilde{b}_t dt + \tilde{\sigma}_t dW_t + a_t dW_t^{\perp}$$

for a Brownian motion $W^{\perp}$ independent of W and locally bounded coefficients $a, \tilde{b}, \tilde{\sigma}$.

Next, we collect the assumptions required for the nonlinear function of the discrete increments of the state process, whose normalized sums are aggregated in the functional central limit theorem:

Assumption 2.7.2. *Let $G : [0,T] \times \mathbb{R} \to \mathbb{R}$ be a function of time and state increments, which satisfies the following conditions for a function g with at most linear growth and a constant $c > \frac{1}{2}$:*

(a) Anti-Symmetry*: $G(t,x) = -G(t,-x)$ for all $(t,x) \in [0,T] \times \mathbb{R}$;*

(b) Smoothness*: $x \mapsto G(t,x) \in C^1$ for all $t \in [0,T]$;*

(c) Linear Growth*: $|G(t,x)| + |\partial_x G(t,x)| \leq g(x)$ for all $(t,x) \in [0,T] \times \mathbb{R}$;*

(d) Time Regularity*: $|G(t,x) - G(s,x)| \leq g(x)|t-s|^c$ for all $(t,s,x) \in [0,T] \times [0,T] \times \mathbb{R}$.*

Under these assumptions, the following functional central limit theorem applies:

Theorem 2.7.3. *Under Assumptions 2.7.1 and 2.7.2, the discrete process*

$$\sqrt{\Delta t^N} \sum_{n=1}^{\lceil Nt \rceil} G\left(t_{n-1}^N, \tfrac{1}{\sqrt{\Delta t^N}} \Delta X_n^N\right)$$

converges stably in law to

$$\int_0^t \left(b_s \tilde{\mathbb{E}}\left[\partial_x G(s, \sigma_s Z)\right] + \frac{\tilde{\sigma}_s}{2} \tilde{\mathbb{E}}\left[(Z^2 - 1)\partial_x G(s, \sigma_s Z)\right] \right) ds + \int_0^t \sqrt{\tilde{\mathbb{E}}\left[G^2(s, \sigma_s Z)\right]} dW'_s.$$

Here, the expectation $\tilde{\mathbb{E}}[\cdot]$ is taken with respect to the law of an independent standard normal random variable Z and W' is a Brownian motion defined on a very good extension

of the initial filtered probability space, whose quadratic covariation with W is[20]

$$[W, W']_t = \int_0^t \frac{\tilde{\mathbb{E}}[ZG(s, \sigma_s Z)]}{\sqrt{\tilde{\mathbb{E}}\left[G^2(s, \sigma_s Z)\right]}} ds.$$

Before outlining how Theorem 2.7.3 follows from (Jacod & Protter, 2012, Theorem 10.3.2(a)), let us briefly explain the intuition for the linearization in the limit. For simplicity, we focus on the terminal time $t = T$ and assume that the function $G(t, x) = G(x)$ only depends on the spatial variable x. Moreover, we suppose that $X_t = X_0 + bt + \sigma W_t$ is a Brownian motion with drift, so that its rescaled increments are of the form $\frac{1}{\sqrt{\Delta t^N}} \Delta X_n^N = b\sqrt{\Delta t^N} + \sigma Z_n^N$, where $Z_1^N, \ldots, Z_N^N$ are iid standard normal. As the drift of the order flow is a small perturbation of its fluctuations, a Taylor expansion yields the linear approximation

$$\sqrt{\Delta t^N} \sum_n G(\tfrac{1}{\sqrt{\Delta t^N}} \Delta X_n^N) = \sqrt{\Delta t^N} \sum_n G(\sigma Z_n^N) + b\Delta t^N \sum_n \partial_x G(\sigma Z_n^N).$$

As $G(x)$ is odd, the terms in first sum are already centered. By the Central Limit Theorem, they therefore converge to a Gaussian random variable with mean zero. To argue analogously for the terms in the second sum, we first have to subtract and add their mean $\mathbb{E}[\partial_x G(\sigma Z])$. The random part of this decomposition in turn converges to another Gaussian random variable with mean zero. The other part is the Riemann sum $\sum_n b\mathbb{E}[\partial_x G(\sigma Z])\Delta t^N$, which converges to $\int_0^T b\mathbb{E}[\partial_x G(\sigma Z])dt$, in line with the first term in Theorem 2.7.3. The second term in Theorem 2.7.3 does not appear here because the state process has no stochastic volatility in our simple example ($\tilde{\sigma} = 0$). Finally, the two centered Gaussians can be aggregated into the noise term in Theorem 2.7.3.

Proof of Theorem 2.7.3. First suppose that $T = 1$. Then, (Jacod & Protter, 2012, Theorem 10.3.2) applies with Hypothesis (a). As G is odd, (Jacod & Protter, 2012, Equa-

[20] The covariation of W' with $W^\perp$ is zero.

tion (10.3.1)) simplifies to

$$\bar{V}'^N = \sqrt{\Delta t^N} \sum_{n=1}^{\lceil Nt \rceil} G\left(t^N_{n-1}, \tfrac{1}{\sqrt{\Delta t^N}} \Delta X^N_n\right).$$

The limit $\bar{V}'(G(X), X)$ from (Jacod & Protter, 2012, Theorem 10.3.2) in turn decomposes into

$$\bar{V}'(G(X), X) = \bar{U}'(G(X), X) + \bar{A}(G(X), X) + \bar{A}'(G(X), X) + \bar{U}(G(X), X).$$

The four terms in this decomposition can be characterized as follows:

(a) Conditionally on the original σ-field $\mathcal{F}$, $\bar{U}'$ is a continuous centered Gaussian process with independent increments, satisfying

$$\begin{aligned}&\mathbb{E}\left[\left(\bar{U}'(G(X), X)_t\right)^2 \middle| \mathcal{F}\right] \\ &= \int_0^t \tilde{\mathbb{E}}\left[\left(G(s, \sigma_s Z) - Z\tilde{\mathbb{E}}\left[G(s, \sigma_s Z) Z\right]\right)^2\right] ds - \int_0^t \tilde{\mathbb{E}}\left[G(s, \sigma_s Z)\right]^2 ds.\end{aligned}$$

As G is odd, this simplifies to

$$\mathbb{E}\left[\left(\bar{U}'(G(X), X)_t\right)^2 \middle| \mathcal{F}\right] = \int_0^t \left(\tilde{\mathbb{E}}\left[G^2(s, \sigma_s Z)\right] - \tilde{\mathbb{E}}\left[Z G(s, \sigma_s Z)\right]^2\right) ds.$$

This is the part of the term $\int_0^t \sqrt{\tilde{\mathbb{E}}\left[G^2(s, \sigma_s Z)\right]} dW'_s$ in the final formula that is uncorrelated with W (and also uncorrelated with the orthogonal Brownian motion $W^\perp$ that appears in the dynamics of the volatility σ of X).

(b) Next, $\bar{U}$ is given by

$$\bar{U}(\bar{G}(X), X)_t = \int_0^t \tilde{\mathbb{E}}\left[Z G\left(s, \sigma_s Z\right)\right] dW_s.$$

In the final formula, this leads to the part of the term $\int_0^t \sqrt{\tilde{\mathbb{E}}\left[G^2(s, \sigma_s Z)\right]} dW'_s$ that is correlated with W (but uncorrelated with $W^\perp$).

(c) The term $\bar{A}$ is the drift $\int_0^t b_s \tilde{\mathbb{E}}\left[\partial_x G(s, \sigma_s Z)\right] ds$.

(d) Finally, the term $\bar{A}'$ is the drift $\int_0^t \frac{\tilde{\sigma}_s}{2} \tilde{\mathbb{E}}\left[(Z^2 - 1)\partial_x G(s, \sigma_s Z)\right] ds$.

Together, these four terms in turn lead to the asserted representation for the limit process. For a general time horizon T, one first reduces to the case of $T = 1$ by a simple linear time change. Then, one argues as above, and finally switches time back to the original clock. □

Remark 2.7.4. *(Jacod & Protter, 2012, Theorem 10.3.2(a)) in fact shows that the same convergence result remains true if the function G not only depends on the increments of the state process ("future state changes") but also on the previous values of the process ("current states"), as long as all assumptions are uniform in this extra argument.*

Analogously, as pointed out in (Jacod & Protter, 2012, Remark 10.3.4), the function G can also depend directly on $\omega \in \Omega$ in an adapted way as long as all regularity condition hold uniformly.

From a microstructure perspective, the second observation extends the result's range to models that have both local nonlinearities G and stochastic parameters, for instance, a time-changed locally concave price impact model.

The functional central limit theorem does not directly apply to discrete nonlinear propagator models (2.1), because the corresponding impact state is not the discretization of a fixed process, but the solution of a difference equation (which depends on the time grid). However, using the explicit solution of the difference equation, Theorem 2.7.3 can be applied in a two-step procedure. To wit, consider the discrete nonlinear propagator model

$$\Delta I_n^N = -\beta I_{n-1}^N \Delta t^N + \sqrt{\Delta t^N} g\left(\tfrac{1}{\sqrt{\Delta t^N}} \Delta F_n^N\right).$$

The explicit solution of this difference equation is

$$I_n^N = \sum_{m=1}^{n} \left(1 - \beta \Delta t^N\right)^{n-m} \sqrt{\Delta t^N} g\left(\tfrac{1}{\sqrt{\Delta t^N}} \Delta F_m^N\right)$$

$$= \left(1 - \beta\Delta t^N\right)^n \sqrt{\Delta t^N} \sum_{m=1}^{n} \left(1 - \beta\Delta t^N\right)^{-m} g\left(\tfrac{1}{\sqrt{\Delta t^N}}\Delta F_m^N\right).$$

Whence, one can first control the gap $(1-\beta\Delta t^N)^{-m} - e^{\beta t_m^N}$ between discrete and continuous-time exponentials. In a second step, one can then apply Theorem 2.7.3 to obtain the convergence

$$\sqrt{\Delta t^N} \sum_{m=1}^{n} e^{\beta t_m^N} g\left(\tfrac{1}{\sqrt{\Delta t^N}}\Delta F_m^N\right) \to e^{\beta t} I_t$$

where I is the linear price impact model from Theorem 2.3.3.

2.7.2 Proofs for Section 2.4

Proof of Lemma 2.4.1. As $\alpha_T = 0$ and $F_0 = 0$, integration by parts gives

$$0 = \alpha_T F_T - \alpha_0 F_0 = \int_0^T \alpha_{t-} dF_t + \int_0^T F_{t-} d\alpha_t + [F, \alpha]_T.$$

Since $S_T - \bar{I}_T$ is square integrable, $\alpha_t = \bar{M}_t - S_t - \bar{I}_t$ for a square-integrable martingale $\bar{M}_t$, which has bounded volatility by assumption. For $F \in \mathcal{L}^2$, this implies that

$$\mathbb{E}\left[\int_0^T F_{t-} d(S_t + \bar{I}_t)\right] = \mathbb{E}\left[\int_0^T -F_{t-} d\alpha_t\right] = \mathbb{E}\left[\int_0^T \alpha_{t-} dF_t + [F, \alpha]_T\right].$$

The goal functional (2.4) can therefore be rewritten as

$$\mathbb{E}\left[\int_0^T \left(\alpha_{t-} dF_t + d[F, \alpha]_t - I_{t-} dF_t - \frac{1}{2} d[I, F]_t\right)\right].$$

Now we use $dF_t = \frac{1}{\lambda_t} dI_t + \frac{\beta}{\lambda_t} I_t dt$ to replace the trader's holdings. This allows us to rewrite (2.4) as

$$\mathbb{E}\left[\int_0^T \left(\frac{\alpha_{t-}}{\lambda_t} dI_t + \frac{\beta\alpha_t}{\lambda_t} I_t dt + \frac{1}{\lambda_t} d[I, \alpha]_t - \frac{I_{t-}}{\lambda_t} dI_t - \frac{\beta I_t^2}{\lambda_t} dt - \frac{1}{2\lambda_t} d[I]_t\right)\right]. \tag{2.7}$$

Integration by parts (using that $\lambda_t = e^{\varphi_t}$ is smooth and $I_0 = 0$), gives

$$\frac{1}{2\lambda_T} I_T^2 = \int_0^T -\frac{\varphi_t'}{2\lambda_t} I_t^2 dt + \frac{1}{\lambda_t} I_{t-} dI_t + \int_0^T \frac{1}{2\lambda_t} d[I]_t$$

as well as

$$0 = \frac{\alpha_T}{\lambda_T} I_T = \int_0^T -\frac{\varphi_t' \alpha_t}{\lambda_t} I_t dt + \int_0^T \frac{\alpha_{t-}}{\lambda_t} dI_t + \int_0^T \frac{I_{t-}}{\lambda_t} d\alpha_t + \int_0^T \frac{1}{\lambda_t} d[\alpha, I]_t.$$

Plugging these two identities into (2.7) in turn yields

$$\begin{aligned} &\mathbb{E}\left[\int_0^T \left(-\frac{I_t}{\lambda_t} d\alpha_t + \frac{\beta + \varphi_t'}{\lambda_t} \alpha_t I_t dt - \frac{2\beta + \varphi_t'}{2\lambda_t} I_t^2 dt\right) - \frac{1}{2\lambda_T} I_T^2\right] \\ &= \mathbb{E}\left[\int_0^T e^{-\varphi_t} \left(-I_t \mu_t^\alpha + (\beta + \varphi_t') \alpha_t I_t - \frac{2\beta + \varphi_t'}{2} I_t^2 dt\right) - e^{-\varphi_T} \tfrac{1}{2} I_T^2\right]. \end{aligned}$$

Under the no price manipulation condition $2\beta + \varphi_t' > 0$, pointwise maximization in turn yields the optimal impact (2.6). □

Chapter 3

A Leland Model for Delta Hedging in Central Risk Books

This chapter is a joint work with Johannes Muhle-Karbe and Kevin Webster.

3.1 Introduction

Using a tractable extension of the Leland (1985) model, we study how a delta-hedging strategy can realistically be implemented using market *and* limit orders in a centralized, automated market-making desk that integrates trading and liquidity provision for both options and their underlyings. In the continuous-time limit, the optimal limit-order exposure can be computed explicitly by a *pointwise* maximization. It is determined by the relative magnitudes of adverse selection, bid-ask spreads, and volatilities. The corresponding option price – from which the option can be replicated using market and limit orders – is characterized via a nonlinear PDE.

Our results highlight the benefit of tactical liquidity provision for contrarian trading strategies, even for a trading desk that is not a competitive market maker. More generally, the chapter also showcases how reduced-form models are competitive with "brute force" numerical approaches to market microstructure. Both the estimation of microstructure parameters, and the simulation of the optimal trading strategy are made concrete and reconciled with real-life high frequency data.

This chapter revisits the Leland (1985) delta-hedging strategy in the modern context of a centralized, automated market-making desk. The motivation of the original Leland paper was to make the Black and Scholes delta-hedging strategy actionable. Rather than just pricing the option via a theoretical replication argument, Leland outlined how the replication strategy could be implemented with actual trades. This involved taking into account some real-life trade-offs trading desks faced, most notably the trade-offs between *trading frequency*, *hedging error* and *transaction costs*. As a by-product, the paper also established how the liquidity of the underlying affected the price of the option.

In this chapter, we update this approach by outlining how a delta-hedging strategy can realistically be implemented in a centralized, automated market-making desk that integrates trading and liquidity provision for both options and their underlyings. More specifically, we model how delta-hedging could practically be implemented in a market-making context. Similarly to the original Leland paper, additional relationships between the liquidity of the underlying and the option price are uncovered as a by-product.

Centralization has been a key force in banks and hedge funds in the last decade. Broadly speaking, the goal is to consolidate related trading activities together for three reasons: facilitating *internalization* (J. Chen, 2021), accelerating *systematic, model-driven trading* and promoting a *big picture understanding of trading exposures*. The last point in particular has led to the creation of "Central Risk Books" (CRBs), which consolidate multiple trading teams into one.[1] Of course, this leverages the trading technology and execution strategies developed for automated, model-driven trading in other contexts before that.[2]

In the context of options market making, centralization means aggregating the trading activity across both the underlying and all related options into one portfolio. This allows two explicit benefits to the delta-hedging of the options positions. First, by netting delta

[1]Finextra (2016) provides the following definition: "The purpose of the CRB in all this is to act as a huge repository for all the firm's positions. Then, someone with a brain the size of a small planet sifts through all this and works out what the net exposure of the bank is."

[2]For example, the SEC's 2020 *Staff Report on Algorithmic Trading in U.S. Capital Markets*(Securities & Commission, 2020) reports: "A common theme echoed by nearly all market professionals, academic researchers, and other students of the securities markets is that that algorithmic trading, in one form or another, is an integral and permanent part of our modern capital markets." (p. 4)

exposures between multiple options positions as well as the market making position in the underlying, the amount of delta-hedging is drastically reduced. Second, sophisticated liquidity provision strategies employing limit orders used for the market making in the underlying can reduce the individual cost of each delta hedge. Both lead to more competitive prices being quoted on the option. The consolidation also can provide benefits to the underlying, which we ignore in this chapter.

The exposure netting benefit is more technological in nature: the centralized team needs a robust real-time trading infrastructure to establish this centralized view of the consolidated exposure.[3] The second benefit is mathematically more interesting, and relates heavily to the market microstructure and market-making literature, cf., e.g., Carmona and Webster (2019); Cartea, Gan, and Jaimungal (2019); Cartea and Jaimungal (2015a, 2015b); Cont and Kukanov (2017); Guéant, Lehalle, and Fernandez-Tapia (2012); Guilbaud and Pham (2013); Herdegen, Muhle-Karbe, and Stebegg (2022); Horst and Naujokat (2014); Kühn and Muhle-Karbe (2015) and the references therein. The benefits of employing liquidity provision strategies similar to those of market makers are qualitatively clear. However, modelling a market-making strategy, its associated exposures and P&L is a difficult task, as by definition a market maker doesn't control exactly when they trade, or what position they end up with when providing liquidity. This is primarily due to the trade-off between capturing the *bid-ask spread*, *adverse selection* and *hedging error*.[4]

Our modern approach to delta-hedging therefore heavily relies on understanding the type of positions a market maker can achieve, the spread they can capture and the adverse selection they are exposed to when providing liquidity on the underlying. In order to obtain tractable results, we do not model the fine structure of the limit order book as a queueing system as in, e.g., Cont, Stoikov, and Talreja (2010). Instead, we propose a simple reduced-form model driven by two correlated Brownian motions, that

[3]See Chapter 12 "The Research Stack" (pp. 401-410) from the book *Algorithmic Trading and Quantitative Strategies* (Velu, Hardy, & Nehren, 2020) for an extensive treatment of the technological challenges involved.

[4]See, e.g., Section 1.3.1 in the book "Trades, Quotes and Prices" (J.-P. Bouchaud, Bonart, Donier, & Gould, 2018) for a definition of adverse selection in the context of market making.

captures the main tradeoffs at hand but still retains most of the tractability of the classical Leland model. In this context, the first Brownian motion drives the price changes of the underlying. The second models the flow of market orders placed by other market participants, which in turn describes how limit order positions are executed. Correlation between the two Brownian motions then models adverse selection, in that (some) other market participants tend to buy using market orders only when they have proprietary information that prices will have a tendency to increase subsequently.

Instead of modelling the complex dynamics of the various queues in the limit order book, our control variable is the aggregate exposure taken with respect to this order flow process.[5] In order to offset any unwanted positions generated in this way and eventually hedge the option position at hand, these limit-order trades are complemented by market orders. As in Leland (1985)'s original model, our baseline model assumes that one rehedges completely to the delta hedge on a discrete time grid. In an extension of the model, we then relax this assumption by introducing a second control variable, the fraction of the deviation from this target position that is offset at each trading time. This leads to a nontrivial tradeoff between spread costs, adverse selection, and risk.[6]

Using limit theorems in the spirit of Jacod and Protter (2012), we show that for frequent trades and small transaction costs, the limiting wealth dynamics can be computed in closed form. For a given target strategy, this in turn allows us to determine the optimal limit-order exposure in closed form by a simple *pointwise* maximization. In a second step, we can then derive a nonlinear partial differential equation that describes the option price from which the option payoff can be replicated in this context. (We show that a smooth solution exists for sufficiently regular payoffs by adapting the results of Ishimura (2010) for the classical Leland model.)

We find that – for market participants for which market making in the underlying

[5] In a pro-rata limit order book, where limit orders are executed without a time-priority rule, full exposure corresponds to posting a very large number of limit orders that capture almost all incoming market order flow. Conversely, no exposure corresponds to posting no new limit orders and canceling all existing ones.

[6] Optimal market and limit orders for option hedging are computed in Agliardi (2016); Cartea et al. (2019); Ellersgaard and Tegnér (2017) by numerically solving the corresponding dynamic programming equations. Agliardi (2016); Ellersgaard and Tegnér (2017) do not consider adverse selection. The model of Cartea et al. (2019) is closest to the one we study in the present chapter.

is not directly profitable – limit orders are only beneficial for "contrarian strategies", which are negatively correlated with the underlying. The intuition is that these can benefit from adverse selection, because limit order execution then tends to be aligned with changes of the strategy. In contrast, for "trend-following strategies" that have positive correlation with the market, executed limit orders typically have to be unwound using market orders, which is never optimal in our model. For contrarian strategies (e.g., delta hedging of positions with negative gamma), the optimal limit order exposure depends on the relation between the adverse selection parameter and the relative average costs of market and limit orders in our model. More specifically, some limit orders are optimal if the bid-ask spread and the adverse selection parameter are neither too large nor too small relative to each other.

When it is optimal to use limit orders, these reduce hedging costs by a constant fraction (unless the option position is so large that even capturing the entire order flow does not lead to an interior maximum). As a consequence, the option price that allows for replication then also lies between its counterparts from the frictionless Black-Scholes model and the Leland model where only market orders are used. In particular, for options with convex payoffs, the shift of the implied volatility suggested by the Leland (1985) model is scaled back by a constant factor for negative gamma positions.

To illustrate the quantitative properties of these results, we implement them using tick-level data from the LOBSTER database. Direct liquidity provision is profitable at the highest trading frequencies, but these profit opportunities can typically only be exploited by the fastest market makers. For other somewhat slower market participants, direct liquidity provision is not profitable but limit orders are still useful for implementing contrarian strategies. For example, for Google data from 2019, this regime applies for realistic trading frequencies with around 25 seconds between trade (10 ticks of the trade clock).

When one always fully rehedges to the delta hedge using market orders at this frequency, the price adjustments relative to the frictionless version of the model are unrealistically large and a substantial fraction of trades is implemented via market orders

rather than limit orders. Both of these results change in the extension of our model where finite risk aversion allows some discretion about displacements from the target hedge. For different values of the risk aversion parameter, the corresponding indifference price then interpolates between the frictionless value (without trading costs) and the model with full rehedging. Moreover, as risk aversion decreases, the fraction of the displacement that is rectified by market orders in each trading round quickly decreases, whereas the optimal limit-order placement remains largely insensitive. As a consequence, trades via limit order quickly account for more than half of the overall trading volume.

Broadly speaking, our empirical analysis highlights two takeaways of the chapter. First, tactical liquidity provision can be beneficial even for non-competitive market makers, at least in the context of contrarian trading strategies. Second, the reduced form model successfully synthesizes the key market microstructure features needed to bring the trading strategy to the data, while remaining tractable and interpretable. This makes the approach competitive and compelling compared to purely numerical simulations of trading strategies at the discrete microstructure level.[7]

The remainder of this article is organized as follows. Section 3.2 introduces the model for trading with market and limit orders on a discrete time grid. In Section 3.3, we then report our first two main results: the continuous-time limits of the corresponding wealth dynamics and the optimal limit-order exposures for a given target hedge. The latter is in turn endogenized in Section 3.4 as the delta hedge for the option price from which perfect replication is possible in the scaling limit of frequent trading and small transaction costs. Section 3.5 implements these results using high-frequency data for limit order books. Finally, the extension of the model with a non-trivial risk-cost tradeoff is discussed in Section 3.6. For better readability, all proofs are delegated to Appendix 3.8.1, Appendix 3.8.2 and Appendix 3.8.4.

[7] Velu et al. (2020) make a compelling case for finding alternatives to microstructure-level brute force simulations: "Creating a realistic microstructure based simulation environment for execution strategies is extremely hard and probably out of reach of most practitioners at this point" (p. 407).

Notation Throughout this chapter, we fix a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{P})$ supporting two independent standard Brownian motions $(W_t)_{t\in[0,T]}$ and $(W_t^\perp)_{t\in[0,T]}$. The adapted processes $(X_t)_{t\in[0,T]}$ satisfying $\mathbb{E}[\int_0^T X_t^2 dt] < \infty$ are denoted by $\mathcal{L}^2$. We first consider trading on a discrete, equidistant time grid

$$t_n^N = nT/N, \quad n = 0, \dots, N,$$

with mesh width

$$\Delta t^N = T/N.$$

Then, we pass to the continuous-time limit $N \to \infty$ in a second step. The samples of any continuous-time process $(X_t)_{t\in[0,T]}$ along this grid and the corresponding discrete differences are denoted by

$$X_n^N = X_{t_n^N}, \qquad \Delta X_n^N = X_n^N - X_{n-1}^N.$$

3.2 Model

3.2.1 Financial Market

We consider a financial market with two assets. The first one is safe, with zero interest rates for simplicity. The second asset is risky, in that its midprice has dynamics

$$dP_t = \mu_t dt + \sigma_t dW_t, \quad \text{for } \mu, \sigma \in \mathcal{L}^2 \text{ and } \sigma_t > 0 \text{ for all } t.$$

3.2.2 Trading with Market Orders

The risky asset can be traded either by market orders or by limit orders. If purchases or sales are implemented via market orders, then these trades are settled immediately at the current best ask- or bid-prices, respectively. For typical financial assets, price changes are of the same order of magnitude as bid-ask spreads. As the discrete price changes ΔP_n^N are of order $\sqrt{\Delta t^N}\sigma_{t_n^N} = O(\sqrt{\Delta t^N})$ for a fine time grid with meshwidth $\Delta t^N = T/N$, we

accordingly model the *bid-ask spread* as

$$s_n^N = \sqrt{\Delta t^N} s_{t_n^N},$$

for a positive continuous-time process $s \in \mathcal{L}^2$. The typical examples are costs proportional to numbers of shares ($s_t = s$) or dollar amounts traded ($s_t = sP_t$).

With this notation, the ask and bid prices at time t_n^N are then given by $P_n^N + s_n^N/2$ and $P_n^N - s_n^N/2$, respectively. A market order of size ΔM_n^N executed at time t_n^N in turn changes the cash account by

$$-P_n^N \Delta M_n^N - \frac{s_n^N}{2} |\Delta M_n^N|. \tag{3.1}$$

Example 3.2.1. *Let us illustrate why the above scaling is a reasonable assumption at high-frequency time scales. For Google shares, the average (absolute) bid-ask spread is* \$0.37 *in LOBSTER data from 2019 (about three basis points of the average stock price in this time period). The average volatility between the* $N = 2552083$ *trading times is about* $\sigma\sqrt{\Delta t^N} = \$0.11$ *(a yearly Black-Scholes volatility of about 20%).*[8] *Whence, for trading strategies that are rebalanced reasonably close to the tick grid (which has a meshwidth of about 2.5 seconds here) the above scaling is very natural. For example, if one trades about every 10 ticks, then the corresponding volatility between trades is about* $\$0.11 \times \sqrt{10} = \$0.35$. *For the entire cross section of S&P500 stocks, the distribution of spreads in units of volatility is plotted in Figure 1.1, showing a rather tight range of similar values across this entire range of stocks.*

3.2.3 Trading with Limit Orders

Market orders allow us to trade the risky asset immediately at the current best bid and ask prices. Alternatively, the risky asset can also be traded using limit orders. We assume that these can only be posted at the best bid and ask prices ("on top of the book"),[9] and

[8] This estimate is obtained by averaging the daily estimates of the volatility between trades.

[9] This is a reasonable simplifying assumption at least for large tick stocks, for which liquidity at the best bid and ask prices is rarely exhausted.

write L_n^N for the cumulative number of shares purchased using limit orders until time t_n^N. Put differently, ΔL_n^N is the number of limit orders that was in the book at time t_{n-1}^N (at the corresponding ask or bid prices $P_{n-1}^N \pm s_{n-1}^N/2$) and is executed at time t_n^N. The corresponding change in the cash account is

$$-P_{n-1}^N \Delta L_n^N + \frac{s_{n-1}^N}{2}|\Delta L_n^N|. \tag{3.2}$$

Together with (3.1) it follows that the total change of the cash account K_n^N from time t_{n-1}^N to t_n^N is

$$\Delta K_n^N = -P_n^N \Delta M_n^N - \frac{s_n^N}{2}|\Delta M_n^N| - P_{n-1}^N \Delta L_n^N + \frac{s_{n-1}^N}{2}|\Delta L_n^N|.$$

The corresponding portfolio value X_n^N (marked-to-market using the midprice of the risky asset) in turn changes by

$$\begin{aligned}\Delta X_n^N =& (K_n^N + (M_n^N + L_n^N)P_n^N) - (K_{n-1}^N + (M_{n-1}^N + L_{n-1}^N)P_{n-1}^N) \\ =& (M_{n-1}^N + L_{n-1}^N)\Delta P_n^N - \frac{s_n^N}{2}|\Delta M_n^N| + \frac{s_{n-1}^N}{2}|\Delta L_n^N| + \Delta L_n^N \Delta P_n^N. \end{aligned} \tag{3.3}$$

Here, the first term on the right-hand side is the change of portfolio value due to price moves, which completely describes the wealth dynamics of self-financing strategies in frictionless markets. With a nonzero bid-ask spread, the second and third terms record the transaction costs paid for market orders and earned by limit orders, respectively. The last term accounts for correlation between the execution of limit orders and price changes. This takes into account that limit orders are at risk of being "adversely selected" – that is, being executed against counterparties with superior information who buy because they accurately predict that prices will go up, for example.

We model this adverse selection in reduced form by introducing an *order flow process* that describes the execution of limit orders:

$$dF_t = \mu_t^F dt + v_t \sqrt{1-\rho_t^2} dW_t^\perp - v_t \rho_t dW_t.$$

Here, $\mu^F, v, \rho \in \mathcal{L}^2$, $v_t > 0$ for all t and ρ takes values in $(0,1)$ to model that positive order flow ΔF_n^N (corresponding to sell orders of other traders hitting the market) is negatively correlated with price changes ΔP_n^N. Therefore, ρ measures the extent of adverse selection.[10] The exact process by which limit orders are posted and executed is very involved. We abstract from this issue and instead describe the main mechanics in reduced form by a control process $(\nu_t)_{t\in[0,T]}$ with values in $[0,1]$. Here, ν_{n-1}^N models the fraction of incoming market order flow captured between time t_{n-1}^N and t_n^N.[11] Then, the number of limit orders executed between time t_{n-1}^N and t_n^N is $\nu_{n-1}^N \Delta F_n^N$ and the self-financing equation (3.3) becomes

$$\Delta X_n^N = (M_{n-1}^N + L_{n-1}^N)\Delta P_n^N - \frac{s_n^N}{2}|\Delta M_n^N| + \frac{s_{n-1}^N}{2}\nu_{n-1}^N|\Delta F_n^N| + \nu_{n-1}^N \Delta F_n^N \Delta P_n^N. \quad (3.4)$$

Remark 3.2.2. *We focus on* symmetric *limit order strategies with the same exposure to the buy or sell orders of other market participants. This is without loss of generality for the continuous-time limit we consider in Theorem 3.3.1, as only the (locally symmetric) martingale parts of the frictionless target strategies (but not a potentially asymmetric drift) contributes there. This would change in the presence of high-frequency signals like the order-flow imbalance, for example.*

3.2.4 Tilted Execution Schemes

The portfolio dynamics (3.4) are controlled by choosing (i) the exposure ν_n^N to the order flow of other market participants via limit orders, and (ii) the market orders ΔM_n^N placed at each time point. An analysis of general market and limit order strategies is beyond our scope here. To obtain tractable results, we instead focus on a class of practically

[10] Cartea et al. (2019) model uncertainty about fills and adverse selection in a similar manner. To wit, they assume that limit orders are filled with some probability, and adversely selected by a price change of one tick with some other probability.

[11] Here, $\nu = 0$ corresponds to posting no new limit orders and cancelling all existing ones. Conversely, in a pro-rata limit order book where limit orders are executed without a time-priority rule, $\nu = 1$ corresponds to posting a very large number of limit orders that capture almost all incoming market order flow. Of course, price impact becomes a major concern when trying to absorb a major fraction of the entire market flow – the focus on the bid-ask spread in the present model is justified if only a reasonably small fraction of the total order flow is targeted.

relevant *tilted execution schemes.*[12]

For these, the limit-order exposure to the order flow of other market participants is chosen freely. After the execution of limit orders is observed, market orders are in turn used to readjust the risky position to an exogenous *target position* with dynamics

$$d\Pi_t = \mu_t^\Pi dt + \gamma_t dW_t, \quad \text{for } \mu^\Pi, \gamma \in \mathcal{L}^2. \tag{3.5}$$

Example 3.2.3. *One standard example are* delta hedging *strategies*

$$\Pi_t = \Pi(t, P_t) = \partial_p C(t, P_t),$$

for a European option with payoff $\varphi(P_T)$ at time T and option price $C(t, P_t)$ at time t in a local volatility model where the price dynamics are of the form $dP_t = \mu(t, P_t)dt + \sigma(t, P_t)dW_t$. In this case, $\gamma_t = \partial_{pp} C(t, P_t)\sigma(t, P_t)$ is determined by the option's "gamma", i.e., the sensitivity of the hedge with respect to price changes of the underlying.

Example 3.2.4. *Another example are investment strategies that hold a constant proportion π of wealth in the risky asset.*[13] *For Black-Scholes price dynamics $dP_t/P_t = \mu dt + \sigma dW_t$, this is optimal for investors with constant relative risk aversion, for example; the corresponding risky position then is of the form*

$$\frac{d\Pi_t}{\Pi_t} = (\pi - 1)(\mu - \sigma^2)dt + (\pi - 1)\sigma dW_t.$$

In particular, $\gamma_t < 0$ if the strategy neither shorts nor leverages the risky asset for $\pi \in (0, 1)$. This means the strategy is always "contrarian", in that it sells after prices rise.

Between time t_{n-1}^N and t_n^N, the target strategy (3.5) changes by $\Delta\Pi_n^N$ and the number of limit orders executed is $\nu_{n-1}^N \Delta F_n^N$. Whence, the market order that needs to be placed

[12] In a trading context, strategies are often classified along an axis with pure liquidity taking strategies on the one end, and pure liquidity provision on the other end. Various algorithms are then said to "tilt" or "lean" towards one end of the spectrum or the other.

[13] The use of limit orders in this context is studied by Kühn and Stroh (2010).

at time t_n^N to offset the discrepancy is

$$\Delta M_n^N = \Delta \Pi_n^N - \nu_{n-1}^N \Delta F_n^N.$$

The portfolio dynamics (3.4) of such a tilted execution scheme in turn are

$$\Delta X_n^N = \Pi_{n-1}^N \Delta P_n^N - \frac{s_n^N}{2} |\Delta \Pi_n^N - \nu_{n-1}^N \Delta F_n^N| + \frac{s_{n-1}^N}{2} \nu_{n-1}^N |\Delta F_n^N| + \nu_{n-1}^N \Delta F_n^N \Delta P_n^N. \quad (3.6)$$

If no limit orders are used ($\nu = 0$), this reduces to the classical model of Leland (1985).

3.3 Optimal Limit-Order Placement

The discrete-time dynamics (3.6) are difficult to optimize over the choice of the tilted execution scheme. However, the continuous time limit $N \to \infty$ (that can be obtained using laws of large numbers for discretized processes from Jacod and Protter (2012)) can be maximized *pointwise* over the limit-order exposure:

Theorem 3.3.1. *As the number of time steps N tends to infinity, the discrete-time portfolio process* (3.6) *converges (uniformly on compacts in probability) to the continuous-time process*

$$dX_t^\nu = \Pi_t dP_t + \left[\frac{s_t}{\sqrt{2\pi}} \left(-\sqrt{\gamma_t^2 + 2\rho_t \gamma_t \nu_t v_t + \nu_t^2 v_t^2} + \nu_t v_t \right) - \rho_t \sigma_t \nu_t v_t \right] dt. \quad (3.7)$$

Each of the terms in the continuous-time portfolio dynamics (3.7) is the limit of the respective discrete counterparts in (3.6): the first records the effect of price changes and the second accumulates the transaction costs paid for market orders. The third term measures the transaction costs earned from limit orders and the final fourth term describes the losses due to adverse selection.

Unlike its discrete counterpart, the continuous-time wealth process can be readily optimized over the choice of the limit orders (controlled through ν_t). Indeed, this control process only affects the drift rate in a pointwise manner. For a given target process

$d\Pi_t = \mu_t^\Pi dt + \gamma_t dW_t$, the optimal limit-order exposure can therefore readily be determined via *pointwise* maximization. To formulate the result in a compact manner, define

$$y_t \coloneqq 1 - \frac{\rho_t \sigma_t \sqrt{2\pi}}{s_t} \in (-\infty, 1).$$

Remark 3.3.2. *The last two terms in the wealth dynamics* (3.7) *show that, without using any market orders, limit order exposure to a fraction ν_t of the flow dF_t generates a drift rate of $\frac{s_t}{\sqrt{2\pi}} \nu_t v_t y_t dt$. Arguing as in the proof of Theorem 3.3.1, one can show that obtaining the same positions $\nu_t dF_t$ via market orders leads to the drift rate $-\frac{s_t}{\sqrt{2\pi}} \nu_t v_t dt$. Whence, y_t measures the relative costs of limit and market orders.*

The empirically most relevant case is $y_t \in (-1, 0)$. This means that limit orders incur a nontrivial cost, but are cheaper than market orders on average. Put differently, adverse selection is large enough relative to the spread to cause a cost, but not so large that it is never optimal to use limit orders.

Example 3.3.3. *For our Google data from 2019, the average correlation measured tick-by-tick on the trade clock is* 22.5% *(cf. Section 3.5 for more details), so that*

$$y = 1 - \frac{\rho \sigma \sqrt{\Delta t^N} \sqrt{2\pi}}{s \sqrt{\Delta t^N}} = 0.83.$$

This suggests that liquidity provision is profitable at the highest frequency. *However, these profits are only attainable for the fastest market makers. For example, at a frequency of once every 10 ticks (about 25 seconds), the realized correlation is* 45.9%, *so that*

$$y = 1 - \frac{\rho \sigma \sqrt{\Delta t^N} \sqrt{2\pi}}{s \sqrt{\Delta t^N}} = -0.14 \in (-\rho, 0).$$

As the trading frequency decreases further, the realized correlation remains relatively stable, cf. Figure 3.1. The corresponding y parameter therefore quickly becomes more negative than $-\rho$, and limit orders are in turn dominated by market orders. Figure 3.3 reports qualitatively similar results across the entire cross section of S&P500 stocks.

Two observations cast some uncertainty on the exact value of ρ to use for our model:

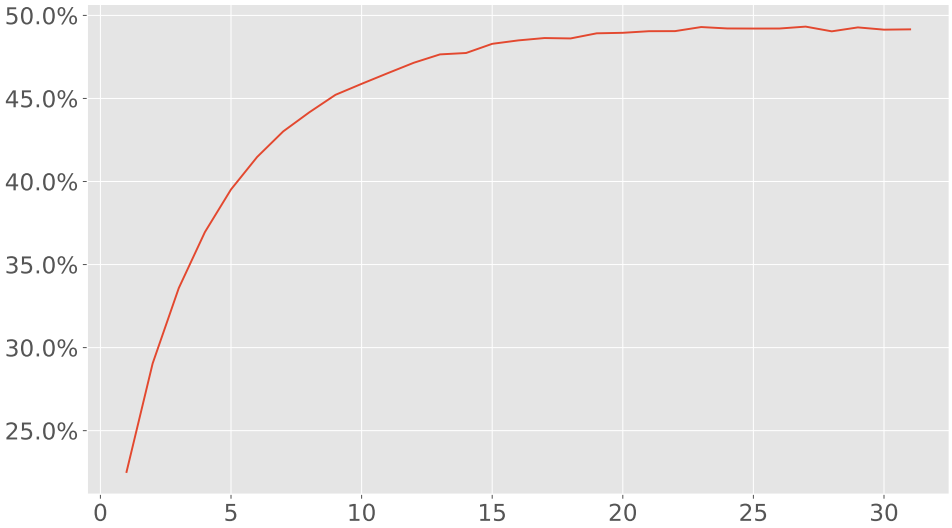

Figure 3.1: Estimates for the adverse selection parameter ρ on various sampling grids on 2019 Google data.

- *The observed correlation could be explained by two different causal models: the price could move* because *of the trade, or* the trade could anticipate the price move. *The former is referred to as price impact. Typically, adverse selection is significantly faster than price impact. This means that some of the increase of ρ is due to price impact, leading the estimate for lower trading frequencies to be biased upward.*

- *The observed correlation is the average correlation across all market participants. This includes very competitive market makers and traders who employ liquidity provision more tactically. Because the former aims to maximize spread capture and minimize adverse selection, more tactical trading strategies will achieve spreads and adverse selection parameters that are less competitive. This means that any estimate of ρ based on the public trading tape is going to be biased downward for a "tactical liquidity provider".*

Both of these biases can be reduced by using proprietary trading data. The above estimates on public data can then be used to situate the team's achieved spreads and adverse selection in comparison to the wider market.

Remark 3.3.4. *Note that while we estimate ρ and s using the full trading tape, not every trade can realistically be captured, and the most profitable trades will be captured by more competitive market makers. Therefore, while the range of trading frequencies for which*

y lies between $-\rho$ and 0 may seem narrow, the large majority of market participants will fall in it. The competitive pressures will allow only the most dedicated, advanced market makers to profit from pure liquidity provision and fall in the positive region of y. Other market participants can see their transaction costs reduced by tactically providing liquidity, but y remaining negative prevents them from employing their execution algorithms as dedicated market making strategies.

With the above notation, the *optimal* limit-order exposure in (3.7) can be described as follows:

Lemma 3.3.5. *The drift rate in* (3.7) *is maximized by*

$$\nu_t^* = \max\left\{\min\{\tilde{\nu}_t^*, 1\}, 0\right\}, \quad \textit{where } \tilde{\nu}_t^* = \begin{cases} -\frac{\gamma_t}{v_t}\left(\rho_t - \operatorname{sgn}(\gamma_t) y_t \sqrt{\frac{1-\rho_t^2}{1-y_t^2}}\right) & \text{if } y_t \in (-1,1), \\ 0 & \text{if } y_t \leq -1. \end{cases} \tag{3.8}$$

Next, we explore in what parameter regimes it is optimal to place at least some limit orders:

Lemma 3.3.6. *It is optimal to use at least some limit orders ($\nu_t^* > 0$) in the following cases:*

(i) If $y_t \in (-1, 0)$, then $\nu_t^* > 0$ if and only if $\gamma_t < 0$ and *$\rho_t > -y_t$.*

(ii) If $y_t \in [0, 1)$, then $\nu_t^* > 0$ if and only if either $\gamma_t < 0$, or *$\gamma_t > 0$ and $\rho_t < y_t$.*

As discussed in Remark 3.3.2, the empirically most relevant case is $y_t \in (-1, 0)$. (Limit orders are cheaper than market orders but still incur a nontrivial cost.) In this case, we see that limit orders are only optimal for "contrarian strategies" with $\gamma_t < 0$, whose sales correlate with positive price shocks. In contrast, limit orders are never optimal for "trend following strategies" with $\gamma_t > 0$ which react to positive price shocks with further purchases. The intuition for this is the following. Due to adverse selection, a positive price shock is more often than not accompanied by negative flow (i.e., purchases of other market participants). Whence, the trend follower will more often than not have

to reverse her limit order executions via market orders and then pay *both* adverse selection and transaction costs. (In the wealth dynamics (3.7), the sum of the second and third terms then is always negative, as is the fourth.) In contrast, for contrarian strategies, one more often then not does *not* have to reverse limit-order executions, so that placing limit orders is potentially profitable.

Whether this is indeed the case depends on the relative sizes of volatility, bid-ask spreads and adverse selection. More specifically, it is optimal to use limit orders in the case $y \in (-1, 0)$ when adverse selection ρ_t and spread s_t are neither too large nor too small:

Corollary 3.3.7. *(i) Fix the bid-ask spread s_t. Then, $y_t \in (-1, 0)$ and it is optimal to use limit orders if and only if $\gamma_t < 0$ and adverse selection is neither too large nor too small, in that*

$$\rho_t \in \left(\frac{s_t}{\sigma_t \sqrt{2\pi}}, \min \left\{ \frac{2s_t}{\sigma_t \sqrt{2\pi}}, \left(\frac{\sigma_t \sqrt{2\pi}}{s_t} - 1 \right)^{-1} \right\} \right);$$

(ii) Fix the adverse selection parameter ρ_t. Then, $y_t \in (-1, 0)$ and it is optimal to use limit orders if and only if $\gamma_t < 0$ and the bid-ask spread is neither too large nor too small, in that

$$s_t \in \left(\frac{\rho_t}{\rho_t + 1} \sigma_t \sqrt{2\pi}, \rho_t \sigma_t \sqrt{2\pi} \right).$$

Figure 3.2 displays how the parameter space for spread s and adverse selection ρ is divided into three regions, where it is either not optimal to use limit orders at all because the spread is too small to compensate for adverse selection, the intermediate regime where it is optimal to use limit orders to implement contrarian strategies, and the case where spreads are large enough that limit orders are always optimal for direct liquidity provision.

A number of studies argue that the bid-ask spread is proportional to volatility, cf., e.g., Madhavan, Richardson, and Roomans (1997); Wyart et al. (2008). In this case, the usefulness of limit orders only depends on whether the adverse selection parameter ρ_t is small, large, or falls in an intermediate range.

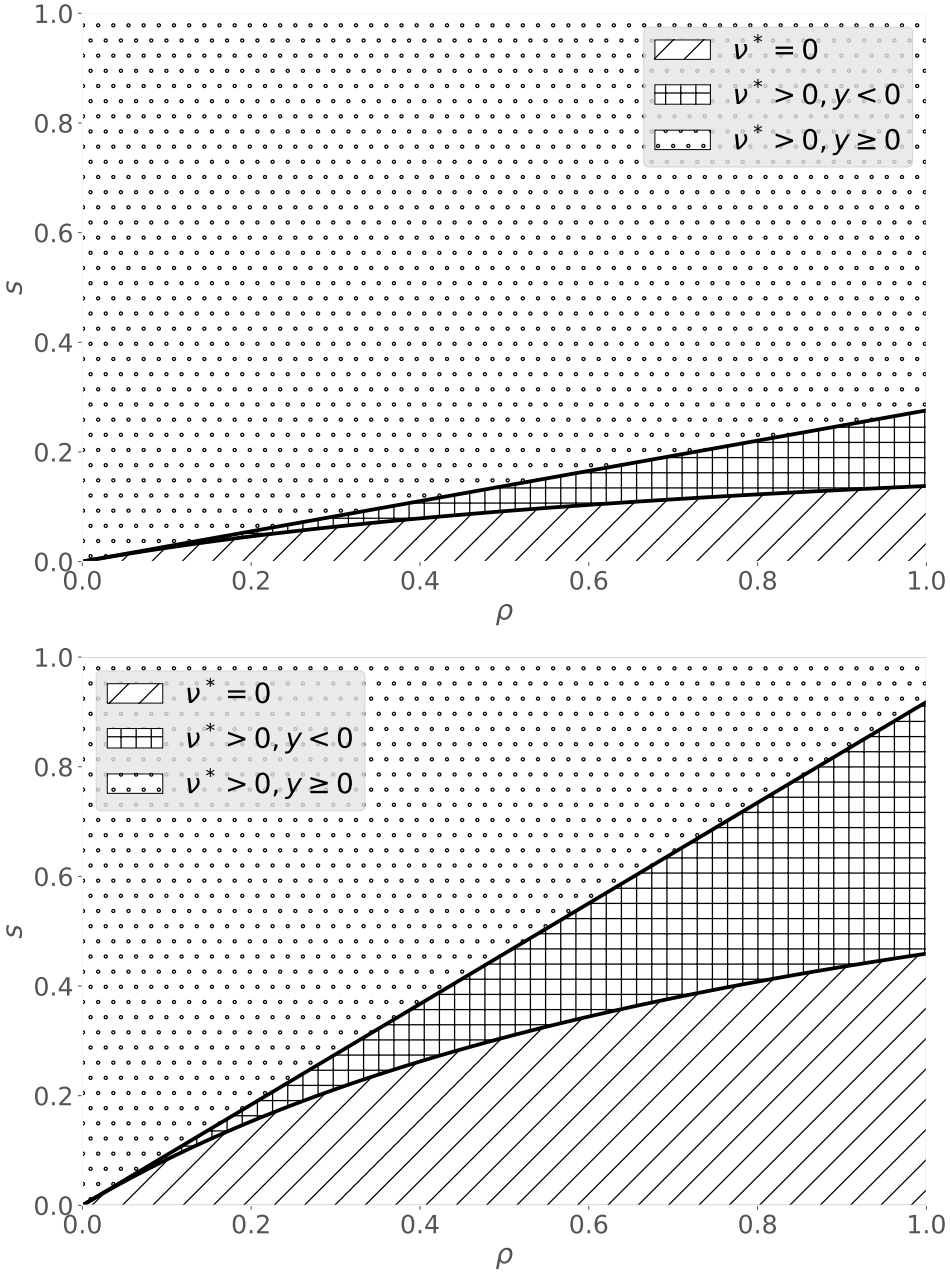

Figure 3.2: The partition of the parameter space (ρ, s) into the regimes where limit orders are never optimal (slashed area), optimal to implement contrarian strategies (crossed area) and optimal for direct liquidity provision for the underlying (dotted area). The upper panel corresponds to trading on the trade clock (with a volatility of price changes of \$0.11). The lower panel displays the corresponding results for trading every 10 ticks (with a volatility of \$0.37 between trades).

The optimally controlled wealth dynamics can be described as follows:

Corollary 3.3.8. *With the optimal tilted execution scheme ν_t^* from Lemma 3.3.5, the wealth process $X_t^{\nu^*}$ has dynamics*

$$dX_t^{\nu^*} = \begin{cases} \Pi_t dP_t - \frac{s_t}{\sqrt{2\pi}} |\gamma_t| \, dt & \text{if } \nu_t^* = 0, \\ \Pi_t dP_t - \frac{s_t}{\sqrt{2\pi}} |\gamma_t| \left(\operatorname{sgn}(\gamma_t) \rho_t y_t + \sqrt{(1-y_t^2)(1-\rho_t^2)} \right) dt & \text{if } \nu_t^* \in (0,1), \\ \Pi_t dP_t + \frac{s_t}{\sqrt{2\pi}} \left(y_t v_t - \sqrt{\gamma_t^2 + 2\rho_t v_t \gamma_t + v_t^2} \right) dt & \text{if } \nu_t^* = 1, \end{cases}$$

where we recall that $y_t = 1 - \rho_t \sigma_t \sqrt{2\pi}/s_t$.

In the first case of Corollary 3.3.8, no limit orders are used and the wealth dynamics collapse to the classical Leland model, where transaction costs accumulate proportionally to the bid-ask spread and the (absolute value of) the volatility of the target strategy Π_t.

In the second case, by capturing a fraction of the flow, the hedging cost can be reduced by a multiplicative factor that depends on y_t and the adverse selection parameter ρ_t. In the empirically relevant case $y_t \in (-1, 0)$ (where this case arises for $\gamma_t < 0$ and $\rho_t > -y_t$), differentiation shows that this factor is decreasing in ρ_t. Whence, hedging costs are reduced by higher adverse selection in this regime. The intuition is that, with strong adverse selection, the sign of the order flow is aligned with the contrarian nature of the target strategy for $\gamma_t < 0$.

In the third case of Corollary 3.3.8, an even larger limit order exposure than the entire flow would be desirable, so this "saturated" case interpolates smoothly between the first and second cases.

3.4 Impact on Option Prices

In Section 3.3, we have studied how to optimally implement delta hedging strategies corresponding to a *given* option price via tilted executions schemes that employ both market and limit orders.

Now, we turn to the question for what option prices this procedure actually *replicates* the option in the continuous-time limit, rather than producing surpluses or losses. To this end, we assume for simplicity that the risky asset has Bachelier dynamics,

$$dP_t = \mu dt + \sigma dW_t, \quad \text{for } \mu \in \mathbb{R} \text{ and } \sigma > 0.$$

For simplicity, we also assume that the bid-ask spread $s > 0$ is constant. Finally, in order to obtain option prices that only depend on time t and the price level P_t but not the cumulative flow F_t, we assume that the latter follows Brownian motion with drift:

$$dF_t = bdt + v\sqrt{1-\rho^2}dW_t^{\perp} - v\rho dW_t, \quad \text{for } b \in \mathbb{R},\ v > 0, \text{ and } \rho \in (0,1).$$

In this setting, the price from which the option can be replicated using market and limit orders via the optimal tilted execution scheme is described by a fully nonlinear PDE:

Lemma 3.4.1. *Consider a European option with payoff function $\varphi(P_T)$ at time T and suppose there is a classical solution $C \in \mathcal{C}^{1,2}([0,T) \times \mathbb{R}) \cap \mathcal{C}([0,T] \times \mathbb{R})$ of the following PDE:*

$$0 = \partial_t C + \frac{\sigma^2}{2}\partial_{pp}C - \sup_{\nu \in [0,1]} \left\{ \frac{s}{\sqrt{2\pi}} \left[\nu v - \sqrt{\sigma^2 \left(\partial_{pp}C\right)^2 + 2\rho\nu v\sigma\partial_{pp}C + \nu^2 v^2} \right] - \rho\sigma\nu v \right\}, \tag{3.9}$$

with terminal condition $C(T,p) = \varphi(p)$. Then, starting from the initial capital $C(0,P_0)$, the optimal tilted execution scheme from Lemma 3.3.5 replicates the option payoff (in probability) in the continuous-time limit.

Note that the option price (3.9) is nonlinear. In particular, the cost of replicating a long and short positions is generally not symmetric. To illustrate this, let us first consider the simplest case for solving the PDE (3.9): if the supremum is always attained by the same limit order exposure ν. This happens, for example, if the option payoff is convex, e.g., if one is trying to replicate a long position in call or put options. Then, limit orders are never optimal and, like in Leland's model, the unique solution of a linear equation

with a modified higher volatility also solves the nonlinear equation (3.9):

$$\partial_t C + \frac{\sigma^2}{2}\left(1 + \sqrt{\frac{2}{\pi}}\frac{s}{\sigma}\right)\partial_{pp} C = 0. \tag{3.10}$$

Indeed, convexity of the terminal condition implies $\partial^2_{pp} C \geq 0$. It in turn follows as in the proof of Lemma 3.3.5 that the supremum in (3.9) is aways attained by $\nu = 0$ for this option price, so that the nonlinear equation (3.9) is also satisfied by the solution of the linear equation (3.10).

Conversely, suppose that the option payoff is concave (e.g., for a short position in calls or puts), $y \in (-\rho, 0)$ (so that limit orders are optimal), and the option position is sufficiently small (to ensure it never reaches the saturated third case in Corollary 3.3.8). Then, the nonlinear PDE (3.9) is solved by the unique solution of another *linear* equation, now with smaller modified volatility:[14]

$$\partial_t C + \frac{\sigma^2}{2}\left(1 - \sqrt{\frac{2}{\pi}}\frac{s}{\sigma}\left(-\rho y + \sqrt{(1-\rho^2)(1-y^2)}\right)\right)\partial_{pp} C = 0. \tag{3.11}$$

Indeed, it follows from standard arguments that C inherits $\partial^2_{pp} C \leq 0$ from the concavity of the terminal condition. For a small option position, one can in turn verify as in Lemma 3.3.5 that the supremum in (3.9) is attained at an interior value $\nu^* \in (0, 1)$ for $\partial_{pp} C < 0$ and by $\nu^* = 0$ for $\partial_{pp} C = 0$. After inserting this maximizer (as well as the linear equation (3.11) for C), it follows that C also solves the nonlinear equation (3.9). We see that, compared to the Leland model (where just the sign of the volatility adjustment in (3.10) is flipped), limit orders reduce the shift in the implied volatility by a factor of $-\rho y + \sqrt{(1-\rho^2)(1-y^2)} \in (0, 1)$ in this case.

We now turn to the general case where the supremum in the nonlinear PDE (3.9) is not necessarily attained by the same value of ν, i.e., where the optimal limit order strategy possibly depends on time and price levels. As all model parameters are constant here, $y = 1 - \rho\sigma\sqrt{2\pi}/s$ is also constant. Whether are not it is optimal to use limit

[14] Like for our wellposedness result for general smooth payoffs in Theorem 3.4.2 below, this solution also requires that the bid-ask spread is small enough relative to the volatility, so that the adjusted volatility remains positive.

orders in turn only depends on time and the price level through the option positions γ_t or, equivalently, it's gamma $\partial_{pp}C(t,p)$. Indeed, if $y \in (-\rho, 0)$, for example, then the optimal limit-order placement policy is $\nu^*(\gamma_t) \equiv 0$ if $\gamma_t = \sigma\partial_{pp}C(t,P_t) \geq 0$ and $\nu^*(\gamma_t) = -\frac{\gamma_t}{v}\left(\rho + y\sqrt{\frac{1-\rho^2}{1-y^2}}\right) \in (0,1)$ if $\gamma_t < 0$.[15] Clearly, the mapping $\nu^*(\cdot)$ is then also the pointwise maximizer of the Hamiltonian in the PDE (3.9) (with $\gamma = \sigma\partial_{pp}C(t,p)$); Equations (3.10) and (3.11) are retrieved as special cases (where $\partial_{pp}C(t,p) \geq 0$ and $\partial_{pp}C(t,p) < 0$ for all $t \in [0,T]$, respectively). With these observations, we can show that a unique solution of the PDE (3.9) exists under the same parameter constraint required for wellposedness of the classical Leland model (Ishimura, 2010):[16]

Theorem 3.4.2. *Suppose $\sigma > s\sqrt{\frac{2}{\pi}}$. Then, there is a unique bounded solution $C \in \mathcal{C}^{1,2}([0,T)\times\mathbb{R}) \cap \mathcal{C}([0,T]\times\mathbb{R})$ of the PDE (3.9) for option payoffs $\varphi \in \mathcal{C}_b^3(\mathbb{R})$.*

3.5 Empirical Illustration

We now illustrate the quantitative properties of our theoretical results using tick data from the LOBSTER database. Since trading at every tick leads to unreasonably large transaction costs, we instead consider the times $\{t_n\}_{n=0}^N$ corresponding to every tenth tick of the trade clock.

3.5.1 Parameter Estimation

Throughout this empirical illustration, we consider a time horizon of $T = 1$ year, suppose that the (absolute) bid-ask spread s is constant, and prices and order flow have Bachelier dynamics:

$$dP_t = \mu dt + \sigma dW_t, \quad dF_t = bdt + v\sqrt{1-\rho^2}dW_t^{\perp} - v\rho dW_t.$$

For Google shares in 2019, we have 2552083 trades; trading every tenth tick thus corresponds to $\Delta t^N = 1/255208.3$. The bid-ask spread s_N can be proxied by the average value

[15] The conclusion that negative gamma enables limit order usage might be obvious, but here we should note that our result is more general.

[16] This parameter constraint ensures that transaction costs are not so high that hedging with market orders would lead to a negative bid-price for the options with a convex payoff

$s^N = s/\sqrt{N} = 0.369$ reported in the LOBSTER database. The volatilities and correlations can be estimated using realized quadratic (co-)variations: by, e.g., (Aït-Sahalia & Jacod, 2014, Theorem 1.14), we have:

$$\sigma^2 \approx \sum_{n=1}^{N} \left(P_{t_n} - P_{t_{n-1}}\right)^2, \quad v^2 \approx \sum_{n=1}^{N} \left(F_{t_n} - F_{t_{n-1}}\right)^2, \quad \rho\sigma v \approx \sum_{n=1}^{N} \left(F_{t_n} - F_{t_{n-1}}\right)\left(P_{t_n} - P_{t_{n-1}}\right).$$

Here, the order flow is proxied by the number of shares traded between the gridpoints. The volume of each individual trade is reported in the LOBSTER database; the signs are estimated using the Lee-Ready algorithm, see Appendix 3.8.3 for more details. For Google data from 2019, this leads to

$$\rho \approx 0.459, \quad \sigma \approx 0.366 \times \sqrt{N}, \quad v \approx 356.468 \times \sqrt{N}.$$

Remark 3.5.1. *The above statistical estimators are set up on the "trade clock" rather than "calendar time". This makes the estimation of all model parameters (volatility, adverse selection) significantly more robust.*[17] *One can, in fact, interpret the whole model under the trade clock. This has the additional advantage that the dynamics of volatility in the volume clock are significantly reduced, circumventing the need for a time-varying or stochastic volatility model, at least on the intraday timescale. The price to pay for this reinterpretation is that the expiry time of the option becomes a random variable under the trade clock. While this affects the hedging strategy over longer durations, it does not have a big impact on the intraday timescale. Furthermore, the volatility contribution of overnight returns, which cannot be captured via continuous-time trading, typically plays a much larger role over the timescales where the option expiry is felt. Accordingly, models in volume time are commonly used in a trading context.*[18]

[17] Aït-Sahalia and Jacod (2014) provide a comprehensive list of statistical results for high frequency data. They treat the case of "Volatility and Irregularly Spaced Observations" in Chapter 9 (p. 299) and focus on time changes in particular in Section 9.1.3 "Irregular Observations Schemes: Time Changes" (p. 309).

[18] See for example Busseti and Lillo (2012), who tackle the estimation of price impact under different clocks. They dedicate their entire Section 5 "The propagator model under other time measures" to this issue and provide in Table 3, p. 14 a direct comparison of the performance of price impact models under the trade and calendar time clocks.

For the above parameter estimates, the (multiplicative) Leland volatility correction when replicating options with a convex payoff is 44.3%. This is still high, but note that this is the value obtained without any netting of delta exposure for option market makers, compare Section 3.5.3.[19] For this trading frequency, the parameter constraints $y < 0$ and $\rho > -y$ are also satisfied, as $y = -0.142 \in (-\rho, 0)$.

To test the robustness of these findings over a larger universe of stocks, we also estimate the model parameters over the entire universe of S&P 500 stocks for 2019 to provide a cross-sectional view of adverse selection and its effect on limit order's profitability. The profitability depends crucially on the trading frequency used for estimation: if the trader is able to capture every tick or every other tick, unconditional market making is typically profitable. Conversely, even a trader capturing every tenth tick will lose money market making. However, this doesn't preclude the use of limit orders to cheapen delta-hedging strategies as long as the limit orders are less unprofitable than the bid-ask spread paid when taking liquidity. Figure 1.1 visualizes the distribution of spread and adverse selection across stocks. Figure 3.3 flattens this distribution into a one-dimensional density describing limit order's profitability as a function of trade frequency.

For limit orders to be useful in delta-hedging an option, Figure 1.1 shows that one looks for stocks with *high adverse selection* and *sizeable spread*:

(a) adverse selection helps limit orders hedge "in the right direction" for the correct gamma.

(b) the bid-ask spread compensates for the adverse selection losses. If the spread is too low, adverse selection costs may make limit orders more costly than market orders.

3.5.2 Optimal Policy

Now suppose that we are about to replicate a short position of 100 at-the-money European call options on Google with time-to-maturity $T = 1$ year. As recapitulated above, the

[19] There is no public data on the average netting of delta exposures for option market makers. However, we can point to the emergence of central risk books as an indication of the appetite for risk netting inside a firm.

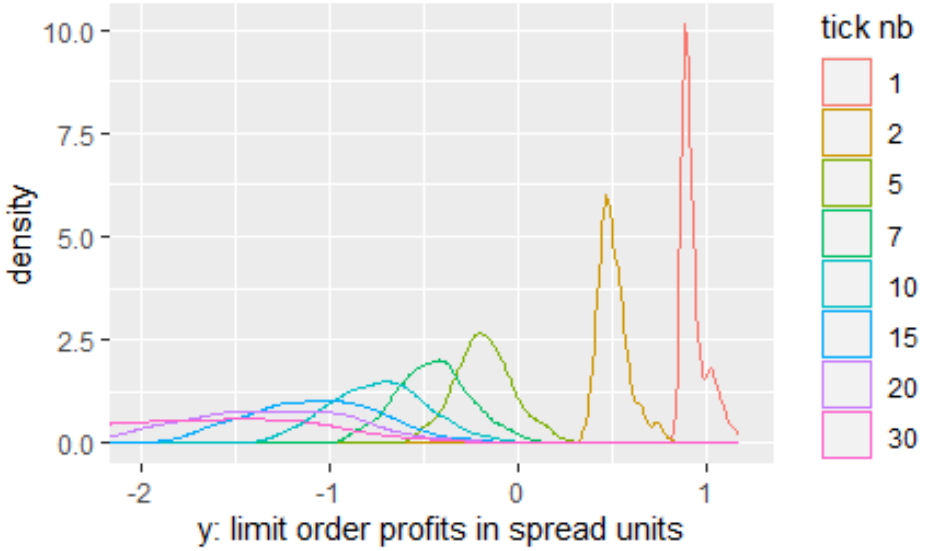

Figure 3.3: Density of y, the profitability of limit orders in spread-units, across the S&P 500 over 2019 for various trading frequencies.

volatility for the option bid price in the classical Leland model with market orders then is about 55.7% lower than in the frictionless version of the model. With both market and limit orders, Corollary 3.3.8 shows that the corresponding volatility correction is smaller, in that the "effective volatility" for pricing and hedging is only reduced by about 50.9%. Converted into option prices, the bid price for the 100 options when hedging with market orders only is \$3271.8453. If the agent hedges with both market and limit orders, this increased to the more competitive bid price \$3623.2647; the corresponding frictionless price is \$7381.8832.

Let us now get a feeling for the trading volumes involved. Recall that the optimal limit order exposure is

$$\nu_t^* = \max\left\{\min\{\tilde{\nu}_t^*, 1\}, 0\right\}, \quad \text{where } \tilde{\nu}_t^* = \begin{cases} -\frac{\gamma_t}{v}\left(\rho - \operatorname{sgn}(\gamma_t) y \sqrt{\frac{1-\rho^2}{1-y^2}}\right), & \text{if } y \in (-1,1), \\ 0, & \text{if } y \leq -1. \end{cases}$$

Taking into account that γ_t is the product of the number of options, the price volatility and the Bachelier at-the-money gamma $\phi(0)/\sigma_{\text{eff}}\sqrt{T} = 1/\sigma_{\text{eff}}\sqrt{2\pi}\sqrt{T}$, this leads to

$$\nu^* = 1.49575 \times 10^{-4}.$$

The expected order flow in each timestep is 284.4, so that the number of shares captured by limit orders is about 0.04254 every 10 ticks, equivalent to 43.08 shares per day. When using limit orders in this manner, the expected number of shares traded by market orders is reduced from 143.97 per day in the classical Leland model to 129.23 per day.

3.5.3 Netting and Aggregation in Central Risk Books

The previous example focuses on a single option for simplicity. To illustrate the additional effects that appear when delta hedging an entire book of different options, we now consider two further examples. The first focuses on netting of long and short positions; the second demonstrates the impact of aggregate payoffs that are neither convex nor concave.

Example 3.5.2. *Consider a market maker for an option with convex payoff, say a call or put. Client buy orders for $Q_a > 0$ options create a short position for the market maker, who in turn needs to replicate a convex payoff for which no limit orders are used. However, client sell orders for $Q_b > 0$ options create a long position for the market maker, for which limit orders help to hedge the corresponding concave payoff.*

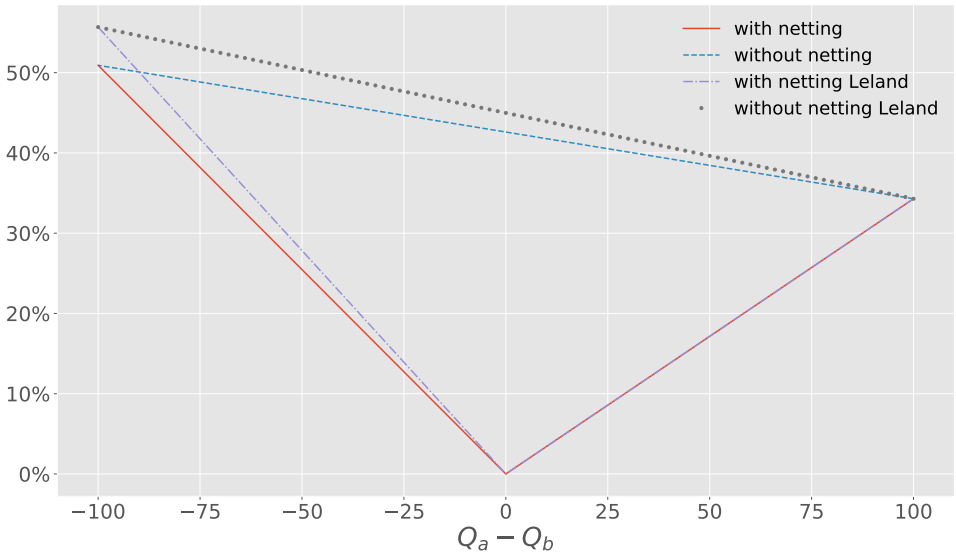

Figure 3.4: Percentage of additional option replications costs relative to the frictionless version of the model. The gross number of options is fixed to $Q_a + Q_b = 100$ throughout.

Without netting in a *central* risk book, the long and short positions of the market maker are delta-hedged separately. As the option payoff is convex, the replication costs

for long and short positions are both linear in the position sizes Q_a and $-Q_b$. However, due to transaction costs, the replication cost C_a per option for client buy orders (i.e., the option ask-price (3.10) with higher volatility) is higher than its frictionless counterpart C. Conversely, the value of the replication portfolio C_b per option for client sell orders is lower than C (the option bid-price with lower volatility (3.11)). Without netting, the additional replication costs due to transaction costs therefore are

$$Q_a(C_a - C) - Q_b(C_b - C).$$

With netting in a central risk book, only the net position $Q_a - Q_b$ is delta hedged. This reduces the additional replication costs to

$$(Q_a - Q_b)\mathbf{1}_{\{Q_a > Q_b\}}(C_a - C) - (Q_b - Q_a)\mathbf{1}_{\{Q_b > Q_a\}}(C_b - C).$$

With only buy or sell orders, this reduces to the formula without netting above. But if client buy and sell orders offset completely, then netting allows to avoid hedging and in turn transactions costs altogether. If client buy and sell orders partially offset each other, then the additional replication costs lie between these upper and lower bounds. Figure 3.4 illustrates this for our Google example when the gross position of options traded $Q_a + Q_b$ is fixed to 100, but the market maker's net position $Q_a - Q_b$ varies from -100 to 100.

Example 3.5.3. *As a concrete example for a* portfolio *of different options, we now consider a "butterfly". This corresponds to a long position of N call options each with strikes $K_1 < K_3$ and a short position of $2N$ call options with strike $K_2 \in (K_1, K_3)$. The resulting payoff function is concave around K_2, but convex outside this neighbourhood. As a result, the nonlinear pricing PDE* (3.9) *no longer has a simple explicit solution but instead needs to be solved numerically. This can be done using finite differences, which also produces the optimal limit-order policy as a byproduct.*

For our Google example with $S_0 = 1000$, $K_1 = 900$, $K_2 = 1000$, and $K_3 = 1100$, and $N = 100$, the replication price of the butterfly in a central risk book then is \$5354.3359. *In contrast, if each of the positions is delta hedged separately, the sum of the three separate*

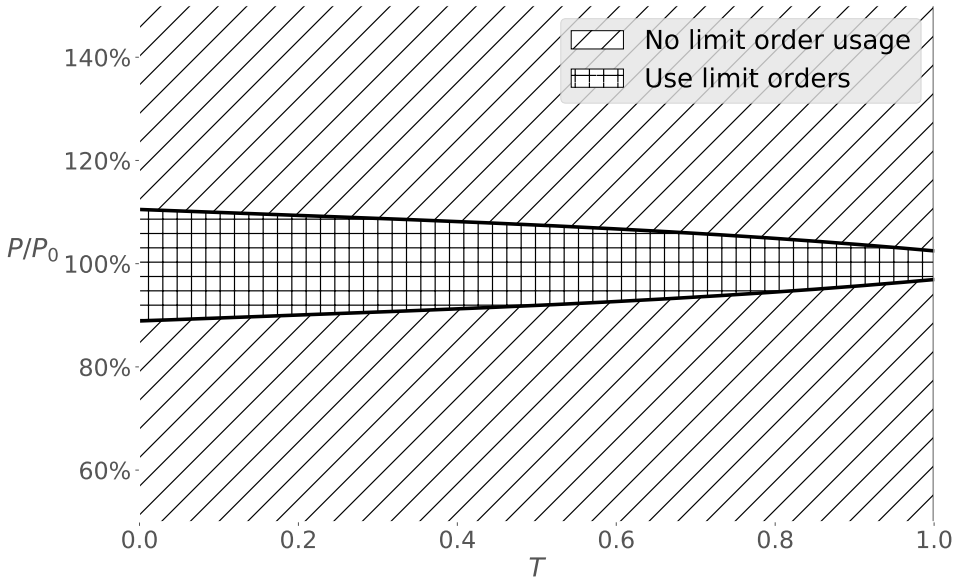

Figure 3.5: Limit-order usage for butterfly option.

replication prices is \$14165.0495. *This is another example for the same netting effect as in the previous example.*

Here, aggregation in a central risk book also affects the optimal limit order policy. To wit, since the aggregate payoff function of the option portfolio is neither convex nor concave, limit orders are now optimal in some states but not in others. For the Google example, this is illustrated in Figure 3.5. We see that limit orders are optimal if the current stock price is near the middle strike K_2, where the payoff of the butterfly is locally concave. Outside of this region, it is optimal to hedge with market orders only.

3.5.4 Simulation Results

Both the optimal choice of limit orders and the analysis of the corresponding option prices studied above crucially depend on the passage to a continuous-time scaling limit. For the model parameters estimated in this section, we now test how well this approximates the discrete version of the model by means of a simulation study.

To this end, we simulate 10,000 price and order flow paths with Bachelier dynamics, using the parameter estimates for Google tick data obtained above. The Bachelier delta hedge has the explicit form $\Pi_t = \Phi(P_t - K)/\sigma_{\text{eff}}\sqrt{T-t}$, for a call option with strike K, time-to-maturity $T - t$, and the "effective" volatility σ_{eff} adjusted for transaction

costs. The discrete P&L corresponding to an option position and the associated hedging portfolio can in turn be computed conveniently for each simulated price path.

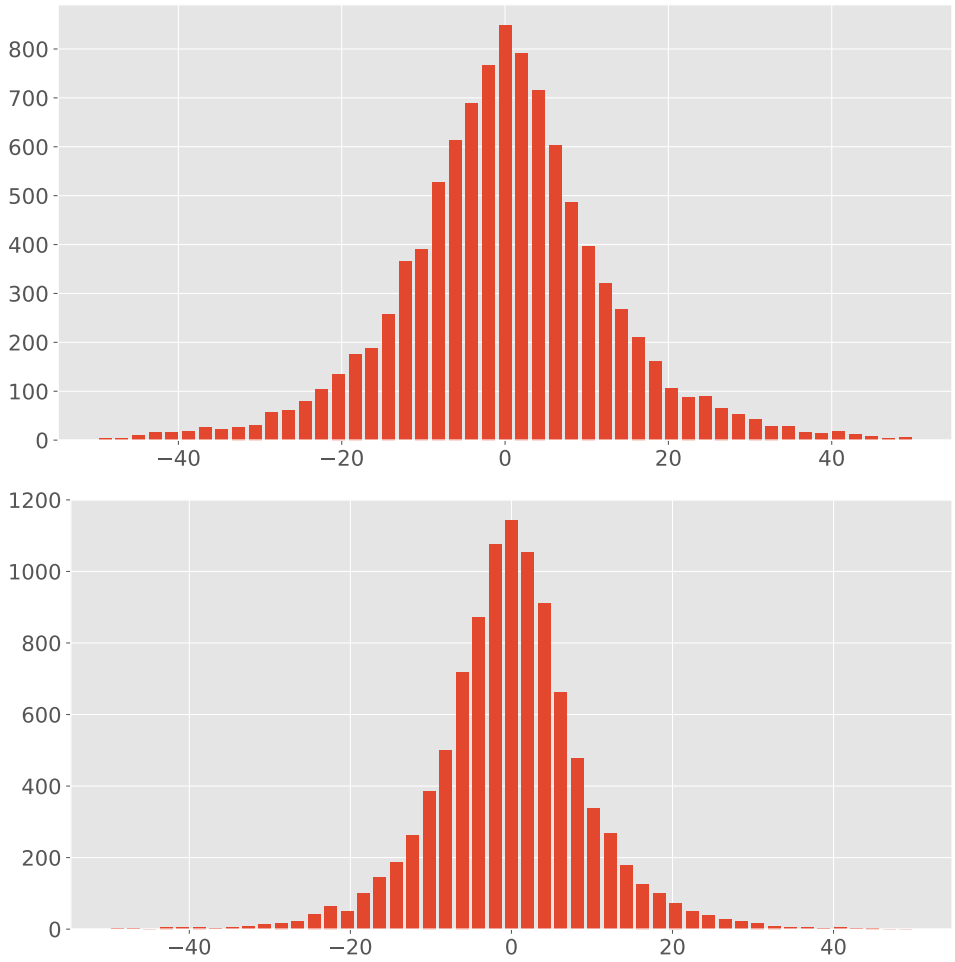

Figure 3.6: Distribution of the hedging error for the classical Leland model (upper panel) and for the extension with market and limit orders (lower panel).

Figure 3.6 displays the distribution of hedging errors for the classical Leland model with market orders only, and for our extension with market and limit orders. For the 100 at-the-money call options we consider, the hedging errors are small in line with the continuous-time replication results, both for the classical Leland model and for our extension with limit orders. More specifically, in the classical Leland model the average hedging error is -0.1255 with a standard deviation of 13.168; for the extended model with limit orders the corresponding values are -0.0986 and 9.3931, respectively. As reported above, the corresponding prices are \$7381.8832 in the frictionless version of the model, \$3271.8453 with market orders only, and \$3623.2647 for hedging with both market and

limit orders.

Finally, Figure 3.7 plots the histogram of cost savings that can be achieved using limit orders.

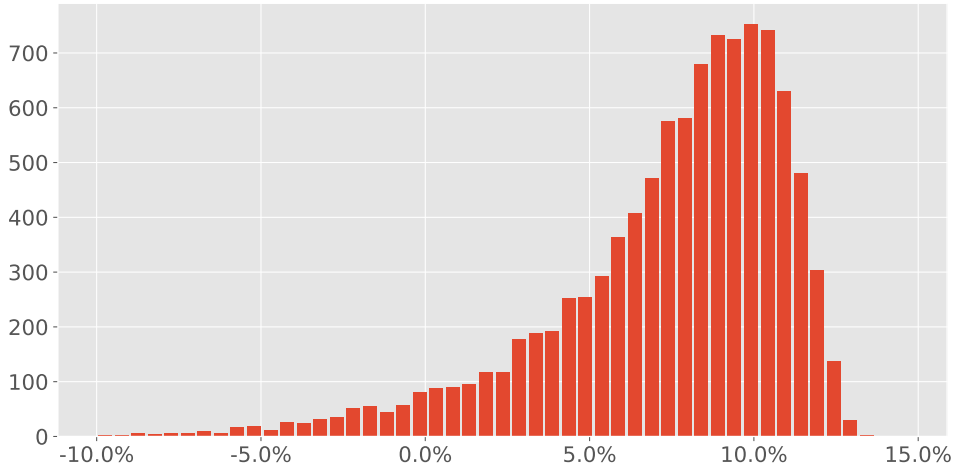

Figure 3.7: Distribution of (relative) savings of replication cost that can be achieved using limit orders.

3.6 Partial Tilted Execution Schemes

In the *tilted execution schemes* discussed so far, market orders are used at each trading time to perfectly rectify the discrepancy between the actual position (after the execution of limit orders) and the target (3.5). This provides a tight control on the associated hedging errors, but generates substantial trading costs since a large amount of trading is still done via market orders.

In order to reduce the amount of transaction costs paid, it is natural to extend this analysis to *partial tilted execution schemes*, where only a fraction of the hedging error is offset at each time step. Put differently, both the exposure to limit orders and the fraction of market orders placed become control variables in this context.[20]

To formalize this, note that before market orders are placed at time t_n^N, the target position has changed by $\Delta\Pi_n^N$ and the hedging position has changed by $\nu_{n-1}^N \Delta F_n^N$ due

[20]The even more general case of a completely free choice of market and limit orders needs to be studied by numerically solving the corresponding dynamic-programming equations (Agliardi, 2016; Cartea et al., 2019; Ellersgaard & Tegnér, 2017). In contrast, the approach we present here retains most of the tractability of the standard Leland model by still reducing optimization to a pointwise problem.

to the execution of limit orders. Accordingly, the previous displacement D^N_{n-1} between actual position and target has changed by $\Delta\Pi^N_n - \nu^N_{n-1}\Delta F^N_n$. Whence, to rectify a fraction $\kappa^N_n \in [0,1]$ of the updated displacement, the market order that needs to be placed at time t^N_n is

$$\Delta M^N_n = \kappa^N_n(D^N_{n-1} + \Delta\Pi^N_n - \nu^N_{n-1}\Delta F^N_n).$$

With this partial tilted execution scheme, the corresponding displacement has the following autoregressive dynamics, controlled by the limit order exposure ν and the market-order fraction κ:

$$D^N_n = (1-\kappa^N_n)(D^N_{n-1} + \Delta\Pi^N_n - \nu^N_{n-1}\Delta F^N_n), D^N_0 = 0. \tag{3.12}$$

The corresponding portfolio dynamics of such a partial tilted execution scheme are

$$\begin{aligned}\Delta X^N_n =&(\Pi^N_{n-1} - D^N_{n-1})\Delta P^N_n - \frac{s^N_n}{2}\kappa^N_n|D^N_{n-1} + \Delta\Pi^N_n - \nu^N_{n-1}\Delta F^N_n| \\ &+ \frac{s^N_{n-1}}{2}\nu^N_{n-1}|\Delta F^N_n| + \nu^N_{n-1}\Delta F^N_n\Delta P^N_n.\end{aligned} \tag{3.13}$$

As for full tilted execution schemes, the first term measures the effect of price changes, the second and third terms records spread costs paid and earned for market and limit orders, respectively, and the last term accounts for adverse selection. As long as the fractions κ^N_n of displacements rectified by market orders are bounded away from zero, the sum of the first terms will still converge to the stochastic integral of the target strategy with respect to the price process in the continuous-time limit. One can in turn trade off the effect of the execution costs described by the other three terms against a functional of the displacement of the actual positions from the target strategy. The most tractable specification is to subtract a multiple of the sum of expected squared displacements:

$$\frac{\alpha}{2}\mathbb{E}\left[\sum_{n=1}^{N}(D^N_n)^2\right]. \tag{3.14}$$

Here, the relative importance of execution costs and displacements is measured by a

risk-aversion parameter $\alpha > 0$.[21]

3.6.1 Continuous-Time Limit

To compute the continuous-time limit of the discrete-time portfolio dynamics (3.13), we need the following mild regularity conditions, which ensure sufficient continuity and integrability of all processes involved:

Assumption 3.6.1. *Suppose $\mu^F, v, \mu, \sigma, s, \rho, \mu^\Pi, \gamma, \nu, \kappa, F, P, \Pi$ are all Itô processes with uniformly bounded drift and diffusion coefficients and, in addition, κ is uniformly bounded away from zero.*

Then, we have the following convergence result:

Theorem 3.6.2. *Suppose Assumption 3.6.1 is satisfied. Then, as the timestep Δt^N between the trading times tends to zero, the expectation of the discrete-time portfolio process* (3.13) *converges to*

$$\mathbb{E}\left[\int_0^T \Pi_t dP_t - \int_0^T \left(\frac{s_t}{2}\sqrt{\gamma_t^2 + 2\rho_t\gamma_t\nu_t v_t + \nu_t^2 v_t^2}\sqrt{\frac{2}{\pi}}\sqrt{\frac{\kappa_t}{2-\kappa_t}} - \frac{s_t\nu_t}{2}\sqrt{\frac{2}{\pi}}v_t + \nu_t\rho_t\sigma_t v_t\right) dt\right]. \tag{3.15}$$

The sum of expected squared displacements (3.14) *converges to*

$$\mathbb{E}\left[\int_0^T \left(\left(\gamma_t^2 + 2\rho_t\gamma_t\nu_t v_t + \nu_t^2 v_t^2\right)\frac{(1-\kappa_t)^2}{\kappa_t(2-\kappa_t)}\right) dt\right]. \tag{3.16}$$

Even though the explicit expressions in Theorem 3.6.2 are more involved than in Theorem 3.3.1, the optimal trading policy that trades off execution costs against average displacements in an optimal manner can thus still be determined by direct *pointwise* minimization of

$$\mathbb{E}\left[\int_0^T H(\nu_t, \kappa_t, \gamma_t, s_t, v_t, \rho_t) dt\right],$$

[21] Note that α is *not* rescaled by the length time steps. Whence, this goal functional corresponds to the regime of small transaction costs and high risk aversion studied by Barles and Soner (1998).

where

$$H(\nu,\kappa,\gamma,s,v,\rho) = \frac{s}{2}\sqrt{\gamma^2+2\rho\gamma\nu v+\nu^2v^2}\sqrt{\frac{2}{\pi}}\sqrt{\frac{\kappa}{2-\kappa}} - \frac{s\nu}{2}\sqrt{\frac{2}{\pi}}v + \nu\rho\sigma v + \frac{\alpha}{2}\left(\gamma^2+2\rho\gamma\nu v+\nu^2v^2\right)\frac{(1-\kappa)^2}{\kappa(2-\kappa)}. \tag{3.17}$$

As a sanity check, first consider the above expression in the limiting case where the risk-aversion parameter α is sufficiently large. Then, the influence of κ_t on the goal functional is dominated by the last term in (3.17). This term is decreasing in κ, and therefore minimized on $[0,1]$ by a full tilted execution scheme with $\kappa_t = 1$. In fact, this regime is not only obtained in the limit $\alpha \to \infty$, but already for large but finite values of α. For such sufficiently high levels of risk aversion, the goal functional (3.17) then collapses to its counterpart (3.7) for full tilted execution schemes and we recover all the results from the previous sections.

Conversely, suppose the risk aversion parameter α is sufficiently small. Then, the impact of κ_t is governed by the first term, which is increasing in κ_t and therefore minimized by $\kappa_t = 0$. As a consequence, sufficiently risk tolerant investors that are not worried about hedging their option position should never place any market orders. Limit orders should also never be used for hedging, but only if direct liquidity provision in the underlying is profitable on average for $y > 0$.

For intermediate values of the risk-aversion parameter α, there is a nontrivial tradeoff between using market orders to reduce displacement, but not doing so too aggressively in order to limit the associated transaction costs. For the parameters estimated from Google data in Section 3.5, Figure 3.8 plots the optimal limit order exposure ν_t and the corresponding market order fraction κ_t against the level α of risk aversion. We see that as risk aversion decreases, the optimal limit order exposure is initially rather insensitive. In contrast the amount of market orders quickly decreases when some deviations from the target position are acceptable. As the risk aversion decreases further, trading with either market or limit orders stops altogether, as their benefits for hedging risk then become smaller and smaller relative to their spread and adverse selection costs.

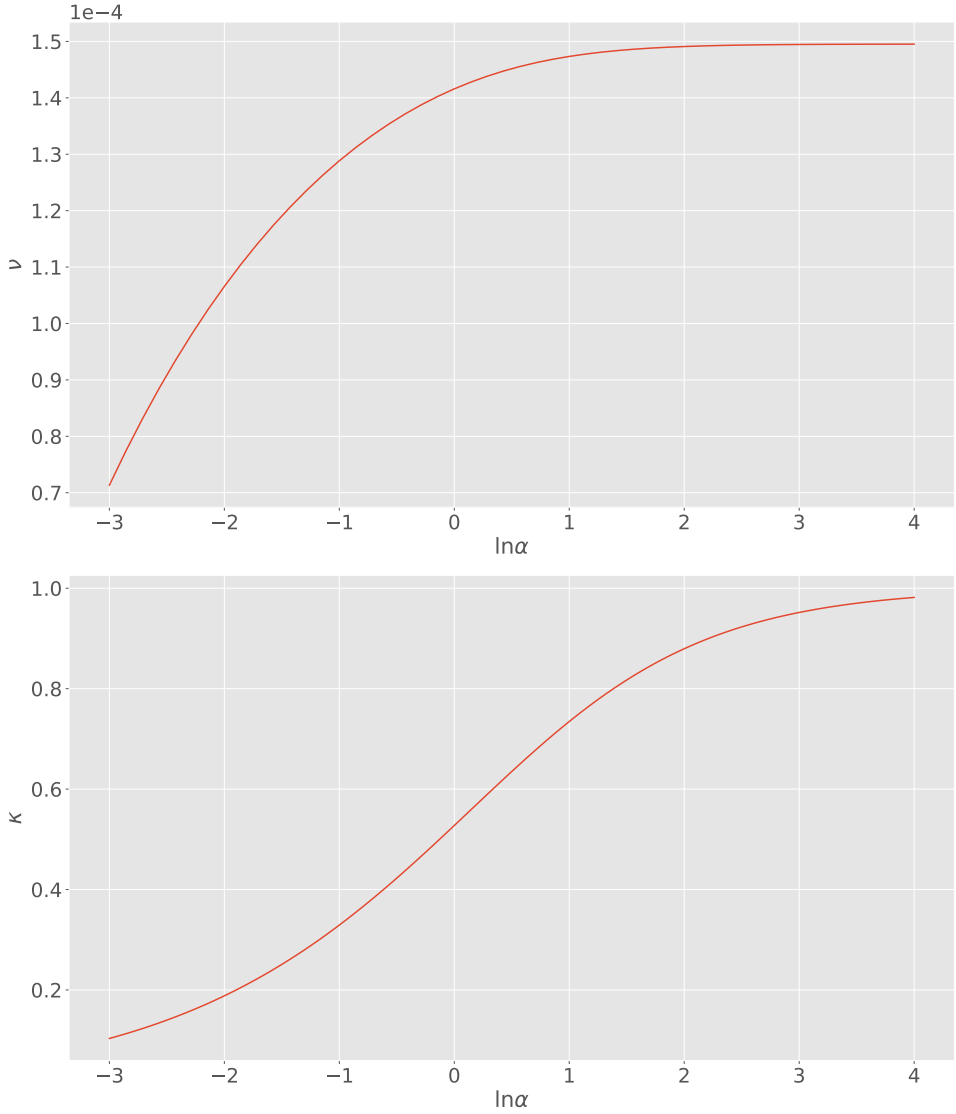

Figure 3.8: Optimal limit order exposure ν_t (upper panel) and optimal market order fraction κ_t (lower panel) plotted against risk aversion α. Other model parameters are as in Section 3.5.

3.6.2 Impact on Option Prices

We now sketch how to endogenize the target strategy in the goal functional (3.17) and determine a corresponding option price. As in Section 3.4, we focus on delta hedges $\partial_p C(t, P_t)$ derived from option prices $C(t, P_t)$ with fixed s, v and ρ. With a nontrivial tradeoff between execution costs and displacements, it is then natural to look for the "indifference price" (Hodges & Neuberger, 1989), for which the optimal performance with the option is the same as without it.

Suppose we have sold one option with payoff $\varphi(P_T)$ and hedge it starting from the initial option price $C(0, P_0)$. If this is implemented using a partial tilted execution scheme (ν, κ) for the delta hedge $\partial_p C(t, P_t)$ then, by Theorem 3.6.2, in the continuous-time limit, the expected payoff of the hedged portfolio is

$$\mathbb{E}\left[-\varphi(P_T) + C(0, P_0) + \int_0^T \partial_p C(t, P_t) dP_t + \int_0^T H\left(\nu_t, \kappa_t, \sigma \partial_{pp} C(t, P_t), s, v, \rho\right) dt\right]. \tag{3.18}$$

For the *indifference* price C no policy should be able to generate expected gains, but there should also be no expected losses for the optimal policy (in our class). Put differently, the expectation (3.18) should be nonpositive for any policy (ν, κ), and zero for the optimal one. Like the replication price with full tilted execution schemes, this indifference price is characterized by a fully nonlinear PDE:

Lemma 3.6.3. *Suppose there is a smooth solution $C \in \mathcal{C}^{1,2}([0, T) \times \mathbb{R}) \cap \mathcal{C}([0, T] \times \mathbb{R})$ of the PDE*

$$0 = \partial_t C + \frac{\sigma^2}{2} \partial_{pp} C - \sup_{\nu, \kappa \in [0,1]} H\left(\nu, \kappa, \sigma \partial_{pp} C, s, v, \rho\right), \qquad C(T, p) = \varphi(p). \tag{3.19}$$

Then, the expected payoff (3.18) *is nonpositive for any policy (ν, κ) and zero for the optimal one, the pointwise optimizer of $H(\nu, \kappa)$.*

We conjecture that a smooth solution of (3.19) exists for sufficiently regular coefficients and terminal conditions just like for the corresponding equation for full tilted execution schemes. However, this is considerably more difficult to prove in the absence of an

explicit expression for the pointwise optimum. We therefore defer further analysis of the PDE (3.19) to future research.

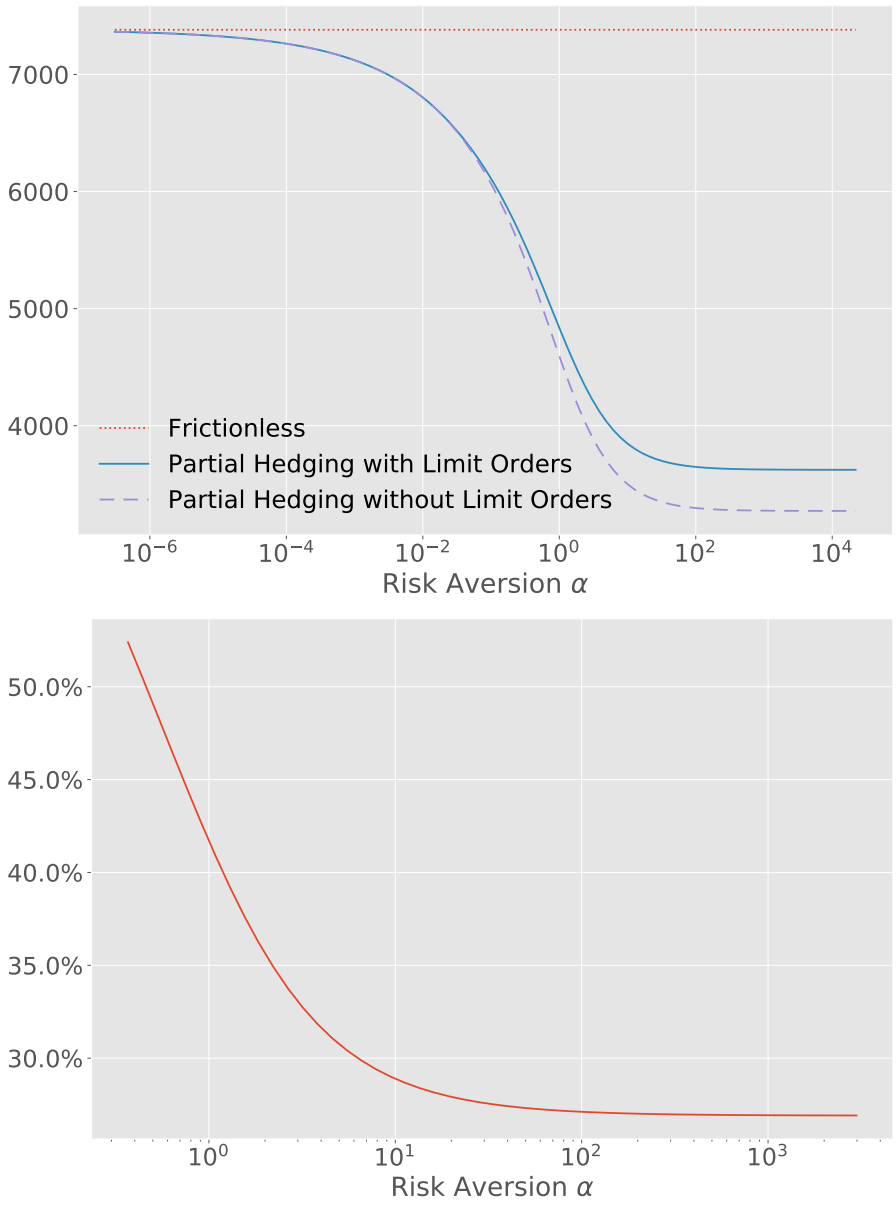

Figure 3.9: Upper panel: Indifference prices $C(0, P_0)$ (with and without using limit orders) plotted against risk aversion α. Lower panel: (expected) fraction of trades implemented via limit orders at time $t = 0$ plotted against risk aversion α. Other model parameters are as in Section 3.5.

3.6.3 Empirical Results

Using standard finite-difference solvers, the PDE (3.19) can be readily solved numerically.

For the parameters estimated from tick data for Google in Section 3.5, the upper panel of

Figure 3.9 shows how the resulting indifference prices with a long call position (i.e., the "bid price" for the option) depend on the risk-aversion parameter α. As discussed above, sufficiently large values of this parameter recover full tilted execution schemes, and the indifference price (3.19) accordingly converges to its counterpart from Section 3.4. If risk aversion is low, the trading frictions matter less and less and the option price in turn converges to its frictionless counterpart. Intermediate levels of risk aversion interpolate smoothly between these limiting cases. For each level of risk aversion, the indifference bid-price with limit orders is sandwiched between its counterpart using market orders only and the frictionless price. For the corresponding ask price, limit orders are never optimal and both indifference prices therefore coincide.

Finally, the lower panel of Figure 3.9 displays how the expected fraction of orders traded via limit orders at the initial time $t = 0$ varies as a function of the risk-aversion parameter. We see that limit orders make up a larger part of the overall trading activity as risk aversion decreases. The intuition for this is the behaviour of the optimal policy from Figure 3.8: as risk aversion decreases, the optimal limit order rate ν remains relatively insensitive, whereas the fraction of displacements κ offset by market orders is reduced much more quickly.

3.7 Conclusion

This chapter studies the derivatives hedging problem from the perspective of a central risk book with reasonable trading frequency, where the risky positions of several trading desks are managed collectively. We propose a reduced-form model that captures the trade-off between spread and adverse selection. Using limit theorems in the spirit of Jacod and Protter (2012), we show that for frequent trades and small transaction costs, the limiting wealth dynamics can be computed in closed form. In the limiting model, the optimization task becomes myopic and we provide the corresponding operationalizable derivatives hedging strategy with both market and limit orders.

We show that limit order is only beneficial for contrarian strategies. More specifically,

it is only beneficial for delta-hedging when the gamma of the risky position is negative. The results of Leland (1985) are generalized in terms of limit order placement (with adverse selection) and the corresponding option price is described by a nonlinear PDE with unique solution. Moreover, we further generalize this result by allowing imperfect hedging and derive another nonlinear PDE characterizing the interplay between transaction costs, adverse selection and risk aversion.

Our empirical analysis demonstrates that tactical liquidity provision is beneficial for non-competitive market makers at reasonable trading frequencies. Additionally, our empirical findings show that the reduced form model summarizes the market microstructural features needed for implementation of the trading strategy, with better tractability and interpretability compared to the purely numerical simulations at the discrete microstructure level.

3.8 Appendix

3.8.1 Proofs for Section 3.3

Proof of Theorem 3.3.1. We separately consider each of the four terms on the right-hand side of (3.6). The first term is a discrete stochastic integral and therefore converges to the continuous-time stochastic integral as $N \to \infty$ (cf., e.g., (Jacod & Shiryaev, 2003, Theorem I.4.31)):

$$\sum_{n=1}^{\lfloor \frac{tN}{T} \rfloor} \Pi_{t_{n-1}^N} (P_{t_n^N} - P_{t_{n-1}^N}) \overset{\text{u.c.p.}}{\Longrightarrow} \int_0^t \Pi_\tau dP_\tau.$$

For the second term on the right-hand side of (3.6), first notice that, by the Law of Large Numbers with normalization for discretized processes (Jacod & Protter, 2012, Theorem 7.2.2(a)):

$$\frac{1}{2\sqrt{N}} \sum_{n=1}^{\lfloor \frac{tN}{T} \rfloor} s_{t_{n-1}^N} \left| \Pi_{t_n^N} - \Pi_{t_{n-1}^N} - \nu_{t_{n-1}^N} (F_{t_n^N} - F_{t_{n-1}^N}) \right| \overset{\text{u.c.p.}}{\Longrightarrow} \int_0^t \frac{s_\tau}{\sqrt{2\pi}} \sqrt{\gamma_\tau^2 + 2\rho_\tau \gamma_\tau \nu_\tau v_\tau + \nu_\tau^2 v_\tau^2} d\tau.$$

We note here that the application of the Law of Large Numbers with normalization for discretized processes (Jacod & Protter, 2012, Theorem 7.2.2(a)) here introduces an extra factor of $1/\sqrt{T}$, which did not appear in the original version in the book Jacod and Protter (2012) in which $T = 1$ is assumed for simplicity.

The triangle inequality and (Jacod & Protter, 2012, Theorem 7.2.2(b)) in turn show that the second term on the right-hand side of (3.6) (where $s_{t_{n-1}^N}$ is replaced by $s_{t_n^N}$ in the above expression) converges to the same limit.

Next, for the third term we have

$$\frac{1}{2\sqrt{N}} \sum_{n=1}^{\lfloor \frac{tN}{T} \rfloor} s_{t_{n-1}^N} \nu_{t_{n-1}^N} |F_{t_n^N} - F_{t_{n-1}^N}| \overset{\text{u.c.p.}}{\Longrightarrow} \frac{1}{\sqrt{2\pi}} \int_0^t s_\tau \nu_\tau v_\tau d\tau,$$

again by (Jacod & Protter, 2012, Theorem 7.2.2(a)).

Finally, another application of (Jacod & Protter, 2012, Theorem 7.2.2(b)) shows that the fourth term on the right-hand side of (3.6) converges to

$$\sum_{n=1}^{\lfloor \frac{tN}{T} \rfloor} \nu_{t_{n-1}^N} (F_{t_n^N} - F_{t_{n-1}^N})(P_{t_n^N} - P_{t_{n-1}^N}) \overset{\text{u.c.p.}}{\Longrightarrow} -\int_0^t \rho_\tau \sigma_\tau \nu_\tau v_\tau d\tau.$$

Together, these four convergence results establish the asserted continuous-time limiting dynamics (3.7) of the discrete self-financing equation (3.6) □

Proof of Lemma 3.3.5. Fix a time t and state $\omega \in \Omega$. For $\ell_t = \nu_t v_t$ we need to maximize

$$G(\ell_t) := \frac{s_t}{\sqrt{2\pi}} \left(\ell_t - \sqrt{\gamma_t^2 + 2\gamma_t \ell_t \rho_t + \ell_t^2} \right) - \rho_t \sigma_t \ell_t, \quad \text{over } \ell_t \in [0, v_t]. \tag{3.20}$$

If $\gamma_t = 0$, this function is strictly decreasing (as $\rho_t \in (0, 1)$) and therefore maximized by $\ell_t = 0$ in accordance with (3.8). For $\gamma_t \neq 0$, differentiation shows that $G_t(\cdot)$ is strictly concave. Therefore, it has a unique maximizer on $[0, v_t]$ that is either characterized by the first-order condition $G_t'(\ell_t) = 0$ or coincides with one of the endpoints of the interval.

We have

$$G_t'(\ell_t) = \frac{s_t}{\sqrt{2\pi}} \left[y_t - \frac{\ell_t + \rho_t\gamma_t}{\sqrt{(1-\rho_t^2)\gamma_t^2 + (\ell_t + \rho_t\gamma_t)^2}} \right]. \tag{3.21}$$

As $\rho_t \in (0,1)$ and in turn $-\frac{\ell_t+\rho_t\gamma_t}{\sqrt{(1-\rho_t^2)\gamma_t^2+(\ell_t+\rho_t\gamma_t)^2}} \leq 1$, the function $G_t(\cdot)$ is strictly decreasing in ℓ_t in the case $y_t \leq -1$. Hence, the corresponding maximizer is $\ell_t = 0$ and the maximizer of (3.7) is $\nu_t^* = 0$ in this case.

We now consider the remaining case $y_t \in (-1,1)$. As $s_t > 0$, the first-order condition $G_t'(\ell_t) = 0$ is then equivalent to

$$y_t = \frac{\ell_t + \rho_t\gamma_t}{\sqrt{(1-\rho_t^2)\gamma_t^2 + (\ell_t + \rho_t\gamma_t)^2}}. \tag{3.22}$$

On the one hand, this requires $\operatorname{sgn}(y_t) = \operatorname{sgn}(\ell_t + \rho_t\gamma_t)$; on the other hand some algebraic manipulations show that we also need

$$\sqrt{1-y_t^2}|\ell_t + \rho_t\gamma_t| = |y_t|\sqrt{1-\rho_t^2}|\gamma_t|.$$

As $\operatorname{sgn}(y_t) = \operatorname{sgn}(\ell_t + \rho_t\gamma_t)$, this can be rewritten as

$$\sqrt{1-y_t^2}(\ell_t + \rho\gamma_t) = y_t\sqrt{1-\rho_t^2}|\gamma_t| \quad \Rightarrow \quad \ell_t = -\rho_t\gamma_t + y_t|\gamma_t|\sqrt{\frac{1-\rho_t^2}{1-y_t^2}}.$$

If this solution of the first-order condition lies in the interval $[0, v_t]$, then it is the unique maximizer of $G_t(\ell_t)$ on this interval as asserted in (3.8). If it lies outside this interval, then the maximizer of this concave function is the nearest endpoint of the interval, again in accordance with (3.8). This completes the proof. □

Proof of Lemma 3.3.6. Fix a time t and state $\omega \in \Omega$.

(i) First suppose that $y_t \in (-1,0)$. Then, if $\gamma_t > 0$, $\rho_t > 0$ implies

$$\tilde{\nu}_t^* = -\frac{\gamma_t}{v_t}\left(\rho_t - y_t\sqrt{\frac{1-\rho_t^2}{1-y_t^2}}\right) < 0. \tag{3.23}$$

As a consequence, we always have $\nu_t^* = 0$ in this case.

If $\gamma_t < 0$, then $\tilde{\nu}_t^* > 0$ (and in turn $\nu_t^* > 0$) is equivalent to

$$-\frac{\gamma_t}{v_t}\left(\rho_t + y_t\sqrt{\frac{1-\rho_t^2}{1-y_t^2}}\right) > 0 \quad \Leftrightarrow \quad \rho_t + y_t\sqrt{\frac{1-\rho_t^2}{1-y_t^2}} > 0 \quad \Leftrightarrow \quad \frac{\rho_t}{\sqrt{1-\rho_t^2}} > \frac{-y_t}{\sqrt{1-y_t^2}}.$$

As the function $x \mapsto x/\sqrt{1-x^2}$ is strictly increasing on $[0,1]$, this is in turn equivalent to $\rho_t > -y_t$ as claimed.

(ii) Now suppose that $y_t \in [0,1)$ (and recall that $\rho_t > 0$). Then, if $\gamma_t > 0$, $\tilde{\nu}_t^* > 0$ is equivalent to

$$-\frac{\gamma_t}{v_t}\left(\rho_t - y_t\sqrt{\frac{1-\rho_t^2}{1-y_t^2}}\right) > 0 \quad \Leftrightarrow \quad \rho_t < y_t\sqrt{\frac{1-\rho_t^2}{1-y_t^2}}. \tag{3.24}$$

As the function $x \mapsto x/(1-x^2)$ is strictly increasing on $[0,1]$, this is in turn equivalent to $\rho_t < y_t$ as claimed.

Finally, if $\gamma_t < 0$, then $\tilde{\nu}_t^*$ (and in turn ν_t^*) is always positive, because

$$\rho_t + y_t\sqrt{\frac{1-\rho_t^2}{1-y_t^2}} > 0 \quad \Rightarrow \quad -\frac{\gamma_t}{v_t}\left(\rho_t + y_t\sqrt{\frac{1-\rho_t^2}{1-y_t^2}}\right) > 0. \tag{3.25}$$

This completes the proof. □

Proof of Corollary 3.3.7. Fix a time t and state $\omega \in \Omega$.

(i) Fix the spread s_t and note that by definition of y_t, $y_t \in (-1,0)$ is equivalent to $\rho_t \in \left(\frac{s_t}{\sigma_t\sqrt{2\pi}}, 2\frac{s_t}{\sigma_t\sqrt{2\pi}}\right)$. This is clearly satisfied if the parameter constraint from the corollary holds. The upper bound in the latter also ensures that $\rho_t > -y_t$. Together with $\gamma_t < 0$, this implies that $\nu_t^* > 0$ by Lemma 3.3.6(i).

Conversely, suppose that $y_t \in (-1,0)$ and $\nu_t^* > 0$. Then, by Lemma 3.3.6(i), we necessarily have $\gamma_t < 0$ and $\rho_t > -y_t$. The second inequality and $y_t \in (-1,0)$ in turn imply that the parameter constraint from the corollary is satisfied.

(ii) This follows along the same lines as (i).

□

3.8.2 Proofs for Section 3.4

Proof of Lemma 3.4.1. The terminal condition for C, Itô's formula, and the PDE (3.9) yield (supressing the arguments of the functions to ease notation):

$$
\begin{aligned}
&\varphi(P_T)\\
&= C(0,P_0) + \int_0^T \left(\partial_t C + \mu\partial_p C + \frac{\sigma^2}{2}\partial_{pp}C\right)dt + \int_0^T \sigma\partial_p C dW_t\\
&= C(0,P_0) + \int_0^T \partial_p C dP_t + \int_0^T \sup_{\nu\in[0,1]}\left\{\frac{s}{\sqrt{2\pi}}\left[-\sqrt{\sigma^2\partial_{pp}C^2 + 2\rho\nu v\sigma\partial_{pp}C + \nu^2v^2} + \nu v\right] - \rho\sigma\nu v\right\}dt\\
&= C(0,P_0) + \int_0^T \partial_p C dP_t + \int_0^T \left(\frac{s}{\sqrt{2\pi}}\left[-\sqrt{\sigma^2\partial_{pp}C^2 + 2\rho\nu^* v\sigma\partial_{pp}C + (\nu^* v)^2} + \nu^* v\right] - \rho\sigma\nu^* v\right)dt\\
&= C(0,P_0) + \int_0^T \Pi dP_t + \int_0^T \left(\frac{s}{\sqrt{2\pi}}\left[-\sqrt{\gamma^2 + 2\rho\gamma\nu^* v + (\nu^* v)^2} + \nu^* v\right] - \rho\sigma\nu^* v\right)dt.
\end{aligned}
$$

Whence, by Theorem 3.3.1, the optimal hedging portfolio from Lemma 3.3.5 (with initial capital $C(0,P_0)$) converges to the option payoff to be hedged in probability in the continuous-time limit $N\to\infty$. □

Proof of Theorem 3.4.2. We focus on the most important case $y\in(-1,0)$ and $\rho > -y$ here; the other cases can be treated along the same lines.

Step 1: Rewrite the PDE. After the change of variable $\tau := \frac{\sigma^2}{2}(T-t)$ and using Lemma 3.3.5, the PDE (3.9) can be rewritten as follows:

$$
\partial_\tau C = \partial_{pp}C + \kappa G\left(\partial_{pp}C\right), \quad C(0,p) = \varphi(p), \tag{3.26}
$$

where $\kappa := 2s/\sigma^2\sqrt{2\pi} > 0$ and G is a piecewise smooth function (plotted in Figure 3.10):

$$
G(z) := \begin{cases}
\sigma z & \text{if } z \geq 0,\\
\sigma\rho y z - \sigma\sqrt{(1-y^2)(1-\rho^2)}z & \text{if } z\in\left(-\frac{v}{\sigma}\left(\rho + y\sqrt{\frac{1-\rho^2}{1-y^2}}\right)^{-1}, 0\right),\\
-yv + \sqrt{(\sigma z+\rho v)^2 + (1-\rho^2)v^2} & \text{if } z \leq -\frac{v}{\sigma}\left(\rho + y\sqrt{\frac{1-\rho^2}{1-y^2}}\right)^{-1}.
\end{cases}
$$

Step 2: Approximate the PDE with a smooth equation. To smooth out the two kinks of the nonlinearity G in (3.26), we convolute it with the heat kernel:

$$G_\varepsilon(z) := \int_{-\infty}^{\infty} G(z + \varepsilon x)\phi(x)dx,$$

where $\phi(\cdot)$ is the probability density function of the standard normal distribution. For small ε, we then consider the following smooth approximations of our original PDE (3.26), illustrated in Figure 3.10:

$$\partial_\tau C_\varepsilon = \partial_{pp} C_\varepsilon + \kappa G_\varepsilon(\partial_{pp} C_\varepsilon), \quad C_\varepsilon(0, p) = \varphi(p). \tag{3.27}$$

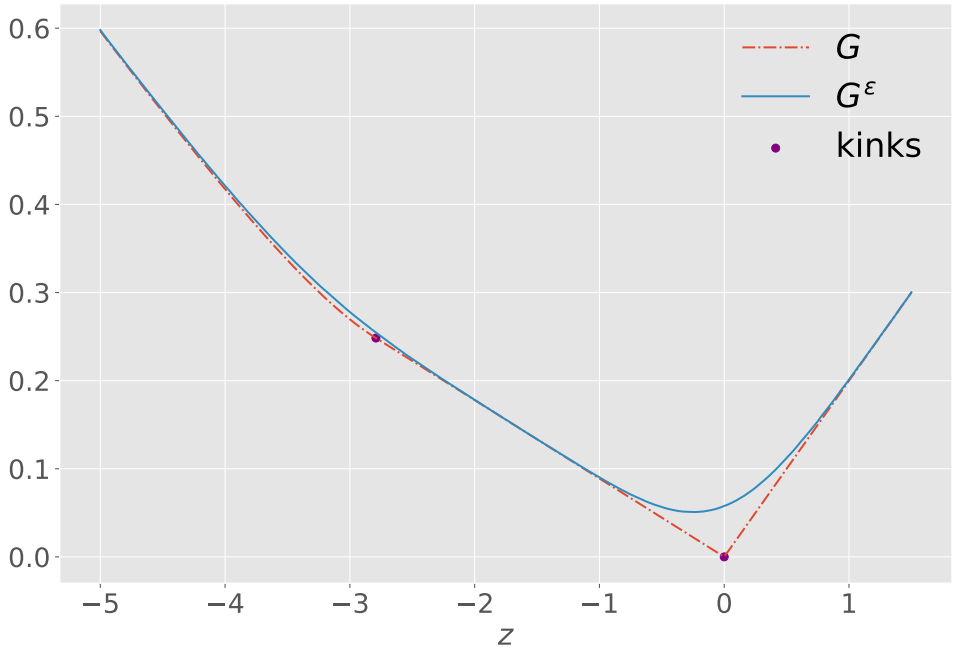

Figure 3.10: The nonlinearity G from (3.26) and its smooth approximation G_ε.

Step 3: Bound the derivative of G_ε. The piecewise function G is differentiable on the interior $R_i \subset \mathbb{R}$, $i = 1, 2, 3$ of each subinterval on which it is defined. In each case, the derivative is bounded from above by $\sigma(1 + \rho)$ and bounded from below by $-\sigma$. Whence, it follows from the dominated convergence theorem that

$$G'_\varepsilon(z) = \sum_{i=1}^{3} \int_{R_i} G'(z + \varepsilon x)\phi(x)dx. \tag{3.28}$$

In particular, the derivative of G_ε is bounded as well: $G'_\varepsilon(z) \in [-\sigma, \sigma(1+\rho)]$.

Step 4: Differentiate the approximating PDE. Formally taking partial derivatives of (3.27) with respect to p, we obtain the following PDEs that should be satisfied by the first, second, and third derivatives of C_ε with respect to this spatial variable:

$$\partial_\tau V_\varepsilon = \partial_{pp} V_\varepsilon + \kappa \partial_{pp} V_\varepsilon G'_\varepsilon(\partial_p V_\varepsilon), \qquad V_\varepsilon(0,p) = \varphi'(p), \tag{3.29}$$

$$\partial_\tau U_\varepsilon = \partial_{pp} U_\varepsilon + \kappa \partial_{pp} U_\varepsilon G'_\varepsilon(U_\varepsilon) + \kappa \left(\partial_p U_\varepsilon\right)^2 G''_\varepsilon(U_\varepsilon), \qquad U_\varepsilon(0,p) = \varphi''(p), \tag{3.30}$$

$$\partial_\tau Q_\varepsilon = \partial_{pp} Q_\varepsilon + \kappa \partial_{pp} Q_\varepsilon G'_\varepsilon(U_\varepsilon) + 3\kappa Q \partial_p Q_\varepsilon G''_\varepsilon(U_\varepsilon) + \kappa \left(Q_\varepsilon\right)^3 G'''_\varepsilon(U_\varepsilon), \quad Q_\varepsilon(0,p) = \varphi'''(p). \tag{3.31}$$

Step 5: Establish uniform parabolicity and in turn existence of the approximating equation. To establish uniform parabolicity in the sense of (Lieberman, 1996, Chapter 14, p. 357) for the approximating PDE (3.27) as well as for the PDEs (3.29) and (3.30) for V_ε and U_ε, we differentiate the right-hand sides of these equations by $\partial_{pp} C_\varepsilon$, $\partial_{pp} V_\varepsilon$ and $\partial_{pp} U_\varepsilon$, respectively. By the lower bound of G'_ε and the parameter constraint $\sigma > s\sqrt{2/\pi}$, these derivatives are all bounded away from zero, so all three equations are uniformly parabolic. As a consequence, all three PDEs have unique bounded classical solutions by (Lieberman, 1996, Theorem 14.7). Next, with U_ε at hand, it follows along the same lines that the PDE (3.31) for Q_ε is also uniformly parabolic and therefore has a unique bounded classical solution as well.

Furthermore, (Lieberman, 1996, Lemma 14.11) applied to the approximating PDE (3.26) shows that C_ε is in fact three-times continuously differentiable with respect to the spatial variable. By parabolic regularity, its time derivative is differentiable in space too, so that (3.29) indeed describes the derivative of the solution of (3.26). The same argument applied to (3.29) and (3.30) in turn yields that (3.30) and (3.31) indeed characterize the second and third spatial derivatives of the solution of (3.26) respectively.

Step 6: Derive uniform bounds for the approximating PDE. We can now apply the maximum principle (Lieberman, 1996, Theorem 14.4) to infer the boundedness of C_ε,

$\partial_p C_\varepsilon = V_\varepsilon$, $\partial_{pp} C_\varepsilon = U_\varepsilon$ and $\partial_{ppp} C_\varepsilon = Q_\varepsilon$, uniformly in the smoothing parameter ε:

$$||C_\varepsilon||_\infty \leq \exp(T)(||\varphi||_\infty + 1),$$
$$||\partial_p C_\varepsilon||_\infty \leq \exp(T)(||\varphi'||_\infty + 1),$$
$$||\partial_{pp} C_\varepsilon||_\infty \leq \exp(T)(||\varphi''||_\infty + 1),$$
$$||\partial_{ppp} C_\varepsilon||_\infty \leq \exp(T)(||\varphi'''||_\infty + 1).$$

Step 7: Apply the Arzela-Ascoli Theorem to construct a solution of the original equation. As $\partial_p C_\varepsilon$, $\partial_{pp} C_\varepsilon$ and $\partial_{ppp} C_\varepsilon$ are all uniformly bounded independent of ε, the families of functions C_ε, $\partial_p C_\varepsilon$ and $\partial_{pp} C_\varepsilon$ are uniformly Lipschitz continuous and hence uniformly equicontinuous. Therefore, we can apply the Arzela-Ascoli Theorem to show that there exists a sequence $\varepsilon_n \to 0$ such that C_{ε_n}, $\partial_p C_{\varepsilon_n}$ and $\partial_{pp} C_{\varepsilon_n}$ converge on a fixed (spatial) interval $[-l, l]$ for every $t \in (0, T)$.

On the larger interval $[-l-1, l+1]$, there is a subsequence ε_k of the sequence ε_n along which C_{ε_k} again converges to a limit. On $[-l, l]$, by uniqueness of the limit, this must be the same limiting function. Continuing like this, we can define a limiting function on the whole real line, for which we have uniform convergence of the functions C_ε, their second spatial derivatives $\partial_{pp} C_\varepsilon$ and, by parabolic regularity, also of their time derivatives $\partial_t C_\varepsilon$ on any interval $[-l, l]$ as the smoothing parameter tends to zero. The limit in turn solves the original PDE (3.9).

Step 8: Uniqueness. By Step 3, the derivative of the nonlinearity in the PDE (3.9) (with respect to $\partial_{pp} C$) is bounded from above by $\sigma(1+\rho)$ and bounded from below by $-\sigma$ on the interior of each of the three subintervals on which it is defined. Together with the parameter constraint $\sigma > s\sqrt{2/\pi}$, this yields the monotonicity of the right-hand side of (3.9) in the second spatial derivative $\partial_{pp} C$. Whence, uniqueness follows from (Lieberman, 1996, Corollary 14.2). □

3.8.3 Details for Empirical Data Analysis

The LOBSTER data we use in the empirical illustration from Section 3.5 is compiled as follows:

1. Collect LOBSTER data and filter out market order execution data (event type 4 or 5);

2. In LOBSTER data, the direction of trades is indicated for limit orders but not hidden orders (which can account for up to 30% in terms of number of trades and also traded volume). The Lee-Ready algorithm is applied to estimate the direction of the hidden order executions: when the execution price is larger than the mid-price, assign the direction as buyer-initiated, and vice versa;

3. Combine executions at the same nanosecond of the same direction;

4. Calculate signed order flow (from the perspective of holders of limit orders) from direction and volume of trades;

5. Calculate price differences and inventory differences;

6. Remove the overnight differences;

7. Compute quadratic covariation between price and inventory, quadratic variation of price, quadratic variation of inventory on the trade clock;

8. Compute estimators for v, σ, s, ρ, y; the spread s_N is estimated by the average of the difference between the best-bid and ask quotes in the market over time.

3.8.4 Proofs for Section 3.6

In this section, we prove Theorem 3.6.2 about the convergence of the partial-hedging portfolio process (3.13) to its continuous-time limit (3.15). We do this under the following weaker but less transparent version of Assumption 3.6.1:

Assumption 3.8.1. *There exists a constant $c > 1$ such that:*

(i) $\kappa_t \geq 1/c > 0$ *for all* $t \in [0,T]$*;*

and, for any process $H \in \{\mu^F, v, \mu, \sigma, s, \rho, \mu^\Pi, \gamma, \nu, \kappa, F, P, \Pi\}$*, we have:*

(ii) $\mathbb{E}[H_t] \leq c$ *for all* $t \in [0,T]$*;*

(iii) $\mathbb{E}[H_{t_1} H_{t_2}] \leq c$ *for all* $t_1, t_2 \in [0,T]$*;*

(iv) $\mathbb{E}\left[H_{t_1} H_{t_2} H_{t_3} H_{t_4}\right] \leq c$ *for all* $t_1, t_2, t_3, t_4 \in [0,T]$*;*

(v) $\mathbb{E}\left[(H_{t_1} - H_t)(H_{t_2} - H_t)\right] \leq c \max(|t_1 - t|, |t_2 - t|)$ *for all* $t_1, t_2, t \in [0,T]$*;*

(vi) $\mathbb{E}\left[(H_{t_1} - H_t)(H_{t_2} - H_t)(H_{t_3} - H_t)(H_{t_4} - H_t)\right] \leq c \max(|t_1 - t|, |t_2 - t|, |t_3 - t|, |t_4 - t|)^2$ *for all* $t_1, t_2, t_3, t_4, t \in [0,T]$.

Using the Cauchy-Schwarz inequality, the triangle inequality and the Itô isometry, it is straightforward to verify that Assumption 3.8.1 is indeed implied by Assumption 3.6.1 in the main body of the chapter.

3.8.4.1 Bounds for the Displacement

Recall that we assume in Assumption 3.8.1(i) that the fractions of the displacement D^N that are rectified using market orders at each time step are uniformly bounded away from zero. This leads to the following bound for the displacement:

Proposition 3.8.2. $(\Delta t^N)^{-1/2+\varepsilon} D^N$ *uniformly converges to zero in* $\mathbb{L}^2$ *for any positive* ε*.*

Proof. By (3.12) and Minkowski's inequality, the second moment of the displacement satisfies the following recursion:

$$\begin{aligned}\mathbb{E}\left[(D_n^N)^2\right] &= \mathbb{E}\left[((1-\kappa_n^N)(D_{n-1}^N + \Delta\Pi_n^N - \nu_{n-1}^N \Delta F_n^N))^2\right] \\ &\leq (1-\tfrac{1}{c})^2 \left[\sqrt{\mathbb{E}\left[(D_{n-1}^N)^2\right]} + \sqrt{4c\Delta t^N}\right]^2.\end{aligned}$$

Here, we have used in the second step that Assumption 3.8.1(ii) implies

$$\mathbb{E}\left[(\Delta\Pi_n^N - \nu_{n-1}^N \Delta F_n^N)^2\right] \leq 2\mathbb{E}\left[(\Delta\Pi_n^N)^2\right] + 2\mathbb{E}\left[(\Delta F_n^N)^2\right] \leq 4c\Delta t^N.$$

As $D_0^N = 0$, it follows that $\mathbb{E}\left[(D_n^N)^2\right]^{1/2} \leq (c-1)\sqrt{4c\Delta t^N}$. Whence, as N tends to infinity (and $\Delta t^N = T/N$ tends to zero), $(\Delta t^N)^{-1/2+\varepsilon} D^N$ converges uniformly to 0 in $\mathbb{L}^2$ for any $\varepsilon > 0$. □

3.8.4.2 Reduction to Piecewise Constant Controls

To ease notation, we henceforth focus on values of N for which $\sqrt{N}$ is an integer. (The intermediate values can be dealt with along the same lines using rounding, but we do not spell this out to ease notation.) Then, we can group the N trading rounds in the discrete model into $\sqrt{N}$ groups of $\sqrt{N}$ trading rounds each. We now argue that the continuous-time limit of the execution costs and displacements does *not* change when the controls ν^N and κ^N as well as the drift and diffusion coefficients of prices and order flow are assumed to be piecewise constant on each subgroup. This approximation in turn allows us to compute the continuous-time limit of the goal functional for each subgroup using conditional normal distributions in Appendix 3.8.4.3 below.

Recall that – in the original model – the execution cost can be decomposed into the spread cost of market order trades, earnings from limit order trades and the impact of adverse selection:

$$C^N := C_M^N - C_L^N - C_A^N, \tag{3.32}$$

where

$$C_M^N := \sum_{j=1}^{N} \frac{s_j^N \kappa_j^N}{2} |D_{j-1}^N + \Delta\Pi_j^N - \nu_{j-1}^N \Delta F_j^N|, \qquad C_L^N := \sum_{j=1}^{N} \frac{s_j^N \nu_j^N}{2} |\Delta F_j^N|,$$

$$C_A^N := \sum_{j=1}^{N} \nu_{j-1}^N \Delta F_j^N \Delta P_j^N.$$

When controls and model coefficients are constant on each subgroup of trading rounds, the total execution cost can be decomposed analogously:

$$\tilde{C}^N := \tilde{C}_M^N - \tilde{C}_L^N - \tilde{C}_A^N. \tag{3.33}$$

Here, like for the original controls,

$$\tilde{C}_M^N := \sum_{j=1}^{N} \frac{\tilde{s}_j^N \tilde{\kappa}_j^N}{2} |\tilde{D}_{j-1}^N + \Delta\tilde{\Pi}_j^N - \tilde{\nu}_{j-1}^N \Delta\tilde{F}_j^N|, \qquad \tilde{C}_L^N := \sum_{j=1}^{N} \frac{\tilde{s}_j^N \tilde{\nu}_j^N}{2} |\Delta\tilde{F}_j^N|,$$

$$\tilde{C}_A^N := \sum_{j=1}^{N} \tilde{\nu}_{j-1}^N \Delta\tilde{F}_j^N \Delta\tilde{P}_j^N,$$

but the drift and diffusion processes and controls are now replaced by their step-function approximations $\tilde{\mu}^F, \tilde{v}, \tilde{\rho}, \tilde{\mu}, \tilde{\sigma}, \tilde{s}, \tilde{\kappa}$ and $\tilde{\nu}$, which keep the value of the original process at time $t_{i\sqrt{N}}^N$ during the whole time period $[t_{i\sqrt{N}}^N, t_{(i+1)\sqrt{N}}^N]$, $i = 0, \ldots, \sqrt{N} - 1$:

$$d\tilde{F}_t = \tilde{\mu}_t^F dt + \tilde{v}_t \sqrt{1 - \tilde{\rho}_t} dW_t^{\perp} - \tilde{v}_t \tilde{\rho}_t dW_t, \quad d\tilde{P}_t = \tilde{\mu}_t dt + \tilde{\sigma}_t dW_t, \quad d\tilde{\Pi}_t = \tilde{\mu}_t^{\Pi} dt + \tilde{\gamma}_t dW_t,$$

$$\tilde{D}_j^N = (1 - \tilde{\kappa}_j^N) \left(\tilde{D}_{j-1}^N + \Delta\tilde{\Pi}_j^N - \tilde{\nu}_{j-1}^N \Delta\tilde{F}_j^N \right). \tag{3.34}$$

As the drift and diffusion coefficients of $\tilde{P}, \tilde{F}$ and $\tilde{\Pi}$ are the step-function approximations of the original coefficients, they also satisfy Assumption 3.8.1(ii-vi).

We now show that as the number of time steps N tends to infinity, the difference between the goal functional for the original model and its piecewise constant approximation vanishes. To establish this, we prove that the squared differences between each of the respective components tend to zero at appropriate rates:

Theorem 3.8.3. *In the continuous-time limit $N \to \infty$, $N^{1/2-\varepsilon}(\tilde{C}_M^N - C_M^N)$, $N^{1/2-\varepsilon}(\tilde{C}_L^N - C_L^N)$, $N^{1/2-\varepsilon}(\tilde{C}_A^N - C_A^N)$ and $N^{3/4-\varepsilon}(\tilde{D}^N - D^N)$ (uniformly) converge to zero in $\mathbb{L}^2$ for any positive ε.*

Proof. The increments of the original order flow process and its approximation satisfy

$$\Delta\tilde{F}_j^N - \Delta F_j^N = \int_{j\Delta t}^{(j+1)\Delta t} (\mu_t^F - \tilde{\mu}_t^F) dt + \int_{j\Delta t}^{(j+1)\Delta t} (v_t - \tilde{v}_t) dW_t.$$

Assumption 3.8.1(iv), the Itô isometry and Fubini's theorem in turn yield the following

upper bound:

$$\begin{aligned}
&\mathbb{E}\left[\left(\Delta\tilde{F}_j^N - \Delta F_j^N\right)^2\right] \\
&\leq 2\mathbb{E}\left[\left(\int_{j\Delta t}^{(j+1)\Delta t}(\mu_t^F - \tilde{\mu}_t^F)dt\right)^2\right] + 2\mathbb{E}\left[\left(\int_{j\Delta t}^{(j+1)\Delta t}(v_t - \tilde{v}_t)dW_t\right)^2\right] \\
&= \mathbb{E}\left[\int_{j\Delta t}^{(j+1)\Delta t}\int_{j\Delta t}^{(j+1)\Delta t}(\mu_t^F - \tilde{\mu}_t^F)(\mu_s - \tilde{\mu}_s^F)dtds\right] + \mathbb{E}\left[\int_{j\Delta t}^{(j+1)\Delta t}(v_t - \tilde{v}_t)^2dt\right] \\
&= \int_{j\Delta t}^{(j+1)\Delta t}\int_{j\Delta t}^{(j+1)\Delta t}\mathbb{E}\left[(\mu_t^F - \tilde{\mu}_t^F)(\mu_s^F - \tilde{\mu}_s^F)\right]dtds + \int_{j\Delta t}^{(j+1)\Delta t}\mathbb{E}\left[(v_t - \tilde{v}_t)^2\right]dt \\
&\leq c(\Delta t)^3\sqrt{N} + c(\Delta t)^2\sqrt{N} = O(N^{-3/2}).
\end{aligned} \tag{3.35}$$

Here, we have used that on any subgroup of $\sqrt{N}$ trading rounds (with total length $\Delta t\sqrt{N}$), the expected squared difference between the process v_t and its piecewise constant approximation $\tilde{v}_t$ is bounded from above by $c\Delta t^N\sqrt{N}$ by Assumption 3.8.1(v). We have also used an analogous estimate for the process b and its approximation $\tilde{b}$. The same arguments also show

$$\mathbb{E}\left[\left(\Delta\tilde{P}_j^N - \Delta P_j^N\right)^2\right] = O(N^{-3/2}), \quad \mathbb{E}\left[\left(\Delta\tilde{\Pi}_j^N - \Delta\Pi_j^N\right)^2\right] = O(N^{-3/2}). \tag{3.36}$$

Impact of adverse selection We first bound the difference between the impacts of adverse selection $\tilde{C}_A^N$ and C_A^N, which can be decomposed as follows:

$$\tilde{C}_A^N - C_A^N = \Delta C_{A,\nu}^N + \Delta C_{A,F}^N + \Delta C_{A,P}^N,$$

where

$$\Delta C_{A,\nu}^N = \sum_{j=1}^N\left(\tilde{\nu}_j^N - \nu_j^N\right)\Delta\tilde{F}_j^N\Delta\tilde{P}_j^N, \qquad \Delta C_{A,F}^N = \sum_{j=1}^N\nu_j^N\left(\Delta\tilde{F}_j^N - \Delta F_j^N\right)\Delta\tilde{P}_j^N,$$

$$\Delta C_{A,P}^N = \sum_{j=1}^N\nu_j^N\Delta F_j^N\left(\Delta\tilde{P}_j^N - \Delta P_j^N\right).$$

The expected squared difference between the adverse selection costs then can be

bounded by

$$\mathbb{E}\left[\left(\tilde{C}_A^N - C_A^N\right)^2\right] \leq 4\mathbb{E}\left[(\Delta C_{A,\nu}^N)^2\right] + 4\mathbb{E}\left[(\Delta C_{A,P}^N)^2\right] + 4\mathbb{E}\left[(\Delta C_{A,F}^N)^2\right].$$

Two applications of the Cauchy-Schwarz inequality yield

$$\begin{aligned}\mathbb{E}\left[(\Delta C_{A,\nu}^N)^2\right] &\leq \sqrt{\mathbb{E}\left[\sum_{j=1}^N \left(\tilde{\nu}_j^N - \nu_j^N\right)^2\right] \mathbb{E}\left[\sum_{j=1}^N (\Delta \tilde{F}_j^N \Delta \tilde{P}_j^N)^2\right]} \\ &\leq \sqrt{cT\sqrt{N}\sum_{j=1}^N \sqrt{\mathbb{E}\left[(\Delta \tilde{F}_j^N)^4\right] \mathbb{E}\left[(\Delta \tilde{P}_j^N)^4\right]}} \\ &\leq c\sqrt{T\sqrt{N} N \sqrt{(\Delta t)^4}} = O(N^{-1/4}).\end{aligned}$$

Here, we have used Assumption 3.8.1(v) with time difference $\Delta t\sqrt{N} = T/\sqrt{N}$ in the second estimate, and Assumption 3.8.1(vi) with time difference Δt in the third step. Arguing along the same lines, with (3.35) and (3.36), we also obtain

$$\mathbb{E}\left[(\Delta C_{A,F}^N)^2\right] = O(N^{-1/4}), \quad \mathbb{E}\left[(\Delta C_{A,P}^N)^2\right] = O(N^{-1/4}).$$

In summary, we therefore have $\mathbb{E}\left[\left(\tilde{C}_A^N - C_A^N\right)^2\right] = O(N^{-1/4})$, so that $N^{1/2-\varepsilon}(\tilde{C}_A^N - C_A^N)$ indeed converges to zero in $\mathbb{L}^2$ for any positive ε.

Displacements Recall that in Proposition 3.8.2 we have already established that each of the displacements and in turn also their difference $D_j^N - \tilde{D}_j^N$ converges to zero in $\mathbb{L}^2$. We now establish a faster rate of convergence, uniformly for all timesteps $1 \leq j \leq N$.

Taking into account the dynamics of the displacements D and $\tilde{D}$ from (3.12) and (3.34), respectively, their difference can be written as

$$\begin{aligned}D_j^N - \tilde{D}_j^N = (\tilde{\kappa}_j^N - \kappa_j^N)\left(\tilde{D}_{j-1}^N + \Delta\tilde{\Pi}_j^N - \tilde{\nu}_{j-1}^N \Delta \tilde{F}_j^N\right) \\ + (1-\kappa_j^N)(D_{j-1}^N - \tilde{D}_{j-1}^N + \Delta\Pi_j^N - \Delta\tilde{\Pi}_j^N + \tilde{\nu}_{j-1}^N \Delta\tilde{F}_j^N - \nu_{j-1}^N \Delta F_j^N).\end{aligned}$$

By the inequalities of Cauchy-Schwarz and Minkowski,

$$\begin{aligned}
&\mathbb{E}\left[(\tilde{\kappa}_j^N-\kappa_j^N)^2\left(\tilde{D}_{j-1}^N+\Delta\tilde{\Pi}_j^N-\tilde{\nu}_{j-1}^N\Delta\tilde{F}_j^N\right)^2\right]\\
&\leq\sqrt{\mathbb{E}\left[(\tilde{\kappa}_j^N-\kappa_j^N)^4\right]\mathbb{E}\left[\left(\tilde{D}_{j-1}^N+\Delta\tilde{\Pi}_j^N-\tilde{\nu}_{j-1}^N\Delta\tilde{F}_j^N\right)^4\right]}\\
&\leq\sqrt{c(\Delta t\sqrt{N})^2\left(\sqrt[4]{\mathbb{E}\left[\left(\tilde{D}_{j-1}^N\right)^4\right]}+\sqrt[4]{\mathbb{E}\left[\left(\Delta\tilde{\Pi}_j^N\right)^4\right]}+\sqrt[4]{\mathbb{E}\left[\left(\tilde{\nu}_{j-1}^N\Delta\tilde{F}_j^N\right)^4\right]}\right)^4} \qquad (3.37)\\
&\leq\sqrt{c(\Delta t\sqrt{N})^2\left(\sqrt[4]{c(\Delta t)^2}+\sqrt[4]{c(\Delta t)^2}+\sqrt[4]{c(\Delta t)^2}\right)^4} \qquad (3.38)\\
&\leq\sqrt{c(\Delta t\sqrt{N})^2\cdot 81c(\Delta t)^2}=O(N^{-3/2}).
\end{aligned}$$

Here, (3.37) follows from Assumption 3.8.1(vi) with time step of at most $\Delta t\sqrt{N}=T/\sqrt{N}$ and the Minkowski inequality. (3.38) follows from an argument similar to Proposition 3.8.2 to bound the fourth moment of $\tilde{D}^N$ and Assumption 3.8.1(vi) with time step Δt.

From a similar argument for 3.36, we can show that

$$\begin{aligned}
\mathbb{E}\left[(1-\kappa_j^N)^2(\Delta\Pi_j^N-\Delta\tilde{\Pi}_j^N)^2\right]&=O(N^{-3/2}),\\
\mathbb{E}\left[(1-\kappa_j^N)^2(\tilde{\nu}_{j-1}^N\Delta\tilde{F}_j^N-\nu_{j-1}^N\Delta F_j^N)^2\right]&=O(N^{-3/2}).
\end{aligned}$$

Another application of Minkowski's inequality in turn yields the following recursion for the second moment of the difference between the displacements:

$$\begin{aligned}
&\sqrt{\mathbb{E}\left[(D_j^N-\tilde{D}_j^N)^2\right]}\\
&\leq\sqrt{\mathbb{E}\left[(\tilde{\kappa}_j^N-\kappa_j^N)^2\left(\tilde{D}_{j-1}^N+\Delta\tilde{\Pi}_j^N-\tilde{\nu}_j^N\Delta\tilde{F}_j^N\right)^2\right]}+\sqrt{\mathbb{E}\left[(1-\kappa_j^N)^2(D_{j-1}^N-\tilde{D}_{j-1}^N)^2\right]}\\
&\quad+\sqrt{\mathbb{E}\left[(1-\kappa_j^N)^2(\Delta\Pi_j^N-\Delta\tilde{\Pi}_j^N)^2\right]}+\sqrt{\mathbb{E}\left[(1-\kappa_j^N)^2(\tilde{\nu}_{j-1}^N\Delta\tilde{F}_j^N-\nu_{j-1}^N\Delta F_j^N)^2\right]}\\
&\leq O(N^{-3/4})+\left(1-\frac{1}{c}\right)\sqrt{\mathbb{E}\left[(D_{j-1}^N-\tilde{D}_{j-1}^N)^2\right]}.
\end{aligned}$$

As $D_0^N=\tilde{D}_0^N=0$, it follows that $\mathbb{E}[(D_n^N-\tilde{D}_n^N)^2]=O(N^{-3/2})$. Whence, as N tends

to infinity, $N^{3/4-\varepsilon}(D^N - \tilde{D}^N)$ indeed converges uniformly to 0 in $\mathbb{L}^2$ for any $\varepsilon > 0$.

Earnings from limit orders Next, the difference between $\tilde{C}_L^N$ and C_L^N can again be decomposed into three parts:

$$\tilde{C}_L^N - C_L^N = \frac{1}{2}\left(\Delta C_{L,s}^N + \Delta C_{L,\nu}^N + \Delta C_{L,F}^N\right),$$

where

$$\Delta C_{L,s}^N = \sum_{j=1}^N (\tilde{s}_j^N - s_j^N)\tilde{\nu}_j^N |\Delta \tilde{F}_j^N|, \qquad \Delta C_{L,\nu}^N = \sum_{j=1}^N s_j^N \left(\tilde{\nu}_j^N - \nu_j^N\right) |\Delta \tilde{F}_j^N|$$

$$\Delta C_{L,F}^N = \sum_{j=1}^N \nu_j^N s_j^N \left(|\Delta \tilde{F}_j^N| - |\Delta F_j^N|\right).$$

The second moments of all three components can be bounded similarly as the components of the adverse selection costs, by recalling that the spread s scales with $\sqrt{\Delta t^N}$:

$$\mathbb{E}\left[(\Delta C_{L,s}^N)^2\right] = O(N^{-1/4}), \quad \mathbb{E}\left[(\Delta C_{L,\nu}^N)^2\right] = O(N^{-1/4}), \quad \mathbb{E}\left[(\Delta C_{L,F}^N)^2\right] = O(N^{-1/4}).$$

It follows that $N^{1/2-\varepsilon}(\tilde{C}_L^N - C_L^N)$ also converges to zero in $\mathbb{L}^2$ for any positive ε.

Transaction costs due to market orders Finally, the difference between $\tilde{C}_M^N$ and C_M^N can once again be decomposed into three parts:

$$\tilde{C}_M^N - C_M^N = \frac{1}{2}\left(\Delta C_{M,s}^N + \Delta C_{M,\kappa}^N + \Delta C_{M,D}^N\right),$$

where

$$\Delta C_{M,s}^N = \sum_{j=1}^N \left(\tilde{s}_j^N - s_j^N\right) \tilde{\kappa}_j^N |\tilde{D}_{j-1}^N + \Delta\tilde{\Pi}_j^N - \tilde{\nu}_j^N \Delta \tilde{F}_j^N|,$$

$$\Delta C_{M,\kappa}^N = \sum_{j=1}^N s_j^N \left(\tilde{\kappa}_j^N - \kappa_j^N\right) |\tilde{D}_{j-1}^N + \Delta\tilde{\Pi}_j^N - \tilde{\nu}_j^N \Delta \tilde{F}_j^N|,$$

$$\Delta C_{M,D}^N = \sum_{j=1}^N s_j^N \kappa_j^N \left(|\tilde{D}_{j-1}^N + \Delta\tilde{\Pi}_j^N - \tilde{\nu}_j^N \Delta \tilde{F}_j^N| - |D_{j-1}^N + \Delta\Pi_j^N - \nu_j^N \Delta F_j^N|\right).$$

The second moments of the first two components can be bounded similarly as in the previous steps:

$$\mathbb{E}\left[(\Delta C^N_{M,s})^2\right] = O(N^{-1/4}), \quad \mathbb{E}\left[(\Delta C^N_{M,\kappa})^2\right] = O(N^{-1/4}).$$

We now bound the third component. By Minkowski's inequality,

$$\begin{aligned}
&\sqrt{\mathbb{E}\left[(\tilde{D}^N_{j-1} - D^N_{j-1} + \Delta\tilde{\Pi}^N_j - \Delta\Pi^N_j - \nu^N_j \Delta F^N_j + \tilde{\nu}^N_j \Delta\tilde{F}^N_j)^2\right]} \\
&\leq \sqrt{\mathbb{E}\left[(\tilde{D}^N_{j-1} - D^N_{j-1})^2\right]} + \sqrt{\mathbb{E}\left[(\Delta\tilde{\Pi}^N_j - \Delta\Pi^N_j)^2\right]} + \sqrt{\mathbb{E}\left[(\nu^N_j \Delta F^N_j - \tilde{\nu}^N_j \Delta\tilde{F}^N_j)^2\right]} = O(N^{-3/4}).
\end{aligned}$$

Here, the second inequality follows from the estimate for $\tilde{D}^N_{j-1} - D^N_{j-1}$ established in the previous step and (3.36). Together with the Cauchy-Schwarz inequality, it follows that the third term in our decomposition above satisfies

$$\begin{aligned}
&\mathbb{E}\left[(\Delta C^N_{M,D})^2\right] \\
&\leq \sqrt{\sum_{j=1}^N \mathbb{E}\left[(\kappa^N_j)^2 (s^N_j)^2\right] \sum_{j=1}^N \mathbb{E}\left[\left(\left|\tilde{D}^N_{j-1} + \Delta\tilde{\Pi}^N_j - \tilde{\nu}^N_j \Delta\tilde{F}^N_j\right| - \left|D^N_{j-1} + \Delta\Pi^N_j - \nu^N_j \Delta F^N_j\right|\right)^2\right]} \\
&\leq \sqrt{cT \cdot \sum_{j=1}^N \mathbb{E}\left[(\tilde{D}^N_{j-1} - D^N_{j-1} + \Delta\tilde{\Pi}^N_j - \Delta\Pi^N_j - \nu^N_j \Delta F^N_j + \tilde{\nu}^N_j \Delta\tilde{F}^N_j)^2\right]} \\
&\leq \sqrt{cT \cdot N(\Delta t^N)^2 \sqrt{N} \cdot 16c} = O(N^{-1/4}).
\end{aligned}$$

Here, the second inequality follows from the triangle inequality and the elementary estimate $-|a-b| \leq |a| - |b| \leq |a-b|$, for $a, b \in \mathbb{R}$. In summary, the difference $N^{1/2-\varepsilon}(\tilde{C}^N_M - C^N_M)$ between transaction costs due to market orders also converges to zero in $\mathbb{L}^2$ for any positive ε. □

3.8.4.3 Goal Functional for Piecewise Constant Coefficients

We now prove that the expected execution costs and the sum of expected squared displacements in the model with piecewise constant controls and model coefficients converges to the integrals from Theorem 3.6.2. Together with the approximation result from the

previous section, this will be used in Appendix 3.8.4.4 below to also establish the limit from Theorem 3.6.2 for the original model with arbitrary controls and model coefficients.

Proposition 3.8.4. *For fixed control processes in the piecewise constant setting, the expected execution cost* $\mathbb{E}[\tilde{C}^N]$ *converges to*

$$\mathbb{E}\left[\int_0^T \left(\frac{s_t}{2}\sqrt{\gamma_t^2 + 2\rho_t\gamma_t\nu_t v_t + \nu_t^2 v_t^2}\sqrt{\frac{2}{\pi}}\sqrt{\frac{\kappa_t}{2-\kappa_t}} - \frac{s_t\nu_t}{2}\sqrt{\frac{2}{\pi}}v_t + \nu_t\rho_t\sigma_t v_t\right) dt\right]. \quad (3.39)$$

The sum of expected squared displacements $\sum_{n=1}^N \mathbb{E}[(\tilde{D}_n^N)^2]$ *has the continuous-time limit* (3.16).

Proof. We split the overall execution cost $\tilde{C}^N$ and the sum of expected squared displacements into $\sqrt{N}$ sub-parts, corresponding to the time periods $[t^N_{i\sqrt{N}}, t_{(i+1)\sqrt{N}}]$, and normalize each of these terms with the duration $T/\sqrt{N}$ of these subintervals[22]:

$$\frac{1}{T/\sqrt{N}}\mathbb{E}\left[\tilde{C}_i^N\right], \qquad \frac{1}{T/\sqrt{N}}\sum_{j=i\sqrt{N}+1}^{(i+1)\sqrt{N}}\mathbb{E}\left[\left(\tilde{D}_j^N\right)^2\right]. \quad (3.40)$$

Here, $\tilde{C}_i^N$ is the execution cost within this time period,

$$\tilde{C}_i^N = \sum_{j=i\sqrt{N}+1}^{(i+1)\sqrt{N}} \frac{s^N\kappa}{2}|\tilde{D}_{j-1}^N + \Delta\Pi_j^N - \nu\Delta F_j^N| - \sum_{j=i\sqrt{N}+1}^{(i+1)\sqrt{N}} \frac{s^N\nu}{2}|\Delta F_j^N| - \sum_{j=i\sqrt{N}+1}^{(i+1)\sqrt{N}} \nu\Delta F_j^N\Delta P_j^N,$$

and to ease notation, we henceforth omit the time index of the controls κ, ν and the parameters $\mu, \sigma, b, v, \rho, \mu^\Pi, \gamma$ that are constant within the i-th time period. We now show that, for each i, the first term of (3.40) converges to the integrand of (3.39) evaluated at $t = t^N_{i\sqrt{N}}$, so that the sum of the sub-execution costs indeed approaches the Riemann sum that converges to the integral (3.39). Similarly, the second term of (3.40), the i-th subsum of expected squared displacements converges to the integrand of (3.16) evaluated at $t = t^N_{i\sqrt{N}}$.

By iterating (3.34), the displacement at the n-th time step of this time period can be

[22]Here the normalization factor of $T/\sqrt{N}$ for each of the terms is inherited from the Riemann sum approximation of the integrals in (3.13) and (3.14).

expressed as

$$\tilde{D}^N_{i\sqrt{N}+n} = \sum_{j=1}^{n}(1-\kappa)^{n-j+1}(\Delta\Pi^N_{i\sqrt{N}+j} - \nu\Delta F^N_{i\sqrt{N}+j}) + (1-\kappa)^n \tilde{D}^N_{i\sqrt{N}}.$$

The distribution of $\tilde{D}^N_{i\sqrt{N}+n}$ conditional on the information available at time $t_{i\sqrt{N}}$ is

$$\tilde{D}^N_{i\sqrt{N}+n}|\mathcal{F}_{t_{i\sqrt{N}}} \sim \mathcal{N}\left(\xi_{n,\kappa}(\mu^\Pi - \nu b)\Delta t^N + (1-\kappa)^n \tilde{D}^N_{i\sqrt{N}}, \delta_{n,\kappa}\hat{\sigma}^2\Delta t^N\right),$$

where

$$\xi_{n,\kappa} = \frac{(1-\kappa)(1-(1-\kappa)^n)}{\kappa}, \quad \delta_{n,\kappa} = \frac{(1-\kappa)^2(1-(1-\kappa)^{2n})}{1-(1-\kappa)^2}, \quad \hat{\sigma}^2 = \gamma^2 + 2\rho\gamma\nu v + \nu^2 v^2.$$

As a consequence, the sum of the conditional expected squared displacements can be expressed as a function of $\tilde{D}^N_{i\sqrt{N}}$,

$$\begin{aligned}
&\frac{1}{T/\sqrt{N}} \sum_{j=i\sqrt{N}+1}^{(i+1)\sqrt{N}} \mathbb{E}[(\tilde{D}^N_j)^2|\mathcal{F}_{t_{i\sqrt{N}}}] \\
&= \frac{1}{T/\sqrt{N}} \sum_{n=1}^{\sqrt{N}} \left(\xi_{n,\kappa}(\mu^\Pi - \nu b)\Delta t^N + (1-\kappa)^n \tilde{D}^N_{i\sqrt{N}}\right)^2 + \delta_{n,\kappa}\hat{\sigma}^2\Delta t^N \\
&= \frac{1}{T/\sqrt{N}} \sum_{n=1}^{\sqrt{N}} \xi^2_{n,\kappa}(\mu^\Pi - \nu b)^2(\Delta t^N)^2 + 2\xi_{n,\kappa}\Delta t^N(\mu^\Pi - \nu b)(1-\kappa)^n \tilde{D}^N_{i\sqrt{N}} \\
&\quad + \frac{1}{T/\sqrt{N}} \sum_{n=1}^{\sqrt{N}} (1-\kappa)^{2n}(D^N_{i\sqrt{N}})^2 + \delta_{n,\kappa}\Delta t^N\hat{\sigma}^2 \\
&= (\mu^\Pi - \nu b)^2\Delta t^N \frac{1}{\sqrt{N}} \sum_{n=1}^{\sqrt{N}} \xi^2_{n,\kappa} + 2(\mu^\Pi - \nu b)\tilde{D}^N_{i\sqrt{N}} \frac{1}{\sqrt{N}} \sum_{n=1}^{\sqrt{N}} \xi_{n,\kappa}(1-\kappa)^n \\
&\quad + (\Delta t^N)^{-1/2}(\tilde{D}^N_{i\sqrt{N}})^2 \frac{1}{\sqrt{T}} \sum_{n=1}^{\sqrt{N}} (1-\kappa)^{2n} + \hat{\sigma}^2 \frac{1}{\sqrt{N}} \sum_{n=1}^{\sqrt{N}} \delta_{n,\kappa}.
\end{aligned}$$

Observe that, as N tends to infinity, we have the following convergences:

$$\frac{1}{\sqrt{N}} \sum_{n=1}^{\sqrt{N}} \xi_{n,\kappa}(1-\kappa)^n = \frac{1}{\sqrt{N}} \frac{1-\kappa}{\kappa} \sum_{n=1}^{\sqrt{N}} \left[(1-\kappa)^n - (1-\kappa)^{2n}\right] \to 0, \tag{3.41}$$

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{\sqrt{N}}\xi_{n,\kappa}^2 = \frac{1}{\sqrt{N}}\frac{(1-\kappa)^2}{\kappa^2}\sum_{n=1}^{\sqrt{N}}\left[1-(1-\kappa)^n\right]^2 \to \frac{(1-\kappa)^2}{\kappa^2}, \tag{3.42}$$

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{\sqrt{N}}\delta_{n,\kappa} = \frac{1}{\sqrt{N}}\frac{(1-\kappa)^2}{1-(1-\kappa)^2}\sum_{n=1}^{\sqrt{N}}\left[1-(1-\kappa)^{2n}\right] \to \frac{(1-\kappa)^2}{1-(1-\kappa)^2}. \tag{3.43}$$

Together with the tower property of the conditional expectation, it follows that the sum of the expected squared displacements has the limit

$$\begin{aligned}
&\frac{1}{T/\sqrt{N}}\sum_{j=i\sqrt{N}+1}^{(i+1)\sqrt{N}}\mathbb{E}[(\tilde{D}_j^N)^2] = \frac{1}{T/\sqrt{N}}\sum_{j=i\sqrt{N}+1}^{(i+1)\sqrt{N}}\mathbb{E}\left[\mathbb{E}[(\tilde{D}_j^N)^2|\mathcal{F}_{t_{i\sqrt{N}}}]\right] \\
&= \Delta t^N\mathbb{E}\left[(\mu^\Pi-\nu b)^2\frac{1}{\sqrt{N}}\sum_{n=1}^{\sqrt{N}}\xi_{n,\kappa}^2\right] + \mathbb{E}\left[2(\mu^\Pi-\nu b)\tilde{D}_{i\sqrt{N}}^N\frac{1}{\sqrt{N}}\sum_{n=1}^{\sqrt{N}}\xi_{n,\kappa}(1-\kappa)^n\right] \\
&\quad + (\Delta t^N)^{-1/2}\mathbb{E}\left[(\tilde{D}_{i\sqrt{N}}^N)^2\frac{1}{\sqrt{T}}\sum_{n=1}^{\sqrt{N}}(1-\kappa)^{2n}\right] + \mathbb{E}\left[\hat{\sigma}^2\frac{1}{\sqrt{N}}\sum_{n=1}^{\sqrt{N}}\delta_{n,\kappa}\right] \\
&= \mathbb{E}\left[\hat{\sigma}^2\frac{(1-\kappa)^2}{2\kappa-\kappa^2}\right] + o(1), \quad \text{as } N\to\infty.
\end{aligned}$$

Here, we have used Proposition 3.8.2, $(\Delta t)^{-1/4}\tilde{D}_{i\sqrt{N}}^N \to 0$ in $\mathbb{L}^2$, (3.41), (3.42) and (3.43), together with Cauchy-Schwarz inequality and Assumption 3.8.1(iii) to show that the first three terms in the second line are of order $o(1)$.

As $N\to\infty$, the difference between the sum of second moments of displacements and the Riemann sum approximating the integral (3.16) therefore vanishes. Whence, the discrete sum of expected squared displacements for the model with piecewise constant controls and coefficients indeed converges to the integral (3.16).

We now discuss the different components of the expected execution costs:

1. Conditional on $\mathcal{F}_{t_{i\sqrt{N}}}$, the market order trade $\tilde{D}_{i\sqrt{N}+n-1}^N + \Delta\Pi_{i\sqrt{N}+n}^N - \nu\Delta F_{i\sqrt{N}+n}^N$ at the $i\sqrt{N}+n$-th time step has the following normal distribution:

$$\mathcal{N}\left((\xi_{n,\kappa}+1)(\mu^\Pi-\nu b)\Delta t^N + (1-\kappa)^n\tilde{D}_{i\sqrt{N}}^N, (\delta_{n,\kappa}+1)\hat{\sigma}^2\Delta t^N\right).$$

By the closed-form expression for the expectation of the folded normal distribution,

the (conditional) *expected* market order at the $i\sqrt{N}+n$-th time step is

$$\hat{\sigma}\sqrt{\delta_{n,\kappa}+1}\sqrt{\frac{2\Delta t^N}{\pi}}\exp\left(-\frac{1}{2}\frac{((1-\kappa)^n\tilde{D}^N_{i\sqrt{N}}+O(\Delta t^N))^2}{(\delta_{n,\kappa}+1)\hat{\sigma}^2\Delta t^N}\right)$$
$$+((1-\kappa)^n D^N_{i\sqrt{N}}+O(\Delta t^N))\mathrm{erf}\left(\frac{(1-\kappa)^n\tilde{D}^N_{i\sqrt{N}}+O(\Delta t^N)}{\sqrt{\delta_{n,\kappa}+1}\hat{\sigma}\sqrt{\Delta t^N}}\right).$$

The inequality $\exp(-x)\geq 1-\sqrt{2x}$, $x\geq 0$ leads to the lower bound

$$\hat{\sigma}\sqrt{\delta_{n,\kappa}+1}\sqrt{\frac{2\Delta t^N}{\pi}}\left(1-\frac{(1-\kappa)^n|\tilde{D}^N_{i\sqrt{N}}|+O(\Delta t^N)}{\sqrt{\delta_{n,\kappa}+1}\hat{\sigma}\sqrt{\Delta t^N}}\right).$$

Conversely, $\exp(-x)\leq 1$ and $\mathrm{erf}(x)\leq 1$, $x\geq 0$ give the upper bound

$$\hat{\sigma}\sqrt{\delta_{n,\kappa}+1}\sqrt{\frac{2\Delta t^N}{\pi}}+(1-\kappa)^n|\tilde{D}^N_{i\sqrt{N}}|+O(\Delta t^N).$$

By the tower property and monotonicity of the conditional expectation, the execution costs due to market-order transactions

$$\frac{1}{T/\sqrt{N}}\sum_{j=i\sqrt{N}+1}^{(i+1)\sqrt{N}}\mathbb{E}\left[\frac{s^N\kappa}{2}|\tilde{D}^N_{j-1}+\Delta\Pi^N_j-\nu\Delta F^N_j|\right],$$

therefore have the lower bound

$$\mathbb{E}\left[\frac{s\kappa\hat{\sigma}}{\sqrt{2\pi N}}\sum_{n=1}^{\sqrt{N}}\sqrt{\delta_{n,\kappa}+1}\right]-\frac{1}{\sqrt{T}}\mathbb{E}\left[\frac{s\kappa}{\sqrt{2\pi}}\xi_{\sqrt{N},\kappa}|\tilde{D}^N_{i\sqrt{N}}|\right]-O(\sqrt{\Delta t^N})\mathbb{E}\left[\frac{s\kappa}{\sqrt{2\pi}}\right] \tag{3.44}$$

and the upper bound

$$\mathbb{E}\left[\frac{s\kappa\hat{\sigma}}{\sqrt{2\pi N}}\sum_{n=1}^{\sqrt{N}}\sqrt{\delta_{n,\kappa}+1}\right]+\frac{1}{\sqrt{T}}\mathbb{E}\left[\frac{s\kappa}{2}\xi_{\sqrt{N},\kappa}|\tilde{D}^N_{i\sqrt{N}}|\right]+O(\sqrt{\Delta t^N})\mathbb{E}\left[\frac{s\kappa}{2}\right]. \tag{3.45}$$

By definition of $\delta_{n,\kappa}$,

$$\frac{1-(1-\kappa)^{(2n+1)}}{\sqrt{\kappa(2-\kappa)}} \leq \sqrt{\delta_{n,\kappa}+1} = \sqrt{\frac{1-(1-\kappa)^{(2n+1)}}{\kappa(2-\kappa)}} \leq \frac{1}{\sqrt{\kappa(2-\kappa)}},$$
$$\frac{1}{\sqrt{N}}\frac{1}{\sqrt{\kappa(2-\kappa)}}\sum_{n=1}^{\sqrt{N}}\left(1-(1-\kappa)^{2(n+1)}\right) \leq \frac{1}{\sqrt{N}}\sum_{n=1}^{\sqrt{N}}\sqrt{\delta_{n,\kappa}+1} \leq \frac{1}{\sqrt{N}}\sum_{n=1}^{\sqrt{N}}\frac{1}{\sqrt{\kappa(2-\kappa)}}.$$

As a consequence,

$$\lim_{N\to\infty}\frac{1}{\sqrt{N}}\sum_{n=1}^{\sqrt{N}}\sqrt{\delta_{n,\kappa}+1} = \frac{1}{\sqrt{\kappa(2-\kappa)}}. \tag{3.46}$$

Both the lower and upper bound (3.44), (3.45) for the execution costs due to market-order transactions converge to the same continuous-time limit as $N \to \infty$:

$$\mathbb{E}\left[\frac{s\hat{\sigma}}{\sqrt{2\pi}}\sqrt{\frac{\kappa}{2-\kappa}}\right] + o(1).$$

More specifically, the convergence of the first terms directly follows from (3.46). The rest of the terms are of order $o(1)$. For the second term in the lower bound (3.44) and the upper bound (3.45), this follows from the $\mathbb{L}^2$ convergence of $\tilde{D}^N_{i\sqrt{N}}$ to zero by Proposition 3.8.2, and Cauchy-Schwarz inequality.

2. As $N \to \infty$, the expected earnings from the limit order trades converge to

$$\begin{aligned}
&\frac{1}{T/\sqrt{N}}\sum_{j=i\sqrt{N}+1}^{(i+1)\sqrt{N}}\mathbb{E}\left[\frac{s^N\nu}{2}|\Delta F^N_j|\right] \\
&\quad= \mathbb{E}\left[\frac{s\nu}{2\sqrt{\Delta t^N}}\left(v\sqrt{\Delta t^N}\sqrt{\frac{2}{\pi}}\exp\left(-\frac{b^2}{2v^2}\Delta t^N\right) + b\Delta t^N \mathrm{erf}\left(\frac{b\Delta t^N}{v\sqrt{\Delta t^N}}\right)\right)\right] \\
&\quad= \frac{s\nu v}{\sqrt{2\pi}} + o(1).
\end{aligned}$$

Here, similar to the previous part about expected transaction costs, we used the inequalities $\exp(-x) \geq 1-\sqrt{2x}, \mathrm{erf}(x) \leq 1, x \geq 0$ to obtain lower and upper bounds for the expected earnings, both of which have the same limit.

3. Finally, the expected adverse selection cost also converges as $N \to \infty$:

$$\frac{1}{T/\sqrt{N}} \sum_{j=i\sqrt{N}+1}^{(i+1)\sqrt{N}} \mathbb{E}\left[\nu \Delta F_j^N \Delta P_j^N\right] = -\mathbb{E}[\nu\rho\sigma v] + o(1).$$

In summary, the (normalized) expected execution costs $\frac{1}{T/\sqrt{N}}\mathbb{E}\left[\tilde{C}_i^N\right]$ converge to

$$\mathbb{E}\left[\frac{s\hat{\sigma}}{\sqrt{2\pi}}\sqrt{\frac{\kappa}{2-\kappa}} - \frac{s\nu v}{\sqrt{2\pi}} + \nu\rho\sigma v\right] + o(1), \quad \text{as } N \to \infty.$$

Similar to the sum of expected square displacements, as $N \to \infty$, the difference between the sum of these sub execution costs and the Riemann sum approximating the integral (3.39) therefore vanishes. Whence, the discrete execution costs $\mathbb{E}\left[\tilde{C}^N\right]$ for the model with piecewise constant controls and coefficients indeed converges to the integral (3.39). □

3.8.4.4 Proof for Theorem 3.6.2

To complete the proof of Theorem 3.6.2, it now remains to combine the results from Appendices 3.8.4.2 and 3.8.4.3 to show that the limits (3.15) and (3.16) also are obtained in the original model with arbitrary (rather than piecewise constant) controls and coefficients, and prove that the first term in (3.13) converges to the stochastic integral in (3.15). By Theorem 3.8.3, we already know that the difference between the execution costs (3.32) and (3.33) is negligible in the continuous-time limit $N \to \infty$.

We now argue that the same is true for the sum of the expected squared displacements. To see this, observe that by the inequalities of Cauchy-Schwarz and Minkowski,

$$\left|\mathbb{E}\left[(D_i^N)^2\right] - \mathbb{E}\left[(\tilde{D}_i^N)^2\right]\right| \leq \mathbb{E}\left[(D_i^N - \tilde{D}_i^N)^2\right]^{1/2} \left(\mathbb{E}\left[(D_i^N)^2\right]^{1/2} + \mathbb{E}\left[(\tilde{D}_i^N)^2\right]^{1/2}\right).$$

By Proposition 3.8.2, the last two terms are of order $O(N^{-1/2})$; by Theorem 3.8.3 the term multiplying them is of order $O(N^{-3/4})$, so that the total order is $O(N^{-5/4})$. In particular, it follows that the difference between the sums of expected squared displacements is of

order $O(N^{-1/4})$. Hence, $\sum_{n=1}^{N} \mathbb{E}\left[(D_n^N)^2\right]$ converges to the expectation of (3.16).

We now check the convergence of $\sum_{n=1}^{N}(\Pi_{n-1}^N - D_{n-1}^N)\Delta P_n^N$ to $\int_0^T \Pi_t dP_t$. This $\mathbb{L}^2$-convergence is equivalent to the $\mathbb{L}^2$ convergence of

$$\sum_{n=1}^{N} D_{n-1}^N \Delta P_n^N \to 0, \quad \text{as } N \to \infty. \tag{3.47}$$

By the tower property and recalling that the second moment of D_{n-1}^N and the second conditional moment of ΔP_n^N are of order $O(\Delta t^N)$, we have

$$\mathbb{E}\left[\left(D_{n-1}^N \Delta P_n^N\right)^2\right] = \mathbb{E}\left[\left(D_{n-1}^N\right)^2 \mathbb{E}\left[\left(\Delta P_n^N\right)^2 | \mathcal{F}_{t_{n-1}^N}\right]\right] \leq c\Delta t^N \mathbb{E}\left[\left(D_{n-1}^N\right)^2\right] \leq (c\Delta t^N)^2. \tag{3.48}$$

This estimate, Assumption 3.8.1(ii), the Cauchy-Schwarz inequality and the fact that the second moment of D^N is of order $O(\Delta t^N)$, in turn yield

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{n=1}^{N} D_{n-1}^N \Delta P_n^N\right)^2\right] &= \sum_{n=1}^{N} \mathbb{E}\left[\left(D_{n-1}^N \Delta P_n^N\right)^2\right] + 2\sum_{i,j=1, i<j}^{N} \mathbb{E}\left[D_{i-1}^N \Delta P_i^N D_{j-1}^N \Delta P_j^N\right] \\
&\leq N(c\Delta t^N)^2 + 2\sum_{i,j=1, i<j}^{N} \mathbb{E}\left[D_{i-1}^N \Delta P_i^N D_{j-1}^N \mathbb{E}\left[\Delta P_j^N | \mathcal{F}_{t_{j-1}^N}\right]\right] \\
&\leq N(c\Delta t^N)^2 + 2c\Delta t^N \sum_{i,j=1, i<j}^{N} \sqrt{\mathbb{E}\left[\left(D_{i-1}^N \Delta P_i^N\right)^2\right] \mathbb{E}\left[\left(D_{j-1}^N\right)^2\right]} \\
&\leq N(c\Delta t^N)^2 + 2c\Delta t^N \sum_{i,j=1, i<j}^{N} \sqrt{(c\Delta t^N)^2 \cdot c\Delta t^N} = O(N^{-1/2}).
\end{aligned}$$

The expectation of $\sum_{n=1}^{N}(\Pi_{n-1}^N - D_{n-1}^N)\Delta P_n^N$ thus indeed converges to the expectation of $\int_0^T \Pi_t dP_t$.

Altogether, we have established that the expectation of the discrete-time portfolio process (3.13) converges to the continuous-time limit (3.15). This completes the proof.

3.8.4.5 Proof of Lemma 3.6.3

Suppose a smooth solution C of the PDE (3.19) exists. Then, the terminal condition for C and Itô's formula yield

$$\varphi(P_T) = C(0, P_0) + \int_0^T \partial_p C(t, P_t) dP_t + \int_0^T \left(\partial_t C(t, P_t) + \frac{\sigma^2}{2} \partial_{pp} C(t, P_t) \right) dt.$$

Plugging this into (3.18) shows that – in the continuous-time limit – the payoff corresponding to an arbitrary partial tilted execution scheme (ν, κ) is

$$\mathbb{E}\left[\int_0^T \left(-\partial_t C(t, P_t) - \frac{\sigma^2}{2} \partial_{pp} C(t, P_t) + H(\nu_t, \kappa_t, \sigma \partial_{pp} C(t, P_t)) \right) dt \right].$$

By the PDE (3.19) for C, the integrand is nonpositive for any partial tilted execution scheme, and zero for the optimizer. Whence, the solution. of the PDE (3.19) is indeed an indifference price as asserted.

Chapter 4

Optimal Execution with Hidden Orders Under Self-Exciting Dynamics

This chapter is a joint work with Chao Zhou, Ge Zhang and Ying Chen.

4.1 Introduction

Hidden liquidity is attracting significant volume share in modern order-driven markets, reducing exposure risk and adverse selection risk. Despite lower pre-trade transparency, non-display orders[1] lose time priority to display limit orders under the price-visibility-time mechanism implemented in most exchanges. Although non-display limit orders (hidden orders therein) allow liquidity providers to mitigate exposure cost, execution only happens after the display orders of the same quoted price are fully executed, lowering the overall execution probability (Nasdaq, 2022)[2].

[1] In general, the market offers three types of instruments with the ability to hide one's order: dark pool, iceberg order and non-display limit order (hidden order). Dark pool is where all trading intentions are completely hidden from the public. Iceberg order is the order-type in lit pools that allows agents to hide the total volume placed, but not the limit price. Non-display limit order is another order-type in lit pools that not only hides the volume placed, but also the limit price. In this chapter, we focus on non-display limit orders (hidden orders therein).

[2] We quote the section 'Order Types and Functionality' in the website: *Nasdaq features a price/time priority model where the execution logic is fair and transparent for all market participants. All displayed limit orders are treated equally and executed in the order in which they were received at the same price.*

As suggested by the U.S. Securities and Exchange Commission(US Securities and Exchange Commission, 2019), hidden order is a critical type of order that reflects the pre-trade transparency on exchanges. The hidden volume ratio on NASDAQ has gradually increased from 10% at 2013 to 28% at 2022, as supported by Figure 1.2. Over all exchanges, it has grown from 9% in 2012 to 15% to 2022(CBOE Insights, 2022). Therefore, it is natural to conclude that hidden order has become an important instrument for on-exchange trading, and should be considered for liquidation tasks.

We develop an optimal liquidation strategy in a continuous-time framework taking into account the trade-off between exposure risk and trading priority, where a risk-neutral agent aims to maximize her terminal wealth with a combination of hidden limit orders and display limit orders over a fixed period, and the remaining shares are sold with a market order. The agent continuously controls the trading rate (order size) and order type (hidden and display) to balance execution cost and time pressure. Our theoretic model suggests that the optimal strategy follows a two-phase pattern, in which the agent changes from a pure-hidden-order phase to a mixed-order phase until termination. The transition takes place at the *switching curve* where the switching time depends on the intensity of the market order arrivals, see Figure 4.1. Before the switching curve, hidden order size increases in time; whereas after the switching curve, hidden order size decreases in time and limit order size increases in time.

Optimal execution in centralized limit order books has received considerable attention in the literature, see Cartea, Jaimungal, and Penalva (2015); Guéant (2016) and references therein. For an agent seeking to liquidate her position within a fixed period, immediate execution of a large order is either impossible or extremely expensive. The early studies on liquidation strategy focus on market orders, in particular how price impact or risk aversion structure affects the optimal liquidation decision. For example, Bertsimas and Lo (1998) derived a time-weighted average price (TWAP) strategy with market orders under the presence of a linear permanent price impact, R. Almgren and Chriss (1999) derived a volume-weighted average price (VWAP) strategy with both permanent and

Non-displayed shares are executed after displayed shares in the order in which they were received at that price.

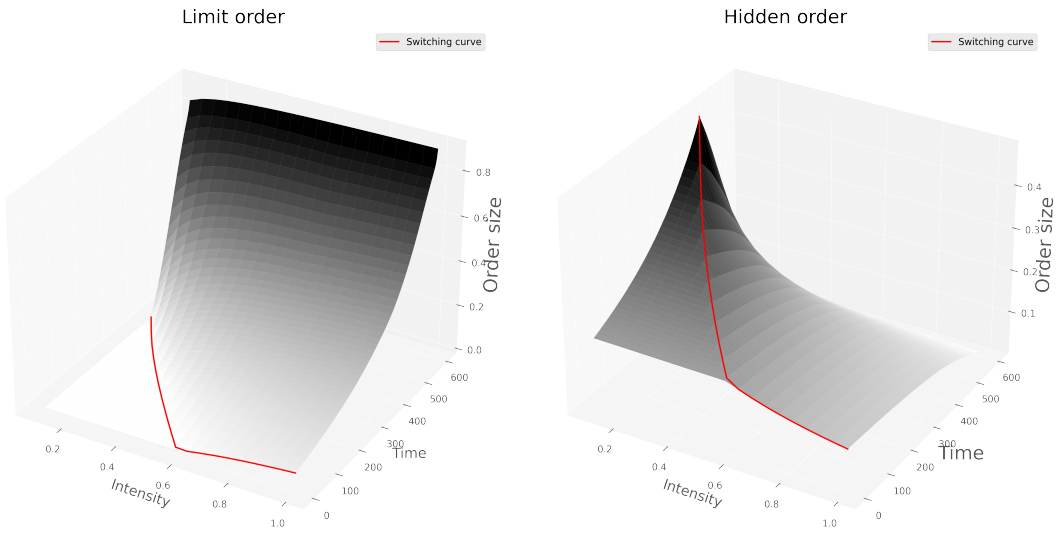

Figure 4.1: Optimal liquidation strategy and switching curve.

immediate execution costs. See also R. Almgren (2012); R. Almgren and Chriss (2001); R. F. Almgren (2003); Belak, Muhle-Karbe, and Ou (2020); Cayé and Muhle-Karbe (2014); Graewe and Horst (2017); Huberman and Stanzl (2005); Lehalle and Neuman (2019); Neuman and Voß (2016, 2022); Obizhaeva and Wang (2013) for extensions in price impact, risk aversion and volatility dynamics.

Compared to market orders, limit orders serve a pivotal role in price discovery, liquidity, and information disclosure. For their greater flexibility in execution price, limit orders are extensively used in trade execution so that agents can control the execution process through controlling either price quote or trading rate. The price quote strategy assumes that order execution follows a certain point process (whose intensity is function of price quote) and the agent controls the price quote to achieve optimal execution outcome, see for example Cartea and Jaimungal (2015a); Cartea and Sánchez-Betancourt (2023); Guéant et al. (2012); Guilbaud and Pham (2013). The trading rate strategy considers uncertainty in the inventory process that propagates to the order fill uncertainty, and the agent controls over trading rate in terms of frequency or order size, see Cheng, Di Giacinto, and Wang (2017) and Bulthuis, Concha, Leung, and Ward (2017).

Limit orders displayed in the book induce exposure risk caused by e.g. the reveal of trader's motive and price impact of future trades. This stimulates liquidity imbalance

and drives the price to a less favorable direction of the agent (Aitken, Berkman, & Mak, 2001; Harris, 1997; Pardo & Pascual, 2012). For example, when an agent places a large sell limit order, the order book imbalance becomes more selling-heavy, where the exposure risk are increasing in the limit order size (Abergel, Anane, Chakraborti, Jedidi, & Toke, 2016; Cao, Hansch, & Wang, 2009; Cartea, Donnelly, & Jaimungal, 2018; Chordia, Roll, & Subrahmanyam, 2002; Jacquier & Liu, 2018; Lehalle & Neuman, 2019; Neuman & Voß, 2022).

Moreover, the derivation of an executable liquidation strategy depends on market microstructure. There has been a significant economic and econometric literature on modeling the stylized fact of order book's microstructure and its interaction with optimal liquidation, see Abergel and Jedidi (2013); Biais, Hillion, and Spatt (1995); Cont and De Larrard (2013); Easley and O'hara (1987); Glosten and Milgrom (1985); Horst and Paulsen (2017) and references therein. Among others, market order arrivals are often assumed to follow the Poisson process in most of the above-mentioned works. However, when exposing to the same news flow, market order arrivals display significant clustering and dispersion effects, where market order submissions are clustered and further stimulate child orders. The self-exciting property in point process of order arrivals can be modelled by Hawkes process (Hawkes, 1971) analogous to the models for trade arrivals in high-frequency markets (Aït-Sahalia, Cacho-Diaz, & Laeven, 2015; Andersen, Bondarenko, Kyle, & Obizhaeva, 2018; Bacry, Delattre, Hoffmann, & Muzy, 2013; Huang, Lehalle, & Rosenbaum, 2015; Lehalle, Mounjid, & Rosenbaum, 2021), volatility process (Bates, 2019; El Euch, Rosenbaum, et al., 2018; Horst & Xu, 2022), limit order book (Horst & Xu, 2019), and market impact and microstructure (Alfonsi & Blanc, 2016a; Bacry, Mastromatteo, & Muzy, 2015; Cartea, Jaimungal, & Ricci, 2014). In the literature of optimal liquidation, Cartea et al. (2014) constructed a high-frequency trading strategy with a Hawkes process model. (Bacry, Iuga, Lasnier, & Lehalle, 2015) considered the mutually exciting property of order arrivals on the best bid or ask price movements. Alfonsi and Blanc (2016a) studied the optimal execution problem under the Hawkes process. Alfonsi and Blanc (2016b) discussed the calibration methods for the proposed

dynamic optimal execution model. Fu et al. (2022) studied a mean-field game where order flow follows a Hawkes process.

In our study, we develop the optimal liquidation strategy for mixed types of limit orders and investigate the effects of time pressure and market order intensity on the optimal liquidation strategy. In our model, both display and hidden limit orders incur the immediate execution cost which is quadratic in executed order size, and the exposure cost only applies to display orders which is linear in both exposure time and display limit order size. We show that the agent's decision on order size is driven by the cost and time trade-off and the decision on order type needs to balance exposure risk and fill probability. We consider two types of market order arrivals dynamics. When market order arrivals follow a homogeneous Poisson process, we are able to derive a closed-form solution. It shows that there is a switch time, at which the agent shifts from hidden orders only to a mixture of display limit orders and hidden orders until termination. In a more general case, where arrivals follow the Hawkes process, a numerical solution is derived with feedback controls. There is a similar two-phase strategy with self-exciting orders, except that the switching time becomes a function (or a switching curve) of the market order intensity. In addition, the theoretical model implies different impact of time pressure on order size under the two phases. Simulation experiments show that the use of hidden orders reduces liquidation cost, particularly in exposure cost, accompanied by an increase in liquidity. In addition, our strategy demonstrates robustness on estimation errors and varying market conditions. Given event-level limit data of NASDAQ 100 stocks, we first estimate the model parameters from the limit order book data, and then test our liquidation strategies using real market order arrivals. The results show that our strategy with mixed order types under the self-exciting dynamics demonstrates superior performance, with cost reduction up to 57% to the pure limit order strategy and 15% to the strategy with mixed order types under the Poisson process.

Our work belongs to the optimal liquidation study considering hidden orders under the self-exciting effect of the market order arrivals. It is inspired by early works but differs in some aspects. For example, Y. Chen, Gao, and Li (2018) develops a discrete time model

for dynamic optimal liquidation with hidden orders in lit pool. Esser and Mönch (2007) considers a continuous-time liquidation problem with iceberg orders. Horst and Naujokat (2014) discusses a similar problem with emphasis on the no-trade region and suggests the agent should use market orders only when the market spread is small enough. Cebiroğlu and Horst (2015) discusses optimal trade-off between market impact and priority gain in liquidity competition using hidden orders. Similar literature in dark pool trading also shares much relevance to our work: Kratz and Schöneborn (2014, 2015) demonstrate that dark pool liquidity is predictable for future price movement, and suggested that it is optimal for the agent to submit orders to a dark pool to survey the information. Crisafi and Macrina (2016) studies an simultaneous trading problem in both lit and dark pools, and takes partial execution of orders into account. Moreover, game-theoretical study of hidden order usage explores the problem from a financial economics perspective. The insider trading problem with iceberg orders is discussed in Moinas (2005). Buti and Rindi (2013) studies an order submission game with traders placing market, limit and iceberg orders. Boulatov and George (2013) analyses the impact of hidden liquidity on market quality in an extended Kyle model (Kyle, 1985, 1989) where an informed trader serves both as liquidity provider and liquidity consumer. Similar to these works, the queuing dynamics is not considered in our scheduling algorithm. To operate at the highest frequency, see Huang et al. (2015); Lehalle et al. (2021).

Among others Y. Chen et al. (2018) is the closest to our work with four major differences:

1. Y. Chen et al. (2018) assumed that the price impact of a limit order is triggered only upon execution of orders, while we consider that the price impact is incurred since limit order submission because of the exposure of trading motif, regardless of the order fills (Cont et al., 2014). This critical difference not only accurately reflects the empirical evidence on price impact of limit order via order book imbalance, but also allows explicit modeling of the benefits of hidden order over limit order in our theoretical framework.

2. Our cost structure considers immediate execution cost of hidden orders, with which

very large sized hidden order is sub-optimal whereas Y. Chen et al. (2018) advocated the usage of extremely large hidden orders.

3. Y. Chen et al. (2018) considered constant market order arrival rates, while our model relaxes the assumption. Specifically, in our model the arrival rate of market orders is non-homogenous and exhibits self-exciting properties. In addition, extensive simulation and real data testing suggest that our model with self-exciting dynamics outperforms.

4. Compared to the discrete-time setting in Y. Chen et al. (2018), we propose a continuous-time control framework to study the joint usage of hidden and limit order.

The contribution of this chapter is multi-fold. Firstly, we extend the optimal liquidation literature by developing a continuous-time dynamic model with hidden orders under self-exciting market order dynamics. Our theoretical model offers a scheduling solution to traders to achieve their liquidation objective in a realistic setting. Secondly, we elaborate on the switching time of the strategies with mixed order types and present the interplay between the switch of execution strategies and market factors. Lastly, we display numerical performance with both simulations and NASDAQ 100 stocks to validate the effectiveness and robustness of the solution, which complements the insights from theoretic models.

The outline of the chapter is as follows: Section 4.2 presents the model setup. Section 4.3 details the theoretical solution with various types of market order arrival processes. Section 4.4 demonstrates numerical performance with simulations. Section 4.5 conducts real data analysis. Section 4.6 concludes.

4.2 Model Setup

We investigate the optimal use of hidden orders along with display limit orders (limit orders thereafter) in share liquidation, where a risk-neutral agent aims to maximize her

terminal wealth. The agent controls both order size and order type to achieve the liquidation target over a fixed period.

4.2.1 Dynamics

We consider an agent who wishes to liquidate Q_0 shares in a lit exchange within a fixed period $[0, T]$. The agent continuously decides the order size of her limit and hidden orders, denoted by $(L_t)_{t\in[0,T]}$ and $(H_t)_{t\in[0,T]}$, respectively. We assume that the agent can only use sell orders (i.e., exclude the possibility of using buy limit order or buy hidden order, which is considered *spoofing* or *price manipulation* here.) At any time t, we have $L_t \geq 0, H_t \geq 0$. [3]

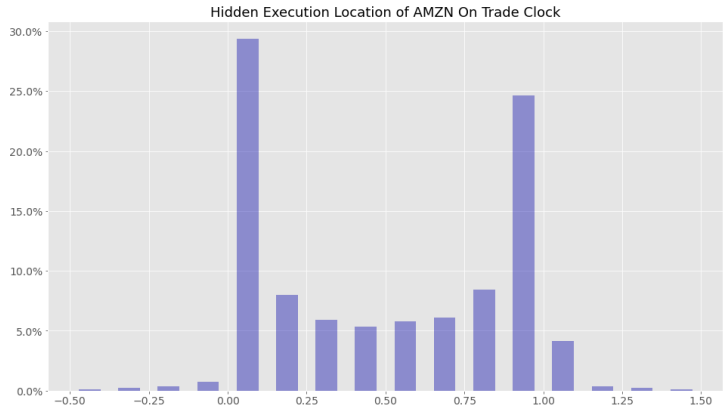

Figure 4.2: Location of executed hidden order with respect to the best bid/ask prices in the transaction data of Amazon in January 2019. 0 represents execution at the best bid price and 1 represents execution at the best ask price.

We assume that the price of both limit and hidden order is pegged at the best ask price. This assumption is supported by the empirical observation, e.g., nearly 100% of the iceberg orders (which is a similar order type to hidden orders) are pegged at the best price for the CME GLOBEX ESH1 futures on January 13, 2011 (Christensen & Woodmansey, 2013). We have also carried out our own empirical analysis on the transaction data of several NASDAQ stocks for the entire January of 2019 and found that for more liquid

[3]Here we note that the difference between this strategy and posting an iceberg order with adjustable peak size is: (1) The hidden part of iceberg order cannot be executed directly. One has to exhaust the visible part of the iceberg order in order to release more from the hidden part to be executed. (2) This strategy allows zero visible part while iceberg order does not.

assets, the execution price of hidden orders are close to the best bid/ask prices for most of the times as indicated in Figure 4.2. This tells us that hidden orders are seldomly placed within the spread. Note that different from the usual assumption that the limit price differs from the best ask price, the agent continuously adjusts her order such that the order size is always currently optimal and the limit price of the orders match the current best ask price. The fluctuation of the asset price is assumed to be driven by the flows of information, e.g., order book events from other market participants, and modeled with Brownian motion. [4] In addition, we assume that limit orders expose the agent's selling intention and thus induce an additional impact on asset price. (Cont et al., 2014) documented that this extra pressure is of almost 'constant' magnitude from the time of order submission to the time of order execution or cancellation. Thus, we consider a fundamental price process S representing the best ask price of the limit order book as follows:

$$dS_t = \sigma dW_t - bL_t dt, \tag{4.1}$$

where W_t is standard Brownian motion, $\sigma > 0$ denotes the volatility of asset, $b > 0$ is the *exposure risk coefficient*, in the same spirit of Chordia et al. (2002). We use an exogenous counting process $(N_t)_{t\in[0,T]}$ to record the number of market order arrivals. Without loss of generality, we assume that N_t follows an **inhomogeneous** Poisson point process whose *intensity* λ_t[5] follows certain known stochastic differential equation. Here, instead of modeling how clustering effects in trading activities are brought by specific agent actions, e.g. informed trading, we directly go from the perspective of an agent where no such information is available and assume that the clustering effects are random. One can choose to improve this model by modelling the correlation between trade imbalance and clustering, which is not the focus of this work and is therefore not discussed here.

[4] We use the Arithmetic Brownian motion model, which theoretically allows the possibility of having a negative price. However note that in short timeframe Arithmetic Brownian motion behaves very similarly to Geometric Brownian motion.

[5] We note here that λ represents intensity of Poisson process in this chapter, which is different from its meaning in the previous chapters. This is because λ is both the standard notation for price impact and intensity.

The agent's inventory process has the following dynamic:

$$dQ_t = -\mathbb{1}_{\mathcal{L}_{1+N_{t-}}}(L_t + \mathbb{1}_{\mathcal{H}_{1+N_{t-}}} H_t)dN_t, \quad t \in [0, T], \tag{4.2}$$

where $\{\mathcal{L}_n\}_{n\in\mathbb{N}}$ and $\{\mathcal{H}_n\}_{n\in\mathbb{N}}$ denote the agent's order fill events (with binary outcomes) for limit and hidden orders, respectively. They are assumed to follow independent and identical Bernoulli distribution with constant *success probability* p_l and p_h, which are independent of any other process. At time t, limit orders are executed at size $\mathbb{1}_{\mathcal{L}_{1+N_{t-}}} L_t$ and hidden orders at size $\mathbb{1}_{\mathcal{L}_{1+N_{t-}}} \mathbb{1}_{\mathcal{H}_{1+N_{t-}}} H_t$. This reflects the fact that hidden orders are executable only after exhaustion of the limit order queue, see Figure 4.3 for graphical illustration of orders queue. As a technical assumption, we disallow orders to be partially filled, i.e., orders are always fully filled. We do not expect this to affect the qualitative results drastically and is therefore not a major concern of this work.

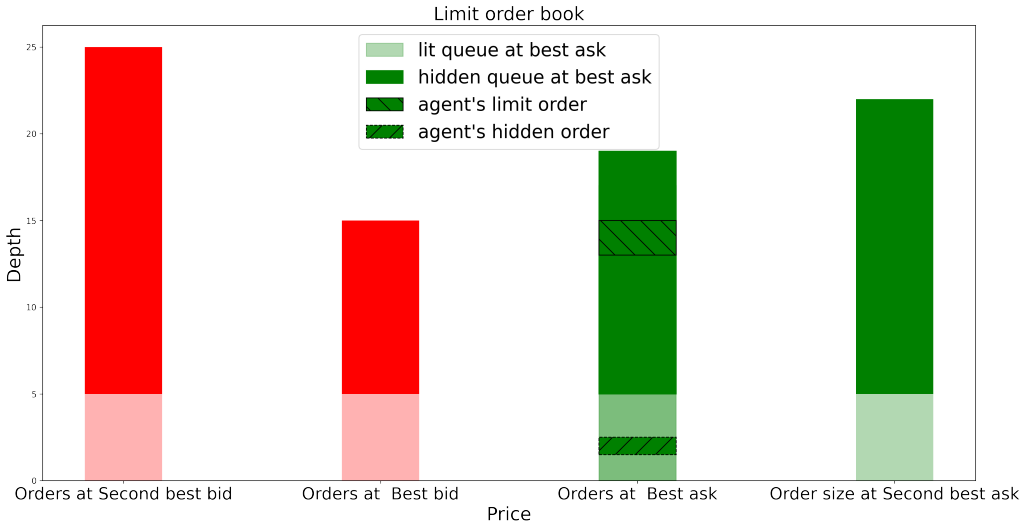

Figure 4.3: A visualization of hidden order in limit order book.

Lastly, cash process X satisfies the following SDE:

$$dX_t = -\hat{S}_t dQ_t, \quad t \in [0, T], \tag{4.3}$$

where $\hat{S}_t$ denotes the transaction price at time t. Assuming that the immediate execution

cost is proportional to the order sizes L and H, the transaction price at time t has the form:

$$\hat{S}_t = S_t - k\mathbb{1}_{\mathcal{L}_{1+N_{t^-}}}(L_t + H_t\mathbb{1}_{\mathcal{H}_{1+N_{t^-}}}), \quad t \in [0,T], \tag{4.4}$$

where $k > 0$ is the *immediate execution cost coefficient*. Given a fixed execution price, immediate execution cost is the opportunity cost of a sizeable sell limit order which may stop the market buy order(s) from walking up the book temporarily, thus resulting in an immediate (opportunity) loss. In this regard, the immediate execution cost is the difference between the execution price and the near-future price, which suggests opportunity cost of agent against if executing less and placing limit orders again when price moving up. Conceptually, this immediate execution cost as part of the price impact for limit and hidden orders also applies for market orders (R. Almgren & Chriss, 2001). It corresponds to the instantaneous price impact caused by the execution of market orders. Note that the price impacts of market orders and passive orders differ mainly in the permanent component, i.e. the order pressure. Market orders exert a much larger persistent price impact compared to limit or hidden orders.

Here we establish a price impact model with the anonymity assumption (Tóth et al., 2011), where the aggressive of trading strategies is only taken into account through the size of the posted order. We acknowledge that the main limitation to this model is that agents can only peg the limit prices to the current best ask price, which is a technical assumption that helps to focus on the effect of order sizes on liquidation, just like the focus on the effect of limit price in the other literature about limit order executions Cartea and Jaimungal (2015a); Guéant et al. (2012); Guilbaud and Pham (2013). This model is not particularly suitable for illiquid markets where the market order arrival is heavily affected by the amount of liquidity. In turn, we tend to focus on a more liquid market where the market order arrival rate is mostly driven by external factors.

4.2.2 Value Function

Throughout this chapter, we work on a completed filtered probability space $(\Omega, \mathbb{P}, (\mathcal{F}_s)_{0\leq t\leq T})$ where the filtration is generated by the processes W_t and N_t. The performance of the agent is measured by the expected terminal wealth, which is the expectation of the terminal cash position and stock position less the price impact cost of the market order at termination T:

$$V(t,x,s,q,\lambda) = \sup_{(L,H)\in\mathcal{A}_{[t,T]}} \mathbb{E}_{t,x,s,q,\lambda}\left\{X_T + Q_T S_T - \alpha Q_T^2\right\}. \tag{4.5}$$

where $\alpha > 0$ denotes the coefficient for *market order price impact* and $\mathbb{E}_{t,x,s,q,\lambda}$ denotes the expectation conditional on $X_{t^-} = x$, $S_t = s$, $Q_{t^-} = q$ and $\lambda_{t^-} = \lambda$. The supremum is taken over all $\mathcal{F}_t$-predictable pairs of processes $(L_\tau, H_\tau)_{t\leq\tau\leq T}$, which satisfies the boundedness condition: $\forall\,\tau \in [t,T), 0 \leq L_\tau, 0 \leq H_\tau$, and $L_\tau + H_\tau \leq Q_\tau$. Note that λ_t is progressively measurable with respect to the natural filtration of N_t.

4.3 Optimal strategy with hidden order

We derive optimal strategies for the liquidation problem under three scenarios. By assuming that market order arrivals follow a homogenous Poisson process, we obtain the Limit-hidden (LH) strategy. When restricting to limit orders only, we derive the Limit-only (LO) strategy, which can be used to evaluate how much cost is reduced from using mixed orders. Finally, we extend the derivation to a Limit-hidden-Hawkes strategy under self-exciting market order arrivals.

4.3.1 Limit-hidden (LH) Strategy

We first consider a simplified case where the market order arrival process is homogenous, i.e., λ_t is constant. In particular, N_t follows a homogeneous Poisson process with a constant intensity $\lambda_t = \lambda$ at any $t \in [0,T]$. As such, the value function V in (4.5) is reduced to a simple value function, denoted by U, where the constant intensity λ is no

longer a state variable but a parameter:

$$U(t,x,s,q) = \sup_{(L,H)\in\mathcal{A}_{[t,T]}} \mathbb{E}_{t,x,s,q}\left\{X_T + Q_T S_T - \alpha Q_T^2\right\}. \tag{4.6}$$

Analogously, the admissible set of controls is still $\mathcal{A}_{[t,T]}$, which stipulates that the total order size cannot exceed current inventory and both order sizes cannot be negative. The expectation $\mathbb{E}_{t,x,s,q}$ is conditional on $X_{t^-} = x, S_t = s, Q_{t^-} = q$ only. In the following, we develop an analytical solution to this control problem.

It implies that a dynamic programming principle holds and the value function U in (4.6), without the dynamics of the intensity process anymore, is the solution of the HJB equation:

$$\begin{aligned} 0 = {} & \frac{\partial U}{\partial t} + \frac{1}{2}\sigma^2 \frac{\partial^2 U}{\partial s^2} \\ & + \sup_{(L,H)\in A(q)} \Bigg\{ \tilde{\lambda}\Big[p_h\big(U(t, x+(L+H)(s-kL-kH), s, q-L-H)\big) \\ & + (1-p_h)U(t, x+L(s-kL), s, q-L) - U\Big] - bL\frac{\partial U}{\partial s}\Bigg\}, \end{aligned} \tag{4.7}$$

with terminal condition

$$U(T,x,s,q) = x + qs - \alpha q^2$$

and

$$A(q) \coloneqq \{(L,H) \mid L \geq 0, H \geq 0, L+H \leq q\}.$$

Note that the execution intensity of limit order is denoted as $\tilde{\lambda} = \lambda p_l$. We adopt an excess value decomposition to break down the optimal liquidation value $U(t,x,s,q)$ into the current cash value x, current stock position qs, and excess value $u(t,q)$. We show that this excess value is purely quadratic in the inventory q, and the feedback control forms of the order sizes can be expressed as a linear function of q, given certain assumption about the value function as follows:

Proposition 4.3.1 (Value Function Decomposition, Feedback Control)**.** *The value function U admits the excess value decomposition $U(t,x,q,s) = x + qs + u(t,q)$ and the excess*

value $u(t,q)$ is purely quadratic in the inventory q: $u(t,q) = q^2 u_2(t)$.

Assuming that u_2 is bounded by $-\alpha$ and 0, the unique feedback controls of the optimal liquidation strategy are:

$$\begin{aligned} L^*(t,Q_t) &= Q_t \cdot \max\left(1 - \frac{k + \frac{b}{2\tilde{\lambda}(1-p_h)}}{k - u_2(t)}, 0\right), \\ H^*(t,Q_t) &= Q_t \cdot \left(1 - \frac{k}{k - u_2(t)}\right) - L^*(t,Q_t). \end{aligned} \tag{4.8}$$

The proof is presented in Appendix 4.7.1.1. With the feedback controls and decompositions, the HJB (4.7) can be reduced into a system of ODEs. Solving the dynamic programming equation through a verification argument, we prove that the unique classical solution to the HJB (4.7) is equal to the value function of the optimal liquidation problem defined in (4.6). Further, the obtained classical solution satisfies the boundedness assumption on u_2 in Proposition 4.3.1.

Theorem 4.3.2 (Limit-hidden Strategy, Solving the HJB Equation, Verification Theorem)**.** *With the decomposition in Proposition 4.3.1, the unique classical solution of the HJB equation* (4.7) *coincides with the value function U of the control problem* (4.6)*.* (4.8) *defines a set of optimal controls in feedback form.*

In particular, $u_2(t)$ is a piecewise function in time:

- *For $t \in (\tilde{t}, T]$ where $\tilde{t}$ is a switching time to be determined afterwards, u_2 satisfies the following equation:*

$$k_1 - t = \frac{1}{\tilde{\lambda}\nu}\left(k + \frac{b}{2\tilde{\lambda}}\right)\tan^{-1}\left(\frac{\tilde{u}_2(t)}{\nu}\right) - \frac{1}{2\tilde{\lambda}}\log\left(\tilde{\lambda}\tilde{u}_2^2(t) + \tilde{\lambda}\nu^2\right), \tag{4.9}$$

  *where $\tilde{u}_2(t) = u_2(t) + \frac{b}{2\tilde{\lambda}}$, $\nu = \frac{b}{2\tilde{\lambda}}\sqrt{\frac{p_h}{1-p_h}}$, k_1 is a constant determined by terminal condition $\tilde{u}_2(T) = -\alpha + \frac{b}{2\tilde{\lambda}}$ and the **switching time** $\tilde{t}$ is given by inverting* (4.9) *with $\tilde{u}_2(\tilde{t}) = -\frac{bp_h}{2\tilde{\lambda}(1-p_h)}$.*

- *For $t \in [0, \tilde{t}]$,*

$$u_2(t) = -\left[W\left(\phi \exp(\phi + \tilde{\lambda}\, p_h(\tilde{t} - t))\right)\right]^{-1}. \tag{4.10}$$

where $\phi = \frac{2\tilde{\lambda}(1-p_h)}{b}$ *is the relative execution cost and the Lambert function* W *is the inverse function of* $g(x) = x\exp(x)$.

We defer the proof to Appendix 4.7.1.2.

Remark 4.3.3. *Optimally, the agent should enter a pure-hidden-order phase before time* $\tilde{t}$*, and convert it to a mixed-orders phase after time* $\tilde{t}$.

For the pure-hidden-order phase, i.e., $u_2(t) \geq -\frac{b}{2\tilde{\lambda}(1-p_h)}$*, the optimal limit order size* L^* *is zero and the optimal hidden order size is proportional to inventory* Q_t*:*

$$H^*(t, Q_t) = Q_t \cdot \left(1 - \frac{k}{k - u_2(t)}\right). \tag{4.11}$$

For the mixed-orders phase, i.e., $u_2(t) < -\frac{b}{2\tilde{\lambda}(1-p_h)}$*, the optimal order sizes are again proportional to inventory* Q_t*:*

$$L^*(t, Q_t) = Q_t \cdot \left(1 - \frac{k + \frac{b}{2\tilde{\lambda}(1-p_h)}}{k - u_2(t)}\right), \quad H^*(t, Q_t) = Q_t \cdot \left(\frac{\frac{b}{2\tilde{\lambda}(1-p_h)}}{k - u_2(t)}\right). \tag{4.12}$$

The feedback control form of total order size does not change throughout the entire period regardless of the phase, i.e.,

$$L^*(t, Q_t) + H^*(t, Q_t) = Q_t \cdot \left(1 - \frac{k}{k - u_2(t)}\right), \tag{4.13}$$

which is decreasing in $u_2(t)$.

In addition, it is only when $\dfrac{b}{2\tilde{\lambda}(1-p_h)} < \alpha$ *that the optimal liquidation strategy has two phases. Otherwise, the agent would use the pure-hidden-order phase for the entire timeframe. To give an intuition about the* ***switching****: During illiquid periods, one should want to hide limit orders because the exposure of trading intention causes rather large impact. This is empirically corroborated in the literature (J.-P. Bouchaud et al., 2018), and the event Flash Crash (CFTC & SEC, 2010).*

Note that the order splitting between passive order (limit or hidden order) and the market order is always present. That is, even if the immediate execution cost k is larger

than the price impact of market order α, the agent should still post passive order throughout the period, instead of staying inactive and waiting to execute all shares using the market order at termination.

Remark 4.3.4. *According to the expression of the function u_2, it is bounded by $[-\alpha, 0]$, thus the assumption in Proposition 4.3.1 is validated. From the proof presented in Appendix 4.7.1.2, we have the ODE governing the dynamics of u_2, with terminal condition $u_2(T) = -\alpha$:*

$$0 = u_2'(t) + \frac{\tilde{\lambda}}{k - u_2(t)} \left[\left(u_2^2(t) + u_2(t)\frac{b}{\tilde{\lambda}} + \frac{b^2}{4\tilde{\lambda}^2(1-p_h)} \right) \mathbb{1}_{\{u_2(t) \leq -\frac{b}{2\tilde{\lambda}(1-p_h)}\}} + p_h u_2^2(t) \mathbb{1}_{\{u_2(t) > -\frac{b}{2\tilde{\lambda}(1-p_h)}\}} \right].$$

From the ODE, u_2 is monotonically decreasing in time t. The value of u_2 directly reflects the level of impatience or aggressiveness of the agent. Over time, the value of u_2 decreases, and the agent becomes more anxious and uses larger-sized orders for liquidation.

Unlike traditional optimal liquidation literature with market orders, our solution does not require the price impact of market orders α to be larger than the immediate execution cost k.

Moreover, it is critical to understand the relationship between dynamics of the value function and the model parameters, through the ODE describing function u_2 in the mixed-orders phase:

$$0 = (k - u_2(t))u_2'(t) + \tilde{\lambda} \left(u_2(t) + \frac{b}{2\tilde{\lambda}} \right)^2 + \frac{b^2}{4\tilde{\lambda}} \frac{p_h}{1-p_h}.$$

One can see that the speed of time decay of u_2 is positively related to $\nu = \frac{b}{2\tilde{\lambda}} \sqrt{\frac{p_h}{1-p_h}}$ and is scaled by $k - u_2(t)$. Economically, it implies that the speed at which the cost aggravates over time is positively related to exposure risk and fill probability hidden orders, and negatively related to immediate execution cost and execution intensity of the limit orders.

Definition 4.3.5. *We define the relative minimal liquidation cost as $c_\alpha(t) \coloneqq -u_2(t)/\alpha, \quad t \in [0, T]$.*

We introduce the relative minimal liquidation cost for the ease of analysis. From

the terminal condition of u_2, we have $c_\alpha(T) = 1$. The agent's value function can be reconstructed using c_α: $U(t, x, q, s) = x + qs - c_\alpha(t)\alpha q^2$, $\forall\, t \in [0, T]$. Here u_2 (c_α) is to quantify the utility (cost) of the agent at a given state, and $c_\alpha(t)$ is the ratio between the minimal liquidation cost and the price impact of market order, which represents the execution cost relative to executing all remained shares using market orders at once. According to Theorem 4.3.2, the relative cost coefficient c_α is bounded between 0, where the cost is almost entirely avoided, and 1, where the execution of all inventory using market orders is almost unavoidable.

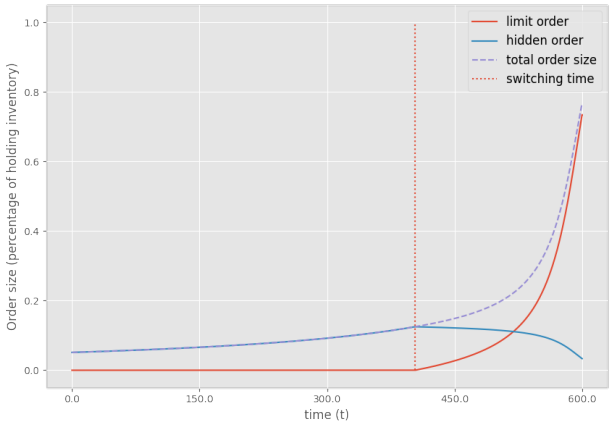

Figure 4.4: Switching time and order size of the Limit-Hidden strategy over time.

Figure 4.4 displays the order size, measured as the percentage value over the remaining inventory. The switching time displayed as a dotted vertical line. It shows that the agent adopts a two-phase strategy. Before the switching time, she enters a pure-hidden-order phase, and switches to the mixed-orders phase after. We observe that the relative minimal liquidation cost $c_\alpha(t)$ grows monotonically with time. The hidden order size increases with time before the switching time and decreases after. The limit order size increases immediately after the switching time. Moreover, the total order size, i.e., the sum of hidden and limit orders, increases monotonically with time. The optimal choice is to use pure hidden orders when the pressure to liquidate is not too strong, and otherwise to use the mixture of hidden orders and limit orders, similar to Y. Chen et al. (2018), . We also conducted analysis of the impact of other model parameters k, α, b, p_l and p_h on the

limit-hidden strategy, which can be found in Appendix 4.7.3.

4.3.2 Limit-only (LO) Strategy

To elaborate how much the agent can benefit from using a mixed strategy of both limit and hidden orders, we consider a pure limit order strategy in liquidation and use it as a benchmark strategy. In the limit-only case, the admissible set of controls further shrinks down to $\mathcal{A}^L_{[t,T]}$, and the set of all $\mathcal{F}_t$-progressively measurable processes $(L_\tau)_{t\leq\tau\leq T}$ satisfies $\forall\tau\in[t,T)$, $0\leq L_\tau\leq Q_\tau$. The corresponding value function in this case, denoted as U^L, is defined as follows:

$$U^L(t,x,s,q)=\sup_{L\in\mathcal{A}^L_{[t,T]}}\mathbb{E}_{t,x,s,q}\left\{X_T+Q_TS_T-\alpha Q_T^2\right\}. \tag{4.14}$$

Corollary 4.3.6 (Limit-only Strategy)**.** *If the market order arrival rate is constant, i.e., $\forall\, t\in[0,T], \lambda_t=\lambda$ and the agent can only use limit orders, the value function is quadratic in the inventory q: $U^L(t,x,q,s)=x+qs+q^2u_2^L(t)$ where u_2^L is a function of time with the following expression:*

1. *If $\alpha>\frac{b}{2\tilde{\lambda}}$, the agent is active:*

$$u_2^L(t)=-\tilde{k}\Big[W\left(\tilde{\phi}\exp(\tilde{\phi}+\tilde{\lambda}(T-t))\right)\Big]^{-1}-\frac{b}{2\tilde{\lambda}}, \tag{4.15}$$

where $\tilde{\lambda}=\lambda p_l$ is the execution intensity and $\tilde{\phi}=\frac{\tilde{k}}{\tilde{\alpha}}$ is the modified relative execution cost with $\tilde{k}=k+\frac{b}{2\tilde{\lambda}}$, $\tilde{\alpha}=\alpha-\frac{b}{2\tilde{\lambda}}$.

The corresponding optimal limit order size L^* has the following expression:

$$L^*(t,Q_t)=Q_t\left[1+W\left(\tilde{\phi}\exp(\tilde{\phi}+\tilde{\lambda}(T-t))\right)\right]^{-1}. \tag{4.16}$$

2. *If $\alpha\leq\frac{b}{2\tilde{\lambda}}$, the agent is inactive: L^* is always zero and for all time t, $u_2^L(t)=-\alpha$.*

We defer the proof to Appendix 4.7.1.3.

The limit order size $L^*(t, Q_t)$ is proportional to the current inventory Q_t. As the Lambert function W is increasing in the positive part of real-axis, the optimal exposure amount is inversely related to $\tilde{\phi} = \left(k + \frac{b}{2\tilde{\lambda}}\right) / \left(\alpha - \frac{b}{2\tilde{\lambda}}\right)$, which is increasing in k and b, and decreasing in α and $\tilde{\lambda}$. It implies that, immediate execution cost and exposure risk lower the size of limit order, while the price impact of market order and market order arrival frequency drive it up. Compared to Theorem 4.3.2, one key difference from the **Limit-hidden** (LH) strategy is the existence of a dormant phase. When the market order price impact is smaller than the exposure risk, $\alpha \leq \frac{b}{2\tilde{\lambda}}$, the **Limit-only** (LO) strategy instructs the agent not to post any limit order but waits until termination T.

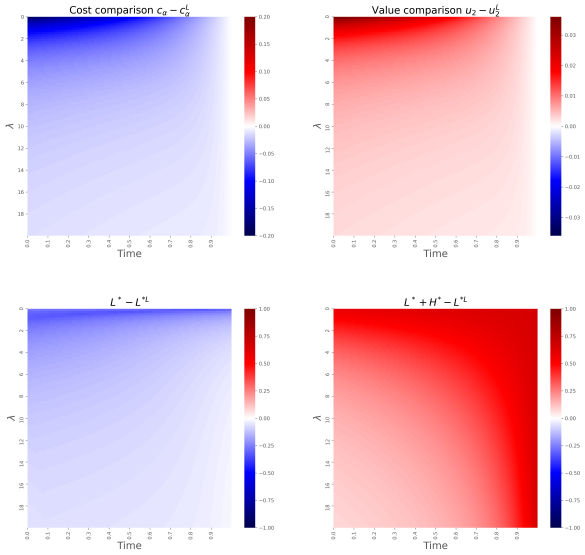

Figure 4.5: Comparison of value/cost (upper panel) and order size (lower panel) between Limit-Hidden strategy and Limit-Only strategy. c_α represents the minimal liquidation cost relative to the impact of market orders using the Limit-Hidden strategy, and c_α^L represents the minimal liquidation cost relative to the impact of market orders using the Limit-Only strategy.

When the agent is constrained to limit orders only, her value function and the total execution cost will be sub-optimal. Figure 4.5 visualizes the intuition, which shows that the limit-hidden strategy is always better than the limit-only in terms of execution cost and value function. The gap between the two strategies is particularly prevalent when the market is less active. For example, when intensity λ approaches 0, there are hardly any trades happening. Here, the priority of the agent would be to hide her trading volume as

much as possible, by e.g., using hidden orders.

Interestingly, it shows that the market also benefits from the use of hidden orders by the agent, with an increase in total liquidity supply. For an easy demonstration, let's assume our agent is the aggregate of all traders in the market. Our model shows that the lit depth of the order book reduces, as shown in the bottom panel of Figure 4.5, and the total liquidity, i.e., the sum of limit and hidden order size[6], increases with the use of hidden order against limit only, with the caveat that things might change with endogenous market order arrival rates.

4.3.3 Limit-hidden-Hawkes Strategy (LHH)

In this section, we study an inhomogeneous Poisson case, where the intensity process λ_t is governed by the following specific SDE:

$$d\lambda_t = -\beta(\lambda_t - \underline{\lambda})dt + \eta dN_t, \tag{4.17}$$

where λ_t follows an Ornstein-Uhlenbeck process with jump, with *decay rate* $\beta > 0$, *jump size* $\eta > 0$ and *base rate* $\underline{\lambda} > 0$. As such, the counting process N_t follows a Hawkes process with exponential kernel. The trade execution depends on not only jumps in market order arrival frequency but also the events $\mathcal{L}$ and $\mathcal{H}$.

Analogously, classical results imply that a dynamic programming principle holds and the value function V (4.5) is the solution of a HJB equation:

$$\begin{aligned} 0 = &\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2\frac{\partial^2 V}{\partial s^2} - \beta(\lambda - \underline{\lambda})\frac{\partial V}{\partial \lambda} + \lambda(1-p_l)\left[V(t,x,s,q,\lambda+\eta) - V\right] \\ &+ \sup_{(L,H)\in A_t}\left\{\lambda p_l\Big[p_h V(t, x+(L+H)(s-kL-kH), s, q-L-H, \lambda+\eta)\right. \\ &\left.+ (1-p_h)V(t, x+L(s-kL), s, q-L, \lambda+\eta) - V\Big] - bL\frac{\partial V}{\partial s}\right\}, \end{aligned} \tag{4.18}$$

with terminal condition $V(T,x,s,q,\lambda) = x + qs - \alpha q^2$. Note that λ is a state variable instead of a model parameter as in (4.7). Moreover, the decay of λ is taken into account

[6]Order size here is regarded as the ratio of optimal order size to the remaining liquidation target.

by the partial derivative, and the self-exciting property of the process is represented by the jump in intensity in the supremum operator. Similar to the Poisson case discussed in Proposition 4.3.1, we can find the feedback control forms for the optimal trade sizes with boundedness assumption on the size of the value function. This leads to the **Limit-hidden-Hawkes** (LHH) strategy, in which the agent continuously estimates the market order arrival intensity and decides appropriate size of limit and hidden orders.

Proposition 4.3.7 (Limit-hidden-Hawkes Strategy)**.** *The value function V admits the excess book value decomposition $V(t,x,s,q,\lambda) = x + qs + v(t,q,\lambda)$. The excess book value function v is purely quadratic in q, admitting this decomposition $v(t,q,\lambda) = q^2 v_2(t,\lambda)$ into the reduced-form value function v_2, governed by the following PDE:*

$$\begin{aligned} 0 = {} & \partial_t v_2 - \beta(\lambda - \underline{\lambda})\partial_\lambda v_2 + \lambda(v_2(t,\lambda+\eta) - v_2(t,\lambda)) \\ & + \frac{\lambda p_l}{k - v_2(t,\lambda+\eta)} \Bigg\{ \left[\left(v_2(t,\lambda+\eta) + \frac{b}{2\lambda p_l} \right)^2 + \frac{b^2 p_h}{4\lambda^2 p_l^2 (1-p_h)} \right] \mathbb{1}_{\{v_2(t,\lambda+\eta) < -\frac{b}{2\lambda p_l(1-p_h)}\}} \\ & + p_h v_2^2(t,\lambda+\eta) \mathbb{1}_{\{v_2(t,\lambda+\eta) \geq -\frac{b}{2\lambda p_l(1-p_h)}\}} \Bigg\}, \end{aligned} \tag{4.19}$$

with terminal condition: $v_2(T,\lambda) = -\alpha$.

Assuming that v_2 is bounded by $-\alpha$ and 0, the feedback controls of the optimal liquidation strategy are:

$$\begin{aligned} L^*(t,q,\lambda) &= q \cdot \max\left(1 - \frac{k + \frac{b}{2\lambda p_l(1-p_h)}}{k - v_2(t,\lambda+\eta)}, 0 \right), \\ H^*(t,q,\lambda) &= q \cdot \left(1 - \frac{k}{k - v_2(t,\lambda+\eta)} \right) - L^*(t,q,\lambda). \end{aligned} \tag{4.20}$$

We defer the proof to Appendix 4.7.1.4.

Remark 4.3.8. *v_2 is called the reduced-form value function as it is the expected liquidation cost at time t with $\lambda_t = \lambda$, due to exposure risk, immediate execution cost, and terminal price impact of market orders, given that the agent always liquidates optimally. The boundedness assumption made about v_2 will be shown in Proposition 4.3.9.*

Indicator functions inside the PDE governing v_2 (4.19) correspond to the two different cases of optimal liquidation strategy, respectively. The reduced-form value function has different rates of time decay when liquidating with mixed orders or pure hidden orders given that negative limit order size is not allowed.

In the general case, the optimal liquidation strategy depends on a ***switching curve*** *in the two-dimensional space of time and intensity, where $v_2(t, \lambda + \eta) = -\dfrac{b}{2\lambda p_l(1-p_h)}$. We should note that the sensitivity of optimal controls depends heavily on the strategy being purely hidden or mixed:*

1. *When $v_2(t, \lambda+\eta) < -\dfrac{b}{2\lambda p_l(1-p_h)}$, the agent shall optimally use a mixture of limit and hidden orders with the following allocation:*

$$\begin{aligned} L^*(t,q,\lambda) &= q \cdot \left(1 - \frac{k + \frac{b}{2\lambda p_l(1-p_h)}}{k - v_2(t,\lambda+\eta)}\right), \\ H^*(t,q,\lambda) &= q \cdot \frac{b}{2\lambda p_l(1-p_h)(k - v_2(t,\lambda+\eta))}, \end{aligned} \tag{4.21}$$

   where the optimal hidden order size H^* is increasing in $v_2(t, \lambda+\eta)$ and the optimal limit order size L^* is decreasing in $v_2(t, \lambda + \eta)$. At a fixed time t, higher intensity causes larger H^* and smaller L^*.

2. *When $v_2(t, \lambda+\eta) \geq -\dfrac{b}{2\lambda p_l(1-p_h)}$, the agent shall optimally employ a pure-hidden-order strategy with the following allocation:*

$$H^*(t,q,\lambda) = q \cdot \left(1 - \frac{k}{k - v_2(t,\lambda+\eta)}\right), \tag{4.22}$$

   where the optimal hidden order size H^* is decreasing in $v_2(t, \lambda+\eta)$. At a fixed time t, higher intensity causes smaller H^*.

There is no knowledge about the boundary value condition of v_2 on λ. In theory, λ can be extremely large and bounded below by $\underline{\lambda}$. This motivates the derivation of certain basic analytical properties about the reduced-form value function v_2. In the constant intensity case studied in Theorem 4.3.2, these properties are immediately attainable from observation of the analytical solution. However, there is no analytical solution in the

general case. We thus use the definition of the original value function V (4.5) to discuss these properties:

Proposition 4.3.9 (Properties of v_2)**.** *The reduced-form value function has the following properties:*

1. ***Monotonicity in*** λ*: For any fixed* t*,* $v_2(t, \lambda)$ *is increasing in* λ*.*

2. ***Boundedness****: For any values of* t *and* λ*,* $-\alpha \leq v_2(t, \lambda) \leq 0$*.*

We defer the proof to Appendix 4.7.2.

The boundedness property indicates that the average cost of liquidation is always positive. Moreover, it never exceeds the cost of selling everything using one market order at terminal time. The increasing property of v_2 in λ shows that as long as market orders arrive at a higher frequency, liquidation cost will be reduced given the agent can split the inventory into smaller-sized orders, provided the uncertainty in market order arrival intensity v_2 does not have the time decay property.

(4.19) is difficult to solve analytically due to its nonlinearity and lack of boundary conditions. We use numerical algorithms to solve for the value of v_2, by applying the finite difference method on a two-dimensional grid of intensity and time. On the upper and lower boundaries for intensity, we assume for a Neumann boundary condition that the partial derivative of v_2 with respect to λ is zero.

- On the upper boundary $\lambda \to \infty$, with economic intuition that there is no execution cost for infinitely fast market order arrivals, the limit of v_2 is zero and that v_2 is continuously differentiable with respect to λ, combining Properties 1 and 3 of Proposition 4.3.9. With this information, the partial derivative of v_2 with respect to λ is zero by contradiction: for any t, suppose that $\lim_{\lambda\to\infty} \partial_\lambda v_2(t, \lambda) = l > 0$. By definition of limit, for any $l > \varepsilon > 0$, there exists some M such that $\forall\, \lambda' > M, |\partial_\lambda v_2(t, \lambda') - l| < \varepsilon$, i.e. $\partial_\lambda v_2(t, \lambda') > l - \varepsilon$. Hence, for $M' = \frac{\alpha}{l-\varepsilon}$, we have

$$v_2(t, M + M') = v_2(t, M) + \int_M^{M+M'} \partial_\lambda v_2(t, \lambda')d\lambda' \geq -\alpha + M'(l - \varepsilon) > 0.$$

This strict positivity of $v_2(t, M+M')$ contradicts the boundedness of v_2. Hence the assumption is wrong, and we have $\lim_{\lambda\to\infty} \partial_\lambda v_2(t,\lambda) = 0$.

- On the lower boundary $\lambda = \underline{\lambda}$, the value of $\partial_\lambda v_2$ does not contribute to the dynamics of v_2 as its multiplier in the PDE (4.19) is zero.

Definition 4.3.10. *We define the relative minimal liquidation cost as* $c^v_\alpha(t,\lambda) \coloneqq -v_2(t,\lambda)/\alpha, \quad t \in [0,T]$.

Similar to the Poisson case, we define a minimal liquidation cost function $c^v_\alpha(t,\lambda)$ here, which is also bounded between 0 and 1. Figure 4.6 displays a heat map of $c^v_\alpha(t,\lambda)$ on a two-dimensional grid of intensity λ_t and time t to demonstrate how the liquidation cost changes with these two variables.

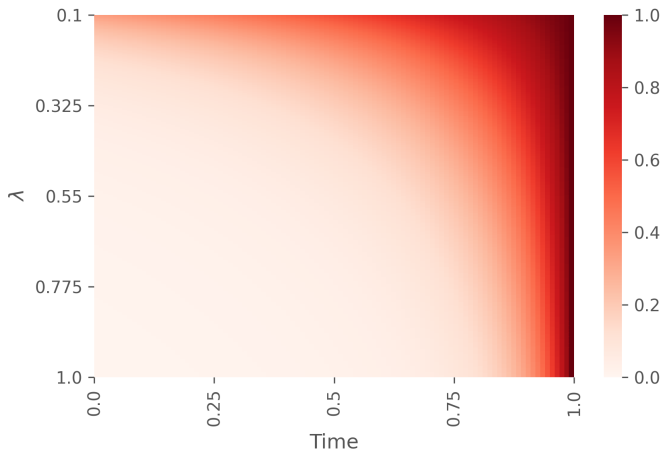

Figure 4.6: Heat map of reduced-form value function $c^v_\alpha(t,\lambda)$ with respect to time t (x-axis) and intensity λ (y-axis) obtained by finite difference method on (4.19).

The heatmap in Figure 4.6 shows that the liquidation cost increases in time and decreases in intensity. The cost reaches its peak at termination (i.e.,$c^v_\alpha(T,\lambda) = 1$). It is worth highlighting that λ_t is one of the state processes of the agent, while in the homogeneous Poisson case $\lambda_t = \lambda$ is a fixed model parameter. In other words, the agent needs to actively monitor the frequency of market order arrivals and determine her order sizes accordingly. As the order sizes are always proportional to the inventory of the agent, we report the order sizes by normalizing the current inventory to one, i.e., $L^*(t,1,\lambda)$ and $H^*(t,1,\lambda)$.

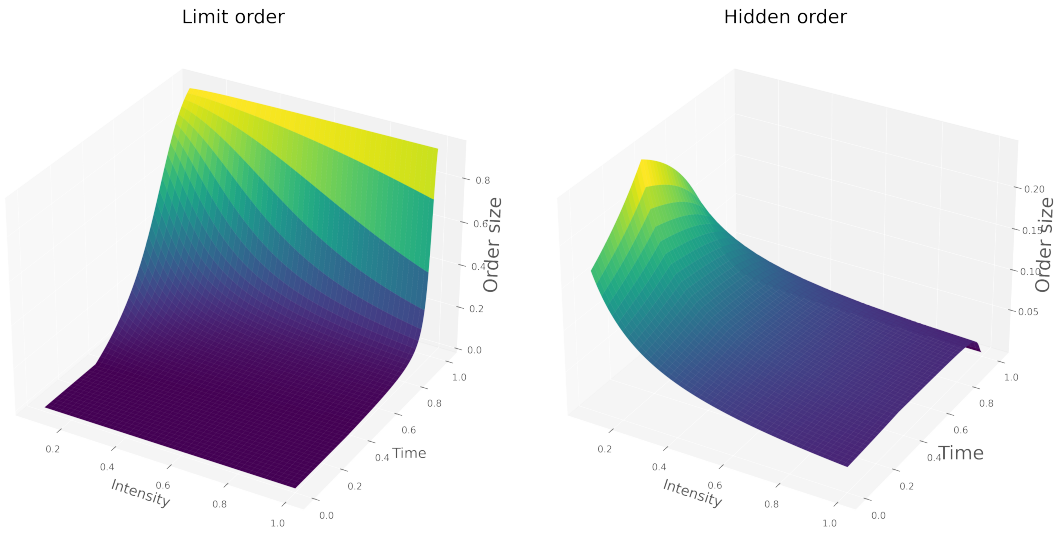

Figure 4.7: Surface plots of limit order size L_t and hidden order size H_t of Limit-Hidden-Hawkes strategy with respect to time t and intensity λ obtained by finite difference method on (4.19).

We also present the surface plots of optimal limit or hidden order sizes L^*, H^* in Figure 4.7. It can be observed that the sensitivity of the optimal controls differs remarkably in the mixed-orders phase and the pure-hidden-order phase separated by the *switching curve*. Despite the lack of an analytical solution for the Hawkes case, we can numerically display the switching curve, see Figure 4.8. It shows that the switching time decreases in the market order arrival intensity as the agent's exposure time is longer for low market order intensity. When the market is relatively quiet, the priority is to prevent exposure of trading intention, and dedicate to a pure-hidden-order strategy. When the market becomes active, the order book imbalance imposed by the agent's ask limit order is neutralized by the incoming market order at a faster rate, making limit orders more desirable.

The switching time can be considered a special case of the switching curve. It directly corresponds to the scenario when $\eta = 0$ and $\beta = 0$, i.e., one cross-section of the two-dimensional space of time and intensity. In that case, there is only one switching time point that divides the optimal decision between mixed-orders phase and pure-hidden-order phase. It is also consistent with economic intuition that in most situations, the switching time is closer to terminal time T for larger intensity, when the agent can

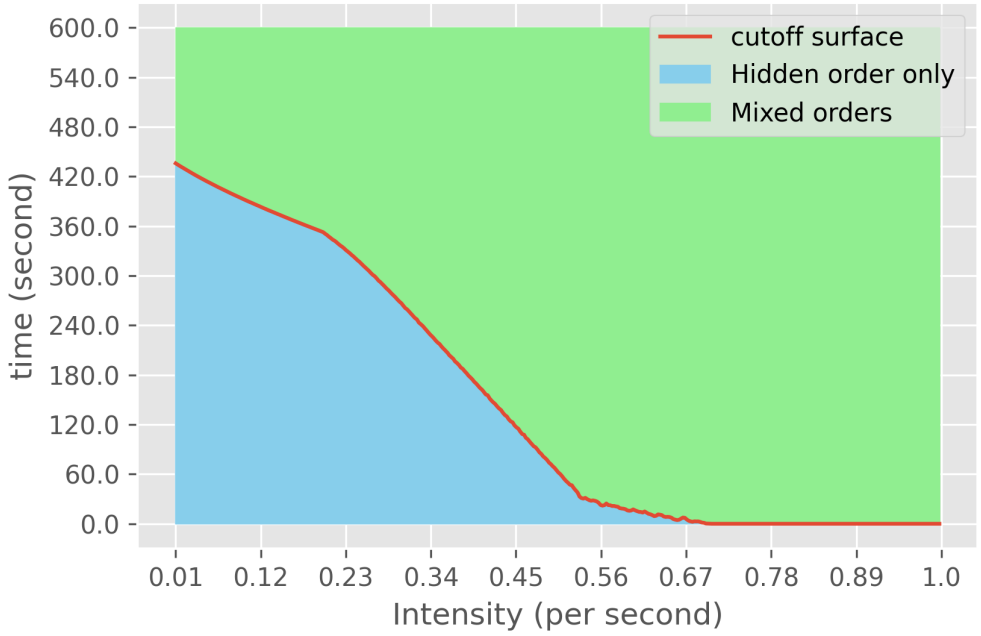

Figure 4.8: Switching curve of the Limit-Hidden-Hawkes strategy in the two dimensional space of time and intensity obtained by finite difference method on (4.19).

liquidate better, consistent with the monotonicity of v_2 in λ.

4.4 Simulation Study

In this section, we demonstrate the optimal solutions of the three trading strategies, i.e., **Limit-Hidden** (LH), **Limit-Only** (LO), and **Limit-Hidden-Hawkes** (LHH), in a simulated environment. We consider two scenarios, where market order arrivals follow either a homogeneous Poisson process or a Hawkes process. Under the known data generating process, we compute the execution costs by implementing different strategies and compare the economic values. Moreover, we conduct a robustness analysis, where we focus on the impact of 1) market arrival intensity and 2) misspecification of parameters on the numerical performance of the strategies.

We break the total execution costs into three components, elaborating respective cost origins. Specifically, the exposure execution cost, denoted as C_b, is induced by exposure risk b; the immediate execution cost C_k is associated with large order execution and

determined by the parameter k; and the terminal cost C_α happens when liquidating the remained shares with market price that is driven by the price impact of market orders α. [7]:

$$\begin{aligned}
\text{Exposure Cost: } C_b &\coloneqq \mathbb{E}\left[\int_0^T bL_tQ_tdt\right], \\
\text{Immediate Cost: } C_k &\coloneqq \mathbb{E}\left[\int_0^T (\hat{S}_t - S_t)dQ_t\right], \\
\text{Terminal Cost: } C_\alpha &\coloneqq \mathbb{E}\left[\alpha Q_T^2\right],
\end{aligned}$$

where the expectation is taken with respect to the randomness in market order arrivals and fill events of limit and hidden orders, with the initial condition $\mathcal{I} = (t_0, Q_0, \lambda_0)$. The agent's value function is the difference between the initial book value Q_0S_0 and the total cost $C_b + C_k + C_\alpha$, regardless the market order arrival dynamics.

Our agent is designated to liquidate her position of one underlying with $Q_0 = 2000$ shares over a fixed period determined by $T = 600$ s. The trading environment is generated with the following parameter values, if not stated otherwise.

$$\begin{aligned}
&\alpha = 2.68 \times 10^{-4}, b = 4.64 \times 10^{-7}s^{-1}, k = 2.31 \times 10^{-5}, p_l = 0.55, p_h = 0.5, \\
&\lambda = 5.76 \times 10^{-2}s^{-1}, \underline{\lambda} = 3.78 \times 10^{-2}s^{-1}, \beta = 2.70 \times 10^{-1}s^{-1}, \eta = 9.26 \times 10^{-2}s^{-1}.
\end{aligned}$$

which are the average values of parameters estimated based on NASDAQ 100 stocks in the later real data analysis[8], except the fill probability of hidden orders p_h. In practice, users can specify the value of p_h according to their own trading record. In the simulation, we take $p_h = 0.5$ as illustration. We generate two scenarios with different market dynamics. In the first scenario, market order arrivals follow the homogeneous Poisson process, where the constant intensity $\lambda = 5.76 \times 10^{-2}s^{-1}$ is used. That is, on average there are 0.0576 market order arrivals per second. In the second scenario, market order arrivals follow the

[7] With a light abuse of notation, we include λ_0 in the state space for unified denotation for C_b, C_k, C_α, knowing that λ_t has no meaning to LH and LO strategy. Moreover, the cost due to volatility of fundamental price σ is zero as an expectation over a Brownian term.

[8] λ is set to $\underline{\lambda}\beta/(\beta - \eta)$ throughout the simulation analysis

Hawkes process with time dependent intensity λ_t driven by $\underline{\lambda} = 3.78 \times 10^{-2} s^{-1}, \beta = 2.70 \times 10^{-1} s^{-1}, \eta = 9.26 \times 10^{-2} s^{-1}$ over time. In the simulation, we use $\lambda_0 = \lambda = 5.76 \times 10^{-2} s^{-1}$ as initial value for the Hawkes process. For a fair comparison in the second scenario, the intensity of the Poisson process should take the same value as the long-term expected intensity of the Hawkes process, where we characterize the long-term mean arrival rates as follows:

Definition 4.4.1. *The long-term expected arrival rate of a Poisson process with intensity* λ *is* $\mathbb{E}[\lambda_\infty] \coloneqq \lim_{t\to\infty} \mathbb{E}[\lambda_t]$.

We align the long-term mean arrival rate between the two processes through the following proposition:

Proposition 4.4.2 (Long-term mean arrival rate of Hawkes process N_t)**.** *For a stationary Hawkes process with exponential kernel* $\phi(t) = \eta \exp(-\beta t)$ *with* $\beta > \eta$*, the long-term expected arrival rate is* $\mathbb{E}[\lambda_\infty] = \lim_{t\to\infty} \mathbb{E}[\lambda_t] = \underline{\lambda} \frac{\beta}{\beta - \eta}$.

Each scenario is tested with 5,000 independent runs with all three strategies. For each run, we first generate a series of market order arrivals $(t_n)_{t\in[0,T]}$ over a fixed time interval $[0, T]$. Upon each market order arrival, we sample two realizations (l, h) independently from the uniform distribution, and use $\mathbb{1}_{p_l > l}$ and $\mathbb{1}_{p_l > h}$ to mimic order fill events as depicted in model setup. We keep the liquidation trajectories consisting of the state (Q_{t_n}, X_{t_n}, t_n) and control variable (L_{t_n}, H_{t_n}) including terminal conditions. The cost associated with each simulation run could be easily derived from the liquidation trajectory. Finally, we compute and compare the average liquidation cost for three strategies LH, LO and LHH for each scenario.

4.4.1 Numerical Results

In this subsection, we compare the numeric performance of three strategies under two settings, namely when market order arrival follows the Poisson and Hawkes process, respectively. In the simulation steps, we assume the agent knows the true parameters for k, α, b, k, p_l, p_h, λ, $\underline{\lambda}$, η and β.

When the generated market arrivals follow the homogeneous Poisson process with $\lambda = 5.76 \times 10^{-2} s^{-1}$, we let the LO and LH take the true value of λ, and the LHH obtains the unbiased estimate of $\underline{\lambda}, \eta$, and β. Specifically, the corresponding parameters of the Hawkes process are $\underline{\lambda} = 5.47 \times 10^{-2} s^{-1}$, $\eta = 7.2 \times 10^{-5} s^{-1}, \beta = 1 \times 10^{-3} s^{-1}$, which are estimated from the data generated with Poisson process for a fair comparison. For small β and η, the estimated Hawkes process behaves very similarly to a Poisson process. In that case, it is expected that the two strategies will behave similarly. However, when LH is calibrated with true data generating process, LO is expected to perform badly as hidden order is not allowed.

When the generated market arrivals follow the Hawkes process, we let the LHH strategy know the true value for $\underline{\lambda}$, η and β. The LO and LH strategies take the average market order arrival intensity as lambda, which is the same as the first scenario.

Table 4.1: Simulation results where market orders are generated with the Poisson process with intensity $\lambda = 5.76 \times 10^{-2} s^{-1}$ (left panel) and inhomogeneous (Hawkes) with exponential kernel $\underline{\lambda} = 3.78 \times 10^{-2} s^{-1}, \beta = 2.70 \times 10^{-1} s^{-1}, \eta = 9.26 \times 10^{-2} s^{-1}$ (right panel), where $\lambda_0 = \lambda = 5.76 \times 10^{-2} s^{-1}$ for both cases. The best performing result is highlighted in bold.

	Poisson			Hawkes		
Metric	LHH	LH	LO	LHH	LH	LO
Q_T/Q_0 (%)	**0.66**	0.87	0.71	**2.15**	4.04	3.59
Exposure cost	**1.31**	1.43	31.23	**2.99**	5.32	45.39
Immediate cost	11.93	11.53	**5.87**	16.89	13.97	**7.89**
Terminal cost	0.54	0.71	**0.49**	**3.87**	8.77	6.19
Total cost	13.78	**13.67**	37.59	**23.76**	28.06	59.46
Average limit order size	12.78	13.70	104.85	20.85	31.75	146.02
Average hidden order size	184.69	182.62	--	254.89	228.40	--
Average total order size	197.46	196.33	104.85	275.74	260.15	146.02

Table 4.1 reports the average execution costs over $5,000$ runs as well as the breakdowns of each strategy under the two scenarios using the estimated parameters.

As expected, the LH strategy delivers the best performance when the market order arrivals follow the homogenous Poisson process, with the lowest total cost of 13.67. The LHH achieves 13.78 and is on par with the LH, and further study shows that LO has the smallest immediate cost C_k, in exchange for a much larger exposure cost C_b for only using

limit orders. The use of hidden orders in the mixed strategies LH and LHH dramatically reduces exposure cost from 31.23 to 1.43 and 1.31, respectively. There are marginal differences in the terminal cost among the three strategies, as the ratios, Q_T/Q_0 indicate only a few remaining shares are liquidated by market orders at termination. Moreover, in general, the implementation of mixed order strategies enhances liquidity in the market, with a smaller limit order size (12.78/13.70/104.85 for LHH/LH/LO) but a significant increase in hidden order size (197.46/196.23 for LHH/LH) pushing up the total order size.

It is worth mentioning that the strategy LHH delivers a similar performance as the optimal strategy, i.e., LH in the scenario of Poisson. Although LHH is based on a mis-specified Hawkes process assumption, it can reasonably describe the stochastic behavior of market arrivals with three parameters. It indicates LHH is robust and applicable for both homogeneous Poisson and Hawkes processes.

When market arrivals are self-exciting, liquidation becomes more expensive in general, and more shares need to be liquidated with market orders at termination, regardless of strategy. Given the uncertainty in arrival intensity, there is a significant increase in the use of limit orders. Nevertheless, the mixed order strategies LH and LHH again display superior performance, where total cost drops from 59.46 to 28.06 for LH and 23.76 for LHH. Among the mixed order strategies, LHH delivers outstanding results, where exposure cost reduces by more than 15-folds compared with LO, and by 43.6% against LH, and terminal cost drops by more than 55.9% compared to LH and by 37.5% against LO. The saving in exposure cost is directly caused by the smaller average limit order size, whereas the lower terminal cost is attributed to timely monitoring of market activity, i.e., considering the time-dependent intensity. By utilizing the dynamic variation of market conditions, LHH seems to ensure a steady inventory trajectory and alleviating the liquidation need of remaining shares with market orders.

4.4.2 Robustness analysis: market intensity

We study the robustness of the liquidation strategies in different market conditions, where the intensity of market order arrivals varies. Specifically, we use Hawkes process as the true data generating process of market order arrivals for its flexibility and implement the three strategies to compute liquidation cost. To facilitate a meaningful comparison, the parameters values λ in the Poisson based strategies LH and LO are assigned to be the long-term expected market order intensity of Hawkes process, i.e., $\underline{\lambda}\beta/(\beta-\eta)$. Moreover, the long-term expected order intensity remains unchanged across the robustness study for consistent comparisons.

We consider four market scenarios by altering the values of i) the initial intensity λ_0 at $t = 0$, ii) the decay rate β and iii) the jump size η in the Hawkes process while keeping the other parameters as the default values. In particular, we design a less active market where the initial intensity λ_0 is scaled down by 10 times and a more active market with the initial intensity is scaled up by 10 times. Simultaneously, the market condition can be influenced by the decay speed and jumps. In the other two scenarios, we mimic a market with a more stable intensity, where β and η are scaled down by a factor of 2, representing a slower decay and lower jump, as well as a market with more turbulence, where β and η are both scaled up by a factor of 2, indicating faster decay and more jumps.

Table 4.2 reports the performance of the three strategies, averaging over $5,000$ independent runs, under the above described specifications.

In general, the relative performance of the three strategies remain the same as in the default case. LHH delivers the lowest execution cost, where the main contribution is from the savings in exposure cost with hidden orders, followed by the cost efficiency at termination with less need for market order liquidation at termination. However, the absolute amount of execution cost monotonically decreases in the initial intensity λ_0, where the total cost drops from 24.30/28.86/60.75 (LHH/LH/LO) to 20.68/25.35/50.72 when the agent moves from a less active market to a more active market at time 0. This suggests that agents should adjust trading behaviors, e.g., to start liquidation at points when market is active for cost efficiency and fast liquidation.

Table 4.2: Simulation results under inhomogeneous (Hawkes process with exponential kernel) market order intensity, with various parameter values. Default values are $\underline{\lambda} = 3.78 \times 10^{-2} s^{-1}, \beta = 2.70 \times 10^{-1} s^{-1}, \eta = 9.26 \times 10^{-2} s^{-1}$, and $\lambda_0 = \lambda = 5.76 \times 10^{-2} s^{-1}$. The best performing result is highlighted in bold. $A \leftarrow cA$ means that parameter A is scaled by a factor of c over its default value.

Metric	LHH	LH	LO	LHH	LH	LO
	$\lambda_0 \leftarrow 0.1\lambda_0$			$\lambda_0 \leftarrow 10\lambda_0$		
Q_T/Q_0 (%)	**2.25**	4.20	3.66	**1.85**	3.73	3.24
Exposure cost	**3.14**	5.51	46.73	**3.50**	4.56	38.61
Immediate cost	17.17	14.17	**8.07**	14.15	12.87	**7.20**
Terminal cost	**4.00**	9.18	5.96	**3.04**	7.92	4.91
Total cost	**24.30**	28.86	60.75	**20.68**	25.35	50.72
Average limit order size	21.45	32.48	148.56	19.87	29.10	134.59
Average hidden order size	260.21	231.81	--	210.03	210.16	--
Average total order size	281.66	264.30	148.56	229.89	239.26	134.59
	$\beta \leftarrow 0.5\beta$ and $\eta \leftarrow 0.5\eta$			$\beta \leftarrow 2\beta$ and $\eta \leftarrow 2\eta$		
Q_T/Q_0 (%)	**2.69**	3.64	3.19	**2.46**	4.40	3.86
Exposure cost	**2.69**	4.76	43.71	**3.40**	5.88	47.55
Immediate cost	15.55	13.66	**7.66**	17.07	14.30	**8.24**
Terminal cost	**4.93**	7.45	5.23	**4.55**	9.80	6.54
Total cost	**23.17**	25.88	56.60	**25.02**	29.98	62.34
Average limit order size	19.03	29.30	140.70	22.84	34.08	151.89
Average hidden order size	245.22	222.73	0.00	261.48	233.93	0.00
Average total order size	264.25	252.03	140.70	284.32	268.01	151.89

Among the self-exciting processes with the same long-term average intensity but different clustering effects, it shows that arrivals with strong clustering with larger jumps (larger η) and faster decay (larger β) lead to more expensive liquidation 25.02/29.98/62.34 (LHH/LH/LO). Meanwhile, when β and η are scaled down, the execution cost is significantly reduced to 23.17/25.88/56.60 (LHH/LH/LO) although both cases have the same long-term average intensity. The average size of both limit and hidden orders increases with respect to the level of clustering. This indicates that a highly clustering/dispersing market presents higher randomness in the intensity, which makes it more difficult to liquidate shares using limit and hidden orders.

4.4.3 Robustness analysis: parameter misspecification

In addition to robustness in market order dynamics, it is also crucial to investigate the numerical performance of the strategies under parameter misspecification due to either

misunderstanding of market state or estimation errors. In this section, we purposely perturb the value of parameters estimated with the simulated data, with ±50% variation from its ground truth. In particular, we use the under-/over-estimated values for five parameters, i.e., immediate execution cost k, exposure risk b, price impact α, and the two fill probabilities p_l and p_h, while keeping the remaining parameters as default.

Table 4.3: Simulation results under inhomogeneous (Hawkes with exponential kernel) market order intensity, with estimation bias. The default values for the parameters are $\alpha = 2.68 \times 10^{-4}, b = 4.64 \times 10^{-7} s^{-1}, k = 2.31 \times 10^{-5}, p_l = 0.55, p_h = 0.5$. The estimation bias is introduced to one parameter at a time. The parameters for market order generating process are $\underline{\lambda} = 3.78 \times 10^{-2} s^{-1}, \beta = 2.70 \times 10^{-1} s^{-1}, \eta = 9.26 \times 10^{-2} s^{-1}$, and $\lambda_0 = \lambda = 5.76 \times 10^{-2} s^{-1}$. The best performing result is highlighted in bold.

		Underestimation (50%)			Overestimation (+50%)		
Para	Metric	LHH	LH	LO	LHH	LH	LO
	Q_T/Q_0 (%)	**1.25**	2.69	2.64	**2.94**	5.09	4.42
	Exposure cost	**1.66**	3.30	45.62	**4.35**	7.20	45.21
k	Immediate cost	20.15	16.15	**8.24**	15.09	12.64	**7.63**
	Terminal cost	**2.36**	6.06	4.56	**5.18**	10.84	7.66
	Total cost	**24.17**	25.51	58.43	**24.61**	30.68	60.50
	Q_T/Q_0 (%)	**1.95**	3.77	3.27	**2.25**	4.17	3.90
	Exposure cost	**6.74**	10.01	45.47	**1.82**	3.64	45.32
b	Immediate cost	14.77	12.40	**8.00**	17.99	14.74	**7.79**
	Terminal cost	**3.27**	7.71	5.64	**4.15**	9.27	6.73
	Total cost	**24.77**	30.12	59.10	**23.96**	27.65	59.84
	Q_T/Q_0 (%)	**4.02**	6.43	5.72	**1.49**	3.08	2.74
	Exposure cost	**2.31**	4.51	44.95	**3.35**	5.71	45.59
α	Immediate cost	15.65	13.05	**7.28**	17.57	14.44	**8.20**
	Terminal cost	**7.54**	14.44	10.13	**2.69**	6.69	4.74
	Total cost	**25.50**	32.00	62.36	**23.61**	26.84	58.54
	Q_T/Q_0 (%)	**0.39**	0.66	0.58	**5.30**	9.40	8.52
	Exposure cost	**1.27**	2.30	46.71	**5.54**	8.56	44.99
p_l	Immediate cost	30.71	23.10	**10.74**	13.49	12.15	**7.92**
	Terminal cost	0.52	0.91	**0.52**	**12.58**	27.56	21.45
	Total cost	32.50	**26.30**	57.97	**31.61**	48.27	74.37
	Q_T/Q_0 (%)	**1.15**	2.15	3.59	4.36	7.59	**3.59**
	Exposure cost	**4.28**	6.63	45.39	**1.94**	4.22	45.39
p_h	Immediate cost	18.73	14.89	**7.89**	15.86	14.17	**7.89**
	Terminal cost	**1.53**	3.26	6.19	10.21	21.92	**6.19**
	Total cost	**24.55**	24.78	59.46	**28.01**	40.31	59.46

Table 4.3 presents the average performance of the three strategies LHH, LH, and LO with the misspecified parameters. In general, the influence of parameter misspecification is mainly on the corresponding cost breakdowns, e.g., k for immediate cost, b for exposure cost, and α for terminal cost. However, the change is marginal in most cases, indicating

that the strategies are robust to parameter misspecification, at least with up to 50% estimation errors. For the case of p_h, there is no influence on the strategy LO, where only limit orders are allowed. Moreover, in both under- and over-estimation scenarios, LHH remains to be the best performing strategy in terms of total cost across the board, except once when LH beats LHH because p_l is under-estimated by 50%, due to much larger immediate execution cost.

4.5 Real data analysis

Significance of hidden liquidity and hidden orders in the financial markets promotes empirical studies of our model: The percentage of U.S. stock trades executed on dark pools and other off-exchange trading venues can be as high as 38.6% in April 2019 (Osipovich, 2019). Hidden orders have obtained significant volume share in a large number of modern order-driven securities markets globally, with empirical evidence on a significant increase of the overall market depth in NASDAQ with the introduction of hidden orders (Tuttle, 2003) and that the percentage of volume traded against hidden orders is more than 20% in October 2010 (Hautsch & Huang, 2012b). Also, studies have shown that hidden orders in the lit pool can serve as substitutes for dark pool trading (Degryse, Karagiannis, Tombeur, & Wuyts, 2021).

In this section, we implement the derived optimal execution strategies to real data. We consider an agent who has to liquidate her position of a single stock with a sizeable number of shares over a fixed period. The agent's interest is to evaluate the cost-benefit of using limit orders only or mixed orders with both limit and hidden as in Section 4.3.1, or mixed orders under self-exciting market order arrivals as in Section 4.3.3.

In the following, we present descriptive statistics of 100 stocks in NASDAQ. We present the estimation methods that have been used in literature to extract the parameters from public data. We detail the estimation model setup and report the numerical performance of the execution strategies. Note that estimation here is for illustration only and there exist other estimation approaches that are appropriate for specific situations.

To check the robustness of the model, we have conducted sensitivity analysis as in the previous section, where we recomputed the execution costs with under- or over-estimated parameters.

4.5.1 Data

We use event-level data of NASDAQ 100 Index data (102 tickers)[9], which are the 100 largest and most actively traded U.S companies listed on the NASDAQ stock exchange. The stocks belong to various industries except for the financial sector. The limit order book data are obtained from LOBSTER from 1 to 15 Feb 2018 (11 trading days). Since each ticker has its own limit order book and is traded separately, our estimation and testing are conducted for each ticker. Table 4.4 presents the summary statistics of the 102 tickers on the ask side of the book during market open. It shows that during the 11 trading days, the average daily trading volume among the 102 tickers is 43,078 shares, yet with a wide range from 7,159 (MELI) to 252,498 (AAPL). The average number of trades per day is 7,180, corresponding to 0.31 trades per second. Again, there is a variation from 1,193 (MELI) to 42,083 (AAPL). The average depth, measured as shares available at the best ask price level of the limit order book, is 803. The average market order arrival rates vary not only from stock to stock, with a range from 0.05 per second (MELI) to 1.8 per second (AAPL), but also over time.

Table 4.4: Summary statistics for NASDAQ 100 Stocks during 1 to 15 Feb 2018.

	Mean	SD	min	P25	P50	P75	max
Average depth at best ask price	803	3,698	79	170	205	386	35,829
Average market order arrival per second	0.31	0.29	0.05	0.14	0.22	0.33	1.80
Average trade price	151	202	6	61	95	166	1414
Average daily trades	7,180	6,679	1,193	3,297	5,031	7,701	42,083
Average daily volume	43,078	40,075	7,159	19,783	30,186	46,207	252,498

[9]There are 102 tickers, with two tickers for Fox Corporation (FOXA, FOX) and Alphabet Inc. (GOOGL and GOOG), respectively.

4.5.2 Implementation and estimation setup

We design a scenario where the agent has to liquidate $Q_0 = 5\%$ [10] of the average daily trading volume of every single stock over $T = 600$ s. Given the average daily trading volume is 43,078 shares, the average liquidation target of the agent is 2,154 shares, for which the individual liquidation volume differs among the 102 tickers. To implement liquidation strategies, i.e., Limit-Hidden-Hawkes (LHH), Limit-Hidden (LH), and Limit-Only (LO), the agent needs to estimate the respective parameters from publicly available data and determine the type and number of shares based on the derived solutions.

The dataset contains both event-level information as well as limit order book price and depth for each event. For each stock, we consider 10:00 AM to 3:30 PM of the trade and quote data, removing the first and last 30 min (9:30 AM to 10:00 AM and 3:30 PM to 4:00 PM) to alleviate microstructural impact in opening and closing. For trade data, we aggregate multiple orders within short intervals, i.e, 100 ms, to avoid capturing the effect of ultra-high frequency market order splitting. The resultant data of 11 days are concatenated to form a single time series. Finally, to utilize as much information from the remaining data as possible, we adopt a trade clock in the real world and fix a set of discrete-time grids $\{t_n\}_{n=1}^{N}$ with t_n being the arrival time of the n-th market order, see (Carmona & Webster, 2019; Cartea et al., 2015; Y. Chen et al., 2018).

Moreover, for a robust evaluation of the liquidation strategies, the agent is asked to liquidate with the same time duration and initial number of shares multiple times with a representative coverage of different days and times in the market. Specifically, they need to liquidate 6 times a day at 10:00-10:10, 11:00-11:10, 12:00-12:10, 13:00-13:10, 14:00-14:10 and 15:00-15:10, respectively, over the 11 consecutive trading days. In the experiments, the actual market order arrivals are used, reflecting the real-life situation, instead of being generated from an agent-specified process. This is the key difference between real data testing and simulation. The remaining testing procedure is similar to that described in the simulation. We conduct 500 independent runs for each of the tickers

[10] We acknowledge that these stocks are traded in multiple exchanges, we only focus on the trading volume captured by the limit order books provided by LOBSTER data. For consistency, summary statistics, estimation and testing are based on the same limit order book data

at each testing session [11].

4.5.3 Model Parameter Estimation

The estimation procedure is conducted for the following model parameters: price impact of market order α, exposure risk b, immediate execution cost k, fill probability of limit orders p_l, together with the Hawkes parameters associated with self-excitement: $\underline{\lambda}$, η and β. For each stock, we concatenate the event-level data over the 11 trading days (10:00 AM to 3:30 PM for each day) for estimation.

The depth of hidden orders is not available in the dataset. In practice, agents can estimate p_h with the actual fill probability of her posted hidden orders from proprietary data. In the experiment, we specify $p_h = 0.5$ to represent the fill probability of hidden orders, the same as the simulation section, and calculate execution cost for each, respectively. For robustness, we also perform a set of tests with $p_h = 0.2$, representing a situation with very deep hidden queues.

Price Impact of Market Orders: α A large market order will consume liquidity on the bid side of the book and thus temporarily push the best bid price down. We construct an order book by taking an average of all the snapshots and execute a hypothetical market order by walking the LOB. For each hypothetical market order, we calculate the price change due to this order with respect to the order book data. Following Cartea and Jaimungal (2015a) and Y. Chen et al. (2018), the price change is assumed to be linearly proportional to the number of executed (hypothetical) shares. In other words, the price impact α can be interpreted as the slope of a linear book:

$$S^{\text{mid}}_{t_n} - S^{\text{bid},Q_i}_{t_n} = \alpha Q_i + \varepsilon^{\alpha}_{i,t_n},$$

where $S^{\text{mid}}_{t_n}$ is the mid-price at the trading time t_n of the n-th market order arrival, $\{Q_i\}_{i=1}^{I}$ is the hypothetical order size up to 1% of the traded volume of the same ticker

[11] Total number of testing runs is $66 \times 102 \times 500$, due to much larger testing demand compared to simulation, we conduct 500 independent runs instead of 5,000.

on the same day, $S_{t_n}^{\text{bid},Q_i}$ is the best bid price after a market sell order of size Q_i is placed and $\varepsilon_{i,t_n}^{\alpha}$ is the error term with mean zero and finite variance. In the estimation, we set $I = 20$ and only consider up to 10 levels of the orderbook book due to the availability of data. This effect is statistically significant as the largest p-value of the Student-t test is 3.3×10^{-274}.

Exposure risk: b Exposure risk is influenced by order book imbalance. When limit orders are posted onto the order book, the order flow imbalance will decrease, thus increasing the downward price pressure. Cont et al. (2014) shows that this extra pressure is almost 'constant' from order submission until order execution or cancellation. We build up a linear regression to estimate the exposure risk,

$$S_{t_{n+1}}^{\text{mid}} - S_{t_n}^{\text{mid}} = b(t_{n+1} - t_n)\text{OBI}_{t_n} + \varepsilon_{t_n}^{b},$$

where the temporal price change is measured at mid-price $S_{\cdot}^{\text{mid}}$ between two consecutive observations at trade time t_{n+1} and t_n, the order book imbalance is computed as $\text{OBI}_t \coloneqq q_t^{\text{ask}} - q_t^{\text{bid}}$, with q_t^{bid} and q_t^{ask} denoting the total order size at best bid and ask levels, respectively. The error term $\varepsilon_{t_n}^{b}$ is assumed to have mean zero and finite variance. Statistical significance of b is strong as all the p-values of the Student-t tests are smaller than 0.2%.

Immediate Execution Cost: k Immediate execution cost is associated with execution of large orders, both limit or hidden. R. Almgren and Chriss (2001) show that the immediate execution cost is due to execution speed and in our case, the limit order size. Hautsch and Huang (2011, 2012a) show that market buy orders temporarily increase the best ask price, while large limit orders prevent walking up the book. Instead of selling a large limit order, the agent can benefit more if they place a smaller order and then other orders after price moves in their favor. We estimate the immediate execution cost by a

linear model between execution price change and executed limit order size:

$$\bar{S}_{t_n}^{\text{exec},d_{t_n}} - S_{t_n}^{\text{exec}} = d_{t_n} k Q_{t_n}^{\text{exec}} + \varepsilon_{t_n}^k,$$

where $S_{t_n}^{\text{exec},\cdot}$ is the execution price at trade time t_n, $\bar{S}_{t_n}^{\text{exec},d}$ is the average future execution price of the next M trades on the side with direction d_{t_n} of the n-th trade. The direction d_{t_n} is dummy, taking 1 for buyer-initiated and -1 for seller-initiated. The executed limit order size is denoted as $Q_{t_n}^{\text{exec}}$[12]. In the estimation, we set $M = 20$ such that the immediate execution cost is neither overwhelmed by the future price trends, nor by the price impact of market orders. Again, $\varepsilon_{t_n}^k$ is the error term with mean zero and finite variance. The immediate execution cost effect is in general statistically significant, in which 92 out of 102 tickers have p-values less than 5% and 86 out of 102 tickers have p-values lower than 1%.

Fill Probability of Limit Orders: p_l The fill probability of limit order is positively correlated to the average market order size and negatively correlated to the displayed market depth. We estimate the fill probability of limit orders p_l with the empirical ratio of market buy order size to order book depth:

$$\hat{p}_l = \frac{\bar{Q}^{\text{mb}}}{\bar{q}^{\text{a}}},$$

where $\bar{Q}^{\text{mb}}$ denotes the average market buy order size and $\bar{q}^{\text{a}}$ represents the average displayed depth at best ask price when a buy market order arrives.

4.5.4 Parameters of Market Order Arrival: $\lambda, \underline{\lambda}$, η and β

We use the maximum likelihood method to estimate the counting processes. The log-likelihood function is of the form:

$$\sum_{n=1}^{N(T)} \log \lambda_{t_n} + \int_t^T \lambda_t dt$$

[12]It may contain a portion of hidden orders if executed by the same trade.

The two strategies, LH and LO, assume that market order arrivals follow the Poisson process. We estimate the constant intensity $\lambda = N(T)/T$, by taking the first-order condition of the log-likelihood function, where $N(T)$ is the total number of market orders. The strategy LHH assumes the Hawkes process with an exponential kernel. We adopt a numeric solver to estimate their values $\underline{\lambda}, \eta$ and β.

Table 4.5: Summary statistics of the parameter estimates of the price impact model fitted on NASDAQ 100 stock data.

Parameter	Mean	SD	min	P25	P50	P75	max
α (10^{-4})	2.68	3.94	0.00	0.39	1.30	2.76	19.67
k (10^{-5})	2.31	3.72	0.01	0.26	0.80	1.98	17.13
b $(10^{-6}s^{-1})$	4.64	7.42	0.00	1.32	3.12	5.33	63.24
p_l (%)	55.76	14.48	5.33	48.53	58.36	64.36	74.39
λ $(10^{-2}s^{-1})$	8.51	8.18	1.30	4.13	6.15	8.72	55.76
$\underline{\lambda}$ $(10^{-2}s^{-1})$	3.78	2.23	0.34	2.20	3.46	4.90	14.47
β $(10^{-2}s^{-1})$	27.00	14.78	1.08	13.34	30.54	38.99	63.54
η $(10^{-2}s^{-1})$	9.26	3.72	0.88	7.67	9.97	11.95	16.08

Estimation Results Table 4.5 presents a summary of the estimated values. The details of estimators for each stock can be found in Appendix 4.7.4. We make the following remarks on the estimation results. Firstly, the price impact parameters α, k, b differ for each ticker, where the ratios between the largest and smallest values go beyond the magnitude of 10^3. Since these parameters are closely related to execution cost, we expect the liquidation costs to vary widely among the tickers. Secondly, the fill probability is scattered around 50% which means that on average limit order pegged at the best price will be successfully executed half of the times upon each market order arrival. Thirdly, the decay β and jump size η are significant compared to the average market order arrival rate λ. This means that the clustering/dispersion effect of market orders is commonly seen across all the tickers, suggesting that the market order intensity is not homogenous.

4.5.5 Execution Cost

We now implement the three strategies LHH, LH, and LO, where LH and LO are derived with constant market order arrival intensity and LHH is with self-exciting intensity.

Meanwhile, LO only submits limit orders, while the other two quote both limit and hidden orders.

Table 4.6: Execution cost of different strategies on NASDAQ 100 stock data.

Metric	LHH	LH	LO	LHH	LH	LO
	$p_h = 0.5$			$p_h = 0.2$		
Q_T/Q_0	**0.05**	0.08	0.09	**0.06**	0.09	0.09
Exposure cost	0.76	**0.46**	13.49	4.26	**2.89**	13.49
Immediate cost	4.12	3.30	**1.91**	4.84	4.34	**1.91**
Terminal cost	**4.21**	6.94	5.93	**5.75**	8.24	5.93
Total cost	**9.09**	10.70	21.33	**14.86**	15.47	21.33

Table 4.6 presents the averages execution cost over 102 tickers and 66 testing sessions. It shows the breakdowns of execution cost, including the exposure cost, immediate cost, and terminal cost, summing up to the total cost. The average remaining shares (Q_T/Q_0) are also reported. As an illustration, the fill probability of hidden order is set to $p_h = 0.5$ for the base case and 0.2 for a robustness check. Due to space limitation, details of the average execution cost for each ticker and that for each testing session can be found in Appendix 4.7.4.

In the base case ($p_h = 0.5$), our experiments show that the use of hidden orders substantially reduces the total execution cost, from 21.33 (LO) to 10.70 (LH) with 50% cost savings. We attribute this to the reduction in exposure cost, dropping from 13.49 (LO) to 0.46 (LH), a direct benefit of hidden orders. Meanwhile, the immediate cost is almost doubled, from 1.91 (LO) to 3.30 (LH), when hidden orders are used, but at a relatively smaller magnitude. The terminal cost of LH (6.94) is higher than LO (5.93), due to the lower execution priority of hidden orders. In the robustness setting where the fill probability of hidden order $p_h = 0.2$, the margin of LH (15.47) over LO (21.33) is narrower because hidden orders become even less likely to execute, which directly leads to an increase in terminal cost (from 6.94 to 8.24) and increase in the remained shares ratio Q_T/Q_0. In addition, the agent uses more limit orders, which results in higher exposure costs than before.

LHH achieves a cost reduction of 57% over LO, from 21.33 (LO) to 9.09 (LHH). By taking into consideration the inhomogeneity of intensity, LHH reduces execution cost

over LH with a cost reduction of 15%, from 10.70 (LH) to 9.09 (LHH), mainly due to the reduction in terminal cost. Compared to LH, LHH actively monitors the market order arrivals and adjusts order type/sizes accordingly, especially when the market is inactive. Moreover, LHH achieves the agent's liquidation target with the lowest remained share ratio, i.e., Q_T/Q_0. The same cost reduction pattern applies when $p_h = 0.2$. It is worth noting that the leftover ratio (Q_T/Q_0) in the real-data testing is larger than that in simulation. Given the market order arrivals are obtained from limit order book data, we consider the empirical arrivals are more chaotic than the generated ones with known parameter values.

We present the average trading cost of strategies over 102 tickers for each of the 66 sessions in Figure 4.9. The execution cost of liquidation follows a similar daily pattern over the testing periods (top), higher around 12 pm to 1 pm and lower approaching market opening and close. LH consistently outperforms LO in all the sessions given the reduction in exposure cost (middle). The Hawkes-based strategy, LHH, further improves the Poisson-based strategy LH, especially during lunch hours when the market is relatively inactive. The cost reduction again is attributed to the saving in terminal cost (bottom).

Figure 4.11 displays the comparison of the average total execution cost over 66 testing sessions for each ticker where the stocks are ordered according to the relative performance which, for plotting purposes, is calculated as $(LH-LHH)/LHH$. The radius of the circle of each ticker represents the absolute amount of the cost difference $|LH - LHH|$. We fill the circle in green if $LH - LHH > 0$ and red otherwise. It shows that LHH is a better strategy for 86 tickers (out of 102 tickers), with cost savings in a range of 0.04-27.42 corresponding to 1%-65% reduction.

We study the relationship between the performance of LHH/LH and the actual market order arrival patterns. Specifically, we examine the ratio between the standard deviation and the mean of the empirical inter-arrival (IA) time of market order for each ticker and use this measure as a proxy for the randomness of market order arrivals. As the inter-arrival time of the Poisson process follows the Exponential distribution, this ratio corresponding to the Poisson process is 1. The larger the ratio, the more random the

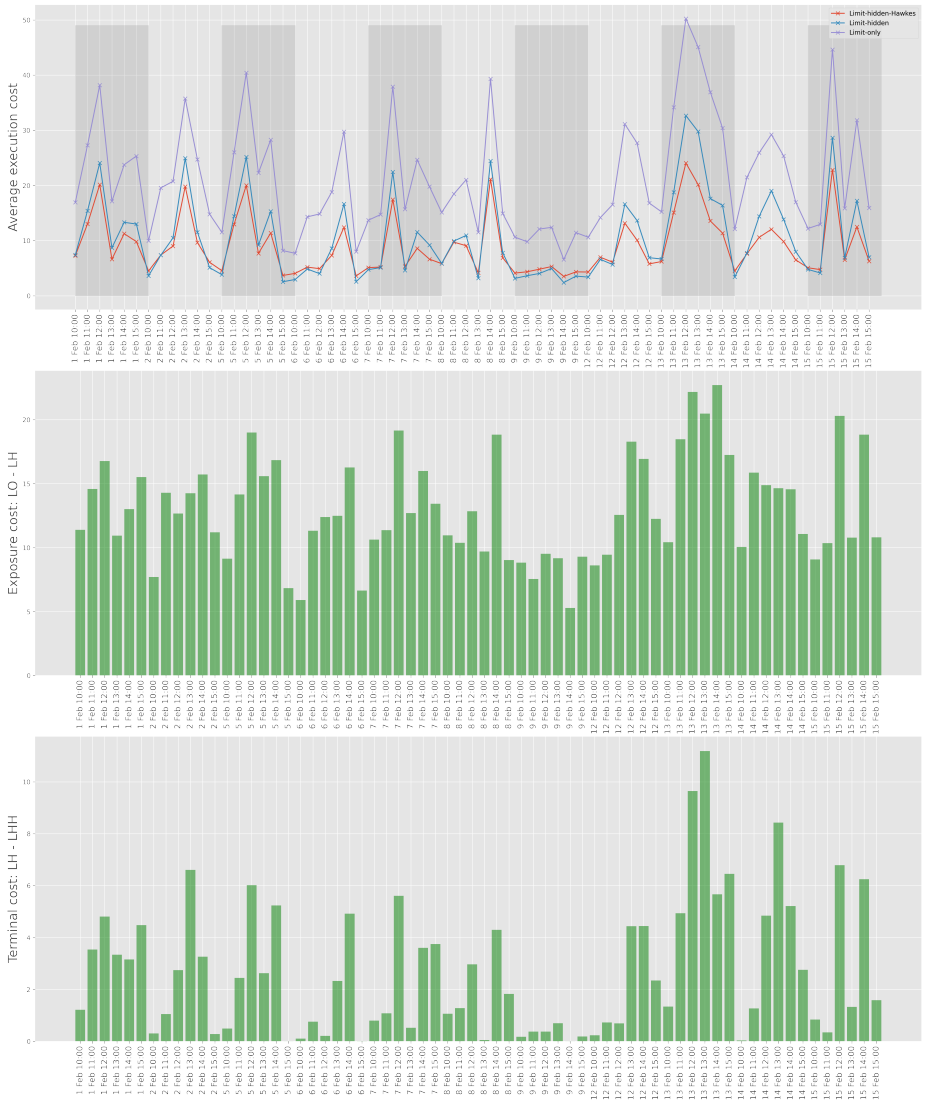

Figure 4.9: Average liquidation cost of NASDAQ 100 stocks for each of the 66 testing sessions. Upper panel: total execution cost of LHH, LH and LO. Middle panel: difference in exposure cost between LO and LH: LO -LH. Lower panel: difference in terminal cost between LH and LHH: LH - LHH.

market order intensity is. We sort the tickers based on the ratio and group them into three baskets. We plot the performance of each basket over the randomness of market order arrivals in Figure 4.10. We see strong correlation between the randomness of market

order arrivals and the margins achieved by LHH (in both relative and absolute sense).

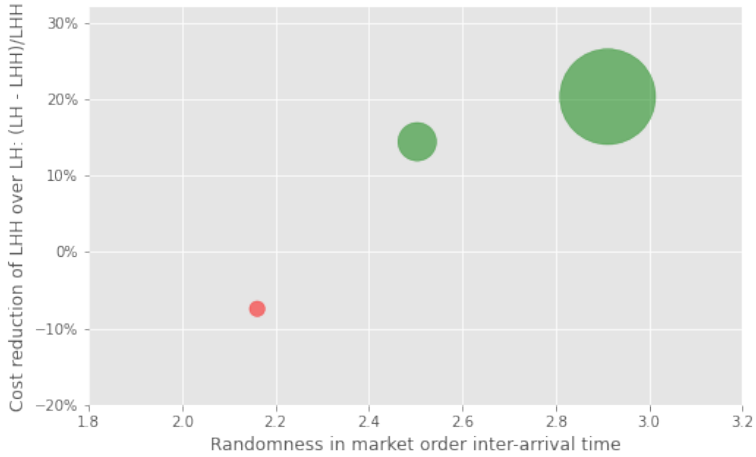

Figure 4.10: Comparison of the Limit-Hidden-Hawkes and Lidden-Hidden strategies over randomness in market order inter-arrival time. The 102 tickers are ranked based on the ratio between SD and mean of market order inter-arrival time, and then grouped into three groups with equal size (i.e., 34). The vertical axis reports the group average cost saving margins where the size of the radius represents the average absolute cost saving for each group. A circle is filled in green if its corresponding group cost saving in LHH is positive and red otherwise. $p_h = 0.5$.

4.6 Conclusion

We study the optimal liquidation strategy with display and hidden limit orders in a limit pool, where the agent is tasked to sell 5% of daily volume within a given period. The theoretical results of our continuous-time model suggest that the optimal liquidation strategy follows a two-phase pattern where it starts with a pure-hidden-order phase, followed by a mixed-order phase. The transition of phases occurs at the switching curve, where the switching time depends on the intensity of market order arrivals. Our simulation results confirm that our strategies are robust to estimation errors and varying market conditions. In addition, the usage of hidden orders helps to reduce the liquidation costs and we show numerically that continuous monitoring of market order arrival frequency further helps to reduce the liquidation costs. Lastly, we test our strategies using event-level data of 102 stocks included in the NASDAQ 100 Index, and find that the strategy using combination of hidden and limit orders while incorporating the self-exciting effect of market order has

the best performance.

Additionally, we leave an important future research direction, that is to extend this work to the multi-asset case see Min et al. (2022), Schneider and Lillo (2019) and Tsoukalas, Wang, and Giesecke (2019). We acknowledge the difficulty in empirically estimating a price-impact matrix from the trades on multiple asset on multiple asset prices. Theoretically, with homogeneous Poisson market order arrivals, it is expected that there will be one switching time for each asset.

4.7 Appendix

4.7.1 Proofs

4.7.1.1 Proof for Proposition 4.3.1

With the excess book value ansatz $U(t,x,q,s) = x + qs + u(t,q)$, the HJB equation is reduced to the following equation about the excess book value function u about time t and inventory q:

$$\begin{aligned} 0 = \frac{\partial u}{\partial t} + \sup_{(L,H)\in A_t} \Bigg\{ \tilde{\lambda} \Big[p_h \big(u(t, q-L-H) - k(L+H)^2 \big), \\ + (1-p_h)(u(t,q-L) - kL^2) - u \Big] - bqL \Bigg\} \\ u(T,q) = -\alpha q^2, \end{aligned} \tag{4.23}$$

where $\tilde{\lambda} = \lambda p_l$ is the execution intensity of limit orders.

Application of the quadratic ansatz $u(t,q) = u_0(t) + qu_1(t) + q^2 u_2(t)$ decomposes u into three functions about time t only. With the ansatz, the Hamiltonian term can be written as:

$$\begin{aligned} \tilde{\lambda} \Big[p_h \big(-(L+H)u_1(t) + ((L+H)^2 - 2q(L+H))u_2(t) - k(L+H)^2 \big) \\ + (1-p_h)(-Lu_1(t) + (L^2 - 2qL)u_2(t) - kL^2) \Big] - bqL. \end{aligned}$$

The optimal order sizes are given by first order conditions over L and H respectively:

$$L^*(t,Q_t) = \frac{u_1(t) + 2Q_t u_2(t) + \frac{bQ_t}{\tilde{\lambda}}}{2(u_2(t)-k)} - p_h H^*(t,Q_t), \tag{4.24}$$

$$H^*(t,Q_t) = \frac{u_1(t) + 2Q_t u_2(t)}{2(u_2(t)-k)} - L^*(t,Q_t). \tag{4.25}$$

If both first order conditions (4.24) and (4.25) hold, we have the following feedback controls by solving the system of linear equations:

$$\begin{cases} L^*(t,Q_t) = -\dfrac{u_1(t)}{2(k-u_2(t))} - Q_t \cdot \dfrac{u_2(t) - \frac{b}{2\tilde{\lambda}(1-p_h)} Q_t}{k-u_2(t)}, & (4.26) \\ H^*(t,Q_t) = \dfrac{-\frac{b}{\tilde{\lambda}(1-p_h)} Q_t}{2(u_2(t)-k)} = Q_t \cdot \dfrac{b}{2\tilde{\lambda}(1-p_h)(k-u_2(t))}. & (4.27) \end{cases}$$

The corresponding optimized Hamiltonian term becomes:

$$\frac{\tilde{\lambda}}{4(k-u_2(t))}\left[u_1^2(t) + 2u_1(t)\left(2u_2(t) + \frac{b}{\tilde{\lambda}}\right)q + \left(4u_2^2(t) + 4u_2(t)\frac{b}{\tilde{\lambda}} + \frac{1}{1-p_h}\left(\frac{b}{\tilde{\lambda}}\right)^2\right)q^2\right].$$

As the optimized Hamiltonian term is also quadratic in q, we can break down the original PDE into a system of ODEs about u_0, u_1 and u_2:

$$\begin{cases} 0 = u_0'(t) + \dfrac{\tilde{\lambda}}{4(k-u_2(t))} u_1^2(t), & u_0(T) = 0, \quad (4.28) \\ 0 = u_1'(t) + u_1(t)\dfrac{\tilde{\lambda}}{2(k-u_2(t))}\left(2u_2(t) + \dfrac{b}{\tilde{\lambda}}\right), & u_1(T) = 0, \quad (4.29) \\ 0 = u_2'(t) + \dfrac{\tilde{\lambda}}{k-u_2(t)}\left[u_2^2(t) + u_2(t)\dfrac{b}{\tilde{\lambda}} + \dfrac{1}{4(1-p_h)}\left(\dfrac{b}{\tilde{\lambda}}\right)^2\right], & u_2(T) = -\alpha. \quad (4.30) \end{cases}$$

As the ODE about u_1 does not have any source term other than u_1 itself, $u_1 \equiv 0$ and the same goes for u_0. With this knowledge, the feedback controls simplify to:

$$\begin{cases} H^*(t,Q_t) = Q_t \cdot \dfrac{b}{2\tilde{\lambda}(1-p_h)(k-u_2(t))}, & (4.31) \\ L^*(t,Q_t) = Q_t \cdot \dfrac{-u_2(t) - \frac{b}{2\tilde{\lambda}(1-p_h)}}{k-u_2(t)}. & (4.32) \end{cases}$$

By non-positivity of u_2, $H^*(t, Q_t)$ in (4.32) is bounded below by 0 and above by Q_t and is therefore admissible. For $L^*(t, Q_t)$ in (4.32), only upper bound Q_t follows from the upper bound of the total order size. Hence the possibility of negative L^* in (4.32) needs to be considered: if the optimal limit order size is negative, then the optimal admissible L should be zero by concavity of the Hamiltonian. When L is selected to be zero, the first order condition on L (4.24) does not hold and only the first order condition on H (4.25) holds.

Hence, the feedback controls of all scenarios are characterized as such:

$$\begin{cases} L^*(t, Q_t) = Q_t \cdot \max\left(1 - \dfrac{k + \frac{b}{2\tilde{\lambda}(1-p_h)}}{k - u_2(t)}, 0\right), & (4.33) \\ H^*(t, Q_t) = Q_t \cdot \left(1 - \dfrac{k}{k - u_2(t)}\right) - L^*(t, Q_t). & (4.34) \end{cases}$$

4.7.1.2 Proof for Theorem 4.3.2

By Proposition 4.3.1, if u_2 is bounded below by $-\alpha$ and above by 0, the feedback controls are as per (4.8). It is worth noting that based on (4.32), $L^* > 0$ if and only if $u_2(t) < -\dfrac{b}{2\tilde{\lambda}(1-p_h)}$. This condition determines the optimal liquidation strategy: when $u_2(t) < -\dfrac{b}{2\tilde{\lambda}(1-p_h)}$, the agent should use both limit orders and hidden orders as per (4.32) and (4.31); otherwise, she should only use hidden orders.

For the pure-hidden-order phase $u_2(t) \geq -\frac{b}{2\tilde{\lambda}(1-p_h)}$, the agent uses such feedback controls:

$$L^*(t, Q_t) = 0, \qquad H^*(t, Q_t) = Q_t \cdot \left(1 - \frac{k}{k - u_2(t)}\right).$$

Plugging the feedback controls into the Hamiltonian term, we have the optimized Hamiltonian:

$$\tilde{\lambda} p_h \left(q u_1(t) \frac{u_2(t)}{k - u_2(t)} + q^2 \frac{u_2^2(t)}{k - u_2(t)}\right)$$

Putting it back into the PDE about the excess value function u (4.23), we can have

a new system of ODEs about u_0, u_1 and u_2:

$$\begin{cases} 0 = u_0'(t), & (4.35) \\ 0 = u_1'(t) + \tilde{\lambda} p_h u_1(t) \dfrac{u_2(t)}{k - u_2(t)}, & (4.36) \\ 0 = u_2'(t) + \tilde{\lambda} p_h \dfrac{u_2^2(t)}{k - u_2(t)}. & (4.37) \end{cases}$$

Note that u_0 and u_1 are constantly zero in the mixed-order phase. As their ODEs in the pure-hidden-order phase also do not have any non-zero source terms, they are always zero in this phase.

Combining with (4.30) describing behaviour of u_2 in the mixed-order phase, the initial (terminal) value problem of u_2 is:

$$\frac{du_2}{dt} = f(t, u_2), \qquad u_2(T) = -\alpha, \tag{4.38}$$

where the function $f : [0, T] \times [-\alpha, 0) \to \mathbb{R}$ is defined as:

$$f(t, y) \coloneqq \mathbb{1}_{\{y < -\frac{b}{2\tilde{\lambda}(1-p_h)}\}} \frac{\tilde{\lambda}}{y - k} \left[\left(y + \frac{b}{2\tilde{\lambda}} \right)^2 + \frac{b^2 p_h}{4\tilde{\lambda}^2 (1 - p_h)} \right] + \mathbb{1}_{\{y \geq -\frac{b}{2\tilde{\lambda}(1-p_h)}\}} \frac{\tilde{\lambda} p_h y^2}{y - k}.$$

Existence and Uniqueness of Solution The piecewise function f is continuous on its domain, and is Lipschitz over y:

- In the region $y \in \left[-\alpha, -\frac{b}{2\tilde{\lambda}(1-p_h)} \right)$,

$$\begin{aligned} \frac{df}{dy} &= -\frac{b^2 + 4bk\tilde{\lambda}(1 - p_h) + 4\tilde{\lambda}^2 y(1 - p_h)(2k - y)}{4\tilde{\lambda}(1 - p_h)(y - k)^2}, \\ \left| \frac{df}{dy} \right| &\leq \frac{b^2 + 4bk\tilde{\lambda}(1 - p_h) + 4\tilde{\lambda}^2 \alpha(1 - p_h)(2k + \alpha)}{4\tilde{\lambda}(1 - p_h)k^2}. \end{aligned}$$

- In the region $y \in \left(-\frac{b}{2\tilde{\lambda}(1-p_h)}, 0 \right)$,

$$\frac{df}{dy} = -\frac{\tilde{\lambda} p_h y(2k - y)}{(y - k)^2},$$

$$\left|\frac{df}{dy}\right| \leq \frac{\tilde{\lambda} p_h \alpha(2k+\alpha)}{k^2}.$$

By Picard–Lindelöf Theorem, the initial value problem (4.38) has unique solution.

Analytical Solution Note that the right hand side of (4.38) is strictly positive, and thus u_2 is decreasing in time. Since the strategy is determined by the value of u_2, the terminal condition for u_2 in the pure-hidden-order phase is $u_2(\tilde{t}) = -\frac{b}{2\tilde{\lambda}(1-p_h)}$. Also, u_0 and u_1 are both uniquely zero, as their associated ODEs do not contain any non-zero source terms.

For the mixed-orders phase case $u_2 \in \left[-\alpha, -\frac{b}{2\tilde{\lambda}(1-p_h)}\right)$, the ordinary differential equation for u_2 with respect to time is as such

$$0 = (k - u_2(t))u_2'(t) + \tilde{\lambda}\left(u_2(t) + \frac{b}{2\tilde{\lambda}}\right)^2 + \frac{b^2}{4\tilde{\lambda}}\frac{p_h}{1-p_h}, \quad u_2(T) = -\alpha. \tag{4.39}$$

This equation can be solved backwards with terminal condition at time T to find the switching time point $\tilde{t}$ where $u_2(\tilde{t}) = -\frac{b}{2\tilde{\lambda}(1-p_h)}$. However, explicit solution does not exist and there is only an implicit equation satisfied by $u_2(t)$:

$$k_1 - t = \frac{1}{\tilde{\lambda}\nu}\left(k + \frac{b}{2\tilde{\lambda}}\right)\tan^{-1}\left(\frac{u_2(t) + \frac{b}{2\tilde{\lambda}}}{\nu}\right) - \frac{1}{2\tilde{\lambda}}\ln\left[\lambda\Big(u_2(t) + \frac{b}{2\tilde{\lambda}}\Big)^2 + \tilde{\lambda}\nu^2\right], \tag{4.40}$$

where $\nu = \frac{b}{2\tilde{\lambda}}\sqrt{\frac{p_h}{1-p_h}}$.

From the terminal condition that $u_2(T) = -\alpha$, the constant k_1 can be backed out:

$$k_1 = T + \frac{1}{\tilde{\lambda}\nu}\left(k + \frac{b}{2\tilde{\lambda}}\right)\tan^{-1}\left[\frac{-\alpha + \frac{b}{2\tilde{\lambda}}}{\nu}\right] - \frac{1}{2\tilde{\lambda}}\ln\left[\tilde{\lambda}\Big(\frac{b}{2\tilde{\lambda}} - \alpha\Big)^2 + \tilde{\lambda}\nu^2\right].$$

As $u_2(\tilde{t}) = -\dfrac{b}{2\tilde{\lambda}(1-p_h)}$, the switching time is given explictly by (4.40):

$$\tilde{t} = k_1 - \frac{1}{\tilde{\lambda}\nu}\left(k + \frac{b}{2\tilde{\lambda}}\right), \tan^{-1}\left[\frac{-bp_h}{2\tilde{\lambda}(1-p_h)\nu}\right] + \frac{1}{2\tilde{\lambda}}\log\frac{b^2 p_h}{4\tilde{\lambda}(1-p_h)^2}.$$

For the pure-hidden-order phase case $t \in [0, \tilde{t}]$, explicit solution of u_2 from (4.37)

exists

$$u_2(t) = -k\left[W\left(k\exp(c_2 - \tilde{\lambda}p_h t)\right)\right]^{-1}.$$

Plugging in the terminal condition $u_2(\tilde{t}) = -\frac{b}{2\tilde{\lambda}(1-p_h)}$,

$$u_2(t) = -k\left[W\left(\phi\exp(\phi + \tilde{\lambda}p_h(\tilde{t}-t))\right)\right]^{-1} \quad \text{for} \quad \phi = \frac{2k\tilde{\lambda}(1-p_h)}{b}. \tag{4.41}$$

The corresponding optimal hidden order size for $t \in (0, \tilde{t}]$ is:

$$H^*(t, Q_t) = Q_t\left\{1 + W\left[\phi\exp\left(\phi + \tilde{\lambda}p_h(\tilde{t}-t)\right)\right]\right\}^{-1}.$$

Finally, we validate our a prior assumption about the boundedness of u_2:

- For the pure-hidden-order phase, from the solution of u_2 (4.41), u_2 is bounded below by threshold $-\frac{b}{2\tilde{\lambda}(1-p_h)}$ and above by 0.
- For the mixed-orders phase, from (4.30) it satisfies, u_2 is monotonically decreasing in time and its terminal value is $-\alpha$. Going backwards in time, as the mixed-orders phase starts once u_2 reaches the threshold $-\frac{b}{2\tilde{\lambda}(1-p_h)} < 0$, u_2 is bounded below by threshold $-\alpha$ and above by $-\frac{b}{2\tilde{\lambda}(1-p_h)}$ in this phase.

Verification Argument In a spirit to apply the traditional verification result, we check the following conditions:

1. **Uniqueness of controlled state processes:** The SDEs governing the joint dynamics of cash process X, inventory process Q and fundamental price process S associated to the optimal liquidation strategy (L^*, H^*) admit unique solutions.
2. **Regularity of the state processes:** The state processes X, Q, S are *corlol*, i.e. right-continuous with left limits. Also, the L^p norm of the state process is bounded above by $C_p(1 + ||(x, q, s)||^p)$ for constant C_p.

3. **Admissibility of the control processes:** The liquidation strategy outlined in 'Analytical Solution' of Appendix 4.7.1.2 is a measurable function of the time and state, and is also adapted, i.e. $\mathcal{F}_t$-measurable.

4. **Quadratic Growth of the value function:** Recall that $u_2 \in [-\alpha, 0)$. Thus, $U(t,x,q,s) = x+qs+q^2u_2(t)$ satisfies the quadratic growth condition $U(t,x,q,s) \leq C(1+||(x,q,s)||^2)$.

5. **Smoothness of the value function:** The function u_2 is at least continuously differentiable in time.[13] As it is the only part in U related to time, U is also continuously differentiable in time and infinitely differentiable in space with the previous decompositions.

Hence, by Fleming and Soner (2006, Theorem 8.1), we can conclude that the liquidation strategy in Appendix 4.7.1.2 is indeed an optimal Markovian control to the optimization problem and the classical solution of the PDE is equal to the value function defined in (4.6).

4.7.1.3 Proof for Corollary 4.3.6: Limit-only Strategy

If the agent can only use limit orders, i.e. $H = 0$, with the excess value decomposition $U^L(t,x,q,s) = x + qs + u^L(t,q)$ the HJB equation is as follows:

$$0 = \frac{\partial u^L}{\partial t} + \sup_{0\leq L\leq Q_t} \Big(\tilde{\lambda}\big(u^L(t,q-L) - kL^2 - u^L\big) - bLq\Big),$$
$$u(T,q) = -\alpha q^2.$$

where the admissible set of controls is $\mathcal{A}^L_{[t,T]}$ which contains all $\mathcal{F}_t$-progressively measurable processes $(L_\tau)_{t\leq\tau\leq T}$ that satisfies $\forall\tau \in [t,T)$, $0 \leq L_\tau \leq Q_\tau$.

With the quadratic ansatz on q, $u^L(t,q) = u^L_0(t) + qu^L_1(t) + q^2u^L_2(t)$, the Hamiltonian term is expressed as $\tilde{\lambda}\big(-kL^2 - Lu^L_1(t) + (L^2 - 2qL)u^L_2(t)\big) - bLq$.

[13] This is due to the fact that we can write $u_2'(t)$ as a continuous function of $u_2(t)$ and that $u_2(t)$ is continuous in t.

The optimal control L^* is given by the first order condition:

$$0 = \frac{\partial}{\partial L}_{L=L^*}\Big[- kL^2 - Lu_1^L(t) + (L^2 - 2qL)u_2^L(t) - \frac{bqL}{\tilde{\lambda}}\Big],$$
$$L^*(t,q) = \frac{-u_1^L(t) - 2qu_2^L(t) - \frac{bq}{\tilde{\lambda}}}{2(k - u_2^L(t))}. \tag{4.42}$$

Given the experience in the mixed-orders phase, L^* in (4.42) *may* not be non-negative (admissible) and investigate the active case in which the agent uses the optimal order size L^* as per (4.42) carefully. In particular, if $u_2^L(t)$ is bigger than $-\frac{b}{2\tilde{\lambda}}$ then the agent should turn dormant and does not place any limit order at all at time t.

Here we solve the problem for the case of $\alpha > \frac{b}{2\tilde{\lambda}}$. Plugging $L = L^*$ into the Hamiltonian term, the optimized Hamiltonian is:

$$\tilde{\lambda}\frac{(u_1^L(t) + 2qu_2^L(t) + \frac{bq}{\tilde{\lambda}})^2}{4(k - u_2^L(t))},$$

and the original HJB can be broken down into the following system of ODEs:

$$\begin{cases} 0 = \dfrac{\partial u_0^L}{\partial t} + \tilde{\lambda}\dfrac{(u_1^L(t))^2}{4(k - u_2^L(t))}, & u_0^L(T) = 0, \quad (4.43) \\ 0 = \dfrac{\partial u_1^L}{\partial t} + \tilde{\lambda}\dfrac{u_1^L(t)\Big(u_2^L(t) + \frac{b}{2\tilde{\lambda}}\Big)}{k - u_2^L(t)}, & u_1^L(T) = 0, \quad (4.44) \\ 0 = \dfrac{\partial u_2^L}{\partial t} + \tilde{\lambda}\dfrac{\Big(u_2^L(t) + \frac{b}{2\tilde{\lambda}}\Big)^2}{k - u_2^L(t)}, & u_2^L(T) = -\alpha. \quad (4.45) \end{cases}$$

Along the lines of 'Existence and Uniqueness of Solution' in Appendix 4.7.1.2, it can be shown that u_2 has a unique solution. Similarly, $u_0^L(t)$ and $u_1^L(t)$ are uniquely zero at any time t. With the change of variable $\chi(t) \coloneqq u_2^L(t) + \frac{b}{2\tilde{\lambda}}$, $\tilde{\alpha} \coloneqq \alpha - \frac{b}{2\tilde{\lambda}}$ and $\tilde{k} \coloneqq k + \frac{b}{2\tilde{\lambda}}$, the ODE about χ is:

$$0 = \chi'(t) + \tilde{\lambda}\frac{\chi^2(t)}{\tilde{k} - \chi(t)}, \chi(T) = -\tilde{\alpha},$$

with solution in the following form:

$$\chi(t) = \frac{-\tilde{k}}{W\left(\tilde{k}\exp\left(c_1 - \tilde{\lambda}t\right)\right)},$$

With the terminal condition $\chi(T) = -\tilde{\alpha}$, we can back out c_1:

$$c_1 = -\log\alpha + \tilde{\lambda}T + \frac{\tilde{k}}{\tilde{\alpha}}.$$

Hence

$$u_2^L(t) = -\tilde{k}\Big[W\left(\tilde{\phi}\exp(\tilde{\phi} + \tilde{\lambda}(T-t))\right)\Big]^{-1} - \frac{b}{2\tilde{\lambda}}, \tag{4.46}$$

where $\tilde{\phi} = \frac{\tilde{k}}{\tilde{\alpha}}$.

We should note that $u_2^L(t)$ is always smaller than $-\frac{b}{2\tilde{\lambda}}$ as per (4.46), and the corresponding optimal control L^* as per (4.42) is then always admissible.

The optimal limit order size L^* has the following expression:

$$\begin{aligned} L^*(t, Q_t) &= Q_t \frac{\tilde{k}\Big[W\left(\tilde{\phi}\exp(\tilde{\phi} + \tilde{\lambda}(T-t))\right)\Big]^{-1}}{k + \frac{b}{2\tilde{\lambda}} + \tilde{k}\Big[W\left(\tilde{\phi}\exp(\tilde{\phi} + \tilde{\lambda}(T-t))\right)\Big]^{-1}}, \\ &= Q_t\left[1 + W\left(\tilde{\phi}\exp(\tilde{\phi} + \tilde{\lambda}(T-t))\right)\right]^{-1}. \end{aligned} \tag{4.47}$$

As $L^*(t, Q_t)$ is always between 0 and Q_t, it is indeed an admissible control. We can prove along the lines of 'Verification Argument' in Appendix 4.7.1.2 to show that the classical solution of the HJB is equal to the value function and L^* is an optimal Markovian control of the optimization problem defined in (4.14).

Regarding the case that $\alpha \leq \frac{b}{2\tilde{\lambda}}$, the optimal limit order size L^* is always zero and for all time t, $u_2^L(t) = -\alpha$.

4.7.1.4 Proof for Proposition 4.3.7: Reduction to Quadratic PDE

Using the ansatz $V(t,x,s,q,\lambda) = x + qs + v(t,q,\lambda)$, we reduce the HJB (4.18) to a PDE about the excess book value u:

$$\begin{aligned} 0 = \frac{\partial v}{\partial t} - \beta(\lambda - \underline{\lambda})\frac{\partial v}{\partial \lambda} + \lambda(1-p_l)\left[v(t,q,\lambda+\eta) - v\right], \\ + \sup_{(L,H)\in\mathcal{A}_q} \Bigg\{ \lambda p_l \Big[p_h\big(v(t,q-L-H,\lambda+\eta) - (L+H)(k_L L + k_H H)\big), \\ + (1-p_h)(v(t,q-L,\lambda+\eta) - k_L L^2) - v \Big] - bqL \Bigg\} \end{aligned} \tag{4.48}$$

$v(T,q,\lambda) = -\alpha q^2$.

Again quadratic ansatz over inventory $v(t,q,\lambda) = q^2 v_2(t,\lambda)$ reduces the Hamiltonian term to:

$$\begin{aligned} \lambda p_l \Big[p_h\big((q-L-H)^2 v_2(t,\lambda+\eta) - k(L+H)^2\big) \\ + (1-p_h)((q-L)^2 v_2(t,\lambda+\eta) - kL^2) - q^2 v_2(t,\lambda) \Big] - bqL. \end{aligned}$$

First order conditions over L and H are

$$\begin{aligned} L^*(t,q,\lambda) &= q\frac{-v_2(t,\lambda+\eta) - \frac{b}{2\lambda p_l}}{k - v_2(t,\lambda+\eta)} - p_h H^*(t,q,\lambda), \\ H^*(t,q,\lambda) &= q\frac{-v_2(t,\lambda+\eta)}{k - v_2(t,\lambda+\eta)} - L^*(t,q,\lambda) \end{aligned} . \tag{4.49}$$

Similar as in Proposition 4.3.1, the first order conditions yield the optimal order sizes

$$\begin{aligned} L^*(t,q,\lambda) &= q \cdot \max\left\{ \frac{-v_2(t,\lambda+\eta) - \frac{b}{2\lambda p_l(1-p_h)}}{k - v_2(t,\lambda+\eta)}, 0 \right\}, \\ H^*(t,q,\lambda) &= q \cdot \frac{-v_2(t,\lambda+\eta)}{k - v_2(t,\lambda+\eta)} - L^*(t,q,\lambda). \end{aligned} \tag{4.50}$$

There are two scenarios in which different optimization results occur:

1. If L^* is positive, i.e. $v_2(t,\lambda+\eta) < -\dfrac{b}{2\lambda p_l(1-p_h)}$, the optimized Hamiltonian term

is:

$$\lambda p_l q^2 \Bigg[\frac{\Big(v_2(t,\lambda+\eta)+\frac{b}{2\lambda p_l}\Big)^2 + \frac{b^2 p_h}{4\lambda^2 p_l^2(1-p_h)}}{k - v_2(t,\lambda+\eta)} + v_2(t,\lambda+\eta) - v_2(t,\lambda) \Bigg].$$

Under this scenario, i.e. $v_2(t,\lambda+\eta) < -\dfrac{b}{2\lambda p_l(1-p_h)}$, (4.48) becomes:

$$\begin{aligned} 0 &= \partial_t v - \beta(\lambda-\underline{\lambda})\partial_\lambda v + \lambda q^2 \left[v_2(t,\lambda+\eta) - v_2(t,\lambda)\right], \\ &+ \frac{\lambda p_l q^2}{k - v_2(t,\lambda+\eta)} \left[\Big(v_2(t,\lambda+\eta)+\frac{b}{2\lambda p_l}\Big)^2 + \frac{b^2 p_h}{4\lambda^2 p_l^2(1-p_h)} \right]. \end{aligned} \tag{4.51}$$

which confirms that under this scenario, $v(t,q,\lambda)$ is still at most quadratic in q.

2. If $L^* = 0$, i.e. $v_2(t,\lambda+\eta) \geq -\dfrac{b}{2\lambda p_l(1-p_h)}$, the optimized Hamiltonian term is:

$$\lambda p_l q^2 \Big[\frac{p_h v_2^2(t,\lambda+\eta)}{k - v_2(t,\lambda+\eta)} + v_2(t,\lambda+\eta) - v_2(t,\lambda) \Big].$$

Under this scenario, i.e. $v_2(t,\lambda+\eta) \geq -\dfrac{b}{2\lambda p_l(1-p_h)}$, (4.48) becomes:

$$0 = \partial_t v - \beta(\lambda-\underline{\lambda})\partial_\lambda v + \lambda p_l q^2 \left[v_2(t,\lambda+\eta) - v_2(t,\lambda)\right] + \frac{\lambda p_l p_h q^2 v_2^2(t,\lambda+\eta)}{k - v_2(t,\lambda+\eta)}, \tag{4.52}$$

which confirms that under this scenario, $v(t,q,\lambda)$ is also still at most quadratic in q.

As we have confirmed that $v(t,q,\lambda)$ is at most quadratic in q in both scenarios and the optimal order sizes are proportional to q, (4.48) about the excess book value function

v can be further reduced to the equation about $v_2(t,\lambda)$:

$$\begin{aligned}
0 &= \partial_t v_2 - \beta(\lambda - \underline{\lambda})\partial_\lambda v_2 + \lambda(1-p_l)\left[v(t,q,\lambda+\eta) - v\right], \\
&+ \sup_{(\hat{L},\hat{H})\in\hat{A}} \Bigg\{ \lambda p_l \Big[p_h\big((1-\hat{L}-\hat{H})^2 v_2(t,\lambda+\eta) - k(\hat{L}+\hat{H})^2\big), \\
&+ (1-p_h)((1-\hat{L})^2 v_2(t,\lambda+\eta) - k\hat{L}^2) - v_2(t,\lambda)\Big] - b\hat{L} \Bigg\}, \\
v_2(T,\lambda) &= -\alpha.
\end{aligned} \tag{4.53}$$

After optimizing the Hamiltonian term, the PDE without supremum appears:

$$\begin{aligned}
0 &= \partial_t v_2 - \beta(\lambda - \underline{\lambda})\partial_\lambda v_2 + \lambda(v_2(t,\lambda+\eta) - v_2(t,\lambda)), \\
&+ \frac{\lambda p_l}{k - v_2(t,\lambda+\eta)} \Bigg\{ \left[\left(v_2(t,\lambda+\eta) + \frac{b}{2\lambda p_l} \right)^2 + \frac{b^2 p_h}{4\lambda^2 p_l^2 (1-p_h)} \right] \mathbb{1}_{\{v_2(t,\lambda+\eta) < -\frac{b}{2\lambda p_l(1-p_h)}\}}, \\
&+ p_h v_2^2(t,\lambda+\eta) \mathbb{1}_{\{v_2(t,\lambda+\eta) \geq -\frac{b}{2\lambda p_l(1-p_h)}\}} \Bigg\}, \\
v_2(T,\lambda) &= -\alpha.
\end{aligned}$$

4.7.2 Proof for Proposition 4.3.9: Properties of Reduced-form value function

Using the definition of v_2 through V (4.5),

$$v_2(t,\lambda) = \frac{V(t,x,q,s,\lambda) - x - qs}{q^2}, \quad \forall\, t,x,q,s,\lambda, \tag{4.54}$$

we show several properties that v_2 satisfies, with arbitrage starting state $X_{t-} = x, S_{t-} = s, Q_{t-} = q$:

1. **Monotonicity in** λ: For any fixed t, $v_2(t,\lambda)$ is increasing in λ.

   We consider the corresponding value function of an intensity-irrelevant liquidation strategy $(L_\tau, H_\tau)_{t\leq\tau\leq T} \in \mathcal{A}_{[t,T]}$, in which at any time τ, $L_\tau = \hat{L}_\tau Q_\tau$ and $H_\tau = \hat{H}_\tau Q_\tau$

are independent of market order arrival intensity.

$$
\begin{aligned}
&\mathbb{E}\left\{X_T + Q_T S_T - \alpha Q_T^2 \mid \mathcal{F}_t\right\} - x - qs \\
&= \mathbb{E}\left\{\int_t^T dX_\tau + \left(q + \int_t^T dQ_\tau\right) S_T - \alpha\left(q + \int_t^T dQ_\tau\right)^2 \mid \mathcal{F}_t\right\} - qs \\
&= \mathbb{E}_t\Big[- \int_t^T \hat{S}_\tau dQ_\tau\Big] \\
&+ \mathbb{E}_t\Big[q\int_t^T dS_\tau + s\int_t^T dQ_\tau + \int_t^T dQ_\tau \int_t^T dS_\tau - \alpha\Big(q + \int_t^T dQ_\tau\Big)^2\Big] \\
&= \mathbb{E}_t\Big[- \int_t^T (S_\tau + k dQ_\tau) dQ_\tau + s\int_t^T dQ_\tau\Big] + \mathbb{E}_t\Big[- bQ_T \int_t^T L_\tau d\tau - \alpha Q_T^2\Big] \\
&= \mathbb{E}_t\Big[b\int_t^T \int_t^\tau L_{\tau'} d\tau' dQ_\tau - k\int_t^T d[Q]_\tau - bQ_T\int_t^T L_\tau d\tau - \alpha Q_T^2\Big] \\
&= \mathbb{E}_t\Big[- b\int_t^T \int_t^\tau L_{\tau'} d\tau' \lambda_\tau (L_\tau + p_h H_\tau) d\tau - k\int_t^T \lambda_\tau (L_\tau + p_h H_\tau)^2 d\tau \\
&- bQ_T\Big(\int_t^T L_\tau d\tau\Big) - \alpha Q_T^2\Big] \\
&= \mathbb{E}_t\Big[- b\int_t^T \int_t^\tau Q_{\tau'} \hat{L}_{\tau'} d\tau' \lambda_\tau Q_\tau (\hat{L}_\tau + p_h \hat{H}_\tau) d\tau - k\int_t^T \lambda_\tau Q_\tau^2 (\hat{L}_\tau + p_h \hat{H}_\tau)^2 d\tau \\
&- bQ_T\Big(\int_t^T Q_\tau \hat{L}_\tau d\tau\Big) - \alpha Q_T^2\Big].
\end{aligned}
$$

For a fixed path Ω, the number of executions N_T^L is increasing with λ_t, and for any $\tau \in [t, T]$, Q_τ is decreasing with λ_t, given a fixed strategy. Hence the quantity $\mathbb{E}[X_T + Q_T S_T - \alpha Q_T^2 \mid \mathcal{F}_t] - x - qs$ is increasing in $\lambda_t = \lambda$. As for each given strategy and given path, this quantity is increasing in λ, the supremum over all possible strategies is also increasing in λ, and so does the reduced form value function $v_2(t, \lambda)$.

2. **Boundedness**: For any values of t and λ, $-\alpha \leq v_2(t, \lambda) \leq 0$.

   For the lower bound $-\alpha$:

$$
\begin{aligned}
v_2(t, \lambda) &= \frac{1}{q^2}\left[V(t, x, s, q, \lambda) - x - qs\right] \\
&= \frac{1}{q^2}\left[\sup_{(L,H)\in\mathcal{A}_{[t,T]}} \mathbb{E}\left\{X_T + Q_T S_T - \alpha Q_T^2 \mid \mathcal{F}_t^{x,s,q,\lambda}\right\} - x - qs\right]
\end{aligned}
$$

$$
\begin{aligned}
&\geq \frac{1}{q^2}\left[\mathbb{E}_{\mathcal{N}_{[t,T]}}\left\{X_T + Q_T S_T - \alpha Q_T^2 \mid \mathcal{F}_t^{x,s,q,\lambda}\right\} - x - qs\right] \\
&= \frac{1}{q^2}\left[\mathbb{E}\left\{X_t + Q_t S_T - \alpha Q_t^2 \mid \mathcal{F}_t^{x,s,q,\lambda}\right\} - x - qs\right] \\
&= \frac{1}{q^2}\left[\mathbb{E}\left[x + q(s + S_T - S_t) - \alpha q^2\right] - x - qs\right] \\
&= \frac{1}{q^2}\left[x + qs - \alpha q^2 - x - qs\right] = -\alpha.
\end{aligned}
$$

For the upper bound 0:

$$
\begin{aligned}
v_2(t,\lambda) &= \frac{1}{q^2}\left[V(t,x,s,q,\lambda) - x - qs\right] \\
&= \frac{1}{q^2}\left\{\sup_{(L,H)\in\mathcal{A}_{[t,T]}} \mathbb{E}\left[X_T + Q_T S_T - \alpha Q_T^2 \mid \mathcal{F}_t^{x,s,q,\lambda}\right] - x - qs\right\} \\
&= \frac{1}{q^2}\left\{\sup_{(L,H)\in\mathcal{A}_{[t,T]}} \mathbb{E}\left[x + \int_t^T dX_\tau + \left(q + \int_t^T dQ_\tau\right) S_T\right] - x - qs\right\} \\
&\leq \frac{1}{q^2}\left\{\sup_{(L,H)\in\mathcal{A}_{[t,T]}} \mathbb{E}\left[x - \int_t^T (S_\tau + k dQ_\tau) dQ_\tau + \left(q + \int_t^T dQ_\tau\right) s\right] - x - qs\right\} \\
&\leq \frac{1}{q^2}\left\{\sup_{(L,H)\in\mathcal{A}_{[t,T]}} \mathbb{E}\left[-\int_t^T S_\tau dQ_\tau + qs + \int_t^T s dQ_\tau\right] - qs\right\} \\
&= \frac{1}{q^2}\left\{\sup_{(L,H)\in\mathcal{A}_{[t,T]}} \mathbb{E}\left[\int_t^T (s - S_\tau) dQ_\tau\right]\right\} \leq 0.
\end{aligned}
$$

4.7.2.1 Proof for Proposition 4.4.2

Proof. The expected intensity is actually a deterministic function of time and its dynamics can be expressed as follows,

$$
d\mathbb{E}[\lambda_t] = -\beta(\mathbb{E}[\lambda_t] - \underline{\lambda})dt + \eta\mathbb{E}[\lambda_t]dt.
$$

Reorganizing, we have the ordinary differential equation

$$
\frac{d}{dt}(\mathbb{E}[\lambda_t] - \frac{\underline{\lambda}\beta}{\beta - \eta}) = (\eta - \beta)(\mathbb{E}[\lambda_t] - \frac{\underline{\lambda}\beta}{\beta - \eta}).
$$

The solution of this ODE is

$$\mathbb{E}[\lambda_t] = \frac{\underline{\lambda}\beta}{\beta - \eta} + C\exp(-(\beta - \eta)t),$$

where $C = \lambda_0 - \frac{\underline{\lambda}\beta}{\beta-\eta}$ is a constant related to the initial value of intensity λ_0.

Hence, taking t to infinity, we have $\mathbb{E}[\lambda_\infty] = \lim_{t\to\infty}\mathbb{E}[\lambda_t] = \underline{\lambda}\frac{\beta}{\beta-\eta}$. □

4.7.3 Model Sensitivity Analysis of the Homogeneous Poisson case

We lay out the dependence of optimal liquidation strategy on market conditions characterized by the model parameters, such as λ, α, k, b and p_h in this section. Note that in the absence of explicit solution, we leverage on numerical results to study the relationship between the optimal strategy and market parameters.

Dependency on immediate execution cost and price impact of market orders: k and α We first look at the how price impact of selling with market orders influences the order sizes in Figure 4.12. The total order size increases with α as seen in Figure b. Facing a higher pressure-to-sell, limit order is more preferable as it is more time efficient to reduce the inventory pressure of the agent, as shown in Figure c. Hidden order size does not heavily depend on α.

Next we examine the relationship between value function / liquidation strategy and the immediate execution cost parameter k in Figure 4.13. Conceptually, higher immediate execution costs prompts the agent to use smaller sized orders and hence on average leaving larger remnant inventory at time T. In addition, we observe across different order types that indeed the order size decreases with k. This is consistent with our discussion in Remark 4.3.4, that the speed of decay of $u_2(t)$ with respect to time is scaled by $k - u_2(t)$ in the ODE.

Dependency on hidden order execution probability and exposure risk: p_h and b Next, we shift our attention to the hidden order fill probability p_h in Figure 4.14.

Generally speaking, higher hidden fill probability benefits the agent, as the agent enjoys a higher overall fill probability as long as she post hidden order to the order book. We can observe the value function increase slightly with p_h in Figure 4.14a. In particular, instead of impacting the total order size, the hidden order fill probability has more pronounced impact on the breakdown between hidden and limit orders. When the hidden order fill probability is high, hidden order is more preferred to take up a higher proportion in the total order size, vice versa for limit order size.

Last but not least, we study the effect of exposure risk coefficient b in Figure 4.15. Recall that posting of limit order is subject to exposure risk due to changes in order imbalance. A bigger b prevents the agent from posting large limit order and therefore increases the likelihood of agent paying a larger terminal execution cost, in turn reducing the agent's value function. This is again consistent with our discussion in Remark 4.3.4, that $\nu = \frac{b}{2\bar{\lambda}}\sqrt{\frac{p_h}{1-p_h}}$ directly controls the speed of decay of $u_2(t)$. Naturally, limit order size decreases with exposure risk coefficient b as reflected in the figure. Facing a more costly limit order, the agent has to rely more heavily on hidden order, thus uses larger sized hidden order. Increasing hidden order size and decreasing limit order size both results in less efficient liquidation, requiring larger total order size as compensation as shown in Figure 4.15.

4.7.4 Empirical Estimation and Testing results

In this section, we display the results from empirical estimation and testing on 100 NASDAQ stocks (102 stocks).

Table 4.7: Parameter estimation results of price impact model for NASDAQ 100 stocks.

ticker	α	k	b	p_l	λ	$\underline{\lambda}$	η	β	**ticker**	α	k	b	p_l	λ	$\underline{\lambda}$	η	β
AAL	4.3E-05	1.9E-06	5.8E-08	0.45	0.074	0.051	0.359	0.111	JBHT	2.1E-04	2.3E-05	3.5E-07	0.59	0.031	0.022	0.459	0.125
AAPL	5.4E-06	8.5E-07	6.3E-08	0.41	0.558	0.145	0.055	0.041	JD	1.0E-05	1.0E-06	1.8E-08	0.42	0.171	0.030	0.016	0.013
ADBE	4.6E-04	2.2E-05	5.9E-07	0.73	0.058	0.041	0.459	0.129	KHC	4.8E-05	3.1E-06	2.4E-07	0.56	0.098	0.067	0.282	0.089
ADI	1.0E-04	1.9E-06	2.6E-07	0.66	0.076	0.048	0.239	0.088	KLAC	1.8E-04	1.0E-06	7.1E-07	0.58	0.036	0.026	0.390	0.107
ADP	2.1E-04	7.1E-06	1.1E-07	0.68	0.047	0.035	0.423	0.113	LBTYA	3.0E-05	1.2E-06	1.8E-08	0.27	0.062	0.020	0.115	0.079
ADSK	1.9E-04	1.0E-05	9.0E-07	0.66	0.055	0.035	0.359	0.133	LBTYK	2.5E-05	2.5E-06	5.4E-08	0.43	0.074	0.014	0.031	0.025
ALGN	1.1E-03	1.2E-04	1.7E-06	0.74	0.031	0.022	0.282	0.083	LRCX	2.5E-04	1.6E-05	1.0E-06	0.70	0.074	0.050	0.423	0.136
ALXN	4.0E-04	1.4E-05	7.3E-07	0.69	0.045	0.029	0.331	0.118	LULU	1.3E-04	5.5E-06	3.8E-07	0.61	0.058	0.024	0.260	0.153
AMAT	1.0E-05	1.5E-06	3.1E-08	0.38	0.211	0.036	0.024	0.020	MAR	1.6E-04	1.6E-05	1.1E-06	0.58	0.056	0.038	0.359	0.116
AMD	9.3E-07	2.8E-07	2.8E-09	0.11	0.070	0.012	0.031	0.026	MCHP	7.3E-05	9.6E-06	6.1E-07	0.60	0.082	0.052	0.331	0.124
AMGN	2.8E-04	2.6E-05	1.0E-06	0.64	0.072	0.050	0.459	0.137	MDLZ	1.4E-05	2.3E-06	2.6E-08	0.38	0.110	0.035	0.065	0.044
AMZN	3.7E-04	1.7E-04	3.5E-06	0.42	0.165	0.077	0.239	0.128	MELI	1.8E-03	1.3E-04	8.6E-07	0.64	0.013	0.010	0.390	0.085
ASML	1.8E-04	7.8E-06	1.8E-07	0.45	0.024	0.005	0.018	0.014	MNST	1.2E-04	1.1E-05	2.5E-07	0.59	0.049	0.036	0.331	0.086
ATVI	5.8E-05	4.1E-06	5.2E-07	0.68	0.130	0.083	0.239	0.087	MSFT	6.7E-06	7.8E-07	1.1E-07	0.41	0.449	0.058	0.034	0.030
AVGO	4.3E-04	3.1E-05	3.2E-07	0.74	0.074	0.049	0.459	0.152	MU	5.6E-06	5.0E-07	1.8E-08	0.26	0.296	0.026	0.021	0.019
BIDU	2.3E-04	1.8E-05	1.7E-07	0.52	0.060	0.040	0.331	0.108	MXIM	1.4E-04	3.3E-06	2.7E-07	0.58	0.060	0.029	0.125	0.065
BIIB	1.0E-03	1.2E-04	6.3E-07	0.66	0.031	0.015	0.115	0.058	MYL	4.2E-05	1.0E-06	3.6E-08	0.50	0.087	0.054	0.187	0.072
BMRN	1.9E-04	1.2E-05	1.5E-07	0.59	0.031	0.022	0.423	0.120	NFLX	1.3E-04	1.3E-05	1.2E-06	0.65	0.135	0.074	0.221	0.099
CDNS	7.3E-05	7.0E-06	1.5E-07	0.53	0.060	0.043	0.203	0.057	NTAP	3.2E-05	4.8E-06	3.3E-07	0.55	0.074	0.033	0.221	0.122
CELG	8.0E-05	4.6E-06	3.9E-07	0.68	0.100	0.070	0.359	0.108	NTES	1.2E-03	9.8E-05	3.2E-07	0.71	0.030	0.023	0.390	0.096
CERN	1.0E-04	7.6E-06	2.3E-07	0.58	0.054	0.037	0.305	0.092	NVDA	2.6E-05	6.2E-06	1.6E-07	0.43	0.217	0.081	0.159	0.100
CHKP	2.1E-04	1.2E-05	4.8E-07	0.63	0.041	0.030	0.423	0.116	NXPI	1.3E-04	3.8E-06	4.3E-08	0.48	0.030	0.017	0.239	0.106
CHTR	1.4E-03	5.1E-05	5.0E-07	0.73	0.032	0.020	0.305	0.113	ORLY	9.9E-04	8.1E-05	3.8E-07	0.72	0.029	0.021	0.359	0.098
CMCSA	8.4E-06	9.1E-07	3.0E-08	0.27	0.153	0.020	0.029	0.025	PAYX	7.6E-05	4.4E-06	2.4E-07	0.61	0.056	0.032	0.239	0.102
COST	4.0E-04	3.2E-05	6.0E-07	0.67	0.050	0.038	0.498	0.124	PCAR	1.4E-04	1.0E-05	3.8E-07	0.64	0.051	0.039	0.359	0.084
CSCO	4.2E-06	9.0E-07	2.0E-08	0.16	0.227	0.018	0.024	0.023	PEP	7.9E-05	8.4E-06	1.8E-07	0.57	0.079	0.056	0.305	0.091
CSX	3.0E-05	4.3E-06	2.0E-07	0.49	0.112	0.049	0.083	0.047	PYPL	1.3E-05	3.5E-06	1.6E-07	0.52	0.176	0.060	0.090	0.059
CTAS	3.7E-04	6.3E-05	7.1E-07	0.57	0.020	0.014	0.282	0.089	QCOM	1.5E-05	2.2E-06	9.3E-08	0.52	0.146	0.051	0.125	0.081
CTRP	4.3E-05	1.1E-06	1.0E-07	0.51	0.069	0.049	0.359	0.102	REGN	1.6E-03	1.3E-04	1.5E-06	0.62	0.029	0.022	0.498	0.121
CTSH	7.0E-05	8.5E-06	2.1E-07	0.55	0.084	0.058	0.305	0.095	ROST	1.3E-04	1.6E-05	4.2E-07	0.64	0.060	0.044	0.390	0.100
CTXS	1.9E-04	2.0E-05	3.1E-07	0.65	0.047	0.034	0.390	0.107	SBUX	1.5E-05	1.4E-06	2.7E-08	0.34	0.137	0.037	0.083	0.061
DLTR	1.7E-04	1.4E-05	4.7E-07	0.64	0.057	0.039	0.390	0.121	SIRI	4.2E-07	1.3E-07	2.7E-10	0.05	0.016	0.003	0.098	0.077
EA	1.3E-04	8.1E-06	7.6E-07	0.72	0.082	0.054	0.423	0.145	SNPS	2.0E-04	1.0E-05	5.4E-07	0.68	0.042	0.031	0.331	0.091
EBAY	1.5E-05	2.4E-06	3.7E-08	0.36	0.150	0.034	0.040	0.031	SWKS	2.0E-04	1.3E-05	2.3E-07	0.52	0.063	0.044	0.390	0.119
EXPE	7.0E-05	1.2E-05	4.2E-07	0.61	0.075	0.038	0.282	0.140	SYMC	1.6E-05	9.1E-07	1.6E-08	0.31	0.082	0.031	0.060	0.037
FAST	9.5E-05	6.6E-06	3.3E-07	0.59	0.059	0.037	0.239	0.089	TMUS	4.8E-05	4.3E-06	1.3E-07	0.57	0.066	0.043	0.221	0.077
FB	6.4E-06	1.8E-06	1.9E-07	0.41	0.275	0.111	0.098	0.058	TSLA	2.4E-04	1.6E-05	3.9E-07	0.57	0.089	0.043	0.305	0.158
FISV	2.8E-04	2.6E-05	4.9E-07	0.60	0.040	0.023	0.282	0.118	TTWO	2.1E-04	1.6E-05	8.7E-07	0.64	0.057	0.030	0.239	0.112
FOXA	9.1E-06	1.3E-06	3.7E-08	0.30	0.111	0.035	0.090	0.062	TXN	5.9E-05	5.1E-06	4.5E-07	0.63	0.135	0.092	0.390	0.125
FOX	1.7E-05	1.3E-06	3.6E-08	0.27	0.064	0.020	0.051	0.035	UAL	1.5E-04	6.1E-06	1.9E-07	0.72	0.051	0.040	0.459	0.103
GILD	3.8E-05	3.4E-06	2.5E-07	0.57	0.131	0.076	0.239	0.100	ULTA	7.9E-04	4.8E-05	2.8E-07	0.64	0.029	0.009	0.239	0.161
GOOGL	9.6E-04	1.0E-04	6.3E-06	0.60	0.079	0.048	0.305	0.117	VRSK	1.4E-04	3.2E-06	2.0E-07	0.66	0.029	0.022	0.498	0.124
GOOG	1.3E-03	6.3E-05	6.4E-07	0.47	0.066	0.039	0.305	0.124	VRSN	3.1E-04	4.0E-05	3.9E-07	0.58	0.027	0.020	0.459	0.124
HAS	2.3E-04	1.7E-05	4.6E-07	0.63	0.045	0.027	0.221	0.085	VRTX	4.3E-04	2.3E-05	1.1E-06	0.69	0.043	0.030	0.359	0.111
HSIC	8.1E-05	7.5E-06	2.8E-07	0.50	0.044	0.027	0.331	0.125	WBA	6.0E-05	6.2E-06	4.7E-07	0.57	0.087	0.059	0.423	0.136
IDXX	6.9E-04	1.0E-04	2.6E-07	0.59	0.016	0.011	0.331	0.093	WDAY	2.9E-04	1.2E-05	2.4E-07	0.55	0.027	0.020	0.459	0.129
ILMN	7.2E-04	5.4E-05	7.8E-07	0.65	0.026	0.019	0.423	0.119	WDC	1.1E-04	2.0E-06	3.7E-07	0.61	0.078	0.058	0.459	0.115
INCY	2.7E-04	9.5E-06	3.2E-07	0.64	0.033	0.024	0.359	0.095	WLTW	6.2E-04	1.1E-04	4.0E-07	0.52	0.014	0.011	0.635	0.137
INTC	6.0E-06	8.7E-07	3.1E-08	0.24	0.223	0.016	0.015	0.014	WYNN	1.3E-04	9.8E-06	5.5E-07	0.68	0.061	0.035	0.203	0.087
INTU	2.8E-04	4.0E-05	5.8E-07	0.64	0.037	0.026	0.359	0.104	XEL	3.0E-05	4.8E-06	6.1E-08	0.39	0.066	0.013	0.011	0.009
ISRG	2.0E-03	1.5E-04	8.5E-07	0.67	0.023	0.016	0.331	0.095	XLNX	8.7E-05	4.2E-06	5.1E-07	0.64	0.067	0.045	0.282	0.091

Table 4.8: Real data testing results (total execution cost) for NASDAQ 100 stocks.

Ticker	LHH	LH	LO	Ticker	LHH	LH	LO	Ticker	LHH	LH	LO
					ph = 0.5						
AAL	**0.7**	0.8	2.0	EXPE	6.0	7.7	18.7	NFLX	10.9	17.3	81.1
AAPL	2.3	**1.5**	11.3	FAST	**1.7**	2.0	6.4	NTAP	**1.4**	1.7	7.7
ADBE	**7.3**	9.3	16.6	FB	2.9	**2.5**	30.5	NTES	**35.3**	38.9	38.9
ADI	**1.0**	1.6	6.1	FISV	**8.9**	9.4	14.2	NVDA	**8.8**	10.3	28.4
ADP	**1.3**	1.5	2.1	FOX	0.8	**0.8**	1.4	NXPI	**2.5**	2.9	3.0
ADSK	**3.6**	5.3	18.9	FOXA	**0.9**	1.0	2.2	ORLY	**20.5**	22.7	23.0
ALGN	**30.0**	33.6	47.0	GILD	**1.4**	1.7	10.3	PAYX	**1.1**	1.4	4.2
ALXN	**5.7**	7.6	16.6	GOOG	**79.0**	102.4	111.5	PCAR	**1.9**	2.1	5.8
AMAT	2.6	**1.4**	3.0	GOOGL	**74.3**	101.7	338.6	PEP	**1.6**	1.8	4.0
AMD	0.4	**0.3**	1.0	HAS	**8.2**	10.1	15.0	PYPL	2.3	**2.1**	10.2
AMGN	**7.9**	8.6	31.7	HSIC	**2.7**	2.9	5.9	QCOM	**1.1**	1.4	4.5
AMZN	**160.9**	182.2	435.6	IDXX	**11.9**	12.8	12.9	REGN	**42.8**	44.3	53.2
ASML	9.4	**9.1**	9.6	ILMN	**13.0**	14.1	18.7	ROST	**2.5**	2.6	7.4
ATVI	**1.3**	1.9	14.1	INCY	**3.7**	4.0	7.1	SBUX	**1.1**	1.6	2.4
AVGO	**14.1**	20.1	22.1	INTC	4.6	**2.1**	4.8	SIRI	**0.0**	**0.0**	**0.0**
BIDU	**9.5**	11.1	12.9	INTU	**10.0**	10.6	14.4	SNPS	**1.4**	1.5	5.6
BIIB	**29.6**	35.8	37.8	ISRG	**36.3**	39.6	39.9	SWKS	**7.0**	8.1	12.8
BMRN	**3.0**	3.4	3.9	JBHT	**4.9**	5.1	7.4	SYMC	**0.7**	0.9	1.1
CDNS	**1.1**	1.2	2.5	JD	1.1	**0.7**	1.2	TMUS	**0.7**	0.8	2.2
CELG	**1.5**	2.0	10.3	KHC	**0.8**	1.0	4.9	TSLA	**16.1**	24.5	36.9
CERN	**1.3**	1.5	3.4	KLAC	**1.1**	1.3	11.2	TTWO	**8.2**	11.4	29.5
CHKP	**2.5**	2.7	6.8	LBTYA	**2.1**	2.5	2.8	TXN	**1.9**	2.5	16.0
CHTR	**19.2**	26.7	27.8	LBTYK	1.4	**1.1**	1.8	UAL	**0.5**	0.6	1.7
CMCSA	3.2	**1.8**	3.4	LRCX	**8.4**	11.9	36.3	ULTA	**13.8**	22.3	22.1
COST	**10.0**	11.5	17.4	LULU	**1.7**	2.8	7.3	VRSK	**1.0**	1.1	2.0
CSCO	6.0	**2.9**	5.5	MAR	**5.8**	6.7	25.8	VRSN	**5.7**	6.2	7.4
CSX	**1.7**	2.2	7.8	MCHP	**5.5**	7.1	21.1	VRTX	**6.5**	7.9	23.0
CTAS	**9.1**	9.2	10.9	MDLZ	**1.0**	1.3	1.7	WBA	**1.7**	1.9	12.3
CTRP	**0.4**	0.4	2.0	MELI	**28.8**	29.2	29.9	WDAY	**4.5**	5.2	6.6
CTSH	**4.2**	5.0	8.3	MNST	**1.3**	1.3	3.1	WDC	**0.5**	0.6	6.6
CTXS	**3.1**	3.2	5.3	MSFT	3.0	**2.1**	15.4	WLTW	**13.4**	13.5	13.9
DLTR	**3.7**	4.5	10.3	MU	3.6	**1.4**	3.3	WYNN	**5.8**	8.5	21.2
EA	**2.5**	3.3	18.7	MXIM	**2.3**	3.3	7.7	XEL	2.0	**1.4**	1.9
EBAY	**1.9**	2.6	3.6	MYL	**0.4**	0.7	1.4	XLNX	**1.0**	1.2	8.2
					ph = 0.2						
AAL	**1.5**	1.5	2.0	EXPE	12.6	12.5	18.7	NFLX	27.8	32.8	81.1
AAPL	6.1	**3.3**	11.3	FAST	4.0	**4.0**	6.4	NTAP	4.1	**3.5**	7.7
ADBE	**13.4**	14.7	16.6	FB	7.2	**6.1**	30.5	NTES	**35.4**	39.0	38.9
ADI	**2.5**	3.0	6.1	FISV	**12.0**	12.9	14.2	NVDA	**18.1**	19.1	28.4
ADP	**1.8**	2.0	2.1	FOX	**1.2**	1.3	1.4	NXPI	**2.6**	3.0	3.0
ADSK	**9.3**	11.6	18.9	FOXA	**1.7**	1.7	2.2	ORLY	**21.0**	23.0	23.0
ALGN	**40.6**	44.4	47.0	GILD	**3.4**	3.6	10.3	PAYX	**2.6**	2.7	4.2
ALXN	**12.9**	13.7	16.6	GOOG	**103.9**	113.9	111.5	PCAR	**3.9**	4.1	5.8
AMAT	4.3	**2.4**	3.0	GOOGL	**170.4**	185.5	338.6	PEP	**2.9**	3.1	4.0
AMD	0.5	**0.4**	1.0	HAS	**12.9**	14.4	15.0	PYPL	5.5	**4.5**	10.2
AMGN	19.1	**17.7**	31.7	HSIC	4.7	**4.6**	5.9	QCOM	**2.4**	2.5	4.5
AMZN	**314.9**	315.2	435.6	IDXX	**12.0**	12.9	12.9	REGN	52.0	**51.1**	53.2
ASML	10.3	**9.5**	9.6	ILMN	**17.1**	17.9	18.7	ROST	5.3	**4.9**	7.4
ATVI	**3.2**	3.7	14.1	INCY	**6.0**	**6.0**	7.1	SBUX	**1.8**	2.2	2.4
AVGO	**19.6**	23.9	22.1	INTC	6.8	**3.5**	4.8	SIRI	**0.0**	**0.0**	**0.0**
BIDU	**12.3**	12.8	12.9	INTU	**13.5**	13.7	14.4	SNPS	**3.3**	3.3	5.6
BIIB	**32.8**	37.5	37.8	ISRG	**37.0**	40.0	39.9	SWKS	**11.8**	12.1	12.8
BMRN	**3.5**	3.9	3.9	JBHT	**6.6**	6.8	7.4	SYMC	**1.0**	1.1	1.1
CDNS	**1.9**	2.0	2.5	JD	1.6	**1.1**	1.2	TMUS	1.5	**1.5**	2.2
CELG	**3.8**	4.1	10.3	KHC	**2.0**	2.2	4.9	TSLA	**30.7**	35.1	36.9
CERN	2.6	**2.6**	3.4	KLAC	6.0	**5.5**	11.2	TTWO	**19.9**	21.6	29.5
CHKP	5.1	**4.7**	6.8	LBTYA	**2.5**	2.7	2.8	TXN	**4.8**	5.2	16.0
CHTR	**23.7**	28.8	27.8	LBTYK	2.0	**1.6**	1.8	UAL	**1.1**	1.2	1.7
CMCSA	4.7	**2.8**	3.4	LRCX	**21.7**	23.4	36.3	ULTA	**14.1**	22.3	22.1
COST	16.8	**16.2**	17.4	LULU	**4.6**	5.3	7.3	VRSK	**1.7**	1.8	2.0
CSCO	8.3	**4.4**	5.5	MAR	13.6	**13.0**	25.8	VRSN	**6.8**	7.1	7.4
CSX	**4.1**	4.3	7.8	MCHP	**11.5**	11.9	21.1	VRTX	**15.9**	16.2	23.0
CTAS	10.7	**10.4**	10.9	MDLZ	**1.4**	1.7	1.7	WBA	4.4	**4.2**	12.3
CTRP	**1.0**	**1.0**	2.0	MELI	**29.3**	29.8	29.9	WDAY	**5.8**	6.4	6.6
CTSH	**7.5**	7.7	8.3	MNST	2.4	**2.3**	3.1	WDC	**1.5**	1.6	6.6
CTXS	4.8	**4.6**	5.3	MSFT	12.1	**4.0**	15.4	WLTW	**13.6**	13.8	13.9
DLTR	**7.7**	8.3	10.3	MU	4.9	**2.3**	3.3	WYNN	**12.8**	14.8	21.2
EA	**6.6**	7.1	18.7	MXIM	**4.7**	5.8	7.7	XEL	2.3	**1.8**	1.9
EBAY	**2.7**	3.4	3.6	MYL	**0.9**	1.3	1.4	XLNX	**2.9**	3.0	8.2

Table 4.9: Real data testing results (total execution cost) for 66 testing sessions.

Time	Strategy	2018-02-01	2018-02-02	2018-02-05	2018-02-06	2018-02-07	2018-02-08	2018-02-09	2018-02-12	2018-02-13	2018-02-14	2018-02-15
						ph = 0.5						
10:00	LHH	**7.3**	4.5	4.6	4.1	5.1	**5.8**	4.1	4.3	**6.2**	4.5	5.1
	LH	7.4	**3.6**	**3.8**	**3.0**	**4.7**	5.9	**3.2**	**3.4**	6.7	**3.4**	**4.8**
	LO	17.0	10.0	11.5	7.7	13.7	15.1	10.6	10.6	15.2	12.1	12.2
11:00	LHH	**13.0**	7.4	**13.0**	5.2	5.2	**9.7**	4.4	7.0	**15.1**	**7.6**	4.7
	LH	15.4	**7.4**	14.4	**4.9**	**5.1**	9.9	**3.6**	**6.6**	18.8	7.8	**4.1**
	LO	27.3	19.6	26.0	14.3	14.7	18.5	9.8	14.2	34.2	21.5	13.0
12:00	LHH	**20.2**	**9.0**	**20.0**	4.9	**17.4**	**9.1**	4.8	6.1	**24.1**	**10.6**	**22.8**
	LH	24.1	10.5	25.2	**4.0**	22.5	10.9	**4.1**	**5.7**	32.7	14.4	28.7
	LO	38.2	20.8	40.4	14.8	37.9	21.0	12.1	16.5	50.3	25.9	44.6
13:00	LHH	**6.6**	**19.8**	**7.7**	**7.3**	5.3	4.2	5.3	**13.2**	**20.2**	**12.1**	**6.5**
	LH	8.7	24.9	9.2	8.6	**4.6**	**3.2**	**4.9**	16.6	29.8	19.0	6.9
	LO	17.1	35.8	22.3	18.8	15.7	11.5	12.4	31.1	45.1	29.2	15.9
14:00	LHH	**11.3**	**9.7**	**11.4**	**12.4**	**8.6**	**21.1**	3.5	**10.1**	**13.6**	**9.8**	**12.5**
	LH	13.3	11.5	15.3	16.6	11.5	24.4	**2.4**	13.6	17.6	13.9	17.2
	LO	23.8	24.7	28.3	29.8	24.7	39.3	6.6	27.7	36.9	25.4	31.8
15:00	LHH	**9.8**	6.1	3.7	3.6	**6.6**	**6.9**	4.3	**5.8**	**11.4**	**6.5**	**6.3**
	LH	13.0	**5.1**	**2.6**	**2.6**	9.2	7.8	**3.6**	6.9	16.4	8.0	7.0
	LO	25.3	14.8	8.2	8.0	19.8	14.9	11.4	16.8	30.4	17.0	15.9
						ph = 0.2						
10:00	LHH	11.6	7.4	7.9	6.3	8.4	9.8	7.0	7.6	**10.0**	7.8	8.8
	LH	**11.4**	**6.1**	**6.3**	**5.1**	**7.5**	**9.2**	**5.6**	**6.0**	10.8	**6.2**	**7.9**
	LO	17.0	10.0	11.5	7.7	13.7	15.1	10.6	10.6	15.2	12.1	12.2
11:00	LHH	**18.8**	12.2	**19.0**	8.9	**8.9**	13.9	7.2	10.3	**25.6**	13.9	8.5
	LH	20.7	**11.4**	19.2	**8.2**	9.0	**13.4**	**6.4**	**9.2**	26.3	**12.1**	**7.4**
	LO	27.3	19.6	26.0	14.3	14.7	18.5	9.8	14.2	34.2	21.5	13.0
12:00	LHH	**29.9**	**14.4**	**31.7**	9.4	**28.4**	**15.0**	8.5	9.7	**36.1**	**19.1**	**34.2**
	LH	31.8	15.1	33.5	**7.1**	30.5	15.8	**6.9**	**9.4**	41.8	21.2	37.2
	LO	38.2	20.8	40.4	14.8	37.9	21.0	12.1	16.5	50.3	25.9	44.6
13:00	LHH	**10.9**	**28.9**	**13.8**	**12.2**	9.7	7.6	8.6	**19.5**	**31.0**	**19.4**	**11.1**
	LH	13.0	33.1	14.5	13.0	**8.1**	**5.7**	**8.2**	21.8	37.7	26.8	11.5
	LO	17.1	35.8	22.3	18.8	15.7	11.5	12.4	31.1	45.1	29.2	15.9
14:00	LHH	**17.4**	**15.9**	**19.6**	**20.9**	**16.5**	31.2	5.5	**17.5**	**23.5**	**17.7**	**24.0**
	LH	19.0	16.6	21.3	25.3	18.0	**29.6**	**4.2**	20.2	25.6	20.3	24.3
	LO	23.8	24.7	28.3	29.8	24.7	39.3	6.6	27.7	36.9	25.4	31.8
15:00	LHH	**16.2**	9.4	6.2	5.8	**12.6**	**10.3**	7.7	**10.4**	**21.4**	**11.3**	**10.9**
	LH	18.1	**7.7**	**4.5**	**4.5**	14.9	11.3	**6.3**	11.0	25.5	13.3	11.1
	LO	25.3	14.8	8.2	8.0	19.8	14.9	11.4	16.8	30.4	17.0	15.9

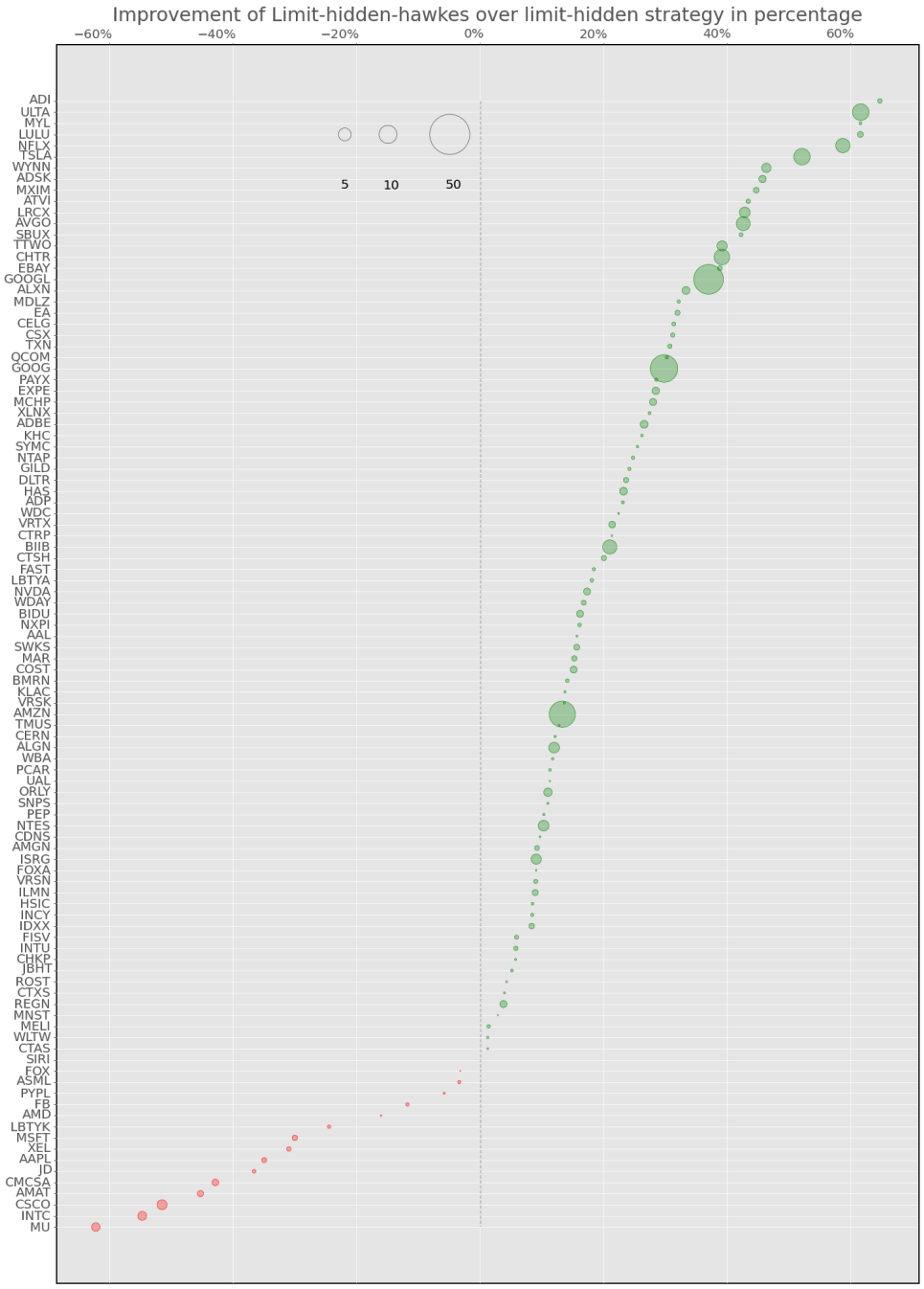

Figure 4.11: Comparison of average total execution cost over 66 testing sessions between the Limit-Hidden-Hawkes and Limit-Hidden strategies for each ticker. The horizontal axis represents the relative performance calculated as $(LH - LHH)/LHH$. The radius of each circle represents the absolute amount of the cost difference $|LH - LHH|$. Circles are filled in green if $LH - LHH > 0$ and red otherwise. The fill probability of hidden order is $p_h = 0.5$.

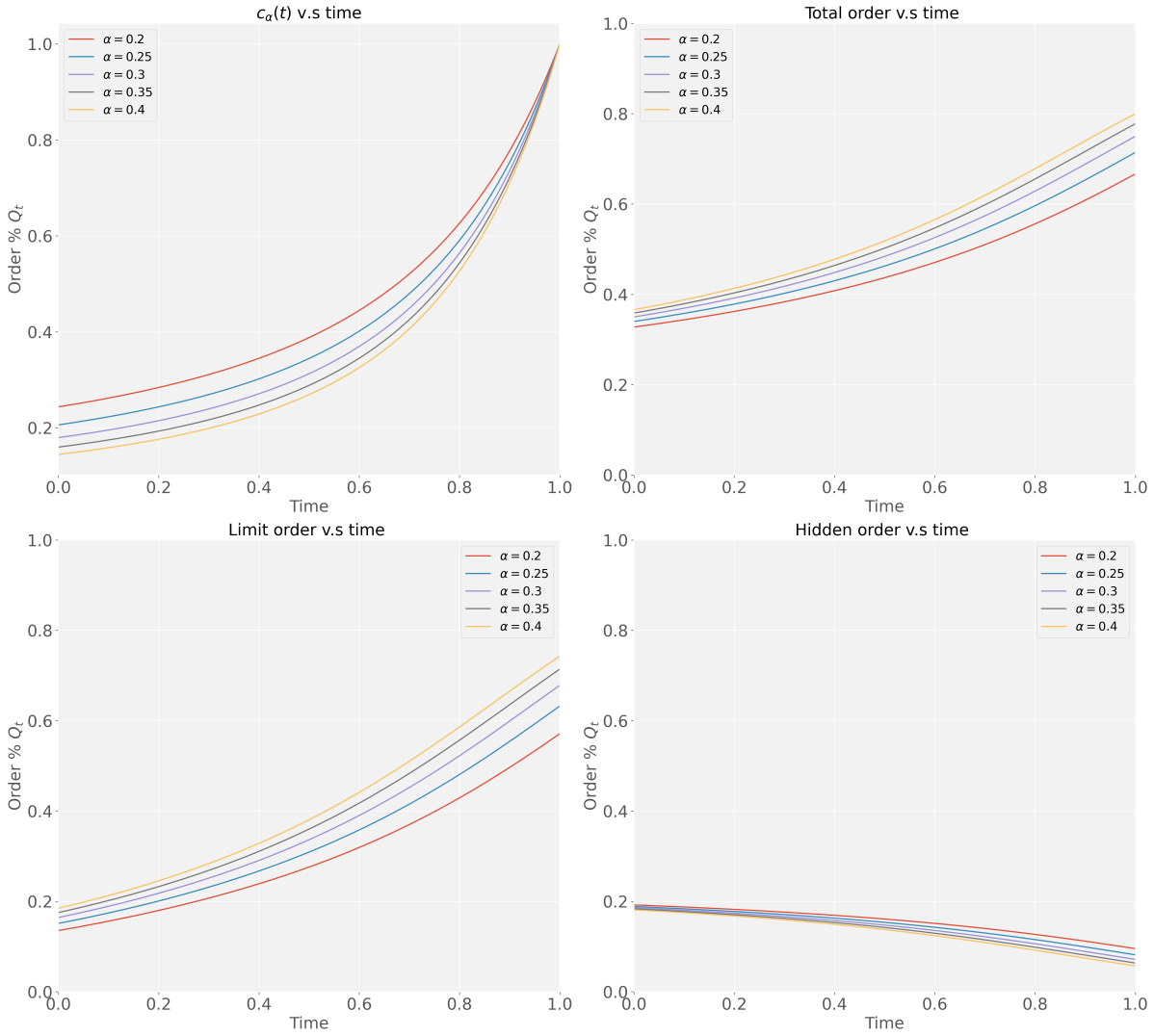

Figure 4.12: $\frac{a|b}{c|d}$ Optimal liquidation strategy and price impact of market orders α: a: c_α, b: total order size $L_t + H_t$, c: L_t, d: H_t.

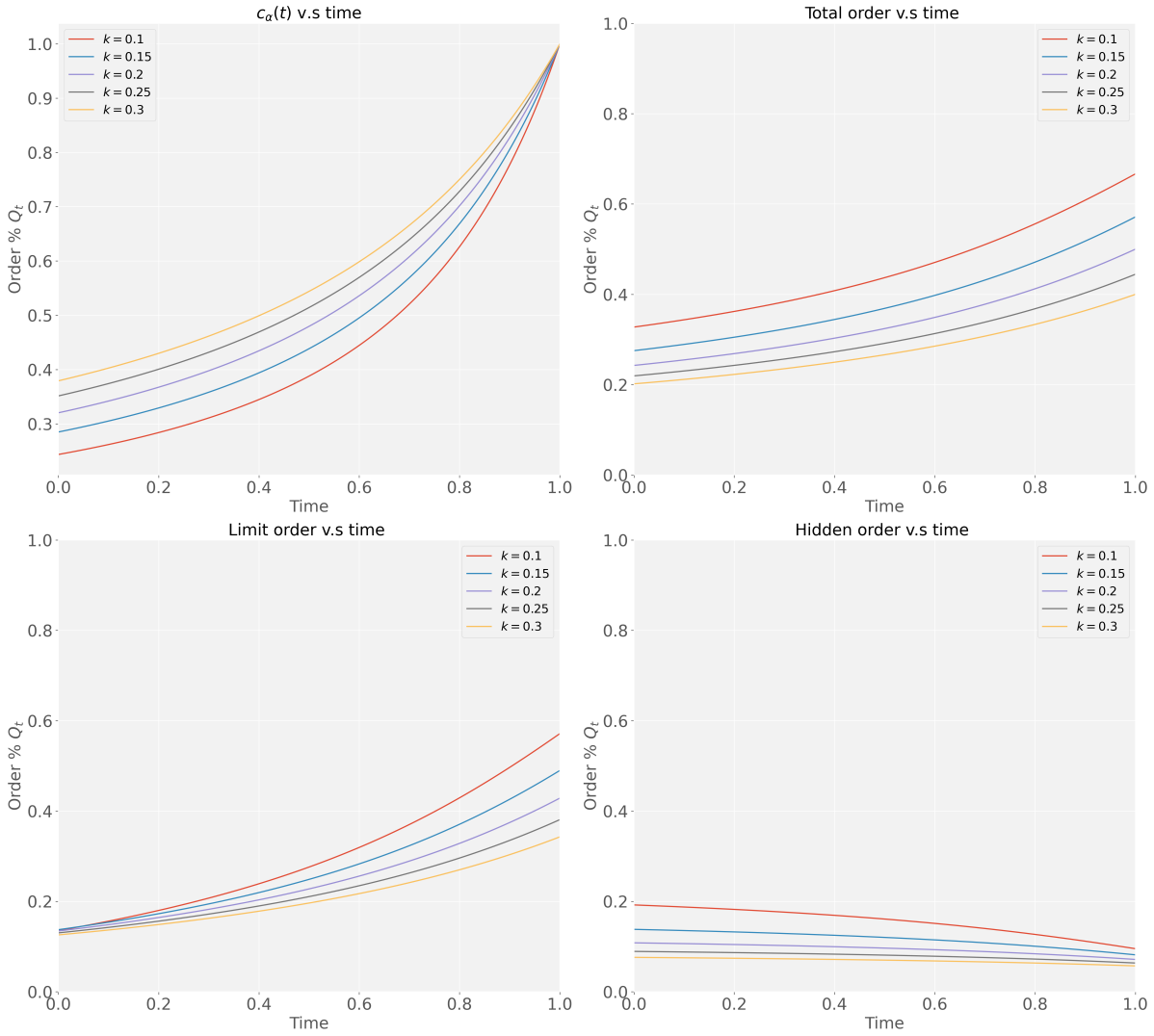

Figure 4.13: $\frac{a|b}{c|d}$ Optimal liquidation strategy and immediate execution cost coefficient k: a: c_α, b: total order size $L_t + H_t$, c: L_t, d: H_t.

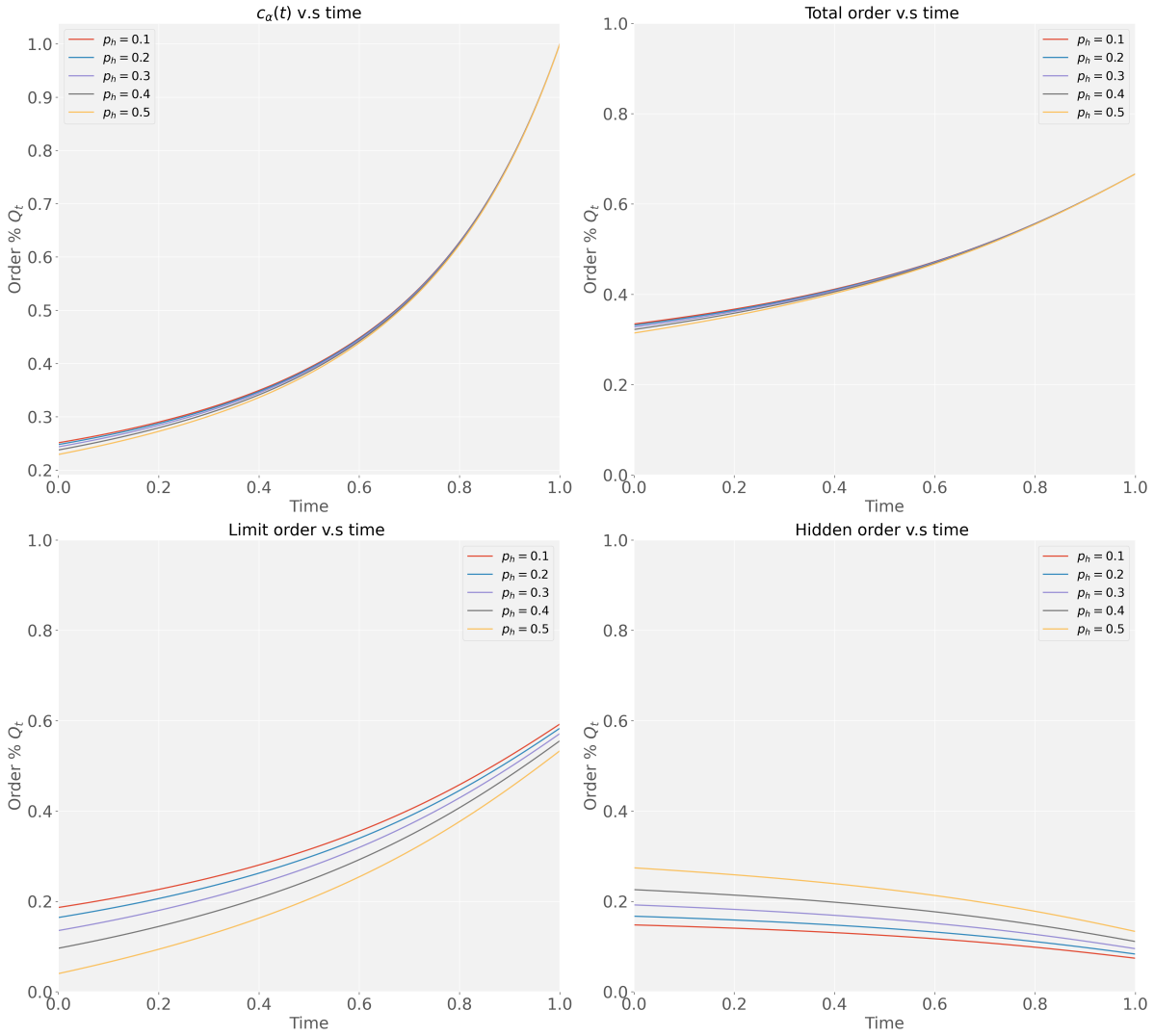

Figure 4.14: $\frac{a|b}{c|d}$ Optimal liquidation strategy and hidden order fill probability p_h: a: c_α, b: total order size $L_t + H_t$, c: L_t, d: H_t.

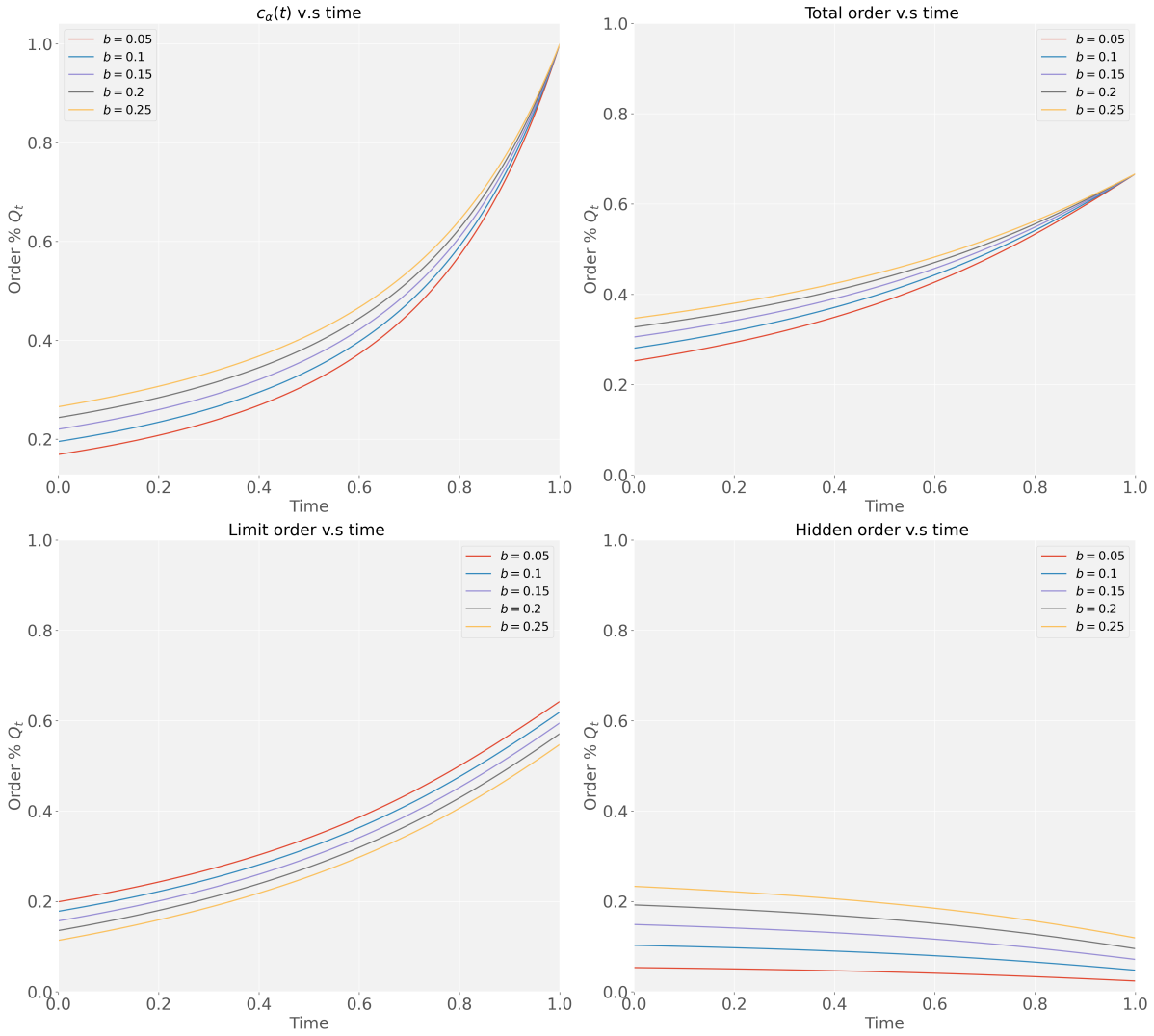

Figure 4.15: $\frac{a|b}{c|d}$ Optimal liquidation strategy and exposure risk coefficient b: a: c_α, b: total order size $L_t + H_t$, c: L_t, d: H_t.

References

Abergel, F., Anane, M., Chakraborti, A., Jedidi, A., & Toke, I. M. (2016). *Limit order books*. Cambridge University Press.

Abergel, F., & Jedidi, A. (2013). A mathematical approach to order book modeling. *International Journal of Theoretical and Applied Finance*, *16*(05), 1350025.

Ackermann, J., Kruse, T., & Urusov, M. (2021). Càdlàg semimartingale strategies for optimal trade execution in stochastic order book models. *Finance and Stochastics*, *25*(4), 757–810.

Agliardi, R. (2016). Optimal hedging through limit orders. *Stochastic Models*, *32*(4), 593–605.

Aitken, M. J., Berkman, H., & Mak, D. (2001). The use of undisclosed limit orders on the australian stock exchange. *Journal of Banking & Finance*, *25*(8), 1589–1603.

Aït-Sahalia, Y., Cacho-Diaz, J., & Laeven, R. J. (2015). Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics*, *117*(3), 585–606.

Aït-Sahalia, Y., & Jacod, J. (2014). *High-frequency financial econometrics*. Princeton, NJ: Princeton University Press.

Alfonsi, A., & Blanc, P. (2016a). Dynamic optimal execution in a mixed-market-impact Hawkes price model. *Finance and Stochastics*, *20*(1), 183–218.

Alfonsi, A., & Blanc, P. (2016b). Extension and calibration of a Hawkes-based optimal execution model. *Market Microstructure and Liquidity*, *2*(02), 1650005.

Alfonsi, A., Fruth, A., & Schied, A. (2010). Optimal execution strategies in limit order books with general shape functions. *Quantitative Finance*, *10*(2), 143–157.

Almgren, R. (2012). Optimal trading with stochastic liquidity and volatility. *SIAM Journal on Financial Mathematics*, *3*(1), 163–181.

Almgren, R., & Chriss, N. (1999). Value under liquidation. *Risk*, *12*(12), 61–63.

Almgren, R., & Chriss, N. (2001). Optimal execution of portfolio transactions. *Journal of Risk*, *3*, 5–40.

Almgren, R., Thum, C., Hauptmann, E., & Li, H. (2005). Direct estimation of equity market impact. *Risk*, *July*.

Almgren, R. F. (2003). Optimal execution with nonlinear impact functions and trading-enhanced risk. *Applied Mathematical Finance*, *10*(1), 1–18.

Andersen, T. G., Bondarenko, O., Kyle, A. S., & Obizhaeva, A. A. (2018). Intraday trading invariance in the e-mini S&P 500 futures market. *Anna A., Intraday Trading Invariance in the E-Mini S&P*, *500*.

Bacry, E., Delattre, S., Hoffmann, M., & Muzy, J.-F. (2013). Some limit theorems for hawkes processes and application to financial statistics. *Stochastic Processes and their Applications*, *123*(7), 2475–2499.

Bacry, E., Iuga, A., Lasnier, M., & Lehalle, C.-A. (2015). Market impacts and the life cycle of investors orders. *Market Microstructure and Liquidity*, *1*(02), 1550009.

Bacry, E., Mastromatteo, I., & Muzy, J.-F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, *1*(01), 1550005.

Bank, P., Ekren, I., & Muhle-Karbe, J. (2021). Liquidity in competitive dealer markets. *Mathematical Finance*, *31*(3), 827–856.

Barles, G., & Soner, H. M. (1998). Option pricing with transaction costs and a nonlinear black-scholes equation. *Finance and Stochastics*, *2*(4), 369–397.

Bates, D. S. (2019). How crashes develop: intradaily volatility and crash evolution. *The Journal of Finance*, *74*(1), 193–238.

Becherer, D., Bilarev, T., & Frentrup, P. (2018). Optimal liquidation under stochastic liquidity. *Finance and Stochastics*, *22*(1), 39–68.

Belak, C., Muhle-Karbe, J., & Ou, K. (2020). Optimal trading with general signals and liquidation in target zone models. *Market Microstructure and Liquidity*, *4*(3),

1950010.

Bernhardt, D., Seiler, P., & Taub, B. (2010). Speculative dynamics. *Economic Theory*, *44*(1), 1–52.

Bershova, N., & Rakhlin, D. (2013). The non-linear market impact of large trades: Evidence from buy-side order flow. *Quantitative Finance*, *13*(11), 1759–1778.

Bertsimas, D., & Lo, A. W. (1998). Optimal control of execution costs. *Journal of Financial Markets*, *1*(1), 1–50.

Biais, B., Hillion, P., & Spatt, C. (1995). An empirical analysis of the limit order book and the order flow in the Paris bourse. *The Journal of Finance*, *50*(5), 1655–1689.

Black, F. (1986). Noise. *Journal of Finance*, *41*(3), 528–543.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *The Journal of Political Economy*, *81*(3), 637-654.

Bouchaud, J., Farmer, J., & Lillo, F. (2009). How markets slowly digest changes in supply and demand. In T. Hens & K. R. Schenk-Hoppé (Eds.), *Handbook of financial markets: Dynamics and evolution* (pp. 57–160). North Holland.

Bouchaud, J., Gefen, Y., Potters, M., & Wyart, M. (2004). Fluctuations and response in financial markets: The subtle nature of random price changes. *Quantitative Finance*, *4*(2), 176–190.

Bouchaud, J., Kockelkoren, J., & Potters, M. (2006). Random walks, liquidity molasses and critical response in financial markets. *Quantitative Finance*, *6*(2), 115–123.

Bouchaud, J.-P., Bonart, J., Donier, J., & Gould, M. (2018). *Trades, quotes and prices: financial markets under the microscope*. Cambridge University Press.

Boulatov, A., & George, T. J. (2013). Hidden and displayed liquidity in securities markets with informed liquidity providers. *The Review of Financial Studies*, *26*(8), 2096–2137.

Bucci, F., Benzaquen, M., Lillo, F., & Bouchaud, J. (2015). Slow decay of impact in equity markets. *Market Microstructure and Liquidity*, *4*(3), 1950006.

Bucci, F., Benzaquen, M., Lillo, F., & Bouchaud, J.-P. (2019). Crossover from linear to square-root market impact. *Physical Review Letters*, *122*(10), 108302.

Bulthuis, B., Concha, J., Leung, T., & Ward, B. (2017). Optimal execution of limit and market orders with trade director, speed limiter, and fill uncertainty. *International Journal of Financial Engineering*, *4*(02n03), 1750020.

Busseti, E., & Lillo, F. (2012). Calibration of optimal execution of financial transactions in the presence of transient market impact. *Journal of Statistical Mechanics*, *2012*(09), P09010.

Buti, S., & Rindi, B. (2013). Undisclosed orders and optimal submission strategies in a limit order market. *Journal of Financial Economics*, *109*(3), 797–812.

Cao, C., Hansch, O., & Wang, X. (2009). The information content of an open limit-order book. *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, *29*(1), 16–41.

Carmona, R., & Leal, L. (2021). *Optimal execution with quadratic variation inventories.* (Preprint, `https://arxiv.org/abs/2104.14615`)

Carmona, R., & Webster, K. (2019). The self-financing equation in limit order book markets. *Finance and Stochastics*, *23*(3), 729–759.

Cartea, Á., Donnelly, R., & Jaimungal, S. (2018). Enhancing trading strategies with order book signals. *Applied Mathematical Finance*, *25*(1), 1–35.

Cartea, Á., Gan, L., & Jaimungal, S. (2019). Hedge and speculate: Replicating option payoffs with limit and market orders. *SIAM Journal on Financial Mathematics*, *10*(3), 790–814.

Cartea, Á., & Jaimungal, S. (2015a). Optimal execution with limit and market orders. *Quantitative Finance*, *15*(8), 1279–1291.

Cartea, Á., & Jaimungal, S. (2015b). Risk metrics and fine tuning of high-frequency trading strategies. *Mathematical Finance*, *25*(3), 576–611.

Cartea, Á., Jaimungal, S., & Penalva, J. (2015). *Algorithmic and high-frequency trading.* Cambridge University Press.

Cartea, Á., Jaimungal, S., & Ricci, J. (2014). Buy low, sell high: A high frequency trading perspective. *SIAM Journal on Financial Mathematics*, *5*(1), 415–444.

Cartea, Á., & Sánchez-Betancourt, L. (2023). Optimal execution with stochastic delay.

Finance Stoch., *27*(1), 1–47.

Cayé, T., & Muhle-Karbe, J. (2014). Liquidation with self-exciting price impact. *Mathematics and Financial Economics*, *10*, 15-28.

CBOE Insights. (2022). Hide-and-Seek: Hidden Liquidity on U.S. Exchanges. https://www.cboe.com/insights/posts/hide-and-seek-hidden-liquidity-on-u-s-exchanges/.

Cebiroğlu, G., & Horst, U. (2015). Optimal order display in limit order markets with liquidity competition. *Journal of Economic Dynamics and Control*, *58*, 81–100.

CFTC, S., & SEC, U. (2010). Findings regarding the market events of may 6, 2010. *Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*, *104*. https://www.sec.gov/news/studies/2010/marketevents-report.pdf.

Chen, J. (2021). Internalization. https://www.investopedia.com/terms/i/internalization.asp. (Accessed 10 November 2021)

Chen, T., Ludkovski, M., & Voß, M. (2022). *On parametric optimal execution and machine learning surrogates.* (Preprint, https://arxiv.org/abs/2204.08581)

Chen, Y., Gao, X., & Li, D. (2018). Optimal order execution using hidden orders. *Journal of Economic Dynamics and Control*, *94*, 89–116.

Chen, Y., Horst, U., & Hai Tran, H. (2019). *Portfolio liquidation under transient price impact – theoretical solution and implementation with 100 NASDAQ stocks.* (Preprint, https://arxiv.org/abs/1912.06426)

Cheng, X., Di Giacinto, M., & Wang, T.-H. (2017). Optimal execution with uncertain order fills in Almgren–Chriss framework. *Quantitative Finance*, *17*(1), 55–69.

Chordia, T., Roll, R., & Subrahmanyam, A. (2002). Order imbalance, liquidity, and market returns. *Journal of Financial Economics*, *65*(1), 111–130.

Christensen, H. L., & Woodmansey, R. (2013). Prediction of hidden liquidity in thelimit order book of globex futures. *The Journal of Trading*, *8*(3), 68–95.

Collin-Dufresne, P., & Fos, V. (2016). Insider trading, stochastic liquidity, and equilibrium prices. *Econometrica*, *84*(4), 1441–1475.

Committee on the Global Financial System. (2014). Market-making and proprietary trading: Industry trends, drivers and policy implications. *CGFS Papers*, *52*.

Cont, R., Cucuringu, M., & Zhang, C. (2021). *Price impact of order flow imbalance: Multi-level, cross-sectional and forecasting.* (Preprint, `https://arxiv.org/abs/2112.13213`)

Cont, R., & De Larrard, A. (2013). Price dynamics in a markovian limit order market. *SIAM Journal on Financial Mathematics*, *4*(1), 1–25.

Cont, R., & Kukanov, A. (2017). Optimal order placement in limit order markets. *Quantitative Finance*, *17*(1), 21–39.

Cont, R., Kukanov, A., & Stoikov, S. (2014). The price impact of order book events. *Journal of Financial Econometrics*, *12*(1), 47–88.

Cont, R., Stoikov, S., & Talreja, R. (2010). A stochastic model for order book dynamics. *Operations Research*, *58*(3), 549–563.

Crisafi, M. A., & Macrina, A. (2016). Simultaneous trading in 'lit'and dark pools. *International Journal of Theoretical and Applied Finance*, *19*(08), 1650055.

Curato, G., Gatheral, J., & Lillo, F. (2017). Optimal execution with non-linear transient market impact. *Quantitative Finance*, *17*(1), 41–54.

Degryse, H., Karagiannis, N., Tombeur, G., & Wuyts, G. (2021). Two shades of opacity: Hidden orders and dark trading. *Journal of Financial Intermediation*, 100919.

Easley, D., & O'hara, M. (1987). Price, trade size, and information in securities markets. *Journal of Financial Economics*, *19*(1), 69–90.

El Euch, O., Rosenbaum, M., et al. (2018). Perfect hedging in rough Heston models. *Annals of Applied Probability*, *28*(6), 3813–3856.

Ellersgaard, S., & Tegnér, M. (2017). Optimal hedge tracking portfolios in a limit order book. *Market Microstructure Liquidity*, *3*(02), 1850003.

Esser, A., & Mönch, B. (2007). The navigation of an iceberg: The optimal use of hidden orders. *Finance Research Letters*, *4*(2), 68–81.

Finextra. (2016). Central Risk Books - the new black for capital markets. `https://www.finextra.com/blogposting/13121/central-risk-books--`

-the-new-black-for-capital-markets. (Accessed 10 November 2021)

Fleming, W. H., & Soner, H. M. (2006). *Controlled markov processes and viscosity solutions* (Vol. 25). Springer Science & Business Media.

Frazzini, A., Israel, R., & Moskowitz, T. J. (2018). *Trading costs.* (Preprint, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3229719)

Fruth, A., Schöneborn, T., & Urusov, M. (2013). Optimal trade execution and price manipulation in order books with time-varying liquidity. *Mathematical Finance*, *24*(4), 651–695.

Fruth, A., Schöneborn, T., & Urusov, M. (2019). Optimal trade execution and price manipulation in order books with stochastic liquidity. *Mathematical Finance*, *29*(2), 507–541.

Fu, G., Horst, U., & Xia, X. (2022). Portfolio liquidation games with self-exciting order flow. *Mathematical Finance*, *32*(4), 1020-1065.

Gabaix, X., Gopikrishnan, P., Plerou, V., & Stanley, H. E. (2006). Institutional investors and stock market volatility. *Quarterly Journal of Economics*, *121*(2), 461–504.

Gârleanu, N., & Pedersen, L. H. (2016). Dynamic portfolio choice with frictions. *Journal of Economic Theory*, *165*, 487–516.

Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, *14*(1), 71–100.

Graewe, P., & Horst, U. (2017). Optimal trade execution with instantaneous price impact and stochastic resilience. *SIAM Journal on Control and Optimization*, *55*(6), 3707–3725.

Guéant, O. (2016). *The financial mathematics of market liquidity: From optimal execution to market making* (Vol. 33). CRC Press.

Guéant, O., Lehalle, C.-A., & Fernandez-Tapia, J. (2012). Optimal portfolio liquidation with limit orders. *SIAM Journal on Financial Mathematics*, *3*(1), 740–764.

Guilbaud, F., & Pham, H. (2013). Optimal high-frequency trading with limit and market orders. *Quantitative Finance*, *13*(1), 79–94.

Harris, L. (1997). Order exposure and parasitic traders. *University of Southern California working paper*.

Hasbrouck, J. (1991). Measuring the information content of stock trades. *Journal of Finance*, *46*(1), 179–207.

Hasbrouck, J., & Seppi, D. J. (2001). Common factors in prices, order flows, and liquidity. *Journal of Financial Economics*, *59*(3), 383–411.

Hautsch, N., & Huang, R. (2011). Limit order flow, market impact and optimal order sizes: evidence from nasdaq totalview-itch data. *SFB 649 Discussion Papers SFB649DP2011-056, Sonderforschungsbereich 649, Humboldt University, Berlin, Germany.*.

Hautsch, N., & Huang, R. (2012a). The market impact of a limit order. *Journal of Economic Dynamics and Control*, *36*(4), 501–522.

Hautsch, N., & Huang, R. (2012b). On the dark side of the market: identifying and analyzing hidden order placements. *Available at SSRN 2004231*.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, *58*(1), 83–90.

Herdegen, M., Muhle-Karbe, J., & Stebegg, F. (2022). Liquidity provision with adverse selection and inventory costs. *Mathematics of Operations Research*.

Hodges, S., & Neuberger, A. (1989). Optimal replication of contingent claims under transaction costs. *Review of Futures Markets*, *8*(2), 222-239.

Horst, U., & Naujokat, F. (2014). When to cross the spread? trading in two-sided limit order books. *SIAM Journal on Financial Mathematics*, *5*(1), 278–315.

Horst, U., & Paulsen, M. (2017). A law of large numbers for limit order books. *Mathematics of Operations Research*, *42*(4), 1280–1312.

Horst, U., & Xu, W. (2019). A scaling limit for limit order books driven by hawkes processes. *SIAM Journal on Financial Mathematics*, *10*(2), 350–393.

Horst, U., & Xu, W. (2022). The microstructure of stochastic volatility models with self-exciting jump dynamics. *The Annals of Applied Probability*, *32*(6), 4568-4610.

Huang, W., Lehalle, C.-A., & Rosenbaum, M. (2015). Simulating and analyzing order

book data: The queue-reactive model. *Journal of the American Statistical Association*, *110*(509), 107–122.

Huberman, G., & Stanzl, W. (2005). Optimal liquidity trading. *Review of Finance*, *9*(2), 165–200.

Ishimura, N. (2010). Remarks on the nonlinear Black-Scholes equations with the effect of transaction costs. *Asia-Pacific Financial Markets*, *17*(3), 241–259.

Jacod, J., & Protter, P. (2012). *Discretization of processes*. Berlin: Springer.

Jacod, J., & Shiryaev, A. (2003). *Limit theorems for stochastic processes* (second ed.). Berlin: Springer.

Jacquier, A., & Liu, H. (2018). Optimal liquidation in a level-i limit order book for large-tick stocks. *SIAM Journal on Financial Mathematics*, *9*(3), 875–906.

Jaisson, T., & Rosenbaum, M. (2015). Limit theorems for nearly unstable Hawkes processes. *Annals of Applied Probability*, *25*(2), 600–631.

Jusselin, P., & Rosenbaum, M. (2020). No-arbitrage implies power-law market impact and rough volatility. *Mathematical Finance*, *30*(4), 1309–1336.

Kolm, P., & Westray, N. (2021). What happened to the rest? A principled approach to clean-up costs in algorithmic trading. *Risk*, *August*.

Kratz, P., & Schöneborn, T. (2014). Optimal liquidation in dark pools. *Quantitative Finance*, *14*(9), 1519–1539.

Kratz, P., & Schöneborn, T. (2015). Portfolio liquidation in dark pools in continuous time. *Mathematical Finance*, *25*(3), 496–544.

Kühn, C., & Muhle-Karbe, J. (2015). Optimal liquidity provision. *Stochastic Processes and their Applications*, *125*(7), 2493–2515.

Kühn, C., & Stroh, M. (2010). Optimal portfolios of a small investor in a limit order market: a shadow price approach. *Mathematics and Financial Economics*, *3*(2), 45–72.

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society*, 1315–1335.

Kyle, A. S. (1989). Informed speculation with imperfect competition. *The Review of*

Economic Studies, *56*(3), 317–355.

Lehalle, C.-A., Mounjid, O., & Rosenbaum, M. (2021). Optimal liquidity-based trading tactics. *Stochastic Systems*, *11*(4), 368-390.

Lehalle, C.-A., & Neuman, E. (2019). Incorporating signals into optimal trading. *Finance and Stochastics*, *23*(2), 275–311.

Leland, H. E. (1985). Option pricing and replication with transactions costs. *The journal of finance*, *40*(5), 1283–1301.

Lieberman, G. M. (1996). *Second order parabolic differential equations*. Singapore: World Scientific.

Lillo, F., Farmer, J., & Mantegna, R. (2003). Econophysics: Master curve for price-impact function. *Nature*, *421*, 129–130.

Lo, A. W., & Wang, J. (2010). Stock market trading volume. In Y. Aït-Sahalia & L. P. Hansen (Eds.), *Handbook of financial econometrics: Applications* (pp. 241–342). Elsevier.

Loeb, T. F. (1983). Trading cost: The critical link between investment information and results. *Financial Analysts Journal*, *39*(3), 39–44.

Lorenz, C., & Schied, A. (2013). Drift dependence of optimal trade execution strategies under transient price impact. *Finance and Stochastics*, *17*(4), 743–770.

Madhavan, A., Richardson, M., & Roomans, M. (1997). Why do security prices change? A transaction-level analysis of NYSE stocks. *The Review of Financial Studies*, *10*(4), 1035–1064.

Mastromatteo, I., Benzaquen, M., Eisler, Z., & Bouchaud, J.-P. (2017). Trading lightly: Cross-impact and optimal portfolio execution. *Risk*, *July*.

Merton, R. C. (1973). Theory of rational option pricing. *The Bell Journal of economics and management science*, 141–183.

Min, S., Maglaras, C., & Moallemi, C. C. (2022). Cross-sectional variation of intraday liquidity, cross-impact, and their effect on portfolio execution. *Operations Research*, *70*(2), 830–846.

Moinas, S. (2005). Hidden limit orders and liquidity in limit order markets. *Available at*

SSRN 676564.

NASDAQ. (n.d.). NASDAQ Equities Market Data. `https://www.nasdaq.com/solutions/nasdaq-equities-market-data-solution`. (Accessed 9 November 2021)

Nasdaq. (2022). Nasdaq order types. `http://www.nasdaqtrader.com/trader.aspx?id=tradingusequities`.

Neuman, E., & Voß, M. (2016). Optimal portfolio liquidation in target zone models and catalytic superprocesses. *Finance and Stochastics*, *20*, 495-509.

Neuman, E., & Voß, M. (2020). Optimal signal-adaptive trading with temporary and transient price impact. *SIAM Journal on Financial Mathematics*, *13*(2), 551-574.

Neuman, E., & Voß, M. (2021). *Trading with the crowd.* (Preprint)

Neuman, E., & Voß, M. (2022). Optimal signal-adaptive trading with temporary and transient price impact. *SIAM Journal on Financial Mathematics*, *13*(2), 551-575.

New York Stock Exchange. (n.d.). Daily TAQ. `https://www.nyse.com/market-data/historical/daily-taq`. (Accessed 9 November 2021)

Obizhaeva, A., & Wang, J. (2013). Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets*, *16*(1), 1–32.

Osipovich, A. (2019, May). 'dark pools' draw more trading amid low volatility. *Wall Street Journal*.

Pardo, A., & Pascual, R. (2012). On the hidden side of liquidity. *The European Journal of Finance*, *18*(10), 949–967.

Patzelt, F., & Bouchaud, J.-P. (2018). Universal scaling and nonlinearity of aggregate price impact in financial markets. *Physical Review E*, *97*(1), 012304.

Schied, A., Strehle, E., & Zhang, T. (2017). High-frequency limit of Nash equilibria in a market impact game with transient impact. *SIAM Journal on Financial Mathematics*, *8*(1), 589–634.

Schneider, M., & Lillo, F. (2019). Cross-impact and no-dynamic-arbitrage. *Quantitative Finance*, *19*(1), 137–154.

Securities, U., & Commission, E. (2020). Staff Report On Algorithmic Trading in

U.S. Capital Markets. `www.sec.gov/files/Algo_Trading_Report_2020.pdf`. (Accessed 9 November 2021)

Tomas, M., Mastromatteo, I., & Benzaquen, M. (2022). How to build a cross-impact model from first principles: Theoretical requirements and empirical results. *Quantitative Finance*, *22*(6), 1017–1036.

Torre, N., & Ferrari, M. (1997). *Market impact model handbook, barra inc., berkeley (1997).*

Tóth, B., Lemperiere, Y., Deremble, C., De Lataillade, J., Kockelkoren, J., & Bouchaud, J.-P. (2011). Anomalous price impact and the critical nature of liquidity in financial markets. *Physical Review X*, *1*(2), 021006.

Tsoukalas, G., Wang, J., & Giesecke, K. (2019). Dynamic portfolio execution. *Management Science*, *65*(5), 2015–2040.

Tuttle, L. A. (2003). Hidden orders, trading costs and information. *Trading Costs and Information (November 29, 2003).*

US Securities and Exchange Commission. (2019). Hidden Volume Ratios. `https://www.sec.gov/marketstructure/research/data-highlight/hidden-volume-ratios`.

Velu, R., Hardy, M., & Nehren, D. (2020). *Algorithmic trading and quantitative strategies* (2nd ed.). Boca Raton, FL: CRC Press.

Vodret, M., Mastromatteo, I., Tóth, B., & Benzaquen, M. (2021). A stationary Kyle setup: Microfounding propagator models. *Journal of Statistical Mechanics*, *2021*(3), 033410.

Waelbroeck, H., Federspiel, F. J., Marchini, S., & Gomes, C. (2013). *Methods and systems related to securities trading.* (US Patent No. 433,645 B1)

Webster, K., Luo, Y., Alvarez, M. A., Jussa, J., Wang, S., Rohal, G., ... Zhao, G. (2015). *A portfolio manager's guidebook to trade execution.* (Deutsche Bank Markets Research, `https://www.slideshare.net/OlivierLedoit/a-portfolio-managers-guidebook-to-trade-execution`)

Wyart, M., Bouchaud, J.-P., Kockelkoren, J., Potters, M., & Vettorazzo, M. (2008).

Relation between bid–ask spread, impact and volatility in order-driven markets. *Quantitative Finance*, *8*(1), 41–57.

Zarinelli, E., Treccani, M., Farmer, J. D., & Lillo, F. (2015). Beyond the square root: Evidence for logarithmic dependence of market impact on size and participation rate. *Market Microstructure and Liquidity*, *1*(2), 1550004.

Zhou, W. (2012). Universal price impact functions of individual trades in an order-driven market. *Quantitative Finance*, *12*(8), 1253–1263.