Stochastic Models of Limit Order Markets

Arseniy Kukanov

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2013

ABSTRACT

Stochastic Models of Limit Order Markets

Arseniy Kukanov

During the last two decades most stock and derivatives exchanges in the world transitioned to electronic trading in limit order books, creating a need for a new set of quantitative models to describe these order-driven markets. This dissertation offers a collection of models that provide insight into the structure of modern financial markets, and can help to optimize trading decisions in practical applications.

In the first part of the thesis we study the dynamics of prices, order flows and liquidity in limit order markets over short timescales. We propose a stylized order book model that predicts a particularly simple linear relation between price changes and order flow imbalance, defined as a difference between net changes in supply and demand. The slope in this linear relation, called a price impact coefficient, is inversely proportional in our model to market depth - a measure of liquidity. Our empirical results confirm both of these predictions. The linear relation between order flow imbalance and price changes holds for time intervals between 50 milliseconds and 5 minutes. The inverse relation between the price impact coefficient and market depth holds on longer timescales. These findings shed a new light on intraday variations in market volatility. According to our model volatility fluctuates due to changes in market depth or in order flow variance. Previous studies also found a positive correlation between volatility and trading volume, but in order-driven markets prices are determined by the limit order book activity, so the association between trading volume and volatility is unclear. We show how a spurious correlation between these variables can indeed emerge in our linear model due to time aggregation of high-frequency data. Finally, we observe short-term positive autocorrelation in order flow imbalance and discuss an application of this variable as a measure of adverse selection in limit order executions. Our results suggest that monitoring recent order flow can improve the quality

of order executions in practice.

In the second part of the thesis we study the problem of optimal order placement in a fragmented limit order market. To execute a trade, market participants can submit limit orders or market orders across various exchanges where a stock is traded. In practice these decisions are influenced by sizes of order queues and by statistical properties of order flows in each limit order book, and also by rebates that exchanges pay for limit order submissions. We present a realistic model of limit order executions and formalize the search for an optimal order placement policy as a convex optimization problem. Based on this formulation we study how various factors determine investor's order placement decisions. In a case when a single exchange is used for order execution, we derive an explicit formula for the optimal limit and market order quantities. Our solution shows that the optimal split between market and limit orders largely depends on one's tolerance to execution risk. Market orders help to alleviate this risk because they execute with certainty. Correspondingly, we find that an optimal order allocation shifts to these more expensive orders when the execution risk is of primary concern, for example when the intended trade quantity is large or when it is costly to catch up on the quantity after limit order execution fails. We also characterize the optimal solution in the general case of simultaneous order placement on multiple exchanges, and show that it sets execution shortfall probabilities to specific threshold values computed with model parameters. Finally, we propose a non-parametric stochastic algorithm that computes an optimal solution by resampling historical data and does not require specifying order flow distributions. A numerical implementation of this algorithm is used to study the sensitivity of an optimal solution to changes in model parameters. Our numerical results show that order placement optimization can bring a substantial reduction in trading costs, especially for small orders and in cases when order flows are relatively uncorrelated across trading venues. The order placement optimization framework developed in this thesis can also be used to quantify the costs and benefits of financial market fragmentation from the point of view of an individual investor. For instance, we find that a positive correlation between order flows, which is empirically observed in a fragmented U.S. equity market, increases the costs of trading. As the correlation increases it may become more expensive to trade in a fragmented market than it is in a consolidated market.

In the third part of the thesis we analyze the dynamics of limit order queues at the best bid or ask of an exchange. These queues consist of orders submitted by a variety of market participants, yet existing order book models commonly assume that all orders have similar dynamics. In practice, some orders are submitted by trade execution algorithms in an attempt to buy or sell a certain quantity of assets under time constraints, and these orders are canceled if their realized waiting time exceeds a patience threshold. In contrast, high-frequency traders submit and cancel orders depending on the order book state and their orders are not driven by patience. The interaction between these two order types within a single FIFO queue leads bursts of order cancelations for small queues and anomalously long waiting times in large queues. We analyze a fluid model that describes the evolution of large order queues in liquid markets, taking into account the heterogeneity between order submission and cancelation strategies of different traders. Our results show that after a finite initial time interval, the queue reaches a specific structure where all orders from high-frequency traders stay in the queue until execution but most orders from execution algorithms exceed their patience thresholds and are canceled. This "order crowding" effect has been previously noted by participants in highly liquid stock and futures markets and was attributed to a large participation of high-frequency traders. In our model, their presence creates an additional workload, which increases queue waiting times for new orders. Our analysis of the fluid model leads to waiting time estimates that take into account the distribution of order types in a queue. These estimates are tested against a large dataset of realized limit order waiting times collected by a U.S. equity brokerage firm. The queue composition at a moment of order submission noticeably affects its waiting time and we find that assuming a single order type for all orders in the queue leads to unrealistic results. Estimates that assume instead a mix of heterogeneous orders in the queue are closer to empirical data. Our model for a limit order queue with heterogeneous order types also appears to be interesting from a methodological point of view. It introduces a new type of behavior in a queueing system where one class of jobs has state-dependent dynamics, while others are driven by patience. Although this model is motivated by the analysis of limit order books, it may find applications in studying other service systems with state-dependent abandonments.

Keywords: limit order markets, price impact, market liquidity, price formation, volatility modeling, high-frequency financial data, queueing models, fluid approximation, high-frequency trading, optimal order execution, order routing, fragmented markets, transaction costs, stochastic approximation.

Contents

List of Figures

List of Tables

Acknowledgments

This dissertation summarizes five years of studies and research, a journey that started with a seemingly simple question: "how do financial markets actually work?" Over the course of these years I met many exceptional scholars and inspiring individuals, to whom I owe most of what I currently know about the subject matter, and who profoundly influenced my scholarly and individual advancement.

First of all, I express my deepest gratitude to Professor Rama Cont for guiding me through this journey. His insightful advice helped me to reach an understanding of the current market microstructure literature, and to build an intuition for this complex subject. His continuous encouragement motivated me to explore new ideas and to constantly ask new questions. Sometimes seeking answers to these questions was not easy, yet Professor Cont always allowed me to work on my own ideas and to develop as an independent researcher, for which I will always be grateful. In times of need, Professor Cont genuinely and patiently supported me, and I thank him being available at these times regardless of his (often difficult) schedule. His scientific curiosity, intellectual breadth and truthfulness have set for me high standards of intellectual inquiry, standards to which I will continue to aspire.

During the latter stages this work, I received a tremendous support from my co-advisor, Professor Costis Maglaras, and I am truly grateful for his aid and his friendship. Our engaging discussions and our fruitful collaboration gave me the energy and inspiration necessary to complete this dissertation. There was not a single time that I left our regular meetings not brimming with excitement and enthusiasm to explore a new idea. The creativity, resourcefulness and spirit with which Professor Maglaras approaches new research questions truly inspired me and I keep wishing to have acquired a fraction of these qualities through our interaction.

This work also benefited from the feedback of Professors Jose Blanchet, Steven Kou and

Tim Leung. I kindly thank them for participating in my thesis defense committee, and for helping me to improve both the results developed in this work and their presentation. I also thank Doctor Timur Davis who carefully revised an early version of this thesis and helped to make its text more coherent and readable. All of the remaining mistakes are my own.

It is said that "a journey of a thousand miles begins with a single step", and my journey began under the guidance of Professor Vladimir Morozov, my undergraduate advisor at the Moscow State University. Although he did not directly participate in this thesis, it would not have been written had he not guided my first steps into the realm of mathematical finance. I am deeply indebted to Professor Morozov for igniting my interest in research, and for encouraging me to explore this subject further.

During my studies at the Columbia University, I had a privilege to interact with its exceptional faculty members - remarkable characters and brilliant thinkers whose lectures deepened my knowledge of optimization, probability, statistics and finance, and who set outstanding examples of personal and professional qualities. I thank Professors Andrew Ang, Jose Blanchet, Rama Cont, Emanuel Derman, Martin Haugh, Garud Iyengar, Michael Johannes, Donald Goldfarb, Ward Whitt, David Yao and Assaf Zeevi for their insightful and engaging courses. I am also sincerely grateful to Martin Haugh, Soulaymane Kachani, Andrei Kirilenko, Mike Lipkin, Ali Sadighian and Sasha Stoikov for their friendly advice, mentorship and support during my studies, all of which greatly helped in my academic and career development.

While working on this thesis, I had a rare opportunity to explore financial markets from a variety of angles - doing research as an academic, but also working as a practitioner at an investment bank, a regulatory agency and a financial exchange. First of all, I cordially thank my co-authors Rama Cont, Costis Maglaras and Sasha Stoikov for sharing their research insights and for being formidable and reliable colleagues. The benefit I received from being a part of the academic community is enormous, and I thank everybody with whom I was fortunate to interact - journal referees and editors for their valuable comments and criticisms, conference participants for engaging conversations and discussions, and all other colleagues in academia for helping me in my discoveries. I thank my colleagues at Cantror Fitzgerald and Morgan Stanley - Jaroslaw Labaziewicz, Jacob Loveless, Paul Schimmel,

Sasha Stoikov, Rolf Waeber, and others - for creating challenging and stimulating research environments during my summer internships that allowed me to experience financial markets firsthand and to develop intuition for how they work. Thanks to Andrei Kirilenko, I had a unique opportunity to participate in numerous research and surveillance projects at the Commodity Futures Trading Commission. This collaboration and numerous discussions that we had there with him, Richard Haynes, Paul Tsyhura, Tugkan Tuzun, Ekaterina Vinkovskaya and other colleagues at the agency allowed me to see financial markets from a regulator's point of view and greatly expanded my perspective. I also had an invaluable opportunity to participate in research projects at a financial exchange during my visits to IMPA in Rio de Janeiro and to BM&F BOVESPA in Sao Paulo. I am grateful to Guilherme Lamacie, Thalita Leite, Luca Mertens and to Professor Jorge Zubelli for being instrumental in organizing this priceless experience, for opening the doors of their institutions and for supporting my research projects there.

I kindly thank Donella Alanwick, Adina Brooks, Jenny Mak, Jaya Mohanty, Darbi Roberts, Jonathan Stark, and others in the department and school administration. Your resourcefulness and friendly help with administrative matters during all these years allowed me to fully focus on my research.

Thanks to my classmates and friends this journey has truly been a lifetime experience. I share many cherished memories of Columbia University, New York City and other amazing places with Andrew Ahn, Fabio D'Andreagiovanni, Nikolay Archak, Carlos Arguello, Serhat Aybat, Anna Barbashova, Amel Bentata, Berk Birand, Jaehyun Cho, Rodrigo Carrasco, Timur Davis, Romain Deguest, David Fournie, Martin Gentile, Elizaveta Golovatskaya, Jonathan Guinegagne, Tulia Herrera, Rouba Ibrahim, Yu Hang Kan, Jinbeom Kim, Song-Hee Kim, Adrien de Larrard, Thiam Hui Lee, Gregoire Lepoutre, Yunan Liu, Alla Markova, Aalok Mehta, Luca Mertens, Amal Moussa, Radka Pickova, Matthieu Plumettaz, Yuriy Pronin, Tony Qin, Johannes Ruf, Marco Santoli, Edson Bastos e Santos, Allison Schwier, Yixi Shi, Dmitriy Smelov, Sasha Stoikov, Roman Sunitskiy, Irene Song, Jingjing Song, Rishi Talreja, Abhinav Verma, Ekaterina Vinkovskaya, Daniela Wachholtz, Rolf Waeber, Lakshithe Wagalath, Xingbo Xu, Cecilia Zenteno, Alexey Zemnitskiy, John Zheng, Haowen Zhong, Yori Zwols, and many others who are not in this list but in my heart. Thank you

for these amazing memories!

A special thanks goes to my family members - my mother Yulia for exciting my interest in sciences from the early age, my father Dmitriy for supporting my decision to pursue a degree in applied mathematics, my step-father Oleg for introducing me to quantitative finance, my grandmothers Nina and Mila for their continuing love and encouragement, and to everybody in my extended family, spread between Los Angeles in the West and Neryungri in the East - thanks for your support.

Finally and most importantly, I thank my wife Olga with all my heart for sharing all happy and challenging moments of my student life. You are the true source of my inspiration and strength.

To Vasilisa

Chapter 1

Introduction

Shareholder: I really must say that you are an ignorant person, friend Greybeard, if you know nothing of this enigmatic business [of a stock exchange] which is at once the fairest and the most deceitful in Europe, the noblest and the most infamous in the world, the finest and the most vulgar on earth. It is a quintessence of academic learning and a paragon of fraudulence; it is a touchstone for the intelligent and a tombstone for the audacious, a treasury of usefulness and a source of disaster, and finally a counterpart of Sisyphus who never rests as also of Ixion who is chained to a wheel that turns perpetually.
Philosopher: Does my curiosity not deserve a short description from you on this deceit and a succinct explanation of this riddle?

Joseph de la Vega, Confusion de confusiones, 1688

1.1 The Brave New Market

The main role of financial markets is to facilitate transactions between buyers and sellers and thus to provide *liquidity*. Markets are also the main vehicle of *price formation* - a process by which prices for commodities, securities and other assets are determined and all relevant information and risks are taken into account. Both liquidity and price formation are outcomes of a dynamic interaction between multiple counterparties who actively trade and provide quotes and other indications of trading interest to the rest of the market.

Market microstructure research is a detailed analysis of how interactions between market participants give rise to liquidity and price formation. The goal of this analysis is to formulate quantitative models of the trading process and to improve this process, leading to a reduction in *transaction costs* and ultimately to a more efficient allocation of goods and capital in the economy. For example, in 1994, an academic paper [24] established that the bid-ask spreads (margins between the best prices for buying and selling securities) on the NASDAQ stock exchange were anomalously wide, suggesting that dealers colluded to purposefully widen these spreads. A subsequent investigation by the U.S. Department of Justice confirmed this observation, leading to a major scandal and an introduction of new trading rules. This new regulation lead to a rapid growth of electronic communication networks (ECNs) - alternative trading platforms that profoundly changed the landscape of the U.S. equity markets, increasing competition and reducing transaction costs.

Trading regulation and trading mechanisms have continuously evolved since the first stock exchange opened in Amsterdam in 1602, but this evolution was never as fast as during the last two decades - the era of *electronic trading*. Computer technology revolutionized financial markets: trading moved from buzzing exchange pits to quiet data centers where racks of co-located computer servers automatically process real-time market information and execute orders. Similarly to other industries altered by technology, financial markets became more cost-efficient as a result of this transformation. Various measures of transaction costs, such as bid-ask spreads, effective spreads, brokerage commissions and order execution delays dramatically improved over time (see [10; 23; 69]). As trading became more transparent and its costs decreased, some previously uneconomical investment strategies became viable and trading volumes skyrocketed.

The standardization and automatization of trading protocols made it possible to implement many trading strategies as computer programs. Electronic trading now dominates liquid financial markets. For example, in the U.S. the share of equity trading volume due to electronic trading increased from 16% in 2000 to 82% in 2009 [71] and continues to grow. Although the rules and costs of trading are the same for all traders, competitive forces have led to a significant specialization among their strategies and algorithms. In particular, it is common to distinguish between *algorithmic trading* in general and its subset *high-frequency*

trading. Algorithmic trading is a general term, but most commonly it is used to refer to trade execution strategies that are typically used by fund managers to buy or sell large amounts of assets. They aim to minimize the cost of these transactions under certain risk and timing constraints. These algorithms adjust how trading is done but typically do not have discretion on whether to trade or not - this decision belongs to a portfolio manager. In contrast, high-frequency trading algorithms have full discretion regarding their trading positions. They opportunistically submit orders to buy and to sell, aiming to profit from the market environment itself by capturing small price differences between their buy and sell orders many times during a trading session.

The transition to electronic trading was a radical shift and it has seen its share of drama, skepticism and controversy. As trading became predominantly electronic and exchange pits closed, thousands of clerks, brokers and other trading specialists lost their jobs [87]. Since then, repetitive periods of extreme volatility and incomprehensible stock price behavior have plagued financial markets. These periods were associated with computer trading algorithms, and their rather frequent occurrence undermines investors' confidence in the current market structure [43; 21]. One of the most well-known episodes of such market turbulence is the *Flash Crash* of May 6, 2010. During a fifteen-minute interval of time between 2:30pm and 2:45pm, all major U.S. equity indices experienced a sudden 5-6% decline followed by an almost complete recovery within the next 30 minutes. Some liquid stocks and exchange-traded funds (ETFs) declined even further during that period, and thousands of trades were executed at prices more than 60% away from their values just minutes prior. The government investigation of that event [109] related it to a computer program executing an exceptionally large order to sell. Empirical evidence presented in this government report and in the subsequent studies [75; 93] shows that complex interactions between the program executing the large sale and other trading programs exacerbated, if not caused, the crash.

Another infamous technological incident occurred with the trading systems of Knight Capital Group on August 1, 2012. Knight Capital was one of the worldwide leaders in automated market-making and a major advocate of electronic trading in general. Its market volume share in U.S. equities alone was more than 16%. According to the company's own official statement "*Knight experienced a technology issue at the open of trading at the NYSE*

... This issue was related to Knight's installation of trading software and resulted in Knight sending numerous erroneous orders in NYSE-listed securities into the market ... , which has resulted in a realized pre-tax loss of approximately $440 million.". According to company's financial statements, this loss is approximately equal to its cumulative net income over the previous 4 years, or about 30% of its previous year book equity value. After just 30 minutes of spurious trading this error brought to an end Knight's 17 year-long history of success, and the firm was subsequently merged with one of its rivals.

These events highlight the fragility of modern electronic markets and the high level of their operational risks. Although a reduction in transaction costs due to automated electronic trading is an empirically established fact (see [60; 90]), the complexity, operational risks and technological costs of maintaining the modern electronic market system are often cited as arguments against it. The incidents cited above are extreme, dramatic consequences of this complexity, but it also presents challenges for the analysis of everyday market behavior. Some of the old questions in the market microstructure literature need to be revisited again for this brave new market. How do changes in supply and demand translate into changes in prices? What drives market volatility? How to quantify market liquidity and what is the connection between liquidity and volatility? There is also a number of new questions that emerged as markets became increasingly automated and fragmented. How can market participants reduce their transaction costs in this challenging environment? What is the effect of high-frequency trading strategies on market dynamics? What benefits does market fragmentation bring to investors? Fortunately, electronic markets also generate enormous amounts of data which help in answering these questions. The availability of data coupled with the simple, transparent rules that underlie electronic trading create exciting opportunities for research and innovation in market microstructure.

1.2 Limit Order Books

> *Merchant: If it is not too great a trouble for our friend, I should like to hear also about the place and the ways of the exchange transactions, how business is done, for, although we know the origin, the innovators, and the confusions of the stock exchange, we do not yet know anything about the kind of business dealings or the site of the contest.*
>
> Joseph de la Vega, Confusion de confusiones, 1688

A major factor that contributed to the transformation of financial markets is the widespread adoption of a unified trading protocol - a *limit order book* - by most exchanges around the world [66; 112]. To trade in a limit order book at an exchange, participants submit *limit orders* - electronic instructions to buy or to sell a certain asset, that specify a quantity to be traded and the worst acceptable price (a limit price). Upon receiving a new order, a matching engine at the exchange compares the order's price and quantity with prices and quantities of pre-existing orders. If an existing order can be matched with the new one (according to their limit prices) they are are executed and a trade is reported. The new order is called *marketable* or *aggressive* in this case because it initiated the trade and the pre-existing order is called *passive*. If the new order's limit price does not allow to execute it right away (i.e. it is *non-marketable*) it joins other unmatched orders in the *order book* and stays there until either a new aggressive order *fills* it or until it is *canceled* by the market participant.

Order matching details vary across exchanges and can sometimes become very involved [67], but in general matching priority is simply given to passive orders that have a better price (e.g. buy orders with a higher price) and to orders that arrived earlier. With this price-time matching priority structure, limit orders with identical prices form *first-in first-out (FIFO) queues* across different price levels as they arrive to the exchange. The queue of orders to buy (sell) at the highest (lowest) price is called *the best bid* (*the best ask*) and only orders at the front of these two queues are matched with aggressive orders. Order queues are depleted by aggressive orders and cancelations, and whenever the best bid or ask queue is depleted, a queue at the next-best price becomes the new best bid or ask. Some

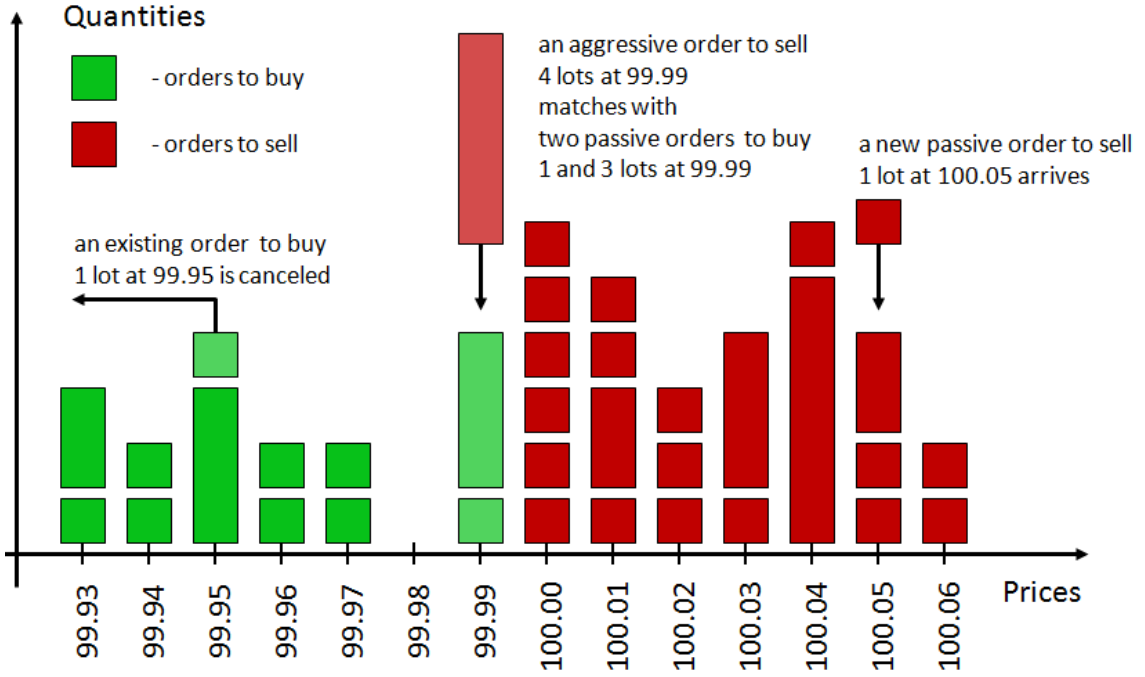

Figure 1.1: An illustration of limit order book dynamics

exchanges provide additional order types, such as *market orders*, pegged orders, iceberg orders, discretionary orders, etc., but most of them can be represented as simple limit order trading strategies.

For markets organized as limit order books liquidity is often associated with passive orders because they provide options to incoming traders to transact with them at their limit prices. If liquidity is viewed as a good, then submission of passive orders is associated with liquidity provision and submission of aggressive orders - with its consumption. In contrast with dealer or specialist markets, limit order markets do not have designated agents whose role is to provide liquidity or to set prices. Instead, liquidity is self-organized in the sense that some market participants send passive limit orders based on their preferences and their experience with the trading process. Similarly, prices are formed by passive orders resting at the *top of the book* - near the best bid and ask prices.

The basic rules of limit order trading described above are quite simple, but limit order *trading strategies* can be very complex. Depending on trader's objectives he can submit limit

orders to buy and/or to sell at different points in time with different prices and quantities and moreover submit orders simultaneously to multiple order books (in a *fragmented market* the same asset can be traded at multiple exchanges). These orders can be subsequently canceled and modified at any time producing a high-dimensional action space of possible strategies. In addition to this, strategies depend on past order submission outcomes and on the actions of other traders. The high flexibility of strategies presents a significant challenge for modeling limit order book markets in an economic equilibrium framework, but some progress has been made in this direction (see [100] for a recent survey). Nevertheless, the formal and well-defined decision environment of a limit order book (as compared to more opaque dealer markets), combined with enormous amounts of data generated and recorded daily by electronic financial markets present a good opportunity for rigorous mathematical modeling of these systems. Their complexity can be overcome and meaningful results can still be obtained from reduced-form models of limit order books using operations research tools such as stochastic processes, statistics and optimization. Models of limit order books have an immediate and important application in the financial markets industry, and also motivate the development of new stochastic modeling techniquies and new statistical tools that could be of interest to a broader academic community.

1.3 Literature Review

One of the major questions in market microstructure analysis is how trading activity leads to price changes. The *price impact* of orders is a fundamental mechanism of price formation because it translates changes in supply and demand into price movements. Early market microstructure literature has described this concept in a setting of specialist markets. In these markets prices are quoted by a centralized intermediary (a specialist) or a group of competitive market-makers. The specialist receives orders from brokers and updates his quotes as he interacts with their order flow. From a broker's viewpoint, the impact of his orders is a cost paid to a market-maker for his continuous availability to accept broker's orders [33], i.e. price impact can be seen as a transaction cost, specifically the cost of immediacy. From a market-maker's viewpoint, the picture is more complex. Some of the

orders are submitted by brokers based on information about future asset payoffs, and a fraction of this information is inferred from their order flow by the market maker. This information becomes permanently impounded in the market-maker's quotes [79] and the *permanent* price impact reflects this learning process. Larger aggressive orders are more frequently used by informed traders than smaller orders, implying larger potential losses for a market-maker who trades against an informed counterparty [34]. Larger orders also create more inventory risk for a market-maker because accepting these orders significantly exposes him to random future price fluctuations. To recoup these utility losses, market-makers quote worse prices for larger orders. The difference between a price that a specific order obtains in the market and the best available quote for a small order is called the *immediate* price impact and it is an increasing function of an order size. The *temporary* price impact can be defined as a difference between the immediate and the permanent impact of an order [73] and it represents a transient price concession required to accommodate a large order.

Although specialist markets have been largely replaced by electronic limit order books, the same price impact terminology is commonly applied to both kinds of markets. However, for limit order markets it is much more difficult to disentangle permanent and temporary price impact because there are no centralized specialists, all traders can send aggressive and passive orders, and all learn from each other's actions. To describe price impact in a theoretical economic model of a limit order market, one needs to detail trader utilities and strategies, and then derive the dynamics of order flows and trader beliefs. Although this is possible to do in some stylized models, they commonly involve unobservable parameters and are rarely tractable (see [51; 100]), which prevents their estimation with real data and limits their practical applicability. Describing the immediate price impact in a limit order book appears, at first, to be a simpler goal because the impact of an aggressive order is fully described by its mechanical effect on the order book state. However this approach is also problematic because it requires a description of the high-dimensional state of an order book. This state and its dynamics can be studied in a simulation framework (see e.g. [41]) but this approach is again infeasible for real-time applications.

Since theoretical market microstructure models give little practical guidance on how to describe price impact in limit order markets, statistical models of price impact have gained

significant popularity. An empirical approach is facilitated by the availability of detailed data from limit order markets, and statistical price impact models are often motivated by applications in algorithmic trading which typically rely on the data. Empirical studies of price impact can be differentiated by the kind data that are used: proprietary data from financial firms regarding their own orders, or public data from an exchange containing all orders but not trader identities.

Proprietary datasets contain orders of a subset of firms (typically just one) but these datasets have more detailed information on each order including *intended* trade sizes that are also called *parent orders* or *meta-orders* in the literature. Using this information one can aggregate smaller orders that were actually submitted to an exchange into intended trades, which is not possible with public data that are lacking trader identifiers. With proprietary order data, one can also directly model parent order transaction cost as a function of its size, direction and other parameters. Most studies of proprietary data agree that the price impact of a parent order is an increasing, concave function of its size, controlling for the average stock trading volume, volatility and an execution time window [7; 115]. The monotonicity and concavity of these price impact functions are confirmed by many studies, but their estimation details vary and their reliance on proprietary data is prone to several general criticisms. First of all, orders are executed over time and trader's discretion to stop an execution in unfavorable circumstances introduces a significant selection bias [96]. There is also a large number of important factors that affect price impact but are difficult to control in a proprietary dataset, such as details of specific algorithms and trading venues used to execute a trade [20]. There are also significant variations in price impact across economies and sectors and across stocks with different properties such as momentum or growth [16]. In summary, modeling the price impact of parent orders is an important practical task, but implementation details make it difficult to draw general conclusions based on proprietary data.

An alternative approach is to rely on anonimized public data from an exchange and to estimate the price impact of all orders that were placed in a market by all traders. Until recently, most studies focused exclusively on *marketable orders* and analyzed correlations between sizes of these orders and subsequent price movements [39; 48; 54; 73; 74; 114; 102;

103]. General conclusions are the same as for the impact of parent orders - the immediate and temporary impact of a single marketable order are both increasing, concave functions of an order size. Unfortunately these results do not at all describe a complete picture of price formation. In most cases, marketable orders constitute about 10% of all orders [57] at an exchange, so a model for their impact gives, at best, a very limited description of market dynamics and price formation. Restricting analysis to marketable orders is also unsatisfactory from a practical point of view because modern algorithmic trading strategies often rely on passive limit orders to reduce trading costs [42]. More recent empirical studies [61; 59; 35] consider the immediate impact of marketable orders, passive orders and order cancelations together with the evolution of their impact through time. Descriptive statistical models used in these studies illuminate rich interactions between order flows in a limit order book and produce a highly complex picture of market dynamics on the level of individual orders. Unfortunately these models also involve hundreds of free parameters which reduces their usefulness for applications.

An important application of price impact models is the *optimal trade execution problem* arising in algorithmic trading. Generally speaking it is a problem of finding a strategy to buy or to sell a large quantity of assets at a low cost within a limited amount of time. The trade quantity (a parent order) is typically much larger than the quantity of passive limit orders available at the top of an order book at any time, so a trade is done with a sequence of small *child orders* that are gradually submitted to an exchange. The problem of finding a good trade execution strategy, i.e. a sequence of child order sizes and submission times is important from both practical and regulatory points of view. In most developed markets trading regulations require brokers to seek "best execution" for their customers[1]. Yet there is no definition of best execution and, until recently, there was no quantitative methodology to evaluate, compare and improve the quality of trade execution strategies.

Different formulations of the optimal trade execution problem vary in terms of their assumptions on asset price dynamics, price impact and mechanisms by which trader's orders are filled in the market. An early treatment of the trade execution problem was given in [15] where the authors used dynamic programming to minimize a mean execution cost

[1] see for example http://www.sec.gov/answers/bestex.htm

objective. This formulation was later augmented in [6; 105] by adding a penalty for the execution price risk carried by price fluctuations over the course of an execution. A trader's aversion to execution risk presses him to trade faster and use larger orders, which also bear higher costs due to temporary price impact. Finding an optimal execution strategy that balances execution risks and price impact costs is an interesting control problem, and it has motivated a large number of theoretical studies.

Studies have noted that certain price impact assumptions make it possible to systematically earn positive returns on round-trip transactions, leading to ill-posed optimal trade execution problems. For example, if price impact is a concave function of an order size and is permanent, one can make systematic profits by selling two shares of a stock in a sequence and then buying them back with one order. A rigorous study of quasi-arbitrage trading strategies is performed in [63] with a conclusion that the permanent price impact needs to be a linear function of an order size to avoid arbitrage. A later study [50] relates the form of an immediate price impact function (which is possibly non-linear) to the rate at which temporary impact decays through time, and establishes no-arbitrage restrictions on these functions for a number of popular price impact models. Later studies [4; 44] found additional restrictions on price impact models that must be imposed to prevent price manipulation or erratic behavior of order execution strategies in a market model.

Practical applications of the optimal trade execution problem have motivated a large number of extensions and alterations of its basic setup. On a macroscopic level, this problem is related to portfolio allocation decisions because trade execution strategies are ultimately used to turn over investors' portfolios. On a microscopic level, each child order generated by a trade execution strategy needs to be placed as a limit or a market order in a limit order book, linking trade execution and order placement decisions. Although it is possible to combine optimal trade execution with optimal portfolio construction [47; 99] or optimal option hedging [86] within a single problem formulation, such problems are rarely analytically tractable, and their solutions require strong assumptions on price impact and on the evolution of asset prices.

Similarly, it is possible to include some details of order placement decisions into an optimal trade execution problem, but such extensions come at the cost of restrictive as-

sumptions on trading strategies and on market dynamics. One approach, followed in [3; 97; 104], restricts a trader to use only marketable orders. The execution cost of a marketable order is explicitly computed by integrating an idealized density function which represents passive orders at the top of the book. After a marketable order is executed, this density restores over time due to order book resiliency [81]. In these models the total cost of a sequence of marketable orders depends on their initial impact and the speed of order book recovery, creating a tradeoff between execution cost and speed. This approach presents an improvement compared to early trade execution models that completely abstracted from mechanisms of order placement. However the restriction to use only marketable orders is not satisfactory. In practice investors rely on passive orders for trade execution [20], thus suggesting that pure marketable order strategies are generally suboptimal. Because the restriction to use marketable orders is imposed ex-ante in these models, they also do not improve our understanding of the order type selection (limit or market) and do not explain what factors contribute to this choice.

More recent studies [14; 53; 65] present optimal trade execution strategies with limit orders. Limit order fills are modeled with a point process whose intensity depends on the distance of a limit order to the best bid or ask price. This distance is within trader's control, and thus a trade execution problem with limit orders can be solved using stochastic control techniques. Market orders can also be included in this setting as impulse controls. A general drawback of this approach is that it requires specifying a joint process of price and limit order book dynamics. Another problem is that even with strong assumptions on this process, the resulting control problems are usually analytically intractable. Finally, the focus of these problems remains on optimizing trade execution across time, and only basic details of order placement decisions (e.g. average limit order execution rates) are included in stochastic control formulations. This leaves out important information about the present state of an order book [111], adverse selection in limit order executions [107], order queue sizes and limit order queueing delays [94] that matter for order timing and order placement decisions in practice.

Another important aspect of trade execution that is rarely addressed in the academic literature is market fragmentation, i.e. the possibility of trading the same asset in multiple

limit order books at the same time. The problem of optimizing order placement across multiple *dark pools* (markets where supply and demand are unobserved) is studied in [49; 82] but there is no similar analysis for limit order books.

In summary, the problem of optimally pricing and placing limit orders in a fragmented market remains relatively unexplored and, until recently, limit order placement optimization was considered only as a complicated extension to the trade execution problem.

The market microstructure literature shows a steady interest in developing tractable models of order-driven markets that can realistically describe these markets, and at the same time can be used in practical applications such as optimal trade execution [97; 104] or market-making [12; 53; 85]. The two main streams of literature - theoretical equilibrium models from financial economics, and statistical order book models from econometrics and econophysics - describe different aspects of limit order markets but do not seem to answer practical needs. Economic models precisely describe various theoretical tradeoffs that are faced by market participants, but usually present them in a simplified setting and rely on unobservable parameters such as individual trader utilities. These models can rarely be estimated with real data which limits their practical value. On the other hand, statistical models of price impact and order book dynamics can fit the data well, but often lack economic structure and involve numerous free parameters which also reduces their usefulness in applications.

Stochastic order book models seem to find a balance between analytical tractability and descriptive power. The distinguishing feature of these models is that they represent aggregate order flows from all traders with a random process, e.g., a Poisson process [31; 41]. With this description a limit order market can be viewed as a stochastic system (a queueing system), where orders randomly arrive, cancel and execute according to order matching rules. This stochastic modeling approach avoids making detailed assumptions on the utilities, beliefs and strategies of individual traders, replacing them with assumptions on measurable statistical properties of aggregate order flows, thus making it straightforward to calibrate a stochastic model to order book data. Stochastic models of limit order books are more flexible than economic equilibrium models, but possess more structure than purely descriptive statistical models.

The analysis of stochastic order book models relies on a rich queueing theory, and their analytical tractability makes them appealing from a practical point of view. For example, order execution costs can be improved in practice by strategically timing order submissions. It is optimal to submit aggressive orders when the probability of an adverse price change implied by the current order book state is high [111]. This probability and other useful distributions for order placement optimization can be easily computed in a birth-death stochastic order book model via Laplace transforms [31; 62]. Analytical results can also be derived from scaling approximations of order book models. Even if processes that govern individual order arrivals and cancelations are very complicated, under suitable conditions their time aggregates can be conveniently approximated with simpler processes using functional laws of large numbers and functional central limit theorems. For instance, a queueing order book model in a heavy-traffic regime is analyzed in [30]. This model is used to derive the distribution of durations between consecutive price changes, as well as the distribution of price increments conditional on the order book state. This analysis creates a link between the order book state and its dynamics on a *microscopic* timescale of a few milliseconds and price fluctuations on longer *mesoscopic* timescales of several minutes or hours. The link is formally established in a subsequent work [29] via a heavy-traffic limit theorem - under suitable conditions microscopic price changes in a queueing order book model converge on a larger scale to a Brownian motion, whose volatility is a function of the average order queue size and the order inter-arrival rate. Other widely used tools in the stochastic modeling literature are fluid approximations, and they have been applied to limit order books as well. A fluid model of a fragmented market with multiple order books is used in [94] to show how the strategic order routing activity of self-interested investors creates coupling between order flows and order queues across market centers.

These are several successful examples of applying classical stochastic modeling techniques to describe limit order markets. However, the rich structure of these markets also motivates the development of entirely new modeling tools. Empirical studies of order data with trader identifiers [1; 13; 75] have exposed complicated interactions and feedback effects between heterogeneous market participants. While the queueing order book models discussed above can describe the behavior of aggregate order flows from all participants in

a limit order book, studying interactions between traders or trader groups seems to be difficult with off-the-shelf stochastic models. So far these interactions have been only explored in the financial economics literature using game theoretical arguments. The stochastic modeling literature offers a large variety of multi-class queueing models for studying systems with dissimilar participants. These models express differences between participants (jobs) in a queue or a network of queues by assigning each class of jobs a distinct distribution of arrival, service or abandonment times (see e.g. [68]). Multi-class queueing models may capture some of the differences between participants in limit order markets, but their use does not seem proper because they have been developed in other contexts such as call center modeling or healthcare. Impatient customers in a service system typically share the same timescale - customers in a call center may wait several minutes, and patients may spend several days or weeks in a ward. In contrast, some participants in financial markets submit and cancel their orders thousands of times faster than other traders [57], responding to changes in the state of the order books. Some orders that queue in a limit order book are driven by patience, similarly to customers in a service system, while others exhibit more complex behavior which depends on the order book state. In summary, modeling limit order books is a new and interesting application of the queueing theory and presents exciting opportunities for developing new models and new theoretical tools.

1.4 Contribution

The rapid transformation of financial markets during the last two decades posed a number of challenges. Regulators and exchanges worldwide recognize the complexity of these markets and work to improve their fairness, transparency and robustness to hazards like the *Flash Crash* event. Practitioners increasingly rely on quantitative models and computer algorithms to hedge their positions, turn over large portfolios, provide liquidity and guide their trading decisions. To better understand and improve the modern marketplace, regulators and practitioners alike need models of limit order books, which on one hand could realistically describe the complicated dynamics exhibited by these markets, and on the other hand would be tractable enough to be useful in applications.

This need has motivated four research goals addressed in this thesis:

- Develop a parsimonious model that relates price dynamics in a limit order market to order submissions and cancelations (order book events) without requiring a complex description of full order book dynamics or references to unobservable parameters.
- Investigate connections between the price impact of order book events, liquidity, trading volume and volatility in limit order markets on intraday timescales.
- Propose a tractable framework for order placement and order routing optimization in a fragmented market with multiple limit order books, and study the tradeoffs that form the order placement decisions of a single investor.
- Study how the heterogeneity in order submission and cancelation strategies across market participants (e.g. algorithmic traders and high-frequency traders) affects the dynamics of limit order queues and order waiting times in these queues.

In Chapter 2 we develop a model that describes price dynamics in limit order books by incorporating the immediate effects of limit orders, market orders and order cancelations on prices. Our model specification (2.7) is detailed enough to describe the immediate price impact of all order book updates, yet it does not require one to describe the complex evolution of a limit order book and involves a single parameter that can be estimated by fitting a linear regression. We follow an empirical approach and test our model on a large sample of high-frequency data. The methodological contribution of this model is that it finds a middle ground between theoretical order book models from financial economics literature and statistical models of price impact. In comparison with theoretical order book models, ours is simpler and can be easily mapped to real data or used in applications. In contrast with statistical price impact models, it is based on simple economic intuition and imposes strong structural restrictions on parameter values. This structure allows us to not only describe the immediate price impact in a limit order book, but also to link different variables of interest: the magnitude of price impact, market depth, volatility and trading volume. Specifically our model predicts an inverse relation between the price impact coefficient and market depth (2.8). This relation and other parameter restrictions posed

by our model are confirmed by empirical analysis in Section 2.3, where we also show that, despite its simplicity, our model successfully explains 65% of variance in short-term price movements (see also Tables 2.2 – 2.4). The link between market depth and price impact predicted by this model allows us to structurally explain intraday variations in volatility with variations in two observable variables - the average market depth and the variance of order flows 2.21. Previous studies of trade data noted that market volatility positively correlates with trading volume, but, in order-driven markets, prices are determined by limit order book dynamics and the effect of trading volume on prices is dubious. Proposition 1 in Section 2.4.3 establishes one channel through which the correlation between volume and volatility can spuriously emerge as a consequence of aggregating high-frequency data. Our model parsimoniously relates price changes, order flows and market liquidity and thus lends itself to multiple applications, from the analysis of intraday volatility dynamics to price impact forecasting. In Section 2.4.1 we develop an application of our results that can improve the quality of limit order executions in practice.

In Chapter 3 we propose a novel framework for order placement optimization in a fragmented market. The main distinction of our approach from the existing literature on optimal trade execution is that we decouple the optimization of a trade execution trajectory in time and individual order placement decisions, focusing on the latter part. This allows us to study order placement decisions in significant detail without sacrificing analytical tractability. In Section 3.2 we describe a realistic model for limit order executions in an order book that considers the length of limit order queues, statistical properties of order flows on multiple exchanges and different order routing incentives (fees and rebates) provided by exchanges. Based on this detailed description of limit order execution mechanics, the search for an optimal order placement policy can be formulated as a convex optimization problem 3.4, which is shown to have an optimal solution in propositions 2,3. An alternative approach to order placement optimization is based on a cost minimization problem 3.6 under execution shortfall constraints, and we show in section 3.2 that the two approaches are connected by duality. Other studies in the optimal trade execution literature have predominantly focused on the risk of price fluctuations during the course of a trade execution, but assumed that each order on a trade execution schedule is filled with certainty. Instead, we focus on the

non-execution risk for an individual order and show that this risk plays a major role in determining an optimal mix of market and limit orders for a trader. In a special case when a single exchange is used for order execution, proposition 4 gives an explicit formula for optimal limit and market order sizes, and shows how the optimal solution shifts to market orders when the non-execution risk outweighs the high cost of these orders.

The fragmentation of financial markets in the U.S. and Europe motivates us to analyze the order placement problem with multiple exchanges. This is an improvement relative to existing studies that typically consider a single exchange for order execution. Proposition 5 shows that the optimal solution in case of fragmented markets is characterized by setting execution shortfall probabilities to specific values computed with transaction cost parameters. In Section 3.5 we propose a stochastic approximation scheme that uses historical data samples to optimize order placement decisions without a need to specify order flow distributions. Our numerical examples in that section illustrate the structure of an optimal solution in a case of fragmented markets and show that substantial cost savings can be realized by applying order placement optimization. The framework developed in Chapter 3 can also be applied to study the costs and benefits of market fragmentation from an investor's perspective. We observe that it is often optimal to oversize the total quantity of limit orders sent to all exchanges (i.e. to overbook) in an attempt to diversify the non-execution risk. However, when order flows on each exchange are strongly positively correlated, this diversification advantage fades and the cost of execution in a fragmented market can become higher than the cost of execution in a consolidated market.

In Chapter 4, we study the dynamics of a limit order queue in a market populated by heterogeneous traders. Our model for this queue, presented in Section 4.2, explicitly distinguishes between orders submitted by trade execution algorithms and high-frequency traders (henceforth, type-1 and type-2 orders). Trade execution algorithms usually operate under time constraints and therefore type-1 orders in our model have finite patience deadlines. In contrast, type-2 orders of high-frequency traders respond to changes in a limit order book state and are not driven by patience. This distinction between order types plays an important role in our analysis and it is a new feature compared to the existing literature on stochastic order book models, which assume identical dynamics for all orders

in the market. In our model, orders of both types arrive at the back of the same FIFO queue at the best bid (ask) and gradually propagate through this queue to match against a contra-side marketable order at the queue front. When the queue size is relatively large, high-frequency traders perceive a low risk of price changes and keep their type-2 orders in the queue, as opposed to type-1 orders that constantly cancel due to impatience. As it takes additional time for marketable orders to clear the extra "workload" introduced by type-2 orders, the queue waiting times for type-1 orders increase, forcing execution algorithms to cancel their orders and trade aggressively at worse prices. This "order crowding" effect has been previously noticed by traders in exceptionally liquid stock and futures markets, but it cannot be captured by models with homogenous order behavior. In contrast, when the queue size becomes small, all type-2 orders immediately cancel due to the risk of a price change, creating a rush of high-frequency order cancelations.

In Section 4.3 we present our results for a fluid model that approximates the evolution of large orders queues over relatively long time intervals. This approximation is analytically tractable and applicable to studying limit order queue waiting times in liquid markets. Propositions 6, 7, 8 analyze the fluid model dynamics and show that a particular queue structure emerges for large queues and for queues that have evolved for a sufficiently long time. These cases correspond to the "crowding" phenomenon because all new type-2 orders remain in the queue until execution, while only a fraction of finitely patient type-1 orders make it to the queue front. Proposition 8 describes the convergence of the fluid model to a steady state and characterizes its limit. Proposition 9 explores the dependence of queue waiting times on the initial queue composition with type-1 and type-2 orders. When all orders in a queue belong to impatient traders, waiting times are expected to be relatively short even if the queue is long because most of its content is eventually canceled. However, the presence of type-2 orders significantly increases queueing delays, and our model prescribes a specific procedure for computing queueing delay forecasts as a function of the initial type-1 and type-2 order quantities. In Section 4.4 these forecasts are tested against an extensive proprietary dataset of realized limit order delays in U.S. equity markets. We find that assuming a homogenous queue composition leads to unrealistic waiting time predictions that can biased up or down by a factor of 10 or more. The model with heterogeneous orders

leads to more realistic predictions that lie between the two extremes (see Figure 4.5) and can potentially be used to predict order queueing delays in practical applications. In summary, our analysis motivates the development of a new generation of stochastic models for order book markets. These models go beyond the description of aggregate order flows and take into account more detailed features of these markets, such as the trader heterogeneity, providing a link with market microstructure models in the financial economics literature. The stochastic model introduced here is also interesting on a stand-alone basis - in comparison with existing multi-class queueing models, it introduces new types of dynamics, induced by the state-dependent behavior of one of the job classes.

Chapter 2

Price impact, liquidity and market volatility

This chapter is based on the paper "The Price Impact of Order Book Events" [28] which is a joint work with Professor Rama Cont and Doctor Sasha Stoikov.

2.1 Introduction

> *Sometimes a quiet state of prices is obtained and the Exchange is influenced by neither favorable nor unfavorable news... Suddenly a cloud appears which portends a storm. The sellers of shares rejoice and start talk about the uncertainties in the situation and the possibilities of disasters. As quick as lightning the bulls hasten forward in order to dam the inundations and to reject this reproach on their wisdom... The skirmishing goes on, and at last the price is higher than before the confusion, because those groups of exchange operators who, suspecting no intrigue, had no thought of fighting and had been pursuing their regular, peaceful practices, have been awakened by the attacks.*
>
> Joseph de la Vega, Confusion de confusiones, 1688

The availability of high-frequency records of trades and quotes has stimulated an extensive empirical and theoretical literature on the relation between order flow, liquidity and

price movements in order-driven markets. A particularly important issue for applications is the impact of orders on prices: the optimal liquidation of a large block of shares, given a fixed time horizon, crucially involves assumptions on price impact (see Bertsimas and Lo [15], Almgren and Chriss [6], Obizhaeva and Wang [97]). Understanding price impact is also important from a theoretical perspective, since it is a fundamental mechanism of price formation.

Various aspects of price impact have been studied in the literature but there is little agreement on how to model it [18], and the only consensus seems to be the intuitive notion that imbalance between supply and demand moves prices. Theoretical studies draw a distinction between instantaneous price impact of orders and its decay through time, and show that the form of instantaneous impact has important implications. Huberman and Stanzl [63] show that there are arbitrage opportunities if the instantaneous effect of trades on prices is non-linear and permanent. Gatheral [50] extends this analysis by showing that if the instantaneous price impact function is non-linear, impact needs to decay in a particular way to exclude arbitrage and if it is linear, it needs to decay exponentially. Bouchaud et al. [19] associated the decay of price impact of trades with limit orders, arguing that there is a "delicate interplay between two opposite tendencies: strongly correlated market orders that lead to super-diffusion (or persistence), and mean reverting limit orders that lead to sub-diffusion (or anti-persistence)". This insight implies that looking solely at trades, without including the effect of limit orders amounts to ignoring an important part of the price formation mechanism.

However, most of the empirical literature on price impact has primarily focused on trades. One approach is to study the impact of "parent orders" gradually executed over time using proprietary data (see Engle et. al [105], Almgren et. al [7]). Alternatively, empirical studies on public data [39; 48; 54; 73; 74; 114; 102; 103] have analyzed the relation between the direction and sizes of trades and price changes and typically conclude that the instantaneous price impact of trades is an increasing, nonlinear function of their size. This focus on trades leaves out the information in quotes, which provide a more detailed picture of price formation [37], and raises a natural question: is volume of trades truly the best explanatory variable for price movements in markets where many quote events can happen

between two trades?

In our view, a price impact model that encompasses limit orders, market orders and cancelations, and relates their impact to the concurrent market liquidity would provide a more detailed description of price formation. Obtaining such model is also desirable from the practical point of view because modern order execution algorithms increasingly use limit orders and incorporate market state variables in their decisions. There is also ample empirical evidence that limit orders play an important role in determining price dynamics. Arriving limit orders significantly reduce the impact of trades [116] and the concave shape of the price impact function changes depending on the contemporaneous limit order arrivals [110]. The outstanding limit orders (also known as market depth) significantly affect the impact of an individual trade ([76]), low depth is associated with large price changes [117; 40], and depth influences the relation between trade sizes and returns [58]. The emphasis in the aforementioned studies remains, however, on trades and there are few empirical studies that focus on limit orders from the outset. Notable exceptions are Engle & Lunde [37], Hautsch and Huang [59] who perform an impulse-response analysis of limit and market orders, Hopman [61] who analyzes the impact of different order categories over 30 minute intervals and Bouchaud et al. [35] who examine the impact of market orders, limit orders and cancelations at the level of individual events.

2.1.1 Relation to the literature

The primary focus of our study is on short-term effects of orders on stock prices, i.e. their price impact, and on the relation between price impact and market liquidity. The immediate price impact of individual orders and aggregate order imbalances (sums of buy order sizes minus sell order sizes) was previously analyzed for marketable orders in [74; 102; 103; 61] and was generally confirmed to be an increasing, concave function of a marketable order size. This can be explained by a typical order book density shape that has more orders near the top of the book and fewer orders deeper in the book [116]. A marketable order consumes liquidity from the top of the book first and then executes against orders that are progressively deeper in the order book. Therefore its impact can be written as an integral of a decreasing order book density, leading to a concave price impact function. Although

the aforementioned studies provide interesting insights into order book properties they do not consider the impact of non-marketable orders or order cancelations which constitute a majority of order book updates, and therefore more detailed models are required to comprehensively describe price dynamics in order-driven markets. Vector autoregression and kernel models were applied to study the price impact of marketable orders [54], non-marketable limit orders and order cancelations [35; 59] and the evolution of their impact through time. These statistical models give a highly detailed description of price impact but involve many parameters which impedes their further application. We propose a more structured model that describes the immediate price impact of all order book events at the top of the order book using a single parameter.

Our model is applied to investigate the relation between market depth (a measure of liquidity), price volatility and price impact. Market depth and volatility are known to be correlated - depth increases in response to an increase in transitory volatility [2], but decreases when informational volatility is high [101], while volatility itself decreases with an increase in market depth. Our model contributes to this discussion by structurally explaining these connections: price volatility is related to variations in order flow, whose effect on prices depends on market depth. We use these links to explain significant intraday variations in volatility, that were previously attributed to information asymmetries by [89] and [55]. In contrast, our model explains these diurnal data features using only observable variables. Multiple studies comment on the positive correlation between trading volume and volatility (see [72] for a review), but we argue that prices in limit order markets are driven by orders and not by trades. However, we make a connection with market volume by showing that a spurious positive correlation between price volatility and volume can emerge as an artifact due to aggregation of high-frequency data. Price impact models are often applied to optimize transaction costs, for instance optimization of limit order executions is studied in [14; 111]. Our findings suggest that monitoring imbalances in order flow can improve the quality of limit order executions.

2.1.2 Summary of main results

We conduct an empirical investigation of the instantaneous impact of order book events – market orders, limit orders and cancelations – on equity prices. Although previous studies give a relatively complex description of their impact, we show that their instantaneous effect on prices may be modeled parsimoniously through a *single* variable, the *order flow imbalance* (OFI). This variable represents the net order flow at the best bid and ask and tracks changes in the size of the bid and ask queues by

- increasing every time the bid size increases, the ask size decreases or the bid/ask prices increase,
- decreases every time the bid size decreases, the ask size increases or the bid/ask prices decrease.

Interestingly, this variable treats a market sell and a cancel buy of the same size as equivalent, since they have the same effect on the size of the best bid queue. This aggregate variable explains mid-price changes over short time scales in a linear fashion, for a large sample of stocks, with an average R^2 of 65%. In contrast, order flows deeper in the order book do not substantially contribute to price changes. Our model based on OFI relates prices, trades, limit orders and cancelations in a simple way: it is *linear*, requires the estimation of a single *price impact coefficient* and it is robust across stocks and across timescales.

Most of variability in the instantaneous price impact, both across time and across stocks is explained by variations in market depth. In fact, we establish an exact inverse relation between the two variables. The coefficient of proportionality in that relation depends dramatically on the depth definition, showing that arbitrary measures of market depth are biased proxies for price impact and may lead to misleading conclusions on market liquidity.

The price impact coefficient exhibits substantial intraday variability coinciding with known intraday patterns observed in spreads, market depth and price volatility [2; 9; 83; 91]. We explain the diurnal effects in price volatility using the volatility of order flow imbalance and market depth, as opposed to unobservable parameters previously invoked in the literature, such as information asymmetry [89] or informativeness of trades [55]. The strong link between price volatility and standard deviation of OFI suggests that our

price impact coefficient is a better estimate of Kyle's λ (a useful metric of liquidity [8; 79]) than traditional estimates based on trades data. We also show that intraday price volatility is mainly driven by OFI and not by trading volume. The positive correlation between price volatility and volume, widely confirmed by empirical studies [72], can be a statistical artifact due to aggregation of data over time, and we establish how such spurious relation can arise in our model.

The OFI variable exhibits positive autocorrelation over short time scales, which can be exploited to improve the quality of order executions. In particular, we show that a limit order fill is more likely to be followed with a price change in the same direction as the order flow imbalance before that fill. For example, a limit sell order is more likely to be adversely selected when order flow imbalance is positive. Monitoring OFI can therefore help reduce adverse selection in limit order fills.

2.1.3 Chapter outline

This chapter is structured as follows. In Section 2.2, we specify a parsimonious model that links stock price changes, order flow imbalance and market depth and motivate it by a stylized order book example. Section 2.3 describes our data and presents estimation results for our model. Section 2.4 discusses potential applications of our results: in 2.4.1 we use order flow imbalance as a measure of adverse selection in limit order executions, in 2.4.2 we demonstrate how diurnal effects in depth and order flow imbalance generate intraday patterns in price impact and price volatility, and in 2.4.3 we show how a spurious relation between volume and the magnitude of price moves emerges as a statistical artifact from our simple model. Section 2.6 presents our conclusions.

2.2 Price impact model

2.2.1 Stylized order book

To motivate our approach we first consider a stylized example of the order book where the instantaneous effect of order book events can be explicitly computed.

Consider an order book in which the number of shares (depth) at each price level beyond the best bid and ask is equal to D. Order arrivals and cancelations occur only at the best bid and ask. Moreover, when bid (or ask) size reaches D, the next passive order arrives one tick above (or below) the best quote, initializing a new best level. Consider a time interval $[t_{k-1}, t_k]$ and denote by L_k^b, C_k^b respectively the total size of buy orders that arrived to and canceled from current best bid during that time interval. Also denote by M_k^b the total size of marketable buy orders that arrived to current best ask, and by P_k^b the bid price at time t_k. The quantities L_k^s, C_k^s, M_k^s for sell orders are defined analogously and P_k^s is the ask price.

In this simple order book model there exists a linear relation between order flows $L_k^{b,s}, C_k^{b,s}, M_k^{b,s}$ and price changes $\Delta P_k^{b,s} = (P_k^{b,s} - P_{k-1}^{b,s})$ (also illustrated on Figures 2.1-2.3):

$$\Delta P_k^b = \quad \delta \left\lceil \frac{L_k^b - C_k^b - M_k^s}{D} \right\rceil \tag{2.1}$$

$$\Delta P_k^s = -\delta \left\lceil \frac{L_k^s - C_k^s - M_k^b}{D} \right\rceil, \tag{2.2}$$

where δ is the tick size[1]. These relations are remarkably simple - they involve no parameters, the impact of all order book events is *additive* and depends only on their net imbalance.

[1]This is easily proven by induction over the number of price changes in $[t_{k-1}, t_k]$. The statement is clearly true when there are no price changes or a single price change of $\pm\delta$. Since any price change of $\pm k\delta$ consists of jumps of size 1, we simply need to sum the order flow imbalances across these jumps on the right side of the equation.

Although all of the subsequent analysis can be carried out separately for bid and ask prices, for simplicity we consider mid-price changes normalized by tick size $P_k = \frac{P_k^b + P_k^s}{2\delta}$:

$$\Delta P_k = \frac{OFI_k}{2D} + \epsilon_k, \tag{2.3}$$

$$OFI_k = L_k^b - C_k^b - M_k^s - L_k^s + C_k^s + M_k^b, \tag{2.4}$$

where OFI_k is the order flow imbalance (or net order flow) and ϵ is the truncation error. We can also rewrite (2.3) as:

$$\Delta P_k = \frac{TI_k}{2D} + \eta_k, \tag{2.5}$$

$$TI_k = M_k^b - M_k^s, \tag{2.6}$$

where TI_k is the trade imbalance and $\eta_k = \frac{L_k^b - C_k^b - L_k^s + C_k^s}{2D} + \epsilon_k$. When limit order activity dominates, i.e. absolute values of terms $|L_k^{b,s}|, |C_k^{b,s}|$ are much larger than $|M_k^{b,s}|$, the correlation of price changes with TI_k is weaker than with OFI_k, because limit order submissions and cancelations manifest as noise in (2.5).

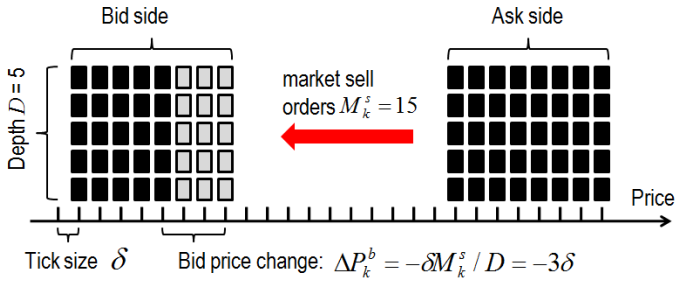

Figure 2.1: Market sell orders remove M^s shares from the bid (gray squares represent net change in the order book).

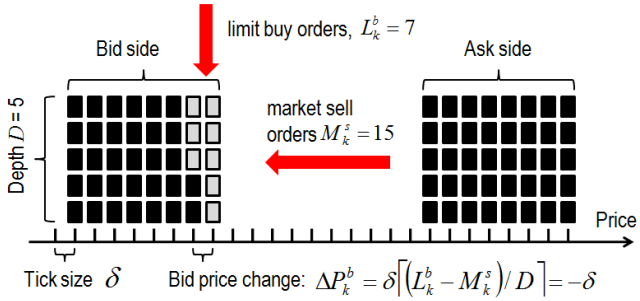

Figure 2.2: Market sell orders remove M^s shares from the bid, while limit buy orders add L^b shares to the bid.

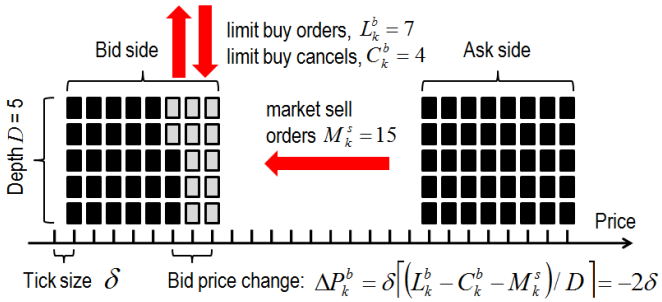

Figure 2.3: Market sell orders and limit buy cancels remove $M^s + C^b$ shares from the bid, while limit buy orders add L^b shares to the bid.

2.2.2 Model specification

Actual order books have complex dynamics: arrivals and cancelations occur at all price levels, the depth distribution across levels has non-trivial features [103; 106; 119], and hidden orders together with data-reporting issues create additional errors [11; 56]. Motivated by the stylized order book example we assume a noisy relation between price changes and OFI, which holds locally for short intervals of time $[t_{k-1,i}, t_{k,i}] \subset [T_{i-1}, T_i]$, where $[T_{i-1}, T_i]$ are longer intervals.

$$\Delta P_{k,i} = \beta_i OFI_{k,i} + \epsilon_{k,i} \tag{2.7}$$

In this model β_i is a *price impact coefficient* for an i-th time interval and $\epsilon_{k,i}$ is a noise term summarizing influences of other factors (e.g. deeper levels of the order book). We allow β_i and the distribution of $\epsilon_{k,i}$ to change with index i, because of well-known intraday seasonality effects. Our discussion from the previous section allows us to interpret $\frac{1}{2\beta_i}$ as an *implied* order book depth. The stylized order book model suggests that price impact coefficient is inversely related to market depth, and we consider the following model:

$$\beta_i = \frac{c}{D_i^{\lambda}} + \nu_i, \tag{2.8}$$

where c, λ are constants and ν_i is a noise term. The stylized order book model corresponds to $c = \frac{1}{2}, \lambda = 1$. We also consider a relation between price changes and trades:

$$\Delta P_{k,i} = \beta_i^T TI_{k,i} + \eta_{k,i}, \tag{2.9}$$

but expect it to be much noisier than (2.7).

The specification (2.7-2.8) may be regarded as a model of instantaneous price impact of order book events, arriving within time interval $[t_{k-1}, t_k]$. An order submitted or canceled at time $\tau \in [t_{k-1}, t_k]$ contributes a signed quantity e_τ to supply/demand. In any given time interval, these contributions are likely be unbalanced, leading to an order flow imbalance OFI_k, which affects supply/demand and leads to a corresponding price adjustment. If an individual order goes in the same direction as the majority of orders ($sgn(e_\tau) = sgn(OFI_k)$),

it reinforces the concurrent order flow imbalance and can affect the price. If the order goes against the concurrent order flow imbalance ($sgn(e_\tau) = -sgn(OFI_k)$), it is compensated by other orders and has an instantaneous impact of zero. In our model all events (including trades) have a linear price impact, on average equal to β_i during the i-th interval. Their realized impact however depends on the concurrent orders.

The idea that the concurrent limit order activity can make a difference in terms of trades' impact was demonstrated in [110], where authors show that the shape of the price impact function essentially depends on the contemporaneous limit order activity. Our approach can also be related to the model proposed in [35], where order book events have a linear impact on prices, which depends on their signs and types[2]. The major difference of our model lies in the aggregation across time and events. As shown in [35], time series of individual order book events have complicated auto- and cross-correlation structures, which typically vanish after 10 seconds. In our data the autocorrelations at a timescale of 10 seconds are small and quickly vanish as well (ACF plots for a representative stock are shown on Figure 2.4). Finally, the model used in [58] for explaining the price impact of trades is similar to (2.9). Although the focus there is on trades, authors allow the price impact coefficient to depend on contemporaneous liquidity factors and change through time.

At the same time, the linear relation (2.7) is different from many earlier models that consider only the effect of transactions [48; 54; 74; 114; 102; 103]. Instead of modeling price impact of trades as a (nonlinear) function of trade size, we show that the instantaneous price impact of a series of events (including trades) is a linear function of their size after these events are aggregated into a single imbalance variable. We will show that, first, the effect of trades on prices is adequately captured by the order flow imbalance and, second, that if one leaves out all events except trades, the relation 2.7 leads to an apparent concave relation between the magnitude of price changes and trading volume.

The next section provides an overview of the estimation results for our model.

[2]Note that in our case all order book events have the same average impact, equal to β_i, regardless of their type. As shown in [35], average impacts of different event types are empirically very similar, allowing to reasonably approximate them with a single number.

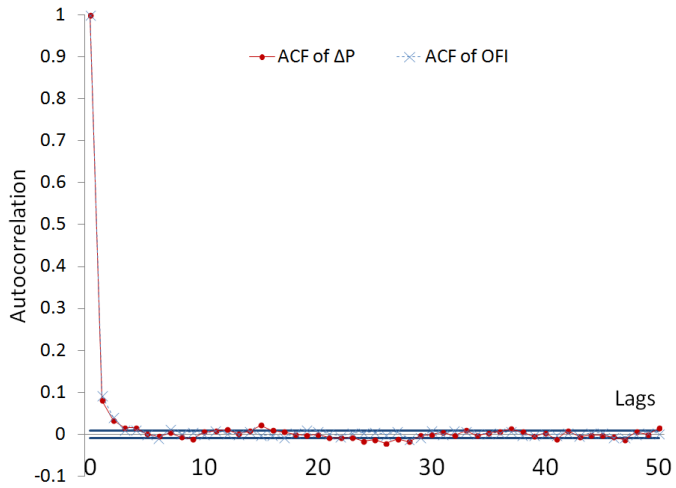

Figure 2.4: ACF of the mid-price changes $\Delta P_{k,i}$, the order flow imbalance $OFI_{k,i}$ and the 5% significance bounds for the Schlumberger stock (SLB).

2.3 Estimation and results

2.3.1 Data

Our main data set consists of one calendar month (April, 2010) of trades and quotes data for 50 stocks. The stocks were selected by a random number generator from S&P 500 constituents, which were obtained from Compustat. The data for individual stocks was obtained from the TAQ consolidated quotes and TAQ consolidated trades databases[3].

Consolidated quotes contain best bid/ask price changes and round-lot changes in best bid/ask sizes. Quote data entries consist of a stock ticker, a timestamp (rounded to the nearest second), bid price and size, ask price and size and various flags including exchange flag. Consolidated trade entries consist of timestamps, prices, sizes and various flags. These two data sets are often referred to as Level 1 data, as opposed to Level 2 data, which includes quote updates deeper in the book, or information on individual orders. The main reason for using TAQ data rather than Level 2 order book data, is that it is far more accessible, yet contains all events in the top order book (best bid and ask updates), except maybe for odd-lot changes. We demonstrate that Level 1 TAQ data can be successfully used to study

[3]The TAQ data were obtained through Wharton Research Data Services (WRDS).

the impact of limit order submissions and cancelations and we hope that more empirical studies of that subject will follow. We find that the ratio between the number of NBBO quote updates and the number of trades is roughly 40 to 1 in our data. Many empirical studies have previously focused exclusively on trades rather than quotes, but the sheer difference in sizes of these data sets suggests that more information may be conveyed by quotes than by trades.

We considered only quotes with timestamps $\in$ [9:30 am, 4:00 pm], positive bid/ask prices and sizes, and quote mode $\notin \{4, 7, 9, 11, 13, 14, 15, 19, 20, 27, 28\}$. Similarly, trades were considered only if they had timestamps $\in$ [9:30 am, 4:00 pm], positive price and size, correction indicator $\leq$ 2 and condition $\notin \{"O", "Z", "B", "T", "L", "G", "W", "J", "K"\}$.

From the filtered quotes data we construct the National Best Bid and Offer (NBBO) quotes. This is done by scanning through the filtered quotes data, while maintaining a matrix with the best quotes for every exchange. When a new entry is read, we check the exchange flag of that entry and update the corresponding row in the exchange matrix. Using this matrix, the NBBO prices are computed at each entry as the highest bid and the lowest ask across all exchanges. The NBBO sizes are simply the sums of all sizes at the NBBO bid and ask across all exchanges. For more details on TAQ dataset we refer the reader to [56], which discusses some particularities of that data, such as possible mis-sequencing of data across exchanges and lack of odd-lot sized orders. With our auxiliary dataset we checked that neither of these issues significantly affects our results. As a robustness check, we also considered using data from one exchange at a time instead of NBBO data and obtained similar empirical results.

After the NBBO quotes are computed, we applied a simple quote test to the NBBO quotes and the filtered trades data. This test matches trades with NBBO quotes and computes the direction of matched trades. A trade is matched with a quote, if:

1. Trade is not inside the spread, i.e.
   (a) Trade price $\geq$ NBBO ask: in this case the trade is considered to be a buy trade.
   (b) Trade price $\leq$ NBBO bid: in this case the trade is considered to be a sell trade.
2. Trade date = quote date.
3. Trade timestamp $\in$ [quote timestamp, quote timestamp + 1 second].
4. If the above conditions allow to match a trade with several quotes, it is matched with the earliest quote.

This matching algorithm cannot identify the direction of trades occurring within the bid-ask spread. By comparing the number of matched trades with the overall number of trades in our sample, we found that 59-95% of trades depending on the stock cannot be matched. Although these percentages appear to be extremely large, the volume percentage of unmatched trades is only 10-39% depending on the stock with an average of 17% across stocks, and we believe that omitting these trades does not affect our results. There are other routines to estimate trade direction, including the tick test and the Lee-Ready rule [84]. Although the latter is used quite frequently, there seems to be no compelling evidence of superiority of either of these heuristics [98; 113]. To test the robustness of our findings to the choice of a trade direction test, we compared our results on a subsample of stocks, applying alternatively the tick test or our quote test and results were virtually the same.

Finally, we removed observations with high bid-ask spreads to filter out "stub quotes" and data errors. To apply this filter coherently across stocks, we computed the 95-th percentile of bid-ask spread distribution for each stock and removed 5% of that stock's quotes with spreads above that percentile. For the representative stock in our sample (SLB), the removed observations fall mostly on the first minutes after market opening: 15.8% of them occur between 9:30 am and 9:35 am, and 42.1% of them occur between 9:30

am and 10:00 am. The average bid-ask spread of the removed quotes is 3.44 cents with a standard deviation 11.98 cents, the average queue size of these quotes is 11.78 round lots with a standard deviation 12.89 lots. The average time interval between two removed quotes is 1.03 seconds with a standard deviation 41.64 seconds. All results in this chapter are generated using the filtered data.

TAQ data has important limitations - the timestamps are rounded to the nearest second, and it may omit odd-lot trades and quotes. To perform several detailed robustness checks we also use an auxiliary data set consisting of NASDAQ ITCH 4.0 messages for the same calendar month (April, 2010) for one representative stock from our main data set (Schlumberger). This data is accessible through LOBSTER website[4] which also provides NASDAQ order book history for the selected stock. We used LOBSTER data for the top five order book levels without any additional pre-processing.

2.3.2 Variables

Every observation of the bid and the ask consists of the bid price P^b, the bid queue size q^b (in number of shares), the ask price P^s and size q^s. We enumerate them by n and compute differences between consecutive observations $(P^b_n, q^b_n, P^s_n, q^s_n)$ as follows:

$$e_n = q^b_n \mathbb{1}_{\{P^b_n \geq P^b_{n-1}\}} - q^b_{n-1} \mathbb{1}_{\{P^b_n \leq P^b_{n-1}\}} - q^s_n \mathbb{1}_{\{P^s_n \leq P^s_{n-1}\}} + q^s_{n-1} \mathbb{1}_{\{P^s_n \geq P^s_{n-1}\}} \tag{2.10}$$

The variables e_n are signed contributions of order book events to supply/demand. When a passive buy order arrives, q^b increases but P^b remains the same, leading to $e_n = q^b_n - q^b_{n-1}$ which is the size of that order. If q^b decreases, we have $e_n = q^b_n - q^b_{n-1}$, representing the size of a marketable sell order or buy order cancelation. If P^b changes, then $e_n = q^b_n$ or $e_n = -q^b_{n-1}$, representing respectively the size of a price-improving order or the last order in the queue that that was removed. Symmetric computations are done for the ask side.

[4] http://lobster.wiwi.hu-berlin.de/Lobster/about/About_WhatIsLOBSTER.jsp

We use two uniform time grids $\{T_0, \ldots, T_I\}$ and $\{t_{0,0}, \ldots, t_{I,K}\}$ with time steps $T_i - T_{i-1} = 30$ minutes and $t_{k,i} - t_{k-1,i} = \Delta t = 10$ seconds[5]. Within each long time interval $[T_{i-1}, T_i]$ we compute 180 price changes and order flow imbalances indexed by k:

$$\Delta P_{k,i} = \frac{P^b_{N(t_{k,i})} + P^s_{N(t_{k,i})}}{2\delta} - \frac{P^b_{N(t_{k-1,i})} + P^s_{N(t_{k-1,i})}}{2\delta}, \tag{2.11}$$

$$OFI_{k,i} = \sum_{n=N(t_{k-1,i})+1}^{N(t_{k,i})} e_n, \tag{2.12}$$

where $N(t_{k-1,i})+1$ and $N(t_{k,i})$ are the index of the first and the last order book event in the interval $[t_{k-1,i}, t_{k,i}]$. The tick size δ is equal to 1 cent in our data. Note that in our empirical study OFI is computed from fluctuations in best bid/ask prices and their sizes according to (2.12), because data on individual orders is not available in our main dataset. If that data is available, OFI can be computed according to (2.4). We believe that a computation based on (2.4) can lead to better empirical results because aggressive order terms M^b, M^s will capture information on hidden orders and unreported odd-lot sized orders within the spread, to the extent that aggressive orders interact with hidden orders. Since TAQ data reports only round-lot sized quote changes, we note that units of OFI are round lots (100 shares), and assume in (2.12) that both sides of the market are equally affected by missing quote updates[6].

We define trade imbalance during a time interval $[t_{k-1,i}, t_{k,i}]$ as the difference between volumes of buyer- and seller-initiated trades during that interval, and also define trading volume within that time interval:

$$TI_{k,i} = \sum_{n=N(t_{k-1,i})+1}^{N(t_{k,i})} b_n - s_n \qquad VOL_{k,i} = \sum_{n=N(t_{k-1,i})+1}^{N(t_{k,i})} b_n + s_n, \tag{2.13}$$

where b_n, s_n are sizes of buyer- and seller-initiated trades (in round lots) that occurred at the n-th quote (equal to zero if no trade occurred at that quote). In contrast with

[5] results for other timescales are reported in Section 2.5

[6] As we demonstrate in Section 2.5, neither missing odd-lot sized observations nor potential mis-sequencing of quote updates across different exchanges during NBBO computation change our qualitative findings.

TI, the OFI measure computed using (2.12) does not hinge on trade classification, which is known to be problematic for TAQ data (see Section 2.5 for more details on matching trades with quotes and trade classification). Whereas previous studies [23; 54; 58; 74; 102; 114] focused on trade imbalance[7], the order flow imbalance is a more general measure. It encompasses effects of all order book events, including trades.

For each interval $[T_{i-1}, T_i]$ we also estimate depth by averaging the bid/ask queue sizes right before or right after a price change, consistently with the definition of depth in the stylized order book model:

$$\begin{aligned} D_i \quad = \quad & \frac{1}{2}\left[\frac{\sum_{n=N(T_{i-1})+1}^{N(T_i)}\left(q_n^b \mathbb{1}_{\{P_n^b<P_{n-1}^b\}} + q_{n-1}^b \mathbb{1}_{\{P_n^b>P_{n-1}^b\}}\right)}{\sum_{n=N(T_{i-1})+1}^{N(T_i)} \mathbb{1}_{\{P_n^b\neq P_{n-1}^b\}}}\right. \\ + \quad & \left.\frac{\sum_{n=N(T_{i-1})+1}^{N(T_i)}\left(q_n^s \mathbb{1}_{\{P_n^s>P_{n-1}^s\}} + q_{n-1}^s \mathbb{1}_{\{P_n^s<P_{n-1}^s\}}\right)}{\sum_{n=N(T_{i-1})+1}^{N(T_i)} \mathbb{1}_{\{P_n^s\neq P_{n-1}^s\}}}\right]. \end{aligned} \tag{2.14}$$

[7] Hopman [61] computes the supply/demand imbalance based on limit orders and trades, but not cancelations.

Table 2.1: Descriptive statistics for 50 randomly chosen U.S. stocks

Name	Ticker	Price	Daily volume, shares	Number of best quote updates	Number of trades	Average Spread, ticks	Maximum spread, ticks	Best quote size, shares
Advanced Micro Devices	AMD	9.61	20872996	417204	6687	1	1	103484
Apollo Group	APOL	62.92	1949337	172942	4095	2	5	1525
American Express	AXP	45.21	8678723	559701	7748	1	24	7918
Autozone	AZO	179.03	243197	43682	1081	9	35	750
Bank of America	BAC	18.43	164550168	1529395	15008	1	1	320801
Becton Dickinson	BDX	78.07	1130362	61029	2968	2	5	1530
Bank of New York Mellon	BK	31.77	6310701	285619	5518	1	1	12199
Boston Scientific	BSX	7.13	25746787	309441	6768	1	1	296501
Peabody Energy corp	BTU	47.14	5210642	298616	7267	1	3	2949
Caterpillar	CAT	67.20	6664891	392499	8224	1	2	3835
Chubb	CB	52.22	1951618	149010	3601	1	2	4251
Carnival	CCL	40.16	4275911	215427	5503	1	2	5330
Cincinnati Financial	CINF	29.41	688914	51373	1528	1	2	4157
CME Group	CME	322.83	418955	38504	1412	31	103	541
Coach	COH	41.91	3126469	176795	4458	1	2	4061
ConocoPhillips	COP	56.09	9644544	426614	8621	1	2	8402
Coventry Health Care	CVH	24.16	1157022	79305	2213	1	2	3838
Denbury Resources	DNR	17.88	5737740	263173	4643	1	1	18622
Devon Energy	DVN	66.98	3260982	177006	5805	2	4	1847
Equifax	EFX	35.34	799505	62957	1945	1	3	3925
Eaton	ETN	78.53	1757136	67989	3580	2	6	1254
Fiserv	FISV	52.56	1038311	58304	2208	1	3	2026
Hasbro	HAS	39.48	1322037	86040	2672	1	2	3438
HCP	HCP	32.63	2872521	213045	4357	1	2	4810
Starwood Hotels	HOT	50.59	3164807	150252	5106	2	4	2174
Kohl's	KSS	56.88	3064821	128196	4936	1	3	2688
L-3 Communications	LLL	94.64	670937	72818	2141	2	6	867
Lockheed Martin	LMT	84.14	1416072	88254	3333	2	5	1495
Macy's	M	23.40	8324639	491756	6469	1	1	17567
Marriott	MAR	34.45	5014098	238190	5499	1	2	6511
McAfee	MFE	40.04	2469324	109073	3561	1	2	4018
McGraw-Hill	MHP	34.90	1954576	102389	3261	1	2	4183
Medco Health Solutions	MHS	63.22	2798098	109382	4680	1	3	2534
Merck	MRK	36.03	13930842	448748	7997	1	1	23137
Marathon Oil	MRO	32.33	5035354	341408	5522	1	1	14259
MeadWestvaco	MWV	26.96	1035547	92825	2312	1	3	3741
Newmont Mining	NEM	53.43	5673718	435295	7717	1	2	3847
Omnicom	OMC	41.17	3357585	150800	4359	1	2	6492
MetroPCS Communications	PCS	7.53	4424560	107967	2901	1	1	52304
Pultegroup	PHM	11.80	6834683	262420	4604	1	1	31856
PerkinElmer	PKI	23.98	1268774	78114	2127	1	2	7163
Ryder System	R	44.01	631889	47422	2085	2	5	1147
Reynolds American	RAI	54.44	773387	56236	2076	1	4	2177
Schlumberger	SLB	67.94	9476060	440839	10286	1	2	3942
Teco Energy	TE	16.52	1070815	70318	1807	1	1	14816
Time Warner Cable	TWC	53.21	1770234	88286	3554	2	3	2174
Whirlpool	WHR	97.73	1424264	134152	3348	4	9	958
Windstream	WIN	11.03	2508830	104887	2937	1	1	79834
Watson Pharmaceuticals	WPI	42.51	895967	63094	2024	1	3	2884
XTO Energy	XTO	48.13	7219436	612804	5040	1	7	22479
Grand mean		51.75	7512376	223232	4552	2	6	22665

Table 2.1 presents average mid-prices, daily transaction volumes, daily numbers of best quote updates, daily numbers of trades, spreads and the depths at the best quotes.

2.3.3 Empirical findings

This section reports detailed results for a representative stock, Schlumberger (SLB) and some average results across stocks. Detailed results for other stocks in our sample are presented in Section 2.5. During the sample period the average price of Schlumberger stock was 67.94 dollars and the average daily volume was 947.6 million shares. The daily average number of NBBO quote updates is about 440 thousands, and the average daily number of trades is around 10 thousands. The average spread is one cent, its 95-th percentile is 2 cents and the average best NBBO quote size is 39 round lots (3900 shares).

The model (2.7) is estimated by an ordinary least squares regression:

$$\Delta P_{k,i} = \hat{\alpha}_i + \hat{\beta}_i OFI_{k,i} + \hat{\epsilon}_{k,i}, \tag{2.15}$$

with separate half-hour subsamples indexed by i. Figure 2.5 presents a scatter plot of $\Delta P_{k,i}$ against $OFI_{k,i}$ for one of such subsamples.

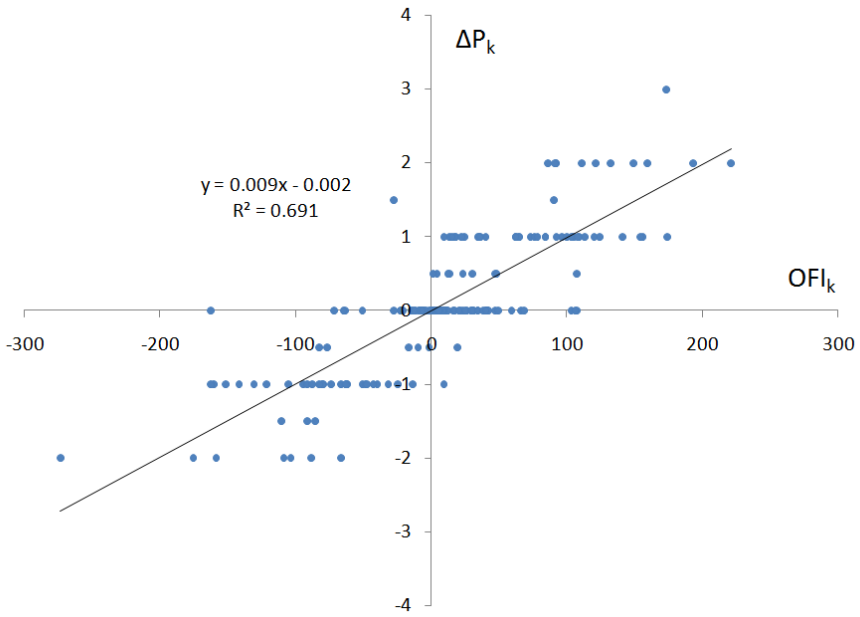

Figure 2.5: Scatter plot of $\Delta P_{k,i}$ against $OFI_{k,i}$ for the Schlumberger stock (SLB), 04/01/2010 11:30-12:00pm.

In general we find that $\hat{\beta}_i$ is statistically significant[8] in 98% of samples, and $\hat{\alpha}_i$ is significant in 10% of samples, which is close to the Type-I error rate. The average t-statistics for $\hat{\alpha}_i, \hat{\beta}_i$ are respectively -0.21 and 16.27 for SLB (cross-sectional averages are -0.02 and 12.08). To check for higher order/nonlinear dependence we estimate an augmented regression:

$$\Delta P_{k,i} = \hat{\alpha}_i^Q + \hat{\gamma}_i OFI_{k,i} + \hat{\gamma}_i^Q OFI_{k,i}|OFI_{k,i}| + \hat{\epsilon}_{k,i}^Q. \tag{2.16}$$

The coefficients $\hat{\gamma}_i^Q$ have an average t-statistic of -0.32 across stocks and are statistically significant only in 17% of our samples. We reject the hypothesis of quadratic (convex or concave) instantaneous price impact, and take this as strong evidence for a linear price impact model (2.7), because other kinds of non-linear dependence would likely be picked up by this quadratic term.

The goodness of fit is surprising for high-frequency data, with an R^2 of 76% for SLB and 65% on average across stocks[9], suggesting that a one-parameter linear model (2.7) performs well regardless of stock-specific features, such as average spread, depth or price level. The definition of R^2 as a percentage of explained variance has an interesting consequence in our case. Since OFI is constructed from order book events taking place only at the best bid/ask, our results show that activity at the top of the order book is the most important factor driving price changes. In Section 2.5 we confirm this by showing that order flow imbalances from deeper order book levels only marginally contribute to short-term price dynamics. Even though large price movements sometimes occur at this timescale, they mostly correspond to large readings of OFI. Figure 2.6 confirms this by demonstrating a relatively low level of excess kurtosis in regression residuals.

When the amount of passive order submissions and cancelations is much larger than the amount of trades, the stylized order book model predicts that trade imbalance TI

[8]Given a relatively large number of observations we use the z-test with a 95% significance level. Since regression residuals demonstrate heteroscedasticity and autocorrelation, Newey-West standard errors are used to compute t-statistics.

[9]We note that OFI includes the contributions e_n of price-changing order book events, leading to a possible endogeneity in the regression (2.15). This problem is inherent to all price impact modeling, because the explanatory variables (events or trades) sometimes mechanically lead to price changes. To test that the high R^2 in our regressions is not due to this endogeneity, we estimated (2.15) on a subsample of stocks, excluding the price-changing events from OFI. With this change the R^2 declined, but remained high, in the 35%-60% region.

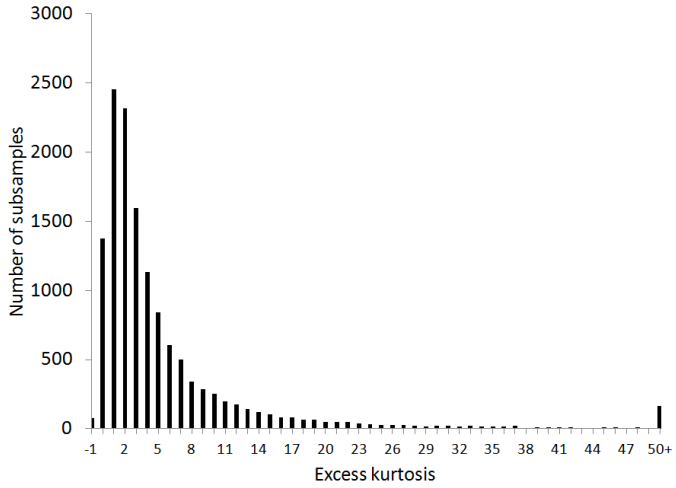

Figure 2.6: Distribution of excess kurtosis in the residuals $\hat{\epsilon}_{k,i}$ across stocks and time.

explains price changes significantly worse than OFI. To empirically confirm this we estimate following regressions using the same half-hour subsamples [10]:

$$\Delta P_{k,i} = \hat{\alpha}_i^T + \hat{\beta}_i^T TI_{k,i} + \hat{\eta}_{k,i} \tag{2.17a}$$

$$\Delta P_{k,i} = \hat{\alpha}_i^D + \hat{\theta}_i^O OFI_k + \hat{\theta}_i^T TI_{k,i} + \hat{\epsilon}_{k,i}^D. \tag{2.17b}$$

When either OFI or TI variable is taken individually, that variable has a statistically significant correlation with price changes. The average t-statistics of slope coefficients in simple regressions (2.15, 2.17a) are, correspondingly 16.27 and 5.31 for SLB (cross-sectional averages are 12.08 and 5.08). The average R^2 for the two regressions are 65% and 32%, respectively, confirming the prediction that relation between price changes and trade imbalance is more noisy. When the two variables are used in a multiple regression (2.17b), the dependence of price changes on trade imbalance becomes much weaker. The average t-statistic of TI coefficient drops to 1.56 for SLB (1.51 across stocks) and it remains statis-

[10] These regressions contain only linear terms, because we found no evidence of non-linear price impacts in our data (for neither OFI nor TI).

tically significant in only 47% of SLB samples (43% of all stock samples). The dependence on OFI remains strong with an average t-statistic 13.91 for SLB (9.53 across stocks), and the coefficient is statistically significant in almost all samples. We conclude that OFI explains price movements better than trade imbalance, and OFI is a more general measure of supply/demand imbalance because it adequately includes the effect of trade imbalance.

Finally, we use time series of D_i and $\hat{\beta}_i$ for each stock to estimate the relation (2.8) with the following two regressions:

$$\log \hat{\beta}_i = \alpha_{\hat{L},i} - \hat{\lambda} \log D_i + \hat{\epsilon}_{L,i}, \tag{2.18}$$

$$\hat{\beta}_i = \alpha_{\hat{M},i} + \frac{\hat{c}}{D_i^{\hat{\lambda}}} + \hat{\epsilon}_{M,i}. \tag{2.19}$$

Both regressions are estimated using ordinary least squares[11]. For SLB we find $\hat{c} = 0.56, \hat{\lambda} = 1.08$ and an R^2 of (2.18) is 92%. The results for all stocks are shown in Table 2.2. We observe that depth significantly correlates with price impact coefficients for the vast majority of stocks, confirming our intuition that $\frac{1}{2\beta_i}$ is the implied order book depth. Interestingly, estimates $\hat{c}, \hat{\lambda}$ across stocks are very close to values predicted by the stylized order book model. With the t-statistics[12] in Table 2.4 the null hypotheses $\{c = 0.5\}$ and $\{\lambda = 1\}$ cannot be rejected for most stocks based on conventional significance levels. The restricted model with $\lambda = 1$ also demonstrates a good quality of fit, making this a good approximation[13]. Figure 2.7 illustrates these results with a log-log scatter plot for D_i and $\hat{\beta}_i$. Some stocks (namely APOL, AZO and CME) have poor fits in regression (2.18), mainly due to outliers in the dependent variable. After removing these outliers and re-estimating the regression, the estimates $\hat{c}, \hat{\lambda}$ for these stocks fell in line with estimates for other stocks.

[11]We note that an estimate $\hat{\lambda}_i$ is used in regression (2.19). This "plug-in" approach leads to potential errors in explanatory variable, and standard errors for $\hat{c}$ may be underestimated. However, the good quality of fit in regression (2.18) with an average R^2 of 76% indicates that $\hat{\lambda}_i$ are estimated with good precision. We believe that errors in variable $\frac{1}{D_i^{\hat{\lambda}}}$ are small and do not affect our results.

[12]Since the residuals of these regressions appear to be autocorrelated, the t-statistics are computed with Newey-West standard errors.

[13]The squared correlation between $\Delta P_{k,i}$ and $\frac{OFI_{k,i}}{2D_i}$ averaged across all subsamples in our data is 0.6523, very close to the average R^2 in (2.15).

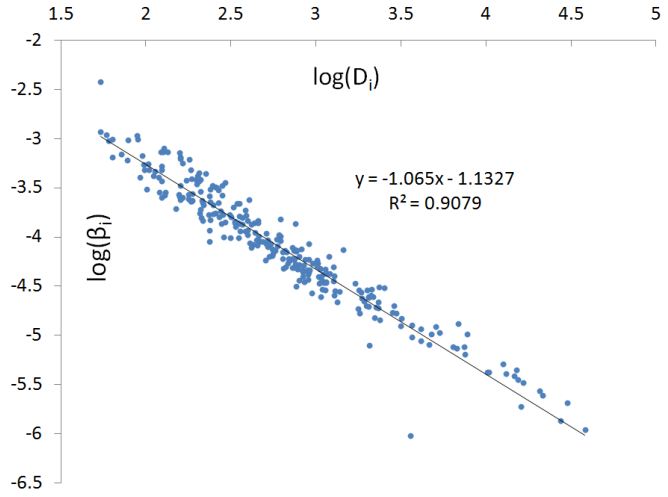

Figure 2.7: Log-log scatter plot of the price impact coefficient estimate $\hat{\beta}_i$ against average market depth D_i for the Schlumberger stock (SLB).

To assess the stability of these findings, we re-estimated (2.18,2.19) with observations pooled across days but not across intraday time intrervals, resulting in 13 estimates $\hat{c}_i, \hat{\lambda}_i$ for each stock. Although these estimates demonstrate some diurnal variability, they are relatively stable and most of variability in price impact coefficients is explained by variations in depth (e.g. see Figure 2.9).

We repeated the analysis with different depth variables, taking D_i to be equal to arithmetic or geometric average of queue sizes over the i-th time interval. Overall, the results were the same, except for the level of $\hat{c}$ estimates, which were about 40% lower across stocks for the arithmetic average depth, and even lower for the geometric average. The systematic difference in these coefficients implies that taking an arbitrary measure of depth (such as arithmetic average of queue sizes) as a proxy of price impact may lead to significant biases, i.e. one would dramatically under- or over-estimate price impact in a given stock. Instead of looking at arbitrary depth measures, we suggest computing price impact coefficients β_i and/or implied depth $\frac{1}{2\beta_i}$ to precisely characterize price sensitivity to order flow.

2.4 Applications

2.4.1 Monitoring adverse selection

Time intervals that are involved in modern high-frequency trading applications are usually so short that price changes are relatively infrequent events. Therefore price changes provide a very coarse and limited description of market dynamics. However, OFI tracks best bid and ask queues and fluctuates on a much faster timescale than prices. It incorporates information about build-ups and depletions of order queues and it can be used to interpolate market dynamics between price changes (see Figure 2.8 for example). Our results confirm that such interpolation is in fact valid because OFI closely approximates price changes over short time intervals (e.g. results for 50 millisecond time intervals are shown in Section 2.5). To study one possible application of OFI for high-frequency trading we turn to our auxiliary dataset, because it contains accurate timestamps up to a millisecond.

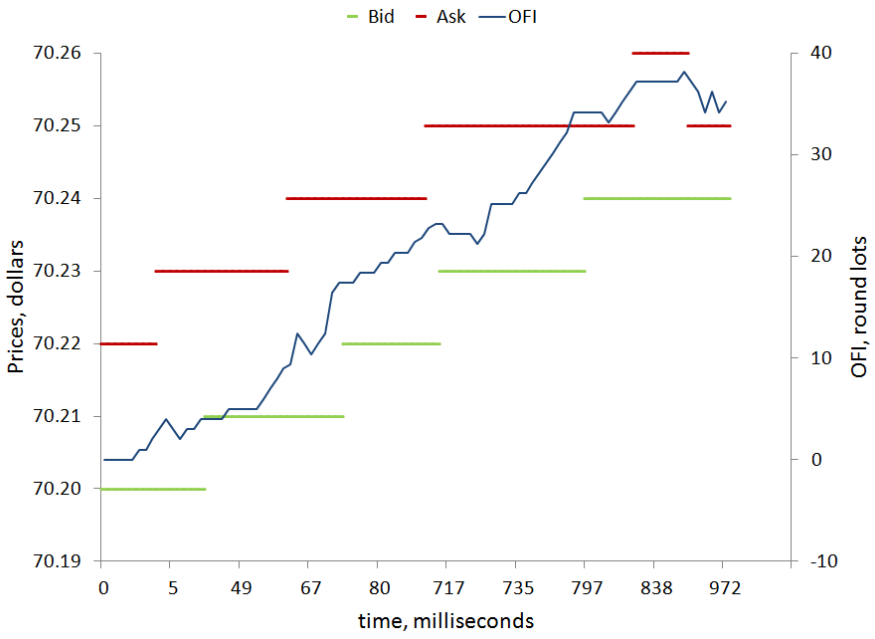

Figure 2.8: Price dynamics and cumulative OFI on NASDAQ for a 1-second time interval starting at 11:16:39.515 on 04/28/2010, Schlumberger stock (SLB).

Given the strong link between OFI and price changes, and the positive autocorrelation of OFI over short time intervals (see Figure 2.4), we propose to use it as a measure of adverse selection in the order flow. For example, when a limit order is filled, and its execution was preceded by positive OFI, a positive price change is more likely to happen after the limit order execution. This is because the pre-execution positive OFI is likely to persist in the future, and can lead to a post-execution positive price change. For a limit sell order a positive post-execution price change implies that the order was executed at a loss, i.e. adversely selected.

To test our hypothesis, we consider all limit order executions in our auxiliary dataset. For each execution we compute the pre-execution order flow imbalance OFI_k^{pre} and the post-execution mid price change ΔP_k^{post}. The pre-execution order flow imbalance is computed from best bid and ask quote updates with timestamps in $[t_k - 200, t_k - 1]$ milliseconds, where t_k is the time of the k-th limit order execution. Similarly the post-execution price change is defined as the difference in mid-quote prices between $t_k + 200$ milliseconds and t_k[14] . Then we consider 30-minute subsamples of data indexed by i, and estimate the following regression:

$$\Delta P_{k,i}^{post} = \alpha_i^p + \beta_i^p OFI_{k,i}^{pre} + \epsilon_{k,i}^p. \tag{2.20}$$

The average R^2 of these regressions across a month is 2.93%, the average t-statistic[15] of β_i^p is 2.68 and this coefficient is significant at a 5% level in 63% of subsamples. The average β_i^p is 0.0105. We conclude that pre-execution OFI are positively correlated with post-execution price changes.

We also estimated regression (2.20) with 50- and 100-millisecond time intervals for pre- and post-execution variables, and obtained similar results, with stronger correlations for smaller time intervals[16]. When we split $OFI_{k,i}^{pre}$ into multiple order flow imbalance variables over non-overlapping subintervals of $[t_k - 200, t_k - 1]$, we find that only the variable closest

[14] If there are multiple quotes with timestamp $t_k + 200$ or t_k, we take the last one.

[15] Here we also use Newey-West standard errors because residuals demonstrate significant autocorrelation.

[16] For instance, with 50-millisecond time intervals the average t-statistic of β_i^p is 3.41 and this coefficient is significant in 75% of samples. The average R^2 becomes 3.32%

to t_k - the execution time - is statistically significant and positively correlated with post-execution price change. These results suggest that limit order traders need to actively monitor order flows and react to emerging order flow imbalances as quickly as possible to avoid being adversely selected.

2.4.2 Intraday volatility dynamics

The link between price impact and market depth established here has important implications for intraday volatility. Market depth is known to follow a predictable diurnal pattern ([2], [83]), and equation (2.8) implies that instantaneous price impact must also have a *predictable* intraday pattern. To demonstrate it, we averaged $\hat{\beta}_i$ for each stock and each intraday half-hour interval across days, resulting in diurnal effects for that stock, normalized these effects by a grand average $\hat{\beta}_i$ for that stock and averaged normalized diurnal effects across stocks. The same procedure was repeated for depths D_i. We also re-estimated (2.18,2.19) with observations pooled across days but not across intraday time intrervals, resulting in 13 estimates $\hat{\lambda}_i, \hat{c}_i$ for each stock. The overall average diurnal effects for these quantities are shown on Figure 2.9.

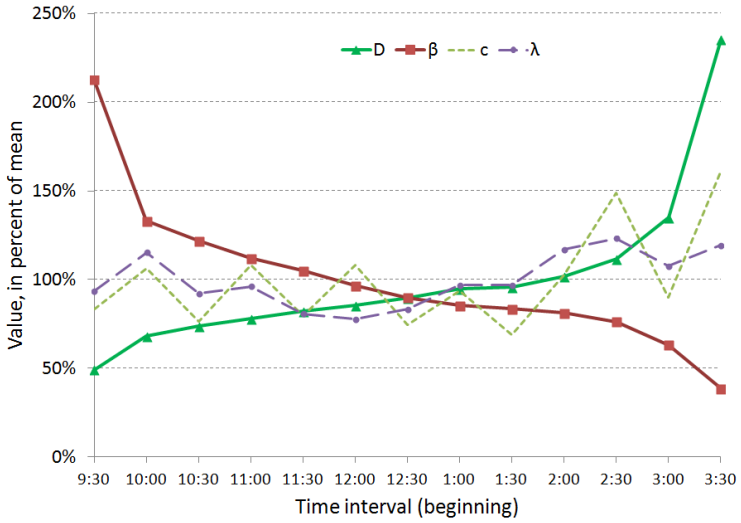

Figure 2.9: Diurnal effects in the price impact coefficient $\hat{\beta}_i$, the average depth D_i and the parameters $\hat{c}_i, \hat{\lambda}_i$. Most of the intraday variation in price impact coefficients comes from variations in depth, while parameters $\hat{c}_i, \hat{\lambda}_i$ are relatively more stable.

We found that between 9:30 and 10am the depth is two times lower than on average, indicating that the market is relatively shallow. In a shallow market, incoming orders can easily affect mid-prices and price impact coefficients between 9:30 and 10am are in fact two times higher than on average. Moreover, price impact coefficients between 9:30 and 10am are five times higher than between 3:30 and 4pm.

The intraday pattern in price impact can be used to explain intraday patterns in price volatility, observed by many studies ([2], [9], [55],[89]). Similarly to the price impact coefficient and the market depth, we computed the intraday patterns in variances of $\Delta P_{k,i}$ and $OFI_{k,i}$, using our half-hour subsamples. Taking the variance on both sides in equation (2.7), we obtain a link between $var[\Delta P_{k,i}]$, $var[OFI_{k,i}]$ and β_i:

$$var[\Delta P_{k,i}] = \beta_i^2 var[OFI_{k,i}] + var[\epsilon_{k,i}] \tag{2.21}$$

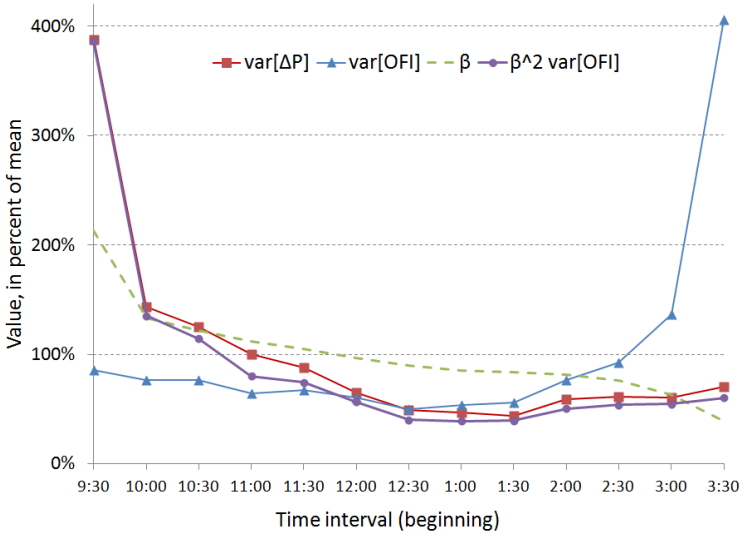

Figure 2.10: Diurnal variability in variances $var[\Delta P_{k,i}]$, $var[OFI_{k,i}]$, the price impact coefficient $\hat{\beta}_i$ and the expression $\beta_i^2 var[OFI_k]_i$.

The average variance patterns are plotted on Figure 2.10. Notice that price volatility has a sharp peak near the market open, while volatility of OFI peaks near the market close. The latter peak is offset by low price impact, which gradually declined throughout the day. For the i-th half-hour interval, equation (2.21) implies that $var[\Delta P_{k,i}] \approx \hat{\beta}_i^2 var[OFI_{k,i}]$

because $var[\epsilon_{k,i}]$ is relatively small, which is also demonstrated[17] on Figure 2.10.

Since the R^2 in regression (2.15) is high, the ratio $\frac{var[\epsilon_{k,i}]}{var[OFI_{k,i}]}$ is small, and we can rewrite (2.21) as $\beta_i \approx \frac{\sigma_{P,i}}{\sigma_{O,i}}$, where $\sigma_{P,i} = \sqrt{var[\Delta P_{k,i}]}$ and $\sigma_{O,i} = \sqrt{var[OFI_{k,i}]}$. This bears strong resemblance to the definition of Kyle's λ (see [79]) - a metric that is used in the asset pricing literature to gauge liquidity risk (see [8] and references therein). This metric is traditionally estimated as a slope β_i^L in regression (2.17a), but our analysis suggests that β_i is a better estimate. Although one could also write $\beta_i^L \approx \frac{\sigma_{P,i}}{\sigma_{T,i}}$, where $\sigma_{T,i} = \sqrt{var[TI_{k,i}]}$, this would be a poorer approximation because $\frac{var[\eta_{k,i}]}{var[TI_{k,i}]} > \frac{var[\epsilon_{k,i}]}{var[OFI_{k,i}]}$ as shown by R^2 values in Table 2.2.

The intraday pattern in price variance was explained in an earlier study [89] using a structural model. The authors argued that price volatility is higher in the morning because of a higher inflow of public and private information. In another study [55] the morning peak of price volatility is explained mostly by higher intensity of public information. Both studies agree that the impact of trades is larger in the morning. Our model contributes to this discussion by explaining the peak of price volatility using tangible quantities, rather than unobservable information variables. Our findings also suggest that price impact and information asymmetry may be, in fact, two sides of the same coin. If there is more private information in the morning than in the evening and if limit order traders are aware of this information asymmetry, their participation will likely diminish in the morning, leading to lower depth near market open. At the same time, low depth implies a higher price impact in our model, making the information advantages harder to realize at the market open.

[17] $\hat{\beta}_i^2 var[OFI_{k,i}]$ was computed from the average patterns of $\hat{\beta}_i$ and $var[OFI_{k,i}]$

2.4.3 Volume and volatility

The positive correlation between magnitudes of price changes and trading volume is empirically confirmed by many authors (see [72] for a review). Recently, trading volume became an important metric for order execution algorithms - these algorithms often attempt to match a certain percentage of the total traded volume to reduce the price impact. However, it remains unclear whether trading volume truly determines the magnitude of price moves and whether it is a good metric for price impact. Casting doubt on this assertion, it was found in [70] that the relation between *daily* volatility and *daily* volume is essentially due to the number of trades and not the volume per se (also see [22] for a following discussion).

We provide further evidence that volume is not a driver of price volatility, now on *intraday* timescales. First, we prove that even when prices are purely driven by OFI and not by volume, a concave relation between magnitude of price changes and transaction volume emerges as an artifact due to aggregation of data across time. Second, we confirm that such relation exists in the data, but it becomes statistically insignificant after accounting for magnitude of OFI.

Comparing the definitions of VOL and OFI we note that both quantities are sums of random variables. As the aggregation time window $[t_{k-1}, t_k]$ becomes progressively larger, the behavior of these sums (under certain assumptions) will be governed by the Law of Large Numbers and the Central Limit Theorem. We consider a general time interval $[0, T]$ and denote by $N(T)$ the number of order book events during that time interval. We also denote by $OFI(T)$ and $VOL(T)$, respectively, the order flow imbalance and the traded volume during $[0, T]$. The following proposition shows a link between $OFI(T)$ and $VOL(T)$ as T grows.

Proposition 1 *Assume that*

1. $\frac{N(T)}{T} \to \Lambda$, *as* $T \to \infty$, *where* Λ *is the average arrival rate of order book events.*

2. $\{e_i\}_{i\geq 1}$ *form a covariance-stationary sequence and have a linear-process representation* $e_i = \sum_{j=0}^{\infty} a_j Y_{i-j}$, *where* Y_i *is a two-sided sequence of i.i.d random variables with* $E[Y_i] = 0$ *and* $E[Y_i^2] = 1$, *and* a_j *is a sequence of constants with* $\sum_{j=0}^{\infty} a_j^2 = \sigma^2 < \infty$. *Moreover,* $cov(e_1, e_{1+n}) \sim cn^{2(H-1)}$ *as* $n \to \infty$, *where* $0 < H < 1$ *is a constant that governs the decay of the autocorrelation function.*

3. $\{w_i\}_{i\geq 1}, w_i = b_i + s_i$ *are random variables with a finite mean* $\mu\pi$, *where* π *is the proportion of order book events that correspond to trades and* μ *is the mean trade size.* $E|w_i|^p < \infty$ *for some* $p > 1$ *and* $\sum_{N\geq 1} \frac{1}{N}(E|\frac{1}{N}\sum_{i\leq N} w_i|^q)^{r/q} < \infty$ *for some* r, q *such that* $0 < r \leq q \leq \infty$ *and* $r/q \leq 1 - 1/p$.

$$\text{Then} \qquad \frac{(\mu\pi)^H}{\sigma}\frac{OFI(T)}{VOL^H(T)} \overset{T\to\infty}{\Rightarrow} \xi \sim N(0,1),$$

where $\Rightarrow$ *denotes convergence in distribution.*

Proof: First, we note that Assumption (1) ensures $N(T) \to \infty$ as $T \to \infty$. With this we can use Assumption (3) and apply the law of large numbers for weakly dependent variables (e.g. see Theorem 7 in [88]) to the traded volume.

$$\frac{VOL(T)}{N(T)} = \frac{\sum_{i=1}^{N(T)} w_i}{N(T)} \to \mu\pi, w.p.1, \text{ as } T \to \infty \tag{2.22}$$

Second, event contributions e_i have a finite variance σ^2 and, using Assumption (2), we apply a central limit theorem for strongly dependent sequences (see Chapter 4.6 in [118]):

$$\frac{OFI(T)}{\sigma N^H(T)} \equiv \frac{\sum_{i=1}^{N(T)} e_i}{\sigma N^H(T)} \Rightarrow \xi, \text{ as } T \to \infty, \tag{2.23}$$

where $\xi \sim N(0,1)$ is a standard normal random variable. Although the denominator $\sigma N^H(T)$ is random, it goes to infinity by assumption (1) and Anscombe's lemma ensures that we can use such a normalization in the central limit theorem [36, Lemma 2.5.8]. Since the function $g(x) = x^H, H > 0, x \geq 0$ is continuous, the convergence in (2.22) takes place

almost-surely and the limit in (2.22) is deterministic, we can combine (2.22) and (2.23) in the following way:

$$\frac{(\mu\pi)^H}{\sigma}\frac{OFI(T)}{VOL^H(T)} \equiv \frac{\frac{\sum_{i=1}^{N(T)} e_i}{\sigma N^H(T)}}{\left(\frac{\sum_{i=1}^{N(T)} w_i}{\mu\pi(N(T))}\right)^H} \Rightarrow \xi,\ as\ T \to \infty \quad \blacksquare \tag{2.24}$$

Proposition 1 implies that as the time interval $[0, T]$ increases and includes a progressively larger enough number of order book events, we can use an approximation for $OFI(T)$:

$$OFI(T) \sim \xi\frac{\sigma}{(\mu\pi)^H}VOL^H(T) \simeq N\left(0, \frac{\sigma^2}{(\mu\pi)^{2H}}VOL^{2H}(T)\right) \tag{2.25}$$

If all time intervals $[t_{k-1,i}, t_{k,i}]$ are large enough to support this approximation then we can substitute (2.25) into (2.7) to obtain

$$\Delta P_{k,i} \sim N\left(0, \frac{\sigma^2\beta_i^2}{(\mu\pi)^{2H}}VOL_{k,i}^{2H} + \sigma_i\right),$$

where $\sigma_i = var[\epsilon_{k,i}]$. Note that even if $\sigma_i = 0$, i.e. even if volume cannot affect price volatility through the residual variance, Proposition 1 predicts a spurious relation between price volatility and volume.

Interestingly, the recent theory of market microstructure invariants (see [80]) also predicts a relation between the volatility of order flow imbalance and trading volume. In their analysis, order flow imbalance is defined differently based on unobservable "bets", however it is natural to assume positive correlation between OFI and the imbalance of "bets", since the latter reach exchanges in form of actual orders.

We can recast this statement in a testable form for the magnitudes (absolute values) of price changes. Assuming $\epsilon_{k,i} \approx 0$, the scaling argument in Proposition 1 together with our linear price impact model imply that

$$|OFI_{k,i}| \approx \frac{\sigma}{(\mu\pi)^H}VOL_{k,i}^H|\xi_{k,i}| \tag{2.26}$$

$$|\Delta P_{k,i}| \approx \frac{\beta_i\sigma}{(\mu\pi)^H}VOL_{k,i}^H|\xi_{k,i}| \tag{2.27}$$

We denote by $\theta_i = \frac{\beta_i \sigma}{(\mu\pi)^H}$ and take logarithms in (2.27) to obtain

$$\log|\Delta P_{k,i}| = \log\hat{\theta}_i + \hat{H}_i \log VOL_{k,i} + \log|\hat{\xi}_{k,i}| \tag{2.28}$$

Based on Proposition 1, we expect this relation to be indirect (i.e. come through $|OFI_{k,i}|$) and noisy. To confirm this empirically, we estimate three regressions[18]:

$$|\Delta P_{k,i}| = \hat{\alpha}_i^O + \hat{\beta}_i^O |OFI_{k,i}| + \hat{\epsilon}_{k,i}^O \tag{2.29a}$$

$$|\Delta P_{k,i}| = \hat{\alpha}_i^V + \hat{\beta}_i^V VOL_{k,i}^{\hat{H}_i} + \hat{\epsilon}_{k,i}^V \tag{2.29b}$$

$$|\Delta P_{k,i}| = \hat{\alpha}_i^W + \hat{\phi}_i^O |OFI_{k,i}| + \hat{\phi}_i^V VOL_{k,i}^{\hat{H}_i} + \hat{\epsilon}_{k,i}^W \tag{2.29c}$$

These regressions are estimated for every half-hour subsample with the exponents $\hat{H}_i$ pre-estimated by (2.28). The averages of $\hat{H}_i$ and their standard deviation for each stock are presented on the left panel in Table 2.5. The exponent varies considerably across stocks and time, but is generally below 1/2 in our data. The average results of regressions (2.29a-2.29c) for each stock are presented on the middle and right panels. We observe that $|OFI_{k,i}|$ explains the magnitude of price moves better than $VOL_{k,i}^{\hat{H}_i}$. Although both variables appear to be statistically significant when taken individually, the t-statistics for $VOL_{k,i}$ drop to marginally significant levels in the multiple regression. Thus, the dependence between absolute values of price moves and traded volume seems to come mostly from correlation between $VOL_{k,i}$ and $|OFI_{k,i}|$. Interestingly, the number of trades variable (suggested in [70]) is also statistically significant on a stand-alone basis, but becomes insignificant when added to (2.29c) as a third variable. Given the recent proliferation of order splitting, the size of most orders is equal to one lot, so $VOL_{k,i}$ is almost the same as the number of trades variable.

[18]Here we estimate linear regressions rather than log-linear ones to directly test whether the effect of VOL is consumed by $|OFI|$ variable

2.5 Robustness checks

2.5.1 Cross-sectional evidence

This section presents various robustness checks for our model. First of all, we re-estimate the regressions from previous sections on a larger sample of 50 randomly selected stocks and find that results fall in line with those for the representative stock. Detailed results for each stock are provided below.

Table 2.2: Relation between price changes and order flow imbalance.

Ticker	Average results							Hypothesis testing		
	$\hat{\alpha}$	$t(\hat{\alpha})$	$\hat{\beta}_i$	$t(\hat{\beta}_i)$	$\hat{\gamma}_i^Q$	$t(\hat{\gamma}_i^Q)$	R^2	$\{\alpha_i \neq 0\}$	$\{\beta_i \neq 0\}$	$\{\gamma_i^Q \neq 0\}$
AMD	-0.0032	-0.24	0.0008	11.10	1.4E-07	0.93	64%	0%	100%	36%
APOL	0.0038	0.13	0.0555	10.74	-2.2E-04	-2.42	63%	17%	96%	6%
AXP	0.0019	0.11	0.0082	14.12	-3.8E-06	-1.37	69%	16%	100%	8%
AZO	0.0101	0.34	0.1619	7.02	-9.3E-04	-1.40	47%	25%	99%	6%
BAC	-0.0018	-0.13	0.0002	19.08	1.9E-09	-0.08	79%	3%	100%	14%
BDX	-0.0008	-0.07	0.0536	10.77	-1.1E-04	-0.74	63%	12%	100%	12%
BK	-0.0078	-0.26	0.0069	15.56	-4.0E-06	-0.89	74%	6%	100%	8%
BSX	0.0000	0.01	0.0003	7.55	7.8E-08	3.51	58%	3%	88%	51%
BTU	0.0048	0.15	0.0242	14.75	-3.5E-05	-2.05	72%	16%	100%	3%
CAT	0.0147	0.30	0.0194	14.80	-1.9E-05	-1.72	71%	19%	100%	5%
CB	-0.0086	-0.05	0.0191	12.61	-3.5E-07	-0.04	64%	10%	100%	18%
CCL	-0.0067	-0.24	0.0140	14.16	-1.2E-05	-1.03	70%	7%	100%	11%
CINF	-0.0030	-0.02	0.0260	11.66	-7.0E-06	0.38	70%	4%	99%	30%
CME	0.0506	0.06	0.6262	5.46	-7.2E-03	-1.66	35%	18%	96%	7%
COH	-0.0221	-0.54	0.0179	13.13	-1.7E-05	-1.18	69%	5%	100%	7%
COP	-0.0008	0.10	0.0084	12.79	-5.8E-06	-1.86	68%	13%	100%	5%
CVH	-0.0034	-0.07	0.0217	11.74	7.6E-06	0.37	65%	7%	99%	20%
DNR	-0.0008	-0.07	0.0045	13.78	-1.3E-07	0.19	69%	5%	99%	22%
DVN	0.0112	0.20	0.0370	12.11	-1.0E-04	-2.72	65%	19%	100%	2%
EFX	-0.0032	-0.06	0.0222	9.47	6.4E-05	0.87	56%	6%	99%	32%
ETN	-0.0076	0.10	0.0712	11.01	-2.3E-04	-1.81	65%	17%	100%	4%
FISV	-0.0002	0.10	0.0397	11.09	-2.3E-05	-0.28	63%	10%	100%	16%
HAS	-0.0031	-0.01	0.0222	12.36	4.7E-06	0.28	67%	6%	100%	23%
HCP	-0.0078	-0.21	0.0150	13.82	-1.4E-05	-0.63	67%	5%	100%	10%
HOT	-0.0012	0.05	0.0345	12.94	-7.2E-05	-2.06	68%	14%	100%	4%
KSS	-0.0030	-0.05	0.0317	14.10	-5.4E-05	-1.38	71%	13%	100%	5%
LLL	0.0160	0.42	0.1000	12.34	-3.8E-04	-1.56	67%	22%	98%	7%
LMT	0.0006	0.00	0.0520	14.14	-1.2E-04	-1.49	72%	17%	100%	4%
M	-0.0010	0.07	0.0043	16.61	8.8E-08	0.15	75%	6%	100%	19%
MAR	-0.0039	0.02	0.0121	15.10	-4.1E-06	-0.43	71%	10%	100%	10%
MFE	0.0087	0.22	0.0205	13.19	-3.8E-05	-0.63	68%	11%	100%	11%
MHP	-0.0073	-0.18	0.0211	12.41	5.8E-06	0.18	68%	5%	99%	24%
MHS	-0.0055	-0.20	0.0334	11.97	-8.3E-05	-1.64	66%	12%	100%	4%
MRK	-0.0065	-0.26	0.0032	13.26	-5.4E-07	-0.61	69%	4%	100%	14%
MRO	0.0018	0.12	0.0058	14.16	-3.6E-07	0.32	69%	8%	100%	23%
MWV	-0.0011	0.02	0.0205	12.55	-1.7E-05	-0.31	68%	9%	100%	17%
NEM	-0.0102	-0.26	0.0170	13.90	-1.9E-05	-2.15	71%	12%	100%	5%
OMC	-0.0099	-0.36	0.0144	12.40	-4.5E-06	-0.19	65%	4%	100%	20%
PCS	-0.0006	-0.05	0.0015	6.52	1.8E-06	3.79	53%	2%	86%	51%
PHM	0.0006	0.02	0.0027	11.27	8.4E-07	1.20	66%	3%	99%	36%
PKI	-0.0004	-0.05	0.0102	7.96	4.1E-05	2.15	53%	3%	96%	51%
R	0.0006	0.03	0.0667	10.90	3.7E-05	-0.21	63%	14%	100%	16%
RAI	-0.0070	-0.10	0.0396	11.39	2.6E-05	-0.03	66%	9%	100%	19%
SLB	-0.0077	-0.21	0.0198	16.27	-1.8E-05	-1.67	76%	10%	100%	2%
TE	0.0011	0.05	0.0049	7.76	1.4E-05	3.27	54%	4%	91%	55%
TWC	-0.0130	-0.15	0.0384	12.24	-5.6E-05	-0.73	64%	12%	99%	9%
WHR	0.0628	0.73	0.1278	11.10	-3.3E-04	-1.44	65%	25%	100%	7%
WIN	-0.0004	-0.04	0.0009	4.32	1.5E-06	3.98	44%	1%	72%	43%
WPI	-0.0090	-0.27	0.0270	11.46	2.9E-05	0.26	66%	5%	99%	23%
XTO	-0.0088	-0.25	0.0029	13.26	2.7E-07	0.48	65%	3%	100%	28%
Average	0.0002	-0.02	0.0398	12.08	-2.0E-04	-0.32	65%	10%	98%	17%

Table 2.2 presents a cross-section of results (averaged across time) for regressions:

$\Delta P_{k,i} = \hat{\alpha}_i + \hat{\beta}_i OFI_{k,i} + \hat{\epsilon}_{k,i}$,
$\Delta P_{k,i} = \hat{\alpha}_i^Q + \hat{\beta}_i^Q OFI_{k,i} + \hat{\gamma}_i^Q OFI_{k,i}|OFI_{k,i}| + \hat{\epsilon}_{k,i}^Q$,

where $\Delta P_{k,i}$ are the 10-second mid-price changes in ticks and $OFI_{k,i}$ are the contemporaneous order flow imbalances. These regressions were estimated using 273 half-hour subsamples (indexed by i) for each stock and their outputs were averaged across subsamples. Each subsample typically contains about 180 observations (indexed by k). The t-statistics were computed using Newey-West standard errors. For brevity, we report the R^2, the average $\hat{\alpha}_i$ and the average $\hat{\beta}_i$ only for the first regression (with a single $OFI_{k,i}$ term). There is almost no difference between averages of estimates $\hat{\beta}_i$ and $\hat{\beta}_i^Q$ and the R^2 in two regressions. The last three columns report the percentage of samples where the coefficient(s) passed the z-test at the 5% significance level.

Table 2.3: Comparison of order flow imbalance and trade imbalance.

Ticker	Order flow imbalance				Trade imbalance				Both covariates					
	R^2	$t(\hat{\beta}_i)$	$\{\beta_i \neq 0\}$	F	R^2	$t(\hat{\beta}_i^T)$	$\{\beta_i^T \neq 0\}$	F	R^2	$t(\hat{\theta}_i^O)$	$t(\hat{\theta}_i^T)$	$\{\theta_i^O \neq 0\}$	$\{\theta_i^T \neq 0\}$	F
AMD	64%	11.10	100%	382	39%	5.06	95%	140	67%	7.64	1.59	99%	45%	214
APOL	63%	10.74	96%	396	30%	5.04	95%	83	66%	8.95	1.58	96%	44%	211
AXP	69%	14.12	100%	449	34%	5.55	92%	101	71%	11.31	1.90	100%	55%	241
AZO	47%	7.02	99%	179	30%	4.88	96%	87	54%	5.78	2.87	98%	81%	118
BAC	79%	19.08	100%	774	45%	7.03	98%	157	80%	13.55	0.80	99%	25%	397
BDX	63%	10.77	100%	362	28%	4.85	92%	79	65%	8.90	1.53	100%	46%	195
BK	74%	15.56	100%	610	36%	5.36	93%	117	75%	11.90	0.80	100%	26%	313
BSX	58%	7.55	88%	338	31%	3.60	71%	106	62%	5.74	0.88	82%	24%	189
BTU	72%	14.75	100%	527	35%	6.03	97%	103	74%	11.96	1.63	100%	44%	277
CAT	71%	14.80	100%	498	33%	5.75	94%	94	72%	12.14	1.55	100%	46%	262
CB	64%	12.61	100%	378	33%	5.47	95%	102	66%	9.41	1.57	99%	44%	202
CCL	70%	14.16	100%	478	32%	5.31	94%	93	71%	11.44	1.17	100%	37%	247
CINF	70%	11.66	99%	552	39%	5.35	96%	141	72%	8.28	1.28	98%	40%	297
CME	35%	5.46	96%	112	24%	4.31	88%	63	44%	4.73	2.78	96%	71%	78
COH	69%	13.13	100%	457	29%	4.75	93%	80	70%	11.06	1.12	100%	31%	238
COP	68%	12.79	100%	450	35%	5.69	92%	107	70%	10.25	1.76	100%	49%	240
CVH	65%	11.74	99%	418	35%	5.05	93%	114	67%	8.43	1.35	97%	37%	222
DNR	69%	13.78	99%	471	32%	4.89	92%	101	70%	10.43	1.27	99%	37%	246
DVN	65%	12.11	100%	414	33%	5.57	95%	96	68%	9.61	2.12	98%	60%	226
EFX	56%	9.47	99%	289	31%	4.75	89%	101	60%	7.13	2.26	98%	55%	167
ETN	65%	11.01	100%	389	25%	4.43	86%	69	67%	9.85	1.47	99%	43%	209
FISV	63%	11.09	100%	380	28%	4.82	93%	79	65%	9.08	1.25	100%	38%	201
HAS	67%	12.36	100%	427	32%	5.15	95%	97	68%	9.67	1.17	100%	34%	223
HCP	67%	13.82	100%	417	31%	5.07	90%	91	68%	10.92	1.33	100%	42%	217
HOT	68%	12.94	100%	438	27%	4.75	88%	74	70%	11.00	1.48	100%	40%	231
KSS	71%	14.10	100%	525	31%	5.16	93%	91	72%	11.86	1.14	100%	37%	274
LLL	67%	12.34	98%	485	36%	6.00	95%	117	70%	9.68	2.14	98%	57%	270
LMT	72%	14.14	100%	516	35%	5.80	96%	105	73%	11.35	1.83	100%	51%	277
M	75%	16.61	100%	640	35%	5.10	93%	108	76%	12.80	1.13	100%	38%	330
MAR	71%	15.10	100%	498	34%	5.54	95%	105	72%	11.41	1.18	100%	36%	258
MFE	68%	13.19	100%	463	31%	4.82	88%	93	69%	10.27	0.89	100%	30%	239
MHP	68%	12.41	99%	489	31%	5.09	93%	96	70%	9.94	1.04	99%	33%	257
MHS	66%	11.97	100%	414	28%	4.81	89%	80	68%	10.03	1.50	99%	40%	218
MRK	69%	13.26	100%	451	31%	4.99	92%	93	70%	10.41	1.02	100%	29%	235
MRO	69%	14.16	100%	465	35%	5.38	96%	104	70%	10.67	1.12	100%	35%	241
MWV	68%	12.55	100%	452	34%	5.30	96%	102	69%	9.66	1.01	100%	33%	237
NEM	71%	13.90	100%	490	34%	5.77	92%	100	72%	11.38	1.90	100%	54%	260
OMC	65%	12.40	100%	411	30%	4.90	93%	88	67%	9.85	1.22	100%	39%	216
PCS	53%	6.52	86%	297	35%	4.08	74%	169	58%	4.47	1.43	81%	35%	195
PHM	66%	11.27	99%	416	35%	4.76	93%	115	68%	8.40	1.22	98%	38%	224
PKI	53%	7.96	96%	263	28%	3.98	82%	89	57%	6.16	1.70	93%	47%	148
R	63%	10.90	100%	352	27%	4.80	96%	71	65%	9.02	1.58	100%	44%	188
RAI	66%	11.39	100%	422	36%	5.60	98%	111	68%	8.64	1.42	100%	43%	224
SLB	76%	16.27	100%	644	32%	5.31	89%	94	77%	13.91	1.56	100%	47%	336
TE	54%	7.76	91%	301	37%	4.65	82%	175	60%	5.27	1.96	86%	45%	200
TWC	64%	12.24	99%	377	31%	5.21	86%	93	66%	9.67	1.70	99%	45%	201
WHR	65%	11.10	100%	394	29%	5.03	95%	85	67%	9.27	1.86	100%	52%	217
WIN	44%	4.32	72%	243	41%	4.74	75%	249	58%	2.60	2.55	58%	47%	206
WPI	66%	11.46	99%	437	32%	4.80	93%	100	68%	8.95	1.35	99%	46%	232
XTO	65%	13.26	100%	399	21%	3.78	78%	54	66%	11.72	1.42	100%	40%	209
Average	65%	12.08	98%	429	32%	5.08	91%	103	67%	9.53	1.51	97%	43%	231

Table 2.3 presents the average results of regressions:
$\Delta P_{k,i} = \hat{\alpha}_i + \hat{\beta}_i OFI_{k,i} + \hat{\epsilon}_{k,i}$,
$\Delta P_{k,i} = \hat{\alpha}_i^T + \hat{\beta}_i^T TI_{k,i} + \hat{\eta}_{k,i}$,
$\Delta P_{k,i} = \hat{\alpha}_i^D + \hat{\theta}_i^O OFI_{k,i} + \hat{\theta}_i^T TI_{k,i} + \hat{\epsilon}_{k,i}^D$,
where $\Delta P_{k,i}$ are the 10-second mid-price changes, $OFI_{k,i}$ are the contemporaneous order flow imbalances and $TI_{k,i}$ are the contemporaneous trade imbalances. These regressions were estimated using 273 half-hour subsamples (indexed by i) for each stock and their outputs were averaged across subsamples. Each subsample typically contains about 180 observations (indexed by k). The t-statistics were computed using Newey-West standard errors. For each of three regressions, Table 2.3 reports the average R^2, the average t-statistic of the coefficient(s), the percentage of samples where the coefficient(s) passed the z-test at the 5% significance level and the F-statistic of the regression.

Table 2.4: Relation between the price impact coefficient and market depth.

Ticker	Parameter estimates						Fit measures		
	$\hat{c}$	$\hat{\lambda}$	$t(\hat{c}=0)$	$t(\hat{c}=0.5)$	$t(\hat{\lambda}=0)$	$t(\hat{\lambda}=1)$	R^2	$corr[\hat{\beta},\hat{\bar{\beta}}]^2$	$corr[\hat{\beta},\hat{\bar{\beta}}^*]^2$
AMD	0.53	1.04	31.06	2.0	22.5	1.0	78%	86%	86%
APOL	0.30	0.38	4.59	-3.2	1.1	-1.8	3%	34%	35%
AXP	0.45	1.01	26.27	-3.2	45.8	0.6	90%	87%	87%
AZO	0.45	0.70	5.47	-0.7	5.2	-2.2	14%	17%	16%
BAC	0.87	1.10	31.80	13.6	19.1	1.7	80%	89%	89%
BDX	0.48	1.03	23.91	-1.2	22.2	0.6	74%	71%	71%
BK	0.47	1.04	28.28	-1.9	68.5	2.5	94%	94%	94%
BSX	0.51	1.02	15.23	0.3	24.1	0.4	73%	81%	81%
BTU	0.58	1.10	45.36	6.2	53.8	5.0	93%	90%	90%
CAT	0.48	1.01	35.12	-1.4	20.7	0.2	91%	91%	90%
CB	0.53	1.09	32.41	2.1	63.6	5.5	93%	91%	91%
CCL	0.45	1.04	35.26	-3.6	41.9	1.5	89%	86%	86%
CINF	0.43	1.03	25.48	-4.2	52.5	1.7	93%	90%	90%
CME	1.21	0.35	2.10	1.2	1.4	-2.7	1%	2%	2%
COH	0.61	1.11	15.35	2.9	44.7	4.3	81%	83%	82%
COP	0.32	0.94	13.77	-8.1	22.6	-1.6	82%	79%	79%
CVH	0.54	1.13	26.92	2.2	37.9	4.2	88%	90%	89%
DNR	0.55	1.10	40.77	3.6	44.9	3.9	92%	90%	90%
DVN	0.34	0.91	16.15	-7.8	19.3	-2.0	48%	61%	61%
EFX	0.43	1.05	19.58	-3.0	27.1	1.2	84%	80%	80%
ETN	0.64	1.11	13.55	2.9	20.8	2.1	65%	63%	63%
FISV	0.47	1.04	25.33	-1.7	34.1	1.3	85%	80%	80%
HAS	0.52	1.08	27.80	1.3	49.8	3.8	90%	86%	85%
HCP	0.37	1.00	33.13	-11.3	64.7	0.0	95%	94%	94%
HOT	0.61	1.13	28.19	5.2	37.7	4.3	87%	87%	87%
KSS	0.59	1.09	28.99	4.3	41.7	3.4	90%	85%	85%
LLL	0.57	1.02	15.30	1.9	14.5	0.3	53%	65%	65%
LMT	0.72	1.17	7.93	2.4	15.6	2.3	69%	63%	63%
M	0.52	1.06	24.92	1.0	52.1	3.0	94%	92%	92%
MAR	0.50	1.06	22.26	0.0	52.7	3.1	92%	89%	89%
MFE	0.47	1.06	22.12	-1.3	45.3	2.7	92%	89%	89%
MHP	0.45	1.02	20.58	-2.1	38.1	0.6	83%	78%	78%
MHS	0.71	1.16	19.88	5.9	39.5	5.3	88%	87%	86%
MRK	0.31	0.94	21.38	-12.8	36.3	-2.3	87%	84%	84%
MRO	0.55	1.09	28.87	2.4	51.9	4.2	94%	94%	94%
MWV	0.54	1.13	28.16	2.2	39.3	4.6	90%	87%	87%
NEM	0.51	1.07	31.09	0.4	39.3	2.6	89%	88%	88%
OMC	0.52	1.04	36.61	1.1	19.5	0.7	86%	90%	90%
PCS	0.43	1.06	22.79	-3.4	18.6	1.1	53%	83%	83%
PHM	0.62	1.10	39.56	7.7	36.6	3.3	87%	92%	92%
PKI	0.49	1.14	29.01	-0.5	34.7	4.4	80%	87%	86%
R	0.50	1.05	17.43	-0.1	15.8	0.7	58%	59%	59%
RAI	0.51	1.07	26.19	0.4	47.3	3.1	88%	79%	79%
SLB	0.56	1.08	23.39	2.5	47.6	3.6	92%	94%	93%
TE	0.35	1.10	12.12	-5.1	25.1	2.2	70%	85%	86%
TWC	0.55	1.07	22.29	1.9	18.9	1.2	73%	85%	84%
WHR	1.09	1.25	12.66	6.9	13.4	2.7	51%	54%	53%
WIN	17.21	1.80	13.95	13.5	12.2	5.4	35%	72%	74%
WPI	0.39	0.99	19.57	-5.6	32.6	-0.4	79%	77%	77%
XTO	0.97	1.19	27.70	13.46	35.64	5.77	88%	91%	90%
Grand mean	0.88	1.05	23.55	0.59	33.40	1.99	76%	79%	79%

Table 2.4 presents the results of regressions:

$\log \hat{\beta}_i = \alpha_{\hat{L},i} - \hat{\lambda} \log D_i + \hat{\epsilon}_{L,i}$,
$\hat{\beta}_i = \alpha_{\hat{M},i} + \frac{\hat{c}}{D_i^{\hat{\lambda}}} + \hat{\epsilon}_{M,i}$,

where $\hat{\beta}_i$ is the price impact coefficient for the i-th half-hour subsample and D_i is the average market depth for that subsample. These regressions were estimated for each of the 50 stocks, using 273 estimates of $\hat{\beta}_i$ for that stock, obtained from (2.15). The second regression uses estimates $\hat{\lambda}$ obtained from the first regression. The t-statistics were computed using Newey-West standard errors. The last three columns provide three alternative fit measures - the R^2 of the linear regression (2.18), the squared correlation between $\hat{\beta}_i$ and fitted values $\hat{\hat{\beta}}_i = \frac{\hat{c}}{D_i^{\hat{\lambda}}}$ and the squared correlation between $\hat{\beta}_i$ and $\hat{\hat{\beta}}_i^* = \frac{\hat{c}}{D_i}$.

Table 2.5: Comparison of traded volume and order flow imbalance.

Ticker	Avg	Stdev	Order flow imbalance				Traded volume				Both covariates					
	$\hat{H}$	$\hat{H}$	R^2	$t(\hat{\beta}_i^O)$	$\beta_i^O \neq 0$	F	R^2	$t(\hat{\beta}_i^V)$	$\beta_i^V \neq 0$	F	R^2	$t(\hat{\phi}_i^O)$	$t(\hat{\phi}_i^V)$	$\phi_i^O \neq 0$	$\phi_i^V \neq 0$	F
AMD	0.06	0.08	63%	11.7	100%	356	14%	4.6	87%	34	63%	10.8	1.2	99%	38%	182
APOL	0.24	0.08	53%	9.1	97%	258	25%	6.9	100%	63	57%	7.6	3.3	94%	86%	144
AXP	0.16	0.08	55%	11.3	100%	249	20%	6.8	100%	48	57%	9.7	2.9	100%	82%	133
AZO	0.43	0.22	39%	6.3	98%	131	32%	5.8	100%	93	50%	5.0	3.9	97%	98%	98
BAC	0.09	0.08	73%	17.6	100%	560	24%	6.0	89%	61	74%	15.3	1.3	97%	40%	285
BDX	0.26	0.10	55%	9.4	100%	261	27%	6.5	100%	71	58%	7.6	3.1	99%	85%	147
BK	0.11	0.07	68%	14.1	100%	437	19%	6.7	97%	46	68%	12.6	2.0	100%	58%	225
BSX	-0.17	2.41	68%	10.3	100%	486	14%	3.4	97%	33	69%	10.1	0.0	99%	13%	246
BTU	0.24	0.07	58%	11.4	100%	283	23%	7.1	99%	57	60%	9.7	2.6	100%	81%	151
CAT	0.22	0.07	56%	11.0	100%	250	19%	6.3	99%	44	57%	9.7	2.3	100%	68%	131
CB	0.19	0.09	56%	11.1	100%	261	23%	6.5	99%	58	58%	9.1	2.8	100%	76%	141
CCL	0.14	0.07	60%	12.2	100%	309	19%	6.7	99%	45	62%	10.8	2.5	100%	77%	162
CINF	0.13	0.12	67%	12.0	100%	505	30%	6.2	98%	85	69%	10.3	2.1	100%	58%	268
CME	0.49	0.24	28%	4.8	98%	78	30%	5.3	100%	83	42%	3.9	4.1	94%	99%	71
COH	0.19	0.07	60%	11.3	100%	299	22%	6.5	99%	52	61%	9.8	2.4	100%	73%	157
COP	0.16	0.07	56%	10.5	100%	277	20%	6.1	97%	49	58%	9.2	2.5	100%	74%	145
CVH	0.18	0.10	62%	11.4	100%	352	27%	6.1	100%	72	64%	9.2	2.4	100%	73%	189
DNR	0.08	0.07	64%	13.4	100%	376	17%	6.4	95%	38	65%	12.0	1.9	99%	57%	193
DVN	0.26	0.07	52%	9.6	97%	236	24%	6.9	100%	59	55%	8.0	3.2	96%	85%	131
EFX	0.20	0.11	52%	9.1	100%	241	26%	5.6	99%	69	56%	7.3	2.8	99%	77%	137
ETN	0.26	0.10	55%	9.1	99%	252	27%	6.6	99%	70	58%	7.6	3.1	98%	85%	142
FISV	0.19	0.11	57%	10.1	100%	284	25%	6.0	100%	65	59%	8.3	2.4	100%	70%	153
HAS	0.20	0.09	61%	11.3	100%	328	26%	6.3	100%	67	63%	9.5	2.5	100%	76%	175
HCP	0.14	0.07	57%	11.8	100%	268	21%	7.1	99%	50	59%	10.0	2.8	100%	80%	143
HOT	0.23	0.08	57%	10.5	99%	263	24%	7.2	100%	60	60%	9.0	3.2	99%	88%	145
KSS	0.24	0.08	60%	11.6	100%	318	25%	6.8	99%	61	62%	9.8	2.6	99%	78%	169
LLL	0.33	0.12	58%	10.3	97%	323	34%	7.2	100%	101	63%	7.9	3.4	96%	92%	188
LMT	0.28	0.09	61%	11.6	100%	327	31%	7.6	100%	85	64%	9.3	3.1	100%	85%	182
M	0.11	0.07	69%	15.2	100%	463	20%	6.3	100%	46	69%	13.5	2.0	100%	63%	238
MAR	0.15	0.07	61%	13.3	100%	324	21%	7.0	99%	50	62%	11.5	2.5	100%	74%	170
MFE	0.16	0.09	60%	11.7	100%	318	24%	7.0	98%	62	62%	9.7	2.6	100%	73%	170
MHP	0.20	0.10	62%	11.6	100%	377	25%	6.1	100%	62	64%	9.7	2.0	100%	56%	199
MHS	0.23	0.08	56%	10.0	100%	258	24%	6.7	100%	58	58%	8.4	2.9	100%	80%	139
MRK	0.10	0.07	62%	12.1	100%	330	17%	5.5	99%	40	63%	10.8	1.9	100%	60%	170
MRO	0.09	0.06	61%	12.7	100%	333	16%	6.4	97%	36	63%	11.5	2.0	100%	56%	172
MWV	0.18	0.10	62%	11.3	100%	330	28%	6.9	100%	75	64%	9.2	2.6	100%	79%	180
NEM	0.20	0.07	56%	10.6	100%	253	20%	6.3	99%	47	58%	9.3	2.6	100%	79%	135
OMC	0.15	0.09	57%	11.0	100%	286	20%	6.4	98%	48	59%	9.4	2.5	100%	75%	151
PCS	0.11	0.18	62%	8.9	100%	411	18%	3.8	98%	54	63%	8.4	0.8	100%	28%	214
PHM	0.07	0.08	64%	11.5	100%	384	15%	5.4	91%	34	65%	10.7	1.2	100%	41%	195
PKI	0.11	0.11	55%	9.0	99%	266	20%	4.8	98%	47	57%	7.8	1.9	98%	55%	141
R	0.27	0.11	56%	9.8	99%	259	28%	6.3	100%	74	59%	7.9	3.1	99%	87%	147
RAI	0.25	0.10	61%	10.6	100%	334	28%	5.9	99%	73	63%	8.8	2.6	100%	75%	182
SLB	0.24	0.07	62%	12.5	99%	330	19%	5.8	97%	46	63%	11.2	1.9	99%	56%	171
TE	0.09	1.69	60%	9.6	100%	371	18%	4.5	84%	48	61%	8.7	1.3	99%	43%	196
TWC	0.25	0.10	55%	10.5	100%	253	27%	6.8	100%	73	58%	8.4	3.1	99%	83%	142
WHR	0.34	0.11	56%	9.2	99%	272	29%	6.6	100%	78	59%	7.5	3.2	98%	88%	156
WIN	0.06	0.26	48%	5.5	86%	340	10%	2.9	50%	34	49%	5.3	0.6	85%	31%	179
WPI	0.22	0.10	61%	11.0	100%	361	28%	5.9	100%	75	64%	9.0	2.4	99%	71%	196
XTO	0.08	0.06	53%	11.3	100%	238	15%	6.6	100%	32	55%	10.0	2.8	100%	82%	125
Average	0.18	0.18	58%	10.9	99%	313	23%	6.1	97%	58	61%	9.3	2.4	99%	70%	168

Table 2.5 presents the average results of regressions:

$$|\Delta P_{k,i}| = \hat{\alpha}_i^O + \hat{\beta}_i^O |OFI_{k,i}| + \hat{\epsilon}_{k,i}^O,$$
$$|\Delta P_{k,i}| = \hat{\alpha}_i^V + \hat{\beta}_i^V VOL_{k,i}^{\hat{H}_i} + \hat{\epsilon}_{k,i}^V,$$
$$|\Delta P_{k,i}| = \hat{\alpha}_i^W + \hat{\phi}_i^O |OFI_{k,i}| + \hat{\phi}_i^V VOL_{k,i}^{\hat{H}_i} + \hat{\epsilon}_{k,i}^W,$$

where $\Delta P_{k,i}$ are the 10-second mid-price changes, $OFI_{k,i}$ are the contemporaneous order flow imbalances and $VOL_{k,i}$ are the contemporaneous trade volumes. The exponents $\hat{H}_i$ were estimated in each subsample beforehand using a logarithmic regression: $\log|\Delta P_{k,i}| = \log\hat{\theta}_i + \hat{H}_i \log VOL_{k,i} + \log|\hat{\xi}_{k,i}|$. These regressions were estimated using 273 half-hour subsamples (indexed by i) for each stock and their outputs were averaged across subsamples. Each subsample typically contains about 180 observations (indexed by k). The t-statistics were computed using Newey-West standard errors. For each of three regressions, Table 2.5 reports the average R^2, the average t-statistic of the coefficient(s), the percentage of samples where the coefficient(s) passed the z-sest at the 5% significance level and the F-statistic of the regression.

2.5.2 Transaction prices

To reconcile our results with earlier studies that operate in transaction time, we repeated regressions (2.15,2.17a,2.17b) with differences between transaction prices $\Delta_L P_k^t = P_k^t - P_{k-L}^t$ for L trades, instead of differences in mid-prices ΔP_k. We picked at random five stocks from our sample (BDX, CB, MHS, PHM and PKI), and computed $\Delta_L P_k^t$ for $L = 2, 5, 10$ trades (we avoided using $L = 1$ because of possible issues with order direction estimation). Using the same inter-trade time intervals we computed concurrent OFI and TI variables. To ensure that there is an ample amount of data for each regression, we pooled data across days for each stock and each intraday time subsample, resulting in 13 samples for each stock over a month of data. The results averaged across time and stocks are presented in Table 2.6 and closely mirror our results for mid prices. The variable OFI_k explains price changes better than TI_k on stand-alone basis. Moreover, the effect of trades on prices seems to be captured by the order flow imbalance, i.e. the variable TI_k loses its statistical significance[19]. when used together with OFI_k in the regression. The increase in R^2 from adding TI_k as an extra regressor is almost nill (0.65%, 0.18%, 0.24% for $L = 1, 2, 5$ respectively).

Table 2.6: Comparison of order flow imbalance and trade imbalance for transaction prices.

Lag	Order flow imbalance				Trade imbalance				Both covariates					
	R^2	$t(\hat{\beta}_i)$	$\{\beta_i \neq 0\}$	F	R^2	$t(\tilde{\beta}_i^T)$	$\{\beta_i^T \neq 0\}$	F	R^2	$t(\tilde{\theta}_i^O)$	$t(\tilde{\theta}_i^T)$	$\{\theta_i^O \neq 0\}$	$\{\theta_i^T \neq 0\}$	F
$L = 2$	14%	15.03	100%	464	1%	2.97	69%	26	15%	14.19	-2.90	100%	71%	245
$L = 5$	38%	16.68	98%	753	8%	4.79	88%	113	39%	15.13	-0.14	98%	14%	379
$L = 10$	51%	14.85	98%	655	13%	4.85	88%	100	51%	13.21	0.70	98%	11%	329

Interestingly, we found that the relation between trade price changes and OFI_k (or TI_k) is sometimes concave. We estimated regressions (2.15) and (2.17a) for trade price changes $\Delta_L P_k^t$ with additional quadratic variable $OFI_k|OFI_k|$ and found that average t-statistics of its coefficient are, respectively -3.02, -4.10 and -3.85 for $L = 1, 2, 5$ trades. The quadratic term is significant at a 5% level in 60%, 74% and 85% of samples for respective values of L, and we did not observe any pattern in these t-statistics, neither across stock nor across time. In the trade imbalance regression the coefficient near quadratic variable $TI_k|TI_k|$ is also significant with average t-statistics -3.64, -5.53, -5.48 for respective lag values and it is

[19]Here we also use Newey-West standard errors because regression residuals have statistically significant autocorrelation

significant in even a larger fraction of samples.

From these results it appears that price impact is concave when prices are sampled at trade times, but it is linear when they are sampled at regular time intervals. This effect may be a consequence of sampling data at special times (i.e. trade times), which introduces systematic biases into the regression. If traders submit large orders when they expect their impact to be minimal, it would lead to a concave (sublinear) price impact. Supporting the idea of a sampling bias, we found that when mid-price changes are sampled at trade times, the price impact of OFI_k is again concave - the quadratic term in the regression is statistically significant in a large fraction of our samples. We also regressed changes in last trade prices sampled regularly at a 1-minute frequency on OFI_k, and observed concave price impact once again. This may again be attributed to a dependent variable bias - since trades are relatively infrequent, for many time intervals the trade prices are going to be stale and trade price changes are equal to zero, while mid price changes are not.

2.5.3 Order flow at higher order book levels

The level of detail in our Level 2 auxiliary data set allows us to analyze contributions of order flows at different price levels to price formation and to confirm our claim that price changes are mostly driven by activity at the top of the order book (thus Level 1 data is sufficient to study the impact of limit orders on prices).

For example, consider the bid side of the order book with 10 shares at the top two levels. Absent any activity on the ask side and the second bid level, an OFI of -11 shares will lead to a bid price change of -1 tick. However, if 9 orders at the second bid level cancel before that order flow happens, the same OFI of -11 shares will lead to a price change of -2 ticks. In other words, if order activity up to second (third, fourth etc) level is important, tracking OFI only at the best prices will give a flawed picture of price dynamics.

To test this assertion, we compute variables $OFI^m, m = 2, \ldots, 5$ from m-th level queue fluctuations similarly to (2.12) and relabel $OFI^1 = OFI$. Then we fit five regressions, similar to (2.15), where variables OFI^m, $m = 2, \ldots, 5$ are added one at a time:

$$\Delta P_{k,i} = \hat{\alpha}_i^M + \sum_{m=1}^{M} \hat{\beta}_i^{m,M} OFI_{k,i}^m + \hat{\epsilon}_{k,i}^M, \quad M = 1, \ldots, 5 \tag{2.30}$$

The average results across time for a representative stock are shown on Figure 2.11. The average increase in explanatory power (measured by R^2) from adding OFI^2 as a regressor is 6.22%, which is quite small compared to the stand-alone R^2 of 70.83% for OFI^1. The effect of $OFI^3 - OFI^5$ is very small, and their coefficients appear to be only marginally significant, in contrast with those of OFI^1 and OFI^2. The cross-time average of coefficients $\hat{\beta}_i^{1,1}$ in the simple regression[20] with OFI^1 is 0.0597. In the multiple regression with OFI^1 and OFI^2 the averages of their respective coefficients are 0.0673 and 0.0406. We conclude that second-level activity, as summarized by OFI^2, has only a second-order influence on price changes, which are mainly driven by OFI^1. The effect of $OFI^3 - OFI^5$ is almost nill.

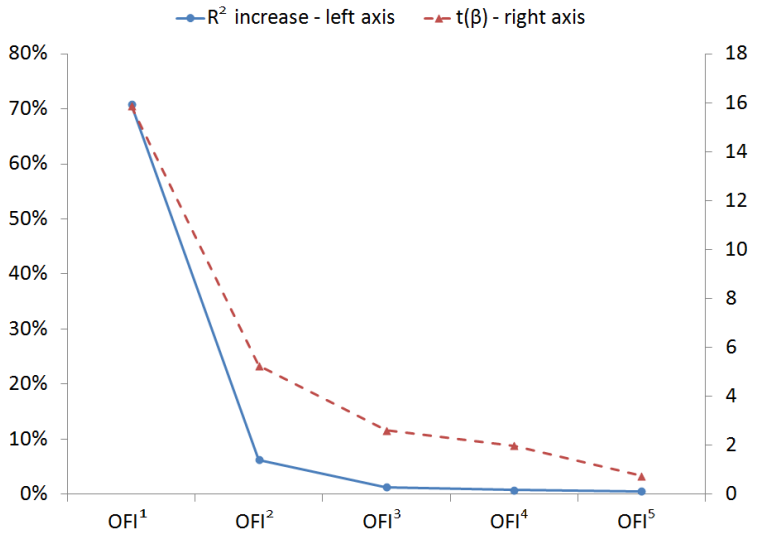

Figure 2.11: Cross-time average increase in R^2 from inclusion of variables $OFI^2 - OFI^5$, and cross-time average Newey-West t-statistics of their coefficient in the regression with all five variables, with NASDAQ ITCH data for the Schlumberger stock (SLB).

2.5.4 Choice of timescale

Using the auxiliary Level 2 dataset, we verify that our results are robust to potential issues in TAQ data, namely odd-lot sized orders at the best bid and offer, and mis-sequencing in quote data across exchanges during NBBO construction. We also compare our results across

[20]This coefficient is higher than the one obtained with NBBO data, because NASDAQ best quote depth is smaller than NBBO depth

a wide range of timescales. The auxiliary data comes from a single exchange (NASDAQ), has information on orders of all sizes and has timestamps up to a millisecond.

We estimate the regression (2.15) for a variety of timescales Δt, ranging from 50 milliseconds to 5 minutes using separate intraday subsamples as before. The size of these samples was different in order to stabilize the number of observations per sample. More precisely, data for the smallest timescales (50, 100 and 500 milliseconds) was separated into 1-minute instead of 30-minute subsamples to make numerical computations feasible. Data for the largest timescales (30 seconds to 5 minutes) was pooled across days preserving separate 30-minute intraday intervals to have a large number of observations per sample. The average R^2 and Newey-West t-statistics for OFI across time for each Δt are presented on Figure 2.12.

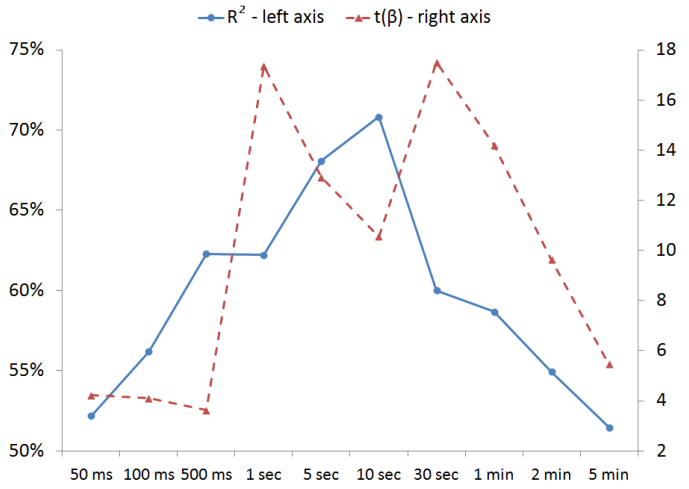

Figure 2.12: Average R^2 and Newey-West t-statistics for OFI coefficient across time for different Δt, with NASDAQ ITCH data for the Schlumberger stock (SLB).

The goodness of fit is stable across Δt, despite pronounced discreteness of data for very short time intervals. The OFI variable is statistically significant at a 95% level[21] in more than 80% of samples for Δt below one second, 100% of samples for Δt between one second and 2 minutes, and 92% of samples for Δt equal 5 minutes.

Notably there are many large price changes even when we consider Δt equal to 50

[21] using Newey-West t-statistics

milliseconds, but they usually correspond to high values of OFI. This is consistent with findings in [57], where authors describe the sporadic character of order activity in modern markets. When a subset of traders reacts to market updates in a matter of several milliseconds, this creates short intervals of increased activity with possibly large price changes and large OFI, and many time intervals with no activity when both variables are equal to zero. From our findings it appears that the simple model (2.7) can capture both of these regimes.

When a quadratic term $\hat{\gamma}_i^Q OFI_{k,i}|OFI_{k,i}|$ is added to the regression, the coefficient $\hat{\gamma}_i^Q$ is significant in a handful of samples (10 out of 871) for Δt bigger or equal to one second. For Δt under one second, the quadratic term is significant in about 16% of samples, and its contribution is marginal (about 3% increase in average R^2). We conclude that the relation between price changes and OFI is linear, irrespective of a timescale.

2.6 Conclusion

We have introduced *order flow imbalance*, a variable that cumulates the sizes of order book events, treating the contributions of market, limit and cancel orders equally, and provided empirical and theoretical evidence for a linear relation between high-frequency price changes and order flow imbalance for individual stocks. We have shown that this linear model is robust across stocks and timescales, and the price impact coefficient is inversely proportional to market depth. These relations suggest that prices respond to changes in the supply and demand for shares at the best quotes, and that the impact coefficient fluctuates with the amount of liquidity provision, or depth, in the market. Moreover, we have demonstrated that order flow imbalance is a more general metric of supply/demand dynamics than trade imbalance, and it can be used to analyze intraday changes in volatility, and monitor possible adverse selection in limit order executions. Trades seem to carry little to no information about price changes after the simultaneous order flow imbalance is taken into account. If trades do not help to explain price changes after controlling for the order flow imbalance, it is highly possible that the relation between the magnitude of price changes, or price volatility and traded volume simply captures the noisy scaling relation between these variables.

Chapter 3

Optimal order placement in limit order markets

This chapter is based on the paper "Optimal Order Placement in Limit Order Markets" [27] which is a joint work with Professor Rama Cont.

3.1 Introduction

> *When [a group of brokers] thinks it is advisable to sell shares, the means for prudently carrying out this purpose are given much thought. The members initiate action only when they can foresee its result, so that, apart from unlucky incidents, they can reckon on rather sure success.*
>
> Joseph de la Vega, Confusion de confusiones, 1688

In today's automated, electronic financial markets, the trading process is divided into several stages, each taking place on a different time horizon: portfolio allocation decisions are usually made on a monthly or daily basis and translate into trades that are executed over time intervals of several minutes to several days. Existing studies on optimal trade execution [15; 6] have investigated how the execution cost of a large trade may be reduced by splitting it into multiple *orders* spread in time. Once this *order scheduling* decision is taken, one still needs to specify how each individual order should be *placed*: this order

placement decision involves the choice of an *order type* (limit order or market order), order size and destination, when multiple trading venues are available. We focus here on this *order placement* problem: given an order which has been scheduled, deciding what type of order –market or limit order– and which trading venue to submit it to.

Orders are filled over short time intervals of a few milliseconds to several minutes and the mechanism through which orders are filled in the limit order book are relevant for such order placement decisions. When trading large portfolios, market participants need to make such decisions repeatedly, thousands of times a day, and their outcomes have a large impact on each participant's transaction cost as well as on aggregate market dynamics.

Early work on optimal trade execution [15; 6] did not explicitly model the process whereby each order is filled, but more recent formulations have tried to incorporate some elements in this direction. In one stream of literature (see [97], [3], [104]) a trader is restricted to using market orders whose execution costs are given by an idealized order book shape function. Another approach is to model the process through which an order is filled as a dynamic random process ([25; 26]) and thus formulate the optimal execution problem as a stochastic control problem: this formulation has been studied in various setting with limit orders ([14], [52]) or limit and market orders [53; 65] but its complexity makes it intractable unless restrictive assumptions are made on price and order book dynamics.

In the present work, we adopt a simpler, more tractable approach: assuming that the trade execution schedule has been specified, we focus on the task of filling each order. Decoupling the scheduling problem from the order placement problem leads to a more tractable approach which is closer to market practice and allows us to incorporate some realistic features which matter for order placement decisions, while conserving analytical tractability.

Individual order placement and order routing decisions play an important role in modern financial markets. Brokers are commonly obliged by law to deliver the best execution quality to their clients and empirical evidence confirms that a large percentage of market orders in the U.S. and Europe is sent to trading venues providing lower execution costs or smaller delays [17; 46]. Market orders gravitate towards exchanges with larger posted quote sizes and low fees, while limit orders are submitted to exchanges with high rebates and lower

execution waiting times (see [94]). These studies demonstrate how investors' aggregate order routing decisions have a significant influence on market dynamics, but a systematic study of the order routing problem from the investor's perspective is lacking. A reduced-form model for routing an infinitesimal limit order to a single destination is used by [94], while [49] and [82] propose numerical algorithms to optimize order executions across multiple dark pools, where supply/demand is unobserved. To the best of our knowledge this work is the first to provide a detailed treatment of investor's order placement decision in a multi-exchange market, unified with the market/limit order choice.

Our key contribution is a quantitative formulation of the order placement problem which takes into account multiple important factors - the size of an order to be executed, lengths of order queues across exchanges, statistical properties of order flows in these exchanges, trader's execution preferences, and the structure of liquidity rebates across trading venues. Our problem formulation is tractable, intuitive and blends the aforementioned factors into an optimal allocation of limit orders and market orders across available trading venues. Order routing heuristics employed in practice commonly depend on past order fill rates at each exchange and are inherently backward-looking. In contrast, our approach is forward-looking - the optimal order allocation depends on current queue sizes and distributions of future trading volumes across exchanges. When only a single exchange is available for execution, this order placement problem reduces to the problem of choosing an optimal split between market orders and limit orders. We derive an explicit solution for this problem and analyze its sensitivity to the order size, the trader's urgency for filling the order and other factors. Similar results are also established in a case of two trading venues under some assumptions on order flow distributions. Finally, we propose a numerical method for solving the order placement problem in a general case and demonstrate its efficiency through examples. Our numerical examples show that the use of our optimal order placement method allows to substantially decrease trading costs in comparison with some simple order placement strategies.

An important aspect of our framework is to account for the *execution risk*, i.e. the risk of not filling an order, through the incorporation of a penalty for such outcomes. This is different from most other studies which focus on the risk of price variations over the course

of a trade execution [6; 64] but assume that orders are always filled, or studies that ignore execution risks altogether [15; 14]. In our framework the penalty for execution risk plays an important role. When it is costly to catch up on the unfilled portion of the order, the optimal allocation shifts from limit to market orders. Although market orders are executed at a less favorable price, their execution is more certain and it becomes optimal to use them when the execution risk is a primary concern. Optimal limit order sizes are strongly influenced by total quantities of orders queueing for execution at each exchange and by distributions of order outflows from these queues. For example, if the queue size at one of the exchanges is much smaller than the expected future order outflow there, it is optimal to place a larger limit order on that exchange. According to a related study [94] such favorable limit order placement opportunities vanish in equilibrium due to competition and strategic order routing of individual traders. However the empirical results in that study also show that short-term deviations from the equilibrium are a norm, and can therefore be exploited in our optimization framework to improve limit order placement decisions. Finally, we find that the targeted execution size plays an important role - limit orders are used predominantly to execute small sizes and market orders are used to execute larger quantities, as long as their cost is less than the penalty for falling behind the target. This is a due to the fact that the amount of limit orders that can be realistically filled at each exchange is naturally constrained by the corresponding queue size and the order outflow distribution, so to execute larger quantities the trader needs to rely on market orders. We find that the optimal order allocation almost always sends limit orders to all available exchanges in an attempt to diversify execution risk, which suggests a benefit in having multiple exchanges. However, when order flows on different exchanges are highly positively correlated, these diversification advantages fade and the cost of execution becomes higher than on a single consolidated exchange.

Section 3.2 describes our formulation of the order placement problem and presents simple and intuitive conditions for the existence of an optimal order placement. In Section 3.3 we derive an optimal split between market and limit orders for a single exchange. Section 3.4 analyzes the general case of order placement on multiple trading venues. Section 3.5 presents a numerical algorithm for solving the order placement problem in a general case

and our simulation results, and Section 3.6 concludes.

3.2 The order placement problem

Consider a trader who has a mandate to buy S shares of a stock within a (short) time interval $[0, T]$. The deadline T may be a fixed time horizon (e.g. 1 minute) or a stopping time (e.g. triggered by price changes or trading volume dynamics). To gain queue priority, the trader can immediately submit K limit orders of sizes L_k to multiple exchanges $k = 1, \ldots, K$ and also market orders for M shares. The trader's *order placement* decision is thus summarized by a vector $X \triangleq (M, L_1, \ldots, L_K) \in \mathbb{R}_+^{K+1}$ of order sizes whose components are non-negative (only buy orders are allowed). Our objective is to define a meaningful framework in which the trader may choose between various possibilities for this order placement decision, for example between sending an order to a single exchange or splitting it in some proportion across K exchanges.

Our focus here is on limit order placement and we assume for simplicity that a market order of any size $M \leq S$ can be filled immediately and with certainty at any exchange. Under this assumption sending market orders to exchanges with high fees is clearly sub-optimal and we therefore consider a single exchange with the smallest liquidity fee f for the purpose of sending a single market order[1] of size M. Limit orders with quantities $(L_1, \ldots, L_K)$ join queues of $(Q_1, \ldots, Q_K)$ pre-existing orders at the best bids of K limit order books, where $Q_k \geq 0$, i.e. empty queues are allowed which corresponds to sending an order inside the bid-ask spread. We assume that all K limit orders submitted by the trader have the same price (the national best bid price), but this can be generalized to having multiple prices at the expense of additional notation[2].

[1]This simplifying assumption is reasonable as long as the target size S is small relative to the prevailing market depth. Otherwise the quantity S can be filled with multiple market orders at exchanges with progressively larger fees $f_1 < f_2 < \cdots < f_K$, and the total cost of these market orders becomes a convex piecewise linear function. Our results extend to this case, but to avoid additional notation we assume that S can be executed with a single market order at a fee f so the market order cost is linear in its size.

[2]In our framework the only difference between submitting a limit order to the best bid price and deeper in the order book is that orders deeper in the book have lower costs and larger queue sizes - i.e. in addition to orders queued at the same price they need to wait for all higher-priced buy orders to clear from the market.

Denote by $(x)_+ \triangleq \max(x, 0)$. If the trader does not modify his limit orders before time T, the amount purchased with each limit order by time T can be explicitly computed as a function of future order flow:

$$\min(\max(\xi_k - Q_k, 0), L_k) = (\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+, \quad k = 1, \ldots, K,$$

where ξ_k is a total outflow from the front of the k-th order queue. The order outflow ξ_k consists of order cancelations that occurred before time T from queue positions in front of an order L_k, and of marketable orders that reach the k-th exchange before T. The mechanics of limit order fills and order outflows are further illustrated on Figure 3.1.

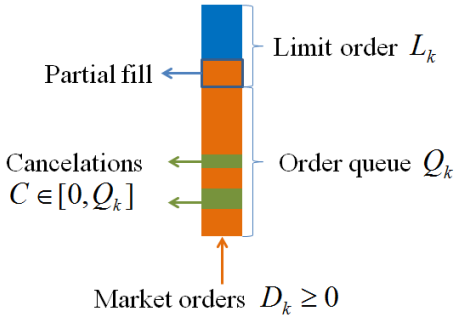

Figure 3.1: Limit order execution on exchange k depends on the order size L_k, the queue Q_k in front of it, total sizes of order cancelations C_k and marketable orders D_k, specifically on $\xi_k = C_k + D_k$.

We note that limit order fill amounts are random because they depend on random future bid queue outflows $\xi = (\xi_1, \ldots, \xi_K)$. The random variable ξ is defined with respect to an execution time horizon T, therefore its distribution depends on a trader's choice of T. Here we do not make any assumptions regarding the distribution of ξ except for illustration purposes. In fact our formulation leads to an intuitive iterative procedure that approximates the optimal order allocation by using historical data and also does not require specifying a distribution for ξ.

The *total amount* of shares $A(X,\xi)$ bought by the trader by time T with his limit and market orders is a function of the order allocation X and an overall bid queue outflow ξ:

$$A(X,\xi) = M + \sum_{k=1}^{K} ((\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+) \tag{3.1}$$

The total price of this purchase is divided into a benchmark cost paid regardless of trader's decisions, computed using a mid-quote price level, and an *execution cost* relative to mid-quote price given by

$$(h+f)M - \sum_{k=1}^{K}(h+r_k)((\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+), \tag{3.2}$$

where h is a half of the bid-ask spread at time 0, f is the lowest available fee for taking liquidity and $r_k, k = 1,\dots,K$ are rebates for adding liquidity on different exchanges. The benchmark price in our formulation is the mid-quote price at time 0, so in (3.2) the trader saves half of the bid-ask spread plus liquidity rebates on his limit orders, and pays half of the spread plus a liquidity fee on his market orders. Limit orders reduce the cost but lead to a risk of falling behind the target quantity S because their fills are random. To capture this *execution risk* we include, in the objective function, a penalty for violations of target quantity in both directions:

$$\lambda_u \left(S - A(X,\xi)\right)_+ + \lambda_o \left(A(X,\xi) - S\right)_+, \tag{3.3}$$

where $\lambda_u \geq 0, \lambda_o \geq 0$ are marginal penalties in dollars per share for, respectively falling behind or exceeding the execution target S. These penalties are motivated by a correlation that exists between limit order executions and price movements (so-called adverse selection, see e.g. [107]). If $A(X,\xi) < S$, the trader has to purchase the remaining $S - A(X,\xi)$ shares at time T with market orders. Adverse selection implies that conditionally on the event $\{A(X,\xi) < S\}$ prices have likely moved up and the transaction cost of market orders at time T is higher than their cost at time 0, i.e. $\lambda_u > h + f$. Alternatively, if $A(X,\xi) > S$ the trader experiences buyer's remorse - conditionally on this event prices have likely moved down and he could have achieved a better execution by sparing some of his market orders.

Besides adverse selection, parameters λ_u, λ_o may reflect trader's execution preferences. For example a trader with a positive forecast of short-term returns may prefer to trade early with a market order and set a larger value for λ_u compared to λ_o. A trader with significant aversion to execution risk would choose high values for both λ_u, λ_o to reflect this aversion.

Problem 1 (Optimal order placement problem) *An optimal order placement is a vector* $X^* \in \mathbb{R}_+^{K+1}$ *solution of*

$$\min_{X \in \mathbb{R}_+^{K+1}} \mathbb{E}[v(X, \xi)] \tag{3.4}$$

where

$$v(X,\xi) := (h+f)M - \sum_{k=1}^{K}(h+r_k)((\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+) + \lambda_u \left(S - A(X,\xi)\right)_+ + \lambda_o \left(A(X,\xi) - S\right)_+ \tag{3.5}$$

is the sum of the execution cost and penalty for execution risk.

We will denote $V(X) = E[v(X, \xi)]$. We begin by assuming certain economically reasonable restrictions on parameter values.

Assumptions

A1 $\min_k\{r_k\} + h > 0$: even if some rebates r_k are negative[3], limit orders reduce the execution cost.

A2 $\lambda_o > h + \max_k\{r_k\}$ and $\lambda_o > -(h+f)$: there is no incentive to exceed the target quantity S regardless of fees and rebates, even if they are negative.

Proposition 2 below shows that it is not optimal to submit limit or market orders that are *a priori* too large or too small (larger than the target size S or whose sum is less than S). Proposition 3 guarantees the existence of an optimal solution.

[3]Some *inverse* exchanges pay for executing marketable orders and charge for executing passive limit orders. For example, on 03/07/2013 a U.S. equity exchange Direct Edge EDGA had a negative rebate $r = -\$0.0006$ per share for passive orders and a negative fee $f = -\$0.0004$ per share for marketabe orders. Another inverse exchange BATS BYX had on that date a rebate $r = -\$0.0002$ and a fee $f = -\$0.0002$. These negative values are typically smaller than the minimal value of $h = \$0.005$ for U.S. equities, justifying our assumptions A1-A2.

Proposition 2 *Consider the compact convex subset of* $\mathbb{R}_+^{K+1}$ *defined by*

$$\mathcal{C} \triangleq \left\{ X \in \mathbb{R}_+^{K+1} \;\middle|\; 0 \le M \le S, \;\; 0 \le L_k \le S - M, k = 1, \ldots, K, \;\; M + \sum_{k=1}^{K} L_k \ge S \right\}.$$

Under assumptions A1-A2 for any $\tilde{X} \notin \mathcal{C}$, $\exists \tilde{X}' \in \mathcal{C}$ *with* $V(\tilde{X}') \le V(\tilde{X})$. *Moreover, if* $\min_k \{\mathbb{P}(\xi_k > Q_k + S)\} > 0$, *the inequality is strict:* $V(\tilde{X}') < V(\tilde{X})$.

Proof: First, for any allocation $\tilde{X}$ that has $\tilde{M} > S$, we automatically have $A(\tilde{X}) > S$ and we can show that the (random) cost and penalty of $\tilde{X}$ are larger than those of $X^{naive} \triangleq (S, 0, \ldots, 0) \in \mathcal{C}$:

$$v(\tilde{X}, \xi) - v(X^{naive}, \xi) = (h + f)(\tilde{M} - S) - \sum_{k=1}^{K} (h + r_k)((\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+) +$$
$$\lambda_o \left(\tilde{M} - S + \sum_{k=1}^{K} ((\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+) \right) =$$
$$(\lambda_o + h + f)(\tilde{M} - S) + \sum_{k=1}^{K} (\lambda_o - h - r_k)((\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+) > 0,$$

which holds for all random ξ. Therefore, $V(\tilde{X}) > V(X^{naive})$. Similarly, for any allocation $\tilde{X}$ with $\tilde{L}_k > S - \tilde{M}$ define a different allocation $\tilde{X}'$ by $\tilde{M}' = \tilde{M}$, $\tilde{L}'_j = \tilde{L}_j, \forall j \ne k$ and $\tilde{L}'_k = S - \tilde{M}$. Then $v(\tilde{X}, \xi) - v(\tilde{X}', \xi) = 0$ on the event $B = \{\omega | \xi_k(\omega) < Q_k + S - M\}$. On its complementary event B^c,

$$v(\tilde{X}, \xi) - v(\tilde{X}', \xi) = -(h + r_k)((\xi_k - Q_k - S + \tilde{M})_+ - (\xi_k - Q_k - \tilde{L}_k)_+)$$

$$+\lambda_o((\xi_k - Q_k - S + \tilde{M})_+ - (\xi_k - Q_k - \tilde{L}_k)_+).$$

Therefore

$$V(\tilde{X}) - V(\tilde{X}') = \mathbb{E}\left[v(\tilde{X},\xi) - v(\tilde{X}',\xi)|B\right]\mathbb{P}(B) + \mathbb{E}\left[v(\tilde{X},\xi) - v(\tilde{X}',\xi)|B^c\right]\mathbb{P}(B^c) =$$
$$0 + \mathbb{E}\left[(\lambda_o - (h+r_k))((\xi_k - Q_k - S + \tilde{M})_+ - (\xi_k - Q_k - \tilde{L}_k)_+)|B^c\right]\mathbb{P}(B^c) \geq 0$$

with a strict inequality if $\mathbb{P}(B^c) > 0$. If $\tilde{X}' \notin \mathcal{C}$, we can continue truncating limit order sizes $\tilde{L}'_j > S - \tilde{M}'$ following the same argument. Each time the truncation does not increase the objective function and finally we obtain $\tilde{X}'' \in \mathcal{C}$, such that $V(\tilde{X}'') \leq V(\tilde{X})$.

Next, if $\tilde{X}$ is such that $\tilde{M} - \sum_{k=1}^{K}\tilde{L}_k < S$ define $s = S - \tilde{M} - \sum_{k=1}^{K}\tilde{L}_k$, take $\tilde{M}' = \tilde{M}, \tilde{L}'_k = \tilde{L}_k, k = 1,\ldots,K-1$ and $\tilde{L}'_K = \tilde{L}_k + s$. Then, on the event $B = \left\{\omega|\xi_K(\omega) < Q_K + \tilde{L}_K\right\}$ we have $v(\tilde{X},\xi) = v(\tilde{X}',\xi)$. However, on the event B^c,

$$v(\tilde{X},\xi) - v(\tilde{X}',\xi) = (h+r_K)((\xi_K - Q_K - \tilde{L}_K)_+ - (\xi_k - Q_k - \tilde{L}_K - s)_+)$$
$$+ \lambda_u((\xi_K - Q_K - \tilde{L}_K)_+ - (\xi_k - Q_k - \tilde{L}_K - s)_+),$$

therefore

$$V(\tilde{X}) - V(\tilde{X}') = \mathbb{E}\left[v(\tilde{X},\xi) - v(\tilde{X}',\xi)|B\right]\mathbb{P}(B) + \mathbb{E}\left[v(\tilde{X},\xi) - v(\tilde{X}',\xi)|B^c\right]\mathbb{P}(B^c) =$$
$$0 + \mathbb{E}\left[(\lambda_u + (h+r_k))((\xi_K - Q_K - \tilde{L}_K)_+ - (\xi_k - Q_k - \tilde{L}_K - s)_+)|B^c\right]\mathbb{P}(B^c) \geq 0$$

with a strict inequality if $\mathbb{P}(B^c) > 0$. ■

Proposition 2 shows that it is never optimal to overflow the target size S with a single order, but it may be optimal to exceed the target S with the sum of order sizes $M + \sum_{k=1}^{K} L_k$. The penalty function (3.3) effectively implements a soft constraint for order sizes and focuses the search for an optimal order allocation to the set $\mathcal{C}$. Specific economic or operational considerations could also motivate adding hard constraints to problem (3.4), e.g. $M = 0$ or $\sum_{k=1}^{K} L_k = S$. Such constraints can be easily included in our framework but absent the aforementioned considerations we do not impose them here.

Proposition 3 *Under assumptions A1-A2, $V : \mathbb{R}_+^{K+1} \mapsto \mathbb{R}$ is convex, bounded from below and has a global minimizer $X^\star \in \mathcal{C}$.*

Proof: First, note that $(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+$ are concave functions of L_k. Therefore, $A(X, \xi)$ is concave as a sum of concave functions. Similarly, the cost term in $v(X, \xi)$ is a sum of convex functions, as long as $r_k \geq -h, k = 1, \ldots, K$ and is itself a convex function. Second, since $S - A(X, \xi)$ is a convex function of X, and the function $w(x) \stackrel{\Delta}{=} \lambda_u(x)_+ - \lambda_o(-x)_+$ is convex in x for positive λ_u, λ_o, so the penalty term $w\left(S - A(X, \xi)\right))$ is also convex.

If $\lambda_o > h + \max\limits_k\{r_k\}$ the function $V(X)$ is also bounded from below since $v(X, \xi) \geq -(h + \max\limits_k\{r_k\})S$.

Finally, since $V(X)$ is convex, it is also continous and reaches a local minimum V_{min} on the compact set $\mathcal{C}$ at some point $X^\star \in \mathcal{C}$. By convexity, V_{min} is a global minimum of $V(X)$ on $\mathcal{C}$. Moreover, since $\lambda_o > h + \max\limits_k\{r_k\}$, Proposition 2 guarantees that $V_{min} < V(\tilde{X})$ for any $\tilde{X} \notin \mathcal{C}$, so V_{min} is also a global minimum of $V(X)$ on $\mathbb{R}_+^{K+1}$. ■

We may also consider an alternative approach to order placement optimization, which turns out to be related to our original formulation by duality. Consider the following problem:

Problem 2 (Alternative formulation: cost minimization under execution constraints)

$$\min_{X \in \mathbb{R}_+^{K+1}} \mathbb{E}[(h+f)M - \sum_{k=1}^{K}(h + r_k)((\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+)], \tag{3.6}$$

$$\textit{subject to:} \quad \mathbb{E}\left[(S - A(X, \xi))_+\right] \leq \mu_u, \tag{3.7}$$

$$\mathbb{E}\left[(A(X, \xi) - S)_+\right] \leq \mu_o \tag{3.8}$$

In this alternative formulation a trader can specify his tolerance to execution risks using constraints on expected order shortfalls and overflows. The goal is to minimize an expectation of order execution costs under the expected shortfall constraints. Problem 2 does not appear to be tractable, but it has a convex objective and convex inequality constraints, and we can easily find its (Lagrangian) dual problem:

Problem 3

$$\max_{\lambda_u \geq 0, \lambda_o \geq 0} \left\{ V^\star(\lambda_u, \lambda_o) - \lambda_u \mu_u - \lambda_o \mu_o \right\}, \tag{3.9}$$

where $V^\star(\lambda_u, \lambda_o)$ is the optimal objective value from Problem 1 given λ_u, λ_o.

We see that Problem 3 is related to our original order placement problem - solving Problem 3 (and therefore, Problem 2) amounts to re-solving Problem 1 for different values of λ_u, λ_o. This discussion also leads to a new interpretation of parameters λ_u, λ_o in Problem 1 as shadow prices for expected shortfall and overflow constraints in the related Problem 2. Hereafter we focus on the (more tractable) Problem 1, but note that the optimal point for Problem 2 can also be found by solving its dual.

3.3 Choice of order type: limit orders vs market orders

To highlight the tradeoff between limit and market order executions in our optimization setup, we first consider a case when the asset is traded on a single exchange, and the trader has to choose an optimal split between limit and market orders. Since $K = 1$, we suppress the subscript 1 throughout this section.

Proposition 4 (Single exchange: optimal split between limit and market orders) *Assume that ξ has a continuous distribution and (A1-A2) hold. The optimal order allocation depends on λ_u:*

If $\lambda_u \le \underline{\lambda_u} \triangleq \dfrac{2h+f+r}{F(Q+S)} - (h+r)$, the optimal strategy is to submit only limit orders: $(M^\star, L^\star) = (0, S)$.

If $\lambda_u \ge \overline{\lambda_u} \triangleq \dfrac{2h+f+r}{F(Q)} - (h+r)$, the optimal strategy is to submit only market orders: $(M^\star, L^\star) = (S, 0)$.

If $\lambda_u \in (\underline{\lambda_u}, \overline{\lambda_u})$, the optimal split between limit and market orders is

$$\begin{cases} M^\star = S - F^{-1}\left(\dfrac{2h+f+r}{\lambda_u+h+r}\right) + Q, \\ L^\star = F^{-1}\left(\dfrac{2h+f+r}{\lambda_u+h+r}\right) - Q, \end{cases} \tag{3.10}$$

where $F(\cdot)$ is a cumulative distribution function of the bid queue outflow ξ.

Proof: By Proposition 2 there exists an optimal split $(M^\star, L^\star) \in \mathcal{C}$ between limit and market orders. Moreover for $K = 1$ the set $\mathcal{C}$ reduces to a line $M^\star + L^\star = S$ so it is sufficient to find $M^\star$. Restricting $L = S - M$ implies that $\{A(X,\xi) > S\} = \varnothing$, $\{A(X,\xi) < S, \xi > Q + L\} = \varnothing$, and we can rewrite the objective function as

$$V(M) = \mathbb{E}\Big[(h+f)M - (h+r)((\xi-Q)_+ - (\xi-Q-S+M)_+) \; + \\ \lambda_u \left(S - M - ((\xi-Q)_+ - (\xi-Q-S+M)_+)\right)\big)_+\Big]. \tag{3.11}$$

For $M \in (0, S)$ the expression under the expectation in (3.11) is bounded for all ξ and differentiable with respect to M for almost all ξ, so we can compute $V'(M) = \frac{dV(M)}{dM}$ by

interchanging the order of differentiation and integration (see e.g. [5], Theorem 24.5):

$$V'(M) = \mathbb{E}\Big[h+f+(h+r)\mathbb{1}_{\{\xi>Q+S-M\}}-\lambda_u\mathbb{1}_{\{\xi<Q+S-M\}}\Big] = 2h+f+r-(h+r+\lambda_u)F(Q+S-M) \tag{3.12}$$

Note that if $\lambda_u \leq \dfrac{2h+f+r}{F(Q+S)} - (h+r)$, then $V'(M) \geq 0$ for $M \in (0,S)$ and therefore V is non-decreasing at these points. Checking that $V(S) - V(0) \geq (h+f-\lambda_u)S + (\lambda_u + h + r)S(1-F(Q+S)) \geq 0$ we conclude that $M^\star = 0$. Similarly, if $\lambda_u \geq \dfrac{2h+f+r}{F(Q)} - (h+r)$, then $v(M) \leq 0$ for all $M \in (0,S)$ and $V(M)$ is non-increasing at these points. Checking that $V(S)-V(0) \leq (h+f-\lambda_u)S+(\lambda_u+h+r)S(1-F(Q)) \leq 0$ we conclude that $M^\star = S$. Finally, if λ_u is between these two values, $\exists \epsilon > 0$, such that $V'(\epsilon) < 0, V'(S-\epsilon) > 0$ and by continuity of V' there is a point where $V'(M^\star) = 0$. This $M^\star$ is optimal by convexity of $V(M)$ and (3.10) solves equations $v(M^\star, \xi) = 0, L^\star = S - M^\star$. ■

In the case of a single exchange, Proposition 2 implies that $M^\star + L^\star = S$, therefore there is no risk of exceeding the target size and λ_o does not affect the optimal solution. The trader is only concerned with the risk of falling behind the target quantity, and balances this risk with the fee, rebate and other market information. The parameter λ_u can be interpreted as trader's urgency to fill the orders, and higher values of λ_u lead to smaller limit order sizes, as illustrated on Figure 3.2. In contrast, the optimal market order size increases with λ_u.

The optimal split between market and limit orders depends on the ratio $\frac{2h+f+r}{\lambda_u+(h+r)}$ which balances marginal costs and savings from a market order. It also depends on the order outflow distribution $F(\cdot)$ and the queue length Q - keeping all else constant, a trader would submit a larger limit order if its execution is more likely and vice versa. The optimal limit order size decreases with λ_u as it becomes more expensive to underfulfill the order and increases with f as market orders become more expensive. Another interesting feature is that $L^\star$ is fully determined by Q, F and pricing parameters h, r, f, λ_u, while $M^\star$ increases with S. The consequence of this solution feature is that as the target size S increases, a larger fraction $\frac{M^\star}{S}$ of it is executed with a market order. The total quantity that can realistically be filled with a limit order is limited by Q and ξ, so to accommodate larger target sizes the trader resorts to market orders. This *bounded capacity* feature of limit orders also appears in our solutions for multiple exchanges. For example, as the number of

available exchanges K increases, the overall prospects of filling limit orders at any of them improve and the fraction $\frac{M^\star}{S}$ decreases. We can see that the solution $(M^\star, L^\star)$ depends on the entire distribution $F(\cdot)$ and not just on the mean of ξ. Limit orders are filled when $\xi \geq Q + L$, so the tail of $F(\cdot)$ affects order executions and is an important determinant of the optimal order allocation. Figure 3.2 shows two order allocations for exponential and Pareto distributions of ξ with equal means.

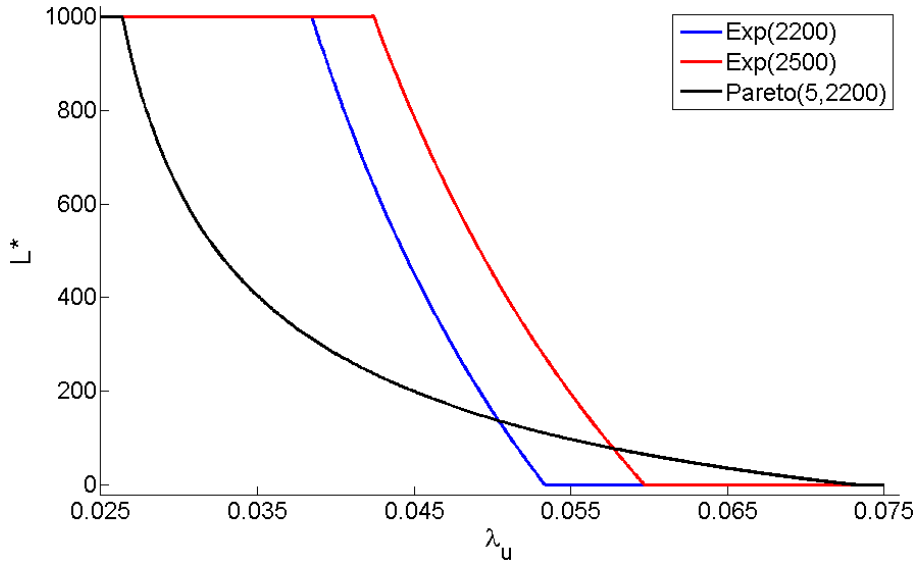

Figure 3.2: Optimal limit order size $L^\star$ for one exchange. The parameters for this figure are: $Q = 2000, S = 1000, h = 0.02, r = 0.002, f = 0.003$. Colors correspond to different order outflow distributions - exponential with means 2200 and 2500 and Pareto with mean 2200 and a tail index 5.

3.4 Optimal routing of limit orders across multiple exchanges

When multiple trading venues are available, dividing the target quantity among them reduces the risk of not filling the order and may improve the execution quality. However, sending too many orders leads to an undesirable possibility of exceeding the target size. Proposition 5 gives optimality conditions for an order allocation $X^\star = (M^\star, L_1^\star, \dots, L_K^\star)$ that balances shortfall risks and costs. The following probabilities play an important role in this balance:

$$p_0 \triangleq \mathbb{P}\left(\bigcap_{k=1}^{K}\{\xi_k \le Q_k\}\right), \quad p_j \triangleq \mathbb{P}\left(\bigcap_{k\neq j}\{\xi_k \le Q_k\}\Big|\xi_j > Q_j\right), \quad j = 1,\dots,K$$

Intuitively, p_0 is a probability that no limit orders will be filled given current queue sizes, and it measures the overall execution prospects for limit orders. Each p_j is a probability of no fills everywhere except the j-th exchange, conditional on a fill at the exchange j, so p_j measure tail dependences between order flows on different exchanges.

Proposition 5 *Assume (A1-A2), also assume that the distribution of ξ is continuous, $\max_k\{F_k(Q_k+S)\} < 1$ and $\lambda_u < \max_k\left\{\dfrac{2h+f+r_k}{F_k(Q_k)} - (h+r_k)\right\}$. Then:*

1. *If*
$$\lambda_u \ge \frac{2h+f+\max_k\{r_k\}}{p_0} - (h+\max_k\{r_k\}), \tag{3.13}$$
then any optimal order placement strategy involves market orders: $M^\star > 0$.

2. *If*
$$p_j > \frac{\lambda_o - (h+r_j)}{\lambda_u+\lambda_o}, \tag{3.14}$$
then any optimal order placement strategy involves submitting limit orders to the j-th exchange: $L_j^\star > 0$.

3. *If* (3.13)-(3.14) *hold for all exchanges $j = 1,\dots,K$: $X^\star \in \mathcal{C}$ is an optimal order*

placement if and only if the following conditions are fulfilled:

$$\mathbb{P}\left(M^\star + \sum_{k=1}^{K}((\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k^\star)_+) < S\right) = \frac{h+f+\lambda_o}{\lambda_u+\lambda_o} \tag{3.15}$$

$$\mathbb{P}\left(M^\star + \sum_{k=1}^{K}((\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k^\star)_+) < S \middle| \xi_j > Q_j + L_j^\star\right) = \frac{\lambda_o - (h+r_j)}{\lambda_u+\lambda_o}, \quad j = 1,\ldots,K \tag{3.16}$$

Proof: Proposition 3 implies the existence of an optimal order allocation $X^\star \in \mathcal{C}$. First, we define $X_M \triangleq (S, 0, \ldots, 0)$ and prove that $X^\star \neq X_M$ by contradiction. If X_M were optimal in problem (3.4) it would also be optimal in the same problem with a constraint $L_k = 0, k \neq j$, for any one j. In other words, the solution $(S, 0)$ would be optimal for any one-exchange problem, defined by using only exchange j. But by our assumption, there exists$J$ such that $\lambda_u < \dfrac{2h+f+r_J}{F_J(Q_J)} - (h + r_J)$ and Proposition 4 implies that $(S, 0)$ is not optimal for the J-th single-exchange subproblem, leading to a contradiction.

The function $v(X, \xi)$ is bounded for $X \in \mathcal{C}$ and for all ξ, differentiable with respect to M and $L_k, k = 1, \ldots, K$ for $X \in \mathcal{C}\backslash\{X_M\}$ for almost all ξ. Applying the same theorem as in the proof of Proposition 4 we conclude that $V(X)$ is differentiable for $X \in \mathcal{C}\backslash\{X_M\}$ and we can compute all of its partial derivatives by interchanging the order of differentiation and integration. The KKT conditions for problem (3.4) and $X \in \mathcal{C}\backslash\{X_M\}$ are

$$h + f - \lambda_u\mathbb{P}(A(X^\star, \xi) < S) + \lambda_o\mathbb{P}(A(X^\star, \xi) > S) - \mu_0 = 0 \tag{3.17}$$

$$-(h + r_k)\mathbb{P}(\xi_k > Q_k + L_k^\star) - \lambda_u\mathbb{P}(A(X^\star, \xi) < S, \xi_k > Q_k + L_k^\star) + \lambda_o\mathbb{P}(A(X^\star, \xi) > S, \xi_k > Q_k + L_k^\star) - \mu_k = 0, \quad k = 1,\ldots,K \tag{3.18}$$

$$M \geq 0, \quad L_k \geq 0, \quad \mu_0 \geq 0, \quad \mu_k \geq 0, \quad \mu_0 M = 0, \quad \mu_k L_k = 0, \quad k = 1,\ldots,K \tag{3.19}$$

Since the objective function $V(\cdot)$ is convex, conditions (3.17)–(3.19) are both necessary and sufficient for optimality. The first result of this proposition follows from considering any $\tilde{X}$ with $\tilde{M} = 0$:

$V(\tilde{X}) \geq \lambda_u S\mathbb{P}\left(\bigcap_k \{\xi_k \leq Q_k\}\right) - (h + \max_k\{r_k\})S\mathbb{P}\left(\overline{\bigcap_k \{\xi_k \leq Q_k\}}\right) \geq (h+f)S = V(X_M)$

and we already argued that $\exists X^\star$ with $V(X^\star) < V(X_M)$, so $X^\star \neq \tilde{X}$ and therefore $M^\star > 0$. Rearranging terms in a j-th equality (3.18) we obtain

$$\mathbb{P}(\xi_j > Q_j + L_j^\star)\left[\lambda_o - (h + r_j) - (\lambda_u + \lambda_o)\mathbb{P}(A(X^\star, \xi) < S | \xi_j > Q_j + L_j^\star)\right] - \mu_j = 0 \quad (3.20)$$

The term in square brackets in (3.20) is negative for any $X \in \mathcal{C} \setminus \{X_M\}$ with $L_j = 0$, because

$\mathbb{P}(A(X,\xi) < S | \xi_j > Q_j + L_j) > \mathbb{P}\left(\bigcap_{k \neq j}\{\xi_k \leq Q_k\} \middle| \xi_j > Q_j\right) > \frac{\lambda_o - (h + r_j)}{\lambda_u + \lambda_o}$

by assumption and since $\mu_j \geq 0$ the condition (3.18) cannot be satisfied with $L_j^\star = 0$. We showed that $M^\star > 0, L_j^\star > 0$ for all $j = 1, \ldots, K$ and therefore, $\mu_0 = \mu_1 = \cdots = \mu_K = 0$ by complimentary slackness. Then the KKT conditions (3.17)–(3.19) reduce to (3.15)–(3.16).

■

Equations (3.15)-(3.16) show that an optimal order allocation equates shortfall probabilities to specific values computed with pricing parameters. This gives yet another interpretation for parameters λ_u, λ_o - a trader can specify his tolerance for execution risk in terms of shortfall probabilities and use the above equations to calibrate these parameters.

When the number of exchanges K is large, the probabilities in (3.15)-(3.16) are difficult to compute in closed-form. However, the case $K = 2$ is relatively tractable and will be analyzed as an illustration. The assumption of independence between ξ_1, ξ_2 is made only in this example and is not required for the rest of our results. In Section 3.5 we study the effect of correlation between order flows on optimal order placement decisions.

Corollary *Consider the case of two exchanges with outflows ξ_1, ξ_2 that are independent and have continuous distributions. If*

1. $\max_{k=1,2}\{F_k(Q_k+S)\} < 1$,
2. $\lambda_u < \max_{k=1,2}\left\{\frac{2h+f+r_k}{F_k(Q_k)} - (h+r_k)\right\}$, $\lambda_u \geq \frac{2h+f+\max_{k=1,2}\{r_k\}}{F_1(Q_1)F_2(Q_2)} - (h+\max_{k=1,2}\{r_k\})$, *and*
3. $F_1(Q_1) < 1 - \frac{h+r_2}{\lambda_o}, F_2(Q_2) < 1 - \frac{h+r_1}{\lambda_o}$,

then there exists an optimal order allocation $X^\star = (M^\star, L_1^\star, L_2^\star) \in int\{\mathcal{C}\}$ and it verifies

$$L_1^\star = Q_2 + S - M^\star - F_2^{-1}\left(\frac{\lambda_o - (h+r_1)}{\lambda_u + \lambda_o}\right) \tag{3.21a}$$

$$L_2^\star = Q_1 + S - M^\star - F_1^{-1}\left(\frac{\lambda_o - (h+r_2)}{\lambda_u + \lambda_o}\right) \tag{3.21b}$$

$$\bar{F}_1(Q_1+L_1^\star)\bar{F}_2(Q_2+S-M^\star-L_1^\star) + \int_{Q_1+S-M^\star-L_2^\star}^{Q_1+L_1^\star} \bar{F}_2(Q_1+Q_2+S-M^\star-x_1)dF_1(x_1) = \frac{\lambda_u-(h+f)}{\lambda_u+\lambda_o}, \tag{3.21c}$$

where $F_1(\cdot), F_2(\cdot)$ are the cdf of ξ_1, ξ_2 respectively.

Proof: Solutions on the boundary of $\mathcal{C}$ are sub-optimal: $M^\star = 0$ and $M^\star = S$ are ruled out by assumption 2, $L_1^\star = S - M$ and $L_2^\star = S - M$ are ruled out by assumption 3 and (3.18). Solutions with $M^\star + \sum_{k=1}^{K} L_k^\star = S$ are ruled out by directly checking (3.18). Finally, $L_1^\star = 0$ and $L_2^\star = 0$ are also ruled out by (3.18). For example if $L_1^\star = 0$, then by Proposition 2 $M^\star + L_2^\star = S$ and in (3.18) $\mu_2 = 0$ by complimentary slackness, $\mathbb{P}(A(X^\star,\xi) < S, \xi_2 > Q_2 + L_2^\star) = \mathbb{P}(A(X^\star,\xi) > S, \xi_2 > Q_2 + L_2^\star) = 0$. But then (3.18) cannot hold because $\mathbb{P}(\xi_2 > Q_2 + L_2^\star) > 0$.

For any $X \in int\{\mathcal{C}\}$, $A(X,\xi) > S$ if and only if all the following three inequalities are satisfied:

$$\xi_1 > Q_1 + S - M - L_2 \tag{3.22a}$$

$$\xi_2 > Q_2 + S - M - L_1 \tag{3.22b}$$

$$\xi_1 + \xi_2 > Q_1 + Q_2 + S - M \tag{3.22c}$$

These inequalities give a simple characterization of the event $\{A(X,\xi) > S\}$ which is directly verified by considering subsets of (ξ_1, ξ_2) forming a complete partition of $\mathbb{R}^2_+$.

Case 1: $\xi_1 > Q_1 + L_1, \xi_2 > Q_2 + L_2$. Since $L_1 + L_2 + M > S$, we have $A(X,\xi) = L_1 + L_2 + M > S$ and at the same time all of the inequalities (3.22a-3.22c) are satisfied, so they are trivially equivalent in this case.

Case 2: $\xi_1 > Q_1 + L_1, Q_2 \leq \xi_2 \leq Q_2 + L_2$. Because of the condition $\xi_1 > Q_1 + L_1$, (3.22a) is satisfied. We have in this case that $A(X,\xi) = L_1 + \xi_2 - Q_2 + M$ and thus $A(X,\xi) > S$ if and only if (3.22b) is satisfied. Finally, $\xi_1 > Q_1 + L_1$ together with (3.22b) imply (3.22c), so $A(X,\xi) > S$ and (3.22a-3.22c) are equivalent in this case.

Case 3: $\xi_2 > Q_2 + L_2, Q_1 \leq \xi_1 \leq Q_1 + L_1$. Similarly to Case 2 we can show that inequalities (3.22a-3.22c) are satisfied if and only if $A(X,\xi) > S$.

Case 4: $Q_1 + S - M - L_2 < \xi_1 \leq Q_1 + L_1, Q_2 + S - M - L_1 < \xi_2 \leq Q_2 + L_2$. This set is non-empty because $0 < S - M - L_1 < L_2$ and similarly for L_1, L_2 reversed. Inequalities (3.22a)–(3.22b) hold trivially, only (3.22c) needs to be checked. We can write $A(X,\xi) = \xi_1 - Q_1 + \xi_2 - Q_2 + M > S$ if and only if (3.22c) holds, so $A(X,\xi) > S$ is equivalent to (3.22a-3.22c).

Case 5: Outside of Cases 1-4, either (3.22a) or (3.22b) is not satisfied. If $\xi_1 \leq Q_1 + S - M - L_2, \xi_2 \leq Q_2 + L_2$, then $A(X,\xi) \leq S - M - L_2 + L_2 + M = S$. The case $\xi_2 \leq Q_2 + S - M - L_1, \xi_1 \leq Q_1 + L_1$ is completely symmetric, and it shows that neither $A(X,\xi) > S$ nor (3.22a-3.22c) hold in this case.

Next, we use inequalities (3.22a-3.22c) to characterize the set $\{A(X,\xi) > S\}$ in the first-order conditions (3.15)–(3.16). We observe that in the two-exchange case

$$\{A(X,\xi) > S, \xi_1 > Q_1 + L_1\} = \{\xi_1 > Q_1 + L_1, \xi_2 > Q_2 + S - M - L_1\}$$
$$\{A(X,\xi) > S, \xi_2 > Q_2 + L_2\} = \{\xi_2 > Q_2 + L_2, \xi_1 > Q_1 + S - M - L_2\},$$

and then use the independence of ξ_1 and ξ_2 to compute

$$\mathbb{P}(A(X,\xi) > S | \xi_1 > Q_1 + L_1) = \bar{F}_2(Q_2 + S - M - L_1)$$
$$\mathbb{P}(A(X,\xi) > S | \xi_2 > Q_2 + L_2) = \bar{F}_1(Q_1 + S - M - L_2)$$

Together with (3.15) and (3.16), this leads to a pair of equations for limit orders sizes:

$$\bar{F}_2(Q_2 + S - M - L_1) = \frac{\lambda_u + h + r_1}{\lambda_u + \lambda_o} \quad \bar{F}_1(Q_1 + S - M - L_2) = \frac{\lambda_u + h + r_2}{\lambda_u + \lambda_o}$$

whose solution is given by $L_1^\star, L_2^\star$ from (3.21a,3.21b). To obtain the equation (3.21c), we rewrite the first equation in (3.15,3.16) using the inequalities (3.22a-3.22c). Then $P(A(X,\xi) > S)$ may be computed as the integral of the product measure $F_1 \otimes F_2$ over the region defined by

$$U(Q,S,M,L_1,L_2) =$$
$$\{(x_1,x_2) \in \mathbb{R}^2,\ x_1 > Q_1 + S - M - L_2,\ x_2 > Q_2 + S - M - L_1,\ x_1 + x_2 > Q_1 + Q_2 + S - M\}.$$

This integral is given by

$$\begin{aligned} &P(A(X,\xi) > S) = F_1 \otimes F_2\left(U(Q,S,M,L_1,L_2)\right) \\ &= \bar{F}_1(Q_1 + L_1)\bar{F}_2(Q_2 + S - M - L_1) + \int\limits_{Q_1+S-M-L_2}^{Q_1+L_1} \bar{F}_2(Q_1 + Q_2 + S - M - x_1)dF_1(x_1) \\ &= \frac{\lambda_u - (h+f)}{\lambda_u + \lambda_o} \end{aligned}$$

■

In the solution (3.21a)–(3.21b) optimal limit order quantities $L_1^\star, L_2^\star$ are linear functions of an optimal market order quantity $M^\star$. When (3.21a)-(3.21b) are substituted into (3.21c) we obtain a (non-linear) equation for $M^\star$, which can be solved for a given distribution of (ξ_1, ξ_2).

3.5 Numerical solution of the optimization problem

Computing the objective function in the order placement problem (Problem 1) or its gradient at any point requires calculating an expectation (a multidimensional integral) which is not analytically tractable aside from special examples. Fortunately, there are numerical methods developed specifically for problems where the objective function is an expectation, and they turn out to be very useful for this problem. Based on these methods we propose a procedure for computing the optimal order placement policy using historical data samples without specifying an order outflow distribution.

Our numerical solution is based on the robust stochastic approximation algorithm of [95]. Consider an objective function $V(X) \triangleq \mathbb{E}[v(X,\xi)]$ to be minimized and denote by $g(X,\xi) \triangleq \nabla v(X,\xi)$ where the gradient is taken with respect to X. The stochastic approximation algorithm tackles the problem of minimizing $V(X)$ in the following way:

1: Choose $X_0 \in \mathbb{R}^{K+1}$ and a step size γ;
2: **for** $n = 1, \ldots, N$ **do**
3: Draw an i.i.d. random variable $\xi^n \in \mathbb{R}^K$ from a distribution F
4: Set $X_n = X_{n-1} - \gamma g(X_{n-1}, \xi^n)$
5: **end**
6: Compute $\hat{X}^\star \triangleq \frac{1}{N}\sum_{n=1}^{N} X_n$

Here ξ^n are independent across n, but the components of each draw ξ^n need not be independent. The iterative algorithm produces an estimate $\hat{X}^\star$, which converges to the optimal point $X^\star$ under some weak assumptions. In particular, it has a performance bound $V(\hat{X}^\star) - V(X^\star) \leq \frac{C}{\sqrt{N}}$, where the constant C depends on K, S and other problem parameters[4]. In general stochastic approximation methods require sampling random variables ξ from the distribution F at each step to compute $g(X,\xi)$. But in our case this function takes

[4]The method assumes that $\min_{X\in\mathcal{X}}\{V(X)\}$ is sought, where $V(X)$ is a well-defined and finite-valued expectation for every $X \in \mathcal{X}$ and $\mathcal{X}$ is a non-empty bounded closed convex set. Moreover $V(X)$ needs to be continuous and convex on $\mathcal{X}$. The optimal step size is $\gamma = \frac{D}{\sqrt{N}M}$ and the constant $C = DM$, where $D = \max_{X,X'\in\mathcal{C}} \|X - X'\|_2$, $M = \sqrt{\max_{X\in\mathcal{C}} \mathbb{E}\left[\|g(X,\xi)\|_2^2\right]}$. We use a step size $\gamma = K^{1/2}S\left(N(h+f+\lambda_u+\lambda_o)^2 + N\sum_{k=1}^{K}(h+r_k+\lambda_u+\lambda_o)^2\right)^{-1/2}$ which scales appropriately with problem parameters. For more details we refer to [78] and [95].

a particularly simple form which makes it unnecessary to specify the distribution F:

$$g(X_n,\xi) = \begin{pmatrix} h+f-\lambda_u \mathbb{1}_{\{A(X_n,\xi)<S\}} + \lambda_o \mathbb{1}_{\{A(X_n,\xi)>S\}} \\ -(h+r_1)\mathbb{1}_{\{\xi_1>Q_1+L_{1,n}\}} - \lambda_u \mathbb{1}_{\{A(X_n,\xi)<S,\xi_1>Q_1+L_{1,n}\}} + \lambda_o \mathbb{1}_{\{A(X_n,\xi)>S,\xi_1>Q_1+L_{1,n}\}} \\ \dots \\ -(h+r_K)\mathbb{1}_{\{\xi_K>Q_K+L_{K,n}\}} - \lambda_u \mathbb{1}_{\{A(X_n,\xi)<S,\xi_K>Q_K+L_{K,n}\}} + \lambda_o \mathbb{1}_{\{A(X_n,\xi)>S,\xi_K>Q_K+L_{K,n}\}} \end{pmatrix}$$

Note that $g(X_n,\xi)$ depends on ξ only through indicator functions, which have simple economic meaning. For example $\mathbb{1}_{\{A(X_n,\xi)<S\}} = 1$ if the trader fell behind the target size and $\mathbb{1}_{\{\xi_k>Q_k+L_{k,n}\}} = 1$ if a limit order L_k was fully executed. As a consequence, the solution is updated on each step in response to order execution outcomes - limit order fills and target size shortfalls or excesses. For example, the first component of $g(X_n,\xi)$ describes market order size updates - on each step this size is decreased by $\gamma(h+f)$ to reduce the cost, increased by $\gamma\lambda_u$ if the trader fell behind the target size, or decreased by $\gamma\lambda_o$ if the target was exceeded. Limit order sizes are updated only when these orders are executed. Their sizes are either increased or decreased depending on whether there was an execution shortfall or excess.

This iterative algorithm gives a specific way to resample past order fill data or historical order flow data and obtain a solution for the order placement problem. Since this method involves only basic arithmetic operations it can also be implemented for on-line order routing optimization in real-time trading applications. Alternatively, one can follow the same steps using a parametric model for F to simulate ξ and compute the optimal order allocation off-line based on a parametric model for order flows.

We apply this algorithm to several numerical examples using typical parameter values for U.S. equity markets. We start by comparing numerical and closed-form solutions in the single exchange case to assess the algorithm stability and convergence. Our example assumes that a trader is buying $S = 1000$ shares of a stock with a deadline $T = 1$ minute with an initial bid queue size $Q = 2000$ shares, order outflow $\xi \sim Pois(\mu T)$ and $\mu = 2200$ shares per minute. With these parameters a small limit order is likely to be executed before T, but a limit order for $S = 1000$ shares is unlikely to be fully

filled. The pricing parameters and penalty costs (all in dollars per share) are set to $h = 0.02, r = 0.002, f = 0.003, \lambda_o = 0.024, \lambda_u = 0.026$. According to (3.10) the optimal allocation with these parameters is $(M^\star, L^\star) = (728, 272)$ shares. Numerical estimates $\hat{X}^\star$ were computed for five initial points X_0 using a different number of samples N. For each choice of X_0 and N the objective function $V(X)$ was approximated with averages $W(X) = \frac{1}{L}\sum_{i=1}^{L} v(X, \xi_i))$ taken over additional $L = 1000$ samples of ξ. Figures 3.3 and 3.4 show that estimates converge to $X^\star$ regardless of the initial point X_0 and moreover when $X_0 = X^\star$ the iterates remain close to the optimal point. Convergence is also quite fast - after as few as 50 samples the algorithm is within 2% of the optimal objective value. In the worst case of initial points on the boundary it can take a few thousand samples to converge. It is also worth noting that convergence in terms of the objective value occurs significantly faster than convergence in terms of the order allocation vector.

We also estimated savings from dividing orders among multiple exchanges in a naive way and according to our solution. Denote a pure market order allocation by $X_M = (S, 0, \ldots, 0)$, a single limit order allocation by $X_L = (0, S, 0, \ldots, 0)$ and an equal split allocation by $X_E = (\frac{S}{K+1}, \frac{S}{K+1}, \ldots, \frac{S}{K+1})$. Table 1 presents numerical solutions with $X_0 = X_E, N = 1000, L = 1000$ for different order sizes S and a varying number of exchanges $K = 1, \ldots, 5$. The parameters $s, f, r, \lambda_u, \lambda_o$ are same as in the previous simulation and K exchanges are identical replicas of each other: $r_k = r, Q_k = Q$ and $\xi_{n,k} \sim Pois(\mu T)$ are i.i.d., $k = 1, \ldots, K, \quad n = 1, \ldots, N$. Optimal order allocations clearly outperform the naive benchmarks, especially when a target quantity S is relatively small. This is because small quantities can be allocated in form of limit orders among available exchanges and fully capture limit order cost savings, while larger quantities are traded with costly market orders even with the optimal allocation. Comparing $W(X_L)$ and $W(X_E)$ we also see that splitting limit orders across multiple exchanges, even in a naive way, can be very advantageous when limit order fills are independent. Since multiple exchanges in this example are copies of each other, the algorithm splits the total limit order amount equally among them. There is however a difference between the equal split allocation X_E and the optimal allocation. The former sets a market order size to $\frac{S}{K+1}$, which may be too big or too small depending on problem parameters. Another interesting feature of the numerical solution $\hat{X}^\star$ is its

tendency to oversize the total quantity of orders: $M + \sum_{k=1}^{K} L_k > S$ for $S = 1000, 5000$ and $K = 4, 5$. This may be a consequence of assumed independence between ξ_k - by submitting large orders to multiple exchanges the algorithm reduces the probability of falling behind the target quantity with a relatively low probability of exceeding it.

K	$(\hat{M}^\star$	$\hat{L}_1^\star$	$\hat{L}_2^\star$	$\hat{L}_3^\star$	$\hat{L}_4^\star$	$\hat{L}_5^\star)/S$	$W(X_M)$	$W(X_L)$	$W(X_E)$	$W(\hat{X}^\star)$
$\bar{S} = 500$										
1	0.481	0.519					11.50	3.36	2.79	2.76
2	0.034	0.601	0.615				11.50	3.48	-2.86	-6.00
3	0.003	0.438	0.433	0.421			11.50	3.39	-5.22	-10.56
4	0.002	0.280	0.277	0.273	0.264		11.50	3.52	-6.44	-10.84
5	0.001	0.198	0.215	0.224	0.206	0.214	11.50	3.35	-7.24	-10.91
$\bar{S} = 1000$										
1	0.713	0.287					23.00	16.30	14.80	14.20
2	0.484	0.343	0.338				23.00	16.43	5.88	5.48
3	0.268	0.338	0.334	0.336			23.00	16.29	-3.16	-3.61
4	0.055	0.316	0.313	0.351	0.333		23.00	16.48	-9.27	-12.12
5	0.003	0.309	0.300	0.300	0.309	0.321	23.00	16.44	-12.88	-19.76
$\bar{S} = 5000$										
1	0.839	0.161					115.00	120.33	112.83	107.76
2	0.747	0.192	0.192				115.00	120.44	105.86	99.65
3	0.693	0.189	0.189	0.189			115.00	120.40	97.35	90.71
4	0.650	0.186	0.186	0.186	0.186		115.00	120.37	88.64	81.89
5	0.614	0.167	0.167	0.167	0.167	0.167	115.00	120.40	79.63	72.93

Table 3.1: Savings from order splitting.

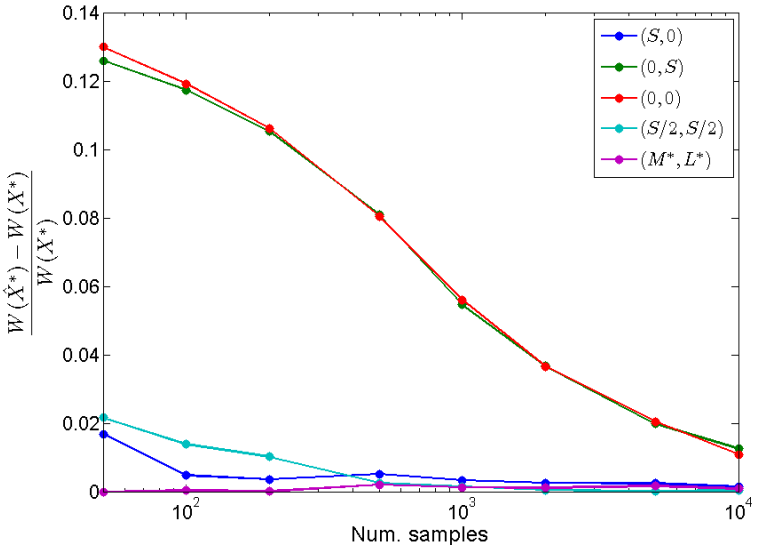

Figure 3.3: Convergence of objective values to an optimal point for different initial points.

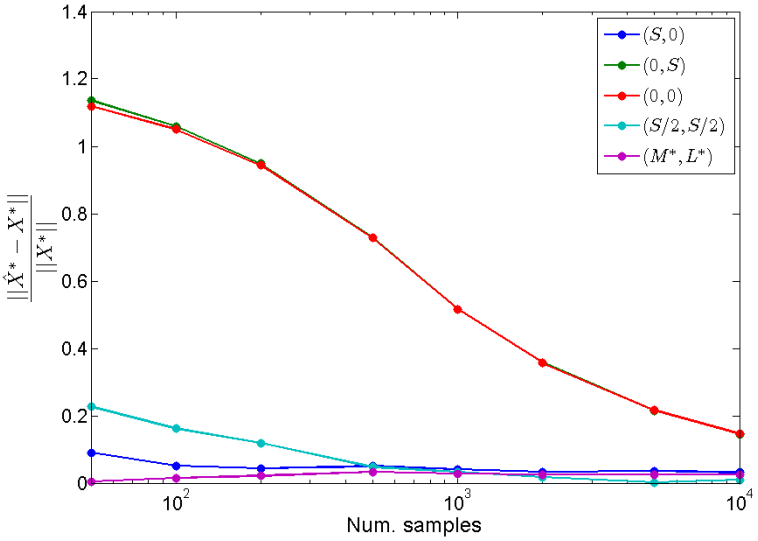

Figure 3.4: Convergence of order allocation vectors to an optimal point for different initial points.

Modern high-frequency trading activity connects different trading venues (see e.g. [92]), so in reality order flows across exchanges are highly positively correlated. To investigate how order flow correlations affect order placement strategies and order execution costs we use an example with two exchanges and a common factor in order flows:

$$\xi_1 = \alpha\xi_0 + (1-\alpha)\epsilon_1, \quad \xi_2 = \alpha\xi_0 + (1-\alpha)\epsilon_2,$$

where $\xi_0, \epsilon_1, \epsilon_2 \sim Pois(\mu T)$ are three i.i.d. random variables and the scalar parameter $\alpha \in [0,1]$ controls the degree of positive correlation between ξ_1 and ξ_2. We set $\mu = 2200$, $(Q_1, Q_2) = (1900, 2000)$ and the rest of the parameters are the same as in the previous example. As the parameter α increases, we can see that the sum of order sizes $M + L_1 + L_2$ decreases to the target quantity $S = 1000$. When $\alpha \to 1$, the second exchange does not provide any benefit in terms of diversifying execution risk, so it is not optimal anymore to oversize the total quantity of orders. This is similar to the case of a single exchange where $M + L = S$ by Proposition 4.

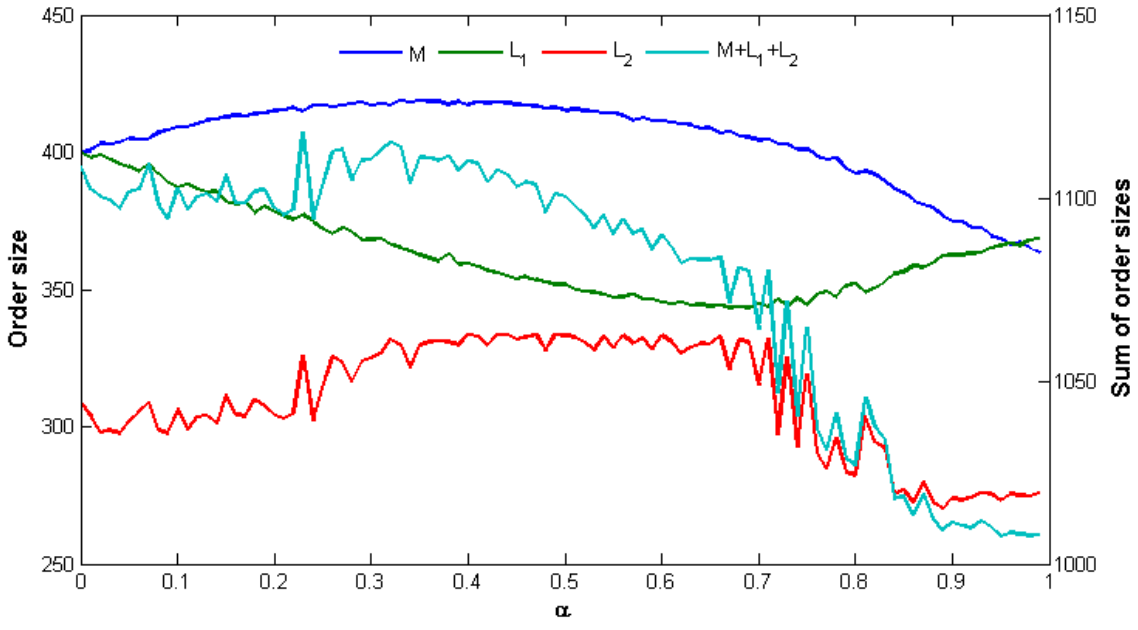

Figure 3.5: Optimal order allocations with correlated order flows.

However, executing orders on two exchanges with order flows (ξ_1, ξ_2) and queue sizes (Q_1, Q_2) is not equivalent to the case of a single exchange with $\xi = \xi_1 + \xi_2$ and $Q = Q_1 + Q_2$, even if the correlation between ξ_1 and ξ_2 is high. Figure 3.6 compares average execution costs for these two cases with an optimal order allocation and different values α. It appears from this example that the availability of multiple exchanges reduces the costs of trading compared to a single consolidated exchange, but only as long as multiple exchanges remain relatively uncorrelated.

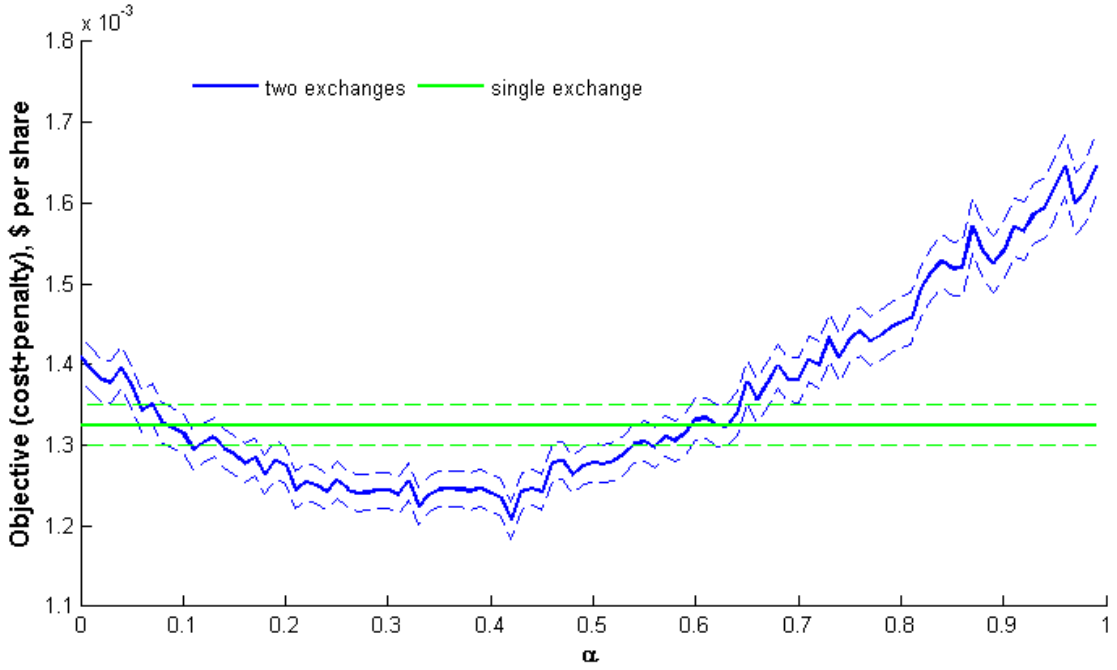

Figure 3.6: Comparison of average order execution costs: two exchanges with correlated order flows against vs a single consolidated exchange. Dashed lines show 95% confidence intervals for averages.

To further illustrate the structure of a numerical solution we performed a sensitivity analysis with $K = 2$ exchanges and parameters $Q_1 = Q_2 = 2000$, $S = 1000$, $\xi_{1,2} \sim Pois(\mu_{1,2}T)$, $\mu_1 = 2600$, $\mu_2 = 2200$, $T = 1$, $h = 0.02$, $r_1 = r_2 = 0.002$, $f = 0.003$, $\lambda_u = 0.26$ and $\lambda_o = 0.24$. Varying some of these parameters one at a time we plot the numerical solution $\hat{X}^\star$ after $N = 1000$ iterations, together with an analytical solution for a single exchange. The results are presented on Figures 3.7 and 3.8. Similarly to the single

exchange case, limit order sizes on two exchanges L_1, L_2 decrease and market order size M increases as the penalty λ_u increases. Increasing the half-spread h, the rebate r_1 or the fee f makes a limit order on exchange number one more attractive, so L_1 increases and M decreases. Because the penalty λ_u is large in this example, execution risk is more important than fees and rebates, therefore the queue size Q_1 and the order outflow mean μ_1 have a much stronger effect on the optimal solution than r_1. Both decreasing the Q_1 and increasing μ_1 make a limit order fill more likely at exchange number one and L_1 increases[5]. Finally, as in the case of a single exchange, the target size S has a strong effect on the optimal order allocation. Only limit orders are used while S is small, but as it becomes larger it is difficult to fill that amount solely with limit orders and the optimal market order size begins to grow to limit the execution risk.

3.6 Conclusion

We have formulated *optimal order placement problem* for a market participant able to submit market orders and limit orders across multiple exchanges as a well-posed optimization problem, and studied the solution of this problem in various configurations. In the case when there is a single exchange we have shown that this leads to an optimal split between limit and market orders. For the general case of K exchanges, we have given a characterization of the optimal order placement strategy and propose a stochastic approximation algorithm for numerically computing it. Using this algorithm, we have explored the properties of an optimal order allocation and showed that an optimal routing of orders across multiple exchanges can lead to a substantial reduction in transaction costs.

[5]The observed drop in L_1 for large μ_1 and small Q_1 appeared only in this example, we were not able to replicate it for other distributions of ξ.

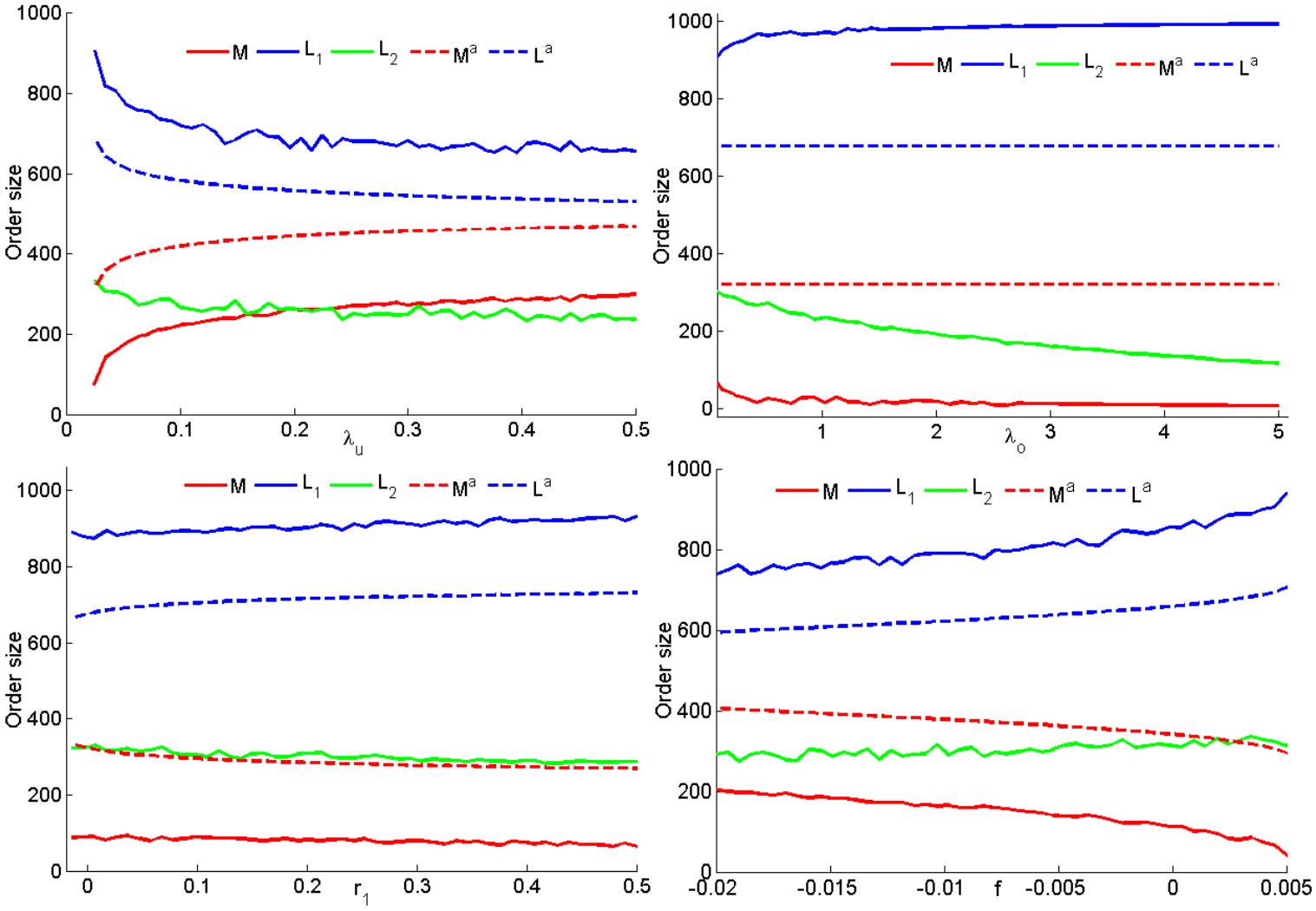

Figure 3.7: Sensitivity analysis for a numerical solution $\hat{X}^\star = (M, L_1, L_2)$ with two exchanges and an optimal solution (M^a, L^a) with the first exchange only.

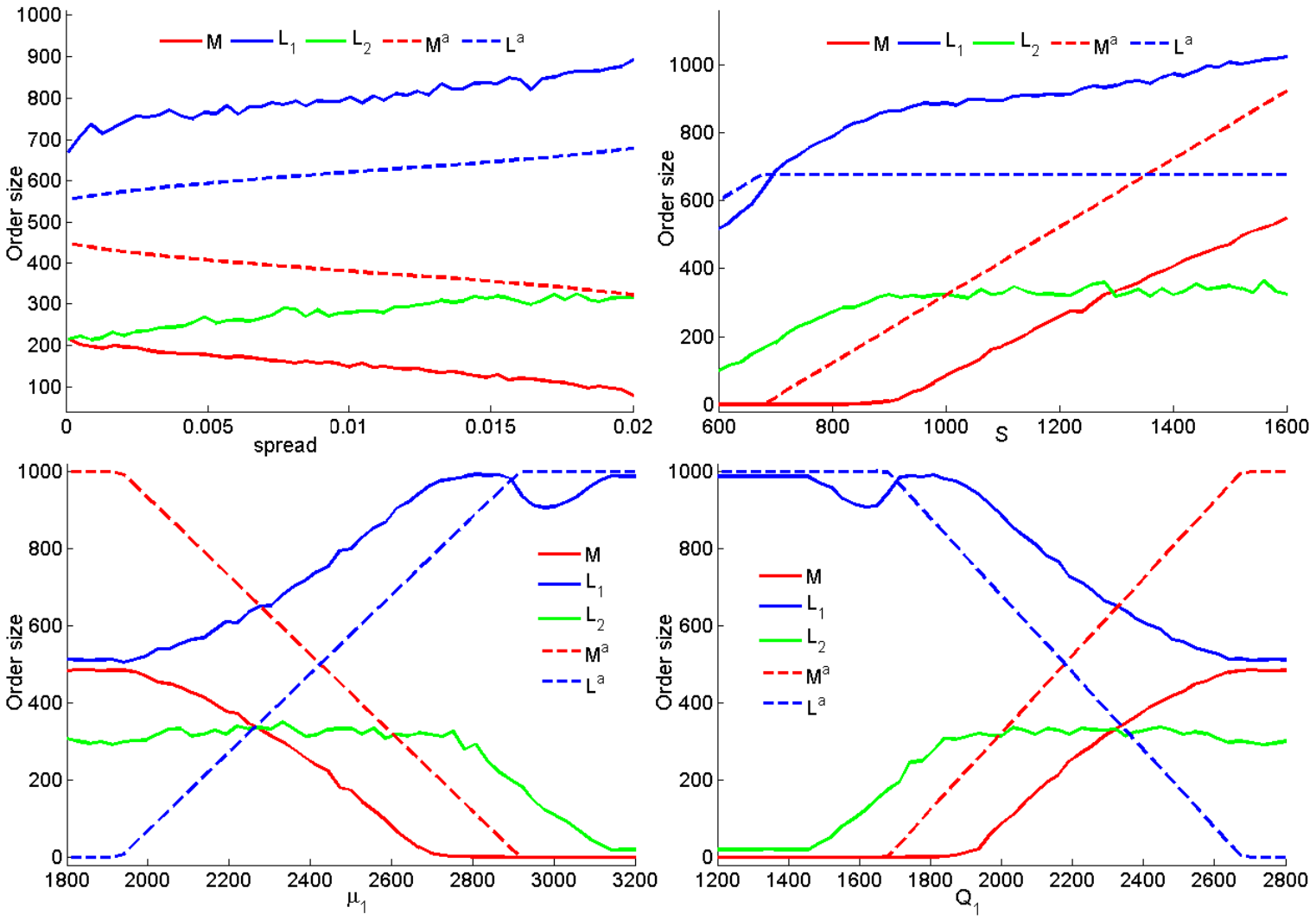

Figure 3.8: Sensitivity analysis for a numerical solution $\hat{X}^\star = (M, L_1, L_2)$ with two exchanges and an optimal solution (M^a, L^a) with the first exchange only.

Chapter 4

Heterogenous traders in a limit order book

This chapter is based on the paper "A Limit Order Queue Model with Heterogenous Traders" [77] which is a joint work with Professor Costis Maglaras.

4.1 Introduction

> *It should be observed that three classes of men are to be distinguished on the stock exchange. The princes of business belong to the first class, the merchants to the second, and the speculators to the last.*
>
> Joseph de la Vega, Confusion de confusiones, 1688

During the last two decades, financial markets worldwide experienced a fundamental transformation fueled by computer technology and new regulation. Evidence shows that automated electronic trading reduced various measures of trading costs (see [60; 69; 90]). But the increasing complexity of electronic financial markets is also stimulating a substantial debate on the overall benefits of their current structure. This discussion together with a practical need to guide trading decision in a complex environment sparked an active interest in *limit order book modeling*.

Most financial markets in the world are organized as limit order books [66; 112]. In this market structure liquidity is provided and prices are formed through a collective activity of a diverse trader population. Strategies of individual traders include decisions on the timing of their order submissions, as well as on order prices and quantities. These strategies are flexible, complex and can depend on the previous activity of other traders, making it difficult to model limit order books in an economic equilibrium framework. Nevertheless useful insights can still be derived from reduced-form models of limit order books (see [100] for a recent survey). Such "engineering" approach is further facilitated by the simplicity and formality of basic rules underlying limit order trading as compared to other, more opaque market structures (e.g. dealer markets).

On the fine level of individual order submissions, a limit order book can be represented as a multiclass queueing system. To buy or sell an asset, traders can use two types of instructions - limit and market orders. Limit orders specify the maximum quantity to be bought or sold and the worst acceptable price, while market orders specify only a quantity. In a standard "price-time" priority scheme, adopted by most financial markets[1], limit orders with identical prices form *first-in first-out queues* across different price levels as they arrive to an exchange. The queue of buy orders with the highest price at a given time is called the *best bid* and the queue of sell orders with the lowest price - the *best ask*. Market orders to buy (sell) are matched only with limit orders at the front of the best ask (bid) queue. From the queueing point of view market orders provide "service" (liquidity) to these limit order queues[2]. A large number of limit orders is canceled and these orders abandon their queues before reaching execution (in U.S. equities markets more than 90% of orders are canceled [57]). When the best bid or ask queue is depleted by market orders and cancelations, the queue at next-best price becomes the new best bid or ask. A new best bid or ask queue can also be initiated by a new limit order with a price between the current best bid and ask prices.

[1]Some markets, for example NASDAQ OMX PSX and certain interest rate futures markets, follow different schemes - see [67] for a discussion on order matching rules.

[2]Aggressive limit orders to buy/sell whose prices are above/below the best ask/bid are similar to market orders. For simplicity we also do not specifically consider derived order types such as discretionary orders and iceberg orders, most of which are equivalent to simple strategies employing limit and market orders.

Limit order books are naturally modeled as dynamic stochastic systems, specifically as queueing systems. An early development in this direction were "zero-intelligence models" in econophysics [41; 108] where instead of making detailed behavioral assumptions regarding traders' strategies authors posited that orders arrive and cancel purely at random (hence the name). Subsequent stochastic queueing models [31; 62] added more realistic features of limit order books - order queues at a discrete price grid, varying order arrival rates and random order sizes. These models are easily calibrated to real data and provide tractable procedures for computing various quantities of practical interest. For instance, the distribution of a next price move, a valuable input for optimal timing of order submissions [111], can be obtained in these models using standard tools from queueing theory such as Laplace transforms [31], or diffusion approximations [11; 29; 30]. Queueing models of limit order books also provide valuable insights into the aggregate behavior of financial markets. They establish a link between limit order flow dynamics on a microscopic scale and price volatility on a larger timescale [30], and allow to describe order flow coupling and queue dynamics in a fragmented market where agents strategically route their orders [94].

A common simplifying assumption made in stochastic order book models is that order arrivals and cancelations are generated by a single mechanism, some random process. However, in reality there are significant differences across strategies of market participants [13; 75] and this heterogeneity plays a key role in classical market microstructure models. Traders differ in terms of their patience [45; 106], information on future asset returns [34; 79], trading objectives and holding periods [57]. This heterogeneity leads to profound differences between properties of their order flows that are hard to capture with a single stochastic process for all order arrivals and cancelations. Another typical assumption is that order arrivals and departures are generated by an exogenous process with a common timescale for all orders. Such process may represent orders of fund managers who come to the market to fulfill their exogenous liquidity needs. But orders of other trading strategies (e.g. high-frequency market-makers) depend mostly on the dynamics of the order book itself, i.e. these orders are generated *endogenously* within the market environment and typically have a *different timescale* for arrivals and cancelations. A recent empirical study [57] contrasts "agency algorithms" that are used by money managers for trading large positions, with

"principal algorithms" whose goal is to profit from the market environment itself. Agency algorithms tend to update their orders based on a fixed time schedule, while principal algorithms operate in "real time" reacting to order book updates almost immediately. The queueing model developed here incorporates this distinction and the endogenous nature of orders generated by market-makers. In our knowledge this is one of the first attempts to unite reduced-form stochastic modeling of limit order books with more detailed market microstructure models.

We propose a two-class queueing model of limit order book dynamics that distinguishes between orders of institutional fund managers and high-frequency traders. Specifically, orders of fund managers arrive to a limit order queue according to an exogenous process, while high-frequency traders submit and cancel their orders in response to queue size changes. Our analysis of this model shows how two classes of orders propagate through a FIFO queue and delivers a procedure for forecasting waiting times of new orders. The model can structurally explain abnormally long waiting times in large limit order queues of large-tick stocks. This phenomenon is related by practitioners to a "crowding" of orders coming from high-frequency traders [42]. Additionally, our model helps to explain limit order cancelation behavior better than the existing models. It makes predictions on the amount of cancelations, their dependence on a queue size and on a queue position of canceled orders. Order waiting time estimates derived from our model can be used in practical order management systems to improve order placement and order routing decisions. From a stochastic modeling viewpoint, our model seems novel and different from the bulk of the literature on that topic that has been largely motivated by the analysis of call centers. In comparison to previous studies of multi-class queues with abandonments (e.g. [68]) our model introduces a new kind of heterogeneity, stemming from an endogenous, state-dependent dynamics of one class of orders. This analysis can be of interest in studying service systems for heterogeneous populations with abandonment. To empirically validate our predictions we use a proprietary data set from a broker providing of order execution services. The data describes actual orders submitted by firm's agency algorithms, their outcomes (executions or cancelations) and their realized queue waiting times. To the best of our knowledge such data has not been previously studied in academic literature. Our empirical results verify

that the two-class model makes realistic predictions for waiting times.

The rest of this paper is structured as follows. In Section 4.2 we introduce a two-class queueing model with state-dependent order behavior. Section 4.3 presents our results for a fluid approximation of this model, Section 4.4 describes empirical tests with order data and Section 4.5 concludes.

4.2 Model description

We develop a two-class queueing model for orders at the best bid (or ask) of an exchange. New limit orders arrive at random times and join at the back of the order queue. Market orders also arrive at random times and match with a limit order currently in front of the queue, if there is any, and all orders have a unit size. We assume that limit and market orders are submitted by a heterogeneous population of traders and that limit orders belong to one of two classes:

Type-1 limit orders are submitted by fund managers whose goal is to satisfy specific needs to buy or sell their assets within certain time constraints. These orders arrive according to an exogenous Poisson process with a rate λ and each has a finite patience deadline, exponentially distributed with a mean $1/\gamma$. Whenever a realized waiting time of a type-1 order reaches its patience deadline, the order is canceled.

Type-2 limit orders are submitted by high-frequency traders. They have no exogenous liquidity needs or time constraints. Instead, their orders react to changes in the current queue state:

- When any limit order joins the queue and its size becomes equal to q, one type-2 limit order may instantaneously join the queue with a probability $F(q)$.
- When any limit order leaves the queue (due to a trade or a cancelation) and the queue size becomes equal to q, each type-2 limit order that is currently in the queue may instantaneously cancel with a probability $G(q)$[3].

[3]Since multiple type-2 orders may cancel at the same epoch, we assume that after an order departure each type-2 order in the queue is selected with a probability $G(q)$, and if more than one order is selected they cancel as a batch. If at least one type-2 order so cancels, the selection and batch cancelation procedure is repeated with the new queue size.

Market orders are submitted by all traders and arrive according to an exogenous Poisson process with a rate μ.

Type-2 order arrivals and cancelations are endogenously generated in this model, i.e. they can be triggered only by another queue update. Since type-2 order arrivals or cancelations can trigger further type-2 orders arrivals or cancelations, these events can cascade. We assume that the function $G(q)$ is non-increasing, and that $F(q)$ is non-decreasing near zero with $F(0) = 0$. These conditions imply that type-2 orders tend to persist (do not cancel) in large queues, but cancel more easily from small queues and also tend to avoid joining small queues.

Such state-dependent queue behavior is observed in practice - fewer orders join a small queue than a large queue, and more orders cancel from a small queue. Figure 4.2 shows how average 10-second arrival, cancelation and trade volumes vary with the initial queue size at the beginning of the 10 second time interval. A possible explanation for this behavior is that small bid (ask) queue size implies a higher probability of a future negative (positive) price change [11; 29]. To avoid trading at a price just before it changes in an adverse direction, traders may prefer submitting orders to large queues where the risk of a price change is small. For the same reason they may also cancel their orders from small queues. In our model, this concerns only type-2 orders, since type-1 orders are driven solely by patience and do not react to queue size fluctuations.

This general model description presented here covers a class of state-dependent queue models with different functions $F(q), G(q)$. For the sake of tractability we pursue a particular choice of $F(q) = \phi \mathbb{1}_{\{q>\theta\}}, G(q) = \mathbb{1}_{\{q\leq\theta\}}$, where $0 < \phi < 1$ is a constant characterizing the participation rate of type-2 orders (e.g. orders of high-frequency traders), and θ is a threshold at which all type-2 orders instantly cancel[4]. We define the average size of a type-2 order arrival cascade $f \stackrel{\Delta}{=} \frac{\phi}{1-\phi}$ and note that non-trivial queue dynamics are observed when $\lambda f < \mu < \lambda(1 + f)$, otherwise the queue tends to build up to infinity or decrease to zero.

[4]as perceived by slower market participants

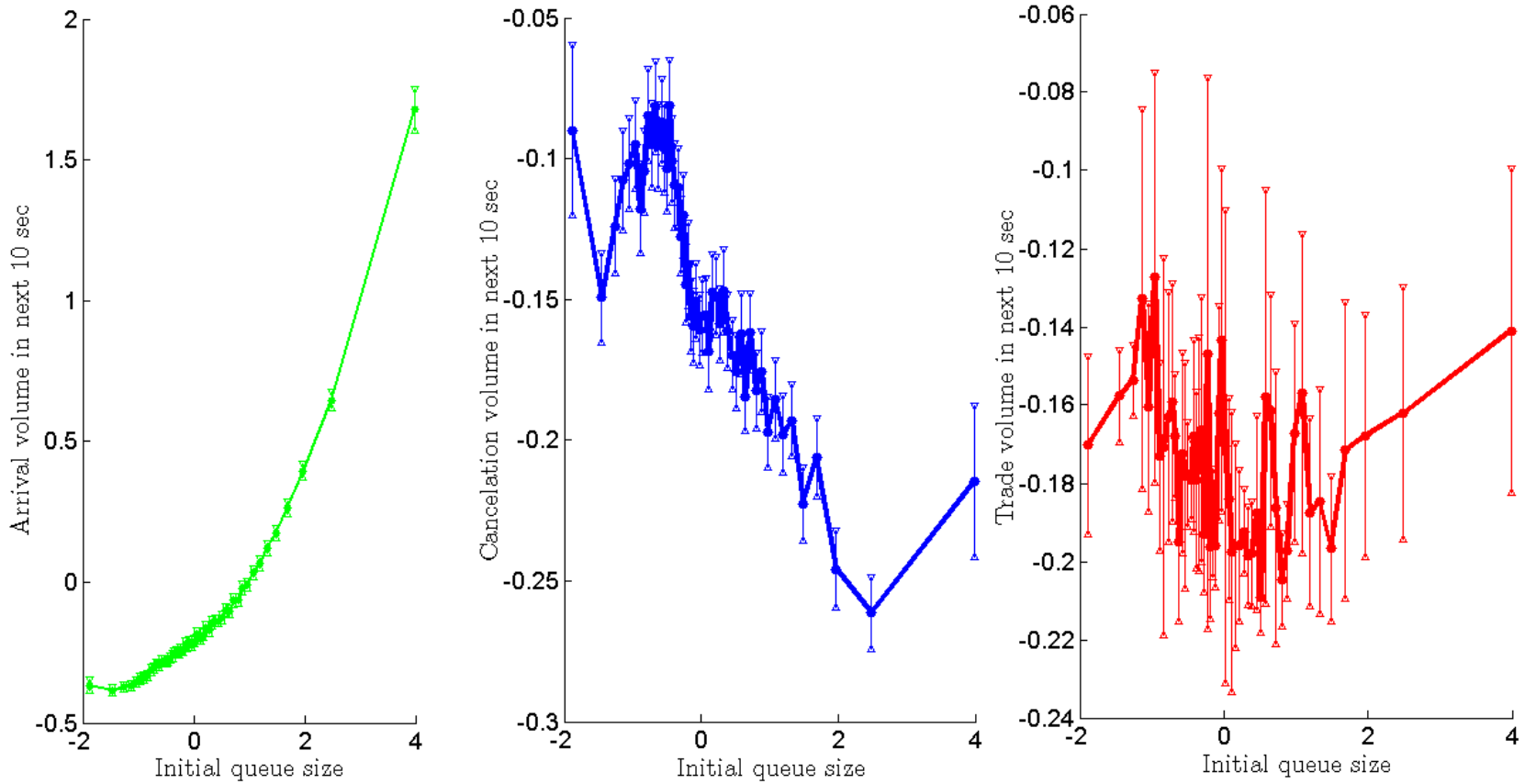

Figure 4.1: Average arrival, cancelation and trade volumes per 10 seconds, with confidence intervals, for different initial queue sizes. This plot is based on trade and quote data for 30 Dow Jones stocks and all U.S. exchanges on 03/08/2012. All volumes and queue sizes were standardized with averages and standard deviations computed separately for each stock, exchange and half-hour interval during the trading day. To avoid the influence of "fleeting quotes" we only considered 10-second time intervals without a price change.

4.3 Fluid model

A direct analysis of the stochastic multiclass queueing model is complicated by event-driven and state-dependent behavior of type-2 orders which motivates us to consider its fluid approximation. The fluid model describes average dynamics of the stochastic model over time intervals that are long in comparison with typical order inter-arrival times. It assumes that the contribution of each order arrival or cancelation is small in comparison with the queue size, and replaces discrete orders with a continuous fluid that is added to the queue by limit order arrivals and is removed by market orders and cancelations. At the back of the queue, infinitesimal quanta of type-1 orders arrive at a constant rate λ. Quanta of type-2 orders also arrive at the back of the queue at a constant rate λf, as long as the queue size is larger than the threshold θ. In front of the queue, orders of both types leave the queue at a rate μ due to executions, and additionally type-1 orders leave the queue due to cancelations.

The basic components of this fluid model are $q_{1,2}(t,y)$ - quantities of type-1 and type-2 fluid that are in the queue at time t and have waited there for at most y units of time. With this notation the total amount of type-1 or type-2 fluid is denoted by $q_{1,2}(t) \stackrel{\Delta}{=} q_{1,2}(t,\infty)$ and the total fluid amount (queue size) is $q(t) \stackrel{\Delta}{=} q_1(t) + q_2(t)$. We can also write $q_{1,2}(t) = q_{1,2}(t,\tau(t))$, where $\tau(t)$ is the head-of-line waiting time, defined as:

$$\tau(t) = \inf_{y\geq 0}\{y|q_1(t,\infty) - q_1(t,y) = 0\} = \inf_{y\geq 0}\{y|q_2(t,\infty) - q_2(t,y) = 0\}$$

We assume that for all t, y functions $q_{1,2}(t,y)$ are differentiable with respect to y, and can be represented as

$$q_{1,2}(t,y) = \int_0^y \zeta_{1,2}(t-u,u)du$$

where $\zeta_{1,2}(t-u,u), u \leq t$ is a density of fluid that arrived at time $t-u$ and waited for exactly u units of time. We are now ready to describe the queue evolution.

Fluid quantity dynamics:

$$\begin{aligned}
\dot{q}_1(t) &= \lambda - \gamma q_1(t) - \alpha(t)\mu \\
\dot{q}_2(t) &= (\lambda f - (1-\alpha(t))\mu)\,\mathbb{1}_{\{q(t)>\theta\}} - q_2(t)\delta\left(\mathbb{1}_{\{q(t)>\theta\}}\right) \\
\alpha(t) &= \frac{\zeta_1(t-\tau(t),\tau(t))}{\zeta_1(t-\tau(t),\tau(t)) + \zeta_2(t-\tau(t),\tau(t))}
\end{aligned} \tag{4.1}$$

where $\delta(x)$ is the delta function.

Fluid density dynamics:

$$\begin{aligned}
&\zeta_1(t,0) = \lambda, t > 0 \\
&\zeta_2(t,0) = \lambda f \mathbb{1}_{\{q(t)>\theta\}}, t > 0 \\
&\frac{\partial \zeta_1(t,u)}{\partial u} = -\gamma \zeta_1(t,u), t \geq 0, 0 < u \leq \tau(t) \\
&\zeta_2(t,u) = \zeta_2(t-u,0)\mathbb{1}_{\left\{\min\limits_{0\leq v\leq u}(q(t-v))>\theta\right\}}, t \geq 0, 0 < u \leq \tau(t)
\end{aligned} \tag{4.2}$$

It easily follows from (4.2) that the density of type-1 orders that arrived after the initial moment, $\zeta_1(t,u) = \lambda e^{-\gamma u}$ and in particular it does not depend on t. However, arbitrary initial densities $\zeta_{1,2}(0,u)$ may introduce a non-linear delay into ordinary differential equations (4.1) through $\alpha(t)$. Differential equations of this kind are generally hard to solve, but this complexity can be avoided by assuming a special queue structure, hereafter called *regular queue structure*.

Assumption R_t: For a fixed $t \geq 0$, $\theta < q_1(t) < \frac{\lambda}{\gamma}, q_2(t) > 0$ *and* $\zeta_1(t,u) = \lambda e^{-\gamma u}, \quad \zeta_2(t,u) = \lambda f, \quad 0 \leq u \leq \tau(t)$

Proposition 6 shows that the regular queue structure is stable - once R_t is true, R_s is also true for all $s > t$. Proposition 7 shows that all sufficiently large queues must be regular, and Proposition 8 shows that the model converges to the regular structure in a finite amount of time. Except for an initial time interval during which arbitrary initial conditions may lead to complicated dynamics, the queue structure is regular and differential equations (4.1) are non-delayed.

Proposition 6 *Assume that* $\frac{\lambda-\mu}{\gamma} > \theta$ *and* R_0 *is true, then* R_t *holds for all* $t > 0$. *Moreover, equations* (4.1) *are non-delayed for all* $t > 0$ *and* $\alpha(t), \tau(t)$ *can be computed as follows:*

$$\alpha(t) = \frac{e^{-\gamma\tau(t)}}{e^{-\gamma\tau(t)} + f} \tag{4.3}$$

$$\tau(t) = \frac{1}{\gamma}\log\left(\frac{\lambda}{\lambda - \gamma q_1(t)}\right) \tag{4.4}$$

Proof: Consider a function $p(t) \triangleq \left(\theta - \frac{\lambda-\mu}{\gamma}\right)e^{-\gamma t} + \frac{\lambda-\mu}{\gamma}$, and note that $p(t) > \theta, t > 0$ because $\frac{\lambda-\mu}{\gamma} > \theta$ and $p(0) = \theta$. Since $q_1(0) > p(0) = \theta$ and $\dot{q}_1(t) \geq \dot{p}(t), t > 0$ we conclude that $q(t) \geq q_1(t) = q_1(0) + \int_0^t \dot{q}(s)ds \geq p(0) + \int_0^t \dot{p}(s)ds > \theta, t > 0$. Since $q(t) > \theta$ for all $t > 0$, type-2 orders never cancel from the queue. Combining this with equations (4.2) and the assumption R_0 we can write $\zeta_1(t,u) = \lambda e^{-\gamma u}, \zeta_2(t,u) = \lambda f$ for all $t \geq 0, 0 \leq u \leq \tau(t)$. Therefore for any $t \geq 0$, $q_1(t) = \int_0^{\tau(t)} \zeta_1(t-u,u)du = \int_0^{\tau(t)} \lambda e^{-\gamma u}du = \frac{\lambda}{\gamma}\left(1 - e^{-\gamma\tau(t)}\right)$, which leads to (4.4). Note that in (4.4) the expression inside logarithm is positive because $q_1(t) < \frac{\lambda}{\gamma}\left(1 - e^{-\gamma t}\right) < \frac{\lambda}{\gamma}$, and therefore $\lambda - \gamma q_1(t) > 0$. Substituting the expressions for $\zeta_{1,2}(t,u)$ and $\tau(t)$ into (4.1) yields (4.3) where $\alpha(t)$ depends solely on $q_1(t)$. Therefore differential equations (4.1) depend only on the current state of $q_1(t), q_2(t)$ and not on their past. ■

Proposition 7 *If* $\frac{\lambda-\mu}{\gamma} > \theta$, $q(t) \geq \bar{Q} \triangleq \frac{\lambda\theta}{\mu+\theta\gamma} + \frac{\lambda f}{\gamma}\log\left(1 + \frac{\theta\gamma}{\mu}\right)$ *and* $\exists T_0, 0 < T_0 < t$ *such that* $q(T_0) < \theta$, *then* R_t *holds.*

Proof: Given that $q(T_0) < \theta$ we can write for some $s > T_0$

$$\dot{q}(s) = \lambda - \gamma q_1(s) - \mu$$

which is solved by $q(s) = \left(q(T_0) - \frac{\lambda-\mu}{\gamma}\right)e^{-\gamma(t-T_0)} + \frac{\lambda-\mu}{\gamma}$. The function $q(s)$ is monotonically increasing towards $\frac{\lambda-\mu}{\gamma} > \theta > 0$. Since it is also continuous there exists a time T_1 such that $q(T_1) = q_1(T_1) = \theta$.

At time T_1 type-2 orders start arriving, but it takes some time for them to reach the front of the queue. Denote by $r_{1,2}(s) = q_{1,2}(s, s - T_1), s > T_1$ the quantities of type-1 and

type-2 fluids that arrived after T_1 and are still in the system at time s. Also denote by $p(s) = q_1(s) - r_1(s)$ the quantity of fluid that arrived before T_1 and is still in the queue. For $s > T_0 > 0$ we can use (4.2) to write $r_1(s) = \int_0^{s-T_1} \lambda e^{-\gamma u} du = \frac{\lambda}{\gamma}\left(1 - e^{-\gamma(s-T_1)}\right)$, $r_2(s) = \lambda f(s - T_1)$. Noting that $\dot{r}_1(s) = \lambda - \gamma r_1(s), r_1(T_1) = 0$ we obtain $\dot{p}(s) = -\gamma p(s) - \mu, \quad p(T_1) = \theta$, which is solved by $p(s) = \left(\theta + \frac{\mu}{\gamma}\right) e^{-\gamma(s-T_1)} - \frac{\mu}{\gamma}$. Setting $p(T_2) = 0$ we find the time $T_2 = T_1 + \frac{1}{\gamma}\log\left(1 + \frac{\theta\gamma}{\mu}\right)$ by which all of the initial single-type queue content has left the system. At that time $q(T_2) = r_1(T_2) + r_2(T_2) = \bar{Q}$. For $s \in [T_1, T_2], q(s) \geq q_1(s) = \left(q(T_0) - \frac{\lambda-\mu}{\gamma}\right) e^{-\gamma(s-T_0)} + \frac{\lambda-\mu}{\gamma} > \theta$, so type-2 orders did not cancel and $q_2(T_2) > 0$. We have established R_{T_2}, which implies R_u for $u > T_2$ via Proposition 6. It remains to show that $q(t) > \bar{Q}$ implies $t > T_2$. The function $q(s) = q_1(s) + q_2(s) = q_1(s) + r_2(s)$ is monotonically increasing for $s \in [T_1, T_2]$ and since $q(T_2) = \bar{Q}$, it implies that $q(s) < \bar{Q}$ for $s \in [T_1, T_2)$, therefore $t \geq T_2$. ■

Proposition 8 *If $\mu > 0$ and $\zeta_{1,2}(0, y) < \infty, y \leq \tau(0) < \infty$, then R_t is true for $t \geq \bar{T} \triangleq \frac{q(0)}{\mu} + T_1 + T_2$, where*

$$T_1 = \frac{1}{\gamma}\log\left(1 + \frac{\theta\gamma}{\lambda - \mu - \theta\gamma}\right), \quad T_2 = T_1 + \frac{1}{\gamma}\log\left(1 + \frac{\theta\gamma}{\mu}\right).$$

Moreover $q_i(t)$ are monotinically increasing (decreasing) if $q_1(T) < q_1^*$ ($q_1(T) > q_1^*$), $q_i(t) \to q_i^*, i = 1, 2$ as $t \to \infty$ and the limit point q_1^*, q_2^* is given by

$$\begin{cases} q_1^* = \dfrac{\lambda(1+f) - \mu}{\gamma}, \\ q_2^* = \dfrac{\lambda f}{\gamma}\log\left(\dfrac{\lambda}{\mu - \lambda f}\right) \end{cases}$$

In the limit $\alpha^* = 1 - \frac{\lambda f}{\mu}$ and the head-of-line waiting time $\tau^* = \frac{1}{\gamma}\log\left(\frac{\lambda}{\mu - \lambda f}\right)$.

Proof: If $q(0) < \theta$ then without loss of generality $q_2(0) = 0$ and we simply follow the argument of Proposition 7 with $T_0 = 0$ to show R_t for $t > \bar{T}$. Otherwise if $q(0) > \theta$, we denote by $r_i(t) = q_i(t, t)$, $p_i(t) = q_i(t) - r_i(t)$ and $p(t) = p_1(t) + p_2(t)$. For $t > 0$ we can write $\dot{r}_1(t) = \lambda - \gamma r_1(t), \dot{r}_2(t) = \lambda f$ and therefore $\dot{p}(t) = -\gamma p_1(t) - \mu \leq -\mu$. Since $p(t) = p(0) + \int_0^t \dot{p}(s)ds \leq p(0) - \mu t$, $\exists T_p$ such that $p(T_p) = 0$ and $T_p \leq \frac{q(0)}{\mu}$. We now consider two alternatives.

If $\exists T_q \in [0, T_p]$, such that $q(T_q) < \theta$, all type-2 orders cancel at the time T_q, and queue dynamics for $t > T_q$ are the same as if the queue were initiated with $q_1(T_q) < \theta$ type-1 orders and no type-2 orders. Following the same steps as in Proposition 7, we see that such queue will reach level θ by time $T_q + T_1 \leq \frac{q(0)}{\mu} + T_1$ and establish R_t for $t \geq T_q + T_1 + T_2$, therefore R_t is true for $t \geq \bar{T}$.

If there is no such T_q and $q(t) \geq \theta, t \in [0, T_p]$, then by the time T_p all initial queue content has left the queue, and R_{T_p} is satisfied by construction. Since $T_p \leq \frac{q(0)}{\mu} + T_1 + T_2$, we also have $R_{\bar{T}}$.

To find the limit state of the queue, we use R_T and note that $q_2(t) = \int_0^{\tau(t)} \lambda f ds = \lambda f \tau(t) = \frac{\lambda f}{\gamma} \log\left(\frac{\lambda}{\lambda - \gamma q_1(t)}\right)$ for $t > T$. This is a continuous monotonically increasing function of $q_1(t)$, so $q_2(t)$ converges to a limit as $t \to \infty$, as long as $q_1(t)$ does. Substituting expressions for $\alpha(t), \tau(t)$ into the equation for $\dot{q}_1(t)$, we can write:

$$\dot{q}_1(t) = \lambda - \gamma q_1(t) - \frac{\lambda - \gamma q_1(t)}{\lambda(1+f) - \gamma q_1(t)}\mu \tag{4.5}$$

Monotonicity of $q_1(t)$ for $q_1(t) > q^*$, $q_1(t) < q^*$ follows directly from (4.5) and $q_1(t) < \frac{\lambda}{\gamma}, t \geq T$. Since $q_2(q_1)$ is also monotonically increasing, this proves monotonicity of $q(t)$. Applying a Lyapunov function $V(x) = x^2$ we find that $q_1(t)$ is globally asymptotically stable with a unique stability point q_1^*, which implies convergence of $q_1(t), q_2(t)$ to q_1^*, q_2^* as $t \to \infty$. The expressions for α^*, τ^*, q_2^* are found by direct substitution of q_1^* into corresponding equations.

■

Remark 1: In the steady state all type-2 fluid must be executed, therefore we must have $(1 - \alpha^*)\mu = \lambda f$. Because of this the steady-state amount of type-1 fluid in the two-class system is the same as the steady-state amount of fluid in a single-class system with a lower market order rate $\mu - \lambda f$. This can also be seen by rewriting $q_1^* = \frac{\lambda(1+f)-\mu}{\gamma} = \frac{\lambda-(\mu-\lambda f)}{\gamma}$.

Remark 2: The steady state fluid quantities depend on model parameters in an intuitive way: both $q_{1,2}^*$ and τ^* increase with λ and f, and decrease with γ and μ.

We now turn to studying transient dynamics of the fluid model. Depending on the initial composition of type-1 and type-2 orders in the queue, the model predicts different limit order waiting times. The following Proposition 9 describes a procedure for computing

the virtual waiting time w - the time that a newly submitted infinitely patient limit order would wait until its execution. The initial queue size is denoted by $q(0) = Q$ and is divided in different proportions into type-1 and type-2 orders.

Proposition 9 *If at $t = 0$ all orders in the queue belong to type 1, i.e. $q_1(0) = Q, q_2(0) = 0$ then $w = \frac{1}{\gamma}\log\left(1 + \frac{\gamma Q}{\mu}\right)$. If $q_1(0) = 0, q_2(0) = Q$ and $Q > \theta$, then $w = \frac{Q}{\mu}$.*

Othewise, if the initial queue structure is given by R_0, w is computed by solving the following system of differential equations:

$$\begin{cases} \dot{p}_1(t) = -\gamma p_1(t) - \alpha(t)\mu \\ \dot{p}_2(t) = -(1-\alpha(t))\mu \\ \dot{q}_1(t) = \lambda - \gamma q_1(t) - \dfrac{\lambda - \gamma q_1(t)}{\lambda(1+f) - \gamma q_1(t)}\mu \\ \alpha(t) = \dfrac{\lambda - \gamma q_1(t)}{\lambda(1+f) - \gamma q_1(t)}, \end{cases} \tag{4.6}$$

with initial conditions $p_1(0) = q_1(0), p_2(0) = q_2(0)$ and terminal conditions

$p_1(w) = p_2(w) = 0$.

The initial quantities $q_1(0), q_2(0)$ solve equations

$$\begin{aligned} q_1(0) + \frac{\lambda f}{\gamma}\log\left(\frac{\lambda}{\lambda - \gamma q_1(0)}\right) &= Q \\ q_2(0) &= Q - q_1(0). \end{aligned} \tag{4.7}$$

Moreover, if $q_1(0) < q_1^*$ then $w \in \left[\dfrac{q_2(0)}{(1-\alpha^*)\mu}, \dfrac{q_2(0)}{(1-\alpha(0))\mu}\right]$, otherwise $w \in \left[\dfrac{q_2(0)}{(1-\alpha(0))\mu}, \dfrac{q_2(0)}{(1-\alpha^*)\mu}\right]$.

Proof: Denote by $p_{1,2} = q_{1,2}(\infty, t) - q_{1,2}(t, t)$ the amounts of type-1 and type-2 fluid that was in the system at time 0 and is still in the queue at time t. The two extreme cases with $q_1(0) = Q, q_2(0) = 0$ and $q_2(0) = Q, q_1(0) = 0$ are trivial. In the first case w is a solution of a differential equation $\dot{p}_1(t) = -\gamma p_1(t) - \mu$ with initial condition $p_1(0) = Q$ and terminal condition $p_1(w) = 0$. In the second case the equation is $\dot{p}_2(t) = -\mu$ with $p_2(0) = Q, p_2(w) = 0$.

If the initial queue structure is given by R_0, $q_2(0) = \lambda f \tau(0)$ and with (4.4) it can be further expressed as a function of $q_1(0)$, which, together with the initial condition $q_1(0) +$

$q_2(0) = Q$ leads to equation 4.7. The equations (4.6) follow from computing the derivatives of $p_{1,2}$ and finding the time w such that $p_1(w) = p_2(w) = 0$. Although they reach zero at the same time it is easier to bound w through type-2 fluid dynamics. Depending on whether $q_1(0) < q_1^*$ or $q_1(0) > q_1^*$, Proposition 8 implies that $q_1(t)$ is monotonically increasing or decreasing, therefore $(1-\alpha(t))$ is either monotonically increasing or decreasing in these two cases. Therefore, it is possible to bound the time when $p_2(t)$ reaches zero by bounding its rate of decrease. In the case when $q_1(0) < q_1^*$, $(1-\alpha(0)) \leq (1-\alpha(t)) \leq (1-\alpha^*)$, therefore $\frac{q_2(0)}{(1-\alpha^*)\mu} \leq w \leq \frac{q_2(0)}{(1-\alpha(0))\mu}$, and vice versa if $q_1(0) > q_1^*$. ■

4.4 Empirical results

We rely on an extensive proprietary dataset of limit order executions to test the results of Section 4.3. Specifically, our aim is to compare queueing delays experienced by limit orders in practice with theoretical predictions of Proposition 9. Our results show that the queue structure significantly affects limit order delays. It is unrealistic to assume that the queue consists only of type-1 or type-2 orders, because this leads to biased delay estimates that under- or overestimate realized delays by a factor of 10. Assuming a mix of order types in a queue leads to more accurate delay predictions that lie between the two extremes.

The dataset for this study was collected by an electronic broker dealer firm and consists of 327,505 orders that were sent by firm's algorithms to various U.S. equity exchanges between 03/01/2012 and 04/30/2012. Each entry describes a single limit order: its submission date and time (up to a millisecond), its destination exchange, a stock symbol, an order direction (buy or sell), a trade execution strategy that generated the order (e.g. VWAP), the order outcome (execution or cancelation) and the order *waiting time* that elapsed between its submission and its execution or cancelation.

Because of a large variation in orders submission times, destination exchanges and stock symbols in our data we chose to estimate fluid model parameters $(\lambda_i, \gamma_i, \mu_i, f_i, Q_i)$ separately for each order $i = 1, \ldots, N$ using trades and quotes that happened before its submission. For order i, we consider a 3 minute time interval before its timestamp (henceforth the i-th time interval) and retrieve trades and quotes at the i-th destination exchange for the i-th

interval from the NYSE TAQ database[5]. The market order rate μ_i (in shares per second) was estimated as the total volume of trades in the i-th time interval divided by 360. The initial queue size Q_i was set to the size of the last bid or ask quote in the i-th interval, depending on the i-th order direction. We also computed a time-weighted average $\bar{Q}_i$ of bid or ask quote sizes during the i-th interval.

TAQ data does not contain neither trader identifiers nor order "types", but parameters λ_i, γ_i and f_i can still be approximately inferred by using structural assumptions of our model. In the stochastic model, Type-2 orders arrive instantaneously after a type-1 order arrival, suggesting that order types can be inferred from the lengths of their inter-arrival periods. In practice, responding to quote updates takes some time even for the fastest traders, for example, the analysis in [57] suggests a lower bound of 2-3 ms. The total number of limit order arrivals in the i-th interval Λ_i is divided into Λ_i^1, the number of new orders whose inter-arrival times were shorter than 20 ms, and $\Lambda_i^2 = \Lambda_i - \Lambda_i^1$. The fraction of "fast" order arrivals Λ_i^1/Λ_i during the i-th interval measures the relative activity of high-frequency traders (i.e. type-2 orders), but it needs to be corrected for the expected number of type-1 orders whose inter-arrivals were short due to randomness. To make this correction we assume in correspondence with the stochastic model that type-1 orders arrive according to a Poisson process and therefore set $\phi_i = \frac{1}{\Lambda_i}\left(\Lambda_i^1 - \left(e^{20/M_i^\Lambda} - 1\right)\Lambda_i^2\right)$, where M_i^Λ is the mean inter-arrival time (in ms) of orders counted in Λ_i^2. Then we calculate the fluid model parameters $f_i = \frac{\phi_i}{1-\phi_i}, \lambda_i = \frac{\Lambda_i}{180(1+f_i)}$.

The parameter γ_i is estimated in a similar way. First we decompose Ω_i - the total number of order departures during the i-th interval - into Ω_i^1 and Ω_i^2 that are correspondingly the number of orders whose inter-departure times were shorter or longer than 20 ms. Then we estimate the total volume of type-1 order cancelations in the i-th interval as $\Gamma_i^1 = \left(e^{20/M_i^\Omega}\Omega_i^2\right)\Gamma_i/\Omega_i$, where M_i^Ω is the mean inter-departure time for orders counted in Ω_i^2 and Γ_i is the total cancelation volume in the i-th interval. Using the steady-state equations (4.5) we can then recover $\gamma_i = \left(\Gamma_i^1 - \lambda_i f_i \log\left(1 - \frac{\Gamma_i^1}{\lambda}\right)\right)/\bar{Q}_i$.

The descriptive statistics and average parameter values for our data are given in Table

[5]For 18% of orders there were no trades or quotes during that interval at their destination exchange. Instead, for these orders we used NBBO quotes and consolidated trades.

4.5, where order dataset is divided into 25 equal-sized subsamples by quintiles of μ_i and $\bar{Q}_i$. There are significant differences in order waiting times across subsamples and we can also note a correlation between average bid-ask spreads and average quote sizes. We find that the bottom group consists of orders for ultra-liquid stocks that have large queue sizes and large tick sizes relative to their prices. Average values of f_i show little variation across subsamples and imply that on average about 40% of orders are of type-2. The values of γ_i decrease with queue size and increase with trading volume, suggesting that type-1 orders are more patient in markets with larger average queues and smaller average trading volumes.

Order data is distributed unevenly among 487 different stocks, 13 exchanges and alternative trading venues, 7 execution strategies and 42 days. To ensure data consistency and at least a moderate number of data points in each group of orders we applied multiple data filters[6]. The filtered dataset contains 109,938 orders for 268 stocks submitted to NASDAQ (48%), NYSE (24%), ARCA (13%), BATS Z (10%) and EDGE X (5%) over the course of 21 days in March and April 2012. As illustrated on Figure 4.2, the top 10 symbols by a number of orders[7] contain more than 30% of observations, with the next 40 symbols contributing another 30%. Orders are submitted throughout the entire trading session and our sample has 5,000 to 20,000 orders in each half-hour time interval between 9:30am and 4:00pm. Most orders are submitted by VWAP strategies (59%), followed by TWAP (31%) and POV (9%). Only 35% of all orders are executed, and the rest is canceled. VWAP and TWAP strategies generate orders with execution rates of 38% and 32% respectively. POV strategies are supposedly less patient with their orders and only 26% of them are executed.

Most orders in our data are canceled (56-85% depending on a subsample) which has important implications for the analysis of limit order execution delays. An order is canceled

[6]Specifically, we filtered out: orders placed behind NBBO prices (deep in the order book) and aggressive orders; trade execution strategies other than TWAP, VWAP and POV; orders sent to alternative trading venues, inverse exchanges and the NASDAQ PSX exchange (due to different order execution mechanics there); orders sent to the National Stock Exchange (it had only 88 orders); trading dates with less than 1000 orders; orders sent during the first five and the last five minutes of trading; orders submitted to markets with wide spreads (the NBBO spread $\geq$ 50 basis points at an order submission time); orders submitted to markets with small queues (average NBB or NBO quote size $<$ 1000 shares during 3 minutes before an order submission); symbols with less than 200 orders in the full dataset or less than 50 orders in the filtered dataset; orders whose lifetime at an exchange includes an anomalous burst of trading volume (defined as a 1-second time interval in the top 5% of 1-second intervals by volume for that stock, date and exchange).

[7]These stocks are MU, GLW, USB, TXN, CVS, ALTR, BAC, GT, COP, LOW.

when its waiting time in a queue exceeds its patience deadline or when the cumulative trading volume since its submission exceeds a threshold value. Orders are also canceled and resubmitted when prices move away from them. Finally, since orders are sent to multiple exchanges at the same time, an order can be canceled when its substitute order at a different trading venue is filled. Regardless of cancelation motives, a hypothetical execution delay of a canceled order (had it remained in its queue) is longer than its actual waiting time until cancelation. In other words, a waiting time of an order until cancelation is a truncated observation of its queueing delay until execution. Ex-post, orders with long execution delays are canceled more frequently than orders with short delays, which can be observed in Table 4.5. Because of this selection bias, it would be incorrect to directly compare a theoretical execution delay estimate for an infinitely patient order with a realized waiting time of a canceled order. To avoid the bias we use four different methods to transform (un-censor) waiting time observations of canceled orders.

First, we fit two parametric models - an exponential and a Pareto distribution to each of 25 order subsamples defined by μ_i and $\bar{Q}_i$. These two distributions are convenient for modeling censored execution delay data because they have non-negative support and are memoryless[8]. The parameters of exponential or Pareto distribution can be directly estimated from censored waiting time data D_i, $i = 1, \ldots, N$, as a fixed-point solution of a corresponding E-M scheme. For an exponential distribution, we estimate for each subsample g a mean parameter $\hat{\nu}_g = \frac{1}{E(g)} \sum_{i=I(g-1)+1}^{I(g)} D_i$, where $E(g)$ is the number of executed orders in subsample g and $I(g)$ is the index of the last observation in subsample g with $I(0) = 0$. For the Pareto distribution, we estimate its scale parameter $\hat{m}_g = \min_{I(g-1)+1<i\leq I(g)} \{D_i\}$ and its tail index $\hat{\beta}_g$ as the positive solution of a quadratic equation $\left[\sum_{i=I(g-1)+1}^{I(g)} \log\left(\frac{\hat{m}_g}{D_i}\right)\right] \beta^2 + (I(g) - I(g-1))\beta + E(g) = 0$. After estimating these distribution parameters separately for each group we use the memorylessness property again and add a random variable drawn from a corresponding distribution to the waiting time of each canceled order.

The other two un-censoring methods are based on a concept of maximum entropy (see

[8]If a delay D has an exponential distribution with a mean μ, then $\mathbb{P}(D > s+t|D > t; \mu) = \mathbb{P}(D > s; \mu)$. If it has a Pareto distribution with a scale parameter m and tail parameter α, then $\mathbb{P}(D > s+t|D > t; m, \alpha) = \mathbb{P}(D > s; t, \alpha)$.

e.g. [32]). The first method assumes a maximum entropy (uniform) distribution for the length of time that each canceled order had to wait until execution, in excess of the time it already spent in a queue. For each such order, this procedure replaces its waiting time with a random variable, uniformly distributed between its waiting time until cancelation and an observation-specific ceiling value Q_i/μ_i. The second method finds a maximum entropy distribution for the entire dataset, subject to fractile constraints of the form $\mathbb{P}(D_i > x) = y(x)$ imposed by the censored data sample. We refer to [38] for a formulation of the maximum entropy distribution problem with fractile constraints. In short, this approach assigns higher values to censored observations so that the resulting distribution of execution delays satisfies fractile constraints and has the highest possible entropy. The maximum entropy distribution has the least possible structure as measured by its descriptive complexity. Having solved for the maximum entropy distribution, we then re-allocate censored observations according to the solution, in addition restricting each un-censored point to be smaller than Q_i/μ_i. To prevent the power and maximum entropy distributions from assigning arbitrarily large values to truncated observations we also impose a constraint that all execution delays are smaller than 30 minutes. Histograms of the resulting execution delay distributions for the entire sample are shown on Figure 4.5. We note substantial differences between some of these histograms - the choice of an un-censoring methodology has a significant effect on the overall distribution of data because a large share of data is censored. For this reason, we view the output of four un-censoring procedures as four different samples of data and test our model on each sample.

For each order i we numerically solve equations 4.6 and compute a delay forecast w_i that assumes a mixture of type-1 and type-2 orders in the initial queue. We also compute forecasts $w_i^0 = \frac{Q_i}{\mu_i}$ that assume no cancelations from initial queues in front of our order, and $w_i^1 = \frac{1}{\gamma_i^1} \log\left(1 + \frac{\gamma_i^1 Q_i}{\mu_i}\right)$ that assume all orders in front of ours to be of type-1. We divide the estimates $(w_i^0, w_i^1, w_i)_{i=1}^N$ into equal-sized bins sorted by un-censored execution delays D_i^u, and compute for each bin the average values of D_i^u, w_i^0, w_i^1, w_i. If estimates were correct, their averages should lie on a 45^o line when plotted against average realized delays. These plots together with 95% confidence intervals for average values are presented on Figures 4.5-4.5. We can observe that for delays D_i^u of 0-10 seconds all three forecasts systematically

overestimate realized delays. A possible explanation is that short delays are experienced when Q_i is small relative to μ_i, i.e. when the volume of market orders is larger than the queue size. This case cannot be reliably described by a fluid model which assumes that on the contrary the queue size is large compared to order sizes, making fluid model-based forecasts inadequate for short realized delays.

As realized delays become longer, single-type forecasts w_i^1 seem to asymptote at around 60 seconds, even as the average delay D_i^u extends to several minutes. Estimates w_i^1 assume that all orders in the initial queue have a finite patience (belong to type-1) so even if the initial queue Q_i is large, most of them are expected to cancel, and therefore their presence does not significantly increase the forecast w_i^1. Mathematically, this can be seen from the expression for $w_i^1(Q_i) \sim O(\log(Q_i))$. On another hand, assuming that all orders in the initial queue belong to type-2 (i.e. never cancel) leads to extremely long delay forecasts. After all, some orders will cancel and realized delays are smaller than forecasts w_i^0.

For delays longer than 60 seconds, we see that two-type forecasts w_i are closer to realized delays, suggesting that a mixture of type-1 and type-2 orders better describes queue dynamics in this range. Although none of three estimates gives a perfect fit, the comparison seems to favor a model with a mixture of order types. We note that the aim of this computation was to compare single-type models with a two-type model *without* fitting any of them to the realized delay data. If we were to calibrate parameters $(\lambda_i, \gamma_i, \mu_i, f_i)$ to realized delays D_i^u we would a priori expect better performance from the model with heterogeneous orders because it has more parameters than its single-type counterparts.

4.5 Conclusion

We propose a multiclass queueing model that describes the evolution of bid and ask quotes in a limit order market. The model distinguishes between orders of fund managers that have finite patience and orders of high-frequency traders that are driven by changes in the order book state. Our model structurally explains long waiting times experienced by limit orders in large queues and bursts of cancelations in small queues. It leads to qualitatively different waiting time estimates that are closer to empirical measurements, and sheds light on the

composition of order queues in limit order markets. Our analysis is one of the first steps in the direction of reconciling naive "zero-intelligence" stochastic models of limit order books with more detailed economic models from the market microstructure literature, and it also motivates a new class of queueing models for studying service systems for heterogeneous populations with state-dependent abandonments.

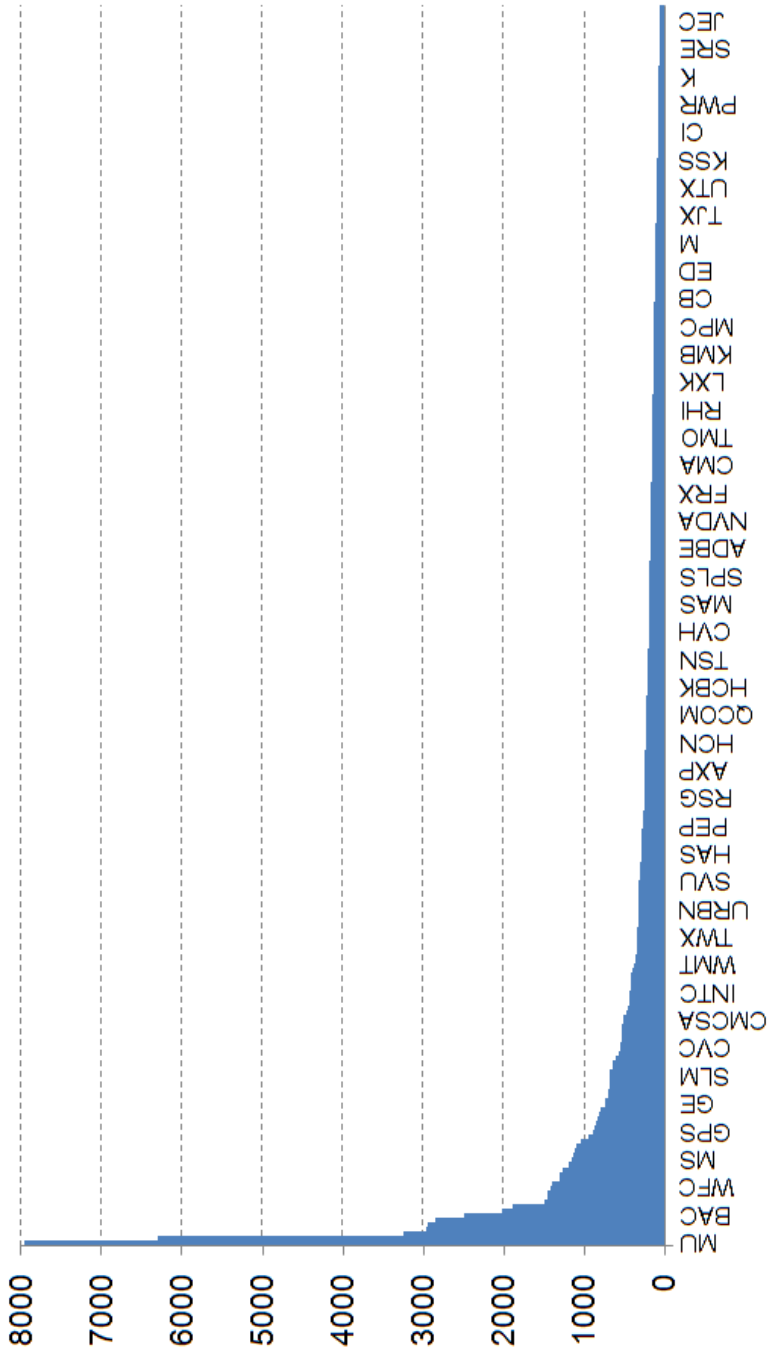

Figure 4.2: Total number of orders per symbol in filtered order data.

Table 4.1: Descriptive statistics for order data

Subsample	Average parameter values					Mean waiting	% of executed	Mean bid-ask	Number
number	Q_i	μ_i	λ_i	γ_i	f_i	time (sec)	orders	spread (bp)	of stocks
1	342	2	22	0.156	1.171	32	34%	4.16	235
2	366	4	31	0.281	1.723	28	37%	3.99	238
3	396	8	57	0.372	1.144	26	38%	4.08	238
4	421	15	77	0.498	1.269	26	36%	4.01	235
5	453	38	127	0.741	1.351	27	38%	4.21	206
6	942	4	50	0.134	1.241	31	34%	4.25	238
7	954	10	78	0.228	1.393	27	36%	4.03	244
8	977	18	120	0.312	1.335	28	38%	4.04	235
9	1000	29	158	0.388	1.136	26	39%	4.17	229
10	1038	70	275	0.665	1.334	24	42%	3.87	210
11	2163	6	98	0.114	1.443	35	31%	4.29	237
12	2197	16	143	0.190	1.755	31	38%	4.13	223
13	2219	28	172	0.280	1.811	28	38%	4.06	225
14	2320	48	296	0.302	1.230	26	40%	3.97	226
15	2391	126	600	0.678	1.298	22	43%	3.85	209
16	6199	6	133	0.064	1.459	44	25%	5.39	214
17	5797	22	240	0.109	1.264	35	31%	5.07	207
18	5772	44	353	0.156	1.171	33	35%	4.82	215
19	6015	78	466	0.233	1.423	30	35%	5.16	204
20	6472	187	986	0.450	1.504	25	41%	4.43	182
21	49285	3	339	0.021	1.325	59	17%	8.20	149
22	41533	20	509	0.038	1.289	47	18%	7.94	172
23	45579	57	984	0.066	1.520	44	24%	7.59	178
24	46354	134	1523	0.130	1.512	38	29%	6.73	171
25	112279	736	5033	0.186	1.437	36	33%	7.64	152

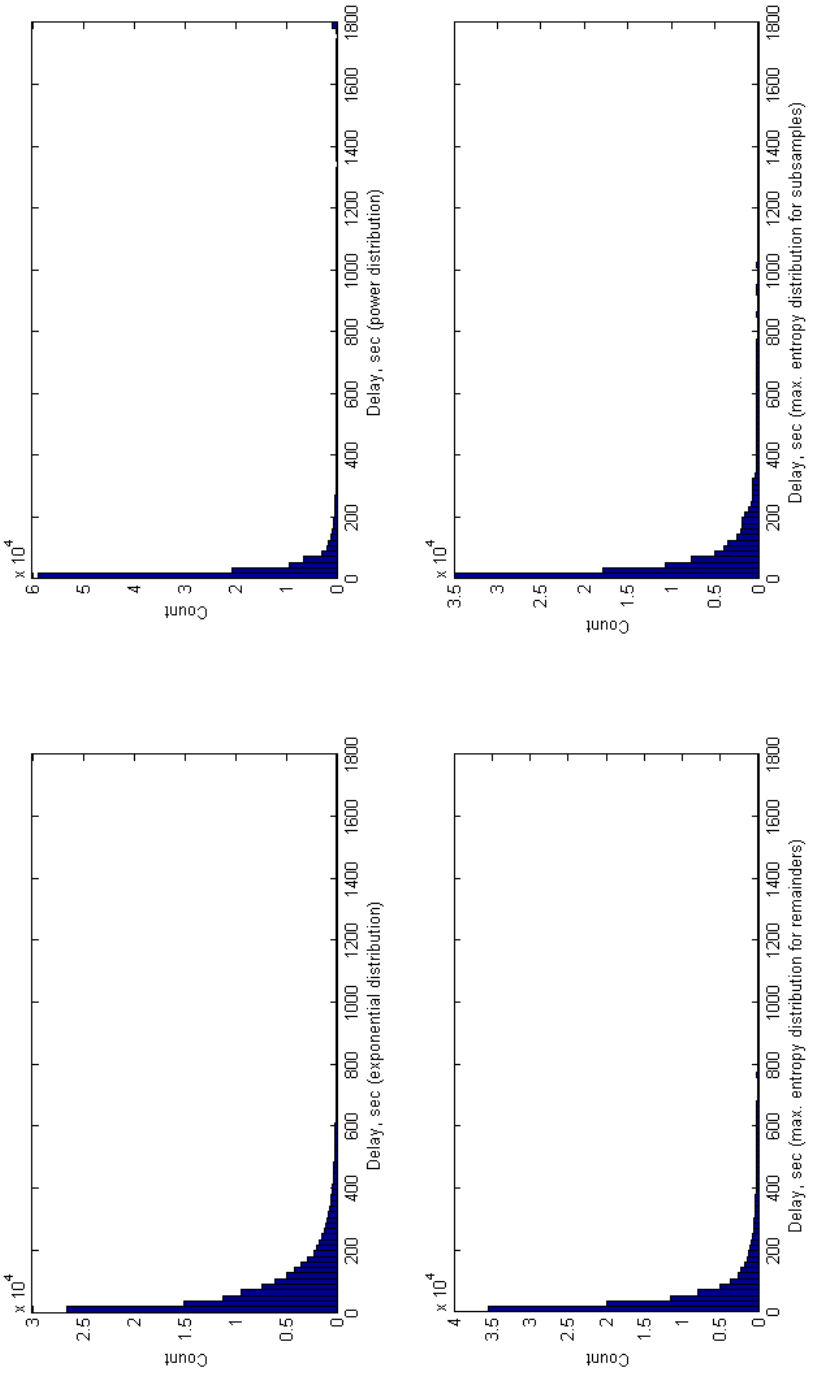

Figure 4.3: Histograms of limit order execution delays for the entire sample.

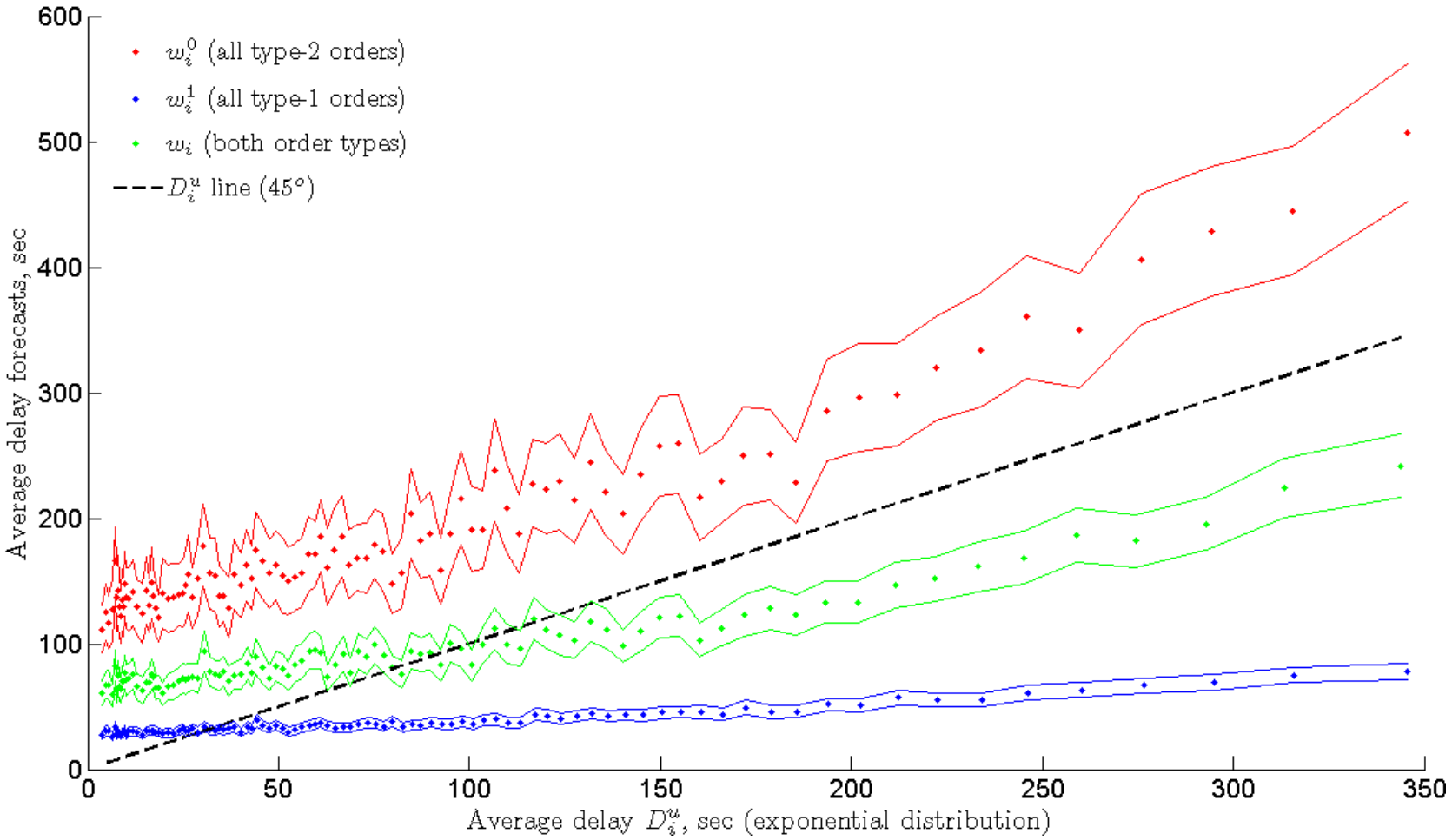

Figure 4.4: Average delay estimates and forecasts, based on the exponential distribution uncensoring.

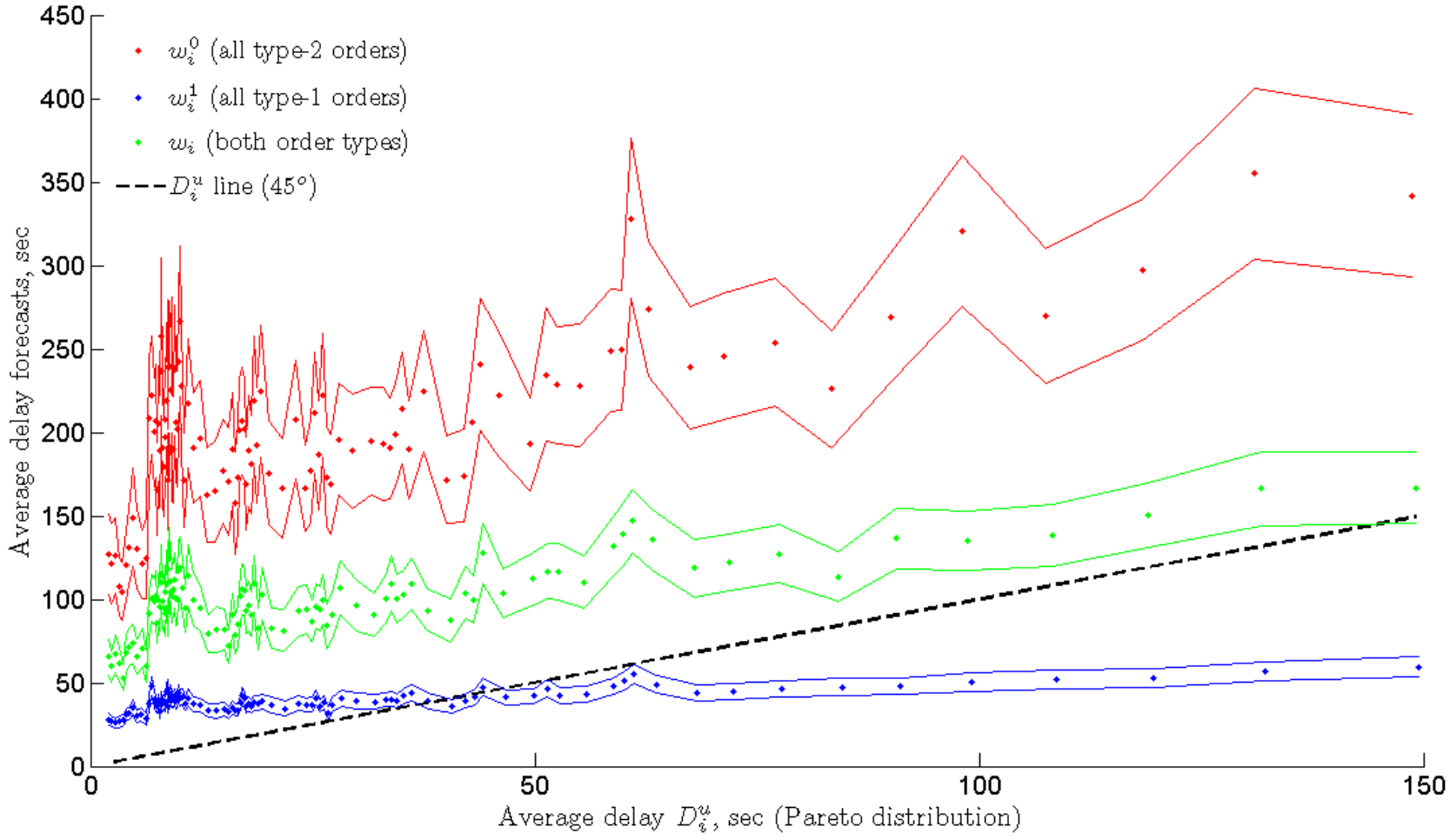

Figure 4.5: Average delay estimates and forecasts, based on the power distribution uncensoring.

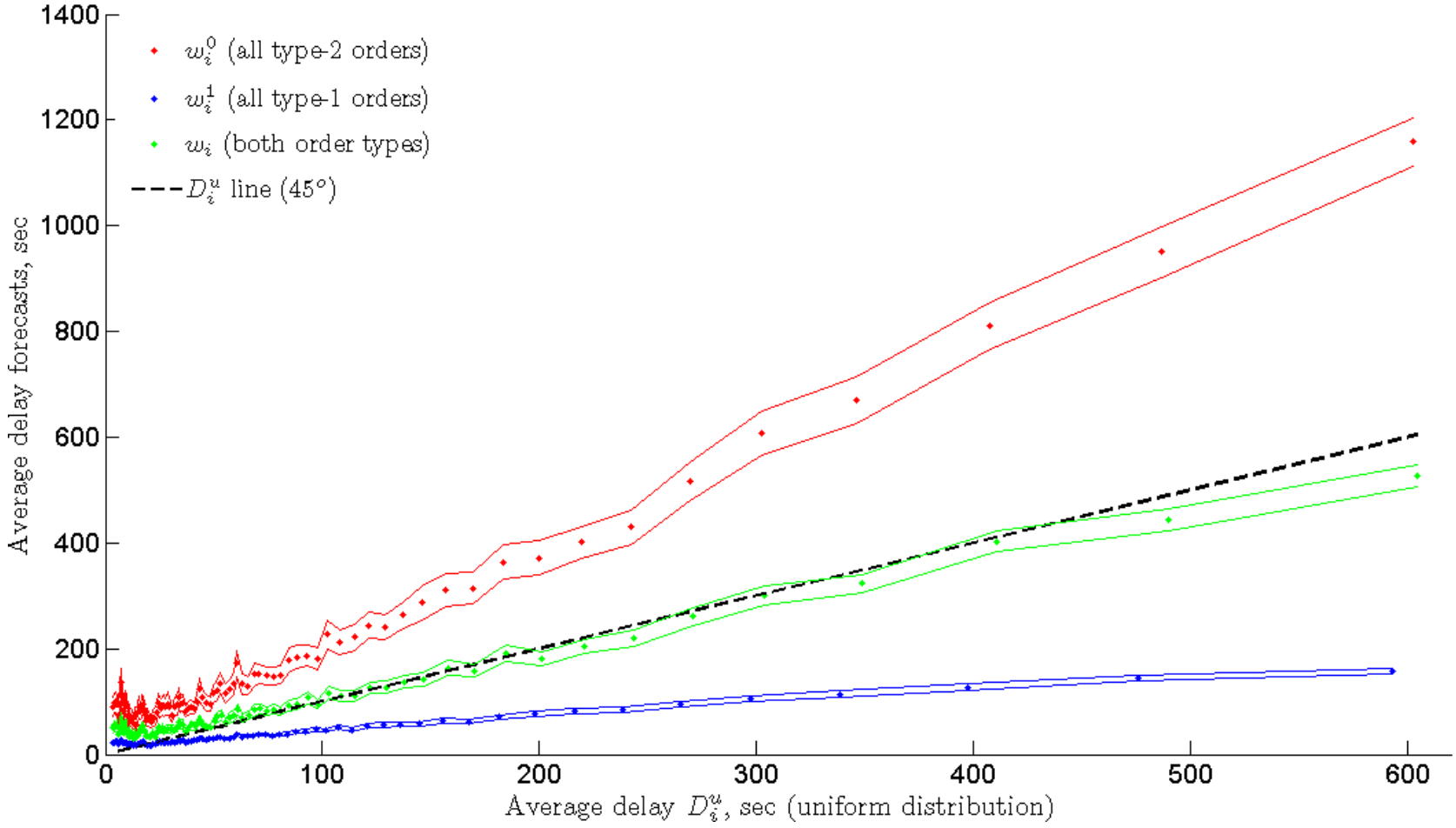

Figure 4.6: Average delay estimates and forecasts, based on the uniform distribution uncensoring.

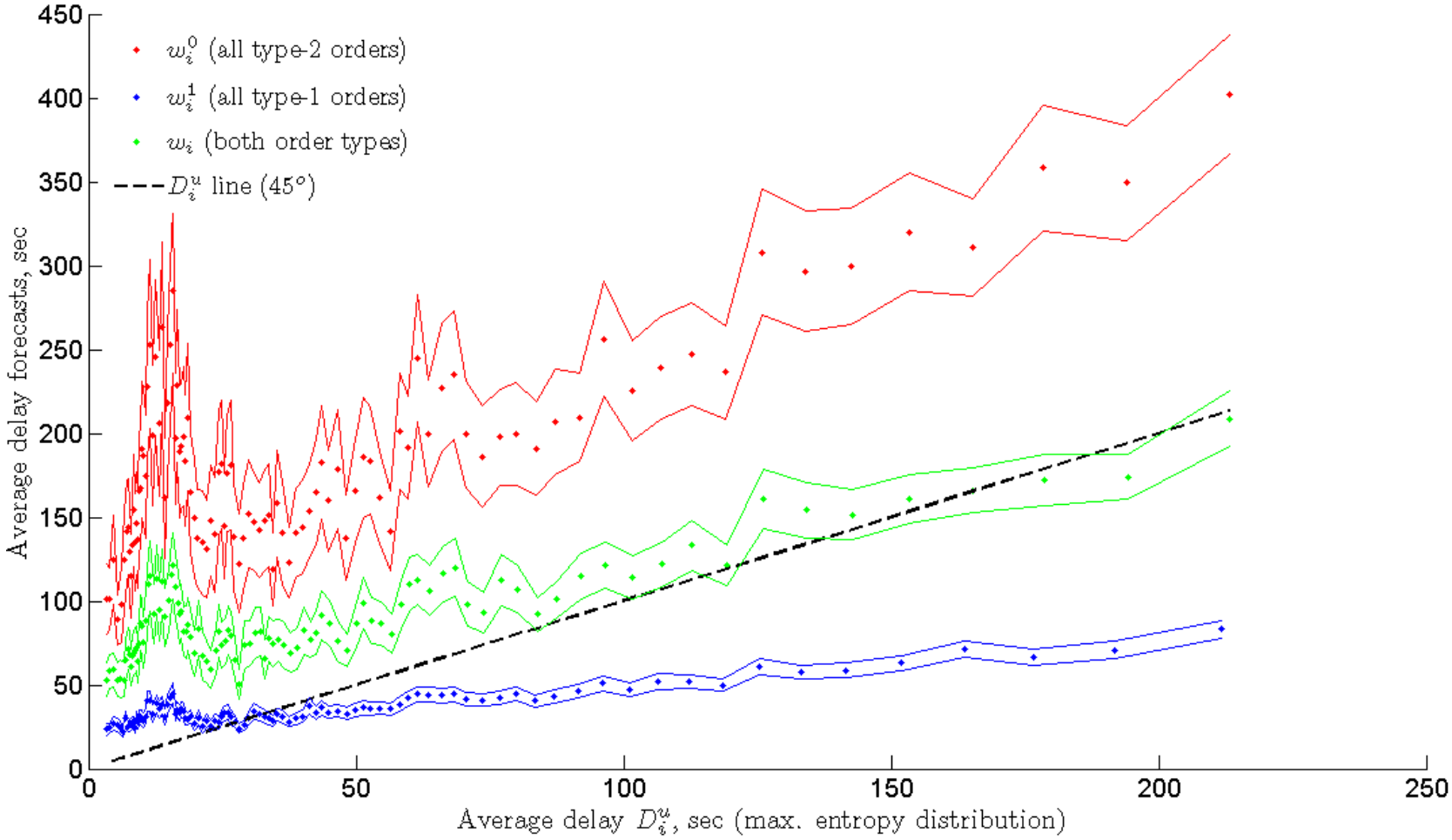

Figure 4.7: Average delay estimates and forecasts, based on the maximum entropy uncensoring.

Bibliography

[1] L. Adamic, C. Brunetti, J. Harris, and A. Kirilenko, *Trading networks*. 2010. 14

[2] H. J. Ahn, K. H. Bae, and K. L. Chan, *Limit orders, depth, and volatility: evidence from the stock exchange of Hong Kong*, Journal of Finance, 56 (2001), pp. 767–788. 24, 25, 46, 47

[3] A. Alfonsi, A. Fruth, and A. Schied, *Optimal execution strategies in limit order books with general shape functions*, Quantitative Finance, 10 (2010), pp. 143–157. 12, 64

[4] A. Alfonsi, A. Schied, and A. Slynko, *Order book resilience, price manipulation, and the positive portfolio problem*, SIAM Journal on Financial Mathematics, 3 (2012), pp. 511–533. 11

[5] C. Aliprantis and O. Burkinshaw, *Principles of Real Analysis*, Academic Press, third ed., 1998. 76

[6] R. Almgren and N. Chriss, *Optimal execution of portfolio transactions*, Journal of Risk, 3 (2000), pp. 5–39. 11, 22, 63, 64, 66

[7] R. Almgren, C. Thum, E. Hauptmann, and H. Li, *Direct Estimation of Equity Market Impact*, Risk, 18 (2005), p. 57. 9, 22

[8] Y. Amihud, H. Mendelson, and L. H. Pedersen, *Liquidity and Asset Prices*, Founations and Trends in Finance, 1 (2006), pp. 269–364. 26, 48

[9] T. Andersen and T. Bollerslev, *Deutsche mark - dollar volatility: intraday activity patterns, macroeconomic announcements, and longer run dependencies*, Journal of Finance, 53 (1998), p. 219. 25, 47

[10] J. Angel, L. Harris, and C. Spatt, *Equity trading in the 21st century*, The Quarterly Journal of Finance, (2011). 2

[11] M. Avellaneda, J. Reed, and S. Stoikov, *Forecasting prices from Level-I quotes in the presence of hidden liquidity*, Algorithmic Finance, 1 (2011), pp. 35–43. 30, 96, 99

[12] M. Avellaneda and S. Stoikov, *High-frequency trading in a limit order book*, Quantitative Finance, 8 (2008), pp. 217–224. 13

[13] M. Baron, J. Brogaard, and A. Kirilenko, *The trading profits of high frequency traders*. 2012. 14, 96

[14] E. Bayraktar and M. Ludkovski, *Optimal Trade Execution in Illiquid Markets*, Mathematical Finance, 21 (2011), pp. 681–701. 12, 24, 64, 66

[15] D. Bertsimas and A. W. Lo, *Optimal control of execution costs*, Journal of Financial Markets, 1 (1998), pp. 1–50. 10, 22, 63, 64, 66

[16] J. Bikker, L. Spierdijk, and P. Vandersluis, *Market impact costs of institutional equity trades*, Journal of International Money and Finance, 26 (2007), pp. 974–1000. 9

[17] E. Boehmer and R. Jennings, *Public disclosure and private decisions: Equity market execution quality and order routing*, Review of Financial Studies, 20 (2007), pp. 315–358. 64

[18] J.-p. Bouchaud, *Price Impact*, in Encyclopedia of Quantitative Finance, Wiley, 2010. 22

[19] J.-P. Bouchaud, Y. Gefen, M. Potters, and M. Wyart, *Fluctuations and response in financial markets: the subtle nature of random price changes*, Quantitative Finance, 4 (2004), pp. 176–190. 22

[20] Y. Brandes, I. Domowitz, B. Jiu, and H. Yegerman, *Algorithms , trading costs and order size*, Trade, (2007). 9, 12

[21] B. Brooksley, J. Brennan, R. Engle, R. Ketchum, M. OHara, S. Philips, D. Ruder, and J. Stiglitz, *Recommendations regarding regulatory responses to the market events of may 6, 2010: Summary report of the joint CFTC-SEC advisory committee on emerging regulatory issues*. 2011. 3

[22] K. L. Chan and W.-M. Fong, *Trade Size, Order Imbalance, and the Volatility-Volume Relation*, Journal of Financial Economics, 57 (2000), pp. 247–273. 49

[23] T. Chordia, R. Roll, and A. Subrahmanyam, *Liquidity and market efficiency*, Journal of Financial Economics, 87 (2008), p. 249. 2, 37

[24] W. G. Christie and P. H. Schultz, *Why do NASDAQ market makers avoid odd-eighth quotes?*, The Journal of Finance, 49 (1994), pp. pp. 1813–1840. 2

[25] R. Cont, *Statistical modeling of high-frequency financial data*, IEEE Signal Processing, 28 (2011), pp. 16–25. 64

[26] R. Cont and A. De Larrard, *Price dynamics in a Markovian limit order market*, SIAM Journal on Financial Mathematics, 4 (2013), pp. 1–25. 64

[27] R. Cont and A. Kukanov, *Optimal order placement in limit order markets*. 2012. 63

[28] R. Cont, A. Kukanov, and S. Stoikov, *Price impact of order book events*, Journal of Financial Econometrics, (Forthcoming). 21

[29] R. Cont and A. D. Larrard, *Order book dynamics in liquid markets : limit theorems and diffusion approximations*. 2011. 14, 96, 99

[30] ——, *Price dynamics in limit order markets: linking volatility with order flow*. 2011. 14, 96

[31] R. Cont, S. Stoikov, and R. Talreja, *A stochastic model for order book dynamics*, Operations Research, 58 (2008), pp. 549–563. 13, 14, 96

[32] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley-Interscience, 2006. 111

[33] H. Demsetz, *The cost of transacting*, The Quarterly Journal of Economics, 82 (1968), pp. 33–53. 7

[34] D. Easley and M. O'Hara, *Price, trade size, and information in securities markets*, Journal of Financial Economics, 19 (1987), pp. 69–90. 8, 96

[35] Z. Eisler, J.-P. Bouchaud, and J. Kockelkoren, *The price impact of order book events: market orders, limit orders and cancellations*, Quantitative Finance, 12 (2012), pp. 1395–1419. 10, 23, 24, 31

[36] P. Embrechts, C. Kluppelberg, and T. Mikosch, *Modelling extremal events for insurance and finance*, Springer, 1997. 50

[37] R. Engle and A. Lunde, *Trades and quotes: a bivariate point process*, Journal of Financial Econometrics, 1 (2003), pp. 159–188. 22, 23

[38] S. Eren and C. Maglaras, *A maximum entropy joint demand estimation and capacity control policy.* 2009. 111

[39] M. Evans and R. Lyons, *Order flow and exchange rate dynamics*, Journal of Political Economy, 110 (2002), p. 170. 10, 22

[40] J. D. Farmer, L. Gillemot, F. Lillo, S. Mike, and A. Sen, *What really causes large price changes?*, Quantitative Finance, 4 (2004), pp. 383–397. 23

[41] J. D. Farmer, P. Patelli, and I. Zovko, *The predictive power of zero intelligence in financial markets.*, Proceedings of the National Academy of Sciences of the United States of America, 102 (2005), pp. 2254–9. 8, 13, 96

[42] E. Fishler, *The difficulty of trading "ultra-liquid" stocks.* 2012. 10, 97

[43] E. Flitter and S. Lynch, *Insight: Chicago fed warned on high-frequency trading, SEC slow to respond*, Reuters, (2012). 3

[44] K. Florian, A. Schied, and Y. Sun, *Price manipulation in a market impact model with dark pool*. 2012. 11

[45] T. Foucault, O. Kadan, and E. Kandel, *Limit Order Book as a Market for Liquidity*, Review of Financial Studies, 18 (2005), pp. 1171–1217. 96

[46] T. Foucault and A. Menkveld, *Competition for Order Flow and Smart Order Routing Systems*, Journal of Finance, 63 (2008), p. 112. 64

[47] P. Fraenkel, *Dynamic portfolio management using optimal control with quadratic costs*. 2009. 11

[48] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. Stanley, *A theory of power-law distributions in financial market fluctuations*, Nature, 423 (2003), p. 267. 10, 22, 31

[49] K. Ganchev, Y. Nevmyvaka, and M. Kearns, *Censored exploration and the dark pool problem*, Communications of the ACM, 53 (2010), p. 99. 13, 65

[50] J. Gatheral, *No-Dynamic-Arbitrage and Market Impact*, Quantitative Finance, 10 (2010), p. 749. 11, 22

[51] R. Goettler, C. Parlour, and U. Rajan, *Equilibrium in a Dynamic Limit Order Market*, Journal of Finance, 60 (2005), p. 2149. 8

[52] O. Gueant and C. Lehalle, *General Intensity Shapes in Optimal Liquidation*. 2012. 64

[53] F. Guilbaud and H. Pham, *Optimal high frequency trading in a pro-rata microstructure with predictive information*. 2012. 12, 13, 64

[54] J. Hasbrouck, *Measuring the information content of stock trades*, Journal of Finance, 46 (1991), pp. 179–207. 10, 22, 24, 31, 37

[55] ——, *The summary informativeness of stock trades: an econometric analysis*, Review of Financial Studies, 4 (1991), pp. 571–595. 24, 25, 47, 48

[56] ——, *The Best Bid and Offer: A Short Note on Programs and Practices.* 2010. 30, 33

[57] J. Hasbrouck and G. Saar, *Low-latency trading.* 2010. 10, 15, 62, 95, 96, 108

[58] J. Hasbrouck and D. Seppi, *Common factors in prices, order flows and liquidity*, Journal of Finance and Economics, 59 (2001), p. 383. 23, 31, 37

[59] N. Hautsch and R. Huang, *The market impact of a limit order*, Journal of Economic Dynamics and Control, 36 (2012), pp. 501–522. 10, 23, 24

[60] T. Hendershott, C. Jones, and A. Menkveld, *Does Algorithmic Trading Improve Liquidity?*, Journal of Finance, 66 (2011), pp. 1–33. 4, 94

[61] C. Hopman, *Do supply and demand drive stock prices?*, Quantitative Finance, 7 (2007), pp. 37–53. 10, 23, 37

[62] H. Huang and A. Kercheval, *A generalized birth-death stochastic model for high-frequency order book dynamics*, Quantitative Finance, 12 (2012), pp. 547–557. 14, 96

[63] G. Huberman and W. Stanzl, *Price manipulation and quasi-arbitrage*, Econometrica, 72 (2004), pp. 1247–1275. 11, 22

[64] G. Huberman and W. Stanzl, *Optimal liquidity trading*, Review of Finance, 9 (2005), pp. 165–200. 66

[65] R. Huitema, *Optimal Portfolio Execution using Market and Limit Orders.* 2012. 12, 64

[66] P. Jain, *Financial market design and the equity premium: Electronic versus floor trading*, The Journal of Finance, LX (2005). 5, 95

[67] K. Janecek and M. Kabrhel, *Matching algorithms of international exchanges.* 2007. 5, 95

[68] O. B. Jennings and J. E. Reed, *An overloaded multiclass FIFO queue with abandonments*, Operations Research, 60 (2012), pp. 1282–1295. 15, 97

[69] C. Jones, *A century of stock market liquidity and trading costs.* 2002. 2, 94

[70] C. Jones, G. Kaul, and M. Lipson, *Transactions, volume, and volatility*, Review of Financial Studies, 7 (1994), pp. 631–651. 49, 52

[71] T. Joyce, J. Nazarali, A. Levitt, F. Tomczyk, L. Harris, C. Spatt, D. Mathisson, G. Katz, P. Stevens, E. Sirri, J. Bethel, D. Niederauer, J. Giesea, and B. Mock, *Current Perspectives on Modern Equity Markets: a Collection of Essays by Financial Industry Experts*, Mighty Media, Inc, 2010. 2

[72] J. Karpoff, *The relation between price changes and trading volume: A survey*, Journal of Financial and Quantitative Analysis, 22 (1987), p. 109. 24, 26, 49

[73] D. Keim and A. Madhavan, *The upstairs market for large-block transactions: Analysis and measurement of price effects*, Review of Economic Studies, 9 (1996), p. 1. 8, 10, 22

[74] A. Kempf and O. Korn, *Market depth and order size*, Journal of Financial Markets, 2 (1999), p. 29. 10, 22, 23, 31, 37

[75] A. Kirilenko, A. S. Kyle, M. Samadi, and T. Tuzun, *The Flash Crash: The Impact of High Frequency Trading on an Electronic Market.* 2011. 3, 14, 96

[76] P. Knez and M. Ready, *Estimating the profits from trading strategies*, Review of Financial Studies, 9 (1996), p. 1121. 23

[77] A. Kukanov and C. Maglaras, *A limit order queue model with heterogenous traders.* 2013. 94

[78] H. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer, New York, 2003. 84

[79] A. S. Kyle, *Continuous auctions and insider trading*, Econometrica, 53 (1985), p. 1315. 8, 26, 48, 96

[80] A. S. Kyle and A. A. Obizhaeva, *Market Microstructure Invariants : Theory and Implications of Calibration.* 2011. 51

[81] J. Large, *Measuring the resiliency of an electronic limit order book*, Journal of Financial Markets, 10 (2007), pp. 1–25. 12

[82] S. Laruelle, C.-A. Lehalle, and G. Pages, *Optimal split of orders across liquidity pools: a stochastic algorithm approach*, SIAM Journal on Financial Mathematics, 2 (2011), pp. 1042–1076. 13, 65

[83] C. Lee, *Spreads, depths, and the impact of earnings information: an intraday analysis*, Review of Financial Studies, 6 (1993), pp. 345–374. 25, 46

[84] C. Lee and M. Ready, *Inferring trade direction from intraday data*, Journal of Finance, 46 (1991), pp. 733–746. 34

[85] C.-A. Lehalle, U. Horst, and Q. Li, *Optimal trading in a two-sided limit order book*. 2013. 13

[86] T. M. Li and R. Almgren, *A fully-dynamic closed-form solution for delta-hedging with market impact*. 2011. 11

[87] A. Lucchetti, *UBS cuts floor staff at NYSE*, Wall Street Journal, (2007). 3

[88] R. Lyons, *Strong laws of large numbers for weakly correlated random variables*, Michigan Mathematical Journal, 35 (1988), pp. 353–359. 50

[89] A. Madhavan, M. Richardson, and M. Roomans, *Why do security prices change? A transaction-level analysis of NYSE stocks*, Review of Financial Studies, 10 (1997), pp. 1035–1064. 24, 25, 47, 48

[90] K. Malinova, A. Park, and R. Riordan, *Do Retail Traders Suffer from High Frequency Traders?* 2012. 4, 94

[91] T. H. Mcinish and R. A. Wood, *An analysis of intraday patterns in bid/ask spreads for NYSE stocks*, Journal of Finance, 47 (1992), pp. 753–764. 25

[92] A. Menkveld, *High frequency trading and the new-market makers*. 2011. 89

[93] A. J. Menkveld and B. Yueshen, *Anatomy of the Flash Crash*. 2013. 3

[94] C. Moallemi, C. Maglaras, and H. Zheng, *Optimal Order Routing in a Fragmented Market*. 2012. 12, 14, 65, 66, 96

[95] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal on Optimization, 19 (2009), pp. 1574–1609. 84

[96] A. Obizhaeva, *Selection bias in liquidity estimates*. 2011. 9

[97] A. A. Obizhaeva and J. Wang, *Optimal trading strategy and supply/demand dynamics*, Journal of Financial Markets, 16 (2012), pp. 1–32. 12, 13, 22, 64

[98] E. R. Odders-White, *On the occurrence and consequences of inaccurate trade classification*, Journal of Financial Markets, 3 (2000), pp. 259–286. 34

[99] B. Park and B. V. Roy, *Adaptive Execution : Exploration and Learning of Price Impact*. 2012. 11

[100] C. Parlour and D. Seppi, *Limit order markets: A survey*, in Handbook of Financial Intermediation and Banking, A. Boot and A. Thakor, eds., 2008. 7, 8, 95

[101] R. Pascual and D. Veredas, *Does the open limit order book matter in explaining informational volatility?*, Journal of Financial Econometrics, 8 (2010), pp. 57–87. 24

[102] V. Plerou, P. Gopikrishnan, X. Gabaix, and H. Stanley, *Quantifying stock-price response to demand fluctuations*, Physical review E: Statistical physics, plasmas, fluids, and related interdisciplinary topics, 66 (2002), p. 27104. 10, 22, 23, 31, 37

[103] M. Potters and J. P. Bouchaud, *More statistical properties of order books and price impact*, Physica A, 324 (2003), pp. 133 – 140. 10, 22, 23, 30, 31

[104] S. Predoiu, G. Shaikhet, and S. Shreve, *Optimal Execution in a General One-Sided Limit-Order Book*, SIAM Journal on Financial Mathematics, 2 (2011), pp. 183–212. 12, 13, 64

[105] E. Robert, F. Robert, and R. Jeffrey, *Measuring and modeling execution cost and risk*, The Journal of Portfolio Management, 38 (2012), pp. 14–28. 11, 22

[106] I. Rosu, *A Dynamic Model of the Limit Order Book*, Review of Financial Studies, 22 (2009), p. 4601. 30, 96

[107] P. Sandås, *Adverse selection and competitive market making: Empirical evidence from a limit order market*, Review of Financial Studies, 14 (2001), pp. 705–734. 12, 69

[108] E. Smith, J. D. Farmer, L. Gillemot, and S. Krishnamurthy, *Statistical theory of the continuous double auction*, Quantitative finance, 3 (2003), pp. 481–514. 96

[109] Staffs of the CFTC and SEC, *Findings Regarding the Market Events of May 6, 2010*. 2010. 3

[110] C. Stephens, H. Waelbroeck, and A. Mendoza, *Relating market impact to aggregate order flow: the role of supply and demand in explaining concavity and order flow dynamics*. 2009. 23, 31

[111] S. Stoikov and R. Waeber, *Optimal Asset Liquidation using Limit Order Book Information*. 2012. 12, 14, 24, 96

[112] P. Swan and P. Westerholm, *Market Architecture and Global Exchange Efficiency: One Design Need Not Fit All Stock Sizes*. 2006. 5, 95

[113] E. Theissen, *A test of the accuracy of the Lee/Ready trade classification algorithm*, Journal of International Financial Markets, Institutions and Money, 11 (2001), pp. 147–165. 34

[114] N. Torre and M. Ferrari, *The Market Impact Model*, BARRA, 1997. 10, 22, 31, 37

[115] B. Toth, Y. Lemperiere, C. Deremble, J. De Lataillade, J. Kockelkoren, and J.-P. Bouchaud, *Anomalous price impact and the critical nature of liquidity in financial markets*, Physical Review X, 1 (2011), p. 021006. 9

[116] P. Weber and B. Rosenow, *Order book approach to price impact*, Quantitative Finance, 5 (2005), pp. 357–364. 23

[117] ——, *Large stock price changes: volume or liquidity?*, Quantitative Finance, 6 (2006), p. 7. 23

[118] W. Whitt, *Stochastic-Process Limits*, Springer, 2002. 50

[119] I. Zovko and J. D. Farmer, *The power of Patience: A Behavioral Regularity in Limit Order Placement*, Quantitative Finance, 2 (2002), pp. 387–392. 30