Essays on Liquidity Risk and Modern Market Microstructure

Kai Yuan

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy under the Executive Committee of the Graduate School of Arts and Science

COLUMBIA UNIVERSITY

2017
ABSTRACT

Essays on Liquidity Risk and Modern Market Microstructure

Kai Yuan

Liquidity, often defined as the ability of markets to absorb large transactions without much effect on prices, plays a central role in the functioning of financial markets. This dissertation aims to investigate the implications of liquidity from several different perspectives, and can help to close the gap between theoretical modeling and practice.

In the first part of the thesis, we study the implication of liquidity costs for systemic risks in markets cleared by multiple central counterparties (CCPs). Recent regulatory changes are transforming the multi-trillion dollar swaps market from a network of bilateral contracts to one in which swaps are cleared through central counterparties (CCPs). The stability of the new framework depends on the resilience of CCPs. Margin requirements are a CCP’s first line of defense against the default of a counterparty. To capture liquidity costs at default, margin requirements need to increase superlinearly in position size. However, convex margin requirements create an incentive for a swaps dealer to split its positions across multiple CCPs, effectively “hiding” potential liquidation costs from each CCP. To compensate, each CCP needs to set higher margin requirements than it would in isolation. In a model with two CCPs, we define an equilibrium as a pair of margin schedules through which both CCPs collect sufficient margin under a dealer’s optimal allocation of trades. In the case of linear price impact, we show that a necessary and sufficient condition for the existence of an equilibrium is that the two CCPs agree on liquidity costs, and we characterize all equilibria when this holds. A difference in views can lead to a race to the bottom. We provide extensions of this result and discuss its implications for CCP oversight and risk management.

In the second part of the thesis, we provide a framework to estimate liquidity costs at a portfolio
level. Traditionally, liquidity costs are estimated by means of single-asset models. Yet such an approach ignores the fact that, fundamentally, liquidity is a portfolio problem: asset prices are correlated. We develop a model to estimate portfolio liquidity costs through a multi-dimensional generalization of the optimal execution model of Almgren and Chriss (1999). Our model allows for the trading of standardized liquid bundles of assets (e.g., ETFs or indices). We show that the benefits of hedging when trading with many assets significantly reduce cost when liquidating a large position. In a “large-universe” asymptotic limit, where the correlations across a large number of assets arise from a relatively few underlying common factors, the liquidity cost of a portfolio is essentially driven by its idiosyncratic risk. Moreover, the additional benefit from trading standardized bundles is roughly equivalent to increasing the liquidity of individual assets. Our method is tractable and can be easily calibrated from market data.

In the third part of the thesis, we look at liquidity from the perspective of market microstructure, we analyze the value of limit orders at different queue positions of the limit order book. Many modern financial markets are organized as electronic limit order books operating under a price-time priority rule. In such a setup, among all resting orders awaiting trade at a given price, earlier orders are prioritized for matching with contra-side liquidity takers. In practice, this creates a technological arms race among high-frequency traders and other automated market participants to establish early (and hence advantageous) positions in the resulting first-in-first-out (FIFO) queue. We develop a model for valuing orders based on their relative queue position. Our model identifies two important components of positional value. First, there is a static component that relates to the trade-off at an instant of trade execution between earning a spread and incurring adverse selection costs, and incorporates the fact that adverse selection costs are increasing with queue position. Second, there is also a dynamic component, that captures the optionality associated with the future value that accrues by locking in a given queue position. Our model offers predictions of order value at different positions in the queue as a function of market primitives, and can be empirically calibrated. We validate our model by comparing it with estimates of queue value realized in backtesting simulations using marker-by-order data, and find the predictions to be accurate. Moreover, for some large tick-size stocks, we find that queue value can be of the same
order of magnitude as the bid-ask spread. This suggests that accurate valuation of queue position is a necessary and important ingredient in considering optimal execution or market-making strategies for such assets.
# Table of Contents

**List of Figures** iv  
**List of Tables** v  

## 1 Introduction 1  
1.1 Hidden Illiquidity and Multiple Central Counterparties 4  
1.2 Portfolio Liquidity Estimation and Optimal Execution 7  
1.3 A Model for Queue Position Valuation 11  

## 2 Hidden Illiquidity with Multiple Central Counterparties 17  
2.1 Introduction 17  
2.2 Background on Central Clearing 18  
2.3 Hidden Illiquidity 22  
2.4 Model 27  
2.5 Linear Price Impact 31  
2.5.1 Equilibrium Characterization 32  
2.5.2 Race to the Bottom 34  
2.5.3 Partitioned Clearing 37  
2.6 Adding Uncertainty 40  
2.7 A Single Instrument with General Price Impact 41  
2.8 Implications and Concluding Remarks 45
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>47</td>
</tr>
<tr>
<td>3.2 Model</td>
<td>49</td>
</tr>
<tr>
<td>3.2.1 Setup</td>
<td>49</td>
</tr>
<tr>
<td>3.2.2 Optimal Strategy</td>
<td>54</td>
</tr>
<tr>
<td>3.3 Examples: Separable Transaction Costs</td>
<td>56</td>
</tr>
<tr>
<td>3.3.1 Zero-Cost Constrained Liquidity</td>
<td>57</td>
</tr>
<tr>
<td>3.3.2 Linear-Cost Constrained Liquidity</td>
<td>60</td>
</tr>
<tr>
<td>3.4 Large Universe</td>
<td>62</td>
</tr>
<tr>
<td>3.4.1 Zero-Cost Constrained Trading</td>
<td>66</td>
</tr>
<tr>
<td>3.4.2 Vanishing Bid-Ask Spread</td>
<td>68</td>
</tr>
<tr>
<td>3.4.3 Linear Transaction Costs</td>
<td>70</td>
</tr>
<tr>
<td>3.4.4 Hedging with Liquidity Bundles</td>
<td>71</td>
</tr>
<tr>
<td>3.5 Empirical Results</td>
<td>73</td>
</tr>
<tr>
<td>3.5.1 Overview of the Data Set</td>
<td>73</td>
</tr>
<tr>
<td>3.5.2 Model Calibration</td>
<td>74</td>
</tr>
<tr>
<td>3.5.3 Results</td>
<td>76</td>
</tr>
<tr>
<td>3.6 Concluding Remarks</td>
<td>78</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>81</td>
</tr>
<tr>
<td>4.2 Model</td>
<td>82</td>
</tr>
<tr>
<td>4.2.1 Order Valuation</td>
<td>84</td>
</tr>
<tr>
<td>4.2.2 Price Dynamics</td>
<td>86</td>
</tr>
<tr>
<td>4.2.3 Limit Order Book Dynamics</td>
<td>88</td>
</tr>
<tr>
<td>4.3 Analysis</td>
<td>91</td>
</tr>
<tr>
<td>4.4 Empirical Calibration</td>
<td>95</td>
</tr>
<tr>
<td>4.4.1 Data Overview</td>
<td>96</td>
</tr>
</tbody>
</table>
4.4.2 Calibrating Parameters ........................................... 96
4.4.3 Observations ......................................................... 100
4.5 Empirical Validation: Backtesting ............................... 102
  4.5.1 Backtesting Simulation .......................................... 102
  4.5.2 Observations ....................................................... 104
  4.5.3 Discussion .......................................................... 105
4.6 Concluding Remarks ................................................. 106

Bibliography ................................................................ 106

A APPENDIX ................................................................ 114
  A.1 Additional Proofs for Chapter 2 ................................. 114
  A.2 Additional Proofs for Chapter 3 ................................. 124
    A.2.1 Proofs for Section 3.2 .......................................... 124
    A.2.2 Proofs for Section 3.3 .......................................... 127
    A.2.3 Proofs for Section 3.4 .......................................... 135
  A.3 Additional Proofs for Chapter 4 ................................. 152
List of Figures

1.1 An illustration of a limit order book. ......................................................... 12

2.1 (a) Payment obligations in an OTC market. (b) Payment obligations after bilateral
netting. (c) Payment obligations in a centrally cleared market. .................... 19

2.2 Variation margin covers the value of a clearing member’s swap portfolio at the time
of default. Initial margin should cover costs the CCP may incur from the time of
default to the completion of the close-out of defaulting member’s portfolio. ... 21


2.4 Histogram of average number of daily CDS trades per reference entity (Q1, 2013). 28

2.5 Histogram of bid-ask spreads for CDS (2013). .................................. 29

2.6 Variance explained by the first 10 principal components of covariance matrices. 32

2.7 Margin requirements combine like resistors connected in parallel. ........ 34

3.1 Liquidity cost as the number of assets for hedging increases. .................. 79

3.2 Convergence of the liquidity costs. ......................................................... 79

4.1 Various Futures Contracts, July–August 2013 (courtesy Rob Almgren) ....... 83

4.2 Model outputs as functions of queue positions on two different trading days (08/09/2013
and 08/20/2013). The red dots represent the average queue length of that trading
day. ............................................................................................................. 101
List of Tables

3.1 Descriptive statistics for the equity holdings of the assets under discussion. The weights and prices are as of 04/01/2016. The average daily volume is calculated through the period 01/01/2012 – 04/01/2016. The volatility is defined as the standard deviation of percentage daily returns. The volume trade through ETF is calculated as $|\gamma_{XLU}/\alpha_j|/\gamma_j$. 74

3.2 Theoretical results for the four trading strategies. 76

3.3 Numerical results for the utility-sector example. 77

4.1 Estimated market parameters for BAC in a month. $\lambda$ is estimated as the price impact in basis points for one percent of daily volume. Note that here we consider only shares traded on NASDAQ. 99

4.2 Descriptive statistics for 9 stocks over the 21 trading days of August 2013. The average bid/ask spread is defined as the time average computed from the ITCH data. The volatility is defined as the standard deviation of percentage daily returns. All other statistics were retrieved from Yahoo Finance. 102

4.3 Estimated model values vs. simulation values. All the values above were calculated as the average across 30 trading days. Touch value refers to the value of orders at the very front of the queue. 105
Acknowledgments

I express my greatest gratitude to Professor Ciamac C. Moallemi for guiding me through the journey. His insightful advice helped shape my way of thinking into the mindset of a researcher. He introduced me to the world of quantitative finance and helped me build an understanding of the current literature. Professor Moallemi kept motivating me to explore novel research ideas and to develop as an independent researcher, for which I am deeply grateful. His unreserved guidance not only benefited my graduate study, but will continue to influence my future development and career.

I am truly grateful to my co-advisor—Professor Paul Glasserman for his illuminating guidance and unparalleled support. He has been an exceptional role model as a researcher for his dedication, scientific curiosity and intellectual breadth. His deep understanding of the financial market has helped me build intuitions for research ideas that contribute to this thesis. I would not have accomplished this without his support and guidance over the years.

I would like to thank Professor Costis Maglaras for being an invaluable source of advice and help during my graduate study. His door is always open for advice on both research and career planning. I am indebted to Professor Alireza Tahbaz-Salehi and Professor Agostino Capponi for many beneficial conversations and to serve on my thesis committee.

This work also benefited from the supportive research environment created by the faculty in Columbia Business School and Department of Industrial Engineering & Operations Research. I would also like to thank my fellow students in Columbia University for all the cheerful moments in the last five years.

I owe my thanks to Professor Minwen Li and Professor Hao Wang, my undergraduate advisors at Tsinghua University. Though they did not directly contribute in this thesis, it would not happen
had they not ignited my interest in research and encouraged me to pursue the graduate degree.

Finally, I am forever indebted to my parents, Chao Yuan and Yan Jiao. Their unconditional love make my accomplishment meaningful. Special thanks to Simiao Chen, for being supportive when I was down, and for being my muse and source of inspiration.
To Yan, Chao and Simiao
Chapter 1

Introduction

Liquidity is an rather broad yet elusive notion. In the most general sense, liquidity relates to “the ability of an economic agent to exchange his or her existing wealth for goods and services or for other assets”. In this thesis, we are particularly interested in the notion of market liquidity which relates to the ability of markets to absorb large transactions of financial assets without much effect on prices. Liquidity risk is then defined as the inability or potential cost of trading with immediacy.

In general, the cost of executing a certain position comes in many ways. The first is the fees (or rebates) charged by the brokers and the exchange for their service. This is the most explicit part of the cost that any investor pays and are often charged at a constant rate. Yet fees only takes up a small component of the potential liquidity cost.

The second is the bid-ask spread paid by investors who take liquidity from the market by, for example, placing market orders to buy or sell. A bid-ask spread is defined as the difference between the ask price and bid price in the market. The economic intuitions behind bid-ask spread has been a important topic in the microstructure literature. Most markets are organized by centralized specialists (as in the traditional dealer markets) or market makers (as in markets operate under electric limit order books) who constantly provide liquidity and set the spread. Generally, the spread has to be large enough to cover potential costs for those liquidity providers. And those

CHAPTER 1. INTRODUCTION

costs may include inventory costs as in [Stoll (1978)] and order-handeling costs as in [Roll (1984)]. More importantly, [Glosten and Milgrom (1985)] characterized bid-ask spread as a result of adverse selection, where liquidity providers charge for the possibility of trading with agents with superior information. In any case, the bid-ask spread reflects the willingness to trade of liquidity suppliers, and often acts as a barometer for liquidity situations in the market.

The third component for liquidity cost is the price impact which is defined as the price movement due to the trading activity. For example, a large buy order can push prices higher, making subsequent purchases more expensive. Similarly, a sell order can push prices lower, reducing revenue from subsequent sales. The concept of price impact are first established in the market microstructure literature, a review of which can be found in [Biais et al. (2005)]. This literature has shown that orders have both a transitory and a permanent impact on prices. In the short term, the order creates imbalance between supply and demand, which prompts the market makers to move price to source more liquidity. This effect is due to the lack of liquidity in the market and often poses no impact on the fundamental value of the asset. Therefore, market price may soon reverse after the order ends. The permanent component, on the other hand, reflects the information inferred from the order flow by market makers. The fact that some orders may come from traders with superior information prompts the market makers to consistently update their quotes to compensate for adverse selection. This information is then permanently incorporated into market price.

Finally, liquidity cost also comes in the form of inventory risk. In order to minimize the overall price impact, large trades are usually split into smaller ones and executed over time. This creates inventory risks as fluctuations in market price can increase the gap between remaining and targeted position.

One closely related problem is that of optimal execution, which tries to find the optimal strategy to unload a position at a low cost within a limit amount of time. Early literature such as [Bertsimas and Lo (1998)] addressed the problem by solving a dynamic programing problem to minimize mean transaction costs. Later formulations led by [Almgren and Chriss (1999)] also accounted for inventory risks and therefore tried to balance the trade-off between risks and costs. A more detailed review of literature on this issue is given in Section 1.2.
CHAPTER 1. INTRODUCTION

On a broader scale, the cost of liquidating large positions, especially in time of distress, can potentially pose great threat to financial stability. For example, liquidity risk played a devastating role in the most recent financial crisis. Highly leveraged institutions panically reduced positions at a time when liquidity was scarce, therefore creating a fire sale which moved prices against them leading to further losses. Pedersen (2008) describes this process as a “liquidity spiral”. After the financial crisis, over-the-counter swap trades are required by law to be central cleared. In Chapter 2, we find that this structure, though designed to mitigate risks by concentrating exposures in central counterparty (CCP), potentially creates a new source of systemic risk related to the resilience of the CCP itself. We show that a lack of coordination between CCPs could lead to a systematically underestimation of liquidity cost, which threatens the stability of the central clearing system. More details can be found in Section 1.1.

Managing liquidity risk is important in portfolio management, as the value of a portfolio depends on its ability to be converted into cash, especially in time of distress. For open-end mutual fund, the ability to meet its redemption request through adequate liquidity management is one of its core responsibilities. As one of the regulators puts it:

“Daily redeemability is a defining feature of mutual funds. This means that liquidity management is not only a regulatory compliance matter, but also a major element of investment risk management, an intrinsic part of portfolio management, and a constant area of focus for fund managers.”

On October 13, 2016, the US Securities and Exchange Commission (SEC) adopted a far-reaching rules requiring all mutual funds and open-end ETF to implement formal liquidity management program. Meeting those standards requires accurate estimation of liquidity cost. In Chapter 3, we provide a novel framework to estimate the liquidity costs in unloading portfolios instead of single assets. Our work contributes to the literature of optimal execution and can help to fill the gap between practice and theoretical modeling.

In Chapter 2 and Chapter 3, we treat liquidity cost as exogenous functionals depending on the

---

2See ICI FSOC Notice Comment Letter, supra note 16.
size of the transaction. But at a micro level, liquidity risk arises from the exchange of liquidity among agents in market places such as limit order books. Chapter 4 contributes to the rich microstructure literature that help determine the micro foundation of liquidity risks. More specifically, we investigate the value of limit orders at different queue positions, which can help solve high-level decision problems such as market making and optimal execution.

The rest of this chapter introduces the following chapters in depth by positing the research questions and objectives along with their connections to the literature. The research in Chapter 2 is a joint work with Professor Paul Glasserman and Professor Ciamac C. Moallemi. The research in Chapters 3 and Chapter 4 resulted from collaborations with Professor Ciamac C. Moallemi.

1.1. Hidden Illiquidity and Multiple Central Counterparties

Swap contracts enable market participants to transfer a wide range of financial risks, including exposure to interest rates, credit, and exchange rates. But swaps themselves can be risky. They create payment obligations that often extend for five to ten years, and they allow participants to take on highly leveraged positions. Indeed, while its proponents see the multi-trillion dollar swap market as an efficient mechanism for risk management and transfer, critics have long seen it as an opaque threat to financial stability.

Regulatory changes are transforming the swap market. Prior to the financial crisis of 2007–2008, nearly all swaps traded over-the-counter (OTC) as unregulated bilateral contracts between swap dealers or between dealers and their clients. In contrast, the 2010 Dodd-Frank Act requires central clearing of all standard swap contracts in the United States, and the European Market Infrastructure Regulation (EMIR) imposes the same requirement in the European Union. The new rules also bring greater price transparency to swaps trading.

In an OTC market, when two dealers enter into a swap contract, they commit to make a series of payments to each other over the life of the swap. Each dealer is exposed to the risk that the other party may default and fail to make promised payments. In a centrally cleared market, the contract between the two dealers is replaced by two back-to-back contracts with a central counterparty
CHAPTER 1. INTRODUCTION

(CCP). The dealers are no longer exposed to the risk of the other’s failure because each now transacts with the CCP.

However, this arrangement takes the diffuse risk of an OTC market and concentrates it in CCPs, potentially creating a new source of systemic risk. So long as all its counterparties survive, the CCP faces no risk from its swaps — its payment obligations to one party are exactly offset by its receipts from another party. But for central clearing to be effective, the CCP needs to have adequate resources to continue to meet its obligations even if one of its counterparties defaults. The disorderly failure of a swap CCP would be a major disruption to the financial system with potentially severe consequences for the broader economy.

As its first line of defense, a CCP collects margin from its swap counterparties in the form of cash or other high quality collateral. Margin — more precisely, initial margin — provides a buffer to absorb losses the CCP might incur at the default of a counterparty. If a dealer defaults, the CCP needs to replace its swaps with that dealer, and it may incur a cost in doing so. The initial margin posted by each counterparty is intended to cover this cost in the event of that counterparty’s default.

Because of limited liquidity in the market, the replacement cost is likely to be larger for a large position by more than a proportional amount. If the CCP needs to replace a $1 billion swap, it may find several dealers willing to trade; but if it needs to replace a $10 billion swap it may find few willing dealers, and those that will quote a price may command a premium to take on the added risk of the position. The consequences of this liquidity effect on margin are the focus of this paper.

An immediate implication of limited liquidity is that a CCP’s margin requirements should be convex and, in particular, superlinear in the size of a dealer’s position. A seemingly obvious but apparently overlooked point is that this is insufficient. The same dealer may have similar positions at other CCPs. If the dealer goes bankrupt, all CCPs at which the dealer participates need to close out their contracts with the dealer at the same time. The impact on market prices is driven by the combined effect from all CCPs. If each CCP sets its margin requirements based only on the positions it sees (as appears to be the case in practice), it underestimates the margin it needs. This is what we call hidden illiquidity. In fact, we show that the very convexity required to capture
CHAPTER 1. INTRODUCTION

illiquidity creates an incentive for dealers to split their trades across multiple CCPs, amplifying the effect.

We next examine the possibility that a CCP can compensate for the impact of positions it does not see by charging higher margin on the positions it does see. We analyze this problem through a model with one dealer, two CCPs, and multiple types of swaps. Given margin schedules from the CCPs, the dealer optimizes its allocation of trades to minimize the total margin it needs to post; given the dealer’s objective, the CCPs set their margin schedules to have enough margin to cover the system-wide price impact should the dealer default. An equilibrium is defined by margin schedules that meet this objective.

We derive our most explicit results when price impact is linear (so that margin requirements are quadratic). We characterize all equilibria and show, in particular, that margin requirements at the two CCPs need not coincide. A CCP with a steeper margin schedule gets less volume and therefore needs to compensate more for the volume it does not see, which it does with its steeper margin. However, we also show that a necessary condition for an equilibrium is that the two CCPs agree on the true price impact. Without this condition, we get “a race to the bottom” in which a CCP that views the true price impact as smaller drives out the other CCP.

We extend this result to allow CCPs to select a subset of swaps to clear. On the subset of swaps cleared by both CCPs, the previous result applies. Equilibrium now imposes a further necessary and sufficient condition precluding cross-swap price impacts between swaps cleared by just one CCP and swaps cleared by the other CCP. We also consider extensions that introduce uncertainty to the model.

We obtain partial results in the case of nonlinear price impact with a single type of swap. We observe that the dealer’s optimization problem combines the convex marginal schedules of the two CCPs into a single effective margin which is the inf-convolution of the individual schedules. A result in convex analysis states that the convex conjugate of an inf-convolution of two convex functions is the sum of the conjugates of these functions. We relate this result to conditions for equilibrium.
CHAPTER 1. INTRODUCTION

1.2. Portfolio Liquidity Estimation and Optimal Execution

Estimation of liquidity costs, those associated with trading a collection of large positions, is an important issue in modern financial markets. In portfolio management, estimation of liquidity costs is important since these costs can be significant. This is particularly true for investors who are very active (and hence incur significant costs by trading frequently) or are very large (and hence incur significant costs through their size). In such settings, effective portfolio construction decisions cannot be made without considering liquidity costs. Similarly, in risk management, assessment of the risk associated with holding a portfolio depends on both the long-term fluctuations in the value of the underlying assets and the short-term ability to convert the portfolio into cash. This latter effect can be especially important in times of distress, and is fundamentally a question of liquidity costs.

A closely related problem is that of optimal execution. In many markets, when an investor seeks to execute a large trade (a so-called “parent order”), it is usually broken into pieces with the help of algorithmic trading systems and executed as a sequence of much smaller trades (“child orders”). Optimal execution problems seek to do this in the most efficient manner by balancing two effects. First, there are transaction costs associated with execution, including, for example, commissions, fees, the bid-ask spread, and (most importantly for large investors) the market impact of the trading itself. Second, by spreading out a large trade over time, investors are exposed to risks associated with the movement of market prices over the execution horizon. Traders must evaluate their trading strategies against the transaction costs and market risks. Those who trade too fast incur high transaction costs from market impact while those who trade too slow are exposed to adverse price movements: both trading strategies could potentially result in more than the expected liquidity cost. This trade-off between cost and uncertainty has given rise to a rich literature on optimal execution in general and optimal liquidation of a single risky asset in particular, starting with the work of [Almgren and Chriss] (2001).

To date, much of the literature on the estimation of liquidity costs and optimal execution has focused on the single-asset setting (with several notable exceptions to be discussed shortly).
contrast, we believe that liquidity is fundamentally a *multi-asset* problem that must be addressed at the portfolio level. This is for several reasons:

(i) Investors make trading decisions seldom in isolation on an asset-by-asset basis, but rather jointly to produce a trade list consisting of a portfolio of trades to be made simultaneously in multiple assets. A simple example would be an open-end fund, which, upon an an inflow or outflow, would in effect trade portfolios to maintain proportional holdings. Since the market risk associated with such a trade depends on the joint distribution of correlated assets, the estimation of its liquidity costs will not decompose across assets, nor can optimal trading schedules be determined by considering assets in isolation.

(ii) Even if an investor seeks to trade only a single asset, he may receive significant benefits from simultaneously trading correlated assets for the hedging purposes. For example, an investor unwinding a position in an illiquid asset may seek to hedge the execution risk by establishing positions in correlated but liquid assets, in order to drive down overall liquidity costs.

(iii) Finally, investors may benefit from the multi-asset approach through the trading of what we call *liquid bundles*. These are collections of assets (in effect, portfolios) existing in many markets that can be directly and atomically traded. For example, in equity markets, investors can directly trade exchange-traded funds (ETFs), which are economically (ignoring creation and redemption issues) equivalent to trading a basket of underlying equities. Similarly, in credit markets, trading credit default swap (CDS) indices is equivalent to taking a simultaneous position in a portfolio of underlying credit entities. In futures markets, spread trades, such as calendar spreads, inter-commodity spreads (e.g., crack spreads), and option spreads, are also portfolio trades. Such portfolio instruments can be important both because they provide another mechanism for trading the constituent assets, and because they are often extremely liquid and have little idiosyncratic risk, which makes them excellent candidates as the hedging instruments.

In Chapter 3, we develop a multi-asset generalization of the model of Almgren and Chriss (2001), building on the work of Guéant (2015), Kim (2014), and Guéant et al. (2015). Going beyond this
earlier work, our model explicitly incorporates the trading of liquid bundles such as ETFs. Our model is easily calibrated and computationally tractable.

The most important contribution of our model, however, is that it enables us to provide a structural analysis of the underlying drivers of liquidity costs. Specifically, we make the assumption of a factor model, where the covariance structure across the universe of tradeable assets decomposes into common, systemic factors (which drive correlations) and individual, idiosyncratic risk. We consider a large-universe asymptotic regime, where a large number of assets are available for trading relative to the number of underlying systemic factors. This large-universe setting is consistent with asset pricing theory, particularly the assumptions made in the arbitrage pricing theory first developed by Ross (1976). It is also consistent with the state of the art in practice, where, for example, commercial risk models for equities (e.g., BARRA) use dozens of factors to explain the covariance structure for thousands of assets.

In this asymptotic large-universe setting, under suitable technical assumptions, we develop simple closed-form approximations for liquidity costs. These approximations are useful for computation, but they also highlight two key structural properties of portfolio liquidity costs. First, liquidity costs are primarily driven by idiosyncratic risk. This is because, in a large-universe setting, systemic risk can be hedged very cheaply and nearly eliminated. Put differently, the benefit from considering optimal execution at the portfolio level roughly corresponds to reducing risk exposure from total risk to only idiosyncratic risk. Second, introducing a liquid bundle (ETF) is approximately equivalent to commensurately increasing the liquidity of each underlying asset by its implied trading volume in the ETF. In other words, liquid high-volume ETFs can offer significant reductions in liquidity costs.

We explore the practical implications of our model in an empirical example consisting of 29 U.S. equities in the utility sector, along with a sector ETF. There, we demonstrate the above-referenced structure effects and illustrate the magnitude of the benefits of our approach. In particular, the portfolio approach to trading single assets in the utility sector can reduce liquidity costs by a factor of up to five. In addition, use of the sector ETF further reduces costs by 10–20%.

Research on optimal execution has been of particular academic interest in the past two decades.
CHAPTER 1. INTRODUCTION

It first started with Bertsimas and Lo (1998), who focused on the minimization of execution costs. The trade-off between transaction cost and market risk was first documented by Grinold and Kahn (2000), and was then used in the seminal papers of Almgren and Chriss (2001) and Almgren and Chriss (1999) to derive the framework of single-asset optimal execution in a mean-variance formulation. Initially in discrete time with linear market impact, the Almgren–Chriss model was extended to continuous time by He and Mamaysky (2005) and Forsyth (2011) using the Hamilton–Jacobi–Bellman approach, and by Almgren (2003) and Guéant (2015) using nonlinear market impact functions. Almgren (2012) further takes into account stochastic volatility and liquidity. Whereas these frameworks are all based on static or deterministic strategies in which the number of shares to be sold at any time is pre-specified, Almgren and Lorenz (2007) improves on them with the more realistic mean-variance formulation of a simple update strategy that accelerates execution when the prices move in favor of the trader. A more detailed discussion of the form of adaptivity is given in Lorenz and Almgren (2011).

Perhaps due to its mathematical difficulties, the portfolio approaches to optimal execution is much less studied. Almgren and Chriss (2001), followed by Engle and Ferstenberg (2007) and Brown et al. (2010), briefly discuss the portfolio approach and provide a solution to a simple case. In recent years, the body of work dedicated to the portfolio approach has grown. Kim (2014) considers the case where market impact is assumed to be minimal and decays sufficiently fast to be negligible in price dynamics. Guéant et al. (2015) present a numerical method to approximate the optimal execution strategy based on convex duality. While the framework used in these two papers is quite similar to that of the present paper, our framework is more general and allows for the trading of liquid bundles. Finally, Tsoukalas et al. (2014) analyze a multi-asset optimal execution problem; however, they confine their attention to the microstructure of cross-asset market impact.

One key observation to be drawn from all these papers is that there are large hedging benefits by using the portfolio approach.

Our research is also related to empirical research that conduct cross-sectional regressions to estimate market impacts. For example, Chacko et al. (2008) provide empirical evidence that the expected market impact is proportional to the square root of the trading size; see also Bouchaud.
et al. (2008). However, this approach has two downsides: it is extremely noisy (because it is hard to estimate transaction costs from actual returns) and from our perspective, it is fundamentally a single-asset approach.

1.3. A Model for Queue Position Valuation

Modern financial markets are predominantly electronic. In modern exchanges, market participants interact with each other through computer algorithms and electronic orders. The image of traders frantically gesturing and yelling to each other on the trading floor has largely given way to impersonal computer terminals. In terms of market structure, the electronic limit order book (LOB) has become dominant for certain asset classes such as equities and futures in the United States. Figure 1.1 illustrates how a limit order book works. It is presented as a collection of resting limit orders, each of which specifies a quantity to be traded and the worst acceptable price. The limit orders will be matched for execution with market orders which demand immediate liquidity. Traders can therefore either provide liquidity to the market by placing these limit orders or take liquidity from it by submitting market orders to buy or sell a specified quantity.

Most limit order books are operated under the rule of price-time priority, that is used to determine how limit orders are prioritized for execution. First of all, limit orders are sorted by the price and higher priority is given to the orders at the best prices, i.e., the order to buy at the highest price or the order to sell at the lowest price. Orders at the same price are ranked depending on when they entered the queue according to a first-in-first-out (FIFO) rule. Therefore, as soon as a new market order enters the trading system, it searches the order book and automatically executes against limit orders with the highest priority. More than one transaction can be generated as the market order may run through multiple subsequent limit orders. In fact, the FIFO discipline suggests that the dynamics of a limit order book resembles a queueing system in the sense that limit orders wait in the queue to be filled by market orders (or canceled). Prices are typically

---

3We do not make a distinction between market orders and marketable limit orders.

4There is an alternative rule called pro-rata, which works by allocating trades proportionally across orders at the same price. Pro-rata is less popular among exchanges and will not be covered here.
discrete in limit order books and there is a minimum increment of price which is referred to as tick size. If the tick size is small relative to the asset price, traders can obtain priority by slightly improving the order price. But it becomes difficult when the tick size is economically significant. As a result, queueing position becomes important as traders prefer to stay in the queue and wait for their turn of execution.

High-level decision problems such as market making and optimal execution are of great interest in both academia and industry. One of the decisions raised by those problems is when to use limit orders as opposed to market orders and how to place limit orders if they are preferred. The key ingredient of that decision is the estimation of the value of a limit order. In Chapter 4, we try to relate the value of a limit order to its queue position. We claim that queue positions are relevant and indeed positions at the front of the queue are very valuable for the following reasons. First of all, good queue positions guarantee early execution and less waiting time. This is particularly important for algorithmic traders who potentially have a large number of trades scheduled to be executed. Additionally, less waiting time can translate to a higher fill rate, because there is less chance that the market price will move away while the limit orders are sitting in the queue. Second,
good queue positions also mean few adverse selection costs. Orders at the end of a large queue will be executed in the next instance only against large trades. On the other hand, orders at the very front of the queue will be executed against the next trade no matter what its size will be. Large trades often originate from informed traders who are confident about the trades’ profitability. In this way, a good queue position acts as a filter on the population of contra-side market orders so that the liquidity provider is less likely to be disadvantaged by trading against informed traders. This relationship between queue positions and adverse selection is first documented in Glosten (1994), which considers a single-period setting.

In practice, we have seen investors expend huge amounts of money trying to take advantage of better queue positions in the limit order book. For example, there has been controversy in recent years over exotic order types on certain exchanges that allow traders to attain priority in the limit order book. These exotic order types “allow high-speed trading firms to trade ahead of less-sophisticated investors, potentially disadvantaging them and violating regulatory rules.”

This shows that there is indeed value in queue positions, as sophisticated investors are interested in paying to get better queue positions. Another example is that there has been an “arms race” between high-frequency traders to invest in technologies for low-latency trading, and part of the driver for low-latency trading is getting good queue positions. In fact, one situation where it is important to trade quickly is the moment right after a price change. For example, when a trade wipes out the current ask and the price is about to tick up, there will be a race to establish queue positions at the new price.

In the literature, some earlier work, such as that of Glosten (1994), has implications about the value of queue positions. Although these models point out the importance of adverse selection, they are fundamentally static models in which the value of the order is assumed to be determined by whether it will be executed by the next trade or not. In the presence of a large queue, the life cycle of the order will not end with the next trade and traders will not cancel and resubmit their limit orders after every single trade. What is more likely to happen is that the order will move up in

---

the queue, if not executed by the next trade. Actually, one way of getting to the front of the queue eventually is to join the queue right now. Therefore there is value in moving up in the queue, and that value may accrue over a number of trades and cancellations. As a result, aside from adverse selection, there should be another dynamic component that can capture the optionality associated with future value that accrues by locking in a given queue position. In order to account for this dynamic component, a multi-period model is needed.

In Chapter 4, we provide a dynamic model for valuing limit orders in large-tick stocks based on their relative queue positions. We appear to be one of the first to study the limit-order-book queueing value through the lens of dynamic multi-period model. Our model identifies two important components of positional value. First, there is a static component that relates to the adverse-selection costs originating from the possibility of information-motivated trades. We capture the fact that adverse selection costs are increasing with queue position. Second, there is also a dynamic component that captures the value of positional improvement that accrues after order-book events such as trades and cancellations.

By making reasonable simplifications, we provide a tractable way to predict order value at different positions in the queue as a function of market primitives. We then empirically calibrate our model in a subset of U.S. equities and find that queue values can be very significant in large-tick assets. Additionally, we validate our model by checking the model-free estimates of queue values using a backtesting technique.

There are many higher-level decision problems that have an ingredient of valuing limit orders. One such example is that market makers need to constantly value limit orders in order to come up with the optimal order-placing strategy. Another example is that in the optimal execution of a large block, algorithmic traders often have to decide between market orders and limit orders. In both cases, we need to value the limit orders and use them as building blocks for the higher-level control problem. What we observe empirically in our model is that queue positions do matter and that positional value is roughly of the same magnitude for large-tick assets. As a result, queue positional value should be an important ingredient downstream of solving optimal control problems with large-tick assets.
Our work builds on the classical financial economics literature on market microstructure that studies the informational motives of trading. Kyle (1985) and Glosten and Milgrom (1985) were among the first to recognize the importance of adverse selection in analyzing the price impact of trades and the spread, by assuming competitive suppliers of liquidity. Both of their models highlight the fact that the possibility of trading against an informed trader creates incentives for liquidity providers to charge additional premiums. However, these models do not consider queueing effect. Glosten (1994) further extended this type of model, with implications for valuing orders in the limit order book. One contribution of the paper is that it states that in cases where the prices are discrete, the queue length should be determined by the fact that the value of the last order in the queue is zero. Basically, the investor putting in the marginal order should be indifferent between joining the queue or not. While the paper does not explicitly model the value of queue positions, it does manage to relate queue length to order values. Moreover, the model in Glosten (1994) is a single-period static model in which the order values are calculated toward the next trade. However, what’s more likely is that an order will move up in the queue if it is not executed. Our model incorporates the dynamic values embedded in the queue position improvement. Additionally, by considering a dynamic model, we are also able to consider order book events such as cancellations. As a result, queue position actually matters in our model, and is clearly correlated with the order values. For example, if the queue position is decreasing, then either there is a trade or people are canceling, and either event conveys information about asset value.

Recently, there has been a growing literature from the financial engineering community on the development of queueing models that solve various kinds of problems regarding limit order books while recognizing that the price-time priority structure in the limit order books can be modeled as a multi-class queueing system. Cont et al. (2010) was the first to model the limit order book as a continuous-time Markov model that tracks the limit orders at each price level. By assuming that order flows can be described as Poisson processes, the authors provided a parametric way to calculate the conditional probability of various order book events such as the probability of executing an order before a change in price. Cont and De Larrard (2013) further modeled the

\footnote{In fact, the paper assumes that competing limit orders in the same queue are executed in a pro-rata fashion.}
CHAPTER 1. INTRODUCTION

order-book events in a Markovian queueing system, and studied the endogenous price dynamics resulting from executions. Lakner et al. (2013) studied a similar setup but focused on the high-frequency regime where the arrival rate of both limit orders and market orders is large. Blanchet and Chen (2013) derived a continuous-time model for the joint evolution of the mid price and the bid-ask spread. Several papers such as Guo et al. (2013), Cont and Kukanov (2013), and Maglaras et al. (2015) have been working on optimizing trading decisions in the context of a queueing model for the limit order book. More specifically, Guo et al. (2013) proposed a model to optimally place orders, given price impact. Cont and Kukanov (2013) derived the optimal split between limit and market orders across multiple exchanges. Maglaras et al. (2015) studied optimal decision making in the placement of limit orders as well as in trying to execute a large trade over a fixed time horizon. Avellaneda et al. (2011) tried to forecast the price change based on order-book imbalance, while in our settings price changes are exogenous. However, the limitation of the queueing literature is that it lacks the informational component of adverse selection. And yet an important ingredient in modeling the positional value of limit orders is the concept of adverse selection, i.e., of a correlation between trades and prices. Our model tries to bridge this gap by considering the economics of adverse selection in a queueing framework.

From the empirical front, there is a significant body of literature conducting empirical analyses of the dynamics of limit order books in major exchanges. Bouchaud et al. (2006) showed that the random-walk nature of traded prices is nontrivial. Biais et al. (1995) and Griffiths et al. (2000) studied the limit-order submission under different market conditions. Hollifield et al. (2004) further stated that optimal order submission depends not only on the valuation of the assets but also on the trade-offs between order prices, execution probabilities, and picking-off risks.

There are several successful examples of modeling the optionality embedded in limit orders. Copeland and Galai (1983) argued that informed traders are willing to pay a “fee” to obtain immediacy in trading with liquidity providers. Chacko et al. (2008) further modeled limit orders as American options that require delivery of the underlying shares upon execution. However, these models are fundamentally static in that they do not explicitly model the queue positions.
Chapter 2

Hidden Illiquidity with Multiple Central Counterparties

2.1. Introduction

The world of swap trading has shifted from unregulated bilateral contracts that traded over-the-counter (OTC) to back-to-back contracts that are cleared by a central counterparty (CCP). In this setup, the CCPs always have a net position of zero by construction, as its payment obligations to one party are exactly offset by its receipts from another party. However, a CCP is still subject to the failure of its counterparties, which may create a source of systemic risk. Therefore, a CCP collects margins from its counterparties to absorb potential losses from the default.

Every time when the market is going up or down, the CCP is collecting variation margins from the clearing memeber to compensate. At the point of the default, the CCP will be holding just enough cash from that clearing member to cover the full value of its portfolio. Since the CCP is not allowed to hold position, it need to find a new counterparty to take over the position of the failing clearing member. This process, however, is often costly. To cover the replacement cost, the CCP charges initial margin according to the clearing positions.

Given limited liquidity in the market, this replacement cost can be enormous and superlinear in the size of the position. The key idea in this analysis is that margin requirement need to cover
the replacement cost, and therefore need to grow superlinearly with position size. In the presence of multiple central counterparties, the very fact that CCPs have to set the right amount of initial margin according to superlinear liquidity charges creates the incentive for dealers to split their positions among multiple CCPs. Therefore, each CCP clears only a fraction of the dealer’s total position. And since each CCP charges margins based on the potential impact from the default of a clearing member and the subsequent liquidation of a large position, swaps dealers are effectively “hiding” potential liquidation costs. We investigate the CCP’s optimal strategy in a systemic way and acknowledge that this will not work if different CCPs have different views on the “right” amount of margin. As a result, a lack of coordination among CCPs can lead to a “race to the bottom” because CCPs with lower perceived liquidation costs can drive competitors out of the market.

The rest of this chapter is organized as follows: Section 2.2 provides some background on central clearing. Section 2.3 introduces the notion of hidden illiquidity. Section 2.4 introduces our model and our definition of equilibrium. Section 2.5 considers the case of linear price impact, including a necessary and sufficient condition for equilibrium and an analysis of what happens when the condition fails to hold. In Section 2.6, we extend the model to include uncertainty. In Section 2.7, we analyze nonlinear price impact in the case of a single type of instrument. Section 2.8 concludes and provides practical implications of our analysis. Most proofs appear in the appendix.

2.2. Background on Central Clearing

Figure 2.1 illustrates the difference between an over-the-counter market and a centrally cleared market. In part (a) of the figure, dealers A, B, and C trade bilaterally. They initiate trades directly with each other, and each pair of dealers manages payments on its swaps.

The numbers in part (a) of the figure show hypothetical payments due between dealers. Dealers may have multiple swaps with each other — indeed, the number of contracts would typically be very large — leading to payment obligations in both directions. The total payments due at any point in time may be viewed as a measure of the total counterparty risk in the system. In the figure, the total comes to 42.
CHAPTER 2. HIDDEN ILLIQUIDITY WITH MULTIPLE CENTRAL COUNTERPARTIES

Figure 2.1: (a) Payment obligations in an OTC market. (b) Payment obligations after bilateral netting. (c) Payment obligations in a centrally cleared market.

Bilateral netting between pairs of dealers can greatly reduce total counterparty risk. Part (b) of Figure 2.1 shows the result of bilateral netting of payment obligations. Total payments have been reduced to 20. In fact, further netting is still possible — in particular, dealer C makes a net payment of zero. However, further netting would require coordination among all three dealers and cannot be achieved bilaterally.

Part (c) of the figure illustrates a market with a central counterparty (CCP). After two dealers agree to enter into a swap, their bilateral contract is replaced by two mirror-image contracts running through the CCP. Whatever payments dealer B would have made to dealer A it makes instead to the CCP. The CCP in turn assumes responsibility for making the payments that A would have received from B. With all the contracts from part (a) of the figure running through a single CCP, central clearing achieves maximal netting in part (c) of the figure, reducing the total payments due

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]

\[ \text{Equation} \]
CHAPTER 2. HIDDEN ILLIQUIDITY WITH MULTIPLE CENTRAL COUNTERPARTIES

to 8. This reduction in system-wide counterparty risk is one of the main arguments for central clearing. Moreover, the CCP theoretically always has a net risk of zero in the sense that the total payments it needs to make on swaps equal the total payments it is owed.

This simple example overstates the benefits of central clearing in several respects. Dealers engaged in different types of OTC swaps — interest rate swaps and credit default swaps, for example — can net bilateral payments across all swaps; so, if different types of swaps are cleared through different CCPs, central clearing can actually reduce the total amount of netting. (See Duffie and Zhu (2011) and Cont and Kokholm (2014) for more on this comparison.) Some of the multilateral netting benefit provided by a CCP can be achieved in an OTC market through third-party trade compression services. In both OTC and centrally cleared markets, dealers provide collateral for their payment obligations, which reduces the counterparty risk that remains from any unnetted exposures. With central clearing, the CCP faces risk from the default of a dealer because of the costs it may incur in replacing or unwinding positions after the dealer fails.

This last point motivates our analysis so we discuss it in further detail. To protect itself from the failure of a clearing member, the CCP collects two types of margin payments from each member on at least a daily basis, variation margin and initial margin. Variation margin reflects daily price changes in a clearing member’s swaps. If the market value of the member’s swaps decreases, the member makes a variation margin payment to the CCP; if the market value increases, the CCP credits the member’s variation margin account. At the time of a clearing member’s default, the variation margin collected by the CCP should offset the value of the clearing member’s position.

Figure 2.2, based on a similar figure in Murphy (2012), illustrates the two types of margin. The figure shows the hypothetical evolution of the value of a clearing member’s swap portfolio over time, from the perspective of the CCP. The value may be positive or negative. In the figure, the clearing member fails at a time when its swaps have positive value to the CCP. The variation margin held by the CCP allows the CCP to recover this value upon the clearing member’s failure.

However, the CCP cannot instantly replace or liquidate the failed member’s positions. Suppose, for example, that dealer B in Figure 2.1 had a single swap, originally entered into with dealer A and subsequently cleared through the CCP. If dealer B fails, the CCP has to continue to meet its
 CHAPTER 2. HIDDEN ILLIQUIDITY WITH MULTIPLE CENTRAL COUNTERPARTIES

Figure 2.2: Variation margin covers the value of a clearing member’s swap portfolio at the time of default. Initial margin should cover costs the CCP may incur from the time of default to the completion of the close-out of defaulting member’s portfolio.

payment obligations to dealer A. In order to do so, it needs to replace the position held by B.

Replacing dealer B’s position may take several days. During this time, the market value of the position will continue to move, as illustrated in Figure 2.2. The value of the CCP’s claim on dealer B is also the value of dealer A’s claim on the CCP. An increase in the market value after B’s failure, as illustrated in the figure, represents a loss to the CCP. The initial margin collected by the CCP is intended to protect the CCP from such losses. Moreover, when the CCP transacts it incurs the cost of the bid-ask spread. This cost should also be covered by the initial margin.

For purposes of illustration, Figure 2.2 shows the change in market value and the bid-ask spread as two separate contributions to the total cost incurred by the CCP. In fact, the two sources of loss are entangled. If the CCP transacts more quickly, buying and selling large positions, it will face lower market risk but incur higher liquidity costs through wider bid-ask spreads. It can try to reduce liquidity costs by breaking the failed member’s positions into smaller pieces and replacing them more slowly. In doing so, it faces greater market risk. See Avellaneda and Cont (2013) for an analysis of a CCP’s optimal liquidation problem.

Larger transactions face wider bid-ask spreads per dollar traded. As a consequence, liquidity costs increase superlinearly in the size of a position. Initial margin must then also grow superlinearly to cover liquidity costs with high probability. Hull (2012) calls this the size effect.

We will argue, however, that superlinear margin requirements create an incentive for a dealer
to split trades across multiple CCPs. If the dealer fails, all CCPs through which it trades will need to replace the dealer’s positions at the same time. Their liquidation costs will be driven by the total size of the dealer’s positions across all CCPs. If each CCP bases its margin requirements solely on the trades it clears, without considering trades by the same dealer at other CCPs, it will underestimate the margin it needs to cover liquidation costs.

In addition to variation margin and initial margin, clearing members make contributions to a CCP’s guarantee fund. If a clearing member defaults, any losses exceeding that member’s margin are first absorbed by the member’s guarantee fund contribution, then by CCP capital, and then by the guarantee fund contributions of surviving members. However, initial margin is required to cover liquidation costs with 99 percent confidence under US regulations (Commodity Futures Trading Commission 2011 p. 69368–69370), or 99.5 percent under EMIR (European Commission 2013 p. 56), so our analysis will focus on the adequacy of the margin collected.

Other work on CCP margins includes Cruz Lopez et al. (2013) and Menkveld (2014), both of which focus on dependence between the trades of members of a single CCP. Amini et al. (2013) consider the impact of central clearing on overall systemic risk. Capponi et al. (2014) examine concentration in CCP membership. Biais et al. (2012) study the incentives created by loss mutualization in a CCP. Pirrong (2009) provides a detailed critique of central clearing.

2.3. Hidden Illiquidity

We contrast margin requirements based solely on market risk with requirements that reflect liquidity costs. We assume that the CCP is able to collect variation margin to cover routine daily price changes, so by “margin” we mean initial margin.

We consider a dealer that is a clearing member of \( K \) identical CCPs. Each CCP clears \( m \) types of swaps. These could be credit default swaps (CDS) on different reference entities or with different terms, or they could be different types of interest rate swaps. A vector \( x \in \mathbb{R}^m \) records the dealer’s swap portfolio, with the \( \ell \)th component of \( x \) measuring the size of a dealer’s position in swaps of type \( \ell \), \( \ell = 1, \ldots, m \).
To clear a vector of swaps \( x \), each CCP collects margin \( f(x) \), for some margin function \( f : \mathbb{R}^m \rightarrow \mathbb{R}_+ \) that is common to all CCPs. We allow the dealer to divide the position vector \( x \) arbitrarily among the \( K \) CCPs, clearing the vector \( x_i \) through the \( i \)th CCP, with \( x_1 + \cdots + x_K = x \). To minimize the total margin it needs to post, the dealer solves

\[
\min_{x_1, \ldots, x_K \in \mathbb{R}^m} \left\{ \sum_{i=1}^{K} f(x_i) \right\} \quad \text{subject to } x_1 + \cdots + x_K = x.
\]  

(2.1)

A margin requirement for market risk alone seeks to cover the 99th or 99.5th percentile of a portfolio’s change in market value between the time of default and the end of the close-out period indicated in Figure 2.2, ignoring liquidity costs. The close-out period is typically assumed to be five to ten days. The percentile can be approximated as a multiple of the standard deviation of the change in value over this period. If we let \( \Sigma \) denote the \( m \times m \) covariance matrix of price changes for the \( m \) types of swaps over the close-out period, then we can define a margin requirement to cover market risk by setting

\[
f(x) \triangleq a(x^\top \Sigma x)^{1/2},
\]

(2.2)

for some multiplier \( a \).

With this choice of \( f \), the dealer could optimally clear the entire portfolio \( x \) through a single CCP. Sending \( x/K \) to each CCP is also optimal, but the dealer receives the full benefit of diversification through a single CCP — there is no incentive for the dealer to split the position. Moreover, if the dealer does split the position, each CCP receives the margin it needs to cover the market risk it faces, assuming \( a \) and \( \Sigma \) are chosen correctly.

The margin function in (2.2) is convex but it scales linearly in position size: for any \( x \in \mathbb{R}^m \) and any \( \lambda \geq 0 \), \( f(\lambda x) = \lambda f(x) \). In other words, this \( f \) is positively homogeneous. As discussed in the previous section, the margin function needs to increase superlinearly in position size to cover liquidity costs. For example, consider

\[
f(x) \triangleq a(x^\top \Sigma x)^{\alpha/2}, \quad \alpha > 1.
\]

(2.3)
This margin function yields \( f(\lambda x) = \lambda^\alpha f(x) \) for any \( x \in \mathbb{R}^m \) and \( \lambda \geq 0 \), so it does indeed grow superlinearly along the direction of any portfolio vector \( x \). In this case, solving (2.1) requires clearing an equal portion \( x/K \) through each CCP. Superlinear margin creates an incentive for the dealer to distribute the position as widely as possible. More generally, we have the following contrast between two types of margin functions.

**Proposition 1.** Suppose the function \( f \) satisfies \( f(0) = 0 \). Then:

(i) If \( f \) has the following two properties,

(a) Subadditivity: \( f(x + y) \leq f(x) + f(y) \), for all \( x, y \in \mathbb{R}^m \),

(b) Positive homogeneity: \( f(\lambda x) = \lambda f(x) \), for all \( x \in \mathbb{R}^m \), \( \lambda \geq 0 \),

then any allocation of the form \( x_i = b_i x \), with \( b_1 + \cdots + b_K = 1 \) and \( b_i \geq 0 \), \( i = 1, \ldots, K \), solves (2.1). In particular, clearing the full portfolio \( x \) through a single CCP is optimal.

(ii) If \( f \) is strictly convex, then an equal split \( x_i = x/K \), \( i = 1, \ldots, K \), is the only optimal solution to (2.1). Furthermore, the margin requirement is superlinear in the sense that \( f(\lambda x) > \lambda f(x) \), for all \( x \in \mathbb{R}^m \), \( x \neq 0 \), and all \( \lambda > 0 \).

**Proof.** For (i) observe that if (a) and (b) hold, then

\[
\sum_{i=1}^{K} f(b_i x) = \sum_{i=1}^{K} b_i f(x) = f \left( \sum_{i=1}^{K} x_i \right) \leq \sum_{i=1}^{K} f(x_i),
\]

for any vector \( b \geq 0 \) satisfying \( b_1 + \cdots + b_K = 1 \) and any \( x_1, \ldots, x_K \in \mathbb{R}^m \) feasible for (2.1).

For (ii) if \( f \) is strictly convex, then for any \( x_1 + \cdots + x_K = x \),

\[
\sum_{i=1}^{K} f(x_i) = K \sum_{i=1}^{K} f(x_i)/K \geq K f \left( \sum_{i=1}^{K} x_i/K \right) = K f(x/K).
\]

The inequality is strict when the vectors \( \{x_i\} \) are not identical. \( \blacksquare \)

We can say more if we specialize to a price impact formulation of liquidity costs. Suppose \( f \) takes the form

\[
f(x) \triangleq x^\top F(x),
\]

(2.4)
where \( F : \mathbb{R}^m \to \mathbb{R}^m \) satisfies \( F(0) = 0 \) and is increasing. Interpret \( F(x) \) as the impact on the market price of closing out a position \( x \). Then, \( x^\top F(x) \) is the cost incurred as a result of this price impact on the portfolio \( x \).

Suppose \( f \) in (2.4) is strictly convex, so the dealer optimally splits its position evenly across CCPs. Each CCP collects \( x^\top F(x/K)/K \) in margin. If the dealer fails and all CCPs liquidate their identical positions, the total price impact is \( F(x) \), so each CCP incurs a cost of \( x^\top F(x)/K \), which is larger than the margin it collected. The strict convexity of \( f \) motivates the dealer to “hide” part of its position from each CCP and, moreover, leaves each CCP with insufficient margin.

If all CCPs have the same margin function, they can eliminate the problem by charging

\[
f(x) \triangleq x^\top F(Kx).
\]

In other words, they can precisely compensate for the hidden illiquidity by overstating the cost of liquidating the positions they clear. Clearing regulations\(^2\) require CCPs to back test their margin requirements against historical data. But this simple result implies that a properly calibrated margin model will underestimate the required margin, unless each CCP considers the simultaneous effects of other CCPs in its analysis. Although they are lengthy and detailed, procedures for swap CCPs adopted by the Commodity Futures Trading Commission (2011) and the European Commission (2013) do not address the need to consider the effect of a member’s default at other CCPs, nor is this point noted in the influential CPSS-IOSCO (2012) principles. In Section 2.5.2, we will see that compensating for the effects of other CCPs may be difficult if the CCPs have different margin models and, more importantly, different views on price impact.

In practice, a dealer faces many considerations in making its clearing decisions, beyond the margin minimization decision reflected in (2.1), including the following:

\begin{itemize}
  \item The dealer faces a sequential allocation problem, with new trades arriving over time and old trades maturing.
  \item Both parties to a swap need to agree on where the swap will be cleared, and their optimal
\end{itemize}

allocations may differ. In order to clear at a given CCP, both parties need to be members of the CCP or trade through members of the CCP.

- Clearing members clear trades for clients as well as for their own accounts, and this limits their ability to subdivide positions.

- Dealers may prefer one CCP over another for reasons unrelated to margin requirements, including, for example, lower clearing fees, greater netting benefits, greater CCP capital to absorb losses, better capitalized clearing members, and differences in regulatory jurisdictions.

Currently, when multiple CCPs clear an instrument, one CCP typically clears a large fraction of the overall volume.

These factors may prevent a dealer from allocating trades uniformly to minimize margin but they do not remove the incentive for the dealer to split positions to the extent possible when margin charges are strictly convex.

The precise margin models used by individual CCPs are proprietary. However, the following excerpt from an industry magazine (Ivanov and Underwood, 2011, p. 32) supports our analysis.

The article describes the margin methodology at ICE Clear Credit, the largest CCP for credit default swaps:

“For portfolio/concentration risks, large position requirements, also known as concentration charges, apply to long and short protection positions that exceed predefined notional threshold levels. The concentration charge threshold reflects market depth and liquidity for the specific index family or reference entity. Positions that exceed selected thresholds are subject to additional, exponentially increasing, initial margin requirements. The accelerated initial margin creates the economic incentive to eliminate large positions.”

Whether the model literally uses an exponential margin function or if this term is used informally to refer to a superlinear increase is unclear.

We should also comment on the degree of liquidity in swaps markets. The most liquid interest rate swaps and index CDS are already centrally cleared. As new types of contracts migrate to CCPs,
they are inevitably less liquid, particularly at the outset. Swaptions and inflation swaps have been proposed for central clearing but are far less liquid than standard interest rate swaps. Even among index CDS, off-the-run indices are significantly less liquid than their on-the-run versions. Each index CDS trades at multiple maturities, and liquidity is much lower at maturities other than five years. Chen et al. (2011) provide a detailed analysis of liquidity in CDS transactions using supervisory data. We make some observations using public data.

Figure 2.3 shows the notional amount outstanding and gross market value of CDS from 2005 to 2013, as reported by the Bank for International Settlements. Both measures show declining liquidity in the CDS market following the financial crisis. Higher bank capital requirements for derivatives have contributed to this trend.

Figure 2.4 shows the distribution of the average number of trades per day for all single-name CDS, as reported by the Depository Trust Clearing Corporation. The figure shows data for the first quarter of 2013. The vast majority of contracts trade at most a few times per day.

Figure 2.5 shows the distribution of bid-ask spreads for one-year and five-year CDS, as reported by Markit Group, Ltd. The figures show the bid-ask spreads for all single-name contracts for all days in 2013, except that we dropped the top ten percent (the widest spreads) in both cases. The distribution for five-year contracts shows large spikes near five and ten basis points. For the one-year contracts the spreads are much wider, reflecting the lower liquidity at that maturity.

2.4. Model

We now turn to a setting with $K = 2$ CCPs. We assume that both CCPs clear a universe of $m$ types of swaps. We consider a dealer that is a clearing member of both CCPs and whose portfolio is described by the vector $x \in \mathbb{R}^m$.

We will measure the liquidation costs associated with a portfolio using price impact functions, defined as follows:

**Definition 1 (Price Impact Function).** A price impact function is a function $F : \mathbb{R}^m \to \mathbb{R}^m$ satisfying the following conditions:
CHAPTER 2. HIDDEN ILLIQUIDITY WITH MULTIPLE CENTRAL COUNTERPARTIES

Figure 2.3: Aggregate CDS market statistics (2005–2013).

Figure 2.4: Histogram of average number of daily CDS trades per reference entity (Q1, 2013).
(i) $F(0) = 0$,

(ii) $F$ is differentiable,

(iii) the map $x \mapsto x^\top F(x)$ is strictly convex over $x \in \mathbb{R}^m$.

Here, $F(x)$ captures the vector of price changes that would occur given the liquidation of the portfolio $x$. Specifically, the $\ell$th component of the vector $F(x)$ represents the price change to swap $\ell$ given the liquidation of a portfolio $x$. Condition (i) requires that if no portfolio is liquidated, then there is no price impact. Condition (ii) will be convenient for technical reasons. Condition (iii) requires that the margin costs associated with the liquidation of a portfolio be increasing with the portfolio size.

We assume that the $i$th CCP believes that price impact is given by a price impact function $G_i: \mathbb{R}^m \to \mathbb{R}^m$. We further assume that the $i$th CCP charges margin as a function of only the portfolio $x_i \in \mathbb{R}^m$ cleared there by the clearing member. This is done according to an alternative price impact function $F_i: \mathbb{R}^m \to \mathbb{R}^m$. In other words, for clearing the portfolio $x_i$, the $i$th CCP charges initial margin according to the schedule

$$f_i(x_i) \triangleq x_i^\top F_i(x_i).$$
The clearing member will divide the overall portfolio $x$ in order to minimize the total initial margin outlay. Given margin schedules $\{f_1, f_2\}$, this involves solving the optimization problem

$$f_{\text{eff}}(x) \triangleq \minimize_{x_1, x_2 \in \mathbb{R}^m} \{ f_1(x_1) + f_2(x_2) \mid \text{subject to } x_1 + x_2 = x \}. \quad (2.5)$$

Here, the optimal value $f_{\text{eff}}(x)$ is the effective margin function experienced a clearing member that optimally divides its portfolio across the CCPs.

Given the liquidation of the portfolio $x$, each CCP should ensure that enough margin is collected to cover liquidation costs. Given that the $i$th CCP believes that the price movement from the liquidation of the overall portfolio will be given by the vector $G_i(x)$, CCP $i$ will incur liquidation costs of $x_i^\top G_i(x)$ on the sub-portfolio $x_i$ it clears. Therefore, for CCP $i$ to collect sufficient margin, it is necessary that

$$x_i^\top F_i(x_i) \geq x_i^\top G_i(x). \quad (2.6)$$

We will assume that the market is competitive, so the CCPs seek to collect no more initial margin than is necessary to cover liquidation costs. In other words, we will replace the inequality in (2.6) with equality.

Combining the various considerations described above, we define an equilibrium between the clearing member, which seeks to minimize its margin requirements, and the CCPs, which seek to collect sufficient margin to cover liquidation costs, as follows:

**Definition 2 (Equilibrium).** Given price impact beliefs $G_1, G_2$ for the two CCPs, an equilibrium $(F_1, F_2, x_1, x_2)$ is defined by

1. allocation functions $x_i : \mathbb{R}^m \to \mathbb{R}^m$, for $i \in \{1, 2\}$,

2. price impact functions $F_i : \mathbb{R}^m \to \mathbb{R}^m$, for $i \in \{1, 2\}$,

satisfying, for each portfolio $x \in \mathbb{R}^m$,

1. $(x_1(x), x_2(x))$ is an optimal solution to the clearing member’s problem (2.5),

2. $x_i^\top F_i(x_i) \geq x_i^\top G_i(x)$ for each CCP $i$.
2. each CCP \(i\) collects initial margin to meet its true price impact beliefs, i.e.,
\[
x_i(x) \trans F_i(x_i(x)) = x_i(x) \trans G_i(x), \quad \text{for } i \in \{1, 2\}.
\]

Definition 2 makes explicit the functional dependence of the allocations \(x_1\) and \(x_2\) on the portfolio \(x\). In what follows, we will sometimes suppress this dependence for notational convenience.

### 2.5. Linear Price Impact

We first consider the case of linear price impact functions, where we require that the price impact functions associated with each CCP satisfy

\[
F_i(x) = F_i x, \quad G_i(x) = G_i x,
\]

for some matrices \(F_i, G_i \in \mathbb{R}^{m \times m}\). Without loss of generality, we will require that the matrices \(F_i, G_i\) be symmetric.\(^3\) Moreover, in order to satisfy Part (iii) of Definition 1, we require that \(F_i, G_i \succ 0\), i.e., that the matrices are positive definite.

Given linear price impact (2.7), the total margin charged by each CCP \(i\) takes the form

\[
f_i(x) = x \trans F_i x,
\]

i.e., the CCP margins charged are quadratic in the position cleared. This is a multivariate version of the Kyle (1985) model, in which price impact is linear and the total liquidation costs are quadratic.

A linear price impact model accommodates cross-price impact: the \((k, \ell)\) entry of a linear price impact matrix captures the effect of liquidating the \(\ell\)th instrument on the price of the \(k\)th instrument. Cross-price impact is important in situations where transactions in one swap propagate to the prices of other swaps. This can occur for supply/demand reasons (e.g., when similar in-

\(^3\)For any matrix \(F \in \mathbb{R}^{m \times m}\), \(x \trans F x = x \trans (F + F^\top) x / 2\) for all \(x \in \mathbb{R}^m\). Hence, if a price impact matrix \(F\) is non-symmetric, we can replace it with its symmetrization \((F + F^\top) / 2\) without changing the resulting margin function.
Instruments function as partial substitutes) or for informational reasons (e.g., when the underlying fundamental values of related instruments are correlated). For example, CDS for different firms in the same sector can be impacted by common liquidity or price shocks, as are CDS for the same reference entity across various tenors, or CDS for different series of a common index.

Direct estimation of price impact functions requires detailed transaction data and can be quite challenging. To get a rough indication of the potential for cross-price impact, we can examine comovements in credit default swaps. Figure 2.6 shows the variance explained by the first 10 principal components of the covariance matrices of daily CDS returns for financial institutions (left) and sovereigns (right). In both cases, a relatively small number of principal components explains a significant fraction of total variance. This suggests significant cross-price impact within each sector.

**2.5.1. Equilibrium Characterization**

In the case of linear price impact functions, the following theorem characterizes possible equilibria:

---

See Fleming and Sarkar (2014) for an analysis of the failure resolution of Lehman Brothers, including its cleared swaps.
Theorem 1. A necessary and sufficient condition for the existence of an equilibrium with linear price impact functions is that the two CCPs have common views on market impact, i.e., that $G_1 = G_2 \triangleq G$.

In this case, all equilibria are determined by the symmetric, positive definite solutions $F_1, F_2 \in \mathbb{R}^{m \times m}$ to the equation

$$G^{-1} = F_1^{-1} + F_2^{-1}. \quad (2.8)$$

Theorem 1 generates two important insights. First, in order for an equilibrium to exist, the CCPs must agree on the true price impact $G$. In Section 2.5.2, we will show that different beliefs about the true price impact can create a “race to the bottom” in which one CCP is driven out of the market.

The second insight of Theorem 1 is that the CCPs need not charge the same margin in equilibrium. There are many possible equilibria, corresponding to solutions of (2.8). To interpret (2.8), note that, in the present setting, the clearing member’s problem takes the form

$$f_{\text{eff}}(x) \triangleq \min_{x_1, x_2 \in \mathbb{R}^m} \left\{ x_1^T F_1 x_1 + x_2^T F_2 x_2 \mid \text{subject to } x_1 + x_2 = x \right\}$$

$$= \min_{x_1 \in \mathbb{R}^m} x_1^T F_1 x_1 + (x - x_1)^T F_2 (x - x_1)$$

$$= x^T \left( F_1^{-1} + F_2^{-1} \right)^{-1} x.$$

Under condition (2.8), then, we have that $f_{\text{eff}}(x) = x^T G x$. In other words, the equilibrium condition is equivalent to the requirement that the effective margin experienced by an optimizing clearing member correspond to the margin that would be charged by a single CCP under the common price impact belief $G$.

A special case of this equilibrium would be

$$F_1 \triangleq \frac{G}{\alpha}, \quad F_2 \triangleq \frac{G}{1 - \alpha}, \quad \alpha \in (0, 1).$$

When $\alpha = 1/2$, each CCP charges according to twice its true belief, and each clears half of the clearing member’s portfolio. This corresponds to the equilibrium discussed in Section 2.3.
\( \alpha < \frac{1}{2} \), CCP 1 will attract less than half of the portfolio because it has a higher margin charge, so it needs to compensate more for the part of the portfolio it does not see, which it precisely accomplishes through its higher margin charge.

Notice that, in our setting, \( G^{-1} \Delta p \) is the size of the portfolio required to achieve a price movement \( \Delta p \in \mathbb{R}^m \). In this way, \( G^{-1} \) is analogous to the “market depth” of Kyle (1985). Thus Theorem 1 can be interpreted as follows: in an equilibrium we require that the two CCPs agree on the true market depth, and that the total depth provided by the two CCPs match the true depth.

Further, the operation \((F_1, F_2) \mapsto (F_1^{-1} + F_2^{-1})^{-1}\) is called the “parallel sum” of matrices in Anderson and Duffin (1969) and a subsequent literature. The name is based on an analogy with how resistors combine when connected in parallel in a circuit. To make the analogy in our setting (see Figure 2.7), identify the price impact used by each CCP with resistance, identify the size of the clearing member’s trade with current, and identify the total price impact with voltage.

With more than two CCPs, the obvious extension of (2.8) remains sufficient for an equilibrium. However, we do not know if agreement on the \( G_i \) remains necessary in that case.

### 2.5.2. Race to the Bottom

Theorem 1 establishes that there can be no equilibrium with linear price impact functions if the CCPs have differing beliefs of price impact. In order to provide intuition for why this is the case, it is useful to analyze the best response dynamics between competing CCPs in this setting.

Specifically, consider a discrete time setting indexed by \( t = 0, 1, \ldots \), where CCPs sequentially ...
CHAPTER 2. HIDDEN ILLIQUIDITY WITH MULTIPLE CENTRAL COUNTERPARTIES

update their margin requirements as follows:

1. At time \( t = 0 \), each CCP \( i \) sets margins according to its initial beliefs by setting \( F_i(0) \triangleq G_i \).

2. At each time \( t \geq 0 \), given margins specified by symmetric, positive definite impact matrices \((F_1(t), F_2(t))\):

   (a) The clearing member computes the optimal allocation \((x_1(t), x_2(t))\) by solving (2.5) assuming price impact matrices \((F_1(t), F_2(t))\) and gets

   \[
   x_1(t) = (F_1(t) + F_2(t))^{-1} F_2(t) x, \quad x_2(t) = (F_1(t) + F_2(t))^{-1} F_1(t) x. \tag{2.9}
   \]

   (b) Given the clearing member’s allocation \((x_1(t), x_2(t))\), CCP 1 sets its price impact matrix \( F_1(t+1) \) for the next period to ensure that it would get sufficient margin for the present allocation by solving

   \[
   x_1(t) \top G_1 x = x_1(t) \top F_1(t+1) x_1(t).
   \]

   Using (2.9), we have that

   \[
   x \top F_2(t)(F_1(t)+F_2(t))^{-1} G_1 x = x \top F_2(t)(F_1(t)+F_2(t))^{-1} F_1(t+1)(F_1(t)+F_2(t))^{-1} F_2(t) x.
   \]

   Since this must hold for all \( x \), and since we require that \( F_i(t+1) \) be symmetric, it must be the case that

   \[
   F_1(t+1) = \frac{1}{2} \left[ G_1 F_2(t)^{-1} (F_1(t)+F_2(t)) + (F_1(t)+F_2(t)) F_2(t)^{-1} G_1 \right]. \tag{2.10}
   \]

   Similarly, for CCP 2,

   \[
   F_2(t+1) = \frac{1}{2} \left[ G_2 F_1(t)^{-1} (F_1(t)+F_2(t)) + (F_1(t)+F_2(t)) F_1(t)^{-1} G_2 \right]. \tag{2.11}
   \]

   First, consider the scalar, single-instrument case \((m=1)\). Suppose the CCPs disagree in their price impact beliefs and, without loss of generality, \( G_1 > G_2 \), so CCP 1 believes the price impact
CHAPTER 2. HIDDEN ILLIQUIDITY WITH MULTIPLE CENTRAL COUNTERPARTIES

is greater than CCP 2 does. Then, for \( t \geq 1 \), the best response dynamics yield

\[
\frac{F_2(t)}{F_1(t)} = \frac{G_2 F_2(t-1)}{G_2 F_1(t-1)} = \left(\frac{G_2}{G_1}\right)^{t+1},
\]

where the first equality follows from (2.10)–(2.11) and the second equality follows by induction. As \( t \to \infty \), we have that \( F_2(t)/F_1(t) \to 0 \), and this implies that

\[
x_1(t) = (1 + F_1(t)/F_2(t))^{-1} x \to 0, \quad x_2(t) = (1 + F_2(t)/F_1(t))^{-1} x \to x.
\]

In other words, asymptotically, CCP 2 clears a larger fraction of the position by charging lower margin. Due to the convexity of the quadratic total margin function, this forces CCP 1 to charge increasingly higher margins in order to cover liquidation costs. Asymptotically, CCP 1 has an infinite initial margin and is thus driven out of the clearing market. We call this a “race to the bottom” because the CCP with the lower price impact ultimately determines margin costs for the entire market.

More generally, we can expand our discussion above to the multidimensional case:

**Proposition 2.** Suppose that the CCPs differ in their price impact belief matrices \( G_1, G_2 \in \mathbb{R}^{m \times m} \). Then:

(i) the matrices \( (F_1(t + 1), F_2(t + 1)) \) defined in (2.10)–(2.11) are positive definite for all \( t \geq 0 \),

(ii) if the spectral radius of \( G_1^{-1}G_2 \) is strictly less than 1, as \( t \to \infty \),

\[
F_2(t)F_1(t)^{-1} \to 0, \quad x_1(t) \to 0, \quad x_2(t) \to x.
\]

Part (i) shows that the best response dynamics suggested earlier are well-defined for all \( t \geq 0 \). Part (ii) states that, if the price impact beliefs of CCP 2 are “smaller” (in the sense of the spectral radius of their ratio) than those of CCP 1, CCP 1 will ultimately be driven out of the clearing market. If \( G_1 \succ G_2 \) in the positive definite ordering, i.e., if the margin required by the matrix \( G_1 \) dominates that of \( G_2 \) for every portfolio, then the spectral radius of \( G_1^{-1}G_2 \) must be less than 1.
2.5.3. Partitioned Clearing

Thus far, we have assumed that both CCPs clear the entire universe of available instruments. But the first decision a CCP makes is which types of instruments to clear. We now extend Theorem 1 by expanding the strategy space for each CCP to include the choice of instruments to clear as well as the initial margin to charge. We continue to suppose that each CCP’s belief about true price impact is given by a symmetric, positive definite matrix \( G_i \in \mathbb{R}^{m \times m} \), where \( m \) is the total number of securities available for clearing.

We assume that a CCP clears all linear combinations of the securities it clears, and does not clear linear combinations that include securities that it does not clear. So, the choice of a subset of security types is a choice of subspace of \( \mathbb{R}^m \). Write \( m = m_1 + m_2 + m_3 \), where:

\[
\begin{align*}
    m_1 &= \text{number of security types cleared only by CCP 1}, \\
    m_2 &= \text{number of security types cleared by both CCPs}, \\
    m_3 &= \text{number of security types cleared only by CCP 2}.
\end{align*}
\]

We also assume that the security types are numbered in this order, so that the first \( m_1 \) types are cleared only by CCP 1, and so on.

The margin matrices \( F_1 \) and \( F_2 \) have dimensions \((m_1 + m_2) \times (m_1 + m_2)\) and \((m_2 + m_3) \times (m_2 + m_3)\), respectively. Denote by \( P_1 \in \mathbb{R}^{(m_1 + m_2) \times m} \) the matrix of the projection of \( \mathbb{R}^m \) onto the first \( m_1 + m_2 \) coordinates corresponding to swap types cleared by CCP 1. Similarly, denote by \( P_2 \in \mathbb{R}^{(m_2 + m_3) \times m} \) the matrix of the projection onto the last \( m_2 + m_3 \) coordinates corresponding to swap types cleared by CCP 2. Finally, let the notation \( 0_k \in \mathbb{R}^k \) denote a zero row vector of length \( k \), and the notation \((x_1^T, 0_{m_3})\) and \((0_{m_1}, x_2^T)\) denote the lifting of vectors \( x_1 \in \mathbb{R}^{m_1 + m_2} \) and \( x_2 \in \mathbb{R}^{m_2 + m_3} \) from the subspaces cleared by the two CCPs to full-length portfolio vectors.

With the above notation in place, we can make the following definition:

---

\(^5\)Without loss of generality, securities cleared by neither CCP can be excluded from consideration.
CHAPTER 2. HIDDEN ILLIQUIDITY WITH MULTIPLE CENTRAL COUNTERPARTIES

Definition 3 (Partitioned Equilibrium with Linear Price Impact). Given price impact belief matrices $G_1, G_2 \in \mathbb{R}^m$ for the two CCPs, a partitioned equilibrium is defined by

1. a partition $(m_1, m_2, m_3)$ of the $m$ swap types,

2. allocation functions $x_1: \mathbb{R}^m \to \mathbb{R}^{m_1+m_2}$ and $x_2: \mathbb{R}^m \to \mathbb{R}^{m_2+m_3},$

3. price impact margin matrices $F_1 \in \mathbb{R}^{m_1+m_2}$, $F_2 \in \mathbb{R}^{m_2+m_3},$

satisfying, for each portfolio $x \in \mathbb{R}^m$,

1. $(x_1(x), x_2(x))$ is an optimal solution to the clearing member’s optimization problem

\[
\min_{x_1 \in \mathbb{R}^{m_1+m_2}, x_2 \in \mathbb{R}^{m_2+m_3}} \left\{ x_1^T F_1 x_1 + x_2^T F_2 x_2 \left| \begin{array}{c} \text{subject to} \ (x_1^T, 0_{m_3}) + (0_{m_1}, x_2^T) = x \end{array} \right. \right\}, \quad (2.12)
\]

2. each CCP $i$ collects liquidity margin based on its true price impact beliefs, i.e.,

\[
x_1(x)^T F_1 x_1(x) = x_1(x)^T P_1 G_1 x, \quad x_2(x)^T F_2 x_2(x) = x_2(x)^T P_2 G_2 x. \quad (2.13)
\]

The following theorem characterizes partitioned equilibria:

Theorem 2. A necessary and sufficient condition for a partitioned equilibrium with linear price impact is that the price impact belief matrices $G_1, G_2$ have a common block diagonal structure

\[
G_i = \begin{pmatrix} G_i(1,1) & & \\ & G_i(2,2) & \\ & & G_i(3,3) \end{pmatrix}, \quad i \in \{1, 2\}, \quad (2.14)
\]

with $G_i(1,1) \in \mathbb{R}^{m_1 \times m_1}$, $G_i(2,2) \in \mathbb{R}^{m_2 \times m_2}$, $G_i(3,3) \in \mathbb{R}^{m_3 \times m_3}$, where the submatrices satisfy

\[
G_1(2,2) = G_2(2,2) \triangleq G(2,2). \quad (2.15)
\]

In this case, CCP 1 clears the first $m_1 + m_2$ swap types, CCP 2 clears the last $m_2 + m_3$ swap
types, and they choose margin matrices

\[
F_1 = \begin{pmatrix} G_1(1,1) \\ F_1(2,2) \end{pmatrix}, \quad F_2 = \begin{pmatrix} F_2(2,2) \\ G_2(3,3) \end{pmatrix}, \tag{2.16}
\]

for any symmetric, positive definite matrices \(F_1(2,2), F_2(2,2) \in \mathbb{R}^{m_2 \times m_2}\) satisfying

\[
F_1(2,2)^{-1} + F_2(2,2)^{-1} = G(2,2)^{-1}. \tag{2.17}
\]

Theorem 2 establishes a number of requirements for partitioned equilibria. Condition (2.15) implies that the two CCPs need to have common beliefs on price impact for the instruments they both clear. The block structure requirement in (2.14) implies that an instrument cleared by only a single CCP cannot have any cross-price impact with any swap clear by the other CCP.

Next, we consider a refinement of the partitioned equilibrium of Definition 3:

**Definition 4 (Stable Equilibrium).** A partitioned equilibrium \((m_1, m_2, m_3, F_1, F_2, x_1, x_2)\) is called stable if it is undominated in the sense that there exists no other equilibrium \((\tilde{m}_1, \tilde{m}_2, \tilde{m}_3, \tilde{F}_1, \tilde{F}_2, \tilde{x}_1, \tilde{x}_2)\) such that

\[
x_1(x)^\top F_1 x_1(x) + x_2(x)^\top F_2 x_2(x) \geq \tilde{x}_1(x)^\top \tilde{F}_1 \tilde{x}_1(x) + \tilde{x}_2(x)^\top \tilde{F}_2 \tilde{x}_2(x), \quad \text{for all } x \in \mathbb{R}^m,
\]

and that the inequality holds strictly for some \(x \in \mathbb{R}^m\).

An equilibrium with the block structure (2.14)–(2.15) may fail to be stable in the following way: Suppose that among the first \(m_1\) instruments (those cleared only by CCP 1) there is some instrument with index \(j\) for which \(G_2(j, j) < G_1(j, j)\), and suppose that \(G_1(j, k) = G_2(j, k) = 0\), for all \(k \neq j\). Then we can construct another equilibrium by moving instrument \(j\) from the set cleared only by CCP 1 to the set cleared only by CCP 2 and reduce the total margin charged.

The following result provides a sufficient condition for stability:
CHAPTER 2. HIDDEN ILLIQUIDITY WITH MULTIPLE CENTRAL COUNTERPARTIES

**Proposition 3 (Stable Partitioned Equilibrium).** A partitioned equilibrium is stable if

\[ G_1(1, 1) \preceq G_2(1, 1), \quad G_1(3, 3) \succeq G_2(3, 3), \]

in the positive definite order.

Proposition 3 states that an equilibrium is stable if each CCP collects less margin for the set of instruments it clears exclusively than the other CCP would. For example, if \( G_1 \succ G_2 \), then having CCP 2 clear all positions alone is the unique stable equilibrium.

2.6. Adding Uncertainty

To this point, we have assumed a completely deterministic model in which each CCP is able to infer a clearing member’s full portfolio vector \( x \) from the portion cleared by that CCP by effectively inverting the solution to the clearing member’s problem. In this section, we extend our results by adding uncertainty. We consider two forms of uncertainty: uncertainty in the CCPs’ inferences about the clearing member’s portfolio, and uncertainty in the CCPs’ beliefs about the true price impact.

To incorporate uncertainty in the CCPs’ beliefs, we take the price impact matrices \( G_1 \) and \( G_2 \) to be stochastic. We assume that these matrices are almost surely symmetric and positive definite. The same is then true of their expectations \( \mathbb{E}[G_i], \ i \in \{1, 2\} \).

We use a simple model of the CCP’s uncertainty about the clearing member’s portfolio. We suppose that when CCP \( i \) clears a portion \( x_i \) of the full portfolio \( x \), it forms an estimate

\[ \hat{x}_i = x + \epsilon_i, \]

of the full portfolio, with \( \mathbb{E}[\epsilon_i] = 0, \ i \in \{1, 2\} \). In other words, a CCP cannot perfectly infer the clearing member’s full portfolio, but it can form an unbiased estimate \( \hat{x}_i \) of the full portfolio.

This model provides a reduced-form description of the many sources of uncertainty that would in practice prevent a CCP from reverse engineering a clearing member’s portfolio. In particular, a CCP
CHAPTER 2. HIDDEN ILLIQUIDITY WITH MULTIPLE CENTRAL COUNTERPARTIES

may not have perfect information about its competitors’ margin functions, and considerations other than margin minimization may influence the clearing member’s allocation. Our key assumption is that these factors do not lead the CCP to systematically misjudge the clearing member’s full portfolio. A more complete model would generate the $\epsilon_i$ endogenously from a more fundamental description of uncertainty. In the absence of such an extension, we proceed with the reduced-form model, recognizing its limitations.

To extend our earlier results to include uncertainty, we suppose that each CCP sets its margin function to collect sufficient margin in expectation. More precisely, we define an equilibrium as in Definition 2 but replacing the last condition given there with the following condition:

$$x_i^\top F_i(x_i) = E \left[ x_i^\top G_i(\hat{x}_i) \right], \quad i \in \{1, 2\}. \quad (2.19)$$

**Proposition 4.** Suppose that for each CCP $i$, $\epsilon_i$ and $G_i$ are uncorrelated. Then a necessary and sufficient condition for equilibrium with linear price impact is that the two CCPs have common views on the mean market impact, i.e., that $E[G_1] = E[G_2] \triangleq G$.

In this case, all equilibria are determined by the symmetric, positive definite solutions $F_1, F_2 \in \mathbb{R}^{m \times m}$ to the equation

$$G^{-1} = F_1^{-1} + F_2^{-1}. $$

**Proof.** Because $G_i$ is uncorrelated with $\epsilon_i$, we have

$$E \left[ x_i^\top G_i(\hat{x}_i) \right] = E \left[ x_i^\top G_i(x + \epsilon_i) \right] = x_i^\top E[G_i](x + E[\epsilon_i]) = x_i^\top E[G_i]x. $$

Thus, (2.19) reduces to $x_i^\top F_i(x_i) = x_i^\top Gx$. The result now follows from Theorem 1. \qed

2.7. A Single Instrument with General Price Impact

In general, it is not easy to solve for equilibrium under nonlinear price impact models. It is, however, possible to characterize the scalar case. In this section, we specialize to the case of a single instrument ($m = 1$) in which the portfolio $x \in \mathbb{R}$ is scalar. Each CCP $i$ has price impact
belief $G_i(x)$ and margin function $f_i(x) = xF_i(x)$.

Suppose that $(F_1, F_2, x_1, x_2)$ form an equilibrium according to Definition 2. Then, first order necessary and sufficient conditions for the clearing member’s problem (2.5) are that

$$F_1(x_1) + x_1F'_1(x_1) = F_2(x_2) + x_2F'_2(x_2).$$

(2.20)

Also, the sufficient margin condition is equivalent to

$$F_i(x_i) = G_i(x_i).$$

(2.21)

In the following, we use

$$f^*(x) \triangleq \sup_{y \in \mathbb{R}} \{xy - f(y)\}$$

to denote the convex conjugate of a function of $f$ on $\mathbb{R}$.

**Theorem 3.** (i) If the CCPs have common beliefs $G_1 = G_2 \triangleq G$, then an equilibrium exists. All equilibria result in proportional allocations $x_1 = \alpha x$ and $x_2 = (1 - \alpha)x$, for some $\alpha \in (0, 1)$, and

$$F_1(x) = G(x/\alpha), \quad F_2(x) = G(x/(1 - \alpha)).$$

(ii) If an equilibrium with proportional allocations exists, then the CCPs have common beliefs $G_1 = G_2$.

(iii) In any equilibrium with common beliefs, $f_{\text{eff}}(x) = g(x) \triangleq xG(x)$, meaning that the effective margin equals the shared view on the required margin. Moreover, the common belief can be recovered from the individual margin functions through the relation

$$g = (f_1^* + f_2^*)^*.$$  

(2.22)

**Proof.** Proof [1] For the existence of an equilibrium, it suffices to show that

$$x_1 = x_2 = x/2, \quad F_1(x) = F_2(x) = G(2x),$$

42
is an equilibrium. This follows from the fact that (2.20) and (2.21) hold in this case.

Next, we establish that all equilibria result in proportional allocations. If $G_1 = G_2 \triangleq G$, (2.21) implies $F_1(x_1) = F_2(x_2)$, so (2.20) implies

$$x_1 F'_1(x_1) = x_2 F'_2(x_2).$$

(2.23)

Differentiating (2.21) with respect to $x$, we get that

$$F'_i(x_i)x'_i = G'_i(x).$$

This yields

$$F'_1(x_1)x'_1 = F'_2(x_2)x'_2.$$  

(2.24)

This implies that $x_1$ and $x_2$ are strictly increasing and therefore strictly positive for $x > 0$. For $x > 0$, combining the (2.23) and (2.24), we get

$$\frac{x'_1}{x_1} = \frac{x'_2}{x_2}.$$  

So $x_2 = cx_1$ for some constant $c > 0$, and the claim holds with $\alpha \triangleq 1/(1 + c)$.

(ii) Suppose $x_1 = \alpha x$ and $x_2 = (1 - \alpha)x$, and define

$$h(x) \triangleq F_1(x_1) - F_2(x_2) = F_1(\alpha x) - F_2((1 - \alpha)x).$$

Differentiating this with respect to $x$, we have

$$h'(x) = \alpha F'_1(\alpha x) - (1 - \alpha)F'_2((1 - \alpha)x).$$

But using the first-order condition (2.20), we can write $h$ as

$$h(x) = -x_1 F'_1(x_1) + x_2 F'_2(x_2) = -\alpha x F'_1(\alpha x) + (1 - \alpha) x F'_2((1 - \alpha)x) = -x h'(x).$$
Then, \( h(x) + xh'(x) = 0 \), which means that \( xh(x) \) is a constant, so we must have \( h(x) \equiv 0 \). In other words, \( F_1(x_1) = F_2(x_2) \), and thus \( G_1 = G_2 \) by (2.21).

(iii) We take the conjugate of the effective margin \( f_{\text{eff}} \) in (2.5). Because \( f_i \) is convex and continuous, we have, by Rockafellar (1997) Theorem 16.4,

\[
f_{\text{eff}}^* = (f_1 \square f_2)^* = f_1^* + f_2^*.
\]

The infimal convolution of convex, continuous functions is also convex and continuous so

\[
f_{\text{eff}} = f_{\text{eff}}^{**} = (f_1^* + f_2^*)^*,
\]

using Theorem 12.2 and Corollary 12.2.1 of Rockafellar (1997). Now, notice that in equilibrium we always have

\[
f_1(x_1) + f_2(x_2) = x_1F_1(x_1) + x_2F_2(x_2) = x_1G_1(x) + x_2G_2(x) = xG(x).
\]

Then by the definition of infimal convolution, we have \( g(x) = xG(x) = f_{\text{eff}}(x) \).

In the case of linear price impact, the total margin functions \( f_1, f_2 \) are quadratic, and (2.22) leads to

\[
g^*(x) = G^{-1}x^2 = f_{\text{eff}}^*(x) = F_1^{-1}x^2 + F_2^{-1}x^2,
\]

for all \( x \in R \), so that

\[
G^{-1} = F_1^{-1} + F_2^{-1}.
\]

This is just the scalar case of Theorem (1).

As another example, suppose the price impact function takes the form \( G(x) \triangleq cx^\beta \), given an exponent \( \beta > 0 \). Theorem (3) yields an equilibrium with \( F_i(x) \triangleq b_ix^\beta, i \in \{1, 2\} \) so long as

\[
b_1^{-1/\beta} + b_2^{-1/\beta} = c^{-1/\beta}.
\]
CHAPTER 2. HIDDEN ILLIQUIDITY WITH MULTIPLE CENTRAL COUNTERPARTIES

To see this, first notice that \( g(x) = cx^{\beta+1} \), hence

\[
g^*(y) = c^{-1/\beta} x^{1+1/\beta} (\beta + 1)^{-1/\beta} \frac{1}{1 + 1/\beta}.
\]

Similarly,

\[
f_i^*(y) = b_i^{-1/\beta} x^{1+1/\beta} (\beta + 1)^{-1/\beta} \frac{1}{1 + 1/\beta}.
\]

Then (2.26) is just a result of applying (2.22). Note that (2.25) is a special case of (2.26) with \( \beta = 1 \).

Theorem 3 leaves open the possibility of an equilibrium in which the CCPs have different views, which would require that the allocations \( x_1, x_2 \) not be proportional.

2.8. Implications and Concluding Remarks

Our analysis has relied on simplifying assumptions and a stylized model of the complex decisions faced by central counterparties and their clearing members. Nevertheless, this analysis has practical implications for the functioning of derivatives markets.

- A CCP’s initial margin requirements should reflect liquidity costs as well as market risk. Liquidity costs increase more than proportionally with position size, so margin requirements should as well. This is a premise of our analysis but it bears repeating. In responding to comments on its proposed rules, the CFTC specifically declined recommendations requiring that position concentration be factored into margin calculations, leaving the matter to the discretion of each CCP; (see Commodity Futures Trading Commission 2011, p. 69366).

- In incorporating liquidity costs into margin requirements, a CCP also needs to consider a clearing member’s positions at other CCPs. If the clearing member defaults, its positions at all CCPs will hit the market simultaneously, so price impact is determined by the clearing member’s combined positions, not its position at a single CCP. Moreover, superlinear margin charges designed to capture liquidity costs create an incentive for clearing members to split positions across CCPs, thus amplifying the effect of hidden illiquidity.
CHAPTER 2. HIDDEN ILLIQUIDITY WITH MULTIPLE CENTRAL COUNTERPARTIES

- To counteract this effect, CCPs and clearing members need to share information about positions across CCPs. If this proves infeasible, given the sensitivity of the information, an alternative approach would be for each CCP to make a conservative assumption about a clearing member’s positions at other CCPs (with a correspondingly conservative margin charge) and create a positive incentive for clearing members to provide this information by offering a potential margin reduction in exchange. A CCP could make a conservative assumption by comparing the positions in a contract it clears with the total outstanding positions in that contract across all participants and CCPs. This type of aggregate data is collected by swap data repositories, as mandated by the Dodd-Frank Act.

- Our analysis also points to the need for CCPs to share information about liquidation costs. The relevant costs would be incurred at the failure of a major swaps dealer and are not easily gleaned from historical data. To better estimate price impacts, CCPs could require their clearing members to regularly provide prices and quantities at which they are committed to buy or sell upon the default of another member.

- A CCP is required to test its default management process, through which a defaulting member’s positions are unwound, at least annually. These default management drills should explicitly account for the actions of other CCPs directly affected by the same member’s default.

- Market participants and regulators have recently called for standardized stress tests for CCPs. Our analysis points to the need for each CCP’s stress scenarios to include the actions of other CCPs. This would be in contrast to the current regulatory stress tests for banks, which treat each bank in isolation.

These recommendations are not necessarily easy to implement. Each of these steps requires further research.
Chapter 3

Portfolio Liquidity Estimation and Optimal Execution

3.1. Introduction

In portfolio management, liquidity is important since the value of a portfolio depends on its ability to convert into cash, especially in a time of distress. Liquidity is even more important for active investors and asset managers who need to unwind significant positions on a frequent basis in order to profit from trading on dynamic predictions of asset returns, as such trading activities could incur huge liquidity costs, especially when the position is large.

We believe that liquidity should be measured at the portfolio level across multiple assets simultaneously, instead of the level of single assets. The reasons are two folds. First of all, many portfolio transitions include trading more than one asset. The simplest example would be the open end funds, whenever they get an inflow or outflow, they have to in effect trade portfolios if they want to maintain proportional holdings. Second, asset prices are often correlated. If you measure liquidity at the level of single assets, you are essentially ignoring the this inter-correlation among assets together with the potential savings by trading correlated assets. Even if you are only trading a single asset, to the extent that optimal execution is about the trade-off of risk versus actual returns, you can potentially do it significantly more cheaply if you are allowed to trade other assets
CHAPTER 3. PORTFOLIO LIQUIDITY ESTIMATION AND OPTIMAL EXECUTION

in a portfolio approach.

Yet there has been a wide disconnect from these intuitions and the practice of liquidation, which still relies primarily on the single-asset optimal execution framework. One reason for this disconnect is that in practice the decision and execution of portfolio liquidation are often separate. Given the challenging task of correctly estimating liquidity costs, executions are often conducted by a specialized team or outsourced to a third party. As a result, the decision of liquidation is made by the portfolio management branch, while the execution are done by a trader or an algorithmic trading system which typically do not have the authority to hedge the liquidation process by establishing new positions.

In this chapter, we propose a tractable multi-dimensional generalization of the Amgren-Chriss model. Our model is built on previous work that allow trading correlated assets such as [Kim 2014]. However, beyond these work, we incorporate the trading of liquid bundles such as ETFs. Our work provides analysis on the underline drivers of the liquidity costs. We specialize our results to the factor model where correlations of returns are driven by common factors. We show the liquidity cost is primarily driven by idiosyncratic risk in the large universe asymptotic regime. Here, large universe refers to the case where there are many assets relative to the number of common underlying factors (see detailed discussion in Section 3.4). This is consistent with the setup of the Arbitrage Pricing Theory first developed by [Ross 1976]. The intuition is that the market risk of the portfolio can be hedged with little costs given the availability of a huge number of assets. Another key question is how the inclusion of standardized liquid bundles affect the optimal liquidity cost in the “large universe” asymptotic regime. By considering a degenerate problem where we only want to liquidate one asset, we show that the benefit of hedging with liquid bundles is essentially equivalent to increasing the liquidity of the individual asset. For the non-asymptotic cases, we manage to provide a good approximation of liquidity cost by exploiting the structural properties we find. In addition, we obtained a bound on the difference between this approximation and exact solutions, and related to the bound to the structural properties of the covariance matrix of asset prices.

The rest of the paper is organized as follows. In Section 3.2 we present our model and characterize the solution of the resulting optimal execution problem. In Section 3.3 we specialize our
results to settings with separable transaction costs that are of particular interest. In Section 3.4, we introduce the large-universe asymptotic regime and establish our main structural results. In Section 3.5, we provide empirical examples calibrated to market data. Section 3.6 concludes. The proofs are provided in the appendix.

3.2. Model

In this section, we describe our general model setup and characterize the solution of the resulting optimization problem.

3.2.1. Setup

**Portfolio and trading strategies.** Consider an agent who wishes to liquidate efficiently a portfolio consisting of positions in up to \( n \) assets. The agent’s initial holdings are specified by the vector \( q \in \mathbb{R}^n \), where component \( q_i \) represents the initial position in asset \( i \) denominated in shares. In order to liquidate this portfolio, the agent can trade \( m \geq n \) possible liquid instruments. The vector \( y_i \in \mathbb{R}^n \) specifies the composition of the \( i \)th instrument in terms of shares of underlying assets. That is, selling one unit of the \( i \)th instrument results in the agent’s portfolio components being adjusted according to the vector \( y_i \). Denote by \( Y \triangleq [y_1, y_2, ..., y_m] \in \mathbb{R}^{n \times m} \) the liquidation matrix that characterizes the available instruments.

In the simplest case, the agent is only allowed to directly trade the underlying assets. Then, the tradeable instruments correspond to the underlying assets \( (n = m) \) and \( Y = I \); i.e., the liquidity matrix is the identity matrix. More generally, our model supports tradeable instruments that are not necessarily individual assets, but can be liquid bundles that are essentially portfolios that can be traded directly. As was discussed in Section 1.2, examples of such liquid bundles include exchange traded funds (ETFs), credit default swap (CDS) indices, and tradeable futures spreads. As an example, consider the following:

\[1\] Strictly speaking, an ETF may not be not exactly equivalent or fungable to its underlying portfolio, but we will assume the existence of efficient creation or redemption mechanisms that make them equivalent for our modeling purposes.
Example 1 (Two-asset ETF). Suppose an agent starts with a portfolio consisting of two stocks, and can trade those stocks directly. In addition, suppose that there exists an ETF of a portfolio consisting of one share of each stock. In this case, the liquidation matrix is given by

\[ Y = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}. \]

We will make the assumption that rank\((Y) = n\), i.e., that \(Y\) is full rank, so that any initial portfolio in \(q \in \mathbb{R}^n\) can be liquidated with the instruments available.

Given a liquidation matrix \(Y\), a trading strategy is characterized by the rate at which each of the liquid instruments (the columns of \(Y\)) are bought or sold. Specifically, a trading strategy is defined by the control process \(u \in L^1([0, \infty); \mathbb{R}^m)\), where \(u_i(t)\) represents the rate at (in shares per unit time) at which instrument \(i\) is traded at time \(t\). We adopt the convention that positive trading rates correspond to selling, while negative trading rates correspond to buying. Given the control \(u\) and the initial position \(q\), the evolution of position over time is given by the position process \(x \in C([0, \infty); \mathbb{R}^n)\), where

\[ x(0) = q, \quad \dot{x}(t) = -Yu(t), \quad \forall \ t \geq 0. \]

Equivalently,

\[ x(t) = q - \int_0^t Yu(s) \, ds, \quad \forall \ t \geq 0. \]

Trading constraints. We consider a constrained liquidity setting where the trading rate of each instrument is bounded according to

\[ |u_i(t)| \leq \gamma_i, \quad \forall \ 1 \leq i \leq m, \ t \geq 0. \quad (3.1) \]

Here \(\gamma_i > 0\) is a bound on the absolute trading rate of instrument \(i\). Such restrictions on the trading rate are very common in practice, for several reasons. First, an excessive trading rate will almost certainly lead to unfavorable execution prices due to market impact. We will momentarily
introduce transaction costs that depend on the trading rate. However, at very high trading rates, the
agent will create a significant supply-demand imbalance in the market, and hence transaction costs
will be dominated by effects such as information leakage and are difficult to estimate. Empirical
evidence on information leakage of large trades is found by [Van Kervel and Menkveld (2015)] where
the authors show that the high frequency traders “prey” on orders that are large. On the other
hand, transaction costs for very low trading rates will be dominated by observable quantities such
as the bid-ask spread and easy to estimate. Hence, transaction cost models typically are accurate
only for a restricted range of trading rates, and the constraint (3.1) can enforce this range. Finally,
observe that constraints of the form (3.1) are very common in practice, and can be easily calibrated
through market parameters. Typically, one might restrict the trading rate to a certain percentage
of the future predicted overall market trading volume for a particular instrument.

**Transaction costs.** We allow for the possibility of trades to be associated with transaction costs.
Such costs may arise from, say, commissions or trading fees, the bid-ask spread, or the distortion
of market prices caused by the agent’s trading. In all of these cases, transaction costs are related
to the trading rate. For example, costs associated per share commissions or the bid-ask spread
accumulate as a linear function of the trading rate. Market impact may take a more complicated
form, but will still be an increasing function of the trading rate.

Though the sources of transaction costs vary, they are all closely related to the trading rate. In
particular, if we look at the rate of transaction cost accumulation, the contributions of commission
fees and the costs from bid-ask spreads are linear as a function of the trading rate, whereas market
impact costs may take a more complicated form such as that studied in [Kyle and Obizhaeva
(2016b)]. Here, we will not seek to decompose the transaction costs and will describe the total rate
of transaction cost accumulation with some functional \( f(\cdot) \) of the trading rate \( u \). We assume that
\( f: \mathbb{R}^n \to \mathbb{R}_+ \) is a non-negative convex function that is symmetric around 0, i.e., \( f(u) = f(-u) \),
for all \( u \in \mathbb{R}^n \). Further, we assume that no costs are incurred by not trading, i.e., \( f(0) = 0 \). By
making these assumptions, we are essentially focusing on only the temporary (or instantaneous)
market impact, which depends only on how fast you trade. We are not considering permanent or
transient market impact, which features the impact of current trade on future execution prices.
CHAPTER 3. PORTFOLIO LIQUIDITY ESTIMATION AND OPTIMAL EXECUTION

Price dynamics. The evolution of the price dynamics is typically determined by a predictable drift component and a random noise component. Since the liquidation process typically happens in a short time horizon, we will neglect the drift and focus only on the unpredictable variations. Specifically, we assume that the prices of the \( n \) assets \( (S(t) \in \mathbb{R}^n) \) follow a multidimensional Brownian motion given by

\[
dS(t) = \Sigma^{\frac{1}{2}} dW(t),
\]

where \( W(t) \in \mathbb{R}^n \) is an \( n \)-dimensional standard Brownian motion, and \( \Sigma \in \mathbb{R}^{n \times n} \) is a symmetric positive definite matrix that characterizes the covariance structure of \( W(t) \). We will also assume that there are no tracking errors for the liquid bundles. As a result, the price process of any instrument \( y_i \) is given by \( y_i^\top S(t) \). We will also make the assumption that the covariance matrix \( \Sigma \) is constant over the period of liquidation. This may be a reasonable approximation since the liquidation process we are considering typically takes a short time horizon ranging from hours to days. It is expected that the covariance structure will not change dramatically over such a short time horizon.

Portfolio value and risk. We now discuss the profit and loss resulting from the liquidation process. For any liquidation process defined by \( (x, u) \), let \( IS_t \) be the implementation shortfall from liquidating the portfolio up to time \( t \). This is defined to be the difference between the value of the initial portfolio at time 0 and the value of the remaining portfolio at time \( t \) (along with any intermediate cashflows resulting from trading between time 0 and time \( t \)). That is,

\[
IS_t \triangleq \int_0^t (S(0) - S(s))^\top dx(s) + \int_0^t f(u(s)) \, ds
= -\int_0^t x(s)^\top dS(s) + \int_0^t f(u(s)) \, ds.
\]

The first term represents the total effect of price changes during the liquidation process up to time \( t \). The second term is the loss due to transaction costs.

The expected value of \( IS_t \) takes the form

\[
E[IS_t] = \int_0^t f(u(s)) \, ds.
\]
Notice that, by construction $x \in C([0, \infty); \mathbb{R}^n)$. It follows immediately that

$$\int_0^t x(s)^\top x(s) \, ds < \infty. \quad (3.5)$$

Then, by Itô’s isometry, we have

$$\text{Var}(IS_t) = \int_0^t x(s)^\top \Sigma x(s) \, ds. \quad (3.6)$$

Let $IS \triangleq \lim_{t \to \infty} IS_t$ denote the implementation shortfall incurred over the entire liquidation process; we have

$$\mathbb{E}[IS] = \int_0^\infty f(u(t)) \, dt, \quad \text{Var}(IS) = \int_0^\infty x(t)^\top \Sigma x(t) \, dt. \quad (3.7)$$

The mean of $IS$ is simply the total transaction costs associated with the liquidation process. The variance of $IS$ provides us with a natural measure of market risk during the liquidation process.

**Optimization problem.** The optimal liquidation problem can be formulated by minimizing the expected implementation shortfall adjusted for the risk according to a mean-variance objective:

$$J^*(q) \triangleq \min_u \int_0^\infty f(u(t)) \, dt + \mu \int_0^\infty x(t)^\top \Sigma x(t) \, dt$$

subject to

$$\dot{x}(t) = -Y u(t), \quad \forall \ t \geq 0,$$

$$|u_i(t)| \leq \gamma_i, \quad \forall \ 1 \leq i \leq m, \ t \geq 0,$$

$$x(0) = q,$$

$$u \in L^1([0, \infty); \mathbb{R}^m). \quad (3.8)$$

Here, $\mu > 0$ is a parameter capturing the degree of the agent’s risk aversion.

The objective value of this dynamic control problem captures an explicit trade-off between transaction costs and market risk. If the agent trades faster, he is more likely to end up with higher transaction costs due to increased market impact; if he trades slower, he will end up facing more market risk over a longer period of time. As we will show in Theorem 4, the optimal liquidation
process always has the finite objective value defined in \((3.8)\). This implies that the asymptotic position is zero as time goes to infinity; otherwise the risk component of the objective value in \((3.8)\) would be infinite. An alternative would be to explicitly impose a exogenous finite time horizon by which the entire position must be liquidated, and this might be more appropriate in a fire sale setting, for example. Many of the results in this paper would hold in such an alternative, but we will opt for the simplicity of an endogenous time horizon.

Note that explicit in the formulation \((3.8)\) is the fact that we are restricting attention to only deterministic strategies; in other words, we are requiring that trading rates for each asset at every time to be specified in advance at time \(t = 0\). In general, there may be adaptive or stochastic strategies that perform better for our mean-variance objective. For example, Almgren and Lorenz (2007) show that stochastic strategies may outperform the best deterministic strategy; see also Lorenz and Almgren (2011). However, the proper economic motivation for mean-variance objective comes from the problem of maximizing expected utility for exponential, or CARA, utility functions. Schied et al. (2010) found that there is no added utility from adaptive strategies for CARA investors with a finite time horizon. Schöneborn (2011) expands this observation to infinite time horizons. As a result, if we believe the mean-variance objective stems from the optimization of CARA utility functions, the deterministic strategy is optimal. In any case, we will restrict our attention to deterministic strategies. This is consistent with much of the rest of the mean-variance optimal execution literature.

### 3.2.2. Optimal Strategy

In this section, we discuss some of the general characteristics of optimal strategies in our formulation.

**Theorem 4 (Existence and Convexity).** The dynamic control problem defined in \((3.8)\) is bounded and an optimal solution \(u^*\) always exists. In addition, the optimal value (the liquidity cost) is convex in initial position.

The proof of the theorem is given in the Appendix and is similar to results of Guéant (2015); Guéant et al. (2015). In our setting, the main technical requirement for existence of an optimal solution is the constrained liquidity assumption \((3.1)\). This helps us to establish the equi-integrability
of the feasible set. Because of this, our result is in some ways simpler than the earlier work. For example, we do not need to impose additional requirements on the transaction cost functional \( f(\cdot) \) beyond the convexity. Note that one implication of the existence theorem is that the optimal objective value is finite. This implies that, as \( t \to \infty \), \( x(t) \to 0 \). In other words, the position will be asymptotically liquidated.

A key element in our framework is that we allow for the direct trading of liquid bundles. As such, we may have more instruments than individual assets \((m > n)\), and it is possible to have more than one trading strategy \( u(\cdot) \) corresponding to any given trajectory of position \( x(\cdot) \). Therefore, the uniqueness of the optimal trading strategy may not be guaranteed. However, the optimal trajectory of position \( x(\cdot) \) must be unique, i.e., all optimal solutions have the same position at any time. Moreover, as established in the following theorem, under an additional convexity assumption the trading strategy must also be unique:

**Theorem 5 (Uniqueness).** All optimal solutions for the dynamic control problem in (3.8) have a unique optimal position trajectory \( x^* \in C([0, \infty); \mathbb{R}^n) \). Moreover, if the transaction cost functional \( f(\dot{\cdot}) \) is strictly convex, the optimal trading strategy \( u^* \in L^1([0, \infty); \mathbb{R}^m) \) must also be unique.

In general, it is difficult to come up with closed-form solutions to the dynamic control problem given in (3.8) (although we will consider some special cases in Section 3.3). We provide sufficient conditions for optimality by exploiting the convexity of the problem in the following:

**Theorem 6 (Sufficiency).** The pair \((x^*, u^*) \in C([0, \infty); \mathbb{R}^n) \times L^1([0, \infty); \mathbb{R}^m)\) form an optimal solution of (3.8) if, for all \( t \geq 0 \),

\[
x^*(t) = q - \int_0^t Y u^*(s) \, ds,
\]

\[
u^*(t) \in \arg\min_{u: -\gamma \leq u \leq \gamma} f(u) - 2 \int_t^\infty x^*(s)^\top \Sigma Y u \, ds. \tag{3.9}
\]

Theorem 6 provides a sufficient condition for the optimal trading strategy. Intuitively, the optimal trading rate at any given time results from a trade-off between the two components in (3.9). The first component represents the instantaneous transaction cost and the second component represents the impact on future risks. Note that Theorem 6 gives a sufficient condition, but not
a necessary one. If, however, the liquidation process takes only finite time, it can be shown that, \eqref{eq:3.9} is also necessary, using Pontryagin’s minimum principle, as we will do later. The necessity is difficult to generalize to an infinite trading horizon, where the corresponding general version of Pontryagin’s minimum principle is often pathological \cite{Halkin1974}.

### 3.3. Examples: Separable Transaction Costs

In the optimization problem \eqref{eq:3.8}, decision making across multiple assets is coupled. This comes from two fronts: the correlation between asset prices and possible cross-asset market impact in the transaction cost functional. However, it is extremely difficult to measure cross-asset market impact if any exists. Although our general framework in \eqref{eq:3.8} allows for the existence of cross-asset market impact, we will assume otherwise in order to maintain the focus on the correlation of assets and on the trade-off between market risk and transaction cost.

Specifically, the class of transaction cost functionals that are of particular interest are what we call **separable transaction costs**. These are transaction cost functionals that take the form of

\[
  f(u) = \sum_j \nu_j \hat{f}(u_j/\gamma_j),
\]

\eqref{eq:3.10}, for \( u \in \mathbb{R}^n \), where \( \hat{f} : \mathbb{R} \to \mathbb{R}_+ \) is a nonnegative convex function symmetric around 0 with \( f(0) = 0 \).

The scaling constant \( \nu_j > 0 \) captures the magnitude of the transaction cost of asset \( j \), as long as \( \gamma_j \) is the maximum trading rate from \eqref{eq:3.1}.

The intuition behind this type of functional is that the transaction cost of each asset is driven by similar mechanisms and depends primarily on the relative trading rate \( (u_j/\gamma_j) \). This corresponds to the fact that assets with higher liquidity (higher \( \gamma_j \)) are expected to have a smaller transaction cost given the same trading rate. Additionally, \eqref{eq:3.10} rules out the possibility of cross-asset market impact. Though advocated by some approaches in the literature such as Tsoukalas \textit{et al.} \cite{Tsoukalas2014}, cross-asset market impact is extremely difficult to estimate. On the other hand, the transaction costs of the form in \eqref{eq:3.10} can be estimated relatively easily from historical transaction data.

In what follows, we focus specifically on two examples of separable transaction costs: zero-cost
CHAPTER 3. PORTFOLIO LIQUIDITY ESTIMATION AND OPTIMAL EXECUTION

constrained liquidity and linear-cost constrained liquidity.

3.3.1. Zero-Cost Constrained Liquidity

The simplest case of a separable transaction cost is where there is a constraint on the trading rate, but trading itself doesn’t incur any cost. In this case we simply assume that \( f(u) = 0 \). Thus, we are capturing a setting where trading costs are minimal relative to risk, e.g., when the agent is very patient and tends to trade slowly and passively. A potential trading strategy is for the agent to trade only passively with mid-point orders where the cost from bid-ask spread is eliminated.

Under zero-cost constrained liquidity, (3.8) is equivalent to

\[
\begin{align*}
J^*(q) \triangleq & \min_{u} \int_0^\infty x^\top(t)\Sigma x(t) \, dt \\
\text{subject to} & \quad \dot{x}(t) = -Yu(t), \quad \forall \ t \geq 0, \\
& \quad |u_i(t)| \leq \gamma_i, \quad \forall \ 1 \leq i \leq m, \ t \geq 0, \\
& \quad x(0) = q, \\
& \quad u \in L^1([0, \infty); \mathbb{R}^m).
\end{align*}
\]

(3.11)

Here we assume that the parameter for risk aversion (\( \mu \)) takes the value of \( \mu = 1 \) without loss of generality. This setup under zero-cost constrained liquidity is very similar to that in Kim (2014). The main difference is that we allow for the trading of liquid bundles and hence \( Y \) does not need to be an identity matrix.

The zero-cost constrained liquidity model has some interesting features. The first is that the problem is scalable in terms of initial position.

**Theorem 7 (Scaling).** If \( u^* \) is optimal for the problem starting from \( q \), then \( \bar{u}(t) = u^*(t/\alpha) \ (\forall t > 0) \) is optimal for the problem starting with \( \alpha q \) with \( \forall \alpha \in \mathbb{R}^+ \), where

\[
J^*(\alpha q) = \alpha^3 J^*(q).
\]

Interestingly, the optimal object value, which is essentially the variance of the liquidation P&L,
Chapter 3. Portfolio Liquidity Estimation and Optimal Execution

Scales with initial position to the power of three. This is consistent with the prediction of inventory risk model (see Chapter 16 in Grinold and Kahn (2000)) that the total trading cost increases to the 3/2 power of the amount traded. First of all, the estimated time before a sufficient number of opposing trades appear to clear out the agent’s inventory is thought to be linear with the position size. The time to clear implies a per-share inventory risk proportional to the square root of the liquidation time (essentially the square root of the initial position). Then, by assuming that the market impact is proportional to the inventory risk, the total cost scales with the 3/2 power of the initial position.

Another interesting property of the zero-cost constrained liquidity model is that the optimal liquidation process requires only finite time. Although we do expect the position to be liquidated eventually, in theory our framework does not guarantee the finiteness of liquidation time. It could be that as the position gets smaller, the transaction cost bypasses the market risk, in which case it makes sense to trade slower and slower to keep the transaction cost small. One example is the case where the transaction cost is quadratic and (3.8) becomes a constrained linear-quadratic control problem, which takes infinite time. However, under zero-cost constrained liquidity, where the finiteness of the liquidation time is guaranteed, the theorem goes as follows.

Theorem 8 (Finite Horizon). For any initial position $q$, the optimal position trajectory $x(t)$ is guaranteed to reach zero in finite time.

Remember that in general the necessary condition of (3.8) is hard to derive. But, given that the optimal liquidation process takes only finite time, Pontryagin’s minimum principle can be used to derive the necessity of (3.9).

Lemma 1 (Optimality). A feasible control $u^*$ is optimal for (3.11) if and only if

$$\forall t \geq 0, \quad u^*(t) \in \arg\max_{u:-\gamma \leq u \leq \gamma} \left( \int_t^\infty (x^*(s))^\top \Sigma Y ds \right) u,$$

where (3.12) suggests what people called a “bang-bang” control or singular control, where $u_j(t)$ takes the upper bound $\gamma_j$ if the $j$th component of $\int_t^\infty (x^*(s))^\top \Sigma Y ds$ is positive and
takes the lower bound \(-\gamma_j\) if it is/they are negative. Properties of singular control problems can be found in [Johnson and Gibson (1963)], among many other works.

In general, it is difficult to come up with a closed-form solution to our problem in high dimensions. However, it is possible to characterize the two-dimensional case. Although it is possible to provide closed-form solutions to all two-dimensional cases, we are particularly interested in what we call high liquidity hedging, where the liquidity of the hedging asset is high. In practice, it is perhaps the most interesting setting, as people tend not to hedge with highly illiquid assets. The following theorem characterizes this setting.

**Theorem 9** (High Liquidity Hedging). In the two-dimensional case where model parameters are given by

\[
\Sigma = \begin{bmatrix}
\sigma_1^2 & \rho \sigma_1 \sigma_2 \\
\rho \sigma_1 \sigma_2 & \sigma_2^2
\end{bmatrix}, \quad Y = \begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix},
\]

if we further assume that

\[
\gamma_2 \geq \left| \rho \right| \frac{\sigma_1 \gamma_1}{\sigma_2},
\]  

(3.13)

then the optimal liquidity cost of portfolio \(q = (q, 0)\) is given by

\[
J^*(q) = \frac{1}{3} \frac{q^2}{\gamma_1} \left( 1 - \frac{\rho^2}{1 + \left| \rho \right| \frac{\sigma_1 \gamma_1}{\sigma_2 \gamma_2}} \right).
\]

(3.14)

Condition (3.13) is what we call the “high liquidity hedging condition”. It requires that the liquidity of the hedging instrument (\(\gamma_2\)) exceed a certain threshold. We can rewrite (3.13) as

\[
\frac{\gamma_2}{\gamma_1} \geq \left| \rho \right| \frac{\sigma_1}{\sigma_2}.
\]

The right-hand side is what we call the optimal hedging ratio. Given a unit of asset 1, it can be shown that the optimal amount of asset 2 needed to minimize total risk is given by \(|\rho| \sigma_1 / \sigma_2\). By examining the proof, one finds that the optimal trading strategy is to trade asset 2 as a hedge before unloading the hedged portfolio. And \(|\rho| \sigma_1 / \sigma_2\) is the optimal quantity of asset 2 needed to hedge every unit of asset 1. Equation (3.14) suggests some very intuitive structural properties of
CHAPTER 3. PORTFOLIO LIQUIDITY ESTIMATION AND OPTIMAL EXECUTION

liquidity costs. The first term \( \frac{q^2}{\gamma_1^2} \sigma_1^2 \) is the fair liquidity cost of trading asset 1 alone without hedging. The second term can then be interpreted as the benefit from hedging. Given that (3.13) holds, it is easy to see that the hedging benefit is increasing in \(|\rho|\), which captures the correlation between the two assets. This indicates that hedging is more efficient when one use highly correlated assets. Additionally, the hedging benefit is increasing in \(\gamma_2 \sigma_2\), a fact that can be interpreted as the rate of risk transferred by trading the hedging asset.

3.3.2. Linear-Cost Constrained Liquidity

Now we consider the case where transaction costs are determined by the following linear function:

\[
f(u) = \sum_j \nu_j |u_j|. \tag{3.15}
\]

Notice that this definition is still consistent with (3.10) if we have

\[
\hat{f}(u) = |u|
\]

and if we define \( \nu_j \) as the coefficient. Now \( \nu_j \) can be viewed as the bid-ask spread of asset \( j \). Basically, then, the agent is a liquidity taker and (3.8) can be written as

\[
J^*(q) \triangleq \min_u \int_0^\infty \nu_j |u_j(t)| dt + \mu \int_0^\infty x^\top(t) \Sigma x(t) dt
\]

subject to \( \dot{x}(t) = -Y u(t), \quad \forall \ t \geq 0, \)

\[|u_i(t)| \leq \gamma_i, \quad \forall \ 1 \leq i \leq m, \ t \geq 0,\]

\[x(0) = q, \]

\[u \in L^1([0, \infty); \mathbb{R}^m).\]  

The scaling property ceases to hold in this case as the transaction costs are typically linear in the position traded, whereas the risk component is at least quadratic. Again, a general closed-form solution is beyond our reach, but we can still explicitly solve the one-dimensional and two-dimensional cases.
Suppose we need to liquidate a certain position in asset 1. In the one-dimensional case where hedging is not possible, the total transaction costs incurred for trading a certain position are fixed and do not depend on the trading rate. As a result, the optimal strategy is to sell the position as fast as possible (at rate $\gamma$). Hence we get the following proposition:

**Theorem 10 (One Asset).** In the one-dimensional case, the cost of liquidating a position of $q$ with parameters $(\sigma, \gamma, \nu)$ is given by

$$J^*(q) = \nu|q| + \mu \frac{|q|^3 \sigma^2}{3\gamma}.$$  

The first term represents the total transaction costs associated with liquidating a position of $q$, and the second term represents the market risk of this liquidating process.

The two-dimensional case is more complicated – yet tractable. We consider only the case of “high liquidity hedging,” where (3.13) holds.

**Theorem 11 (Two Assets).** In the two-dimensional case where model parameters are given by

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix}, \quad Y = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad q = (q, 0)^\top,$$

if we further assume that

$$\gamma_2 \geq |\rho| \frac{\sigma_1 \gamma_1}{\sigma_2},$$

then the asset 2 will only be used to hedge if and only if

$$q^2 \geq \frac{2\gamma_1 \nu_2}{\mu \gamma_2 \rho \sigma_1 \sigma_2}.$$

If (3.17) is satisfied, then the optimal liquidity cost of the portfolio is given by

$$J^*(q) = \frac{1}{3} \frac{q^3}{\gamma_1} \sigma_1^2 \left(1 - \frac{\rho^2}{1 + |\rho| \frac{\sigma_1 \gamma_1}{\sigma_2 \gamma_2}}\right) + \nu_1 q + 2\nu_2 q \frac{\gamma_2}{\gamma_1} \frac{|\rho| \frac{\sigma_1 \gamma_1}{\sigma_2 \gamma_2}}{1 + |\rho| \frac{\sigma_1 \gamma_1}{\sigma_2 \gamma_2}} - \frac{4 \nu_2}{\sigma_2} \sqrt{\frac{2\nu_2 \rho \sigma_1 \gamma_1}{\sigma_2}} \frac{1}{3(1 + |\rho| \frac{\sigma_1 \gamma_1}{\sigma_2 \gamma_2})}.$$  

In this case, hedging with other assets comes with transaction costs, which are proportional to the hedging position acquired. Intuitively, if the transaction costs strictly dominate the market
CHAPTER 3. PORTFOLIO LIQUIDITY ESTIMATION AND OPTIMAL EXECUTION

risk, the agent will find hedging unattractive. Theorem 11 indicates that hedging is only optimal when the position passes the threshold defined in (3.17). This provides intuition about the trade-off between hedging benefits and their associated transaction costs. Specifically, hedging is less likely to be beneficial when:

1. The position size is small.
2. The covariance between the two assets is small.
3. The transaction costs for the hedging asset (asset 2 in this case) are large.
4. The agent is not risk averse (i.e., there are smaller \( \mu \)).

Additionally, the structure of (3.18) is interesting. We can see that the first term of (3.18) is exactly the liquidity cost given in (3.14) in Theorem 9 which is the liquidity cost for the case of a zero transaction cost. The second and third terms are the transaction costs associated with the trading strategy given in Theorem 9. The last term is the penalty that results from the fact that we expect less hedging in the presence of transaction costs. Interestingly, the penalty term is some constant that does not depend on the liquidating position. Thus, if the position is too small, hedging is not worthwhile; otherwise the optimal liquidity cost is the cost associated with the optimal strategy in the zero transaction cost case minus a constant that does not depend on position size.

3.4. Large Universe

Although deriving a closed-form solution for (3.8) proves to be difficult, it is not hard to see that the optimal liquidity cost is determined primarily by two factors: the covariance structure of prices and transaction cost functionals. We have discussed several transaction cost models in Section 3.3; now we consider the covariance structure of prices. Throughout this section, we will assume separable transaction costs.

In the most straightforward case where all asset prices are independent of each other, the liquidation problem of a portfolio consisting of \( n \) assets will degenerate to \( n \) one-dimensional sub-
problems where each asset is liquidated on its own. But if the asset prices are correlated, the story is more complicated. First of all, decisions regarding the liquidation of assets across a portfolio are coupled. Second, it might be beneficial to hedge a position’s market risk by acquiring some assets that are negatively correlated with the liquidating portfolio, as long as the extra transaction costs are acceptable. However, since the complexity of covariance structure grows with asset numbers, it becomes extremely difficult to provide an intuitive analysis of the liquidity costs. Unless we can somehow decrease the dimensions of the problem, very little can be said. By the same logic, the widely accepted idea that the variations of the prices of a large number of assets can be modeled by a small number of systemic factors becomes appealing.

Various models have been developed in finance to model the structure of the covariance matrix of asset prices. Here, we consider the multi-factor risk model first developed by [Ross (1976)] and then generalized by [Chamberlain and Rothschild (1983)]. The multi-factor risk is the basis for the arbitrage pricing theory which is well-studied in the finance literature. The main idea is that covariance across asset prices can be decomposed into two components: a systemic one and an idiosyncratic one. The systemic component is then modeled through various systemic factors that characterize different sources of systemic risks. This type of model has been widely used in the industry to predict risk structure in the solution of practical investment problems, e.g., the BARRA model from MSCI.

We define $F(t) \in \mathbb{R}^K$ to be the $K$-dimensional factor process. Without loss of generality, we assume that the factors are orthonormal and follow a standard $K$-dimensional Brownian motion. If the factors are correlated, we can always find a new set of $K$ orthonormal factors with a change of coordinate, as long as the covariance matrix of factors is full rank. Under a continuous time version of the multiple-factor model, the price dynamics of asset $i$ can be written as

$$dS_i(t) = l_i^T dF(t) + \varsigma_i dz_i(t), \quad 1 \leq i \leq n, t \geq 0,$$  

(3.19)

where $z_i(t)$ is a standard Brownian motion representing the idiosyncratic shocks for asset $i$. $l_j \in \mathbb{R}^K$ is the loading vector for asset $j$. $\varsigma_i$ is the magnitude of the asset’s idiosyncratic risk and hence
CHAPTER 3. PORTFOLIO LIQUIDITY ESTIMATION AND OPTIMAL EXECUTION

$\zeta_i z_i(t)$ represents the idiosyncratic disturbances that are zero-mean and independent across assets. In addition, we assume that $z_i(t)$ is independent of the factors.

Usually, the number of assets is much larger than the number of underlying factors. For example, in BARRA’s equity multi-factor model, the covariance structure of thousands of U.S. equities is explained by 60 industry factors, 12 style factors, and one country factor. This inspires us to explore the large-universe setting.

Now consider a sequence of problems with an increasing universe of securities, where the $n$th problem contains the first $n$ assets. The $n$th problem is characterized by asset price covariance matrix $\Sigma^{(n)}$. Now we can see that the definition in (3.2) is equivalent to (3.19) if the following decomposition holds:

$$\Sigma^{(n)} = L^{(n)}(L^{(n)})^\top + \Xi^{(n)},$$

(3.20)

where $L^{(n)} = (l_1, ..., l_n)$ is the factor loadings of the assets and

$$\Xi^{(n)} \triangleq \text{diag}(\varsigma_1^2, \varsigma_2^2, ..., \varsigma_n^2)$$

captures the idiosyncratic risk contribution.

Now we define $\lambda_{\text{min}}^{(n)}$ to be the smallest eigenvalues of $L^{(n)}(L^{(n)})^\top$. The notion of large universe is defined as follows:

**Definition 5 (Large Universe).** The sequence of problems is said to satisfy the large-universe property if the following conditions hold:

1. The magnitude of the idiosyncratic risk for each asset is bounded above,

$$\sup_j \varsigma_j^2 < \infty.$$ 

2. The smallest non-zero eigenvalue of $L^{(n)}(L^{(n)})^\top$ goes to infinity as $n$ goes to infinity:

$$\liminf_{n \to \infty} \lambda_{\text{min}}^{(n)} = \infty.$$
3. We will assume that the trading rate of each asset is lower bounded by $\gamma > 0$:

$$\inf_j \gamma_j = \gamma > 0.$$ 

The first condition indicates that the idiosyncratic risk of each asset is upper bounded by some positive number. This condition basically says that the idiosyncratic risk for each asset is small and hence can be diversified away.

The second condition has two interpretations: firstly, the factors are pervasive, in the sense that each factor affects almost all of the assets; secondly, there have to be enough variations in the factor loadings; otherwise, some factors may become redundant as their loadings can be explicitly calculated from the loadings of other factors. This condition can be linked to the conditions of arbitrage pricing theorem of Chamberlain and Rothschild (1983). The intuition is that we can potentially approximate the return of each factor after diversifying away the idiosyncratic risks. This condition is also related to the literature on estimating factor models, for example Fan et al. (2013). Those works typically make a stronger assumption, which requires the smallest non-zero eigenvalue to be linear on $n$, in order to asymptotically estimate the factor decomposition.

The third condition requires that there be non-vanishing liquidity for each asset.

**Proposition 5 (Factor Replicating Portfolio).** If the large-universe conditions hold, then for each factor $F_i(t)$, there exists a series of portfolios $\{p^{(i,n)}(t)\}$ defined by weights $\{\beta^{(i,n)}_j\}$ where

$$p^{(i,n)}(t) \triangleq \sum_{j=1}^{n} \beta^{(i,n)}_j S_j(t),$$

such that

1. The portfolio $p^{(i,n)}(t)$ has unit exposure on factor $F_i(t)$:

$$p^{(i,n)}(t) - F_i(t) = \epsilon^{(i,n)}(t),$$

where $\epsilon^{(i,n)}(t)$ is zero mean and independent of all factor-price processes, and has variance
2. The sum of the squares of the weights converge to 0:

\[ \lim_{n \to \infty} \sum_{j=1}^{n} (\beta_{j}^{(i,n)})^2 = 0. \]

This proposition indicates that in the large-universe regime, we can construct a sequence of well-diversified portfolios that eventually converge to the factor returns. The intuition is that as the number of tradeable assets increases, we can potentially take a small position in each asset and the idiosyncratic risks will be canceled out due to diversification. The proposition also provides an upper bound on the idiosyncratic risks for the factor portfolios, which is given by the ratio between the maximum idiosyncratic variance and the smallest non-zero eigenvalue of \( L(n) (L(n))^\top \). On one hand, if the assets have larger idiosyncratic risks, diversification becomes more difficult. On the other hand, achieving perfect diversification also depends on the assumption that the smallest non-zero eigenvalue of \( L(n) (L(n))^\top \) goes to infinity, which is guaranteed by the large-universe conditions.

The second part of the proposition implies that \( \beta_{j}^{(i,n)} \to 0 \) as \( n \to \infty \); hence we can construct those portfolios without trading too much of any asset. Combined with the third condition for large-universe regime, this suggests that factor portfolios can be traded very quickly.

### 3.4.1. Zero-Cost Constrained Trading

To start with, we will adopt the zero-cost constrained trading model where the transaction cost of each asset is represented by a constraint on its maximum trading rate. For simplicity, we assume that the \( Y \) matrix is just the identity matrix; hence only single assets are traded. We will later expand the results to the case of liquid bundles.

Now, consider the \( n \)th problem where there are \( n \) tradeable assets. Suppose we want to liquidate a portfolio \( q \in \mathbb{R}^n \) with positions in at most the first \( m \) assets, i.e.,

\[ q_j = 0, \quad \forall j > m. \]
Further, define $J_n^*(q)$ to be the optimal liquidity cost of portfolio $q$. If we only consider the idiosyncratic risks, this will result in less risk for the portfolio and hence should provide a lower bound for liquidity costs. As there is no correlation between assets, the problem will also be separable and can be solved asset by asset. By applying the results from Section 3.3.1, we have the following:

**Theorem 12 (Lower Bound of Hedging Benefits).** If we are allowed to trade other assets during the liquidation process, the liquidity cost is lower bounded by

$$J_n^*(q) \geq \sum_{j=1}^{m} \frac{\gamma_j^2}{3\gamma_j}|q_j|^3.$$  \hspace{1cm} (3.22)

The lower bound in Theorem 12 captures the situation where the portfolio has zero exposure to any of the risk factors. In this case, no other assets are needed for hedging and hence the liquidity cost consists only of idiosyncratic risks of assets already in the portfolio. Since this situation is of rare occurrence, the question is whether the lower bound is informative. In the following theorem, we try to prove that the lower bound in Theorem 12 is tight under the large-universe regime.

Consider the sequence of problems, indexed by $n$, discussed in the previous section. As we expand the set of assets that can be used for hedging, the liquidity cost should go down simply because we have more choices for hedging.

**Theorem 13 (Large Universe).** If the large-universe property is satisfied, then, asymptotically, the liquidity cost of any portfolio consisting of finitely many assets will be driven purely by idiosyncratic risks. More specifically, we have

$$J_\infty^*(q) = \lim_{n \to \infty} J_n^*(q) = \sum_{j=1}^{m} \frac{\gamma_j^2}{3\gamma_j}|q_j|^3,$$  \hspace{1cm} (3.23)

where $q$ is defined in (3.21), and $J_n^*(q)$ represents the optimal costs of liquidating $q \in \mathbb{R}^A$ in $A_n$.

Theorem 13 guarantees the convergence of liquidity cost when the number of tradeable assets goes to infinity. In showing what drives liquidity costs, Theorem 13 is important for two reasons. Firstly, from a risk perspective, only the idiosyncratic risks matter. Secondly, from a computational
perspective, we can simply use (3.23) to approximate actual cost if the large-universe setting is valid, instead of solving some complex dynamic control problem as in (3.11).

However, this only answers part of the question: (3.23) is still impractical if the convergence is too slow. The following theorem addresses this problem by explicitly bounding the rate of convergence.

**Theorem 14 (Convergence Speed).** Asymptotically, the difference between the liquidity cost and the theoretical limit converges at rate $1/\sqrt{\lambda_{\min}^{(n)}}$:

$$\limsup_{n \to \infty} \sqrt{\lambda_{\min}^{(n)}} |J_n^*(q) - J_\infty^*(q)| < \infty. \quad (3.24)$$

Theorem 14 says that the liquidity cost converges to the theoretical value roughly at the speed of one of $1/\sqrt{\lambda_{\min}^{(n)}}$. For a concrete example, let’s consider a simple case where the factor loadings of assets are drawn independently from a certain distribution. We then have the following theorem.

**Theorem 15 (Random factor loading).** If the asset factor loadings are drawn independently from a $K$-dimensional distribution (with a finite second moment), then, asymptotically, we have

$$\frac{\lambda_{\min}^{(n)}}{n} \overset{a.s.}{\to} C, \quad (3.25)$$

where $C$ is some constant that depends on only the distribution of factor loadings, and, therefore,

$$\limsup_{n \to \infty} \sqrt{n} |J_n^*(q) - J_\infty^*(q)| < \infty, a.s. \quad (3.26)$$

Theorem 15 shows that if the asset factor loadings are i.i.d., the liquidity cost converges to the large-universe approximation at a rate of at least $1/\sqrt{n}$.

### 3.4.2. Vanishing Bid-Ask Spread

So far, we have explored the asymptotic features of liquidity costs for the model of zero-cost constrained trading. Notice that in this case the only cost incurred by hedging comes from the
CHAPTER 3. PORTFOLIO LIQUIDITY ESTIMATION AND OPTIMAL EXECUTION

extra idiosyncratic risks added by trading other assets. The intuition here is that if there are many assets to choose from, we can construct a perfect hedging portfolio by trading only a small amount of each asset. By assuming a certain covariance structure, the large-universe conditions ensure the availability of such portfolios.

In reality, however, hedging with other assets is almost always associated with additional costs originating from commission fees, bid-ask spreads, and possibly price impact. In such cases it would be more interesting if the results for the large-universe regime could be extended to models with non-zero transaction costs. Fortunately, it can be shown that similar results as in (3.10) can be extended to a class of models with a separable transaction cost. More specifically, we define the vanishing bid-ask spread condition as follows.

Definition 6 (Vanishing Bid-Ask Spread). The vanishing bid-ask spread condition holds if the transaction cost functional is twice differentiable with

\[ \hat{f}'(0) = 0, \quad \hat{f}(0) = 0. \]

The idea is that there is no transaction cost for trading very small quantities. One such example is the case of quadratic transaction cost, which is documented in Gârleanu and Pedersen (2013).

Again, we consider the case where \( Y \) is the identity matrix. Here, the optimization problem we are considering becomes

\[
J^*(q) \triangleq \min_u \int_0^\infty \sum_{j=1}^m \nu_j \hat{f}(u_j(t)) \, dt + \mu \int_0^\infty x^\top(t) \Sigma x(t) \, dt
\]

subject to

\[
\dot{x}(t) = -Yu(t), \quad \forall \ t \geq 0,
\]

\[
|u_i(t)| \leq \gamma_i, \quad \forall 1 \leq i \leq m, \ t \geq 0,
\]

\[
x(0) = q,
\]

\[
u \in L^1((0, \infty); \mathbb{R}^m).
\]

For any liquidation model specified by (3.27), let’s consider the one-dimensional case \((n = 1)\), where only one asset of size \(q\) is traded. Additionally, we assume that the asset has a transaction
cost parameter $\nu$ and a liquidity parameter $\gamma$. If we consider only the idiosyncratic risk of this asset, which is $\varsigma^2$, the corresponding optimal liquidity cost is given by

$$J^*(q, \varsigma, \nu, \gamma) \triangleq \min_u \mu \varsigma^2 \int_0^\infty x^2(t) \, dt + \nu \int_0^\infty \hat{f}(u(t)) \, dt$$

subject to

$$\dot{x}(t) = -u(t), \quad \forall \ t \geq 0,$$

$$|u(t)| \leq \gamma, \quad \forall \ t \geq 0,$$

$$x(0) = q,$$

$$u \in L^1([0, \infty); \mathbb{R}).$$

(3.28)

It is easy to check that (3.28) satisfies the conditions in 4 and 5. Then the optimal liquidity cost is just a function of $q, \varsigma, \gamma, \nu$, and we denote it by $J^*(q, \varsigma, \nu, \gamma)$. The specific form of $J^*(q, \varsigma, \nu, \gamma)$ depends on the corresponding transaction cost function $\hat{f}(\cdot)$, and can be solved through HJB equations.

**Theorem 16 (Generalization).** For any liquidation model specified in (3.27), if the transaction cost functional is twice differentiable with

$$\hat{f}'(0) = 0, \quad \hat{f}(0) = 0,$$

the extended results of Theorem 13 still hold. More specifically, we have

$$\lim_{n \to \infty} J^*_n(q) = \sum_{j=1}^m J^*(q_j, \varsigma_j, \nu_j, \gamma_j),$$

(3.29)

where $J^*_n(q)$ represents the optimal liquidity costs for a portfolio $q$ with assets in $A_n$.

Theorem 16 says that the liquidity cost of any portfolio consisting of only finitely many assets is equal to the cost of liquidating each asset individually but with only idiosyncratic risks.

### 3.4.3. Linear Transaction Costs

The previous theorem depends on the assumption that the average transaction cost diminishes when the position is very small. Now we consider the case where transaction costs are determined
by the following linear function:

\[ f(u) = \sum_j \nu_j |u_j|. \]

In this case, \( \nu_j \) can be viewed as the bid-ask spread of asset \( j \). Without hedging, the optimal strategy is to sell the position as soon as possible (at trading rate limit \( \gamma \)), as is illustrated in Proposition 10.

Unlike in the setup of the previous model, here hedging with other assets is not cost-free. With linear transaction costs, the agent can no longer make the transaction cost vanish by trading small positions. It is always costly to incur any other positions.

Theorem 11 says that hedging is desirable if the position is large. Now, in the case of a large position, the question is, will a similar version of Theorem 13 hold asymptotically?

**Theorem 17** (Linear costs).

\[
\lim_{n \to \infty} \lim_{||q|| \to \infty} \frac{\tilde{J}_{LC}^*(q)}{\tilde{J}_{LC,n}^*(q)} = 1
\]  

where

\[
\tilde{J}_{LC}^*(q) = \mu \sum_{j=1}^m \frac{s_j^2}{3\gamma_j} |q_j|^3.
\]

This theorem suggests that when the position is extremely large, the market-risk contribution strictly dominates that of the transaction cost, and the best strategy is to fully hedge.

### 3.4.4. Hedging with Liquidity Bundles

So far we have studied the case of trading only individual assets. Let us now expand the result to allow for the trading of liquid bundles. For tractability reasons, we restrict our attention to the model of zero-cost constrained trading and consider trading only one liquidity bundle, such as an ETF. Without loss of generality, suppose this ETF covers the first \( m \) assets. Then the liquidation
matrix is given by
\[
Y = \begin{bmatrix}
1 & 0 & \ldots & 0 & \alpha_1 \\
0 & 1 & \ldots & 0 & \alpha_2 \\
0 & \ldots & \ldots & \ldots & \ldots \\
0 & \ldots & \ldots & 1 & \alpha_m \\
0 & \ldots & \ldots & \ldots & 0 \\
\end{bmatrix}
\]  

(3.31)

In addition, we assume that
\[
|\alpha_i \gamma_{\text{ETF}}| < \gamma_i, \forall 1 \leq i \leq m.
\]

This assumption suggests that the liquidity of asset \(i\) from trading the ETF (\(|\alpha_i \gamma_{\text{ETF}}|\)) should be less than the liquidity from trading asset \(i\) itself (\(\gamma_i\)). The assumption is generally true in practice and enables us to bypass technical difficulties.

In a further attempt to keep things simple, we consider liquidating the position of a single asset.

**Theorem 18 (ETF).** If the large-universe property is satisfied, then, asymptotically, the cost for liquidating \(q_j\) shares of asset \(j \leq m\) is given by
\[
\lim_{n \to \infty} J_{\text{ETF},n}^*(q) = \frac{\gamma_j^2}{3(|\alpha_1 \gamma_{\text{ETF}} + \gamma_j|)} q_j^3,
\]
where \(q_i = 0, \forall i \neq j\).

We can see that (3.32) is very similar to the one-dimensional case of (3.23). First of all, only idiosyncratic risk matters in the large-universe context. Second, the structure of the optimal liquidity cost is the same except for different denominators. In particular, \(|\alpha_1 \gamma_{\text{ETF}}|\) can be viewed as the liquidity from trading the ETF, while \(\gamma_j\) is the liquidity from trading asset \(j\) itself. This theorem provides the intuition that when asset space is large enough, adding an ETF is equivalent to directly increasing the liquidity of the asset.
CHAPTER 3. PORTFOLIO LIQUIDITY ESTIMATION AND OPTIMAL EXECUTION

3.5. Empirical Results

So far, we have built the framework for estimating liquidity costs for portfolios, and discussed the implications for the drivers of liquidity costs. However, many of our theoretical results rely heavily on assumptions about the structure of price covariance matrix, liquidity level of assets, and so on. For example, one wonders whether the conditions of a large universe are necessarily easy to satisfy in the real world. Also, it would be interesting to illustrate some of our main findings with concrete examples. In order to demonstrate these questions, in the remainder of this paper we will calibrate our model with a small subset of U.S. equities (29 stocks in the utility sector). Specifically, we will fit the factor model using historical returns.

3.5.1. Overview of the Data Set

As candidates for our calibration, we restrict our attention to the 29 stocks in the Utilities Select Sector Index, which is one of the eleven Select Sector Indices in S&P 500 that track major economic segments and are highly liquid. All the stocks included are from the following industries: electric utilities, water utilities, multi-utilities, independent power producers and energy traders, and gas utilities. We also take into account the Utilities Select Sector SPDR Fund (or XLU), which is an ETF seeking to track the performance of the Utilities Select Sector Index. As a result, the universe of instruments is comprised of 29 individual stocks and one ETF. The market parameters, including prices, daily returns, and daily volume, are obtained from Yahoo Finance for all trading days from January 1, 2012, to April 1, 2016.

Summary statistics of XLU and its asset holdings (as of April 1, 2016) are given in Table 3.1. As we can see, the weights of assets are somewhat close. This is a result by construction. Quarterly rebalancing ensures that no stock is allowed to have a weight greater than 25%, and that the sum of the stocks with weight greater than 4.8% cannot exceed 50% of the total index weight. All the individual assets are actively traded, with a daily volume ranging from $7.3 \times 10^5$ shares to $6.8 \times 10^6$ shares. In particular, the ETF (XLU) is highly liquid with an average daily volume of $1.6 \times 10^7$ shares.

2Prices and returns are adjusted for dividends.
shares, which is much larger than any individual stocks. The average daily volatility of return for XLU is much smaller than its underlying assets, which was to be expected from the diversification it brought. For each asset, the volume traded through the ETF is quite significant and accounts for a sizable portion of the total daily volume ranging from 7.82% for NRG to 31.49% for NEE. We also observe large correlations between individual stocks.

<table>
<thead>
<tr>
<th>Name</th>
<th>Identifier</th>
<th>Weight (%)</th>
<th>Price ($)</th>
<th>Average Daily Volume (Shares, × 10^6)</th>
<th>Total Risk (Daily, %)</th>
<th>Volume Trade Through ETF (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utilities Select Sector SPDR Fund</td>
<td>XLU</td>
<td>-</td>
<td>49.81</td>
<td>16.08</td>
<td>0.85</td>
<td>-</td>
</tr>
<tr>
<td>Ameren Corporation</td>
<td>AEE</td>
<td>1.93</td>
<td>50.54</td>
<td>1.68</td>
<td>1.33</td>
<td>19.93</td>
</tr>
<tr>
<td>American Electric Power Company Inc.</td>
<td>AEP</td>
<td>5.24</td>
<td>67.01</td>
<td>2.86</td>
<td>1.32</td>
<td>23.90</td>
</tr>
<tr>
<td>AES Corporation</td>
<td>AES</td>
<td>1.26</td>
<td>11.57</td>
<td>6.22</td>
<td>2.04</td>
<td>15.26</td>
</tr>
<tr>
<td>American Water Works Company Inc.</td>
<td>AWK</td>
<td>2.15</td>
<td>69.50</td>
<td>3.30</td>
<td>1.51</td>
<td>8.19</td>
</tr>
<tr>
<td>CMS Energy Corporation</td>
<td>CMS</td>
<td>1.89</td>
<td>42.71</td>
<td>2.50</td>
<td>1.31</td>
<td>15.46</td>
</tr>
<tr>
<td>CenterPoint Energy Inc.</td>
<td>CNP</td>
<td>1.55</td>
<td>21.20</td>
<td>4.62</td>
<td>1.59</td>
<td>13.89</td>
</tr>
<tr>
<td>Dominion Resources Inc.</td>
<td>D</td>
<td>7.04</td>
<td>75.39</td>
<td>2.56</td>
<td>1.26</td>
<td>31.86</td>
</tr>
<tr>
<td>DTE Energy Company</td>
<td>DTE</td>
<td>2.64</td>
<td>91.05</td>
<td>1.01</td>
<td>1.20</td>
<td>25.00</td>
</tr>
<tr>
<td>Duke Energy Corporation</td>
<td>DUK</td>
<td>8.25</td>
<td>81.13</td>
<td>3.43</td>
<td>1.32</td>
<td>25.91</td>
</tr>
<tr>
<td>Consolidated Edison Inc.</td>
<td>ED</td>
<td>3.60</td>
<td>76.99</td>
<td>1.82</td>
<td>1.36</td>
<td>22.41</td>
</tr>
<tr>
<td>Edison International</td>
<td>EIX</td>
<td>3.80</td>
<td>71.94</td>
<td>2.10</td>
<td>1.44</td>
<td>22.01</td>
</tr>
<tr>
<td>Evergy Resource Energy</td>
<td>ES</td>
<td>2.96</td>
<td>58.80</td>
<td>1.77</td>
<td>1.30</td>
<td>24.87</td>
</tr>
<tr>
<td>Entergy Corporation</td>
<td>ETR</td>
<td>2.25</td>
<td>79.97</td>
<td>1.26</td>
<td>1.39</td>
<td>19.48</td>
</tr>
<tr>
<td>Exelon Corporation</td>
<td>EXC</td>
<td>5.38</td>
<td>35.66</td>
<td>6.80</td>
<td>1.63</td>
<td>19.38</td>
</tr>
<tr>
<td>FirstEnergy Corp.</td>
<td>FE</td>
<td>2.46</td>
<td>36.03</td>
<td>3.71</td>
<td>1.62</td>
<td>16.09</td>
</tr>
<tr>
<td>AGL Resources Inc.</td>
<td>GAS</td>
<td>1.33</td>
<td>65.51</td>
<td>0.73</td>
<td>1.41</td>
<td>24.51</td>
</tr>
<tr>
<td>NextEra Energy Inc.</td>
<td>NEE</td>
<td>8.98</td>
<td>118.71</td>
<td>2.10</td>
<td>1.31</td>
<td>31.49</td>
</tr>
<tr>
<td>NiSource Inc.</td>
<td>NI</td>
<td>1.22</td>
<td>23.82</td>
<td>2.24</td>
<td>1.46</td>
<td>20.06</td>
</tr>
<tr>
<td>NRG Energy Inc.</td>
<td>NRG</td>
<td>0.77</td>
<td>12.77</td>
<td>6.77</td>
<td>3.31</td>
<td>7.82</td>
</tr>
<tr>
<td>PG&amp;E Corporation</td>
<td>PCG</td>
<td>4.77</td>
<td>59.83</td>
<td>2.57</td>
<td>1.39</td>
<td>27.08</td>
</tr>
<tr>
<td>Public Service Enterprise Group Inc.</td>
<td>PEG</td>
<td>3.89</td>
<td>47.32</td>
<td>3.07</td>
<td>1.38</td>
<td>23.41</td>
</tr>
<tr>
<td>Pinnacle West Capital Corporation</td>
<td>PNW</td>
<td>1.37</td>
<td>75.49</td>
<td>0.83</td>
<td>1.31</td>
<td>19.12</td>
</tr>
<tr>
<td>PPL Corporation</td>
<td>PPL</td>
<td>4.16</td>
<td>38.19</td>
<td>3.93</td>
<td>1.31</td>
<td>24.25</td>
</tr>
<tr>
<td>SCANA Corporation</td>
<td>SOG</td>
<td>1.60</td>
<td>71.10</td>
<td>0.99</td>
<td>1.34</td>
<td>19.85</td>
</tr>
<tr>
<td>Southern Company</td>
<td>SO</td>
<td>7.58</td>
<td>51.70</td>
<td>4.53</td>
<td>1.36</td>
<td>28.30</td>
</tr>
<tr>
<td>Sempra Energy</td>
<td>SRE</td>
<td>4.05</td>
<td>105.92</td>
<td>1.27</td>
<td>1.41</td>
<td>26.29</td>
</tr>
<tr>
<td>TECO Energy Inc.</td>
<td>TE</td>
<td>1.12</td>
<td>27.56</td>
<td>2.46</td>
<td>1.56</td>
<td>14.47</td>
</tr>
<tr>
<td>WEC Energy Group Inc.</td>
<td>WEC</td>
<td>3.02</td>
<td>59.97</td>
<td>1.88</td>
<td>1.34</td>
<td>23.40</td>
</tr>
<tr>
<td>Xcel Energy Inc.</td>
<td>XEL</td>
<td>3.39</td>
<td>41.94</td>
<td>3.09</td>
<td>1.32</td>
<td>22.86</td>
</tr>
</tbody>
</table>

Table 3.1: Descriptive statistics for the equity holdings of the assets under discussion. The weights and prices are as of 04/01/2016. The average daily volume is calculated through the period 01/01/2012–04/01/2016. The volatility is defined as the standard deviation of percentage daily returns. The volume trade through ETF is calculated as $|\gamma_{XLU}\alpha_j|/\gamma_j$.

3.5.2. Model Calibration

The main parameters involved in our model are the liquidation matrix $Y$, covariance matrix of asset prices $\Sigma$, factor loading matrix $L$, and liquidity parameters $\gamma_j$.

Liquidation matrix and liquidity constraints. The liquidation matrix in the example takes the same form as in (3.31). The key is to determine parameters $\{\alpha_i\}$, where $\alpha_i$ represents the number
CHAPTER 3. PORTFOLIO LIQUIDITY ESTIMATION AND OPTIMAL EXECUTION

of shares of asset $i$ contained in one share of the ETF (XLU). More specifically, they are given by the following the formula:

$$
\alpha_i = \frac{S_{ETF}w_i}{S_i},
$$

where $S_{ETF}$ is the price of the ETF (XLU), $S_i$ is the price of asset $i$, and $w_i$ is the dollar weight of asset $i$ in the ETF. XLU is subjected to quarterly rebalances after the close of business on the second to last calculation day of March, June, September, and December. As a result, the weights for each stock can be modified accordingly. But in our analysis, since the liquidation process takes place only in a short time period, we may safely assume that the structure of the ETF does not change over time; in other words, $\{\alpha_i\}$ is fixed. It is worth noticing that the weights of all the individual stocks do not sum up to 100%. The reason is that the ETFs often put away a small percentage of money in cash. In our analysis, we will neglect those terms as they pose no risk whatsoever.

In our model of zero-cost constrained trading, the liquidity constraint is defined as the maximum rate one can trade without incurring any transaction cost. In general, there is no good way to estimate the threshold without proprietary trading data. For simplicity, we set the threshold at 10% of market trading rate. As we can see from Theorem 13, the exact level of liquidity constraints does not affect the properties of the solutions. Another issue that could complicate the analysis is that the market trading rate can be changing over time. Typically, more trading activities are expected to happen around open and close, and fewer are expected at noon. For tractability, we will just assume that the liquidity constraint is fixed during the liquidation period we are looking at. For example, when AEE is trading at an average daily volume of $1.68 \times 10^6$ shares per day, then we set $\gamma_{AEE} = 1.68 \times 10^5$ shares per day.

**Covariance structure.** We fit a single-factor model (since all the stocks are in the same sector) using historical daily returns from January 1, 2012 to April 1, 2016. In our analysis, it is achieved using the principal component method. An in-depth discussion of this method is given in Chapter 9.4 in Tsay (2005).
3.5.3. Results

We consider liquidating \( q \) shares of one individual stock within the asset universe. Now for each specific stock \( j \), we can consider the following four trading strategies:

1. **No Hedging**: trade stock \( j \) only.

2. **Hedging with ETF**: trade stock \( j \) and hedge with the ETF (XLU).

3. **Hedging with Basket**: trade stock \( j \) and hedge with all other individual stocks.

4. **Hedging with All Assets**: trade all assets including the ETF.

Without loss of generality, for each stock \( j \) we consider the liquidation of a position by 10\% of its daily volume.

<table>
<thead>
<tr>
<th>Trading Strategy</th>
<th>Theoretical Liquidity Cost in Closed Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Hedging</td>
<td>( \frac{\sigma_j^2}{3\gamma_j}q^3 )</td>
</tr>
<tr>
<td>Hedging with ETF</td>
<td>( \frac{1}{3}\gamma_j^2\sigma_j^2 \left(1 - \frac{\rho^2}{1 +</td>
</tr>
<tr>
<td>Hedging with Basket</td>
<td>( \frac{\varsigma_j^2}{3\gamma_j}q^3 )</td>
</tr>
<tr>
<td>Hedging with All Assets</td>
<td>( \frac{\varsigma_j^2}{3(</td>
</tr>
</tbody>
</table>

**Table 3.2**: Theoretical results for the four trading strategies.

Table 3.2 provides the theoretical liquidity costs associated with the four strategies, where the results for the last two strategies are obtained as in the large-universe asymptotic limit. If we compare the strategy of no hedging with that of hedging with basket, we can see that the former is proportional to the total variance \( \sigma_j^2 \), whereas the latter is proportional to the idiosyncratic variance \( \varsigma_j^2 \).

Table 3.3 shows the numerical results of applying the estimated market parameters. For strategy 3 and strategy 4, we provide two sets of numerical results: the one we call exact is calculated by
CHAPTER 3. PORTFOLIO LIQUIDITY ESTIMATION AND OPTIMAL EXECUTION

<table>
<thead>
<tr>
<th>Identifier</th>
<th>No Hedging</th>
<th>Hedging with ETF Only</th>
<th>Hedging with Basket Exact</th>
<th>Hedging with Basket Approximate</th>
<th>Hedging with All Assets Exact</th>
<th>Hedging with All Assets Approximate</th>
</tr>
</thead>
<tbody>
<tr>
<td>AEE</td>
<td>3.90</td>
<td>1.24</td>
<td>1.06</td>
<td>1.00</td>
<td>0.88</td>
<td>0.83</td>
</tr>
<tr>
<td>AEP</td>
<td>4.31</td>
<td>1.52</td>
<td>1.16</td>
<td>1.00</td>
<td>0.92</td>
<td>0.81</td>
</tr>
<tr>
<td>AES</td>
<td>1.68</td>
<td>1.15</td>
<td>1.01</td>
<td>1.00</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>AWK</td>
<td>2.04</td>
<td>1.20</td>
<td>1.05</td>
<td>1.00</td>
<td>0.96</td>
<td>0.92</td>
</tr>
<tr>
<td>CMS</td>
<td>5.09</td>
<td>1.35</td>
<td>1.11</td>
<td>1.00</td>
<td>0.95</td>
<td>0.87</td>
</tr>
<tr>
<td>CNP</td>
<td>2.42</td>
<td>1.22</td>
<td>1.04</td>
<td>1.00</td>
<td>0.90</td>
<td>0.88</td>
</tr>
<tr>
<td>D</td>
<td>3.29</td>
<td>1.28</td>
<td>1.10</td>
<td>1.00</td>
<td>0.82</td>
<td>0.76</td>
</tr>
<tr>
<td>DTE</td>
<td>5.37</td>
<td>1.31</td>
<td>1.10</td>
<td>1.00</td>
<td>0.87</td>
<td>0.80</td>
</tr>
<tr>
<td>DUK</td>
<td>3.61</td>
<td>1.44</td>
<td>1.16</td>
<td>1.00</td>
<td>0.91</td>
<td>0.79</td>
</tr>
<tr>
<td>ED</td>
<td>3.21</td>
<td>1.14</td>
<td>1.07</td>
<td>1.00</td>
<td>0.86</td>
<td>0.82</td>
</tr>
<tr>
<td>EIX</td>
<td>2.89</td>
<td>1.22</td>
<td>1.07</td>
<td>1.00</td>
<td>0.87</td>
<td>0.82</td>
</tr>
<tr>
<td>ES</td>
<td>4.36</td>
<td>1.29</td>
<td>1.09</td>
<td>1.00</td>
<td>0.86</td>
<td>0.80</td>
</tr>
<tr>
<td>ETR</td>
<td>2.82</td>
<td>1.16</td>
<td>1.05</td>
<td>1.00</td>
<td>0.87</td>
<td>0.84</td>
</tr>
<tr>
<td>EXC</td>
<td>2.10</td>
<td>1.25</td>
<td>1.07</td>
<td>1.00</td>
<td>0.89</td>
<td>0.84</td>
</tr>
<tr>
<td>FE</td>
<td>2.12</td>
<td>1.14</td>
<td>1.04</td>
<td>1.00</td>
<td>0.89</td>
<td>0.86</td>
</tr>
<tr>
<td>GAS</td>
<td>1.25</td>
<td>1.03</td>
<td>1.00</td>
<td>1.00</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>NEE</td>
<td>3.55</td>
<td>1.44</td>
<td>1.15</td>
<td>1.00</td>
<td>0.87</td>
<td>0.76</td>
</tr>
<tr>
<td>NI</td>
<td>2.49</td>
<td>1.12</td>
<td>1.02</td>
<td>1.00</td>
<td>0.85</td>
<td>0.83</td>
</tr>
<tr>
<td>NRG</td>
<td>1.35</td>
<td>1.13</td>
<td>1.01</td>
<td>1.00</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>PCG</td>
<td>2.68</td>
<td>1.18</td>
<td>1.06</td>
<td>1.00</td>
<td>0.82</td>
<td>0.79</td>
</tr>
<tr>
<td>PEG</td>
<td>3.55</td>
<td>1.34</td>
<td>1.10</td>
<td>1.00</td>
<td>0.88</td>
<td>0.81</td>
</tr>
<tr>
<td>PNW</td>
<td>4.33</td>
<td>1.18</td>
<td>1.06</td>
<td>1.00</td>
<td>0.88</td>
<td>0.84</td>
</tr>
<tr>
<td>PPL</td>
<td>3.22</td>
<td>1.28</td>
<td>1.08</td>
<td>1.00</td>
<td>0.86</td>
<td>0.80</td>
</tr>
<tr>
<td>SCG</td>
<td>4.97</td>
<td>1.18</td>
<td>1.07</td>
<td>1.00</td>
<td>0.89</td>
<td>0.83</td>
</tr>
<tr>
<td>SO</td>
<td>3.39</td>
<td>1.26</td>
<td>1.11</td>
<td>1.00</td>
<td>0.86</td>
<td>0.78</td>
</tr>
<tr>
<td>SRE</td>
<td>2.73</td>
<td>1.22</td>
<td>1.05</td>
<td>1.00</td>
<td>0.83</td>
<td>0.79</td>
</tr>
<tr>
<td>TE</td>
<td>1.67</td>
<td>1.08</td>
<td>1.01</td>
<td>1.00</td>
<td>0.88</td>
<td>0.87</td>
</tr>
<tr>
<td>WEC</td>
<td>4.41</td>
<td>1.27</td>
<td>1.10</td>
<td>1.00</td>
<td>0.88</td>
<td>0.81</td>
</tr>
<tr>
<td>XEL</td>
<td>5.20</td>
<td>1.40</td>
<td>1.13</td>
<td>1.00</td>
<td>0.91</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table 3.3: Numerical results for the utility-sector example.

solving the discretized version of the optimalization problem as in (3.11); the other one we call approximated is calculated using the closed-form equations in the large-universe limit as shown in Table 3.2. From the scaling property in Theorem 7, it is expected that the size of the position we are liquidating should not affect the comparisons between different trading strategies. To better illustrate the results, we normalize the results by setting the approximated liquidity cost of strategy 3 (in large-universe asymptotic limit) as a benchmark.

First of all, the benefit from hedging is quite substantial. By hedging with the ETF alone, we see a significant decrease in all assets, with the ratio of reduction ranging from 16% for NRG to 76% for SCG. This can be explained by the huge liquidity of XLU and the high correlation between XLU and the individual assets.
Secondly, hedging with a basket of individual assets is even better than hedging with the ETF alone for all the assets, though the size of the benefit varies among assets. For AEP, trading with individual assets further reduces the liquidity cost by about 38% from that of hedging with ETF only; by contrast, the number is only 3% for GAS.

Thirdly, for strategy hedging with basket, we can see that the approximate values are very close to the exact ones obtained from solving the dynamic control problem. This shows that the conditions for the large-universe regime should be satisfied here and (3.23) is indeed a good approximation of the actual liquidity cost.

Finally, we see that the benefit of adding ETF to the hedging basket is sizable. In most cases, the reduction of liquidity cost is close to that predicted by (3.32). This shows that in the large-universe regime, trading ETF is equivalent to providing additional liquidity, since the portfolio’s market risk exposure has been almost perfectly hedged by the basket of individual stocks.

Finally, we consider the liquidation of a certain position in a representative stock: AEE. To do so, we add other stocks one by one into the stock basket in alphabetic order. Figure 3.1 shows how the liquidity cost changes as more and more individual stocks are allowed to be used for hedging. The convergence of the liquidity cost to the large-universe asymptotic limit is very fast. Figure 3.2 further shows the evolution of $\sqrt{\lambda_{\min}^{(n)}|J_n^*(q) - \tilde{J}^*(q)|}$, as defined in Theorem 14. As expected, the quantity gradually converges to some constant, which shows that the convergence rate of the liquidity cost is roughly converging at the rate of $1/\sqrt{\lambda_{\min}^{(n)}}$.

3.6. Concluding Remarks

Accurately estimating liquidity cost is of central importance in portfolio management, and is especially crucial when portfolio managers need to unwind large positions. Additionally, liquidity risk premia can be used to penalize illiquid assets in portfolio construction. We provide a framework to address the multi-asset optimal execution problem, which is far from being a simple extension to the single-asset approach currently adopted in practice.

Our results suggest that managing execution at the portfolio level can substantially reduce
CHAPTER 3. PORTFOLIO LIQUIDITY ESTIMATION AND OPTIMAL EXECUTION

Figure 3.1: Liquidity cost as the number of assets for hedging increases.

Figure 3.2: Convergence of the liquidity costs.

liquidity cost by taking advantage of the inter-correlation of asset prices. The complex interaction between asset prices can have a substantial impact on the aggregate portfolio execution cost and risk. We find that traders can improve execution efficiency by hedging the market risk by trading correlated assets simultaneously. This advanced strategy is also true even for the execution of single assets.
An even more compelling takeaway is that in the large-universe setting where the covariance structure of asset prices can be explained by only a handful of factors, the liquidity cost is almost purely driven by idiosyncratic risks. This implies that portfolio managers need to pay more attention to an asset’s idiosyncratic risk as it not only impacts the risk of security return but also plays a key role in the liquidation process. Additionally, we are able to provide a good closed-form approximation of the liquidity costs in non-asymptotic situations. This can potentially save one the trouble of solving a large-scale dynamic control problem.

Finally, our results signify the importance of trading liquid bundles such as ETFs in optimal liquidation. While previous works are mainly focusing on the hedging benefits of trading liquid bundles, we are the first to recognize its contribution in terms of liquidity provision. In fact, the contribution of liquidity provision is often larger than that of hedging risks. In the large-universe context, we manage to show that trading liquid bundles is almost equivalent to providing an additional source of liquidity to the underlying asset.
Chapter 4

A Model for Queue Position Valuation

4.1. Introduction

The way people trade in the financial market has changed fundamentally in the past decades. Computer technology has revolutionized the financial market: most liquid financial market are now dominated by electronic trading in central limit order books. Problems around modeling the limit order books are of paramount importance both in academia and in practice. Most limit order books are operated under the rule of “price-time priority”, in which limit orders are prioritized first based on price, and then on their arrival time. This structure naturally fit in the queuing models, and has therefore being studied extensively in the literature.

In this chapter, we study the economic value of a limit order as a function of its position in the queue. Specifically, we focus on “large-tick asset” where queuing effect is important. We identify two components that driven the value embedded in queue positions. The first is a informational component which relates to the adverse selection costs incurred in trading. The second is a dynamic component, which accrues over time as you move up in the queue. We are able to develop a tractable framework that can be easily calibrated with market parameters.

We calibrated our model to a set of U.S. equity data and obtain predictions for order values at different queue positions. We then validate those predictions by backtesting or simulating those orders in the real data. We can hypothetically imagine inserting synthetic simulated orders,
and track how those orders evolve in the limit order books versus real orders. Therefore, we get nonparametric estimates of the order values and find that they line up very well with our model predictions. We also find that, for many stocks, the value of queue position are enormous and are comparable to the spread.

The rest of this chapter is organized as follows: Section 4.2 provides an overview of our approach and describe the dynamic of the order book. In Section 4.3 we provide closed-form solution for the value function. In Section 4.4 we applies the model to trading data from NASDAQ. Section 4.5 describes the procedure of backtesting and compare the backtesting results with model predictions. Section 4.6 concludes and provides practical implications of our analysis. Most proofs appear in the appendix.

4.2. Model

In the modern equity market, while the price per share differs substantially across assets, the tick size is artificially fixed. For example, all stocks traded at NYSE have a minimum increment of $0.01. Large-tick assets, which according to Eisler et al. (2012), are such that “the bid-ask spread rarely exceeds the minimum tick size”. These are the assets where the tick size is economically significant, and therefore they are typically traded with the bid-spread equal to the tick size. Another important characteristic of large-tick assets is that they tend to have large queues in the limit order book. The reason is that the cost of raising the price by one tick will be very significant. Hence, instead of competing through price, people tend to queue up. Figure 4.1 shows the relationship between bid-ask spread and displayed liquidity for various future contracts, and we can see a clear pattern that queueing effect is more prominent for large-tick assets. In this paper, we will restrict our attention to the large-tick assets where queueing is important.

For simplicity, we will assume that over the time scale of our model, the bid and ask prices do not change as the tick size is so big. Also, we will assume that the bid-ask spread is constant and equal to the tick size (which is almost always true for large-tick assets). Without loss of generality, we will normalize prices so that the tick size (and hence, the bid-ask spread) is 1. We will focus
only on the ask side of the market, where limit orders are posted to sell the asset and wait to be executed against market orders from buyers. The case for the bid side can be derived similarly. We will also consider a single-exchange setup to avoid the complications of merging limit order books from different exchanges.

As we are interested in situations where the queue length is large, we ignore the integrality issues and assume that the queue position is continuous. In particular, we are interested in modeling the positional value in placing orders of infinitesimal size. We are concerned with short intraday time horizons over which an order might get executed. Over this short time period, we assume that the risk-free rate is zero since it doesn’t cost anything to borrow money intraday. Also, we assume that the agent is risk neutral. Risk neutrality is appropriate for several reasons. First of all, we are looking at a single order here, which is relatively small compared to the agent’s wealth. Therefore we can assume that the agent’s utility function is linear for this particular order. Second, we expect the agent to submit many orders to accumulate a large position. Then the law of large numbers will kick in, making the agent effectively risk neutral.
4.2.1. Order Valuation

Our goal is to estimate the value of a limit order, especially as it relates to the queue position of the order, in a dynamic multi-period setting. To this end, we consider a stylized problem where an agent arrives seeking to provide liquidity by selling an infinitesimal quantity of an asset via a limit order. The order is placed at time $t = 0$, at the best ask price $P^A$, and remains in the order book until either it transacts (i.e., is filled) or until its price changes and (by assumption) the order is canceled.

To understand the value of the order, it is necessary to develop a model for the value of the underlying asset. To this end, we assume that the asset will be liquidated at a random future time $T$, and at that time will realize a (random) cash flow $P$. The cash flow $P$ can be viewed as the fundamental value of the asset. $T$ should be viewed as the time when all information regarding the price of the underlying asset has been made public. Denote by $\{\mathcal{F}_t\}$ the filtration that represents the information possessed by the agent, at each time $t \geq 0$, and define latent efficient price process $P_t$, for $t \geq 0$, according to

$$P_t \triangleq \mathbb{E}[P|\mathcal{F}_t].$$

We will further assume that the filtration is right-continuous in the sense that $\mathcal{F}_t = \mathcal{F}_{t+}$ for all $t \geq 0$. By construction, $P_t$ is a right-continuous Doob martingale.

Now consider the case where the agent places an infinitesimal order on the ask side. At each time $t$, the agent will be willing to sell the asset at prices above $P_t$, and buy the asset at prices below $P_t$. The order will stay in the queue until it is either filled or canceled.

Now, define $\tau^* \in [0, T)$ to be the $\mathcal{F}_t$-measurable stopping time when the order is either filled or canceled. Notice that since we have defined $T$ to be the time when all information is revealed, the order should be traded or canceled before then. If the order is filled, the agent is paid $P_A$ in exchange for a short position with (eventual) fundamental value $P$. Therefore (assuming risk-neutrality and

---

1This is without loss of generality, since the buying case is symmetric.
a zero risk-free rate), the value of this order to the agent is given by

\[ V_t \triangleq \mathbb{E} \left[ (P_A - P) \mathbb{1}_{\{\text{fill}\}} \mid \mathcal{F}_t \right], \]

for all \( t \geq 0 \). For \( t \in [0, \tau^\ast) \), since \( P_t \) is a right-continuous martingale, we can apply the optional stopping theorem as in Theorem 3.22 of Karatzas and Shreve (2012):

\[ V_t = \mathbb{E} \left[ (P_A - P_t) \mathbb{1}_{\{\text{fill}\}} - (P - P_t) \mathbb{1}_{\{\text{fill}\}} \mid \mathcal{F}_t \right] \]

\[ = \alpha_t \left( \delta_t - \text{AS}_t \right), \quad (4.1) \]

where

\[ \alpha_t \triangleq \mathbb{P} (\text{fill} \mid \mathcal{F}_t), \]

\[ \delta_t \triangleq P_A - P_t, \]

\[ \text{AS}_t \triangleq \mathbb{E} \left[ (P_{\tau^\ast} - P_t) \mid \mathcal{F}_t, \text{fill} \right]. \]

These stochastic processes have natural interpretations at each time \( t \in [0, \tau^\ast) \):

- \( \alpha_t \) is the fill probability of the order.
- \( \delta_t \) captures the difference between the order’s posted price \( P_A \) and the latent efficient price \( P_t \); we call this the liquidity premium or liquidity spread earned by the order.
- \( \text{AS}_t \) measures the revision of the agent’s estimate of the asset’s fundamental value from the present time \( P_t \) to the time of a fill \( P_{\tau^\ast} \), conditional on a fill. Note that \( \text{AS}_t = 0 \) if fills are independent of the efficient price process. However, in realistic settings with asymmetrically informed traders, one typically expects that \( \text{AS}_t > 0 \). This is because of the possibility that the contra-side trader, who is demanding liquidity and paying associated spread costs by buying at the ask price, is motivated by private information about the fundamental value of the asset. Hence, trades and innovations of the efficient price process are dependent in a way.

\[ \text{For example, if } P_t \text{ happened to coincide with the mid-market price, } \delta_t \text{ would equal a half-spread. The quantity } \delta_t \text{ generalizes the notion of a “half-spread” to situations where the expected value of the asset differs from the mid-market price.} \]
that is to the detriment of the liquidity provider and, accordingly $\text{AS}_t$ is known as adverse selection. Adverse selection is an important issue in evaluating the value of limit orders, and has been noted in many studies, such as Glosten and Milgrom (1985) and Kyle (1985).

The decomposition in (4.1) can be interpreted informally as an accounting identity that breaks down the expected profitability of liquidity provision at the level of an individual order as follows:

$$\text{order value} = \text{fill probability} \times (\text{liquidity spread premium} - \text{adverse selection cost}).$$

Hollifield et al. (2004) used a similar decomposition to (4.1) to describe the agent’s expected pay-off in placing the order. Their approach is slightly general as they included an error term to represent the trader’s private value for the assets. In our model, we are looking from the perspective of competitive market makers who have no private information. As a result, the private values are assumed to be zero. Hollifield et al. (2004) do not explicitly consider queue positions, and the fill probabilities are estimated in a non-parametric way for different price levels. In fact, their approach finding the trader’s optimal submission strategy across different price levels is fundamentally static, whereas our approach estimating values of orders at different queue positions uses a dynamic model.

### 4.2.2. Price Dynamics

We assume that innovations in the latent efficient price process are driven by two types of discrete exogenous events, trades and price jumps. Trades correspond to the arrival of an impatient buyer (resp., seller), who demands immediate liquidity and is matched with a seller (resp., buyer) at the best ask (resp., bid) price. For the $i$th trade, denote its arrival time by $\tau_i^u > 0$ and its signed trade size by $u_i \in \mathbb{R}$. Price jumps, on the other hand, represent an instant in time at which price levels across the board shift up (resp., down) due to the arrival of new information. In an upward (resp., downward) jump, we assume that all orders at the best ask (resp., best bid) price are filled. We denote the arrival time of the $k$th jump by $\tau_k^J$ and its size by $J_k$.

---

3 The case where $u_i < 0$ represents a market order to sell, while $u_i > 0$ represents a market order to buy.
We posit the following dynamics for the latent efficient price:

\[
P_t = P_0 + \lambda \sum_{i: \tau^u_i \leq t} u_i + \sum_{k: \tau^J_k \leq t} J_k,
\]

or, equivalently, for liquidity premium:

\[
\delta_t = \delta_0 - \lambda \sum_{i: \tau^u_i \leq t} u_i - \sum_{k: \tau^J_k \leq t} J_k,
\]

for \( t \in [0, \tau^*) \). Accordingly, we make the following assumptions:

- **Linear price impact.** The \( i \)th trade impacts the latent efficient price by \( \lambda u_i \); i.e., there is a permanent linear price impact. The quantity \( \lambda > 0 \) captures the sensitivity of prices to trade size. This is consistent with the strategic model of [Kyle (1985)], where such price impact results from asymmetrically informed traders. Although our model is reduced form in that the price impact is specified exogenously, the spirit of it is that large trades are more likely to be due to informed traders, and hence have a greater impact on the posterior beliefs of the trader.

- **Poisson trade arrivals.** We will assume that the trade times \( \{\tau^u_i\} \) are Poisson arrivals with rate \( \mu > 0 \).

- **I.i.d. trade sizes.** We will assume that the trade sizes \( \{u_i\} \) are independent and identically distributed with probability density function \( f(\cdot) \) over \( \mathbb{R} \). In order to ensure that \( P_t \) is a martingale, we will require that \( \mathbb{E}[u_i] = 0 \). To avoid technicalities, we further assume that \( f(\cdot) \) is continuous and \( f(u) > 0 \) for all \( u \in \mathbb{R} \); i.e., the support of the distribution is all of \( \mathbb{R} \).

- **Poisson jump arrivals.** We will assume that the jump times \( \{\tau^J_k\} \) are Poisson arrivals with rate \( \gamma > 0 \).

- **I.i.d. jump sizes.** We will assume that the jump sizes \( \{J_k\} \) are independent and identically distributed. In order to ensure that \( P_t \) is a martingale, we require that \( \mathbb{E}[J_k] = 0 \).
CHAPTER 4. A MODEL FOR QUEUE POSITION VALUATION

We require that arrival times, trade sizes, and jump sizes be \( \mathcal{F}_t \)-measureable, so that \( P_t \) is an \( \mathcal{F}_t \)-adapted process with sample paths that are right continuous with left limits (RCLL) — in fact, \( P_t \) is a piecewise constant pure jump process.

In equation (4.2), we are assuming that a price moves for two reasons. First of all, a price moves when trades empty the queue at a certain price level. Second, a price moves not because of trading but because of the arrival of new information; this is what we call a jump. For example, a correlated asset move or a corporate news release may move the price through cancellation of existing orders or placement of new orders at other price levels. Another reason why we are explicitly modeling jumps is that price changes are better modeled by jumps than by other forms of continuous noise for short periods of time, see Barndorff-Nielsen and Shephard (2004).

Note that the dynamics of \( P_t \) are determined by the arrival rate parameters \((\lambda, \mu, \gamma) \in \mathbb{R}^3_+\) and the distributions of trade sizes and jump sizes. An application of the law of total variance yields, for \( t \in [0, T) \),

\[
\text{Var}(P_t) = \left(\mu \lambda^2 \sigma_u^2 + \gamma \sigma_J^2\right) t,
\]

where \( \sigma_u^2 \triangleq \text{Var}(u) \) is the variance of trade sizes and \( \sigma_J^2 \triangleq \text{Var}(J) \) is the variance of jump sizes. Expressing this as a per-unit time price volatility of the asset \( \sigma_P \), we have

\[
\sigma_P \triangleq \sqrt{\text{Var}(P_t)/t} = \sqrt{\mu \lambda^2 \sigma_u^2 + \gamma \sigma_J^2}.
\]

4.2.3. Limit Order Book Dynamics

The limit order is placed at the best ask price \( P_A \), and remains in the order book either until it is filled, or until its price changes and (by assumption) it is canceled. Moreover, during the time that is active, the order moves toward the front of its position, as orders with greater queue priority are filled or cancelled, according to price-time priority rules.

Specifically, subsequent to its placement, denote the queue position of the limit order by \( q_t \in Q \triangleq \mathbb{R}_+ \cup \{\text{FILL, CANCEL}\} \). Specifically, at each time \( t \in [0, \tau^*) \) at which the order has not been
filled or canceled, \( q_t \in \mathbb{R}_+ \) and the quantity \( q_t \) of asset shares\(^4\) available for sale at the best ask price, is of greater priority than the limit order. If the order has been filled (resp., canceled) prior to time \( t \), then \( q_t = \text{FILL} \) (resp., \( q_t = \text{CANCEL} \)). Until the order is filled or canceled, the queue position \( q_t \) evolves according to a sequence of arrivals of one of the following types of events at each event time \( \tau > 0 \):

1. A trade occurs with size \( u_i \in \mathbb{R} \). As per equation (4.2), the liquidity spread evolves according to

\[
\delta_\tau = \delta_{\tau-} - \lambda u_i.
\]

For the evolution of the queue position, there are three cases:

(a) \( u_i \in [q_{\tau-}, \infty) \). In this case, the quantity of shares is purchased at the best ask price that exceeds the limit order queue position; hence, the order is filled and realizes a final expected value of

\[
V_\tau = \mathbb{E}[P_A - P | \mathcal{F}_\tau] = \delta_\tau = \delta_{\tau-} - \lambda u_i,
\]

where, for the last inequality, we apply the price dynamics of equation (4.2).

(b) \( u_i \in [0, q_{\tau-}) \). In this case, the quantity of shares is purchased at the best ask price but it is insufficient to result in a fill; however, the order position improves according to

\[
q_\tau = q_{\tau-} - u_i > 0.
\]

(c) \( u_i \in (-\infty, 0) \). In this case, the quantity of shares is purchase; hence the queue position \( q_\tau \) remains fixed.

2. A price jump occurs with size \( J_k \in \mathbb{R} \). As per equation (4.2), the liquidity spread evolves according to

\[
\delta_\tau = \delta_{\tau-} - J_k.
\]

For the evolution of the queue position, there are two cases:

\(^4\)Here we ignore integrality issues as we are considering large queue length.
(a) $J_k > 0$. Under a positive price jump, the order is assumed to be filled and realizes a final expected value of

$$V_{\tau} = \mathbb{E}[P_A - P | \mathcal{F}_{\tau}] = \delta_{\tau} = \delta_{\tau-} - J_k.$$ 

(b) $J_k > 0$. Under a negative price jump, the price levels shift down and the order is assumed to be canceled, realizing a final value of $V_T = 0$.

3. The next event is the cancellation of a quantity of higher priority at the best ask price level. We will describe the underlying assumptions of cancellation model shortly, but for now it suffices to note that the $i$th cancellation event is associated with a proportion $\ell_i \in [0, 1]$, and therefore a fraction $1 - \ell_i$ of the shares with higher priority at the best ask price level are canceled. Hence,

$$q_{\tau} = \ell_i q_{\tau-}.$$ 

While the impact of trades is easy to model with the FIFO rule, cancellations can happen at any position in the queue. Moreover, we are interested only in the cancellations that happened in front of the current position. In order to model cancellations, we introduce the two assumptions below.

Proportional and Uniform Cancellations. We will assume that after each cancellation on the ask side, the ask queue is homogeneously contracted by a certain proportion $\ell$, where $\{\ell_i\}$ are i.i.d. with continuous p.d.f. $g(\cdot)$ over $[0, 1]$. Further, cancellations occur on the ask side at times associated with a Poisson process of rate $\eta^+$. Additionally, we assume that the cancellation happens uniformly across different queue positions. Under this assumption, the queue position of a limit order will be updated from $q$ to $\ell_i q$ after the $i$th cancellation.

Uninformed Cancellations. We assume that cancellations happen randomly and possess no extra information. Some empirical work, such as Cont et al. (2014), has argued that there is a correlation between price moves and cancellations; however, the market impact of cancellations should be much smaller than that of market orders and hence we will neglect this effect due to its
CHAPTER 4. A MODEL FOR QUEUE POSITION VALUATION

All things being equal, we expect cancellations to be larger when the queue is larger. Therefore, instead of modeling both the size and the position of cancellations, we assume proportional cancellations with a specific distribution fitted from the data. The order dynamics with cancellations is then presented as follows:

1. If the cancellation happens on the ask side with cancellation fraction \( \ell \), then the queue position of the order (currently \( q \)) is assumed to shrink to \( \ell q \).

2. If the cancellation happens on the bid side, then the referenced order is not affected at all.

4.3. Analysis

Now we consider the value of the queue position from the perspective of the agent. In Section 4.2.1, we defined the value of a limit order. In this section, under the dynamics described in Section 4.2.2 and Section 4.2.3, we will characterize this value. In what follows, we assume that the agent places his order at time 0.

Naturally, the value of a limit order is determined by the price at which the order is placed \( (P_A) \), the latent efficient price \( (P_t) \) at the time it is executed (resp., canceled), and the probability of execution. Because our price dynamics do not depend on price levels, we can consider prices relative to the ask price of time zero \( (P_A) \), which is denoted by \( \delta_t \). In addition, the probability of execution is a function of queue position according to the order dynamics in our model. Hence the value of a limit order can be uniquely determined by the state variable \( (\delta, q) \). Given that all the events in our model (trades, price jumps, and cancellations) are assumed to have Poisson arrival times, the evolution of state variable \( (\delta, q) \) over time can be viewed as a continuous-time Markov chain. By setting the uniformization parameter as \( \zeta = \mu + \gamma + \eta^+ \), we can transfer the continuous-time Markov chain to a discrete-time Markov chain (Chapter 5.8 in [Ross 1996]). Following our discussion in Section 4.2.3, the transitions of states are as follows:

- With probability \( \frac{\mu}{\zeta} \), the next event will be a trade. Suppose that the trade size is \( u \).
1. If $u < 0$, the state will be updated to $(q, \delta - \lambda u)$.

2. If $0 \leq u < q$, the state will be updated to $(q - u, \delta - \lambda u)$.

3. If $u \geq q$, the order value is realized at $\delta - \lambda u$.

- With probability $\frac{\gamma}{\zeta}$, the next event will be a price jump, with jump size $J$.
  
  1. If $J > 0$, the order value is realized at $\delta - J$.
  
  2. If $J \leq 0$, the order value is realized at 0.

- With probability $\frac{\eta}{\zeta}$, the next event will be a cancellation, with cancellation fraction $\ell$. The state will be updated to $(\ell q, \delta)$, where $\ell$ is the proportion that remains after the cancellation.

Putting together all of the above, we have the following lemma.

**Lemma 2.** The order value process $V_t$ takes the form

$$V_t = V(q_t, \delta_t),$$

for $t \in [0, \tau^*)$, where $V(\cdot)$ is the unique solution of the equation

$$V(q, \delta) = \frac{\mu}{\zeta} \mathbb{E}_u \left[ \mathbb{I}_{\{0 \leq u < q\}} V(q - u, \delta - \lambda u) + \mathbb{I}_{\{u \geq q\}} (\delta - \lambda u) + \mathbb{I}_{\{u < 0\}} V(q, \delta - \lambda u) \right]$$

$$+ \frac{\gamma}{\zeta} \mathbb{E}_J \left[ \mathbb{I}_{\{J > 0\}} (\delta - J) \right]$$

$$+ \frac{\eta}{\zeta} \mathbb{E}_\ell \left[ V(\ell q, \delta) \right],$$

for all $(q, \delta) \in \mathbb{R}_+ \times \mathbb{R}$.

In what follows, define the quantities

$$p_u^+ \triangleq \mathbb{P}(u > 0), \quad \bar{u}^+ \triangleq \mathbb{E}[u \mathbb{I}_{\{u > 0\}}], \quad p_J^+ \triangleq \mathbb{P}(J > 0), \quad \bar{J}^+ \triangleq \mathbb{E}[J \mathbb{I}_{\{J > 0\}}].$$

**Theorem 19 (Value Function for Market Maker).** The value function $V(q, \delta)$ is linear in $\delta$; that is,
it takes the form
\[ V(q, \delta) = \alpha(q)\delta - \beta(q), \]  
where the functions \( \alpha: \mathbb{R}_+ \to \mathbb{R} \) and \( \beta: \mathbb{R}_+ \to \mathbb{R} \) are uniquely determined by the integral equations

\[ \alpha(q) = \frac{\mu}{\mu \rho^+ + \gamma + \eta^+} \left\{ \rho^+ \int_0^q (\alpha(q-x)-1)f(x)\,dx \right\} + \frac{\gamma p^+}{\mu \rho^+ + \gamma + \eta^+} + \frac{\eta^+}{\mu \rho^+ + \gamma + \eta^+} \int_0^1 \alpha(\ell q)g(\ell)\,d\ell, \]  

\[ \beta(q) = \frac{\mu}{\mu \rho^+ + \gamma + \eta^+} \left\{ \int_0^q \beta(q-x)f(x)\,dx + \lambda \int_0^q (\alpha(q-x)-1)xf(x)\,dx \right\} - \lambda \bar{u}^+ (\alpha(q)-1) + \frac{\gamma \bar{J}^+}{\mu \rho^+ + \gamma + \eta^+} + \frac{\eta^+}{\mu \rho^+ + \gamma + \eta^+} \int_0^1 \beta(\ell q)g(\ell)\,d\ell, \]  

for \( q > 0 \), with boundary conditions

\[ \alpha(0) = \frac{\mu \rho^+ + \gamma_+}{\mu \rho^+ + \gamma}, \quad \beta(0) = \frac{\mu[\gamma(1-p^+)]}{(\mu \rho^+ + \gamma)^2} \lambda \bar{u}^+ + \frac{\gamma}{\mu \rho^+ + \gamma} \bar{J}^+. \]  

Theorem 19 shows that the value function is quasi-linear on the premium \( \delta \) while the coefficients are determined by the queue position. Specifically, if the order is executed, the agent will earn the premium \( \delta \) but incur cost \( \beta(q) \); if the order is not executed, the order value is just zero. Note that the Volterra integral equation (4.6) can be readily solved numerically.

In order to estimate the value function, the following parameters need to be obtained from data:

1. \( \gamma/\mu \), ratio of arrival rate of jumps to arrival rate of trades.
2. \( \eta^+/\mu \), ratio of arrival rate of cancellations to arrival rate of trades.
3. \( f(\cdot) \), distribution of trade size.
4. \( \lambda \), price impact coefficient.
5. \( p^+_j = P(J_i > 0) \), probability that a price jump is positive.
6. \( \bar{J}^+/p^+_j = E[J_i|J_i > 0] \), expected value of a positive jump.
CHAPTER 4. A MODEL FOR QUEUE POSITION VALUATION

Notice that the value function is determined by the ratio of arrival rates rather than their absolute value. Intuitively, ratios of arrival rates determine whether an order is executed, while their absolute values determine when that happens. As we do not consider the value of time, absolute values of arrival rates do not change the value of the order. Additionally, we need only the first moment of price jumps rather than their distribution. This is because the size of a price jump is used only to calculate the expected order value at the time that the price jump happens. The distribution of trade size is important as it helps to determine the optionality of an order that has been executed. The price impact coefficient captures the adverse selection cost due to trading, and hence appears only in the expression of $\beta(\cdot)$.

We can now establish the following properties of $\alpha(\cdot)$ and $\beta(\cdot)$.

**Theorem 20.**
1. Compared with equation (4.1), we have
   
   $$\alpha_t = \alpha(q), \quad AS_t = \frac{\beta(q)}{\alpha(q)}.$$

2. The probability of execution $\alpha(q)$ is non-increasing in queue position.

3. The adverse selection is positive
   
   $$\beta(q)/\alpha(q) > 0.$$

4. With no cancellations ($\eta = 0$), we have
   
   $$\lim_{q \to \infty} \alpha(q) = p_f^+, \quad \lim_{q \to \infty} \beta(q) = \bar{J}^+.$$

The first statement provides the intuition for the two coefficients. A by-product of the proof shows that the quasi-linear form of the value function in equation (4.6) is a general result that does not require a Poisson arrival of events.

The second statement shows that the probability of execution is smaller for orders with a larger queue position. This is expected due to the FIFO rule.

The third statement suggests that the adverse selection cost is always positive, which is in
line with intuition. Specifically, adverse selection can be broken down into two parts. The first part originates from price jumps, and the second comes from the asymmetric information between liquidity takers and liquidity providers.

The last statement provides the asymptotic behavior of the value function when there is no cancellation. Intuitively, if the queue position is extremely large, it is unlikely that the order will be executed by trades. Hence the probability of execution ($\alpha(q)$) is just the probability of a positive price jump. The case with cancellations is technically complicated as we assume that cancellations cause a shrinking of the queue length.

While in general it’s difficult to obtain close-form solutions to Volterra integral equations, some special cases can be solved using Laplace transform. Theorem 21 provides such an example.

**Theorem 21 (Exponential Trade Sizes).** Suppose there are no cancellations and that the trades sizes follow the exponential distribution with parameter $\theta > 0$, i.e.,

$$f(u) \triangleq \frac{\theta}{2} e^{-\theta |u|},$$

for $u \in \mathbb{R}$. Then, the value function is given by $V(\delta, q) = \alpha(q) \delta - \beta(q)$, where

$$\alpha(q) = p^+ J + \frac{\mu(1 - p^+ J) e^{-bq}}{\mu + 2\gamma},$$

$$\beta(q) = J^+ (1 - \frac{\mu}{\mu/2 + \gamma} e^{-bq}) + \frac{\lambda \mu \gamma (p^+ J - 1)}{2(\gamma + \mu/2)^2 \theta} e^{-bq} + \frac{\lambda (\gamma - \mu) \gamma (p^+ J - 1)}{2(\gamma + \mu/2)^3} q e^{-bq},$$

for all $q \geq 0$, where $b \triangleq \frac{(\gamma + \zeta) \theta}{\mu/2 + \gamma}$.

### 4.4. Empirical Calibration

Having laid the framework, we now test our model using NASDAQ ITCH data for large-tick U.S.m stocks with high liquidity. NASDAQ ITCH data is a so-called market-by-order data feed. As opposed to market-by-level data, which displays orders accumulated on price, market-by-order data contains all order-book events including limit order postings, trades, and limit order cancellations.
CHAPTER 4. A MODEL FOR QUEUE POSITION VALUATION

Market-by-order data makes it possible to reconstruct the limit order book at any given time and hence can be used to view queue position and size of individual orders at a price while remaining anonymous.

One advantage of our model is that it offers predictions of order value at different positions in the queue as a function of market primitives, and hence can be easily calibrated. In this section, we will take Bank of America (BAC) as an example to illustrate our estimation process and model results. We will first describe the calibration of our model parameters, and then solve for the predicted queue position values using the market primitives obtained.

4.4.1. Data Overview

Our attention is restricted to large-tick assets, where the queueing effect is large. Bank of America (BAC) is one of the most liquid stocks traded, with an average daily volume of 88 million shares in August 2013. The bid-ask spread is almost always equal to one tick and is large (about 7 basis points) relative to its price. Hence BAC qualifies as a large-tick asset.

A stock can be traded on multiple exchanges simultaneously. To avoid the complexity of aggregating multiple limit order books, we consider only the NASDAQ order book by using ITCH data, which provides historical data for full order depth. ITCH enables us to track the status of each order from the time it is placed to the time it is either executed or canceled. We use the database of Yahoo Finance for daily closing prices.

4.4.2. Calibrating Parameters

The main parameters involved in our model are: distribution of order size, trade arrival rate $\mu$, price jump arrival rate $\gamma$, cancellation arrival rate $\eta$, market impact $\lambda$, and jump size $J$. These parameters exhibit significant day-to-day heterogeneity as some days are more active than others. In what follows, these parameters will be estimated on a daily basis and we will see how their heterogeneity changes order values.

Price jumps are instances when the ask or the bid price changes. A trade happens when a market order (or a marketable limit order) is executed with existing limit orders. Sometimes trades
and price jumps can coincide. This happens when an execution is large enough to eliminate the entire queue and cause a price jump. In the following analysis, trades will refer to executions that do not cause price moves, while executions that are large enough to deplete the queue will be counted as price jumps. As a result, a price jump can come in the form of an order being executed with arbitrary size.

**Price Jumps.** In our settings, the size of price jumps is defined by changes in the latent efficient price. Since the latent efficient price is not observable, we assume that the price $\Delta t$ later is an unbiased estimate of the latent price after a jump.\(^5\) The intuition here is that the market will take some time ($\Delta t$) to digest and factor in the information. Hence, the size of a price jump is calculated as the price change $\Delta t$ after the price moves. $\Delta t$ is expected to differ among stocks due to differences in factors such as liquidity. Here, we take $\Delta t$ to be proportional to the expected time interval between price jumps. Notice that in this case the jump size can be smaller than one tick when a reversion happens within $\Delta t$. The number of price jumps is counted separately for both the ask side and the bid side, and then the average is taken. The arrival rate for price jumps is calculated simply by counting price jumps.

**Trades.** In our model, trade size is defined as the size of an aggressive market order. In electronic markets, once an aggressive market order comes, it is matched with the very first limit order in the queue. If, however, the aggressive market order is too large to be filled with a single limit order, it may trade with multiple resting limit orders, resulting in multiple individual fills. Notice that what we observe from the ITCH data feeds are individual fills, and therefore it is necessary to combine these fills to reconstruct the size of the original market order. We take a time window of two milliseconds, and calculate the order size by putting together the trades of the same side within that time window. If the price changes during that time, we consider the execution to be a price jump.

Our empirical results show that the shape of order size distribution closely resembles a log-normal distribution, which is consistent with findings in [Kyle and Obizhaeva (2016a)]. In particular, we obtained the MLE estimate of the mean and standard deviation. We obtained the arrival rate,\(^5\) In our analysis, $\Delta t$ is set at one minute.
however, in a much more straightforward manner, we simply counted the number of trades.

**Cancellations.** With *market-by-order* data, we can keep close track of the position and size of every canceled order. As we mentioned in Section 4.3, we view each cancellation as a contraction of the whole queue. In other words, we assume that each cancellation decreases the queue size uniformly by a certain proportion $l$. We then fit the cancellation proportion $l$ using Beta distribution. Note that this requires an underlying assumption that the positions where cancellations happen follow a uniform distribution. In reality, this doesn’t always hold as cancellations tend to concentrate at the end of the queue. However, due to the technicality of taking into account cancellation positions, we have to trade some accuracy for an empirically solvable model.

**Market Impact.** The calibration of market impact has always been of great interest in the market microstructure literature. Kyle (1985) formulated a linear market impact in a continuous-time theoretical model. He argued that the price impact of one unit of asset is determined by the fundamental volatility and variance of order-flow imbalance. Other researchers, such as Breen et al. (2002), took a purely empirical approach by regressing the price changes on order-flow imbalances. In this paper, we derive the market impact parameter by following the market invariant approach of Kyle and Obizhaeva (2016a). Specifically, Kyle and Obizhaeva (2016a) proposed a model in which the market impact parameter $\lambda$ is given by the following equation:

$$\lambda = C(P\sigma)^\frac{4}{3}V^{-\frac{2}{3}},$$  \hspace{1cm} (4.11)

where $C$ is some known constant, $P$ is the asset price, $\sigma$ is the asset’s volatility of daily return, and $V$ is the daily trading volume (in shares).

**Liquidity Premium.** In reality, the latent price is not observable. We will assume that on average it can be approximated by the mid-price. In other words, we will assume that the liquidity premium is a half-spread. However, we will make an adjustment in order to factor in a liquidity rebate of 0.3 ticks offered by NASDAQ. The rebate is offered by the exchange in order to encourage

---

$^6$C = 0.0156 according to Kyle and Obizhaeva (2016a).
market participants to provide liquidity. Hence the liquidity premium is just
\[
\delta_0 = \text{(half-spread)} + \text{(rebate)} = 0.8 \text{(ticks)}.
\]

Table 4.1 provides the estimated parameters for Bank of America over 22 trading days. As we can see, the average jump size is very close to one tick, which means that the price process is driven primarily by single tick jumps. Note that the jump size can be less than one tick as we approximate it as the price change \(\Delta t\) after the price moves. Our empirical findings show that the order size distribution is roughly consistent across trading days. The market impact parameter \(\lambda\) too is subject to very little variation across trading days. The only parameters with much variation from day to day are the ratios between arrival rates \((\gamma/\mu, \eta/\mu)\), which, as will see, are the driving force of intraday heterogeneity.

<table>
<thead>
<tr>
<th>Date</th>
<th>(\bar{\mu})</th>
<th>Average Trade Size (shares)</th>
<th>Average Cancellation Size (shares)</th>
<th>(\gamma)</th>
<th>Average Jump Size (ticks)</th>
<th>(\lambda)</th>
<th>Average Queue Size (shares)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8/30/13</td>
<td>1.43</td>
<td>2270</td>
<td>4793</td>
<td>1971</td>
<td>0.85</td>
<td>0.86</td>
<td>3.91</td>
</tr>
<tr>
<td>8/29/13</td>
<td>1.25</td>
<td>2635</td>
<td>6535</td>
<td>2103</td>
<td>0.57</td>
<td>0.86</td>
<td>3.97</td>
</tr>
<tr>
<td>8/28/13</td>
<td>1.55</td>
<td>2526</td>
<td>4463</td>
<td>2140</td>
<td>0.79</td>
<td>1.12</td>
<td>4.41</td>
</tr>
<tr>
<td>8/27/13</td>
<td>2.25</td>
<td>2435</td>
<td>5395</td>
<td>2049</td>
<td>1.05</td>
<td>0.91</td>
<td>4.79</td>
</tr>
<tr>
<td>8/26/13</td>
<td>1.27</td>
<td>2481</td>
<td>5441</td>
<td>2114</td>
<td>0.51</td>
<td>0.82</td>
<td>3.53</td>
</tr>
<tr>
<td>8/23/13</td>
<td>1.11</td>
<td>2058</td>
<td>3922</td>
<td>2340</td>
<td>0.56</td>
<td>0.85</td>
<td>3.68</td>
</tr>
<tr>
<td>8/22/13</td>
<td>1.20</td>
<td>2998</td>
<td>6082</td>
<td>1835</td>
<td>0.53</td>
<td>0.92</td>
<td>3.84</td>
</tr>
<tr>
<td>8/21/13</td>
<td>1.74</td>
<td>2091</td>
<td>4329</td>
<td>2134</td>
<td>1.39</td>
<td>1.14</td>
<td>3.93</td>
</tr>
<tr>
<td>8/20/13</td>
<td>1.38</td>
<td>3669</td>
<td>7157</td>
<td>1970</td>
<td>0.95</td>
<td>1.03</td>
<td>4.19</td>
</tr>
<tr>
<td>8/19/13</td>
<td>1.21</td>
<td>1979</td>
<td>3676</td>
<td>1868</td>
<td>0.72</td>
<td>0.84</td>
<td>4.87</td>
</tr>
<tr>
<td>8/16/13</td>
<td>1.26</td>
<td>2223</td>
<td>4512</td>
<td>1737</td>
<td>0.67</td>
<td>0.94</td>
<td>5.37</td>
</tr>
<tr>
<td>8/15/13</td>
<td>1.94</td>
<td>1582</td>
<td>3037</td>
<td>1698</td>
<td>0.82</td>
<td>0.75</td>
<td>5.23</td>
</tr>
<tr>
<td>8/14/13</td>
<td>1.37</td>
<td>2034</td>
<td>4449</td>
<td>1455</td>
<td>0.90</td>
<td>0.95</td>
<td>4.51</td>
</tr>
<tr>
<td>8/13/13</td>
<td>1.54</td>
<td>2327</td>
<td>5008</td>
<td>1628</td>
<td>0.72</td>
<td>1.09</td>
<td>4.68</td>
</tr>
<tr>
<td>8/12/13</td>
<td>1.08</td>
<td>2426</td>
<td>4977</td>
<td>1887</td>
<td>0.49</td>
<td>0.92</td>
<td>4.39</td>
</tr>
<tr>
<td>8/9/13</td>
<td>1.65</td>
<td>2762</td>
<td>5986</td>
<td>1875</td>
<td>0.71</td>
<td>0.91</td>
<td>4.81</td>
</tr>
<tr>
<td>8/8/13</td>
<td>1.29</td>
<td>2303</td>
<td>4351</td>
<td>1764</td>
<td>0.95</td>
<td>1.25</td>
<td>5.16</td>
</tr>
<tr>
<td>8/7/13</td>
<td>1.99</td>
<td>2494</td>
<td>5106</td>
<td>2303</td>
<td>1.01</td>
<td>0.91</td>
<td>5.99</td>
</tr>
<tr>
<td>8/6/13</td>
<td>1.47</td>
<td>2610</td>
<td>5692</td>
<td>1930</td>
<td>0.55</td>
<td>0.93</td>
<td>5.19</td>
</tr>
<tr>
<td>8/5/13</td>
<td>0.81</td>
<td>1598</td>
<td>2941</td>
<td>1502</td>
<td>0.44</td>
<td>0.78</td>
<td>4.59</td>
</tr>
<tr>
<td>8/2/13</td>
<td>1.57</td>
<td>2039</td>
<td>4545</td>
<td>1511</td>
<td>0.53</td>
<td>0.86</td>
<td>5.05</td>
</tr>
<tr>
<td>8/1/13</td>
<td>1.58</td>
<td>2853</td>
<td>7978</td>
<td>1854</td>
<td>0.71</td>
<td>0.85</td>
<td>5.52</td>
</tr>
</tbody>
</table>

Table 4.1: Estimated market parameters for BAC in a month. \(\lambda\) is estimated as the price impact in basis points for one percent of daily volume. Note that here we consider only shares traded on NASDAQ.
4.4.3. Observations

Given the market parameters estimated above, the main output of our model is the value function of queue position, which can be obtained by numerically solving equation (4.6) and (4.7) in Section 4.3. Figure 4.2 provides the plots of the value function, execution probability, and adverse selection for BAC on two representative trading days (8/9/2013 and 8/20/2013).

First, as predicted by Theorem 20, the probability of execution is decreasing with queue length and becomes quite flat when the queue length is large. Intuitively, when the queue length is extremely large, the order on the ask side can be executed only by positive price jumps. Hence, the execution probability should converge toward the probability of a positive price jump ($p_J^+$ as in Theorem 20). Second, the adverse selection cost remains positive and is increasing with queue length. Intuitively, this is because orders at the end of a large queue are more likely to be executed against a large trade. With our assumption of linear price impact, large trades translate to higher adverse selection costs. Third, the order value curve is decreasing as the queue gets longer. From equation (4.1), we can see that the decreasing value curve is due to a combined effect of decreasing execution probability and increasing adverse selection cost. Fourth, the value difference between an order placed at the very front of the queue and an order placed in a queue length of average was about 0.26 ticks on 8/9/2013 and 0.21 ticks on 8/20/2013, which is comparable to the bid-ask spread. This shows that the queue’s positional value cannot be neglected in higher-level control problems such as optimal execution and market making. Finally, Figure 4.2 provides comparisons of model outputs on two different trading days. We can see that orders in the same queue position were worth less on 8/20/2013, and had a lower fill probability. This is because the ratio of arrival rate $\gamma/\mu$ was significantly higher on 8/20/2013 (0.69) than on 8/9/2013 (0.43). Intuitively, large $\gamma/\mu$ means that the order is less likely to be executed against a trade before the price changes, and hence translate to a lower fill probability.
Figure 4.2: Model outputs as functions of queue positions on two different trading days (08/09/2013 and 08/20/2013). The red dots represent the average queue length of that trading day.
4.5. Empirical Validation: Backtesting

In the previous section, we calibrated a parametric model to estimate the positional value of limit orders using market data. Now we want to verify these predictions using a non-parametric model based on backtesting. The difficulty is that the order value cannot be measured by the profitability of the orders in the limit order book, since actual orders may have private information. Therefore, instead of actual orders, we have to use randomly placed artificial orders.

Market-by-order data enables us to simulate the life-span of each artificial order in the limit order book. We can then calculate various statistics such as order value and fill probability for orders at different positions. We then compare the backtesting results with the parametric estimations. More specifically, we restrict our attention to 9 highly liquid U.S. equities or ETFs with a bid/ask spread close to 1 tick. A list of the stocks and their descriptive statistics are given in Table 4.2.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Exchange</th>
<th>Low ($)</th>
<th>High ($)</th>
<th>Bid-Ask Spread ($</th>
<th>Average Volatility (daily)</th>
<th>Average Daily Volume (shares, $10^6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bank of America</td>
<td>BAC</td>
<td>NYSE</td>
<td>14.11</td>
<td>14.95</td>
<td>1.017</td>
<td>1.2%</td>
</tr>
<tr>
<td>Cisco</td>
<td>CSCO</td>
<td>NASDAQ</td>
<td>23.31</td>
<td>26.38</td>
<td>0.996</td>
<td>1.0%</td>
</tr>
<tr>
<td>General Electric</td>
<td>GE</td>
<td>NYSE</td>
<td>23.11</td>
<td>24.70</td>
<td>1.002</td>
<td>0.9%</td>
</tr>
<tr>
<td>Ford</td>
<td>F</td>
<td>NYSE</td>
<td>15.88</td>
<td>17.50</td>
<td>1.005</td>
<td>1.4%</td>
</tr>
<tr>
<td>Intel</td>
<td>INTC</td>
<td>NASDAQ</td>
<td>21.90</td>
<td>23.22</td>
<td>1.005</td>
<td>1.1%</td>
</tr>
<tr>
<td>Pfizer</td>
<td>PFE</td>
<td>NYSE</td>
<td>28.00</td>
<td>29.37</td>
<td>1.007</td>
<td>0.7%</td>
</tr>
<tr>
<td>Petroleo Brasileiro</td>
<td>PBR</td>
<td>NYSE</td>
<td>13.39</td>
<td>14.98</td>
<td>1.010</td>
<td>2.6%</td>
</tr>
<tr>
<td>iShares MSCI Emerging Markets</td>
<td>EEM</td>
<td>NYSE</td>
<td>37.35</td>
<td>40.10</td>
<td>1.006</td>
<td>1.2%</td>
</tr>
<tr>
<td>iShares MSCI EAFE</td>
<td>EFA</td>
<td>NYSE</td>
<td>59.17</td>
<td>62.10</td>
<td>1.021</td>
<td>0.7%</td>
</tr>
</tbody>
</table>

Table 4.2: Descriptive statistics for 9 stocks over the 21 trading days of August 2013. The average bid/ask spread is defined as the time average computed from the ITCH data. The volatility is defined as the standard deviation of percentage daily returns. All other statistics were retrieved from Yahoo Finance.

4.5.1. Backtesting Simulation

The technique of backtesting is widely used in the financial industry to test a predictive model with existing historical data. Our paper benefited from the advantage of accessing ITCH data, a source of market-by-order data provided by NASDAQ. With full information on historical order/trade data, we were able to construct a simulator to backtest our proposed valuation model. Backtesting
CHAPTER 4. A MODEL FOR QUEUE POSITION VALUATION

with artificial orders poses one challenge: real orders may influence other market participants. Here we will assume that all the artificial orders are of infinitesimal size and hence have no market impact. This is actually in accordance with our model setup. First, the historical data will be used to create the dynamics of order books; then artificial orders will be placed and processed according to market rules; finally, the value of the artificial orders will be calculated.

**Placement of Artificial Orders.** We start by defining two types of artificial orders based on the position at which they are inserted.

- **Regular orders** are orders that are appended to the end of the queue at the current best price. The name *regular orders* comes from the fact that these orders are placed according to the FIFO rule.

- **Touch orders** are orders that are inserted at the very front of the queue at the current best price. These orders are used to evaluate the value of being placed at the front of the queue. Comparing *touch orders* with *regular orders* will help to illustrate the magnitude of the effect of queue positions.

In the simulation, we associate each real limit order with an *entry-time stamp* to keep track of the time that the order entered the order book. The side (bid or ask) of each artificial order is randomly picked. Suppose that it is an ask order; then its evolution in the limit order book will be as follows.

- The artificial order wakes up at a random time, and is inserted in the queue according to its type.

- The process of artificial orders follows the market rule of price/time priority. We start updating the limit order book according to the real data until one of the following events occurs.

1. **New order arrival:** If a new limit order is added to the same side at a better price (lower for the ask side, higher for the bid side) than that of the artificial order, then the artificial order will no longer be at the best price, and we will assume that it is canceled immediately.
2. **Fill**: If a limit order arrives after the artificial order is filled, we will assume that the artificial order is also filled.

3. **Cancellation**: If the price moves because all other orders in the queue are canceled (which is rare), we will assume that the artificial order is canceled as well.

We ignore the first and last half hours of the trading day because the market tends to be very volatile during these two periods. Accordingly, we pick 1000 time points uniformly at random between 10:00 and 15:30 on each trading day on the random side of the market.

**Order Valuation.** If the artificial order is canceled then it possesses no value. If, however, the artificial order is filled then its value will be the difference between the execution price and the fundamental value of the asset. In order to backtest order values at different positions, we need to determine the fundamental value. Since the fundamental value cannot be observed directly in the historical data, we need to calibrate it through a tractable valuation process. In this paper, we assume that the mid-price one minute after the order’s execution is an unbiased estimate of its fundamental value. This is certainly a noisy approximation and lots of data are needed for a reasonably accurate estimate, which is why we choose to estimate the average order value over 30 trading days instead of using a shorter period.

### 4.5.2. Observations

Table 4.3 shows the comparison of the results from backtesting and model outputs. The order value measures the value of *regular orders* that are placed at the end of the queue, while the touch value measures the value of *touch orders* placed at the very front of the queue.

We can see that the values estimated from our model are very close to the backtesting results. Further, if we break down the value into fill probability and adverse selection cost, we can see that the values are still close. This shows that our model provides a good approximation of the value of queue positions.

Notice that the difference between the value of *touch orders* and the value of *regular orders* provides good intuitions about the magnitude of the value of queue positions. First of all, the value of orders placed at the front of the queue is always larger than the value of orders placed at the
CHAPTER 4. A MODEL FOR QUEUE POSITION VALUATION

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Order Value Model (ticks)</th>
<th>Order Value Simulation (ticks)</th>
<th>Fill Probability Model</th>
<th>Fill Probability Simulation</th>
<th>Adverse Selection Model (ticks)</th>
<th>Adverse Selection Simulation (ticks)</th>
<th>Touch Value Model (ticks)</th>
<th>Touch Value Simulation (ticks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAC</td>
<td>0.14</td>
<td>0.14</td>
<td>0.62</td>
<td>0.60</td>
<td>0.57</td>
<td>0.57</td>
<td>0.36</td>
<td>0.31</td>
</tr>
<tr>
<td>CSCO</td>
<td>0.08</td>
<td>0.07</td>
<td>0.63</td>
<td>0.59</td>
<td>0.68</td>
<td>0.68</td>
<td>0.24</td>
<td>0.21</td>
</tr>
<tr>
<td>GE</td>
<td>0.08</td>
<td>0.09</td>
<td>0.62</td>
<td>0.60</td>
<td>0.67</td>
<td>0.65</td>
<td>0.19</td>
<td>0.23</td>
</tr>
<tr>
<td>F</td>
<td>0.13</td>
<td>0.15</td>
<td>0.65</td>
<td>0.64</td>
<td>0.60</td>
<td>0.53</td>
<td>0.24</td>
<td>0.23</td>
</tr>
<tr>
<td>INTC</td>
<td>0.11</td>
<td>0.09</td>
<td>0.64</td>
<td>0.61</td>
<td>0.63</td>
<td>0.56</td>
<td>0.28</td>
<td>0.23</td>
</tr>
<tr>
<td>PFE</td>
<td>0.12</td>
<td>0.11</td>
<td>0.63</td>
<td>0.58</td>
<td>0.62</td>
<td>0.61</td>
<td>0.16</td>
<td>0.21</td>
</tr>
<tr>
<td>PBR</td>
<td>-0.03</td>
<td>-0.04</td>
<td>0.57</td>
<td>0.53</td>
<td>0.85</td>
<td>0.89</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>EMM</td>
<td>0.07</td>
<td>0.08</td>
<td>0.63</td>
<td>0.63</td>
<td>0.69</td>
<td>0.64</td>
<td>0.21</td>
<td>0.15</td>
</tr>
<tr>
<td>EFA</td>
<td>0.03</td>
<td>0.04</td>
<td>0.57</td>
<td>0.53</td>
<td>0.74</td>
<td>0.73</td>
<td>0.06</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Table 4.3: Estimated model values vs. simulation values. All the values above were calculated as the average across 30 trading days. Touch value refers to the value of orders at the very front of the queue.

end. This shows that better queue position does carry an advantage. Second, the magnitude of the gap differed between symbols. For some symbols, such as BAC and CSCO, the gap can be very large and comparable to the bid ask spread (> 0.1 ticks). For others, such as PFE and PBR, the gap is less prominent (< 0.1 ticks).

4.5.3. Discussion

In this section, we provide a framework based on backtesting to estimate the value of queue positions. This non-parametric approach enables us to test the accuracy of our model. But if a non-parametric model is available, why do we still need a parametric one, such as the one discussed in this paper? The reasons are as follows. In the backtest, artificial orders are placed randomly across time to simulate a real situation. As a result, it can be used only to estimate the average value across time. However, market parameters, such as arrival rates of order book events, are constantly changing and backtesting cannot capture that variation. Additionally, the estimates from our model are conditional on market primitives and hence provide more precise predictions in real time.
4.6. Concluding Remarks

In this paper, we exhibited a dynamic model for valuing queue position in limit order books. We provided analytic evidence for sizable difference in values for orders at different queue positions. We specifically quantified the disadvantage of bad queue positions that originate from decreasing execution probability and increasing adverse selection costs.

The formulation of the model is based entirely on observable quantities so that the parameters can be estimated from market data. This tractability allowed us to calibrate our model empirically. We further validated the model by comparing the outputs with results from backtesting simulations.

This analysis has practical implications for both market participants and regulators.

1. For large tick-size assets, queueing effects can be very significant.

2. Accounting for queue position cannot be ignored when solving market making or algorithmic trading problems. This gives rise to various exotic order types that enable traders to jump to better queue positions. If we look from the other direction, we may conclude that trades need to respond faster to jump to the front of the queue. In this respect, our analysis partly explains the “speed competition” between high-frequency trading firms.

3. The value embedded in the queue position rewards the trading speed of high-frequency firms. This creates a disadvantage for individual traders who have less or no access to fast-trading technologies. From a regulatory level, an important question is whether this time-price priority rule is a good mechanism for organizing exchanges of large-tick assets.

4. One possible future research direction is to expand the model to accommodate other market properties such as volatility.
Bibliography


BIBLIOGRAPHY


Appendix A

A.1. Additional Proofs for Chapter 2

**Theorem 1.** A necessary and sufficient condition for the existence of an equilibrium with linear price impact functions is that the two CCPs have common views on market impact, i.e., that \( G_1 = G_2 \equiv G \).

In this case, all equilibria are determined by the symmetric, positive definite solutions \( F_1, F_2 \in \mathbb{R}^{m \times m} \) to the equation

\[
G^{-1} = F_1^{-1} + F_2^{-1}.
\]

**Proof.** We will make frequent use of the fact that our definitions require the matrices \( F_i \) and \( G_i \) to be symmetric and positive definite.

**Necessity.** Suppose \((x_1, x_2, F_1, F_2)\) defines an equilibrium.

The first-order conditions for the clearing member’s optimization problem (2.9) yield

\[
x_1 = (F_1 + F_2)^{-1} F_2 x.
\]

(A.1)

The sufficient margin condition for CCP 1 implies

\[
x_1^\top G_1 x = x_1^\top F_1 x_1,
\]
for all $x$. We can use (A.1) to write this as

$$x_1^\top G_1(F_1 + F_2)F_2^{-1}x_1 = x_1^\top F_1x_1.$$  \hspace{1cm} (A.2)

We need this to hold for all $x_1 \in \mathbb{R}^m$ because from (A.1) we see that $x_1$ ranges over all of $\mathbb{R}^m$ as $x$ does. Thus, the matrices on the two sides of (A.2) must have the same symmetric parts. Applying the same argument to CCP 2, this yields

$$F_1 = \frac{1}{2}(G_1F_2^{-1}F_1 + F_1F_2^{-1}G_1) + G_1,$$

$$F_2 = \frac{1}{2}(G_2F_1^{-1}F_2 + F_2F_1^{-1}G_2) + G_2.$$

We can rewrite these equations as

$$F_1 = \frac{1}{2}(I + F_1F_2^{-1})G_1 + \frac{1}{2}G_1(I + F_2^{-1}F_1),$$  \hspace{1cm} (A.3)

$$F_2 = \frac{1}{2}(I + F_2F_1^{-1})G_2 + \frac{1}{2}G_2(I + F_1^{-1}F_2).$$  \hspace{1cm} (A.4)

Each of these equations has the form

$$B = AX + X^\top A^\top$$

According to Braden (1998, Theorem 1), the solutions to (A.3) and (A.4) take the following form: for some skew-symmetric matrices $Q_1, Q_2$,

$$G_1 = (I + F_1F_2^{-1})^{-1}F_1 + \frac{1}{2}Q_1(I + F_1F_2^{-1}),$$

$$G_2 = (I + F_2F_1^{-1})^{-1}F_2 + \frac{1}{2}Q_2(I + F_2F_1^{-1}).$$

Making the substitutions

$$(I + F_1F_2^{-1})^{-1} = F_2(F_2 + F_1)^{-1}, \quad (I + F_2F_1^{-1})^{-1} = F_1(F_2 + F_1)^{-1},$$

A square matrix $A$ is skew-symmetric if it satisfies the condition $-A = A^\top$.
we get

\[ G_1 = F_2(F_2 + F_1)^{-1}F_1 + \frac{1}{2}Q_1(I + F_1F_2^{-1}), \]  
\[ G_2 = F_1(F_2 + F_1)^{-1}F_2 + \frac{1}{2}Q_2(I + F_2F_1^{-1}). \]  
\[ \text{(A.5)} \]
\[ \text{(A.6)} \]

Next observe that for any symmetric, invertible \( A, B \),

\[ A(A + B)^{-1}B = A(I + B^{-1}A)^{-1} \]
\[ = A \left[A^{-1}(A^{-1} + B^{-1})^{-1}\right] \]
\[ = (A^{-1} + B^{-1})^{-1}. \]

Thus, we can write \([\text{A.5})-\text{(A.6)}\) as

\[ G_1 = (F_1^{-1} + F_2^{-1})^{-1} + \frac{1}{2}Q_1(I + F_1F_2^{-1}), \]  
\[ G_2 = (F_1^{-1} + F_2^{-1})^{-1} + \frac{1}{2}Q_2(I + F_2F_1^{-1}). \]  
\[ \text{(A.7)} \]
\[ \text{(A.8)} \]

We will show that \( Q_1 = Q_2 = 0 \). It will then follow that

\[ G_1 = (F_1^{-1} + F_2^{-1})^{-1} = G_2 \triangleq G \]

and therefore

\[ G^{-1} = F_1^{-1} + F_2^{-1}. \]  
\[ \text{(A.9)} \]

It remains to show that \( Q_1 = Q_2 = 0 \). Observe that the first term on the right side of \([\text{A.7})\]
and \([\text{A.8})\) is symmetric, so the last term must be symmetric as well. Also, because the \( F_i \) are positive definite, \( F_1F_2^{-1} \) and \( F_2F_1^{-1} \) have positive eigenvalues (even though they are not necessarily positive definite). Thus, it suffices to show that if \( Q \) is skew-symmetric and \( X = F_1F_2^{-1} \) has positive eigenvalues, then \( Q(I + X) \) cannot be symmetric unless \( Q = 0 \).
APPENDIX A. APPENDIX

If \( Q(I + X) \) is symmetric, \( Q + QX = -Q + X^TQ^T \) and

\[
2Q = (X^TQ^T - QX). \tag{A.10}
\]

Any skew-symmetric matrix \( Q \) can be written in the form \( Q = U\Lambda U^T \), where \( U \) is orthogonal, and

\[
\Lambda = \begin{pmatrix}
0 & \lambda_1 \\
-\lambda_1 & 0 \\
& \ddots \\
0 & \lambda_{m-k} \\
-\lambda_{m-k} & 0
\end{pmatrix},
\]

where \( 0_{k \times k} \) is a block of zeros, for some \( k \). We always have \( m - k \) even, and \( k \) may be zero if \( m \) is even. We can write (A.10) as

\[
2U\Lambda U^T = (X^TU\Lambda^T U^T - U\Lambda U^T X)
\]

and then

\[
2\Lambda = (U^T X^TU\Lambda^T - \Lambda U^T XU) = (\tilde{X}^T\Lambda^T - \Lambda\tilde{X}),
\]

where \( \tilde{X} \) has the same eigenvalues as \( X \). So, it suffices to consider (A.10) in the case \( Q = \Lambda \),

\[
2\Lambda = (X^T\Lambda^T - \Lambda X). \tag{A.11}
\]

With \( \Lambda \) as given above, we claim that \( X \) must have a block decomposition

\[
X = \begin{pmatrix}
A & 0_{m-k \times k} \\
C & B
\end{pmatrix}. \tag{A.12}
\]

If \( k = 0 \), there is nothing to prove, so suppose \( k \geq 1 \). Consider any \( X_{ij} \) with \( i \leq m - k \) and
\( j > m - k \). Denote by \( \Lambda_{\ell i} \) the unique nonzero entry in the \( i \)th column of \( \Lambda \). Then if (A.11) holds,

\[
0 = 2\Lambda_{\ell j} = (\Lambda X)_{\ell j} - (\Lambda X)_{\ell j} = \sum_m \Lambda_{jm}X_{m\ell} - \Lambda_{\ell i}X_{ij} = -\Lambda_{\ell i}X_{ij},
\]

so \( X_{ij} = 0 \), which confirms (A.12). As a consequence of (A.11) and (A.12), we have

\[
2\lambda_1 = 2\Lambda_{12} = (\Lambda A)_{21} - (\Lambda A)_{12} = -\lambda_1 A_{11} - \lambda_1 A_{22},
\]

so \( A_{11} + A_{22} = -2 \). The same calculation applies for all \( \lambda_2, \ldots, \lambda_{(m-k)/2} \), so the trace of \( A \) is negative (in fact, equal to \( -(m-k) \)), so \( A \) must have at least one negative eigenvalue. But from (A.12) we see that every eigenvalue of \( A \) is an eigenvalue of \( X \), and we know that \( X \) has only positive eigenvalues. We conclude that the only solution to (A.11) is \( \Lambda = 0 \).

**Sufficiency.** Suppose the CCPs have common views on market impact \( G_1 = G_2 = G \), and suppose \( F_1, F_2 \) satisfy (2.8). Then (A.1) and (A.2) hold, and \( F_1, F_2 \) define an equilibrium. ■

**Proposition 2.** Suppose that the CCPs differ in their price impact belief matrices \( G_1, G_2 \in \mathbb{R}^{m \times m} \). Then:

(i) the matrices \((F_1(t + 1), F_2(t + 1))\) defined in (2.10)–(2.11) are positive definite for all \( t \geq 0 \),

(ii) if the spectral radius of \( G_1^{-1}G_2 \) is strictly less than 1, as \( t \to \infty \),

\[
F_2(t)F_1(t)^{-1} \to 0, \quad x_1(t) \to 0, \quad x_2(t) \to x.
\]

**Proof.** In order to establish part (ii) we will prove the following by induction: for all times \( t \geq 0 \),

\[
F_i(t) > 0, \quad i \in \{1, 2\}, \quad (A.13)
\]

\[
F_1(t)F_2(t)^{-1} = (G_1G_2^{-1})^{t+1}. \quad (A.14)
\]

Clearly (A.13)–(A.14) hold when \( t = 0 \).
Suppose they hold for $t$. Then, substituting (A.14) in (2.10)–(2.11),

\[
F_1(t + 1) = \frac{1}{2} \left[ G_1 \left( G_2^{-1} G_1 \right)^{t+1} + 2G_1 + \left( G_1 G_2^{-1} \right)^{t+1} G_1 \right] = G_1 + G_1 \left( G_2^{-1} G_1 \right)^{t+1},
\]

\[
F_2(t + 1) = \frac{1}{2} \left[ G_2 \left( G_1^{-1} G_2 \right)^{t+1} + 2G_2 + \left( G_2 G_1^{-1} \right)^{t+1} G_2 \right] = G_2 + G_2 \left( G_1^{-1} G_2 \right)^{t+1}.
\]

Then, since $G_1, G_2 \succ 0$, clearly (A.13) holds at time $t + 1$. Further,

\[
F_1(t + 1) F_2(t + 1)^{-1} = G_1 \left[ I + \left( G_2^{-1} G_1 \right)^{t+1} \right] \left[ I + \left( G_1^{-1} G_2 \right)^{t+1} \right]^{-1} G_2^{-1}
\]

\[
= G_1 \left( G_2^{-1} G_1 \right)^{t+1} G_2^{-1}
\]

\[
= \left( G_1 G_2^{-1} \right)^{t+2},
\]

establishing (A.14) at time $t + 1$.

For part (ii), since the spectral radius of $G_1^{-1} G_2$ is less than 1,

\[
\lim_{t \to \infty} \left( G_1^{-1} G_2 \right)^{t} = 0.
\]

This implies that

\[
\lim_{t \to \infty} x_2(t) = \lim_{t \to \infty} \left( F_1(t) + F_2(t) \right)^{-1} F_1(t) x
\]

\[
= \lim_{t \to \infty} \left[ I + F_1(t)^{-1} F_2(t) \right]^{-1} x
\]

\[
= \lim_{t \to \infty} \left[ I + (G_1^{-1} G_2)^{t} \right]^{-1} x
\]

\[
= x,
\]

Further,

\[
\lim_{t \to \infty} x_1(t) = \lim_{t \to \infty} x - x_2(t) = 0.
\]

\[\square\]

**Theorem 2.** A necessary and sufficient condition for a partitioned equilibrium with linear price...
APPENDIX A. APPENDIX

impact is that the price impact belief matrices $G_1, G_2$ have a common block diagonal structure

$$G_i = \begin{pmatrix} G_i(1,1) \\ G_i(2,2) \\ G_i(3,3) \end{pmatrix}, \quad i \in \{1, 2\}, \quad \quad (2.14)$$

with $G_i(1,1) \in \mathbb{R}^{m_1 \times m_1}, G_i(2,2) \in \mathbb{R}^{m_2 \times m_2}, G_i(3,3) \in \mathbb{R}^{m_3 \times m_3}$, where the submatrices satisfy

$$G_1(2,2) = G_2(2,2) \triangleq G(2,2). \quad (2.15)$$

In this case, CCP 1 clears the first $m_1 + m_2$ swap types, CCP 2 clears the last $m_2 + m_3$ swap types, and they choose margin matrices

$$F_1 = \begin{pmatrix} G_1(1,1) \\ F_1(2,2) \end{pmatrix}, \quad F_2 = \begin{pmatrix} F_2(2,2) \\ G_2(3,3) \end{pmatrix}, \quad \quad (2.16)$$

for any symmetric, positive definite matrices $F_1(2,2), F_2(2,2) \in \mathbb{R}^{m_2 \times m_2}$ satisfying

$$F_1(2,2)^{-1} + F_2(2,2)^{-1} = G(2,2)^{-1}. \quad (2.17)$$

**Proof.** Sufficiency. Let the number of rows (and columns) in the three blocks be $m_1, m_2,$ and $m_3$. We claim that we get an equilibrium if CCP 1 clears the first $m_1 + m_2$ security types, CCP 2 clears the last $m_2 + m_3$ security types, and they choose margin matrices

$$F_1 = \begin{pmatrix} G_1(1,1) \\ F_1(2,2) \end{pmatrix}, \quad F_2 = \begin{pmatrix} F_2(2,2) \\ G_2(3,3) \end{pmatrix}, \quad \quad (A.15)$$

for some symmetric $F_1(2,2), F_2(2,2)$ satisfying

$$F_1(2,2)^{-1} + F_2(2,2)^{-1} = G(2,2)^{-1}. \quad (A.16)$$
To show that this holds, for any $x \in \mathbb{R}^m$ we can write
\[
x = \begin{pmatrix} u \\ v \\ w \end{pmatrix}, \quad x_1 = \begin{pmatrix} u \\ v_1 \end{pmatrix}, \quad x_2 = \begin{pmatrix} v - v_1 \\ w \end{pmatrix},
\]
u \in \mathbb{R}^{m_1}, \ v, v_1 \in \mathbb{R}^{m_2}, \text{ and } w \in \mathbb{R}^{m_3}. \text{ The minimization over } (x_1, x_2) \text{ in (2.12) reduces to a}
\text{minimization over } v_1 \text{ with solution}
\]
v_1 = (F_1(2, 2) + F_2(2, 2))^{-1} F_2(2, 2) v.
\]
To verify the first condition in (2.13) observe that
\]
x_1^T F_1 x_1 = u^T G_1(1, 1) u + v_1^T F_1(2, 2) v_1 \tag{A.17}
\]
and
\]
x_1^T P_1 G_1 x = u^T G_1(1, 1) u + v_1^T G(2, 2) v. \tag{A.18}
\]
But (A.16) implies that
\]
G(2, 2) = (F_1^{-1}(2, 2) + F_2^{-1}(2, 2))^{-1} = F_1(2, 2)(F_1(2, 2) + F_2(2, 2))^{-1} F_2(2, 2)
\]
so (A.17) and (A.18) are equal. A similar argument verifies the second condition in (2.13).

Necessity. We now show that if $(G_1, G_2)$ admit an equilibrium $(F_1, F_2, m_1, m_2, m_3)$, then $(G_1, G_2)$
\text{have the block structure in (2.14)–(2.15).}

First consider any securities $i$ and $j$ cleared only by CCPs 1 and 2, respectively. Write $c_1(i, j)$
for the $(i, j)$ entry of $F_1$, and write $\bar{c}_1(i, j)$, $\bar{c}_1(i, i)$ for the corresponding entries of $G_1$. Consider a
portfolio holding $u$ units of $i$ and $w$ units of $j$. Condition (2.13) requires
\]
u^2 c_1(i, i) = u(\bar{c}_1(i, i)u + \bar{c}_1(i, j)w)
\]
for all \( u \) and \( w \). The case \( w = 0 \) implies that \( c_1(i, i) = \tilde{c}_1(i, i) \), and then any \( w \neq 0 \) implies \( \tilde{c}_1(i, j) = 0 \). Thus, the block \( G_1(1, 3) = G_1(3, 1) \) is identically zero. By the same argument, \( G_2(1, 3) = G_2(3, 1) = 0 \).

Now suppose security \( j \) is cleared by both CCPs and consider a portfolio holding \( u \) units of \( i \) and \( v \neq 0 \) units of \( j \), with \( v_1 \) units cleared through CCP 1 and \( v - v_1 \) units cleared through CCP 2. To solve (2.12), the clearing member chooses \( v_1 \) to minimize

\[
 u^2 c_1(i, i) + 2uv_1 c_1(i, j) + c_1(j, j) v_1^2 + c_2(j, j) (v - v_1)^2, 
\]

which yields

\[
 v_1 = \frac{c_2(j, j) v - c_1(i, j) u}{c_1(j, j) + c_2(j, j)}. \tag{A.19} 
\]

To satisfy (2.13), we need to have

\[
 u^2 c_1(i, i) + 2uv_1 c_1(i, j) + c_1(j, j) v_1^2 = u^2 \tilde{c}_1(i, i) + u(v_1 + v) \tilde{c}_1(i, j) + vv_1 \tilde{c}_1(j, j).
\]

We have already established that \( c_1(i, i) = \tilde{c}_1(i, i) \), so this entails

\[
 c_1(j, j) \frac{v_1^2}{v^2} - \tilde{c}_1(j, j) \frac{v_1}{v} = \tilde{c}_1(i, j) u \left[ \frac{1}{v} + \frac{v_1}{v^2} \right] - 2 \tilde{c}_1(i, j) u \frac{v_1}{v^2}. \tag{A.20} 
\]

If neither \( c_1(i, j) \) nor \( \tilde{c}_1(i, j) \) is zero, then \( v_1 = 0 \) in (A.19) at some \( u \neq 0 \) but not in (A.20). So, suppose \( c_1(i, j) = 0 \). Then \( v_1/v \) in (A.19) is a constant, independent of \( u \). But for the same to hold in (A.20) we must have \( \tilde{c}_1(i, j) = 0 \). We conclude that \( G_1(1, 2) = 0 \), and the same argument shows \( G_1(3, 2) = 0 \).

\[\square\]

**Proposition 3 (Stable Partitioned Equilibrium).** A partitioned equilibrium is stable if

\[
 G_1(1, 1) \preceq G_2(1, 1), \quad G_1(3, 3) \succeq G_2(3, 3), \tag{2.18} 
\]

in the positive definite order.
Proof. First write \( \{G_1, G_2\} \) in the same block diagonal structure with \( k \) as large as possible, such that

\[
G_i = \begin{pmatrix}
G_i(1,1) & & \\
& G_i(2,2) & \\
& & \ddots \\
& & & G_i(k,k)
\end{pmatrix}
\]

where \( G_i(j,j) \in \mathbb{R}^{m_j \times m_j} \), \( \sum_{j=1}^{k} m_j = m \), and for \( \mathcal{B} \cup \mathcal{F}_1 \cup \mathcal{F}_2 = \{1, 2, \ldots, k\} \) the following hold:

1. for \( j \in \mathcal{B} \), \( G_1(j,j) = G_2(j,j) \)

2. for \( j \in \mathcal{F}_1 \), \( G_2(j,j) \succeq G_1(j,j) \) and \( G_1(j,j) \neq G_2(j,j) \)

3. for \( j \in \mathcal{F}_2 \), \( G_1(j,j) \succeq G_2(j,j) \) and \( G_1(j,j) \neq G_2(j,j) \)

This means that the two CCPs disagree for security classes in \( \mathcal{F} = \mathcal{F}_2 \cup \mathcal{F}_2 \) and agree on security classes in \( \mathcal{B} \). There are no cross impacts between securities in different security classes.

Let \( E_1 \) denote an equilibrium in Definition 4. From Theorem 2, we know that in any partitioned equilibrium, CCPs can only jointly clear security classes for which they have the same market beliefs.

For equilibrium \( E_1 \), we assume that CCP 1 clears security classes in \( \mathcal{S}_1 \), and CCP 2 clears security classes in \( \mathcal{S}_2 \). Then we have \( \mathcal{F} \cap \mathcal{S}_1 = \mathcal{F}_1 \), \( \mathcal{F} \cap \mathcal{S}_2 = \mathcal{F}_2 \) and \( \mathcal{S}_1 \cap \mathcal{S}_2 \subseteq \mathcal{B} \).

For a partitioned equilibrium \( E_2 \) other than \( E_1 \), we assume that CCP 1 clears security classes in \( \mathcal{S}_1 \), and CCP 2 clears security classes in \( \mathcal{S}_2 \). We have:

\[
\mathcal{S}_1 \cap \mathcal{S}_2 \subseteq \mathcal{B}, \quad \mathcal{S}_1 \cup \mathcal{S}_2 = \{1, 2, \ldots, k\}
\]

For any position \( x^\top = (x^\top(1), \ldots, x^\top(k)) \), with \( x(j) \in \mathbb{R}^{m_j} \), by definition the total margin collected in equilibrium \( E_1 \) is

\[
x_1^\top F_1 x_1 + x_2^\top F_2 x_2 = \sum_{j \in \mathcal{S}_1 \cap \mathcal{F}} x^\top(j) G_1(j,j) x(j) + \sum_{j \in \mathcal{S}_2 \cap \mathcal{F}} x^\top(j) G_2(j,j) x(j) + \sum_{j \in \mathcal{B}} x^\top(j) G_1(j,j) x(j)
\]

(A.21)
APPENDIX A. APPENDIX

The total margin collected in equilibrium $E_2 (\tilde{x}_1, \tilde{x}_2, \tilde{F}_1, \tilde{F}_2)$ is

$$\tilde{x}_1^\top \tilde{F}_1 \tilde{x}_1 + \tilde{x}_2^\top \tilde{F}_2 \tilde{x}_2 = \sum_{j \in \tilde{S}_1 \cap F} x^\top (j) G_1(j,j)x(j) + \sum_{j \in \tilde{S}_2 \cap F} x^\top (j) G_2(j,j)x(j) + \sum_{j \in B} x^\top (j) G_1(j,j)x(j) \tag{A.22}$$

Taking the difference between (A.21) and (A.22), we get

$$x_1^\top F_1 x_1 + x_2^\top F_2 x_2 - \tilde{x}_1^\top \tilde{F}_1 \tilde{x}_1 - \tilde{x}_2^\top \tilde{F}_2 \tilde{x}_2$$

$$= \sum_{j \in \tilde{S}_2 \cap F} \left( x^\top (j) G_1(j,j)x(j) - x^\top (j) G_2(j,j)x(j) \right)$$

$$+ \sum_{j \in \tilde{S}_1 \cap F} \left( x^\top (j) G_2(j,j)x(j) - x^\top (j) G_1(j,j)x(j) \right) \tag{A.23}$$

$$\leq 0$$

Thus, equilibrium $E_1$ is stable. ■

A.2. Additional Proofs for Chapter 3

A.2.1. Proofs for Section 3.2

Theorem 4 (Existence and Convexity). The dynamic control problem defined in (3.8) is bounded and an optimal solution $u^*$ always exists. In addition, the optimal value (the liquidity cost) is convex in initial position.

Proof. Given $q$ and $u$, $x$ is then given by

$$x(t) = q - \int_0^t Y u(s) \, ds. \tag{A.24}$$

By substituting (A.24) in (3.8), we obtain the reduced optimal control problem

$$J_{\text{red}}(u) \triangleq \min_u \quad \int_0^\infty f(u(t)) \, dt + \mu \int_0^\infty (q - \int_0^t Y u(s) \, ds)^\top \Sigma (q - \int_0^t Y u(s) \, ds) \, dt \tag{A.25}$$

subject to $|u_j(t)| \leq \gamma_j$, $\forall j$. 124
APPENDIX A. APPENDIX

It is easy to see that the functional $J_{\text{red}}$ is convex in $u$. Now define

$$J^* \triangleq \inf_u J_{\text{red}}(u).$$

Notice that $J^*$ is well defined since $J_{\text{red}}(u)$ is lower bounded by 0. Then there exists a sequence of feasible controls $S = \{u^{(i)} | i = 1, 2, \ldots\}$ such that

$$J_{\text{red}}(u^{(i)}) \to J^*.$$  \hfill (A.26)

Without loss of generality, we assume that $J_{\text{red}}(u^{(i)}) < \infty$. Combined with the fact that $u^{(i)}$ is bounded, this suggests that $x^{(i)}$ must be bounded.

Next we will prove that $S \subset L^1([0, \infty); \mathbb{R}^m)$ is equi-integrable. Since $u_j(t)$ are bounded, $\|u(t)\|_1$ must be bounded, and we can simply take constant $C$ such that

$$\|u(t)\|_1 \leq C, \quad \forall u \in L^1([0, \infty); \mathbb{R}^m), t \geq 0.$$

Now consider any measurable set $A \subset [0, \infty)$ such that

$$\int_A \|u(s)\|_1 \, ds \leq C \mu(A), \forall u \in S.$$

The equi-integrability then follows trivially. Next, by the Dunford–Pettis theorem (Chapter II, Theorem T25 in Dellacherie and Meyer (2011)), $S \subset L^1([0, \infty)$ is relatively compact for the weak topology. Then there exists a weakly convergent subsequence $\{\hat{u}^{(i)}\}$ of $\{u^{(i)}\}$ that converges to some $u^* \in S$ such that

$$\hat{u}^{(i)} \overset{w}{\to} u^*, \quad \hat{u}^{(i)} \in S.$$  \hfill (A.27)

Due to its convexity, the reduced functional $J_{\text{red}}$ is lower semicontinuous with respect to the weak topology and hence

$$\liminf_{i \to \infty} J_{\text{red}}(\hat{u}^{(i)}) \leq J_{\text{red}}(u^*),$$
which allows us to conclude that $u^*$ is a minimizer.

\[\square\]

**Theorem 5 (Uniqueness).** All optimal solutions for the the dynamic control problem in (3.8) have a unique optimal position trajectory $x^* \in C([0, \infty); \mathbb{R}^n)$. Moreover, if the transaction cost functional $f(\cdot \dot{\ })$ is strictly convex, the optimal trading strategy $u^* \in L^1([0, \infty; \mathbb{R}^m)$ must also be unique.

**Proof.** Denote $J(u, x)$ to be the liquidity cost associated with feasible trading rate $u$ and position process $x$.

Suppose the optimal solution is not path-unique; then there exist $(u_1, x_1), (u_2, x_2)$ that are both optimal and such that

$$J(u_1, x_1) = J(u_2, x_2) = J^*, \quad x_1 \neq x_2.$$ 

Now consider

$$u_3 = \frac{u_1 + u_2}{2}, \quad x_3 = \frac{x_1 + x_2}{2}.$$ 

It is easy to see that $(u_3, x_3)$ is also feasible for (3.8).

Then we have

$$J(u_3, x_3) = \int_0^\infty f(u_3(t))dt + \mu \int_0^\infty x_3(t)^\top \Sigma x_3(t)dt$$

$$= \int_0^\infty f(\frac{u_1(t) + u_2(t)}{2})dt + \mu \int_0^\infty (\frac{x_1(t) + x_2(t)}{2})^\top \Sigma (\frac{x_1(t) + x_2(t)}{2})dt$$

$$< \int_0^\infty (f(u_1(t)) + f(u_2(t))) dt/2 + \mu \int_0^\infty (\frac{x_1(t)^\top \Sigma x_1(t) + x_2(t)^\top \Sigma x_2(t)}{2}) dt/2$$

$$= J(u_1, x_1)/2 + J(u_2, x_2)/2 = J^*.$$ 

The strict inequality is provided by the strict convexity of $x^\top \Sigma x$. Notice that this contradicts the optimality of $(u_1, x_1), (u_2, x_2)$.

If $f(\cdot)$ is strictly convex, we will also require $u_1 = u_2$ such that

$$\int_0^\infty f(\frac{u_1(t) + u_2(t)}{2})dt = \int_0^\infty (f(u_1(t)) + f(u_2(t))) dt/2.$$ 

This provides the uniqueness of the optimal solution.
Theorem 6 ( Sufficiency). The pair \((x^*, u^*)\) ∈ \(C([0, \infty); \mathbb{R}^n) \times L^1([0, \infty); \mathbb{R}^m)\) form an optimal solution of (3.8) if, for all \(t ≥ 0\),
\[
x^*(t) = q - \int_0^t Y u^*(s) \, ds,
\]
\[
u^*(t) ∈ \arg\min_{u: \ -\gamma ≤ u ≤ \gamma} f(u) - 2 \int_0^\infty x^*(s) \Sigma Y u \, ds.
\]

Proof. Define \(J_{\text{red}}(u)\) to be the optimal liquidity cost associated with trading strategy \(u\). Now suppose \(u^*\) satisfies (3.9) but is not optimal. Then there must exist some feasible trading strategy \(\tilde{u} \neq u^*\) such that
\[
J_{\text{red}}(\tilde{u}) < J_{\text{red}}(u^*).
\]
By algebra, we also have
\[
J_{\text{red}}(\tilde{u}) = J_{\text{red}}(u^*) + \int_0^\infty \int_0^t (\tilde{u}(s) - u^*(s)) Y^\top \Sigma Y (\tilde{u}(s) - u^*(s)) \, ds \, dt
\]
\[
- 2 \int_0^\infty \int_0^t (x^*(t)) Y^\top \Sigma Y (\tilde{u}(s) - u^*(s)) \, ds \, dt + \int_0^\infty (f(\tilde{u}(t)) - f(u^*(t))) \, dt
\]
\[
≥ J_{\text{red}}(u^*) + \int_0^\infty \left( f(\tilde{u}(t)) - f(u^*(t)) - 2 \int_s^\infty (x^*(s)) Y^\top \Sigma Y (\tilde{u}(s) - u^*(s)) \, ds \right) \, dt.
\]

Since \(u^*\) satisfies (3.9), it is easy to see that the second term is always positive. Hence we reach a contradiction.

A.2.2. Proofs for Section 3.3

Theorem 7 (Scaling). If \(u^*\) is optimal for the problem starting from \(q\), then \(\tilde{u}(t) = u^*(t/\alpha) \forall t > 0\) is optimal for the problem starting with \(\alpha q\) with \(\forall \alpha ∈ \mathbb{R}^+\), where
\[
J^*(\alpha q) = \alpha^3 J^*(q).
\]

Proof. Under trading strategy \(\tilde{u}\),
\[
\tilde{x}(t) = \alpha q - \int_0^t Y \tilde{u}(s) \, ds = \alpha q - \alpha \int_0^{t/\alpha} Y u(s) \, ds = \alpha x(t/\alpha).
\]
APPENDIX A. APPENDIX

Suppose that $\tilde{x}$ is not the optimal path for the problem starting with $\alpha q$; then there exists a better path $y$ such that
\[
\int_0^\infty (y(t))\top \Sigma y(t)dt < \int_0^\infty (\tilde{x}(t))\top \Sigma \tilde{x}(t)dt.
\]
Then we have
\[
\int_0^\infty (y(\alpha t))\top \Sigma y(\alpha t)dt < \int_0^\infty (x(t))\top \Sigma x(t)dt.
\]
It is easy to see that $y_\alpha(t) = y(\alpha t)$ is feasible for the problem starting with $q$, a contradiction:
\[
J^*(\alpha q) = \int_0^\infty (\tilde{x}(t))\top \Sigma \tilde{x}(t)dt = \alpha^3 \int_0^\infty (x(t))\top \Sigma x(t)dt = \alpha^3 J^*(q).
\]

Theorem 8 (Finite Horizon). For any initial position $q$, the optimal position trajectory $x(t)$ is guaranteed to reach zero in finite time.

Proof. First of all, define the norm $||\cdot||_\Sigma$ as
\[
||q||_\Sigma \triangleq \sqrt{q\top \Sigma q}, \quad \forall q \in \mathbb{R}^n.
\]
By Theorem 7, we know that it suffices to prove the theorem for $\forall q \in \mathbb{R}^n$ with $||q||_\Sigma = 1$.

Now, for any $q$, let $x^*(t)$ be the position associated with its optimal execution strategy at time $t$. Clearly, we have $x^*(0) = q$. Now define $T(q)$ to be the first time that the norm of the position is less than $1/2$:
\[
T(q) \triangleq \inf\{t : ||x^*(t)||_\Sigma \leq \frac{1}{2} ||q||_\Sigma\}.
\]
Given that $Y$ is full rank, we know that the set $\mathcal{A} = \{Yu|u \in \mathbb{R}^m, |u_i| \leq \gamma_i\}$ is an $n$-dimensional polytope in $\mathbb{R}^n$. Notice that $0 \in \mathcal{A}$. Therefore, there must exist $\epsilon > 0$ such that $\{q \in \mathbb{R}^n||q||_\Sigma \leq \epsilon\} \subset \mathcal{A}$. 128
APPENDIX A. APPENDIX

Now consider any $q$ such that $||q||_{\Sigma} = 1$, and let $x^*(t)$ be its position at time $t$ in the optimal execution strategy. Define

$$\tau_i = \inf\{t \geq 0 : ||x^*(t)||_{\Sigma} \leq \frac{1}{2^i}\}.$$

Lemma 3 shows that $\tau_1 - \tau_0 \leq T^*$. By applying Theorem 7 we know that

$$T(q) \leq T^*/2, \quad \forall ||q|| \leq 1/2.$$

It follows that

$$\tau_{i+1} - \tau_i \leq \frac{T^*}{2^i}.$$

Then

$$\lim_{i \to \infty} \tau_i = \lim_{i \to \infty} \left[ \tau_0 + \sum_{j=1}^{i} (\tau_j - \tau_{j-1}) \right] \leq 0 + \lim_{i \to \infty} \sum_{j=1}^{i} \frac{1}{2^{j-1}} T^* = 2T^*.$$

Then, $\{\tau_i\}$ is increasing and bounded from above and so there exists a limit $\tau^*$ such that

$$\tau^* = \lim_{i \to \infty} \tau_i \leq 2T^*.$$

By continuity of $x^*(t)$, we have $x^*(\tau^*) = 0$. Then, it must be that

$$x^*(t) = 0, \quad t \geq \tau^*.$$

otherwise the liquidity cost could be reduced.

Lemma 3.

$$T^* \triangleq \sup\{T(q) : ||q||_{\Sigma} = 1\} < \infty.$$
APPENDIX A. APPENDIX

Proof. \( \forall q \) such that \( ||q||_\Sigma = 1 \), consider a trading strategy where we set

\[
u(t) = \begin{cases} u^*, & t \leq \frac{1}{\epsilon} \\ 0, & t > \frac{1}{\epsilon} \end{cases},
\]

where \( Y u^* = \epsilon q \). Notice that this strategy is clearly feasible.

The liquidity cost of this strategy \( J(q) \) is easily given by

\[
J(q) = \int_0^{\frac{1}{\epsilon}} ||q - t\epsilon q||_\Sigma dt < \frac{1}{\epsilon}.
\]

Now, in the optimal trading strategy, we have

\[
\frac{1}{2} T(q) \leq \int_0^{T(q)} ||x^*(t)||_\Sigma dt < \int_0^{\infty} ||x^*(t)||_\Sigma dt \leq J(q) < \frac{1}{\epsilon},
\]

which leads to

\[
T(q) < \frac{2}{\epsilon}.
\]

Since the upper bound does not depend on \( q \), we have

\[
T^* < \frac{2}{\epsilon} < \infty.
\]

Lemma 1 (Optimality). A feasible control \( u^* \) is optimal for (3.11) if and only if

\[
\forall t \geq 0, \quad u^*(t) \in \text{argmax}_{u: -\gamma \leq u \leq \gamma} \left( \int_t^\infty (x^*(s))^\top \Sigma Y ds \right) u,
\]

where \( x^* \) is uniquely determined by \( u^* \) and \( q \) through the control function.

Proof. The sufficiency is given by Theorem \( \boxed{6} \). By Theorem \( \boxed{8} \), given the initial position \( q \), there exists some \( T \) such that the optimal execution ends before time \( T \). As a result, (3.11) is equivalent to the following:
APPENDIX A. APPENDIX

\[ J^*(q) \triangleq \min_u \int_0^T x^\top(t) \Sigma x(t) \, dt \]

subject to
\[
\dot{x}(t) = -Y u(t), \quad \forall \, t \geq 0,
\]
\[
|u_i(t)| \leq \gamma_i, \quad \forall \, 1 \leq i \leq m, \, t \geq 0,
\]
\[
x(0) = q, \quad u \in L^1([0, \infty); \mathbb{R}^m). \quad (A.30)
\]

We can obtain the necessity through Pontrjagin’s minimum principle. The convexity assumption and regularity assumption are satisfied trivially in this case due to the linear control. The Hamiltonian function of (A.30) can be written as
\[
H(x, u, p) = x^\top \Sigma x - p^\top Y u. \quad (A.31)
\]

Suppose that \( x^*(t), u^*(t) \) is optimal; then there must exist an optimal adjoint state \( p^*(t), t \in [0, \infty) \) that satisfies the following:
\[
\dot{p}^*(t) = -\nabla_x H(x^*(t), u^*(t), p^*(t)) = -2\Sigma x^*(t),
\]
\[
p^*(T) = 0.
\]

We can then solve for \( p^*(t) \) as
\[
p^*(t) = 2\Sigma Y \int_t^T x^*(s) \, ds. \quad (A.32)
\]

Moreover, we have
\[
u^*(t) \in \arg\min_{u: \, -\gamma \leq u \leq \gamma} \quad H(x^*(t), u(t), p^*(t)), \quad (A.33)
\]

which can also be written as
\[
u^*(t) \in \arg\min_{u: \, -\gamma \leq u \leq \gamma} \left( -2 \int_t^T (x^*(s))^\top \Sigma Y u \, ds \right), \quad \forall \, t \in [0, T].
\]
Since we know that \( x^*(t) = 0 \) for \( \forall t \geq T \), the condition above can be written as

\[
u^*(t) \in \arg\max_{u: -\gamma \leq u \leq \gamma} \left( \int_t^\infty (x^*(s))^\top \Sigma Y u ds \right), \forall t \in [0, \infty).
\]

\[\blacksquare\]

**Theorem 9 (High Liquidity Hedging).** In the two-dimensional case where model parameters are given by

\[
\Sigma = \begin{bmatrix}
\sigma_1^2 & \rho \sigma_1 \sigma_2 \\
\rho \sigma_1 \sigma_2 & \sigma_2^2
\end{bmatrix}, \quad Y = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},
\]

if we further assume that

\[
\gamma_2 \geq |\rho| \frac{\sigma_1 \gamma_1}{\sigma_2},
\]

then the optimal liquidity cost of portfolio \( q = (q, 0) \) is given by

\[
J^*(q) = \frac{1}{3} \frac{q^3}{\gamma_1^2} \left( 1 - \frac{\rho^2}{1 + |\rho| \frac{\sigma_1 \gamma_1}{\sigma_2 \gamma_2}} \right).
\]

**Proof.** Without loss of generality, we assume that \( \rho > 0 \). In order to simplify the notations and to provide better intuitions, we define the following:

\[
\begin{bmatrix} a & b \\ b & c \end{bmatrix} = Y^\top \Sigma Y,
\]

where \( a, b, c \) can be easily determined by the model parameters:

\[
a = \sigma_1^2, \quad b = \rho \sigma_1 \sigma_2, \quad c = \sigma_2^2.
\]

Now consider the following trading strategy:

1. For \( 0 \leq t \leq \frac{b}{b+c} \frac{a}{\gamma_1} \), trade at rate \( u^*(t) = (\gamma_1, -\gamma_2)^\top \).

2. For \( \frac{b}{b+c} \frac{a}{\gamma_1} \leq t \leq \frac{a}{\gamma_1} \), trade at rate \( u^*(t) = (\gamma_1, \frac{b}{c} \gamma_1)^\top \).
APPENDIX A. APPENDIX

The assumption that $\gamma_2 \geq |\rho| \frac{\gamma_1}{\sigma_2}$ guarantees the feasibility of this strategy.

The liquidation path can now be calculated as

$$x^*(t) = \begin{cases} (q - \gamma_1 t, \gamma_2 t) \top, & 0 \leq t \leq \frac{b}{b+c} \frac{q}{\gamma_1} \\ (q - \gamma_1 t, -\frac{b}{c} (q - \gamma_1 t) \top), & \frac{b}{b+c} \frac{q}{\gamma_1} \leq t \leq \frac{q}{\gamma_1} \\ (0, 0) \top, & t > \frac{q}{\gamma_1} \end{cases}.$$  \hspace{1cm} (A.34)

Then it's easy to justify that (A.34) satisfies the optimality condition in Lemma 1.

Hence the optimal solution is given by

$$J^*(q) = \int_0^{\infty} (x^*(t)) \top \Sigma x^*(t) dt = \frac{1}{3} \frac{q^3}{\gamma_1} \left( 1 - \frac{\rho^2}{1 + \rho^2 \frac{\sigma_1 \gamma_1}{\sigma_2 \gamma_2}} \right).$$

\hspace{1cm} ■

Theorem 10 (One Asset). In the one-dimensional case, the cost of liquidating a position of $q$ with parameters $(\sigma, \gamma, \nu)$ is given by

$$J^*(q) = \nu |q| + \mu \frac{|q|^3 \sigma^2}{3 \gamma}.$$

Proof. In the one-dimensional case, it is easy to see that the transaction cost is only a function of the total position traded and does not depend on the trading rate. As a result, the optimal trading strategy is simply to unload the position as fast as possible.

With this in mind, the agent should trade the asset at a constant rate $\gamma$ if $q > 0$ and $-\gamma$ if $q \leq 0$. Then the liquidity cost is given by:

$$J^*(q) = \nu |q| + \mu \frac{|q|^3 \sigma^2}{3 \gamma}.$$ 

\hspace{1cm} ■
Theorem 11 (Two Assets). In the two-dimensional case where model parameters are given by

\[
\Sigma = \begin{bmatrix}
\sigma_1^2 & \rho \sigma_1 \sigma_2 \\
\rho \sigma_1 \sigma_1 & \sigma_2^2 \\
\end{bmatrix}, \quad Y = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad q = (q, 0)^\top,
\]

if we further assume that

\[
\gamma_2 \geq |\rho| \frac{\sigma_1 \gamma_1}{\sigma_2},
\]

then the asset 2 will only be used to hedge if and only if

\[
q^2 \geq \frac{2\gamma_1 \nu_2}{\mu \gamma_2 \rho \sigma_1 \sigma_2}. \tag{3.17}
\]

If (3.17) is satisfied, then the optimal liquidity cost of the portfolio is given by

\[
J^*(q) = \frac{1}{3} \frac{q^3}{\gamma_1} \sigma_1^2 \left( 1 - \frac{\rho^2}{1 + |\rho| \frac{\sigma_1 \gamma_1}{\sigma_2 \gamma_2}} \right) + \nu_1 q + 2 \nu_2 q \frac{\gamma_2}{\gamma_1} \left( 1 + |\rho| \frac{\sigma_1 \gamma_1}{\sigma_2 \gamma_2} \right) - 4 \frac{\nu_2}{\sigma_2} \sqrt{\frac{2\nu_2 \rho \sigma_1 \gamma_1}{\sigma_2}} \frac{\gamma_2}{\gamma_1} \left( 1 + |\rho| \frac{\sigma_1 \gamma_1}{\sigma_2 \gamma_2} \right). \tag{3.18}
\]

Proof. Without loss of generality, we assume that \( \rho > 0, q > 0 \), in which case the two assets are positively correlated. Notice that Theorem 9 can be viewed as a special case where the transaction cost parameters \( \nu_1, \nu_2 \) are zero. Accordingly, we first establish a short position in asset 2 in order to hedge the market risk. Since there is no transaction cost, asset 2 is traded at full rate \( \gamma_2 \) until the ratio of position in the two assets reaches \( \rho \frac{\sigma_1 \gamma_1}{\sigma_2 \gamma_2} \), and this ratio is maintained till the end of the liquidation process. Intuitively, by shorting asset 2 we are hedging the market risk but introducing another source of idiosyncratic risk, and the ratio \( \rho \frac{\sigma_1 \gamma_1}{\sigma_2 \gamma_2} \) is the balance point of such a trade-off. Now we have transaction costs for trading asset 2, and it is expected that the agent will trade asset 2 less.

Now consider the following trading strategy:

- If \( q^2 < \frac{2\gamma_1 \nu_2}{\mu \gamma_2 \rho \sigma_1 \sigma_2} \):
  1. Sell asset 1 as fast as possible until the entire position is unloaded.

- If \( q^2 \geq \frac{2\gamma_1 \nu_2}{\mu \gamma_2 \rho \sigma_1 \sigma_2} \):


APPENDIX A. APPENDIX

1. For $0 \leq t \leq T_1$, trade at rate $u^*(t) = (\gamma_1, -\gamma_2)^\top$.

2. For $T_1 < t \leq T_2$, trade at rate $u^*(t) = (\gamma_1, 0)^\top$.

3. For $T_2 < t \leq \frac{2}{\gamma_1}$, trade at rate $u^*(t) = (\gamma_1, \frac{\rho_2}{\sigma_2^2}\gamma_1)^\top$,

where

$$T_1 = \frac{q\rho\gamma_1\sigma_2 - \sqrt{2\nu_2\rho\gamma_1\sigma_2\gamma_2}}{\rho\gamma_1\sigma_2\gamma_2 + \sigma_2^2\gamma_2},$$

$$T_2 = \frac{q\gamma_1\rho\sigma_1 + \sqrt{2\nu_2\gamma_1\gamma_2\sigma_2}}{\gamma_1^2\rho\sigma_1 + \gamma_1\gamma_2\sigma_2}.$$

The cost induced by this strategy is given by (3.18). This strategy can be shown to be optimal by checking (3.9) in Theorem 6. We omit the details here as the algebra is cumbersome.

A.2.3. Proofs for Section 3.4

**Proposition 5** (Factor Replicating Portfolio). If the large-universe conditions hold, then for each factor $F_i(t)$, there exists a series of portfolios $\{p^{(i,n)}(t)\}$ defined by weights $\{\beta^{(i,n)}_j(t)\}$ where

$$p^{(i,n)}(t) \triangleq \sum_{j=1}^{n} \beta^{(i,n)}_j S_j(t),$$

such that

1. The portfolio $p^{(i,n)}(t)$ has unit exposure on factor $F_i(t)$:

$$p^{(i,n)}(t) - F_i(t) = \epsilon^{(i,n)}(t),$$

where $\epsilon^{(i,n)}(t)$ is zero mean and independent of all factor-price processes, and has variance upper bounded by

$$\text{Var}(\epsilon^{(i,n)}(t)) \leq \frac{\sup_j S_j^2}{\lambda_{\min}} t.$$
APPENDIX A. APPENDIX

2. The sum of the squares of the weights converge to 0:

\[
\lim_{n \to \infty} \sum_{j=1}^{n} (\beta_{ij}^{(i,n)})^2 = 0.
\]

Proof. For each factor \(i\), we want to find the portfolio that has unit exposure on factor \(i\) and has minimum idiosyncratic risks. This can be done by solving the following optimization problem:

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} w^\top \Xi^{(n)} w \\
\text{subject to} & \quad (L^{(n)})^\top w = e_i, \\
& \quad w \in \mathbb{R}^n.
\end{align*}
\]

(A.35)

where \(e_i\) is the \(i\)th column of the \(K \times K\) identity matrix.

For simplicity, we will assume that the idiosyncratic risk for each individual asset is strictly positive \((\varsigma_j^2 > 0)\), and will simply ignore the assets with \(\varsigma_j^2 = 0\).

Now denote \(z = (\Xi^{(n)})^{1/2} w\), and consider the singular value decomposition of \(L^{(n)}\):

\[
L^{(n)} = (U^{(n)})^\top \Lambda^{(n)} V^{(n)}.
\]

(A.36)

Then (A.35) is equivalent to

\[
\begin{align*}
\min_z & \quad \frac{1}{2} z^\top z \\
\text{s.t.} & \quad (V^{(n)})^\top \Lambda^{(n)} U^{(n)} (\Xi^{(n)})^{-1/2} z = e_i, \\
& \quad z \in \mathbb{R}^n.
\end{align*}
\]

(A.37)

The lagrangian of (A.37) is

\[
L = \frac{1}{2} z^\top z - \mu^\top \left( (L^{(n)})^\top (\Xi^{(n)})^{-1/2} w - e_i \right).
\]
The optimal solution $z^*$ is given by solving
\[
\begin{align*}
z^* - (\Xi^{(n)})^{-1/2}L^{(n)} \mu &= 0, \\
L^{(n)}(\Xi^{(n)})^{-1}L^{(n)} \mu &= e_i.
\end{align*}
\tag{A.38}
\]

Then we have
\[
\begin{align*}
(w^*)^\top \Xi^{(n)} w^* &= (z^*)^\top z^* = \mu^\top (\Xi^{(n)})^{-1}L^{(n)} \mu \\
&= e_i^\top \left( (L^{(n)}(\Xi^{(n)})^{-1}L^{(n)})^{-1} \right) e_i \\
&\leq \lambda_{\text{max}} \left( (L^{(n)}(\Xi^{(n)})^{-1}L^{(n)})^{-1} \right) \\
&\leq \sup_{j \leq n} \varsigma_j^2 \lambda_{\text{min}}^{(n)}.
\end{align*}
\tag{A.39}
\]

We can now set $\beta^{i,n}$ as $w^*$ solved from above. As $n$ goes to infinity, we have
\[
\lim_{n \to \infty} \sum_{j=1}^{n} \beta^{i,n}_j \leq \lim_{n \to \infty} \frac{(w^*)^\top \Xi^{(n)} w^*}{\sup_{j \leq n} \varsigma_j^2} \leq \lim_{n \to \infty} \frac{1}{\lambda_{\text{min}}^{(n)}} = 0.
\]

\[\blacksquare\]

**Lemma 4.** Consider two liquidation problems that differ only in their covariance matrices ($\Sigma_1$ and $\Sigma_2$, respectively). Suppose that their optimal liquidity costs are given by $J_1^{*}(q)$ and $J_2^{*}(q)$. If we have
\[
\Sigma_1 \preceq \Sigma_2,
\]
where $\preceq$ is the positive definite ordering, then
\[
J_1^{*}(q) \leq J_2^{*}(q), \quad \forall q \in \mathbb{R}^n.
\]

**Proof.** Suppose that $u^{(2)} \in L_1([0, \infty); \mathbb{R}^m)$ is the optimal solution to problem 2, and $x^{(2)} \in C([0, \infty); \mathbb{R}^n)$ is the corresponding position process. Since problems 1 and 2 differ only in their covariance matrices, $(u^{(2)}, x^{(2)})$ is also feasible for problem 1. If we denote $J_1^{*}(q)$ as the optimal
liquidity cost for problem 1 and $J^*_2(q)$ as that for problem 2, then we have

$$J^*_2(q) = \int_0^\infty f(u^{(2)}(t))dt + \mu \int_0^\infty x^{(2)}(t)\Sigma_2 x^{(2)}(t)dt$$

$$\geq \int_0^\infty f(u^{(2)}(t))dt + \mu \int_0^\infty x^{(2)}(t)\Sigma_1 x^{(2)}(t)dt$$

$$\geq J^*_1(q),$$

(A.40)

where the first inequality comes from the fact that $\Sigma_1 \preceq \Sigma_2$.  

\textbf{Theorem 12} (Lower Bound of Hedging Benefits). If we are allowed to trade other assets during the liquidation process, the liquidity cost is lower bounded by

$$J^*_n(q) \geq \sum_{j=1}^m \frac{\xi_j^2}{3\gamma_j}|q_j|^3.$$  

(3.22)

\textbf{Proof.} Consider the following problem where we replace the covariance matrix $\Sigma^{(n)}$ with $\Xi^{(n)}$:

$$\tilde{J}_n^*(q) = \min_{u} \int_0^\infty x^\top(t)\Xi^{(n)}x(t)dt = \sum_{i=1}^n \int_0^\infty \xi_i^2x_i^2(t),dt$$

subject to

$$\dot{x}(t) = -Yu(t), \quad \forall \ t \geq 0,$$

$$|u_i(t)| \leq \gamma_i, \quad \forall \ 1 \leq i \leq m, \ t \geq 0,$$

$$x(0) = q,$$

$$u \in L^1([0, \infty); \mathbb{R}^m).$$

(A.41)

Since there are no correlations, it is easy to see that the optimal execution strategy in this case is to liquidate each asset separately at full rate. Hence the optimal solution to the above problem is given by

$$\tilde{J}_n^*(q) = \sum_{j=1}^m \frac{\xi_j^2}{3\gamma_j}|q_j|^3.$$  

Notice that $\Xi^{(n)} \preceq \Sigma^{(n)}$; then, by applying Lemma 4 stated above, we always have

$$J_n^*(q) \geq \tilde{J}_n^*(q) = \sum_{j=1}^m \frac{\xi_j^2}{3\gamma_j}|q_j|^3.$$
Theorem 13 (Large Universe). If the large-universe property is satisfied, then, asymptotically, the liquidity cost of any portfolio consisting of finitely many assets will be driven purely by idiosyncratic risks. More specifically, we have

\[
J^*_\infty(q) = \lim_{n \to \infty} J^*_n(q) = \sum_{j=1}^{m} \frac{2}{3\gamma_j} |q_j|^3, \tag{3.23}
\]

where \(q\) is defined in (3.21), and \(J^*_n(q)\) represents the optimal costs of liquidating \(q \in \mathbb{R}^A\) in \(A_n\).

Proof. The key to the proof is finding a trading strategy that converges to the lower bound asymptotically. We assume that the chosen factor portfolios don’t contain assets in the liquidation portfolio. By Proposition 5, for each factor \(F_i(t)\) there exists a sequence of portfolios \(\{p(i,n)(t)\}\) characterized by \(\{\beta(i,n)\}\) such that \(p(i,n)(t) \to F_i(t)\). More specifically, we have \(\sum_{j=1}^{m} (\beta(j,n))^2 \to 0\).

Now, for each asset \(j\) in the portfolio to be liquidated, we construct a sequence of portfolios \(\{z(j,n)\}\) characterized by \(\{\beta(j,n)\}\) such that

\[
\beta(j,n) = \sum_{i=1}^{K} l_{ji} \beta(i,n), \tag{A.42}
\]

where \(l_{ji}\) is the factor exposure of asset \(j\) on factor \(i\):

\[
\sum_{k=1}^{n} (\beta_k(j,n))^2 \leq \sum_{k=1}^{n} \left( \sum_{i=1}^{K} l_{ji} \beta_k(i,n) \right)^2 \leq \bar{l}^2 K \sum_{i=1}^{n} \sum_{k=1}^{K} (\beta_k(i,n))^2 \leq \bar{l}^2 K^2 \frac{1}{\lambda_{\min}^{(n)}} \to 0,
\]

where \(\bar{l} = \max_{i \leq m, j \leq K} l_{ij}\).
APPENDIX A. APPENDIX

The exposure of $z^{(j,n)}$ on factor $F_i(t)$ is given by:

$$l_{ji} \sum_{k=1}^{n} \beta_{k}^{(i,n)} l_{ki} = l_{ji}.$$ 

Essentially, we have created a sequence of portfolios that has the same factor exposure as that of asset $j$, but whose idiosyncratic risk converges to 0.

Further, define

$$N(n) = \sqrt{\frac{\lambda_{\text{min}}^{(n)}}{2K^2}}.$$ 

(A.43)

Notice that for each $n$ and $1 \leq j \leq m$, we have

$$|\hat{\beta}_{k}^{(j,n)}| \leq \sqrt{n \sum_{i=1}^{n} (\hat{\beta}_{k}^{(j,n)})^2} \leq \frac{1}{N(n)}, \quad \forall k \leq n.$$ 

Intuitively $\frac{1}{N(n)}$ can be viewed as the upper bound of the weight of each asset in every portfolio we constructed. Given that $\sum_{k=1}^{n} (\hat{\beta}_{k}^{(j,n)})^2 \to 0$, we have $N(n) \to \infty$.

We consider the following “dumb” strategy for the problem indexed by $n$:

1. For $0 \leq t \leq \frac{|q_j|}{\gamma_j}$, buy asset $j$ at a rate of $-\frac{q_j}{|q_j|} \gamma_j$.
2. For $0 \leq t \leq \sum_{j=1}^{m} \frac{|q_j|}{N(n)^2}$, buy portfolio $z^{(j,n)}$ at a rate of $\frac{-q_j}{\sum_{i=1}^{m} |q_i|} N(n) \gamma_j$. Do it for all $1 \leq j \leq m$.
3. For $\sum_{j=1}^{m} \frac{|q_j|}{N(n)^2} \leq t \leq \frac{|q_j|}{\gamma_j}$, buy $z^{(j,n)}$ at a rate of $\frac{q_j}{|q_j|} \gamma_j$. Do it for all $1 \leq j \leq m$.

Let’s first try to understand this “dumb” strategy. First of all, we notice that $z^{(j,n)}$ approximates the factor risk exposure of asset $j$, and so step 2 is the hedging process. Basically, we acquire a certain amount of portfolio $z^{(j,n)}$ in order to hedge the factor risks contributed by asset $j$. Step 3 is the liquidation process: we sell each asset together with its hedging portfolio as soon as possible.

We still need to check whether this strategy violates the liquidity constraints. Consider a particular asset $k$ whose weight in each portfolio is at most $1/N(n)$. In step 2, its trading rate is upper bounded by

$$\frac{1}{N(n)} \sum_{j=1}^{m} \frac{-q_j}{\sum_{i=1}^{m} |q_i|} N(n) \gamma_j \leq \gamma.$$
Hence the liquidity constraint is satisfied in step 1.

In step 3, the trading rate for asset \( j > m \) is upper bounded by

\[
\sum_{i=1}^{m} \gamma_i / N(n) < \gamma
\]
given \( n \) is large enough. As a result, given \( n \) is large enough, all the liquidity constraints are satisfied.

Following this trading strategy, the risk of the position at time \( t \) \((V_n(t))\) is

\[
V_n(t) = \begin{cases} 
\sum_{j=1}^{m} (q_j - \gamma_j t)(l_j - \frac{t}{t_n} l_{z(j,n)})^T \sum_{j=1}^{m} (q_j - \gamma_j t)(l_j - \frac{t}{t_n} L_{z(j,n)}) \\
+ \sum_{j=1}^{m} (q_j - \gamma_j t)^2 \varsigma_j^2 + \sum_{j=1}^{m} (q_j - \gamma_j t_n)^2 \frac{t^2}{t_n^2} \varsigma_{z(j,n)}^2, & 0 \leq t \leq t_n, \\
\sum_{j=1}^{m} \frac{(t - T_j)^+}{T_j} (q_j - \gamma_j t)(l_j - l_{z(j,n)})^T \sum_{j=1}^{m} \frac{(t - T_j)^+}{T_j} (q_j - \gamma_j t)(l_j - l_{z(j,n)}) \\
+ \sum_{j=1}^{m} \left( \frac{(t - T_j)^+}{T_j} (q_j - \gamma_j t) \right)^2 (\varsigma_j^2 + \varsigma_{z(j,n)}^2), & \text{otherwise.}
\end{cases}
\]

where

\[
t_n = \sum_{j=1}^{m} |q_j| / N(n) \gamma, \quad T_j = \frac{|q_j|}{\gamma_j},
\]

and \( l_{z(j,n)} \in \mathbb{R}^K \) is the factor exposure of portfolio \( z^{(j,n)} \), and \( \varsigma_{z(j,n)}^2 \) is its idiosyncratic risk exposure.

The liquidity cost \( J_n(q) \) in this case can be written as

\[
J_n(q) = \int_0^{t_n} V_n(t) dt + \int_{t_n}^{\infty} V_n(t) dt.
\] (A.45)

Notice that by construction, we have the following:

\[
l_{z(j,n)} = l_j, \quad \lim_{n \to \infty} \varsigma_{z(j,n)}^2 = \lim_{n \to \infty} \sum_{i=1}^{n} \varsigma_i^2 (\tilde{\beta}_k^{(j,n)})^2 \leq \sup_n \varsigma^2 \lim_{n \to \infty} \sum_{i=1}^{n} (\tilde{\beta}_k^{(j,n)})^2 = 0.
\] (A.46)
By using (A.46), it is easy to show that
\[
\int_0^{t_n} V_n(t) dt = \int_0^{t_n} \left( \sum_{j=1}^{m} (q_j - \gamma_j t) \right)^2 l_j^\top l_j (1 - t/t_n)^2 dt \\
+ \int_0^{t_n} \left( \sum_{j=1}^{m} (q_j - \gamma_j t)^2 s_j^2 + \sum_{j=1}^{m} (q_j - \gamma_j t_n)^2 t_n^2 \varsigma_{z(j,n)}^2 \right) dt,
\]
(A.47)
and that
\[
\int_{t_n}^{\infty} V_n(t) dt = \int_{t_n}^{\infty} \sum_{j=1}^{m} \left( \frac{(t-T_j)^+}{T_j} (q_j - \gamma_j t) \right)^2 \left( s_j^2 + \varsigma_{z(j,n)}^2 \right) dt.
\]
(A.48)
If we define
\[
\tilde{J}^*(q) = \sum_{j=1}^{m} \frac{s_j^2}{3\gamma_j} |q_j|^3,
\]
then we have
\[
J_n(q) = \int_0^{t_n} V_n(t) dt + \int_{t_n}^{\infty} V_n(t) dt \\
\leq \frac{1}{3} \left| \sum_{j=1}^{m} q_j l_j \right|_2^2 t_n + \int_{t_n}^{\infty} \sum_{j=1}^{m} \left( \frac{(t-T_j)^+}{T_j} (q_j - \gamma_j t) \right)^2 \left( s_j^2 + \varsigma_{z(j,n)}^2 \right) dt \\
= \frac{1}{3} \left| \sum_{j=1}^{m} q_j l_j \right|_2^2 t_n + \sum_{j=1}^{m} \frac{s_j^2}{3\gamma_j} |q_j|^3 + \sum_{j=1}^{m} \frac{s_j^2}{3\gamma_j} |q_j|^3.
\]
(A.49)
By using (A.46) and the fact that \( \lim_{n \to \infty} t_n = 0 \), we have
\[
\lim_{n \to \infty} J_n(q) \leq \tilde{J}^*(q).
\]
Since the optimal cost should be less than or equal to any feasible trading strategy, we have
\[
J^*_n(q) \leq J_n(q).
\]
Thus, we have proved that
\[ \lim_{n \to \infty} J_n^*(q) \leq \lim_{n \to \infty} J_n(q) \leq \tilde{J}^*(q). \]

Combined with Theorem 12, this yields
\[ \tilde{J}^*(q) \leq \lim_{n \to \infty} J_n^*(q) \leq \tilde{J}^*(q), \]
and the proof of the theorem is complete. ■

**Theorem 14 (Convergence Speed).** Asymptotically, the difference between the liquidity cost and the theoretical limit converges at rate \( 1/\sqrt{\lambda_{\min}^{(n)}} \):

\[ \lim_{n \to \infty} \sup \sqrt{\lambda_{\min}^{(n)}} |J_n(q) - J_\infty^*(q)| < \infty. \tag{3.24} \]

**Proof.** According to (A.39), (A.43), (A.46), and (A.49), we have
\[ J_n(q) - \tilde{J}^*(q) \leq \frac{1}{3} \left\| \sum_{j=1}^{m} q_j l_j \right\|_2^2 t_n + \sum_{j=1}^{m} \frac{\zeta_2^{(j,n)}}{3 \gamma_j} |q_j|^3 \\
= \frac{1}{3} \left\| \sum_{j=1}^{m} q_j l_j \right\|_2^2 \frac{\sum_{j=1}^{m} |q_j|}{N(n) \bar{\gamma}} + \sum_{j=1}^{m} \frac{\zeta_2^{(j,n)}}{3 \gamma_j} |q_j|^3 \tag{A.50} \]
\[ \leq A \frac{1}{\sqrt{\lambda_{\min}^{(n)}}} + B \frac{1}{\lambda_{\min}^{(n)}}, \]
where \( A, B \) are constants that are not related to \( n \):
\[ A = \frac{1}{3} \left\| \sum_{j=1}^{m} q_j l_j \right\|_2^2 \frac{\sum_{j=1}^{m} |q_j|}{\bar{\gamma}} \sqrt{t^2 K^2}, \quad B = \sum_{j=1}^{m} \frac{|q_j|^3}{3 \gamma_j}. \]

Given that \( \lambda_{\min}^{(n)} \to \infty \), we have
\[ \lim_{n \to \infty} \sup \sqrt{\lambda_{\min}^{(n)}} |J_n(q) - \tilde{J}^*(q)| \leq \lim_{n \to \infty} \sup \sqrt{\lambda_{\min}^{(n)}} |J_n(q) - \tilde{J}^*(q)| \leq A. \]
Theorem 15 (Random factor loading). If the asset factor loadings are drawn independently from a $K$-dimensional distribution (with a finite second moment), then, asymptotically, we have

$$\frac{\lambda_{\min}(n)}{n} \xrightarrow{a.s.} C,$$

where $C$ is some constant that depends on only the distribution of factor loadings, and, therefore,

$$\limsup_{n \to \infty} \sqrt{n}|J^*_n(q) - J^*_\infty(q)| < \infty, a.s.$$ (3.26)

Proof. Now suppose that the factor loadings are i.i.d., and define $\hat{G} \in \mathbb{R}^{K \times K}$ as

$$\hat{G}_{ij} = \begin{cases} 
E[l_{ki}l_{kj}], & i \neq j, \\
E[l^2_{ki}], & i = j.
\end{cases}$$

Given the matrix $G^{(n)} = (L^{(n)})^\top L^{(n)}$, we have

$$G^{(n)}_{ij} = \begin{cases} 
\sum_{k=1}^n l_{ki}l_{kj}, & i \neq j, \\
\sum_{k=1}^n l^2_{ki}.
\end{cases}$$

Then, by adopting the strong law of large numbers, we have

$$\frac{G^{(n)}}{n} \xrightarrow{a.s.} \hat{G}.$$

Suppose that $\hat{\lambda}_{\min}$ is the smallest eigenvalue of $\hat{G}$. It is easy to see that $det(\hat{G}) > 0$ if there is no perfect linearity in the factor loadings. As a result, we have

$$\frac{\lambda^{(n)}_{\min}}{n} \xrightarrow{a.s.} \hat{\lambda}_{\min}.$$ (A.51)

The theorem is proved by plugging (A.49) into Theorem 14.
APPENDIX A. APPENDIX

**Theorem 16 (Generalization).** For any liquidation model specified in (3.27), if the transaction cost functional is twice differentiable with

\[ \hat{f}'(0) = 0, \quad \hat{f}(0) = 0, \]

the extended results of Theorem 13 still hold. More specifically, we have

\[ \lim_{n \to \infty} J^*_n(q) = \sum_{j=1}^m \tilde{J}^*(q_j, \varsigma_j, \nu_j, \gamma_j), \quad (3.29) \]

where \( J^*_n(q) \) represents the optimal liquidity costs for a portfolio \( q \) with assets in \( A_n \).

**Proof.** We first prove the following lemma.

**Lemma 5.** Suppose that \( f \) is a convex function:

\[ f : [-1, 1] \to \mathbb{R}_+ \cup \{+\infty\}. \]

If \( f(\cdot) \) is twice differentiable and

\[ f(0) = 0, \quad f'(0) = 0, \quad f''(0) > 0, \quad (A.52) \]

then there exists \( \delta > 0 \) such that for any \( |\beta| < \delta \),

\[ f(\beta x) \leq C\beta^2 f(x), \quad \forall |x| \leq 1, \quad (A.53) \]

where \( C \) is some constant.

**Proof.** The case for \( x = 0 \) is trivial. Turning to the case where \( x \neq 0 \), by convexity we know that \( f(x) > 0 \):

\[ \lim_{\beta \to 0} \frac{f(\beta x)}{(\beta x)^2} = f''(0). \quad (A.54) \]
Hence there exists $\delta$ such that for $\forall \beta < \delta$ we have
\[
\frac{1}{2} f''(0) \leq \frac{f(\beta x)}{(\alpha x)^2} \leq \frac{3}{2} f''(0).
\]

Denote
\[
C_1 = \max_x \left(\frac{x^2}{f(x)}\right).
\]

Then, $\forall \alpha < \delta$, we have
\[
\frac{f(\alpha x)}{\alpha^2 f(x)} = \frac{f(\alpha x)}{\alpha^2} \frac{x^2}{f(x)} \leq \frac{3}{2} f''(0)C_1.
\]

This proves the lemma.

Following the proof in Theorem 13 we construct a series of portfolios $\{z^{(j,n)}\}$ for each asset $j \leq m$. Moreover, we denote $\tilde{u}_j^*(t)$ to be the optimal trading strategy of liquidating asset $j$ alone without hedging with other assets and only considering its idiosyncratic risk.

To simplify the notations, we denote $\check{\beta}_k^{(j,n)} = \sum_{j=1}^{m} \frac{-q_j}{|q_j|} \tilde{\beta}_k^{(j,n)}$, where $\tilde{\beta}_k^{(j,n)}$ is defined in (A.42). It is easy to see that we also have $\sum_{k=1}^{n} (\check{\beta}_k^{(j,n)})^2 \to 0$.

Define
\[
N(n) = \frac{1}{\sqrt{\sum_{k=1}^{n} (\check{\beta}_k^{(j,n)})^2}};
\]

hence we have $N(n) \to \infty$.

Now consider the following trading strategy:

1. For $0 \leq t \leq t_n$, buy portfolio $z^{(j,n)}$ at a rate of $\frac{-q_j}{|q_j|} \sqrt{N(n)}$, where $t_n$ is given by
\[
t_n = \frac{\sum_{j=1}^{m} |q_j|}{\sqrt{N(n)}}.
\]

2. For $t > t_n$, trade asset $j$ at a rate of $\tilde{u}_j^*(t - t_n)$ and trade $z^{(j,n)}$ at a rate of $-\tilde{u}_j^*(t - t_n)$. Do it for all $1 \leq j \leq m$.

In this case, the total cost is made up of two parts, namely, transaction costs and market risks. Let us look at them separately.
First, following this trading strategy \((V_n(t))\), the market risk contribution of the position at time \(t\) is given by

\[
V_n(t) = \left\{ \begin{array}{ll}
\left[ \sum_{j=1}^{m} q_j (l_j - \frac{t}{t_n} l_{z,(j,n)}) \right]^\top \left[ \sum_{j=1}^{m} q_j (l_j - \frac{t}{t_n} l_{z,(j,n)}) \right] + \sum_{j=1}^{m} (q_j - \gamma_j t)^2 \varsigma_j^2 + \sum_{j=1}^{m} (q_j - \gamma_j t_n)^2 \frac{t^2}{t_n^2} \varsigma_{z,(j,n)}^2, & 0 \leq t \leq t_n, \\
\sum_{j=1}^{m} \left( q_j - \int_0^{t-t_n} \tilde{u}_j^* (s) ds \right) (l_j - l_{z,(j,n)}) \right\}^\top \left[ \sum_{j=1}^{m} \left( q_j - \int_0^{t-t_n} \tilde{u}_j^* (s) ds \right) (l_j - l_{z,(j,n)}) \right] + \sum_{j=1}^{m} \left( q_j - \int_0^{t-t_n} \tilde{u}_j^* (s) ds \right)^2 \left( \varsigma_j^2 + \varsigma_{z,(j,n)}^2 \right), & \text{otherwise.}
\end{array} \right.
\]

where \(\varsigma_{z,(j,n)}^2, l_{z,(j,n)}\) are defined in the same way. Further, we have

\[
l_{z,(j,n)} = l_j, \quad \lim_{n \to \infty} \varsigma_{z,(j,n)}^2 = 0, \quad \lim_{n \to \infty} t_n = 0. \tag{A.55}
\]

Similarly to the proofs in Theorem 13, it can be shown that

\[
\lim_{n \to \infty} \int_0^\infty V_n(t) dt = \sum_{j=1}^{m} \int_0^\infty \left( q_j - \int_0^t \tilde{u}_j^* (s) ds \right)^2 \varsigma_j^2 dt.
\]

Now let’s consider the transaction costs. The transaction costs at time \(t\) \((T_n(t))\) are given by

\[
T_n(t) = \left\{ \begin{array}{ll}
\sum_{k=1}^{n} \nu_k \hat{f} \left( \sum_{j=1}^{m} \frac{q_j \sqrt{N(n)}}{\sum_{j=1}^{m} |q_j|} \beta_{j,(n)} / \gamma_k \right) = \sum_{k=1}^{n} \nu_k \hat{f} \left( \beta_{j,(n)} / \sqrt{N(n)} / \gamma_k \right), & 0 \leq t \leq t_n, \\
\sum_{j=1}^{n} \nu_j \hat{f} \left( - \sum_{k=1}^{m} \beta_{j,(n)} / \gamma_k \right) + \sum_{j=1}^{m} \nu_j \hat{f} \left( \tilde{u}_j^* (t-t_n) / \gamma_j + \sum_{k=1}^{m} \beta_{j,(n)} / \gamma_k \right), & \text{otherwise.}
\end{array} \right.
\]

Notice that \(|\beta_{j,(n)}| \leq \sqrt{\sum_{k=1}^{m} (\beta_{j,n,k})^2} = 1/N(n)\); hence we have \(\beta_{j,(n)} / \sqrt{N(n)} \to 0\). Then we can take the Taylor expansion of \(\hat{f}(\cdot)\) around 0:

\[
\hat{f} \left( \beta_{j,(n)} / \sqrt{N(n)} / \gamma_k \right) = \frac{1}{2} N(n)(\beta_{j,(n)})^2 / \gamma_k^2 + o(N(n)(\beta_{j,(n)})^2).
\]
We have

\[
\lim_{n \to \infty} \int_0^{t_n} T(t) dt = \lim_{n \to \infty} \int_0^{t_n} \sum_{k=1}^{n} \nu_k \hat{f} \left( \hat{\beta}^{(j,n)}_k \sqrt{N(n)/\gamma_k} \right) dt
\]

\[
= \lim_{n \to \infty} \int_0^{t_n} \sum_{k=1}^{n} \nu_k \left( \frac{1}{2} N(n) (\hat{\beta}^{(j,n)}_k)^2 / \gamma_k^2 + 2\alpha N(n)(\hat{\beta}^{(j,n)}_k)^2 \right) dt
\]

\[
\leq \lim_{n \to \infty} \frac{\tilde{\nu}}{2\gamma^2} t_n \left( N(n) \sum_{k=1}^{n} (\hat{\beta}^{(j,n)}_k)^2 \right) (A.56)
\]

\[
= \lim_{n \to \infty} \frac{\tilde{\nu}}{2\gamma^2} t_n
\]

\[
= 0.
\]

Then,

\[
\lim_{n \to \infty} \int_{t_n}^{\infty} T(t) dt = \lim_{n \to \infty} \int_{t_n}^{\infty} \sum_{k=1}^{n} \nu_k \hat{f} \left( -\sum_{j=1}^{m} \hat{\beta}^{(j)}_{nk} \tilde{u}^*_j(t - t_n)/\gamma_k \right) dt
\]

\[
+ \lim_{n \to \infty} \int_{t_n}^{\infty} \sum_{j=1}^{m} \nu_j \hat{f} \left( \tilde{u}^*_j(t - t_n)/\gamma_j + \sum_{k=1}^{m} \hat{\beta}^{(k)}_{nj} \tilde{u}^*_k(t - t_n)/\gamma_j \right) dt. (A.57)
\]

Notice that the second term is

\[
\lim_{n \to \infty} \int_{t_n}^{\infty} \sum_{j=1}^{m} \nu_j \hat{f} \left( \tilde{u}^*_j(t - t_n)/\gamma_j + \sum_{k=1}^{m} \hat{\beta}^{(k)}_{nj} \tilde{u}^*_k(t - t_n)/\gamma_j \right) dt = \int_0^{\infty} \sum_{j=1}^{m} \nu_j \hat{f}(\tilde{u}^*_j(t)/\gamma_j) dt. (A.58)
\]

Now it remains to show that the first term converges to 0:

\[
\lim_{n \to \infty} \int_{t_n}^{\infty} \sum_{k=1}^{n} \nu_k \hat{f} \left( -\sum_{j=1}^{m} \hat{\beta}^{(j)}_{nk} \tilde{u}^*_j(t - t_n)/\gamma_k \right) dt
\]

\[
\leq \tilde{\nu} \sum_{j=1}^{m} \lim_{n \to \infty} \int_{t_n}^{\infty} \sum_{k=1}^{n} \hat{f} \left( \hat{\beta}^{(j)}_{nk} \tilde{u}^*_j(t - t_n)/\gamma_k \right) dt
\]

\[
\leq \tilde{\nu} \sum_{j=1}^{m} \lim_{n \to \infty} \int_{t_n}^{\infty} \left( \hat{\beta}^{(j)}_{nk} \right)^2 \int_{t_n}^{\infty} \hat{f} \left( \tilde{u}^*_j(t - t_n)/\gamma_k \right) dt
\]

\[
= 0.
\]

148
The first inequality is a direct application of $E$ and the last equality is due to the fact that $\sum_{k=1}^{n}(\hat{\beta}_{nk})^2 \rightarrow 0$.

Notice that the total liquidity cost of this strategy ($J_n(q)$) is given by

$$J_n(q) = \int_0^\infty T_n(t)dt + \mu \int_0^\infty V_n(t). \quad (A.60)$$

Given that this cost should never be smaller than the optimal cost, together with (A.56), (A.57), (A.58), (A.59), and (A.60), we have

$$\lim_{n \to \infty} J_n^*(q) \leq \lim_{n \to \infty} J_n(q) = \int_0^\infty \sum_{j=1}^{m} \nu_j \hat{f}(\tilde{u}_j^*(t)/\gamma_j)dt + \mu \sum_{j=1}^{m} \int_0^{\infty} \left( q_j - \int_0^t \tilde{\gamma}_j u_j^*(s)ds \right)^2 \varsigma_j^2$$

$$= \sum_{j=1}^{m} \tilde{J}^*(q_j, \varsigma_j, \nu_j, \gamma_j).$$

**Theorem 17 (Linear costs).**

$$\lim_{n \to \infty} \lim_{||q||_\infty \to \infty} \frac{\tilde{J}_{LC}^*(q)}{J_{LC,n}^*(q)} = 1 \quad (3.30)$$

where

$$\tilde{J}^*_{LC}(q) = \mu \sum_{j=1}^{m} \frac{\varsigma_j^2}{3\gamma_j}|q_j|^3.$$

**Proof.** From Theorem 12 we know that the term in the denominator is the lower bound of the liquidity cost of the portfolio when we neglect the transaction costs. Thus, it also has to be the lower bound of $J_n^*(q)$:

$$\frac{\mu \sum_{j=1}^{m} \frac{\varsigma_j^2}{3\gamma_j}|q_j|^3}{J_n^*(q)} \leq 1. \quad (A.61)$$

Now consider the trading strategy given in the proof of Theorem 13, the only difference here is that we also have to calculate the transaction cost from this strategy. The transaction cost comes from two sources: the transaction cost of selling the position in the portfolio that is given by $\sum_{j=1}^{m} \nu_j |q_j|$ and the transaction cost of establishing and liquidating the hedging positions. More
specifically, the trading cost of trading portfolio \( z^{(j,n)} \) is given by

\[
\nu_{z^{(j,n)}} = \sum_{k=1}^{n} |\hat{\beta}_k^{(j,n)}| \nu_k. \tag{A.62}
\]

By Cauchy–Schwarz inequality, we have

\[
\left( \sum_{k=1}^{n} |\hat{\beta}_k^{(j,n)}| \right)^2 \leq n \sum_{k=1}^{n} |\hat{\beta}_k^{(j,n)}|^2 \to 0, \quad n \to \infty, \tag{A.63}
\]

Combining (A.62) and (A.63), we know that when \( n \) is large enough,

\[
\nu_{z^{(j,n)}} < \tilde{\nu} \sqrt{n}. \tag{A.64}
\]

The total transaction costs for the \( n \)th problem are given by

\[
TC_n \leq 2 \sum_{j=1}^{m} \nu_{z^{(j,n)}} |q_j| + \sum_{j=1}^{m} \nu_j |q_j| < \sum_{j=1}^{m} (2\tilde{\nu} \sqrt{n} + \nu_j) |q_j| < m(2\tilde{\nu} \sqrt{n} + \nu_j) ||q||_{\infty}.
\]

Thus we have

\[
\lim_{||q||_{\infty} \to \infty} \frac{TC_n}{||q||_{\infty}} = 0.
\]

\[
\lim_{n \to \infty} \lim_{||q||_{\infty} \to \infty} \frac{\mu \sum_{j=1}^{m} \frac{s_j^2}{\sigma_j^2} |q_j|^3}{J_{LC,n}^*(q)} \geq \lim_{n \to \infty} \lim_{||q||_{\infty} \to \infty} \frac{\mu \sum_{j=1}^{m} \frac{s_j^2}{\sigma_j^2} |q_j|^3}{J_n(q) + TC_n} \geq \lim_{n \to \infty} \frac{\mu \sum_{j=1}^{m} \frac{s_j^2}{\sigma_j^2} |q_j|^3}{J_n(q)} = 1. \tag{A.65}
\]

Combining (A.61) and (A.65), we complete the proof of the theorem.

\[\blacksquare\]

**Theorem 18 (ETF).** If the large-universe property is satisfied, then, asymptotically, the cost for
liquidating $q_j$ shares of asset $j \leq m$ is given by
\[
\lim_{n \to \infty} J_{\text{ETF},n}^*(q_j) = \frac{\varsigma_j^2}{3(|\alpha_1|\gamma_{\text{ETF}} + \gamma_j)} q_j^3,
\]
(3.32)
where $q_i = 0, \forall i \neq j$.

**Proof.** We start with the following observations:

1. For asset $j$, the fastest trading rate attainable is $|\alpha_j|\gamma_{\text{ETF}} + \gamma_j$. It is obtained by selling asset $j$ and the ETF at full rate, but at the same time buying back other assets in the ETF, so that the net liquidity contribution from trading the ETF is just $|\alpha_j|\gamma_{\text{ETF}}$. The feasibility of this strategy is guaranteed by the assumption that $|\alpha_i\gamma_{\text{ETF}}| < \gamma_i$.

2. At any time point during the liquidation process, the total risk is made up of three components: the idiosyncratic risk of asset $j$, which is given by $\varsigma_j^2 x_j^2(t)$; idiosyncratic risks of the hedging positions; and the entire portfolio’s market risk. Since the latter two terms are nonnegative, the total risk term is always greater than or equal to $\varsigma_j^2 x_j^2(t)$.

From these two observations, we have
\[
J_{\text{ETF},n}^*(q_j) \geq \int_0^\infty \varsigma_j^2 x_j^2(t) dt \geq \frac{\varsigma_j^2}{3(|\alpha_j|\gamma_{\text{ETF}} + \gamma_j)} q_j^3.
\]
As a result, the right-hand side of (3.32) is actually the lower bound of the liquidity cost.

As discussed above, it is possible to trade asset $j$ at rate $|\alpha_j|\gamma_{\text{ETF}} + \gamma_j$ by essentially trading a portfolio $p$ containing the ETF and all the assets in it (this portfolio has only a net position in asset $j$). By viewing this portfolio as a single asset and using the results in Theorem [13] we have
\[
\lim_{n \to \infty} J_{\text{ETF},n}^*(q_j) = \frac{\varsigma_j^2}{3\gamma_p} q_j^3,
\]
where $\gamma_p = |\alpha_1|\gamma_{\text{ETF}} + \gamma_j$, and this completes the proof of the theorem.
A.3. Additional Proofs for Chapter 4

Theorem 19 (Value Function for Market Maker). The value function $V(q, \delta)$ is linear in $\delta$; that is, it takes the form

$$V(q, \delta) = \alpha(q)\delta - \beta(q),$$

where the functions $\alpha : \mathbb{R}_+ \to \mathbb{R}$ and $\beta : \mathbb{R}_+ \to \mathbb{R}$ are uniquely determined by the integral equations

$$\alpha(q) = \frac{\mu}{\mu p_u^+ + \gamma + \eta^+} \left\{ \int_0^q (\alpha(q-x) - 1)f(x)dx + \frac{\gamma p_f^+}{\mu p_u^+ + \gamma + \eta^+} \right\}$$

$$+ \frac{\eta^+}{\mu p_u^+ + \gamma + \eta^+} \int_0^1 \alpha(\ell q)g(\ell)d\ell,$$

$$\beta(q) = \frac{\mu}{\mu p_u^+ + \gamma + \eta^+} \left\{ \int_0^q \beta(q-x)f(x)dx + \lambda \int_0^q (\alpha(q-x) - 1)xf(x)dx ight.$$ 

$$- \lambda \bar{u}^+ (\alpha(q) - 1) \right\} + \frac{\gamma \bar{J}^+}{\mu p_u^+ + \gamma + \eta^+} + \frac{\eta^+}{\mu p_u^+ + \gamma + \eta^+} \int_0^1 \beta(\ell q)g(\ell)d\ell,$$

for $q > 0$, with boundary conditions

$$\alpha(0) = \frac{\mu p_u^+ + \gamma p_f^+}{\mu p_u^+ + \gamma}, \quad \beta(0) = \frac{\mu [\gamma (1 - p_f^+)]}{(\mu p_u^+ + \gamma)^2} \lambda \bar{u}^+ + \frac{\gamma}{\mu p_u^+ + \gamma} \bar{J}^+.$$

Proof. First of all, we solve for the solution to equation (4.4). The boundary condition can be verified by setting $q = 0$ in (4.4), which gives

$$V(0, \delta) = \frac{\mu}{\zeta} \mathbb{E} \left[ \mathbb{I}_{\{u \geq 0\}}(\delta - \lambda u) + \mathbb{I}_{\{u < 0\}} V(q, \delta - \lambda u) \right]$$

$$+ \frac{\gamma}{\zeta} \mathbb{E} \left[ \mathbb{I}_{\{J > 0\}}(\delta - J) \right]$$

$$+ \frac{\eta^+}{\zeta} V(0, \delta).$$

Notice that it’s an integral equation with a linear drift on $\delta$. Hence the solution of $V(0, \delta)$
should also be linear on $\delta$. The equation above thus boils down to

\[
\frac{\mu + \gamma}{\mu} (\alpha(0)\delta - \beta(0)) = E \left[ \mathbb{I}_{\{u>0\}}(\delta - \lambda u) \right] + E \left[ \mathbb{I}_{\{u\leq 0\}} (\alpha(0)(\delta - \lambda u) - \beta(0)) \right] + \gamma(\delta - \bar{J}^+)/\mu \\
= p^+\delta - \lambda \int_0^{+\infty} uf(u)du + \int_0^0 (\alpha(0)(\delta - \lambda u) - \beta(0)) f(u)du + \gamma(\delta - \bar{J}^+)/\mu.
\]

Solving the equation above for $\alpha(0)$ and $\beta(0)$, we obtain the boundary condition:

\[
\alpha(0) = \frac{\mu p_u^+ + \gamma p_J^+}{\mu p_u^+ + \gamma}, \quad \beta(0) = \frac{\mu[\gamma(1 - P_J^+)]}{(\mu p_u^+ + \gamma)^2} \lambda \bar{u}^+ + \frac{\gamma}{\mu p_u^+ + \gamma} \bar{J}^+.
\]

For $q > 0$, the integral equation still has a linear drift, which provides the linearity of the solution:

\[
\frac{\mu + \gamma + \eta^+}{\mu} (\alpha(q)\delta - \beta(q)) = E \left[ \mathbb{I}_{\{u>q\}}(\delta - \lambda u) \right] + E \left[ \mathbb{I}_{\{u\leq 0\}} (\alpha(q)(\delta - \lambda u) - \beta(q)) \right] \\
+ E \left[ \mathbb{I}_{\{0<u\leq q\}} (\alpha(q-u)(\delta - \lambda u) - \beta(q-u)) \right] \\
+ \gamma(\delta - \bar{J}^+)/\mu + \eta^+ E [\alpha(\ell q)\delta - \beta(\ell q)] \\
= \int_q^{+\infty} (\delta - \lambda u) f(u)du + \int_q^0 (\alpha(q-u)(\delta - \lambda u) - \beta(q-u)) f(u)du \\
+ \int_0^0 (\alpha(q)(\delta - \lambda u) - \beta(q)) f(u)du \\
+ \gamma(\delta - \bar{J}^+)/\mu + \frac{\eta^+}{\mu} \int_0^1 (\alpha(\ell q)\delta - \beta(\ell q))d\ell.
\]

Solving the equation for $\alpha(q)$ and $\beta(q)$, we obtain the solution:

\[
V(q, \delta) = \alpha(q)\delta - \beta(q).
\]

Now we would like to prove the uniqueness and existence of the solution to equations (4.6) and (4.7). Notice that $\alpha(\cdot)$ is defined on $\mathbb{R}^+$. Then the expression of $\alpha(q)$ is a Volterra integral equation of the second kind. We can rewrite the expression as follows:

\[
\alpha(q) = k_1(q) + \int_0^q k_2(x, q, \alpha(x))dx, \quad \forall q \in \mathbb{R}^+.
\]
where
\[ k_1(q) = \frac{\mu}{\mu \rho_u + \gamma + \eta^+} (p_u^+ - \int_0^q f(x) \, dx) + \frac{\gamma p_j^+}{\mu \rho_u + \gamma + \eta^+}, \quad \forall q \in \mathbb{R}^+, \]
\[ k_2(x, q, z) = \{ \frac{\mu f(q - x)}{\mu \rho_u + \gamma + \eta^+} + \frac{\eta^+ g(x/q)/q}{\mu \rho_u + \gamma + \eta^+} \} z, \quad \forall q \in \mathbb{R}^+, x \in \mathbb{R}^+, z \in \mathbb{R}. \]

Given the continuity of \( f(\cdot), g(\cdot) \), we have \( k_1 \in C(\mathbb{R}^+) \), \( k_2 \in C(\mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}) \). Also, it is trivial that \( k_2 \) satisfies the following Lipschitz condition:
\[ |k_2(x, q, z) - k_2(x, q, z')| \leq L(x, q)|z - z'|, \text{ for some } L \in C(\mathbb{R}^+ \times \mathbb{R}^+). \]

Hence by Theorem 2.1.1 of Hackbusch (1995), there is exactly one solution of the integral equation (A.66). Additionally, the solution \( \alpha(\cdot) \) is continuous on \( \mathbb{R}^+ \).

The existence and uniqueness of \( \beta(\cdot) \) can be established in a similar way. Specifically, we can write equation (4.7) in the following form:
\[ \beta(q) = k_3(q) + \int_0^q k_4(x, q, \beta(x)) \, dx, \quad \forall q \in \mathbb{R}^+, \quad (A.67) \]

where
\[ k_3(q) = \frac{\mu}{\mu \rho_u + \gamma + \eta^+} \left\{ \lambda \int_0^q (\alpha(q - x) - 1) x f(x) \, dx - \lambda \bar{a}^+ (\alpha(q) - 1) \right\} + \frac{\gamma \bar{J}^+}{\mu \rho_u + \gamma + \eta^+}, \quad \forall q \in \mathbb{R}^+, \]
\[ k_4(x, q, z) = k_2(x, q, z), \quad \forall q \in \mathbb{R}^+, x \in \mathbb{R}^+, z \in \mathbb{R}. \]

Hence, by a similar analysis, there is exactly one solution to integral equation (4.7), and that solution is continuous.

\[ \text{Theorem 20. \ 1. Compared with equation (4.1), we have} \]
\[ \alpha_t = \alpha(q), \quad AS_t = \frac{\beta(q)}{\alpha(q)}. \]
2. The probability of execution $\alpha(q)$ is non-increasing in queue position.

3. The adverse selection is positive

$$\beta(q)/\alpha(q) > 0.$$ 

4. With no cancellations ($\eta = 0$), we have

$$\lim_{q \to \infty} \alpha(q) = p_{J}^{+}, \quad \lim_{q \to \infty} \beta(q) = J^{+}.$$ 

**Proof.**

1. Consider an order placed on the ask side at position $q$ at time $0$; denote $\tau^{*}$ to be the time that it is filled or canceled.

$$V(q, \delta) = \mathbb{E} \left[ (P^{A} - P) I_{\{\text{FILL}\}} \right]$$

$$= \mathbb{E} \left[ (P^{A} - P_{0}) - (P - P_{0}) \right] I_{\{\text{FILL}\}}$$

$$= \mathbb{E} \left[ (\delta - (P - P_{0})) I_{\{\text{FILL}\}} \right]$$

$$= \mathbb{P}(\text{FILL}) \delta - \mathbb{P}(\text{FILL}) \mathbb{E} [P_{\tau^{*}} - P_{0}] | \text{FILL}].$$

Notice that $\mathbb{E} [P_{\tau^{*}} - P_{0}] | \text{FILL}$ represents the opportunity cost conditional on executing the order, which coincides with the definition of adverse selection.

Compared to the notations in equation (4.5), it is easy to see that

$$\alpha(q) = \mathbb{P}(\text{FILL}) \quad \beta(q)/\alpha(q) = \mathbb{E} [P_{\tau^{*}} - P_{0}] | \text{FILL}].$$

Hence $\alpha(q)$ is exactly the probability of the order being executed, and $\beta(q)/\alpha(q)$ represents the adverse selection cost.

2. It suffices to show that $\forall 0 \leq q_{0} < q_{1}$, we have $\alpha(q_{0}) \geq \alpha(q_{1}).$

Consider an infinitesimal order $A_{0}$ with a queue position $q_{0}$, and let $\mathcal{E}_{0}$ be the set of events
APPENDIX A. APPENDIX

that the order is eventually filled. Then we have

$$\alpha(q_0) = \mathbb{P}(E_0).$$

Notice that in our model, the value of an order does not depend on the orders that follows it in the queue. Hence it is possible to couple the order $A_0$ with an infinitesimal order $A_1$ in the exact same queue but with a position $q_1$. Similarly, we define $E_1$ to be the set of events that $A_1$ is eventually executed.

Notice that since the size of the orders is infinitesimal, the marginal probabilities $\mathbb{P}(E_0), \mathbb{P}(E_1)$ should be intact with coupling. There are two scenarios where $E_1$ can happen:

- $A_1$ is executed by a trade. Then, in our setup, there can be no price jump before this event. As $A_0$ is placed in front of $A_1$, it should be executed already.

- $A_1$ is executed by a positive price jump. In our setup, there can be no negative price jump before this event. Hence $A_0$ can be executed either by this positive price jump or by an earlier trade.

The above analysis shows that $\{E_1\} \subseteq \{E_0\};$ hence

$$\alpha(q_1) = \mathbb{P}(E_1) \leq \mathbb{P}(E_0) = \alpha(q_0).$$

3. Since $\forall q > 0, 0 < \alpha(q) < 1$, it suffices to show that $\forall q > 0, \beta(q) > 0$.

We have already proved that $\alpha(q)$ is increasing in $q$; hence $\forall 0 \leq x < q, \alpha(q - x) \geq \alpha(q)$.

According to equation (4.6), we have

$$\beta(q) \geq \frac{\mu}{\mu p_{\alpha}^+ + \gamma + \eta^+} \left\{ \int_0^q \beta(q - x)f(x) \, dx - \lambda \int_q^\infty (\alpha(q) - 1)xf(x) \, dx \right\} + \frac{\gamma J^+}{\mu p_{\alpha}^+ + \gamma + \eta^+} + \frac{\eta^+}{\mu p_{\alpha}^+ + \gamma + \eta^+} \int_0^1 \beta(\ell q)g(\ell) \, d\ell. \tag{A.69}$$

Notice that we have $\beta(0) > 0$ and that $\beta(\cdot)$ is continuous. Now suppose that $\beta(q)$ is not
always positive for \( q \geq 0 \); then there must exist \( q_0 \) such that \( \beta(\cdot) \) attains a value of zero for the first time. By continuity, we have \( \beta(q) > 0 \) for \( q \in [0, q_0) \). Notice that at \( q_0 \), the right-hand side of equation (A.69), is strictly positive; hence it is impossible that \( \beta(q_0) = 0 \). As a result, it must be that \( \beta(\cdot) \) is positive for all \( q \geq 0 \).

4. Given that p.d.f. of trade size \( f(\cdot) \) is assumed to be continuous over \([0, +\infty)\), we have

\[
\int_0^\infty e^{-st} f(t) dt \leq \int_0^\infty f(t) dt = p^+_u, \forall s \geq 0.
\]

Hence the Laplace transform of \( f(\cdot) \) exists on \([0, +\infty)\). Let \( P(s) \) denote the Laplace transform of p.d.f. \( f(\cdot) \) of trade size and define \( C(q) = \alpha(q) - 1 \). We have

\[
C(q) = \frac{\mu}{\mu p_u^+ + \gamma} \int_0^q C(q-x) f(x) dx + \frac{\gamma(p^+_u - 1)}{\mu p_u^+ + \gamma}.
\]  

(A.70)

Now take the Laplace transform on both sides of equation (A.70): we have

\[
\mathcal{L}\{C\}(s) = \frac{\mu}{\mu p_u^+ + \gamma} \mathcal{L}\{C\}(s) P(s) + \frac{\gamma(p^+_u - 1)}{s(\mu p_u^+ + \gamma)},
\]

\[
\Rightarrow \frac{\mu p_u^+ + \gamma - \mu P(s)}{\mu p_u^+ + \gamma} \mathcal{L}\{C\}(s) = \frac{\gamma(p^+_u - 1)}{s(\mu p_u^+ + \gamma)}.
\]

Given the fact that \( \forall s \geq 0, P(s) \leq p_u^+ \), we have \( \mu p_u^+ + \gamma - \mu P(s) > 0 \). As a result, the Laplace transform of \( C(q) \) is well defined on \([0, +\infty)\) and takes the form

\[
\mathcal{L}\{C\}(s) = \frac{\gamma(p^+_u - 1)}{s(\mu p_u^+ + \gamma - \mu P(s))}.
\]  

(A.71)

Hence, the Laplace transform for \( \alpha(q) \) is

\[
\mathcal{L}\{\alpha\}(s) = \mathcal{L}\{C\}(s) + 1/s = \frac{\gamma(p^+_u - 1)}{s(\mu p_u^+ + \gamma - \mu P(s))} + 1/s.
\]  

(A.72)
By the final value theorem of Laplace transform, we have
\[
\lim_{q \to \infty} \alpha(q) = \lim_{s \to 0} s\mathcal{L}\{\alpha\}(s) = \frac{-\gamma(p_J^+ - 1)}{\mu p + \gamma - \mu P(0)} + 1 = p_J^+.
\]
Similarly, it is easy to see that the Laplace transform of \(\beta(q)\) is also well defined on \([0, +\infty)\); hence we have
\[
\mathcal{L}\{\beta\}(s) = -\mu \frac{\lambda s\mathcal{L}\{C\}(s)P'(s) + \lambda \bar{u} + s\mathcal{L}\{C\}(s) - \gamma \bar{J}^+/(s\mu)}{\mu p + \gamma - \mu P(s)}.
\]
Then by the finite value theorem of Laplace transform:
\[
\lim_{q \to \infty} \beta(q) = \lim_{s \to 0} s\mathcal{L}\{\beta\}(s)
= \lim_{s \to 0} -\frac{\mu}{\mu p + \gamma - \mu P(s)} \left[ \lambda s\mathcal{L}\{C\}(s)P'(s) + \lambda \bar{u} + s\mathcal{L}\{C\}(s) - \gamma \bar{J}^+/(s\mu) \right]
= \bar{J}^+.
\]

**Theorem 21 (Exponential Trade Sizes)**. Suppose there are no cancellations and that the trades sizes follow the exponential distribution with parameter \(\theta > 0\), i.e.,
\[
f(u) \triangleq \frac{\theta}{2}e^{-\theta |u|},
\]
for \(u \in \mathbb{R}\). Then, the value function is given by \(V(\delta, q) = \alpha(q)\delta - \beta(q)\), where
\[
\alpha(q) = p_J^+ + \frac{\mu(1 - p_J^+)}{\mu + 2\gamma} e^{-bq},
\]
\[
\beta(q) = \bar{J}^+ (1 - \frac{\mu}{\mu/2 + \gamma} e^{-bq}) + \frac{\lambda \mu \gamma (p_J^+ - 1)}{2(\gamma + \mu/2)^2} e^{-bq} + \frac{\lambda (\gamma - \mu) (p_J^+ - 1)}{2(\gamma + \mu/2)^3} q e^{-bq},
\]
for all \(q \geq 0\), where \(b \triangleq \frac{(\gamma + \zeta)\theta}{\mu/2 + \gamma}\).

**Proof**. First denote \(P(s)\) as the Laplace transform of the truncated p.d.f. of trade size on the
positive domain \( f(u) = \frac{\theta}{2} e^{-\theta u} \). We have
\[
P(s) = \frac{\theta}{2(s + \theta)} \quad \bar{u}^+ = \frac{1}{2\theta}.
\] (A.75)

Plugging equation (A.75) into (A.72), we obtain the Laplace transform of \( \alpha(q) \):
\[
\mathcal{L}\{\alpha\}(s) = \frac{\gamma(p_j^+ - 1)}{s(\mu p_j^+ + \gamma - \mu P(s))} + 1/s.
\] (A.76)

Then, by taking the inverse Laplace transform, we get
\[
\alpha(q) = p_j^+ + \frac{\mu(1 - p_j^+)}{\mu + 2\gamma} e^{-\frac{\gamma\theta}{\gamma + \mu/2}}.
\] (A.77)

Similarly, we can plug equation (A.75) into (A.73) to obtain the Laplace transform of \( \beta(q) \):
\[
\mathcal{L}\{\beta\}(s) = \frac{\bar{J}^+}{s} - \frac{\mu}{\mu + 2\gamma} \frac{\bar{J}^+}{s + b} + \frac{\lambda\mu\gamma(p_j^+ - 1)}{2(\gamma + \mu/2)^2\theta(s + b)} + \frac{\lambda(\gamma - \mu)\gamma(p_j^+ - 1)}{2(\gamma + \mu/2)^3(s + b)^2}.
\] (A.78)

where \( b = \frac{\gamma\theta}{\gamma + \mu/2} \). Taking the inverse Laplace Transform, we get
\[
\beta(q) = \bar{J}^+(1 - \frac{\mu}{\mu + 2\gamma} e^{-bq}) + \frac{\lambda\mu\gamma(p_j^+ - 1)}{2(\gamma + \mu/2)^2\theta} e^{-bq} + \frac{\lambda(\gamma - \mu)\gamma(p_j^+ - 1)}{2(\gamma + \mu/2)^3} q e^{-bq}.
\] (A.79)