7-24-2020 10:00 AM

Exploration Of Stock Price Predictability In HFT With An Application In Spoofing Detection

Andrew Day, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Applied Mathematics

Abstract

Today many brokerage firms use computer algorithms to make trade decisions, submit orders, and manage orders after submission. This algorithmic trading is required to maximize execution speed and so minimize the cost, market impact and risk associated with trading large volumes of securities. Traders place orders to buy or sell a given amount of a security for a specific price on an exchange. These buy and sell orders accumulate in the 'order book' until they either find a counter-party for execution or are canceled. All participants can also issue market orders to buy or sell at the best available prices; these orders are immediately executed on a 'first come first serve' basis.

Using high frequency trading (HFT) data on the Toronto Stock Exchange, provided by the TMX Group, we explore a data driven model to detect a form of high frequency price manipulation – known as spoofing. A spoofer manipulates prices by placing limit orders which they do not intend to be executed in order to mislead other traders about the available volume of shares. The hope is that this will cause prices to move in their favour. We show that a generalized form of volume imbalance is associated with price movements and this can be manipulated by spoofing strategies. The literature argues spoofing strategies are detrimental to the integrity of markets and new models are necessary for regulators to combat them.

The size of the data sets we use definitely qualify for the moniker 'Big Data'. The limit order book must be constructed each time an order arrives for a particular stock. This process is implemented on a distributed data system using Pyspark since it would be impossible to do so, efficiently, on a local machine. We discuss some issues and complications that arise from working with very large data sets of this type.

We define a generalized volume imbalance as the weight in a convex combination of two price change distributions which forms our price change model. Price changes for different stocks happen at different time scales. We remedy this issue by comparing stocks on time intervals over which they all have the same variance in their price change distributions. Statistical and goodness of fit tests using Cramer's V statistic and Kullback–Leibler divergence, respectively, are implemented to validate our model across a large collection of stocks. The model is then used to test the sensitivity of the limit order book to spoofing and derive relationships between the spoofer's constraints and their optimal decisions. These results could then be implemented by regulators as a way to flag periods of the trading day where market conditions make spoofing possible as a means to improve market surveillance.

Keywords: limit order book, price manipulation, spoofing, data analysis, optimization, financial modelling, high frequency trading

Summary for Lay Audience

Price manipulation is detrimental to the integrity of financial markets. Price manipulation strategies have always existed, but, since the adoption of computer systems, new forms of price manipulation are emerging. In the past traders manipulated prices by injecting false or misleading information into the market in order to capitalize from resulting price movements and high frequency trading is not immune to these tactics. Traders can 'spoof' the market by strategically commiting specific orders to an exchange to buy or sell a set number of shares while actually never intending to allow their order to be executed. The idea is that other traders can see these spoofing orders, act on this misleading information, and move prices in the spoofer's favour.

Using high frequency order data on the Toronto Stock Exchange, provided by the TMX Group, we explore a data driven stock price model which is influenced by the orders arriving to the exchange. From our model we can calculate the average costs associated with a spoofer's optimal decisions to manipulate the market. We analyze this decision process to gain insights into how regulators can combat this type of illegal trade behaviour.

Acknowlegements

I would like to thank my supervisor, Dr. Matt Davison, for all of his support over the years. His knowledge and expertise have been instrumental in shaping the way I approach mathematical modelling. It was not easy transitioning away from physics, but he gave me all the opportunities and encouragement to make it as painless as possible. I would also like to thank Dr. Alex Buchel for his understanding and support when I decided to completely change research directions. I learned so much from both of you.

This project never would have happened without the Field's Institute's industrial problem solving workshops. They were an incredible place to learn new tools, and meet new friends and colleagues. Through these workshops I started my collaboration with the TMX group – which ultimately led to this work. I have to thank Abe Chan, and the rest of the people at TMX I had the pleasure of working with, for allowing me to pursue this project.

I would also like to thank the staff of applied mathematics, statistical and actuarial sciences, and the dean of science for making all of this just a little easier over the years.

Finally, I thank my friends and family for their unwavering support. I would not be where I am now without my parent's dedication to my education. It has been a long journey, and they have always been my greatest advisors and financiers. Most importantly, I have to thank Dr. Daisy Wong. It was not easy, but we both made it through together. Your passion for research has always inspired me.

Contents

List of Figures

List of Tables

List of Appendices

List of Abbreviations, Stocks, and Nomenclature

Abbreviations

HFT	High frequency trading
LOB	Limit order book
MAP	Maximum a posteriori estimation
MLE	Maximum likelihood estimation
TSX	Toronto stock exchange (Owned by TMX Group)

Stocks

ABX	Barrick Gold Corp
AEM	Agnico Eagle Mines Ltd
ARX	ARC Resources Ltd
BMO	Bank of Montreal
CNR	Canadian National Railway
HFU	BetaPro S&P/TSX Capped Financials 2× Daily Bull ETF
HSU	BetaPro S&P 500 2× Daily Bull ETF
PAAS	Pan American Silver Corp
PPL	Pembina Pipeline Corp
XEG	iShares S&P/TSX Capped Energy Index ETF

Nomenclature

Best ask	Best available price to buy stock
Best bid	Best available price to sell stock
Interarrival time	Time between successive orders
Limit order	Order to buy/sell a volume of stock at a specific price
Market order	Order to buy/sell a volume of stock regardless of price
Spoofing	Placement of limit orders with no intention of being executed
Spread	Difference between best ask and best bid
Tick	Increments of a penny
Touch	The best bid and/or the best ask, simultaneously

Chapter 1

Introduction

1.1 The Stock Price and Manipulation

Ownership of publicly traded companies is divided into shares of stock (or securities). These shares of stock naturally carry a monetary value which fluctuates over time to form a price time series – sequence of prices indexed in time order. Everyone has seen a stock price time series such as the one shown in Figure 1.1 where we see the price of Bank of Montreal (BMO) stock change over the entire trading day on April 17, 2017.

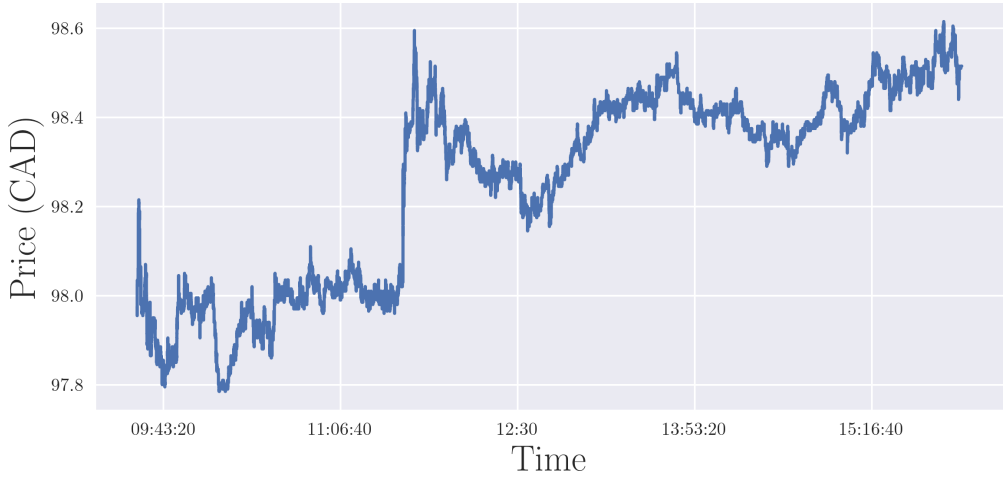
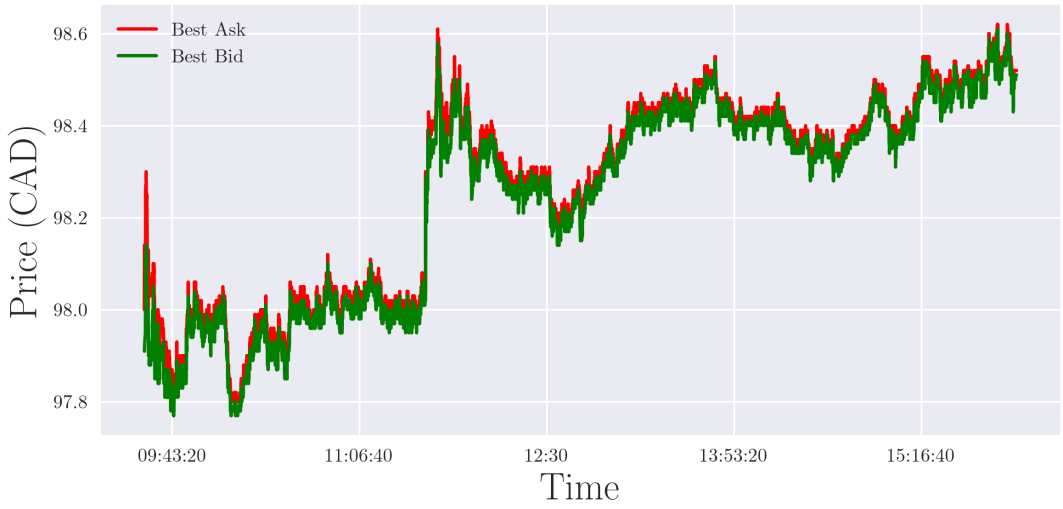


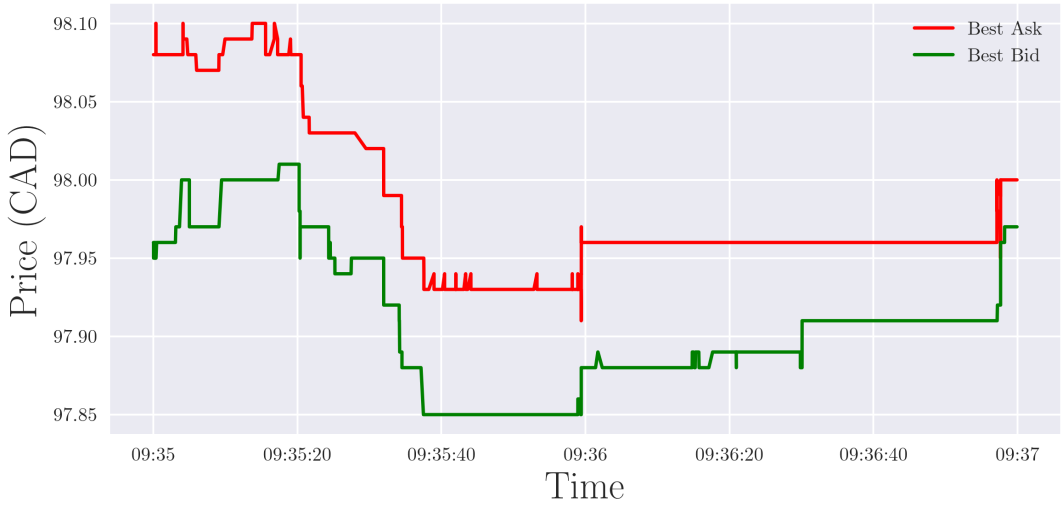
Figure 1.1: Mid point price for BMO stock on April 17, 2017 for the entire trading day.

One of the first topics covered when one begins to learn mathematical finance is modelling stock prices using continuous time stochastic processes. This is covered in detail by any mathematical finance text book, such as [1]. These processes assume the stock price changes continuously over time. Such stock price models have been effective tools since the initial work by Bachelier [2] in 1900 which culminated in the

Nobel prize winning research by Black, Scholes, and Merton, on pricing derivatives using these underlying stochastic processes [3]. Similarly, modelling stock price changes is instrumental in Markowitz' classical work in portfolio optimization [4]. However, there are key aspects of stock price dynamics which are not captured by these models. Namely, changes in stock prices are not continuous and do not change continuously through time. Also, there is no such thing as 'the' stock price.



(a) Entire Trading Day



(b) Two Minute Snapshot

Figure 1.2: Best ask and best bid for BMO stock on April 17, 2017 for the entire trading day and a two minute snapshot.

One wishing to purchase a share of stock would find there are different prices quoted for buyers and sellers and, in fact, there can be many different prices available to both

groups. The best available price to buy/sell a share of stock is called the best ask/bid and Figure 1.2 (a) displays the corresponding best ask and bid for BMO on April 17, 2017 over the entire trading day. The price shown in Figure 1.1 was the midpoint price – the average between the best bid and best ask.

Figure 1.2 (b) also shows how different things look over smaller time scales for the stock price. Here we see changes in the price are in increments of a penny, also refered to as a tick, and the prices change nonuniformly over time. However, these discontinuous price changes have been modelled in mathematical finance with jump diffusion processes [5] which are discontinuous at finitely many time points.

Prices are not driven by random movements, but by the actions of traders. It is the aggregate behaviour of traders in the market which drives the stock price. In most cases this is caused by traders buying and selling stocks over time with the bid and ask prices reflecting the current supply and demand of the stock. However, there has been a history of stock price dynamics being caused by manipulation. Two examples from Allen and Gale [6] are: 1) during the Napoleonic Wars there was stock and bond price manipulation on the London Stock Exchange where traders conspired with newspapers to spread false information about the state of the war to profit from the resulting price changes, and 2) in 1901 the managers of American Steel and Wire Company shorted[1] their stock then closed its steel mills causing the stock price to fall and earning themselves a profit after reopening the mills. The Securities Exchange Act of 1934 was an attempt to make manipulation more difficult in the United States by regulating information disclosure and monitoring the trading activities of firm insiders. However, manipulation in financial markets still exists and has been actively studied.

Literature exists on studying whether a financial model allows for trade-based market manipulation from buying and selling stocks [7–9], but a considerable amount of work has gone into the impacts of insider trading and 'pump and dump' strategies [6, 10–15]. So called 'pump and dump' strategies involve a trader buying stock and then injecting false information into the market to profit from increased price movements. A form of this strategy rose to prominance with the adoption of computers and the internet – email spam [16–18]. A manipulator will buy shares of stock and use email spam or social media platforms to spread false information in an attempt to get others to buy the same stock, increasing its price, so the manipulator can sell it at a profit. Manipulation has also been studied in more exotic cases with derivative markets [19] and dark pools [20]. In all these cases a key component is the injection of false or misleading information into

[1] A trader shorts a stock by borrowing shares and immediately selling them with the hope the price drops and they can buy the stock back and return it at a profit.

financial markets to influence the behaviour of traders and, by extension, the stock prices themselves.

Today, stock price dynamics are largely driven by electronic trading and the speed of execution which modern computers bring to the market. Like the emergence of email spam as a form of stock price manipulation, the age of high frequency trading (HFT) has opened up new avenues for manipulators to take advantage of. In the following section we introduce the mechanisms in which traders can interact with the stock market and how these interactions lead to price dynamics. In section 1.4 we go into detail on the form this high speed price manipulation takes – called spoofing.

Our first goal is to motivate and build a data-driven price change model from the aggregate trader behaviour. The second goal is to then apply this model to the application of high speed stock price manipulation in order to develop tools which could be used to aid in the detection of such behaviour. However, we need to present the mechanisms involved in electronic trading which will be important to incorporate into our model.

1.2 Electronic Trading

Today all brokers use computer systems to submit and manage orders. Algorithmic (algo) traders use them to make decisions. The reason for using computers is to minimize the cost, market impact and risk associated with trading large volumes of securities.

There are many different types of traders participating in the market at any given time. This naturally leads to traders that operate on different time scales. Some traders buy and hold stock for years as investors, some look to buy quotas of stock for multiple days, and some buy and sell rapidly at nanosecond timescales. All of these different traders simultaneously operate in the stock market, but each has different objectives and different time horizons. This is outside the scope of our work, but how these traders come together to form the dynamics we see in the stock market is an active field of study (see for example: [21–24]).

Participants place buy or sell orders to an exchange. These show the intention to buy or sell a given amount of a security for a specific price. Since all the available stock to be bought or sold is spread across many different prices there is no such thing as 'the' stock price. There is a limited amount that can be bought/sold at any given price, and traders are actively looking to buy/sell at the best possible prices.

The buy order with the highest price is called the best bid, while the sell order with the lowest price is called the best offer. The difference between the best bid and best offer is the spread. The mid price is defined as the average between the best bid and best

ask. These buy and sell orders accumulate in what is called the limit order book (LOB) until they find a counter-party for execution or are cancelled. We give the fictitious order book in Figure 1.3 as an example. We will use this example to illustrate how traders can interact with it and the effect they have. For the Figure 1.3 example LOB the best bid is 19.24, the best ask is 19.26, the spread is 0.02, and the mid price is 19.25.

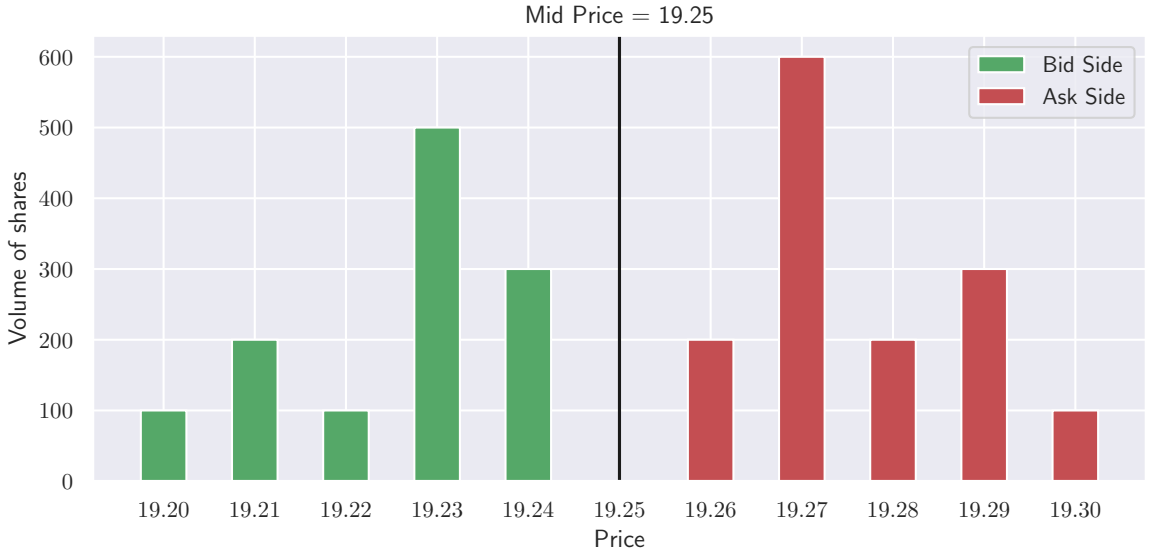


Figure 1.3: Example limit order book.

Dealer markets exist where dealers (individual people or firms) post a price at which they will buy and another at which they will sell a specific stock in order to make money from the spread. This is analogous to a foreign exchange kiosk – buying and selling currency at a different price in order to make money from the difference. Alternatively, one could buy or sell stock in an auction market (such as the Toronto or New York stock exchange) where there are many such people or firms posting prices in which they are willing to buy or sell stock. Typically, one refers to the auction market as the stock market. There exist brokers, known as 'market makers', which provide a significant amount of the new buy/sell orders to the market. These market makers act like the dealers in the dealer market, but for the auction market. These participants are looking to take advantage of the difference in price that stocks are bought and sold, called the spread, by constantly buying and selling without the purpose of holding onto the stock as an investment. In theory, the faster you can process information and react, the better your chances of profit. Some brokers pay a premium to an exchange for a faster connection which gives them a time advantage over other participants. For a discussion of broker behaviour, see Appendix A.

The increased speed in which brokers can interact with the LOB gives room for unfair or outright illegal behaviour by attempting to manipulate markets. This topic will be discussed later, in section 1.4, but is central to the goal of this work. Once the neccessary background has been established, we want to build a stock price dynamics model which can be influenced by a broker's (potentially) manipulative behaviour so that we can analyze the conditions under which it is profitable to employ such a strategy and what kind of payoff they can expect. A better understanding of how and why the limit order book would be manipulated would provide insights to regulators on how to better detect it and punish the offending party.

Market participants can interact with the LOB in three ways:

1. Limit orders: as buy/sell limit orders arrive they are added to the ask/bid side of the LOB. If a buy and sell limit order are placed at the same price they will be matched and executed.

2. Cancellation: traders may cancel their existing limit orders which are then removed from the LOB.

3. Market orders: a buy/sell market order will remove shares from the best available prices on the ask/bid side of the LOB until their total order is completed. Market orders may end up removing shares deep into the LOB at worse and worse prices. This is known as 'walking the book'.

The order book changes constantly as buy/sell orders arrive, buy/sell market orders arrive, and orders are cancelled. Depending on the stock these changes could be happening at time scales of a fraction of a nanosecond. The LOB also has a queuing system aspect. The first limit orders placed at a particular price will be executed first when matched to an opposite limit or market order - known as 'first in, first out'.[2] This 'battle' over queue position is an area of research in its own right [25].

The top panel in Figure 1.4 shows the outcome of the placement of a sell limit order of 100 shares at 19.28 with our example LOB from Figure 1.3. There are no limit orders to match, so the sell limit order is added to the existing orders at 19.28. However, this new order will be executed only after the existing limit orders since it is last in the queue. The bottom panel shows there are now 300 shares available to be bought at 19.28.

When an order is cancelled that volume of stock at that price is removed from the order book. The top panel of Figure 1.5 shows a trader cancelling their sell limit order of

[2]There can be a preference given to certain brokers so that their orders are executed ahead of others regardless of their position in the queue.

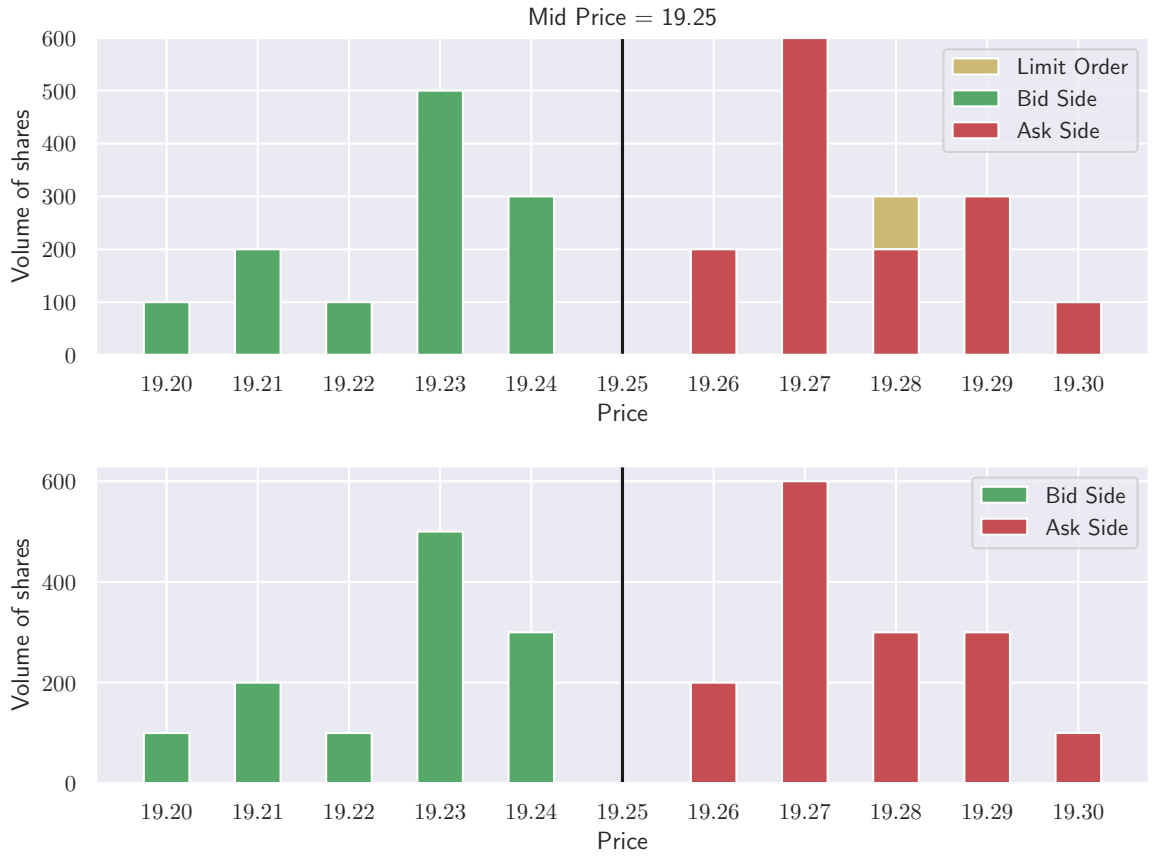


Figure 1.4: Sell limit order of 100 shares placed at 19.28 for our example limit order book. After the limit order is placed there are 300 total shares available to be bought for 19.28. The best bid and ask remain unchanged, so the spread and mid price are unaffected.

Figure 1.5: Cancellation of sell limit order of 200 shares for our example limit order book. Since this was the only sell limit order at 19.28 there will be no shares available to be bought for that price. The best bid and ask remain unchanged, so the spread and mid price are unaffected.

200 shares at price 19.28. The bottom panel of Figure 1.5 shows that, since there were no other sell limit orders at 19.28, there are no longer any available shares to be bought at this price. Traders may, at no cost, cancel their limit orders at any time while the market is open. Usually these limit orders are cancelled and placed again at a different price to reflect the changing attitudes of traders about where the stock prices will be moving.

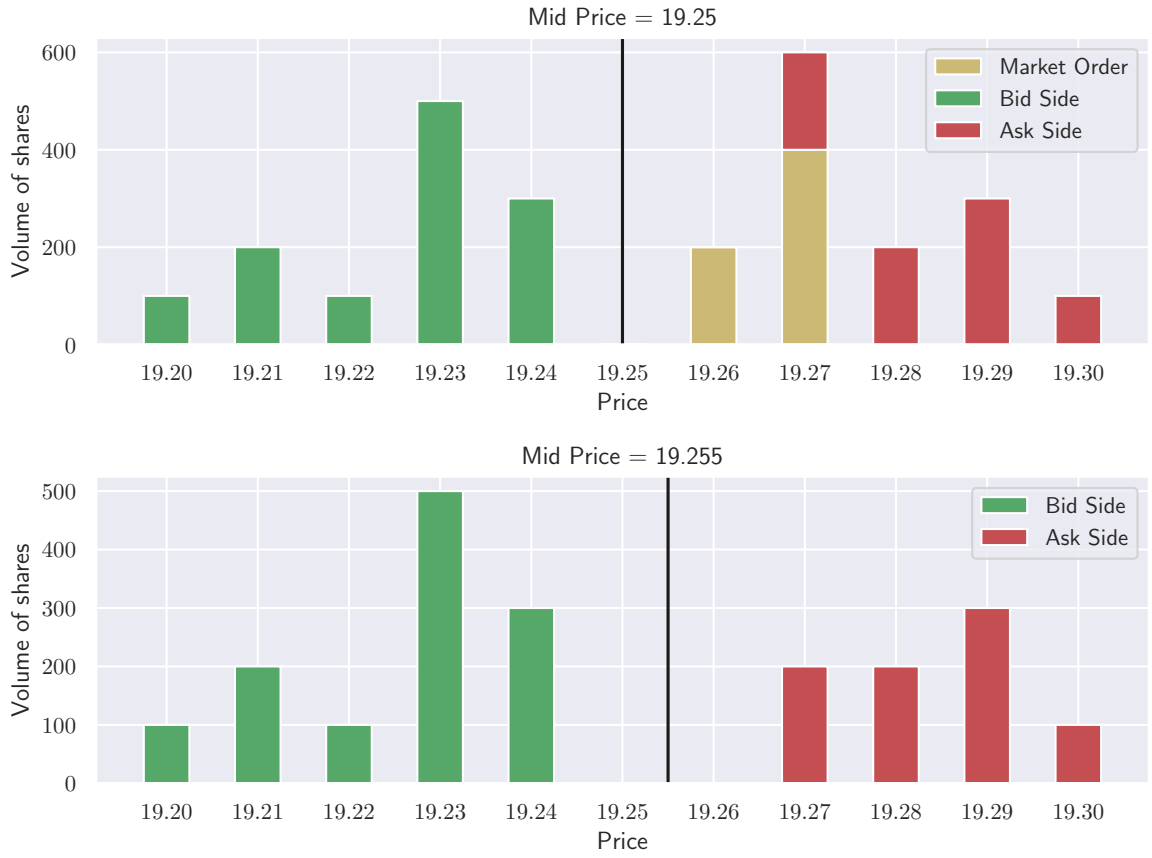


Figure 1.6: Buy market order of 600 shares is placed on our example limit order book. 600 shares are removed from the ask side of the book at the best available prices. The best bid is unchanged, but the best ask increases by 0.02. The spread and mid price increase to 0.03 and 19.255, respectively.

A market order will be matched to the best price and however much stock was ordered will be removed from the book. The top panel of Figure 1.6 shows a trader placing a buy market order of 600 shares. Since it is a buy market order it removes shares from the ask side of the LOB at the best available price until no shares remain at that price and then begins removing shares at the next available price. In this case 200 shares are bought for 19.26 per share and the remaining 400 are bought for 19.27 per share. The

second panel of Figure 1.6 then shows the best ask has moved up to 19.27 because of this market order. This example illustrates the idea of price impact on the LOB caused by market orders which remove large amounts of shares from the market. The price impact of a small enough order may only be temporary and new sell limit orders are placed back at 19.26 – returning the price back to where you started after a short recovery period. However, the price impact may also be permanent from a large order as prices move up and volume is not replaced with new limit orders below the best ask or the best bid increases to tighten the spread – causing the best ask to not return to where it started. Since we could not get all 600 shares at the best ask we had to 'walk the book' with our market order. This caused us to pay more per share than if we had been able to fill the whole order at the best ask. Traders will look to minimize this extra cost of walking the book by breaking large market orders up into smaller orders which they execute over some fixed time horizon.

The order and price dynamics of limit order books have been studied before in the context of general mixture models [26], Markov chains [27], Hawkes processes [28], and stochastic partial differential equations [29], to name a few. These models do not make use of the limit order book volumes directly and instead use continuous time stochastic processes for the prices themselves or to model the intensities and distributions of the different order types which drive a price process.

The study of optimal placement of limit orders is also of particular interest [8, 9, 30, 31]. These papers also explore whether or not trade-based manipulation can occur in their limit order book model, but they approximate the volumes in the book with a continuous shape function of the quoted bid and ask prices. For our model, we aim to maintain the discrete nature of the book in both prices and share volumes at each price.

1.3 Level 1 and 2 Data

Limit order book data is labeled as either level 1 or level 2 data. Level 1 order book data consists of just the best bid price, quantity at the best bid, best ask price, quantity at the best ask, the last traded price, and the last traded quantity. Level 1 data provides a surface level look at the trading activity of a particular stock ticker. Level 1 data is publicly available and easy to access, but to know more market details one must examine level 2 data.

Level 2 order book data requires a paid subscription to the exchange, but comprises considerably more information. The level 2 data consists of all the orders placed on the exchange as well as which broker is placing the orders. With the level 2 data one can see

exactly how many shares are available to be bought or sold at any given price and who would be the counter-party to the trade.

A sample of the level 2 data provided by TMX is shown in Table 1.1. This particular collection of data is for the stock AEM, but orders for all stocks are normally mixed together in the full dataset. The collected data includes all orders placed on the TSX over any given day with each row corresponding to a single order. The information provided in each row tells us everything we need to know about the order.

From Table 1.1 we see that columns are time, broker ID, side, book change, price, order id, seq, reason, other broker id, other order id, best bid, best offer, best bid size, best offer size, and market state. There is normally another column indicating the stock ticker for the order, but this was suppressed for readability. Definitions of each column are provided below.

time

Date and time stamp for order placement in nanoseconds.

broker id

Internal ID code for each broker with the exchange.

side

Side of the limit order book being interacted with.

book change

Change in number of shares.

price

Price placement of order

order id

Assigned ID code for the order

seq Internal sequence of received orders, included because two orders may have the same time stamp.

reason

Either 'TRADE' (matched limit order or market order), 'BOOKED' (limit order), or 'CANCELLED' (cancelled limit order).

other broker id

If reason is 'TRADE' then this is the broker id of the order on the other side of the trade.

other order id
: If reason is 'TRADE' then this is the order id of the order on the other side of the trade

best bid
: Current best bid price for the given stock ticker.

best offer
: Current best ask/offer price for the given stock ticker.

best bid size
: Current quantity of shares at the best bid price for the given stock ticker.

best offer size
: Current quantity of shares at the best ask/offer price for the given stock ticker.

market state
: Current state of market ('Open', 'Closed', 'Beginning of day', etc).

As one can see from Table 1.1, the limit order book needs to be generated from this much larger dataset of individual orders. For a single stock the number of orders placed on a given day can be in the hundreds of thousands resulting in raw datasets of several hundred megabytes for a single day. Generating the limit order book can therefore be a computationally expensive exercise. The dataset is far too large to interact with on a single personal computer, so we make use of an Amazon Web Service (AWS) account provided by TMX to do our analysis. See Section 1.5 for further details of the distributed data system. Sampling from the dataset does not help alleviate this issue since we need to know all orders placed up to time t to know what the limit order book looks like at time t.

One can reproduce the limit order book at a given time by simply summing the book changes at each price for each side of the limit order book up to that time. The result will be a collection of tuples (p, q) for each side of the book indicating the number of shares q at each price p. However, some prices p may be missing from the dataset if no one has placed a limit order at them. This is fine for display purposes like Figure 1.7, but when using the data in modelling we will need to know the quantity of shares sitting at every price p for both sides of the book.

Figure 1.7 displays an actual LOB for ABX stock on April 4, 2017 just after 3:00 PM that was built from data provided from the Toronto stock exchange.

	time	broker	side	book change	price	order id	seq	reason	other broker id	other order id	best bid	best offer	best bid size	best offer size	market state
1650	2017-04-17 09:31:24.055621321	1	Buy	-100	62.15	1—B20170417000000204	766150	TRADE	1	1—S20170417000000206	62.15	62.24	100	600	Open
1651	2017-04-17 09:31:24.055621327	2	Buy	-100	62.15	2—B20170417000000193	766151	TRADE	1	1—S20170417000000206	62.14	62.24	300	600	Open
1652	2017-04-17 09:31:24.055793062	2	Sell	100	62.28	2—S20170417000000197	766153	Booked	NaN	NaN	62.14	62.24	300	600	Open
1653	2017-04-17 09:31:24.055915806	2	Buy	-100	62.14	2—B20170417000000192	766154	CANCELLED	NaN	NaN	62.14	62.24	200	600	Open
1654	2017-04-17 09:31:24.055935694	2	Buy	-100	62.13	2—B20170417000000191	766156	CANCELLED	NaN	NaN	62.14	62.24	200	600	Open
1655	2017-04-17 09:31:24.056082876	2	Buy	100	62.14	2—B20170417000000198	766157	Booked	NaN	NaN	62.14	62.24	300	600	Open
1656	2017-04-17 09:31:24.056089293	2	Buy	100	62.13	2—B20170417000000199	766159	Booked	NaN	NaN	62.14	62.24	300	600	Open
1657	2017-04-17 09:31:24.056901174	3	Buy	-200	62.14	3—B20170417000000003	766161	CANCELLED	NaN	NaN	62.14	62.24	100	600	Open
1658	2017-04-17 09:31:24.060729084	1	Sell	-100	62.24	1—S20170417000000199	766169	CANCELLED	NaN	NaN	62.14	62.24	100	500	Open
1659	2017-04-17 09:31:24.060735133	1	Sell	-100	62.26	1—S20170417000000189	766171	CANCELLED	NaN	NaN	62.14	62.24	100	500	Open
1660	2017-04-17 09:31:24.060737312	1	Sell	-100	62.33	1—S20170417000000038	766172	CANCELLED	NaN	NaN	62.14	62.24	100	500	Open
1661	2017-04-17 09:31:24.061368993	2	Sell	-100	62.28	2—S20170417000000197	766173	CANCELLED	NaN	NaN	62.14	62.24	100	500	Open
1662	2017-04-17 09:31:24.066693881	1	Buy	100	62.15	1—B20170417000000207	766194	Booked	NaN	NaN	62.15	62.24	100	500	Open
1663	2017-04-17 09:31:24.066844825	1	Buy	-100	62.13	1—B20170417000000198	766196	CANCELLED	NaN	NaN	62.15	62.24	100	500	Open
1664	2017-04-17 09:31:24.070886384	4	Buy	-100	62.13	4—B20170417000000045	766202	CANCELLED	NaN	NaN	62.15	62.24	100	500	Open
1665	2017-04-17 09:31:24.070886387	4	Buy	100	62.14	4—B20170417000000049	766203	Booked	NaN	NaN	62.15	62.24	100	500	Open
1666	2017-04-17 09:31:24.086019171	1	Sell	100	62.24	1—S20170417000000208	766223	Booked	NaN	NaN	62.15	62.24	100	600	Open
1667	2017-04-17 09:31:26.282657689	4	Buy	-100	62.14	4—B20170417000000049	769200	CANCELLED	NaN	NaN	62.15	62.24	100	600	Open

Table 1.1: Sample of market data used to reconstruct the limit order book. Each line corresponds to a single order and provides all of the details of that order. Orders for different stock tickers are mixed together in this data, but the data provided in the above table has been pulled for the stock AEM only. Broker ID has been anonymised in the 'broker' and 'order id' columns for privacy.

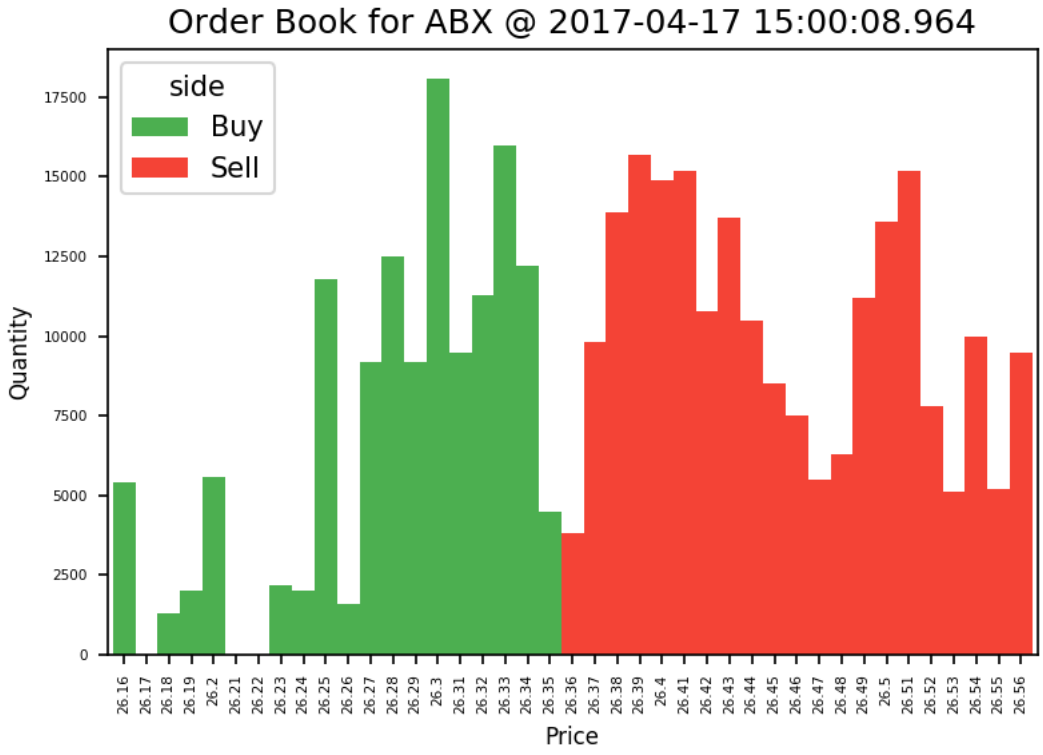


Figure 1.7: Order book for American Barrick (ABX) stock.

1.4 Spoofing

In Section 1.2 we covered the various ways traders can interact in the stock market through limit, market, and cancellation orders. Although traders are assumed to be placing limit orders in good faith, they may actually be trying to manipulate the order book in their favour by a process called 'spoofing' or 'layering'. Spoofing and layering are terms often, but not always, used interchangeably to refer to the act of placing an ordered sequence of buy/sell orders with no intention of allowing them to be executed. Spoofing can also refer to placing a single order with no intent to execute as opposed to layering which is placing multiple orders instead of one [32]. The Financial Industry Regulatory Authority, Inc. (FINRA)[3] instead refers to spoofing as entering limit orders on one side of the book with the intent of moving the market for a beneficial execution on the opposite side of the book. However, FINRA refers to layering as enticing other market participants to join the same side of the book as your spoofing limit order and trading against them favourably [33].

The reason to place these limit orders – which you have no intention of executing

[3]FINRA is a private corporation and the largest independent regulator in the United States.

– is so other traders will see the increased liquidity in the book and react believing (mistakenly) these orders were placed in good faith. That is, a trader may place their own limit orders or market orders under the assumption all other trader's orders are there to be bought or sold. In essence, the spoofer is tricking other traders into making decisions they might otherwise not do and then attempt to capitalize on the consequences of these resulting trades.

To illustrate how this works suppose our spoofer, Alice, wishes to purchase 200 shares of stock. As shown in Figure 1.8, she could place a market order to the exchange and collect the shares now.

Figure 1.8: Alice places a market order to buy her 200 shares. She gets her shares now, but with no improved price.

Instead, what if Alice attempted to use spoofing to lower the best ask price to get her 200 shares for less money? Figure 1.9 shows the following example of a positive payoff. Alice places a sell limit order of 400 shares at $19.28 – two ticks above the best ask price so other traders believe there is a huge demand to sell shares at the current best ask price. This might lead another trader, Bob, to believe that this increased liquidity to sell

is because other traders believe the stock price is going to drop and they want to sell their shares now before it does. Bob wants to sell his shares now before the price drops so he places a sell limit order at \$19.25 – one tick below the best ask price so Bob can have his limit order executed ahead of everyone. Other traders like Bob follow suit and the best bid drops to \$19.25 with 400 total shares available to be purchased. Alice sees that her plan has worked and places her market order to buy 200 shares at \$19.25 and cancels her limit order at \$19.28. Alice has now made a small profit over having just placed a market order like Figure 1.8. Since this would all be happening at nanosecond time scales Alice is able to place spoofing orders and cancel them at extremely high speeds making use of computer algorithms.

However, the price could move against Alice as shown in Figure 1.10. In Figure 1.10, Alice places the same sell limit order of 200 shares at \$19.28, but now Bob and the other traders place market orders which remove 500 shares from the ask side of the order book. This may be because they see a nice wall of stock at \$19.28 and believe their market orders will not go too deep in the book. This causes the best ask to increase to \$19.27 when Alice places her market order and cancels her limit order. Unfortunately, Alice has now paid more for her 200 shares than had she just placed her original market order as in Figure 1.8.

Things can get even worse for Alice. In Figure 1.11 Alice places her limit order of 200 shares too close to the best ask. Like before, the best ask price increases because of the market orders from Bob and other traders which causes Alice's sell limit order to be executed. Now Alice needs 200 shares to cover her executed limit order on top of the 200 shares she originally wanted to purchase. Alice places a market order of 400 shares to cover this and ends up paying significantly more than if she had just placed her market order as in Figure 1.8. From this example it should be obvious how dangerous it is for Alice to place her spoofing orders close to the touch[4] – such orders are more likely to be executed by market orders which require her to buy them back at a higher price. No trader wants to buy high and sell low.

The strategy for the spoofer is then to determine how large their limit orders need to be and at what price to place them in order to manipulate the price in their favour while also placing the limit orders deep enough into the order book so that they will not be executed.

Clearly this is a conveniently devised example and this strategy may not always work out in Alice's favour. The increased spread could lead to more buy limit orders which increase the best bid instead and Alice is back where she started. Spoofing is not

[4]The 'touch' refers to the best bid and/or the best ask, simultaneously.

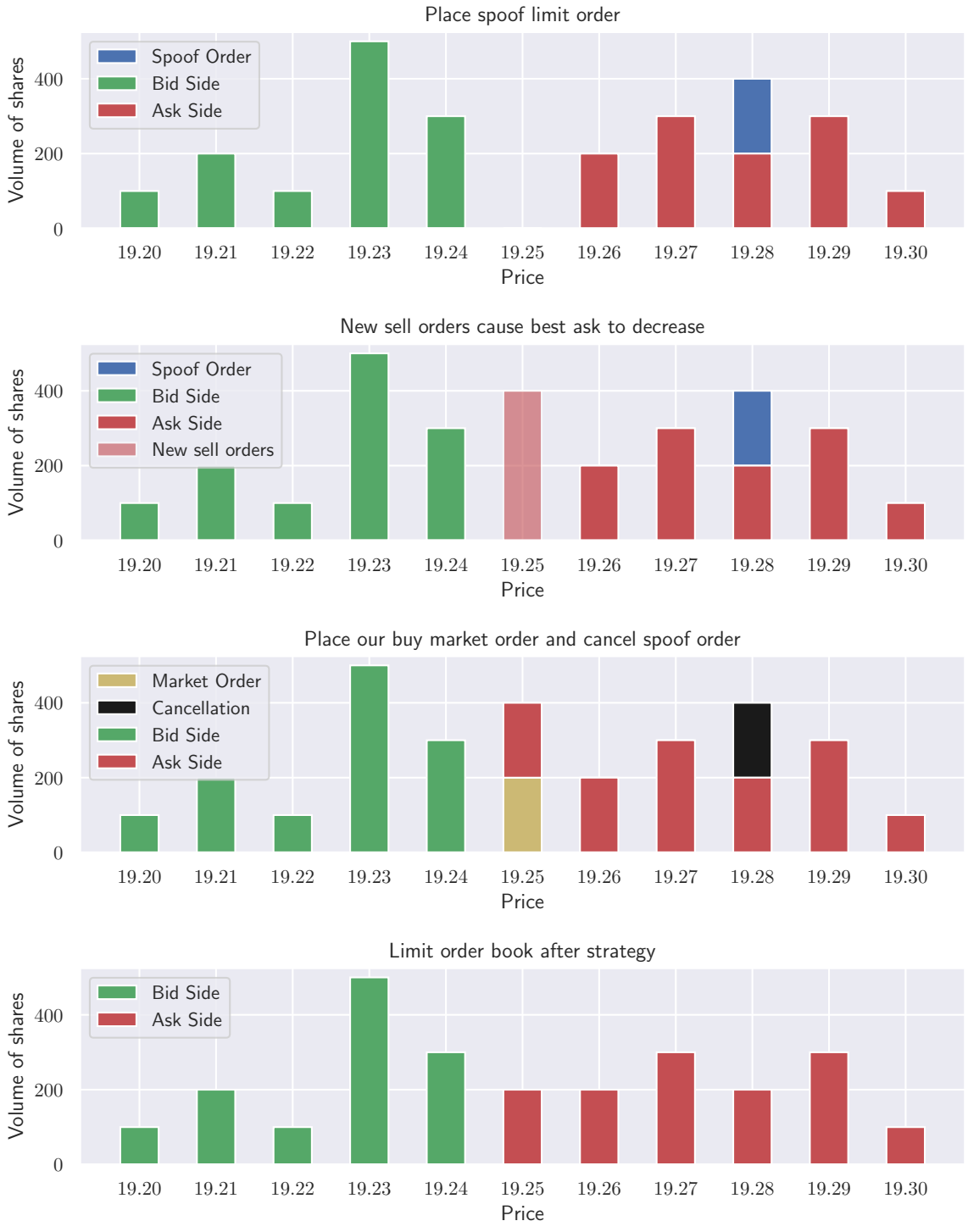


Figure 1.9: Case where spoofing has positive payoff. Alice places a spoofing sell limit order of 200 shares at \$19.28. Bob and other traders enter sell limit orders totaling 400 shares at \$19.25 – 1 tick below the best ask. The best ask has dropped so Alice cancels her spoof order and places her market order of 200 shares which are lifted at \$19.25. The final result is Alice has acquired her 200 shares, saving \$0.01 per share.

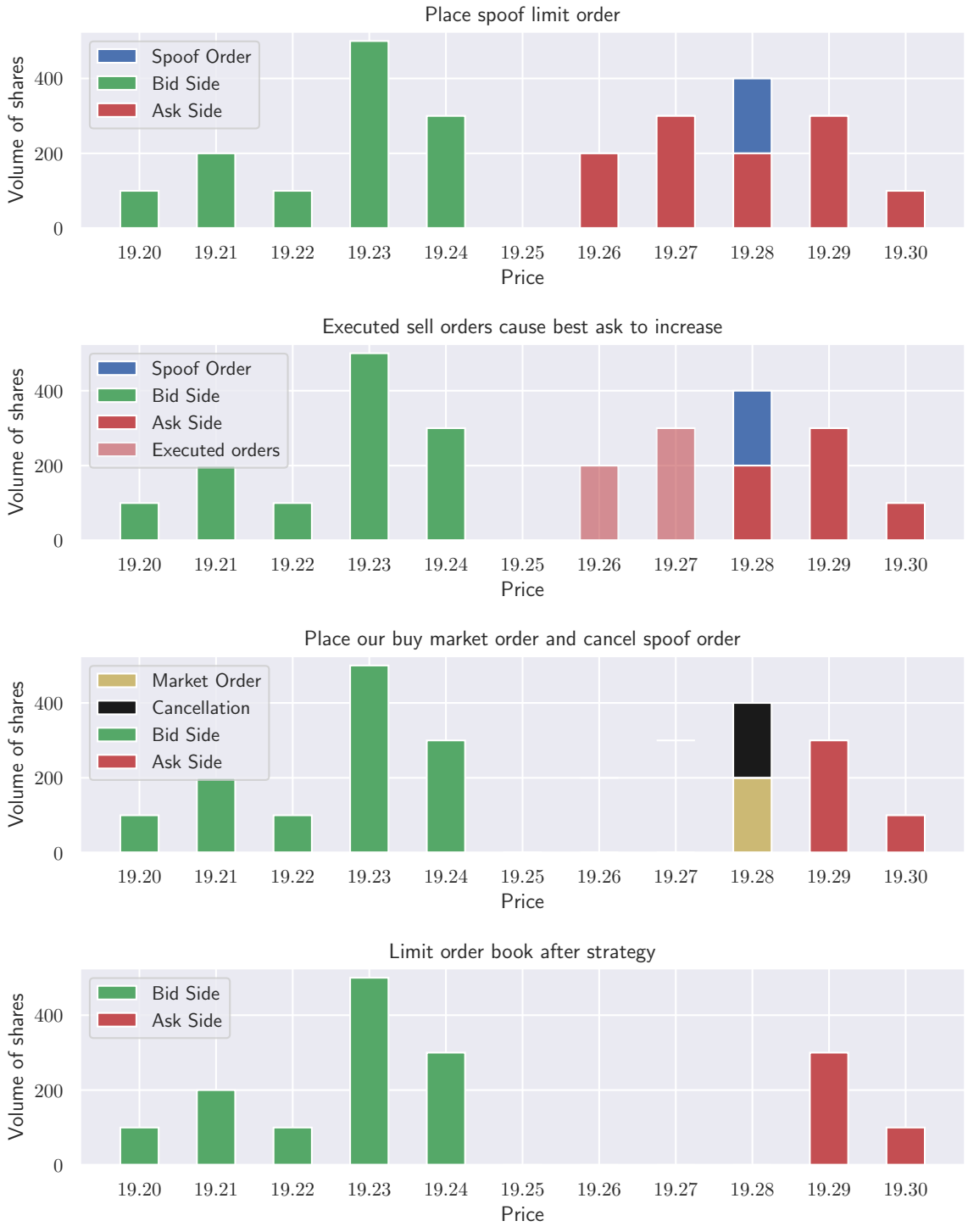


Figure 1.10: Case where spoofing has negative payoff. Alice places a spoofing sell limit order of 200 shares at \$19.28. Bob and the other traders place market orders which remove 500 shares from the ask side of the order book. This causes the best ask to increase to \$19.27 when Alice cancels her spoofing limit order and places her market order of 200 shares which are lifted at \$19.28. Alice has still acquired her 200 shares, but she spent \$0.02 per share more than she needed to.

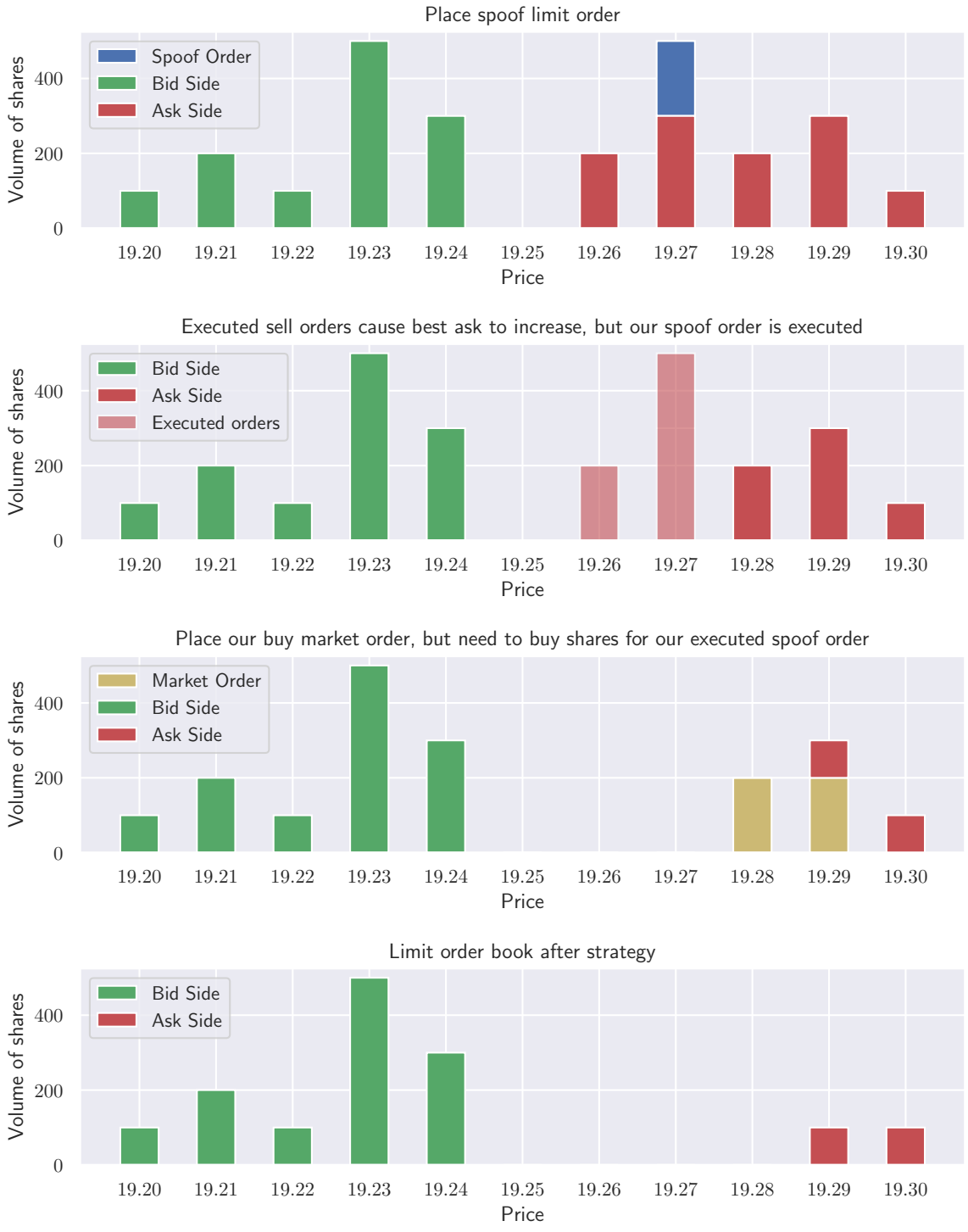


Figure 1.11: Worst case scenario for spoofing. Alice places a spoofing sell limit order of 200 shares at \$19.27. Bob and the other traders place market orders which remove 700 shares from the ask side of the order book. Alice's sell limit order is executed before she can cancel it – now she must buy her 200 shares plus another 200 from the executed trade. Alice places a market order for 400 shares which walk the book to \$19.29. Alice has paid more for her original 200 shares than she needed to, but she also had to buy shares back at a higher price than she sold them. The executed spoofing order cuts further into Alice's profits.

guaranteed to work for the spoofer – whose goal is to increase the odds of favourable price movements and capitalize when they happen. In what follows we provide a list of documented instances of spoofing. We see that spoofers are doing this over years on specific stocks, futures, options, or commodities at high frequencies.

Stocks with large spreads and volatility would be obvious targets for spoofing orders as their impact would be larger [34, 35]. The large spread would allow traders to place limit orders below the best ask more easily – this is exactly the desire of a spoofer wishing to buy shares. Very tight spreads would mean that shares need to be lifted from the best bid before the best ask can drop. This can obviously happen, but the steps needed in depleting the volume at the best bid (through cancellations or sell market orders) is more involved than simply placing a single limit order to lower the best ask.

In the wake of the 2008 financial crisis the United States passed the (2010) Dodd-Frank act to overhaul regulations on the financial industry. It became "against the law to spoof, or post requests to buy or sell futures, stocks and other products in financial markets without intending to actually follow through on those orders." [36].

In 2014, the US Commodity Futures Trading Commission (CFTC) told the Chicago Mercantile Exchange & Chicago Board of Trade (CME Group Inc.) to continue developing spoofing detection [37]. Since then a number of individuals and groups have been accused and sentenced for spoofing. That same year Michael Coscia was arrested for spoofing commodity futures markets in the first application of the relevant sections of the Dodd-Frank Act and was found guilty in 2015 [38]. Each spoofing count brings a 10 year jail sentence plus $1 million in fines. Coscia was sentenced to 3 years in prison for his crimes. In 2015, Navinder Singh Sarao of the UK was arrested for contributing to the 2010 flash crash – he pleaded guilty in 2016 [39]. A Canadian, Aleksandr Milrud, pleaded guilty to US Federal prosecutors for spoofing to earn profits of $1.9 million [40].

Staff of the Ontario Securities Commission (OSC) settled their first spoofing case in 2015 against Oasis World Trading [41]. In this case 0.14% of Oasis' total trades and 0.04% of their trade volume was found to be spoofing orders. It was a small number relative to their total trades, but "a very high proportion" of the total orders and volume of stocks targeted by Oasis. That is, very little of Oasis' trading activity was manipulative on the market except for the stocks they targeted for spoofing where they made up a significant amount of the orders and volume on the book for that stock. This may imply it is easier to detect manipulative practices on an individual stock rather than looking at the aggregate of orders by a trader.

The OSC has also made allegations in 2018 against K2 Investment Fund for spoofing options on the Montreal Exchange [42]. Fines and bans were placed on K2 and its upper

management after the investigation. That same year three major financial institutions, Deutsche Bank, UBS, and HSBC, paid $47 million collectively for manipulating futures markets [43]. In 2019 a former Bear Stearns and Bank of Nova Scotia trader, Corey Flaum, pleaded guilty for spoofing precious metal markets [44] and JPMorgan Chase has also been accused by US Federal prosecutors of spoofing precious metals [45].

The problem facing prosecutors when charging traders with spoofing is proving intent. A former UBS previous metals trader, Andre Flotron, was acquitted by a jury in Connecticut in 2018 on spoofing charges [46]. The main issue is that most orders on individual stocks go unfilled (see Appendix A, for example) and canceling orders is not illegal, so its hard to classify an order as spoofing [34]. What worked in Andre Flotron's favour was that the two traders that testified against him recieved non-prosecution agreements which called into question their trustworthiness, he did not make use of computer algorithms and allowed orders to stay on the book for up to a minute, and his victims were other large financial firms that would not come across as sympathetic victims. After only a few hours of deliberation, Mr. Flotron was found not guilty. Detecting potentially illegal orders does not prove a trader's intent to manipulate markets, but methods for detection are needed to even bring these cases to the courts. Moreover, sophisticated detection methods may lead to stronger evidence that suspicious orders were placed in bad faith.

All of these cases have come from increased scrutiny of the financial markets. As the above list of incidents makes clear – spoofing is not limited to just stocks, but has been detected in futures, options, and commodity markets. Financial regulators have increased their efforts to detect manipulation and academic research has expanded into this territory as well in recent years. Machine learning techniques have been applied to detecting anomalies in limit order book timeseries [47, 48]. Also, we previously discussed the work in trade-based manipulation for a limit order book model [8, 9], but research into modelling limit order book spoofing still remains 'scant' [49].

Lee et al. [34] take an empirical approach studying the impacts of spoofing on the Korean Exchange (KRX) using intraday trading data taken from November 2001 - February 2002. This time period was chosen because up to the end of 2001 the KRX disclosed the volume of shares on both sides of the limit order book in addition to the prices and volumes 5 ticks from the best ask and best bid. At the start of 2002 the KRX, in an attempt to stop price manipulation, stopped disclosing the total volume on both sides of the book, but increased the visible portion of the book from 5 to 10 ticks from the best ask and best bid. The authors define a spoofing order as a limit order at least twice the size of the previous day's average order size placed at least 6 ticks away from the best ask/bid, followed by a market order on the opposite side of the book and cancellation

of the spoofing order. The author's goal was to test, empirically, if there was a statistically significant decrease in spoofing orders after the information disclosure change on the KRX. They show that during the initial 2 month period the average spoofing order is 5.6 times larger than the typical limit order and almost all spoofing orders are placed more than 10 ticks away from the touch. The spoofer places an extremely large order outside the disclosed portion of the book (so far from the touch that it is very unlikely to be executed) so that other traders cannot see where it is placed, but they can see the spoofing order's impact on the total volume posted on that side of the book. This gives other traders the impression of an imbalance between the number of shares available to buy and sell and moves the price and achieving substantial net profits of 67 - 83 basis points[5] over the course of 45 minutes. They also show that spoofers target stocks with higher return volatility, lower market capitalization, lower price level, and lower managerial transparency, and trading volume is not a significant determinant of targeting by spoofers. The opening and closing of the market was also the most active spoofing period. In addition, 96% of the spoofing orders are placed by individuals instead of institutional investors which they argue is likely because: 1) a large firm would need to spoof considerably to earn profits that interest them which would raise flags to regulators, and 2) it is not likely for compliance departments within the firms to allow the use of potentially illegal trading strategies. When the information disclosure change came into effect on January 2, 2002, the proportion of spoofing orders located 11 ticks and below decreased from 89% to 40% – succeeding in reducing the total number of spoofing orders on the KRX.

Wang [35] employs a similar empirical analysis of spoofing in the Taiwan's index futures market (TAIFEX). TAIFEX discloses the first 5 ticks on each side of the book, but not the total volume beyond. For this reason they use a different definition of spoofing – any order within the disclosed 5 ticks with a size larger than the prior volume on the 5th tick which is followed by a market order on the opposite side of the book, and then a cancellation of the spoofing order. They find spoofing is associated with all types of traders, but mostly with individuals, and that spoofing is likely to occur when volume, volatility, and prices are high. Like Lee et al. [34] they find that spoofing is most common during the first and last hour of the trading day and is profitable in the majority of cases. Wang also argues that spoofing destabilizes the market by increasing trade volume along with the spread and volatility.

Cartea et al. [49] take a mathematical modelling approach to spoofing a limit order book. They continue the idea in Lee et al. [34] that volume imbalance in the book

[5] A percent of a percent – 0.01%

impacts price movements which a spoofer can take advantage. In their model they take on the role of a spoofer that wishes to liquidate a collection of shares using spoofing orders at the best bid to incite traders to execute the spoofer's sell limit order at the best ask. This is favourable to the spoofer over liquidating their shares by selling at the best bid as the spoofer will earn the spread in profit. They use a Markov chain to model the volume imbalance which moves the market between buy heavy, neutral, and sell heavy regimes. Intensities of the various order types during each regime are calibrated using Nasdaq high frequency data for a collection of individual stocks. The spoofer can then influence which regime the market is in using their spoofing orders. The stock's price dynamics are then dictated by the different order intensities which change with the imbalance regime. They also incorporate into their model the possibility of their spoofing limit orders being executed causing them to purchase shares which would add to the inventory they need to liquidate. The authors then formulate the maximization of their profit over a finite time horizon as a dynamic programming problem. This problem can then be converted into a partial differential equation by standard methods and solved to yield the optimal spoofing strategy. Cartea et al. [49] find that spoofing increases the performance of their execution strategy and considerable profit can be made. They argue spoofing strategies are detrimental to the integrity of markets and their model can be employed to understand these strategies better in order regulators to combat them.

In chapter 2 we detail the volume imbalance employed by Cartea et al. [49] and use a generalized form of this statistic in our price manipulation model covered in chapter 3.

1.5 Distributed Data

We briefly touched on some aspects of the data system earlier in this chapter. We illustrated the type of data in Table 1.1, we described the large size of the raw datasets, and sketched some of the tools we need to use to work with them. In this section we will provide more detail about the issues working with very large and complicated data sets. This section will be particularly useful for readers without a background in big data.

All data was hosted on an Amazon Web Service (AWS) cluster with the ability to interact with the data by executing scripts in an Apache Zeppelin notebook. Apache Zeppelin is a front-end for the cluster which allows one to execute code from various back-end languages such as R and Python. Apache Zeppelin is also fully integrated with Apache Spark - an open-source distributed general-purpose cluster-computing framework. With Apache Spark and Zeppelin we can create whatever data sets we want, regardless of size, and work with them as if they were any small data set loaded into R or Python's

Pandas, for example. The advantages are that any calculations done on the data are done in parallel across the cluster's nodes and we can load data sets into memory that would normally crash an ordinary personal computer.

The cluster itself starts with a collection of tables like Table 1.1 which are accessed through 'Structured Query Language' (SQL) queries. Since this is a distributed framework we need to pass our SQL queries through Presto to load the massive data sets across many executor nodes which are all passed instructions from the driver node. Presto is a SQL query engine for big data which allows for easy data visualization and investigation. However, the generated data frame from the Presto SQL query cannot be passed on for deeper analysis - say optimization or statistical tests. Instead, we pass our SQL queries to Pyspark - an application programming interface (API) which exposes the Spark programming model to Python. With Pyspark we can use our SQL queries to generate Pyspark dataframes that can be manipulated with Python-esque scripts.

Anyone with a background in Python's data analysis library (Pandas) will find it, syntactically, very similar to Pyspark. In Pyspark the the data manipulation is passed to the executor nodes and the results are passed back to the driver and collected. This can be very time consuming if the dataset is large. We are also limited in what we can do with Pyspark on the dataframe. The built in functions are for handling simple statistics, groupings, windows, etc, but progress is being made on integration with machine learning packages. However, we can use user-defined functions (UDFs) to do more complicated operations by row or column in the dataframe. Again, this can be very time consuming based on the dataframe size and how complicated the UDF is. To avoid this problem we can take advantage of Pyspark's ability to convert Pyspark dataframes into Pandas dataframes so we have access to all Python libraries to carry out any analysis we want. For example, the first optimization algorithm we tried would take over 2 hours to run using UDFs and less than 5 minutes if converted to a Pandas dataframe. The drawback is that if the data set is too large we risk crashing the entire cluster when we read it into memory on the driver. This is because it was originally spread out across many nodes.

Size issues with this conversion can mostly be remedied by careful use of assigning data types to each column. For example, we could convert a column with only a few different entries into categorical variables. The 'Side' column in the data set stores the string 'Buy' or 'Sell' which is expensive in a dataframe of several hundred thousand rows. Since there are only two possible values in each row we can convert all of these strings to a categorical variables 0 and 1 by 'Buy' $\rightarrow$ 0 and 'Sell' $\rightarrow$ 1. This effectively turns each multi-byte entry into a single bit. Many columns can also be efficiently stored as integers instead of floating point numbers. We can even convert our timestamps to

integers instead of strings by using unix time (the number of seconds that has passed since midnight UTC on Jan. 1, 1970) and splitting nanosecond portion of the time stamp into a separate column – also converted to an integer. These easy steps can reduce the size of our data frame by as much as 75%.

We have outlined the process we use to generate the data sets needed for our analysis in later chapters. With that in mind, in the following subsection we can discuss some examples of problems that arose from this project that are typical for data science projects and the solutions used to fix them.

1.5.1 Examples of Issues Arising in Data Science

Example 1: Missing data One of the most common problems in data science is missing data. This could be, for example, missing rows in a time series where data was not collected or missing columns where only partial data was collected. The process of cleaning data sets is often the most time consuming part of data analysis and in industry there are usually entire groups dedicated to making sure data is complete and consistent. Our problem with encountering missing data is that we cannot modify or control how TMX collects and designs its data sets. Also, since we need to pass SQL queries to the cluster to generate our data sets we cannot know if we have missing data until after it has been collected. It would be impractical to read the entire data set of all trading data across all stocks for years into memory - even distributed over a large cluster. We need to generalize the code for our analysis enough that we can modify it easily when we encounter these problems without actually knowing when or how they will appear.

An example of working with missing data happened in our work when generating our data set for the limit order book on June 15, 2017 for all stocks we analyzed. The problem was that there were multiple orders, with time stamps after the trading day had concluded, that had a 'not a number' (NaN) stored for the price the order was placed at. These orders had no reason to be there and it makes no sense to have no price listed for the orders anyway. When attempting to store these NaN values in the data frame the code would crash because our price column was set to unsigned integer data types (after converting dollars to pennies). This problem was never encountered for any other day. Its likely these entries in the order book table were present because of internal TMX bookkeeping. The NaN values themselves did not raise suspicion because TMX uses NaN's to fill in entries where there is no value – for example, the 'other broker id' column in Table 1.1 for 'Booked' or 'CANCELLED' orders. While TMX is using 'NaN' as a placeholder entry in their data sets this is against the Institute of Electrical

and Electronics Engineers (IEEE) standard of using 'NaN' specifically for undefined or unrepresentable values such as $\log(0)$ or $\frac{1}{0}$, for example.

Its important to have a solid knowledge of what your data is and how you are using it in order to solve these kinds of problems when they show up. Since these orders with missing prices appear after the regular trading time we can just drop these entries from the data set - we do not need trade data after the market closes. Knowing the hours of when the TSX is active presents this as a solution to our problem. If we knew nothing about our data we would not expect to be able to throw problematic entries away nearly so easily.

Example 2: Breaking existing data structure It is frequently helpful to build new code from an existing code base. For us, our initial code to produce aggregate limit order book data was written by TMX and gave us a skeleton code to modify for producing what we needed. Our own code base for our order book analysis was entirely built from these initial few scripts and never presented a problem until our code would crash for different stocks on different days with a generic error message attempting to create a data frame from empty arrays.

In the early hours of the morning, TMX updates the order book for the day by rolling over any orders that were left from the previous trading day. These orders come in one at a time at the same time every day. The sell orders appear first until they have all been added to the book and then the buy orders are added. The initial code by TMX would begin generating the limit order book a few seconds after this process would start. There are often very few orders which roll over and both buy and sell orders have filled the book by the time the code starts preparing data frames to fill with this data. However, on some days with some stocks the sell orders had not finished rolling in by the time the data frame tries to collect both sides of the book. It would find an empty buy side and return errors - crashing the code.

This was never an issue for the original purpose of the code, but became an issue for us when we modified it for our purposes. Again, like the first example, our knowledge of how the data is organized gave us an easy solution - just start building and storing the limit order book shortly before the regular trading time starts. All orders would have rolled over hours before this happens. Like the previous example this seems like an easy and obvious solution, but in data science we are often not so lucky to have the problematic data entries appearing in parts of the data set we would be removing anyway.

Example 3: Distributing data by incorrect columns When creating a distributed data frame you need to select what columns of the data frame to partition over the different nodes. A natural choice would be to partition by time, date, side of the order book, or stock symbol. For time series outside of finance, say daily rain or snow fall, there is usually never a problem with time not being unique to each data entry. However, in the age of electronic algorithmic trading we cannot precisely give a unique time stamp to an order even down to nano-second timescales. Our problem then arose by partitioning over time of day and date since for some entries the time was not unique. When the data set was collected back to the driver node the data set was completely out of order with column entries from different rows being swapped. Since time is not unique in the data set we also cannot just reorder it afterwards. No errors would be produced and the distribution of the change in the best ask would not look out of the ordinary, so you would never know there was an issue until you plotted, say, the best ask over the trading day. Always check that your results are consistent when working with data. You may never know you are using incorrectly generated data if you do not have safety checks set up throughout your analysis.

Conveniently, TMX already thought of this problem and had a solution built into the data structure. Looking back to Table 1.1 we have the 'seq' column which tracks the order that all orders arrive to the exchange. Partitioning and ordering by the 'seq' column solves our issues with the distributed data frame and sorting in time. If time is not guaranteed to be unique for your data you always need a way to know how the rows are ordered. The growing theme across these examples is that having a solid knowledge of what your data is and how you are going to use it allows you to handle potential problems.

Example 4: Outliers Another thing that can happen with time series data is that you can encounter specific days or time periods where you get vastly different results than usual. Ordinarily, you might expect you simply made a mistake in your code or your model does not present the consistent results you were expecting. However, with financial data there can be alternative explanations for what you see.

In chapter 4 we investigate the calibration of the models we present in chapter 3 and find unusual results on the same dates across multiple stocks. In this example, the dates are May 29, 2017 and July 4, 2017. If we were unable to extract exact dates (or were never provided dates) we would never know these days correspond to holidays in the United States where their own exchanges and financial institutions are closed. Our unusual results are caused by significantly less trading activity because the American

traders are on holiday when the Canadian traders are still active. We also find similarly strange results for specific stocks which corresponded to dates surrounding their record dividend date. This is the date in which the owner of the share is guaranteed to receive a dividend payment usually a month later.

Since our data is properly labeled and timestamped these kinds of issues are possible to solve with a little extra work. In our case these outlier points were not a bug, but a feature of the model – we are able to detect unusual days in the time series from our model parameters.

1.6 A Look Ahead

Price manipulation has existed in financial markets for a long time, but the introduction of computer systems and the move towards high frequency trading (HFT) has introduced new ways to manipulate. Like the pump and dump strategies of the past, spoofing is a new form of price manipulation through introducing misleading information into the market. The literature on detecting spoofing is relatively small, so it is important to develop mathematical models which can combat these strategies.

So far we have covered how stocks are traded electronically through the limit order book, how this data is stored, and how a spoofer could manipulate prices in the limit order book. In recent years there has been a considerable up-tick in prosecution of spoofers, but research needs to be done on better and more efficient detection methods to catch these people before they do damage to financial markets – either through monetary losses or distrust in the system. Our goal over the course of this work is to build a model which can be used to explore the costs associated with spoofing that we saw in Figures 1.8, 1.9, 1.10, and 1.11 in order to determine when and how a spoofer can best profit from the state of a limit order book. From this model we can develop tools to explore when a limit order book is most vulnerable to spoofing and give regulators a way to narrow their search for price manipulation caused by spoofing.

In the following chapters we build up our model for spoofing order detection. When we talk about price change distributions we are naturally determining a time scale associated with those changes. In chapter 2 we discuss the implications and complications related to time scales in the limit order book. Information about volume imbalance has been a common component in the study of spoofing. We continue with this idea by condensing the limit order book into a single variable, the volume imbalance ratio, which we analyze with statistical tests for its predictive power in price changes. These statistical tests will be an important tool for us to compare models later this work.

In chapter 3 we introduce our mathematical notation for the limit order book and derive the equations needed to quantify the payoffs associated with the actions we saw in Figures 1.8, 1.9, 1.10, and 1.11. We then develop a model for the price change distribution dependent on the volume imbalance ratio before giving the final optimization problem for the spoofer on where to place their limit orders for maximum impact. The ultimate result is a starting point to look for manipulation in the book.

In chapter 4 we discuss the calibration of our model to the limit order book data we presented in this chapter. The resulting model parameters are then analyzed against each other as well as with data from the market activity itself. We also return to the statistical tests conducted in chapter 2 to see which stocks found improvement under the calibrated model parameters. A goodness of fit of our price change model is also presented.

In chapter 5 we apply our spoofing cost model to multiple stocks to determine when spoofing is the optimal decision for purchasing a fixed number of shares. This will provide an idea of when our model would predict the book is sensitive to price manipulation. We also explore including risk in the decision making process as spoofing may be profitable, but the risk of your spoofing limit orders being executed against you may make spoofing undesirable. Incorporating risk into the decision making provides excellent clusters which allows us to study changes to the decision boundary as we change the spoofer's constraints: the volume of shares they are willing to spoof with, and the number of shares they wish to purchase. We also study the optimal spoofing strategy of a specific stock and find the shape of book plays a key role in the spoofer's decisions.

Finally, in chapter 6, we end with a summary of the work and possible directions and extensions we can take to improve our model.

Figure 1.12 shows a breakdown of the key concepts covered in each chapter as well as how these concepts are connected. We use arrows between sections to highlight how we will be returning to specific ideas in later chapters once we have developed the machinery necessary to discuss these topics. For example, in chapter 2 we define the volume imbalance ratio and conducted statistical tests to show its association with future price movements. We generalize this definition in chapter 3 and use this definition to calibrate our model in chapter 4. Afterwards we return to the statistical tests in chapter 2 using our new model parameters to show improvements in the model over the classical definition of the imbalance. We also develop our price change model in chapter 3 and analyze the goodness of fit in chapter 4. At least, the results from our model calibration and generalized imbalance weights in chapter 4 are fed into our analysis of spoofing detection in chapter 5.

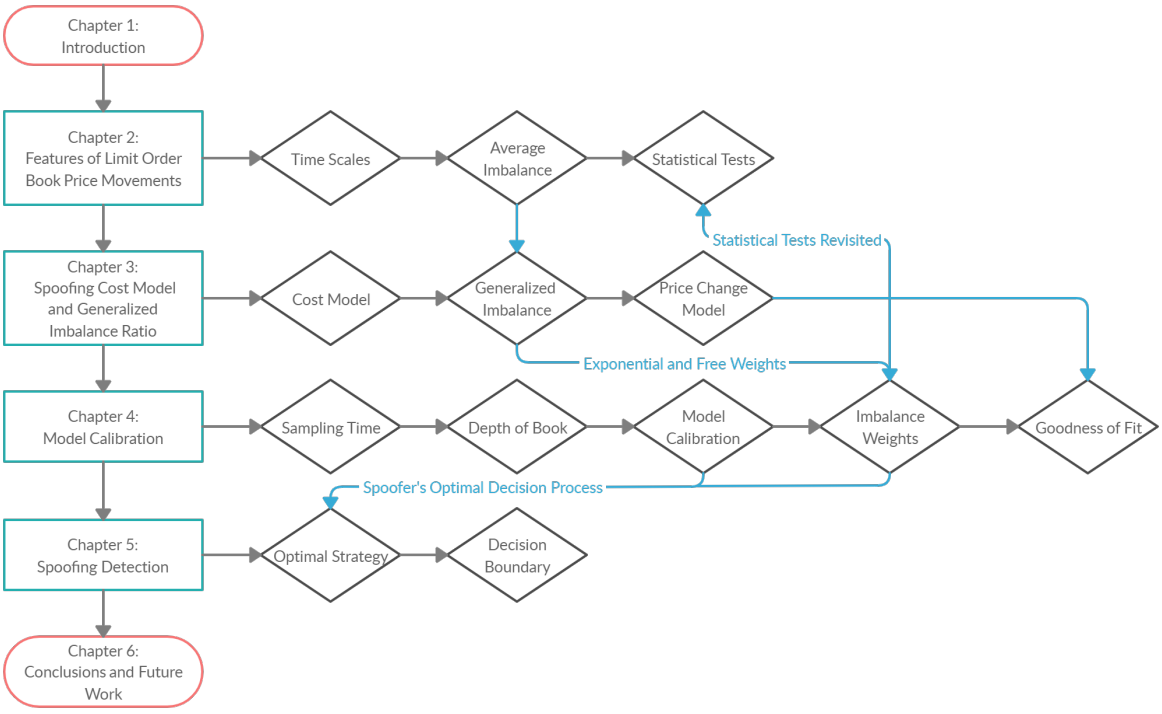


Figure 1.12: Breakdown and connections of main concepts of the thesis.

Chapter 2

Features of Limit Order Book Price Movements

2.1 Introduction

Now that we have covered how stocks are traded electronically by means of the limit order book we can can explore some deeper features of the book. Our main focus in this work will be on the features which impact the changes in the best ask price and how can we write down a model which incorporates these features into the price changes. We focus on the ask side of the limit order book, but the same methodology can be applied to the bid side by just flipping replacing ask $\rightarrow$ bid in all relevant equations and definitions since there is a conceptual symmetry between the two sides of the book.

Extensive literature exists on features of the limit order book impacting future price movements. The two most common features used are the current prices, volumes, and order numbers on both sides of the limit order book. Cont et al. [50] show that the difference between the order flows on the bid and ask sides of the book, net order flow imbalance, is an excellent predictors of future prices. Similar work using net order flow and future price prediction was conducted in [51, 52]. The volumes in the book have been used frequently in the literature as a predictor variable of future price movements in machine learning algorithms [53–55] and more standard financial modelling settings [56, 57]. More recently, studies have been done using the volume imbalance ratio as a predictor variable which aggregates the limit order book volumes down to a single number. Cartea et al. [58] being the first to explore this idea and apply it to enhancing algorithmic trading strategies[1]. Since then a number of papers have expanded on this

[1]Their paper on spoofing [49], which we discussed in chapter 1, uses the volume imbalance ratio for its regime switching mechanism.

work, [59] for example. However, information about depths in the book beyond the best ask and bid has not been studied in the context of the volume imbalance ratio. Following in the footsteps of Cartea et al. [49], in chapter 3, we generalize the classical definition of the volume imbalance in order to include volume information deeper in the book which a spoofer could influence, and, in turn impact future price movements. Past studies [51, 56] have shown the depth of book plays a role in determining future price movements and we aim to continue this study using the volume imbalance ratio.

In this chapter we discuss the time dependence that arises from limit order book data on price movements as well as one particular feature that has been explored in the literature in a different context – the volume imbalance ratio. The volume imbalance ratio is a way of aggregating the volumes in the limit order book at all available prices into a single feature variable. This variable, defined more precisely later in this chapter, has been connected by the past literature with limit/market/cancel order frequency on both sides of the book; the result of these, mediated through the logic of the order book, leads to price impact. Instead of taking this approach we look directly at volume imbalance and how it impacts the distributions of price movements as a way of describing how the shape of the limit order book determines possible prices. When referring to the shape of the limit order book we follow, for example [8], to mean how the quantities of shares q at each price p can be thought of as some function $q(p)$ on each side of the book. That is, a general term to encompass the value, slope, curvature, etc, of $q(p)$ for both sides of the book.

Before we can discuss the volume imbalance we need to tackle the importance of some of the properties of the limit order book over time. This includes both the time of day and the time interval over which the limit order book is observed. From there we can start to build our model.

2.2 Sampling Time and Price Movements

Care must be taken when investigating the distribution in change of best bid or ask prices throughout the trading day for a particular stock. Computing a price change distribution implicitly requires a time step or sampling frequency. This could be when any order arrives, when a trade occurs, or every 5 or 10 seconds. The choice of frequency will yield potentially very different results for each distribution.

For example, Figure 2.1 shows the distribution of best ask prices in ticks (of pennies) for AEM stock on April 17, 2017 for different sampling intervals. We chose AEM stock because it was the most actively traded stock on April 17, 2017. We observe the peak at

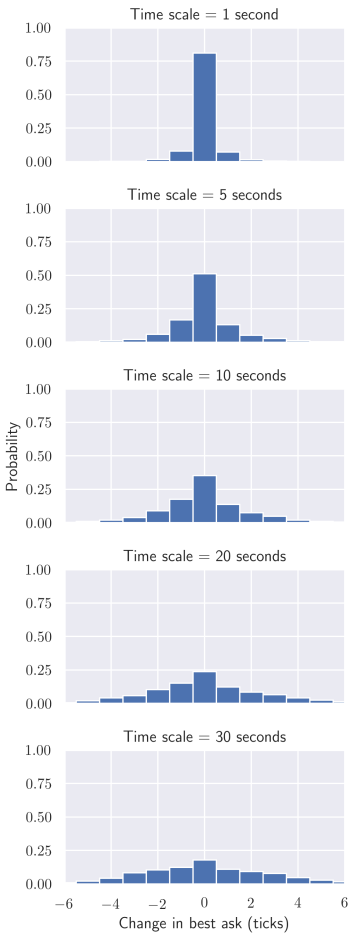
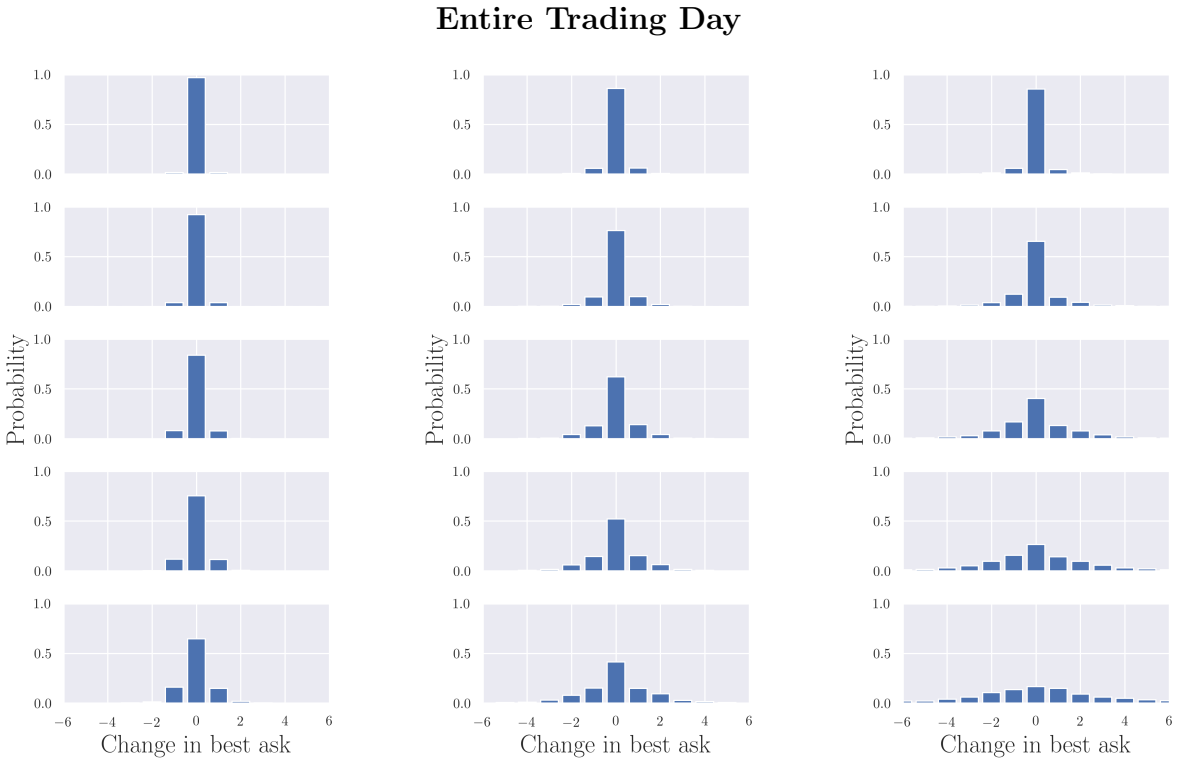


Figure 2.1: Probability of change in best ask price (in pennies or ticks) for AEM stock on April 17, 2017. Each subplot shows the distribution of prices when sampling every 1, 5, 10, 20, and 30 seconds, respectively.

0 decreases with an increase in sampling interval. This is because prices have a higher probability of moving over longer time frames. Not surprisingly, we observe a larger range of price movements over longer time periods.

The price distributions will also heavily depend on the activity of the stock. The more active the stock, the larger the price movements we observe at a given sampling frequency. For less traded stocks we may need to look at significantly larger sampling times to produce pictures like Figure 2.1. When performing any kind of modelling of price distributions with limit order book data one needs to be mindful of the sampling intervals used.

Figure 2.2 compares the distribution in the best ask price for three different stocks over 5 time intervals. We see each stock exhibits a different time dependency on its price movements. Even after 60 seconds, XEG resembles HFU at 5 seconds. Similarly, HFU



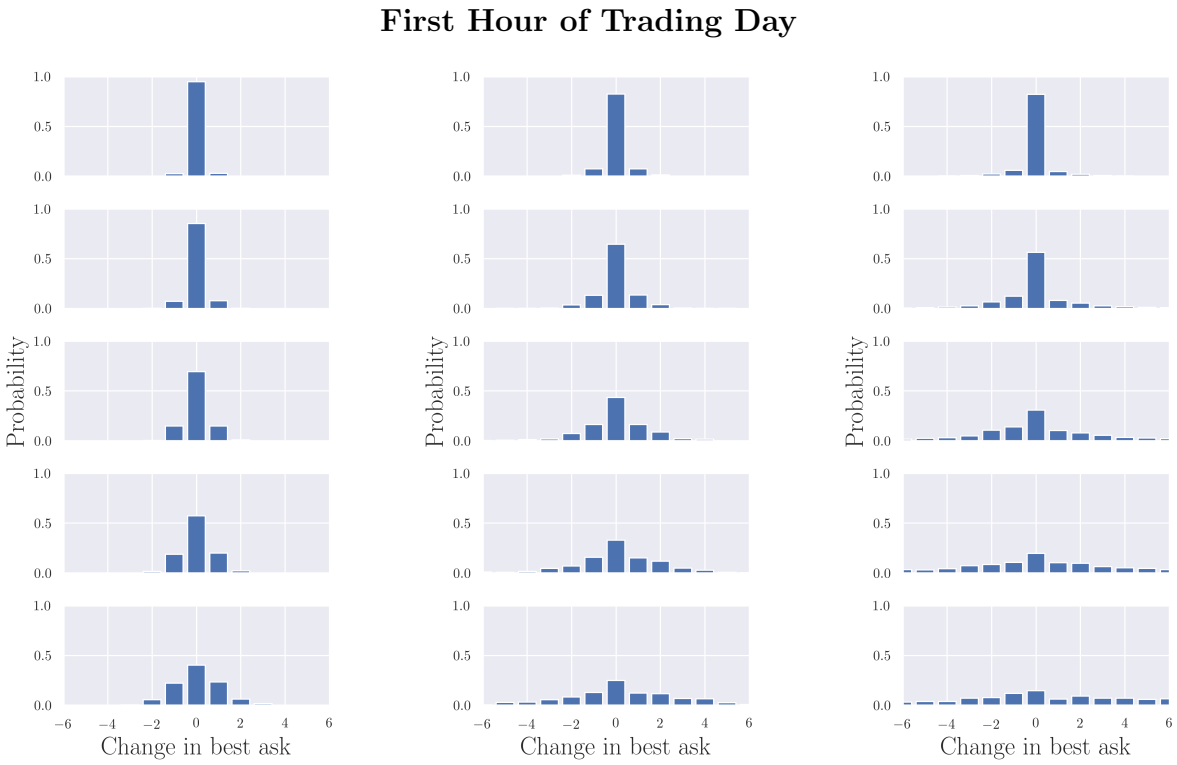
(a) XEG (b) HFU (c) CNR

Figure 2.2: Distribution of change in best ask price for XEG, HFU, and CNR stocks on June 1 - 8, 2017. Each subplot row from top to bottom is the distribution after sampling 1, 5, 15, 30, 60 second time intervals, respectively. Data used is from the entire trading day for each day.

at 60 seconds resembles CNR at 15 seconds. Also, among these three stocks you would likely say CNR is most like AEM, shown in Figure 2.1.

In addition to the sampling time problem we also have that different periods of the trading day display different behaviour. It is well known from past literature that the first hour of the trading day exhibits a larger number of orders and more frequent price movements than other hours of the trading day [34, 49, 50, 56, 57, 59]. This is usually dealt with by excluding data from the first 30 to 60 minutes of the trading day. Comparing Figure 2.3 to Figure 2.2 we can see that the first hour of the trading day for our three stocks see more frequent price movements over the same time interval and the movements you do see are larger. This is odd and possibly due to our small selection of stocks. A more expansive analysis will be done in the next chapter.

We also have another issue appearing in the final subplot of Figure 2.3 for CNR stock. The same plot is shown blown up in Figure 2.4. If we take sampling times long enough we start to lose the structure and symmetry of the price distribution we have at smaller



(a) XEG (b) HFU (c) CNR

Figure 2.3: Distribution of change in best ask price for XEG, HFU, and CNR stocks on June 1 - 8, 2017. Each subplot row from top to bottom is the distribution after sampling 1, 5, 15, 30, 60 second time intervals, respectively. Data used is from the first hour of the trading day for each day.

time intervals. It could be because the distribution in Figure 2.4 is a mixture of two or more distributions that emerges over longer time intervals. This is outside the scope of this work, but would be an interesting direction. We will end up up limiting ourselves to time intervals over which this behaviour is not observed.

In contrast to the first hour of the day, taking a look at the final hour of the trading day in Figure 2.5 we cannot see too much difference from the patterns in the entire trading day in Figure 2.2. Although Figure 2.5 does not support this, the final 30 to 60 minutes of the trading day are often excluded from other works. This is to exclude the period of the day where, on top of the continuous trading, traders are canceling their orders in order to prevent them from staying on the book overnight. The TSX allows people to cancel and place market orders after the market closes, but not to place new limit orders.

This would suggest that stocks have their own time scales determined by their price movements. If we want to investigate collections of stocks and analyze the characteristics

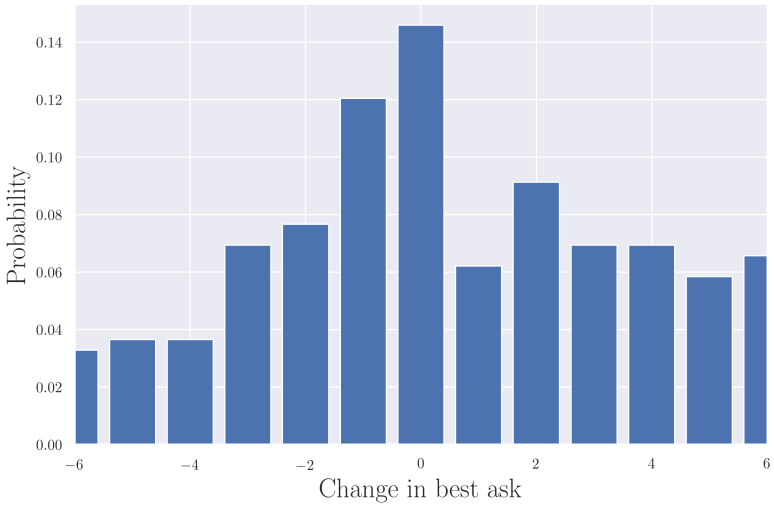


Figure 2.4: Distribution of change in best ask price for CNR stock over 60 second intervals during first hour of trading day. Zoomed in picture of the final subplot for CNR stock in Figure 2.3.

of their price movements we need to be able to compare them 'apples to apples'. Over what time scales would XEG and CNR look most similar? We have to fix the time scale and determine an appropriate way of determining what we mean by 'most similar'. This will be expanded on in the next chapter.

2.3 Volume Imbalance Ratio, Prices, and Time

A much studied predictor of limit order book price movements is the volume imbalance ratio. The volume imbalance ratio is a variable which quantifies the mismatch in the quantity of shares available to be bought versus sold. Its a function of the shape of the book and the collection of ratios over the trading day forms a distribution. Cartea et al. [58] show that the volume imbalance has a strong impact on the frequency of market orders placed on a particular side of the limit order book. The increased frequency of market orders would cause prices to move up (down) for the best ask (bid) as liquidity is removed and orders walk the book. If the volume imbalance can be a predictor for market order frequency then it might be a predictor of the distribution of price movements. A relationship between these two distributions would allow us to investigate the question - Does the shape of the limit order book have a hand in determining future price movements?

If at time t, we can compute the volume of shares at the best bid $V_{\text{bid}}(t)$ and the volume of shares at the best ask $V_{\text{ask}}(t)$ we can calculate the volume imbalance ratio $I(t)$

(a) XEG (b) HFU (c) CNR

Figure 2.5: Distribution of change in best ask price for XEG, HFU, and CNR stocks on June 1 - 8, 2017. Each subplot row from top to bottom is the distribution after sampling 1, 5, 15, 30, 60 second time intervals, respectively. Data used is from the last hour of the trading day for each day.

by equation 2.3.1.

$$I(t) = \frac{V_{\text{bid}}(t) - V_{\text{ask}}(t)}{V_{\text{bid}}(t) + V_{\text{ask}}(t)} \tag{2.3.1}$$

It is clear from equation 2.3.1 that the volume imbalance ratio $I(t) \in [-1, 1]$ and $I(t) \to 1$ as $V_{bid}(t) \to \infty$ while $I(t) \to -1$ as $V_{ask}(t) \to \infty$ holding the other volume fixed. That is, an imbalance of 1 (-1) means that all volume is on the bid (ask) side of the book. We refer to $I(t)$ in equation 2.3.1 as the instantaneous volume imbalance at time t.

For price movements from order to order we can use the instantaneous imbalance as a predictor, but when looking at price movements over a specific time interval we need a way to aggregate the instantaneous imbalance. A simple way to do this would be to just take the mean of the instantaneous imbalances over the time interval. We start by choosing equally spaced sample times $t_1 < t_2 < \cdots < t_M$. For a time interval $\Delta t = t_m - t_{m-1}$

we have orders at times $s_1 < s_2 < \cdots < s_N$ where $t_{m-1} \leq s_1$ and $s_N \leq t_m$. From the instantaneous imbalances $I(s_1), I(s_2), \ldots, I(s_N)$ we define the average imbalance I_{avg} as

$$I_{\text{avg}} = \frac{1}{N} \sum_{i=1}^{N} I(s_i) \tag{2.3.2}$$

Keep in mind the number of orders N will likely be different for each time interval. Another approach is to weight each instantaneous imbalance over the time interval. One would expect that more weight should be applied to the imbalance the longer it goes without changing. Many orders enter the book and are immediately cancelled and these orders should have less impact on prices since they do not exist in the book long enough for people to act upon them. We can incorporate this idea by weighting each instantaneous imbalance $I(s_i)$ by the time between successive orders, $s_{i+1} - s_i$. Instead of equation 2.3.2 we can define the average time weighted imbalance I_{avg} between times t_{m-1} and t_m as

$$I_{\text{avg}} = \frac{1}{s_N - s_1} \sum_{i=1}^{N-1} I(s_i)(s_{i+1} - s_i) \tag{2.3.3}$$

We exclude the instantaneous imbalance at s_N since the next order would be outside the time interval and we have no associated next price movement. Figure 2.6 shows our one period model between times t and $t + \Delta t$ and how we aggregate our instantaneous imbalances into an aggregated average imbalance I_{avg}.

Figures 2.7 and 2.8 are the time series and histograms of the average imbalance of AEM stock on April 17, 2017 over 5 second intervals. We can think of the definition from equation 2.3.2 as having no time weighting as all orders are treated as equal. In this chapter we use time weighting when we refer to the average imbalance, but compare results using both methods in a later chapter.

It is difficult to tell what the difference is between the subplots in Figure 2.7. Figures 2.9 and 2.10 depict the mean and variance, respectively, of the 5 second average imbalance over 10 minute time intervals. The mean is roughly the same for both methods and they alternate over which is greater throughout the day, but the variance is almost always greater with time weighting.

The difference in the mean and variance of the average imbalance between the two aggregation choices may be due to a change in sample size caused by the introduction of the time weights in equation 2.3.3. Any order that arrives and is immediately cancelled would be given a weight of approximately zero – effectively removing it from the calculation and lowering the sample size within that time interval. We can compare the sample

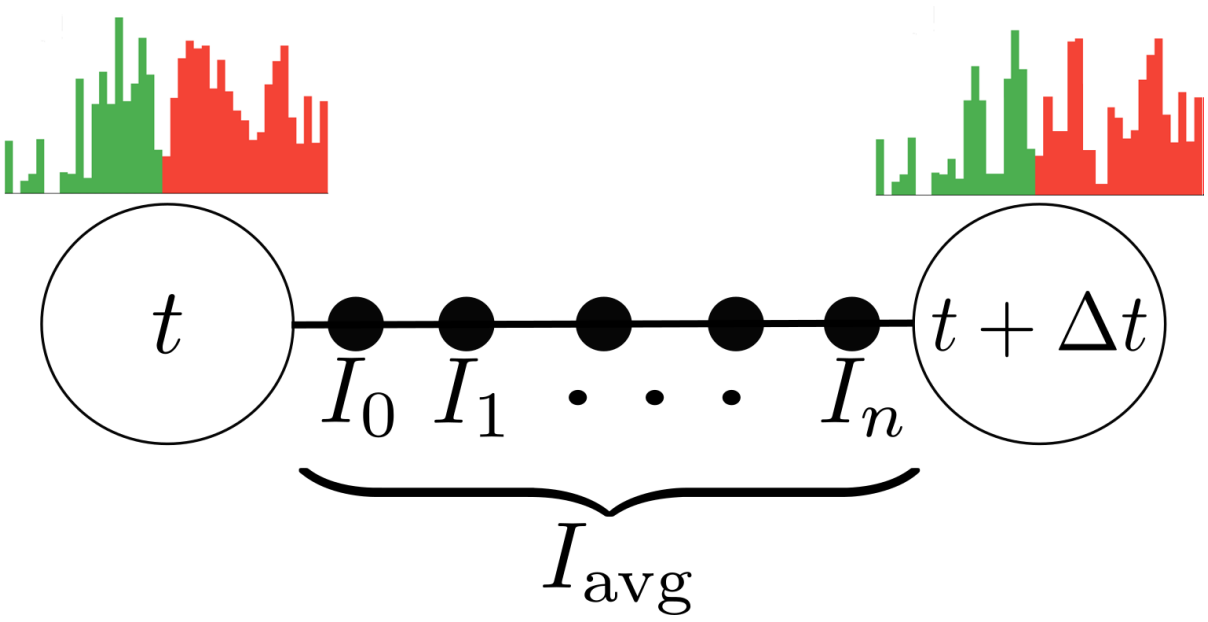
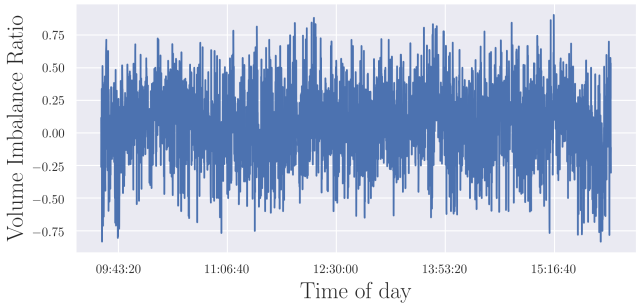
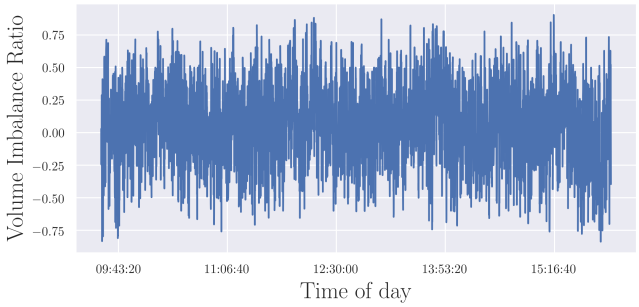


Figure 2.6: One period model for aggregating instantaneous imbalances into the average imbalance I_{avg}. We use the notation $I_i = I(s_i)$ for the instantaneous imbalances between time t and $t + \Delta t$.

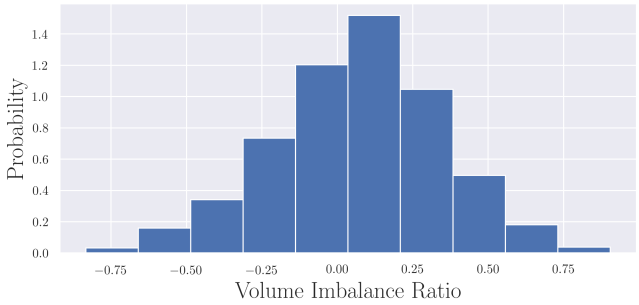


(a) Average imbalance (simple mean)

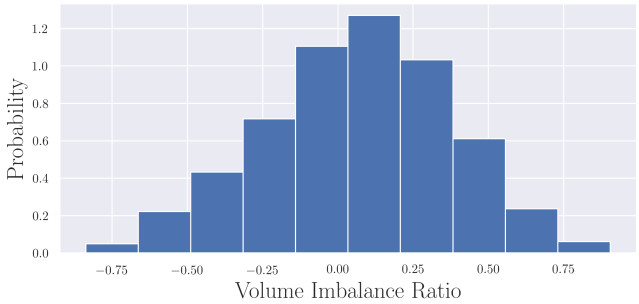


(b) Average imbalance (time weighted mean)

Figure 2.7: Time series of the volume imbalance ratio for AEM stock on April 17, 2017. Sampling was done every 5 seconds



(a) Average imbalance (simple mean)



(b) Average imbalance (time weighted mean)

Figure 2.8: Probability density of volume imbalance ratio for AEM stock on April 17, 2017. Sampling was done every 5 seconds.

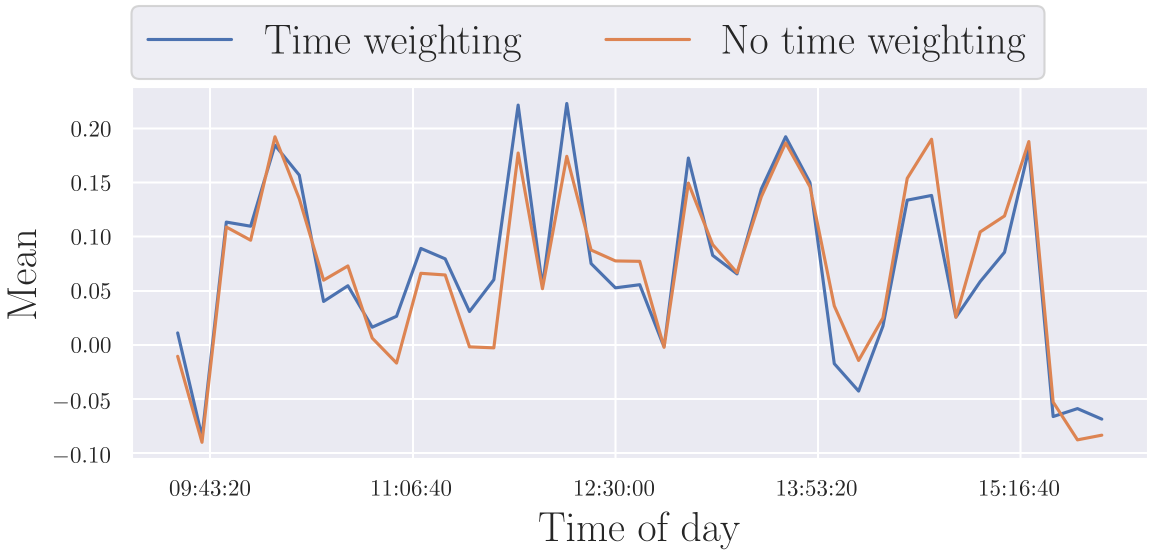


Figure 2.9: Mean of time series in Figure 2.7 over 10 minute intervals. Mean varies slightly with both methods having periods being greater than the other.

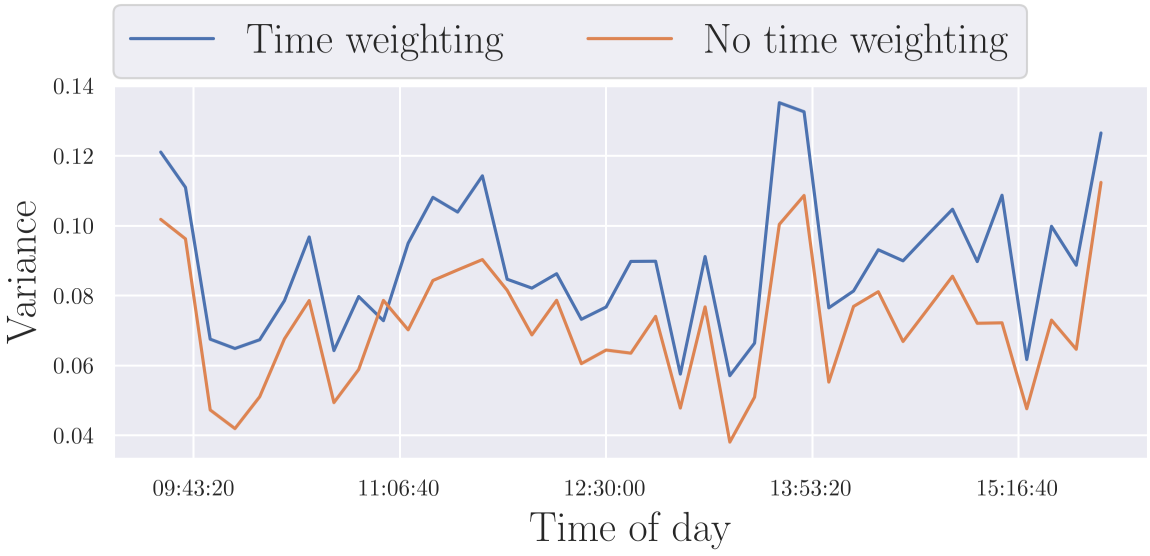


Figure 2.10: Variance of time series in Figure 2.7 over 10 minute intervals. Variance is almost always greater with time weighting.

size in each 5 second bin to Kish's effective sample size [60] when the mean of the data is calculated with weights w_i. The effective sample size is defined as

$$n_{\text{eff}} = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2} \tag{2.3.4}$$

Here the weights w_i are taken to be the time differences $s_{i+1} - s_i$ between orders in equation 2.3.3. Figure 2.11 compares the sample size of each 5 second bin to the effective sample size over the trading day. The smaller effective sample size is a possible explanation for the increased variance when using time weighting. We also see that when we use time weighting we need significantly fewer samples to achieve roughly the same mean. This would suggest that there are a few orders in each 5 second bin which dominate the calculation for the average imbalance.

Figures D.1 and D.2 show the result of replacing the instantaneous imbalance with random draws from the standard normal distribution to see if it is solely the weighting which causes the increased variance. We see the same increase in variance as we would expect (although the gap is much smaller), but the mean for the average imbalance with both weightings is much closer than we see from the standard normal. Figure D.6 is a visual test to see if the imbalance is independent and identically distributed. We can see that this is not the case over short time scales, but is true over longer time scales and could explain why we see the approach taken to calculate I_{avg} has little impact on the mean. We also see that there is an interesting grid appearing based on specific imbalance

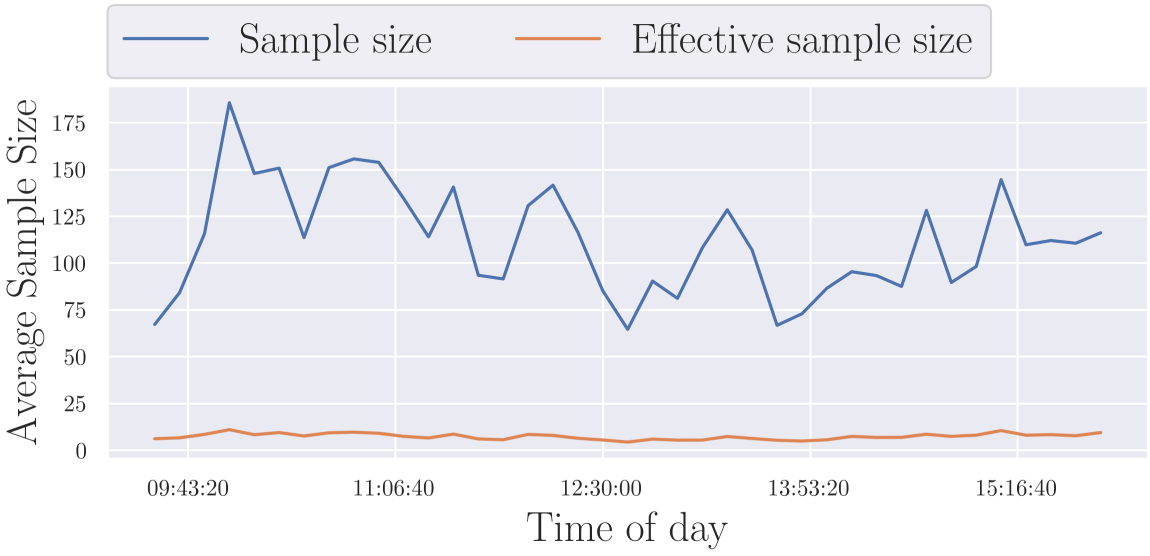


Figure 2.11: Average sample size in 5 second bins over 10 minute intervals. Sample size is the number of orders in each 5 second bin and the effective sample size is calculated from the time weights. Effective sample size is also known as Kish's effective sample size. Weights are the same as Figure 2.7.

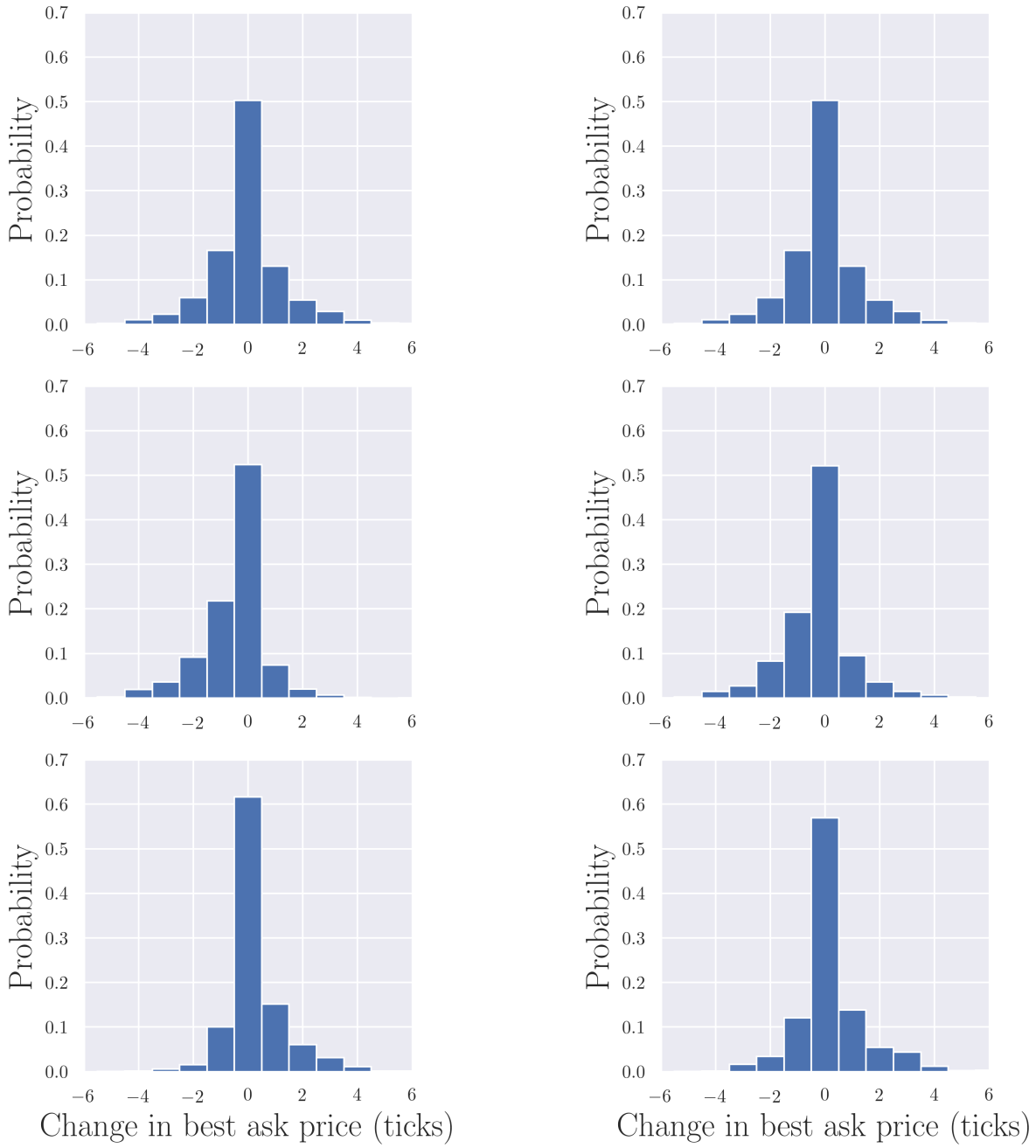
values. It looks like the imbalance is bound between nice rational number values (ex. 1/3, 2/5, 3/4, etc). This may just be an artefact of how algorithms place limit orders at the best bid/ask over the trading day. Cartea et al. [61] find a similar pattern in the number of executed orders at particular cents in the stock price. Their interpretation is "that for some reason (rational or not) there is a preference for providing liquidity at prices that end in round cent values". We may be seeing something similar in the volume imbalance by traders placing orders to move the imbalance to nice round numbers as well.

Figures D.3, D.4, D.5, and D.7, show the same characteristics as AEM on the same day to show this is not an isolated case.

Our two different ways of calculating the average imbalance have only a small impact on the mean while the variance increases from the smaller sample size. The preceding figures were only for a single stock on a single day and would just be anecdotal evidence for what we see in Figures 2.9 and 2.10, but the real test will be which weighting scheme provides better predictive power of price movements.

The impact of the volume imbalance on the distribution of price movements for AEM stock is shown in Figure 2.12. The effect of imbalance is not to translate the distribution, but to skew it. A heavy positive (negative) imbalance skews the distribution to the right (left). This would mean it is more likely to see an up (down) price movement when the

imbalance is positive (negative).



(a) Average imbalance from mean

(b) Average imbalance from time weighting

Figure 2.12: Probability of change in best ask price (in pennies or ticks) for AEM stock on April 17, 2017. Sampling was done every 5 seconds. The top plots show the distribution of prices unconditional on the volume imbalance. The center is conditional on the volume imbalance being less than the 25% quantile. The bottom is conditional on the volume imbalance being greater than the 75% quantile.

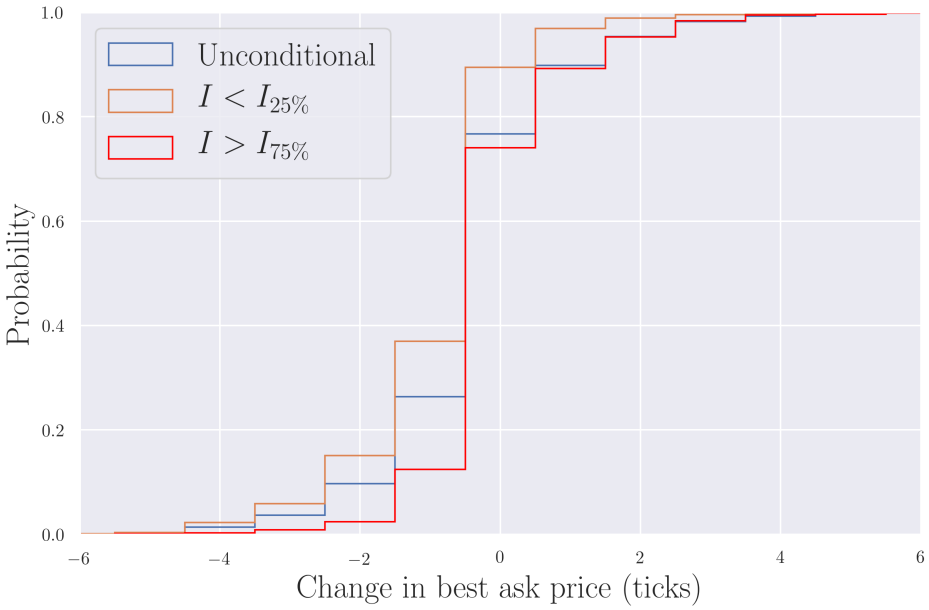
Since AEM was already skewed regardless of the imbalance on April 17, 2017 it is hard to tell just from looking at the distributions what is really happening. In Figure 2.13 we plot the cumulative distribution function (cdf) for each subplot in Figure 2.12 on top of each other. We can see that the cdf is skewed left when the imbalance is less than the 25% quantile and skewed right when greater than the 75% quantile. We can also see that the time weighting has helped separate the different imbalance regimes from the unconditional case for the positive price movements, but this is just a single stock and we will need look at some measure over a larger collection to see how these two ways of calculating the average imbalance differ.

In these examples we condition on the imbalance taking its more extreme values in order to magnify the effect, but we will need to use hypothesis testing to determine if conditioning on the imbalance has a statistically significant impact on price movements and how large that impact is if it exists. We then conduct two tests for statistical independence between the price movements and the imbalance.

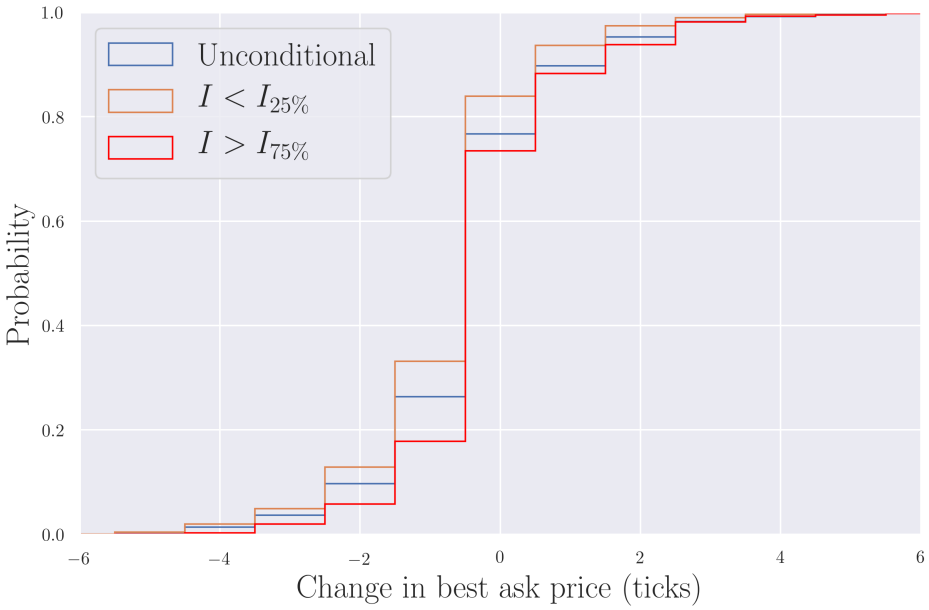
2.4 Statistical Tests of Volume Imbalance

In this section we explore two tests to provide evidence whether the impact of the volume imbalance ratio is a predictor of changes in the best ask price. We will focus on the instantaneous imbalance as a predictor of order-by-order price movements – the current instantaneous imbalance as a predictor of the price movement caused by the next order. We also perform the same tests for AEM stock on April 17, 2017 over 5 second intervals as described in the previous sections of this chapter. So, the average imbalance over 5 seconds as a predictor of the overall price movement from the beginning to the end of the time interval. We will explore the imbalance and price movements over time intervals in the next chapter, but give a taste for what is to come in this section. Our examples in the next subsections will be using the time weighting when calculating the average imbalance, but we will present the results from both methods in chapter 4.

The purpose of the first test is to determine if the difference between the distribution of the change in best ask price when the imbalance is positive or negative is statistically significant. We will call the imbalance coarse because it is binary (positive or negative) while the price movements are fine because we take the full distribution itself (probability of seeing each individual tick). For the second test we bin the imbalance into 4 groups to see if the probability of price movement being up or down dependent on the imbalance bin is statistically significant. Here things are reversed where the imbalance is fine (broken into bins) and the price movements are binary (up or down).



(a) Average imbalance from mean



(b) Average imbalance from time weighting

Figure 2.13: Cumulative probability distribution for the three subplots shown in Figure 2.12.

For our statistical tests we will have two collections of categorical count data to which we apply Pearson's chi-squared test for statistical independence. The null hypothesis is that the two collections are independent of each each other and the test gives us a p-value for accepting or rejecting the null hypothesis. In our case the statistical population

is the collection of all time-indexed orders placed on the exchange for a specific stock over a given time period. At each order we know the instantaneous imbalance prior to its arrival and the best ask price. Once a time interval Δt has been specified we can sample from this population by randomly drawing a time-indexed order and all other orders Δt seconds in the future. We can then aggregate the instantaneous imbalances into the average imbalance via equation 2.3.3 and determine the change in the best ask price between the first and last order in Δt.

An issue which can arise with statistical tests is that very large sample sizes can produce arbitrarily small p-values. To compensate for this we also provide the Cramer's V measure [62] from the chi squared value of the test which is independent of sample size and bound between 0 and 1. A Cramer's V score of 1 means the two collections are identical to each other. The measure gives a strength of association between the two count collections. We also apply a bias correction to the Cramer's V to get a more conservative estimate of the strength of association [63]. Details of how p-values and the Cramer's V are calculated for Pearson's chi-square test for statistical independence can be found in Appendix C.

In the following two subsections we detail the two count collections used in each test and provide the results for a larger sample of stocks.

2.4.1 Test 1: Coarse Imbalance and Fine Price Movements

For our first test we look at the effect of the sign of the imbalance on the individual price movement counts order-by-order and over 5 second intervals. Our categorical data is the number of individual price movements we see when the imbalance is positive or negative. A common rule for the chi square test is to have 5 or more counts in each cell for larger contingency tables like Tables 2.1 and 2.2. We then take the largest range of price movements which satisfy this rule. Order-by-order we take price movements between -3 and 3 ticks and over 5 second time intervals we take the price movements between -5 and 5 ticks. We will use AEM stock on April 17, 2017 for our examples. For the time interval data we sample 30000 five second intervals throughout the day which can overlap. If the 5 second interval has 1 or fewer orders we discard it.

Table 2.1 summarizes the order-by-order count data. Table 2.2 summarizes the 5 second interval count data. In both cases we see that there is a significantly larger number of price movements when the imbalance was positive, but there were more overall downward price movements over the day.

Dividing the counts in each row of Tables 2.1 and 2.2 by the row's total count we get

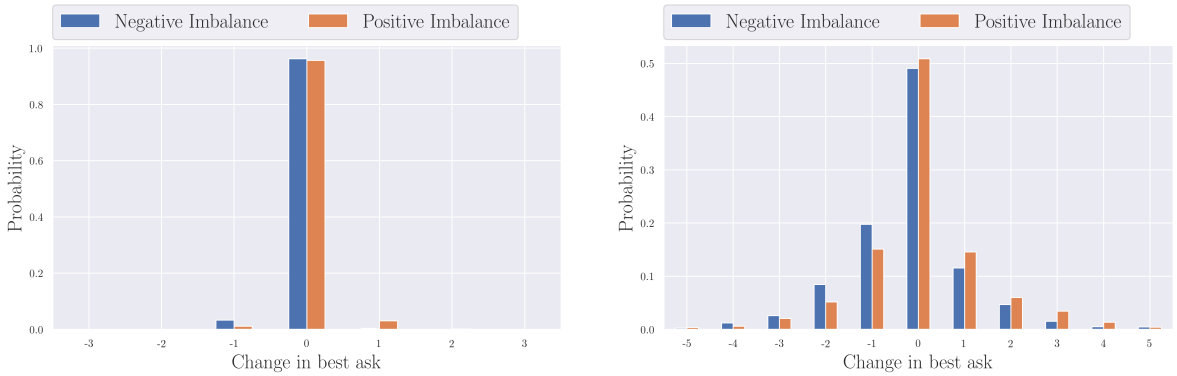
	Δp_a							
Imbalance	-3	-2	-1	0	1	2	3	Total by Imbalance
$I > 0$	37	113	2859	248568	7936	257	71	259841
$I < 0$	22	84	6193	179684	697	9	7	186696
Total by Δp_a	59	197	9052	428252	8633	266	78	446537

Table 2.1: Counts for order-by-order price movements conditioned on sign of imbalance. Data for AEM stock on April 17, 2017.

	Δp_a											
Imbalance	-5	-4	-3	-2	-1	0	1	2	3	4	5	Total by Imbalance
$I > 0$	49	132	329	912	2625	9410	2493	1114	638	243	84	18029
$I < 0$	34	136	339	945	2284	5529	1274	463	209	83	66	11362
Total by Δp_a	83	268	668	1857	4909	14939	3767	1577	847	326	150	29391

Table 2.2: Counts for 5 second interval price movements conditioned on sign of imbalance. Data for AEM stock on April 17, 2017. Average imbalance calculated using time weights.

the distributions shown in Figure 2.14. Figure 2.14 shows us that, when conditioning on a negative imbalance, we see more downward price movements and vice versa for the positive imbalance. This is the case order-by-order and also over the 5 second intervals.



(a) Order-by-order

(b) 5 second intervals

Figure 2.14: Distribution of change in best ask price order-by-order and over 5 second intervals conditioned on the average imbalance being positive or negative. Data from AEM stock on April 17, 2017.

This is all just initial observation from the count data, but now we can use Tables 2.1 and 2.2 as contingency tables and perform our chi square test on each under the null hypothesis that the change in the best ask price is independent of the imbalance. The results of the tests are shown in Table 2.3.

The p-values in both cases imply that we can easily reject the null hypothesis at any significance level, but the Cramer's V tells us the strength of the relationship between

	Cramer's V	p-value	Correlation
Order-by-order	0.122	0.0	0.123
5 second intervals	0.114	7.33e-76	0.129

Table 2.3: Summary of chi square test for coarse imbalance and fine price movements. Correlation is taken between change in best ask price and the imbalance.

the imbalance and the count data is meaningful and the correlation between the two is positive. We report the same test for a large sample of stocks in Table 2.4. We only run the same test on order-by-order data for other stocks as we still need a way of comparing stocks which have very different behaviour over the same time interval. This will be discussed in the next chapter.

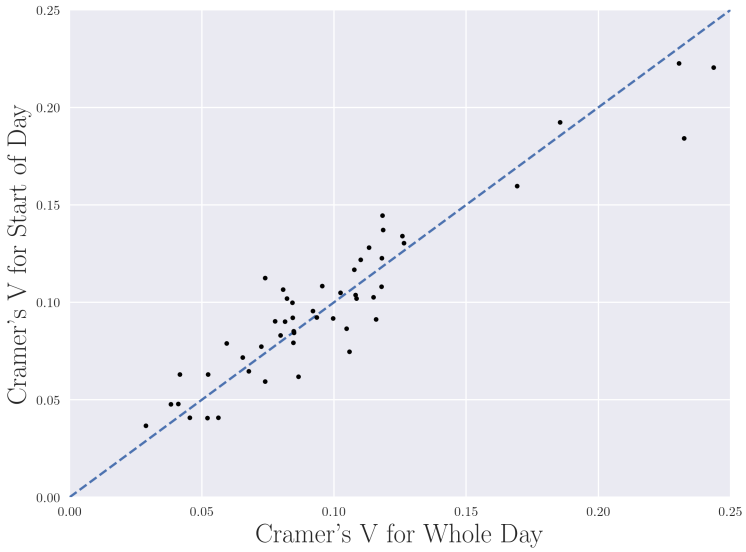


Figure 2.15: Cramer's V for whole day and start of day plotted against each other. Data is from Table 2.4. The dashed line would represent a totally linear relationship between the two statistics.

Table 2.4 summarizes our first test across a large sample of stocks using data from the entire trading day as well as the first hour of the trading day. We see that all p-values imply we can reject the null hypothesis at any significance level and we have a fairly strong association between our counts and the sign of the imbalance across all stocks. Note that some stocks have a much larger Cramer's V than others - HQU, HSU, VFV, for example. These stocks have a stronger relationship between the imbalance and price movements. From Figure 2.15 we see roughly the same Cramer's V is found in both time periods giving the impression that order-by-order the imbalance has roughly the same association to changes in the best ask price regardless of time of day.

ticker	Cramer's V
HQU	0.244
HSU	0.232
VFV	0.231
HVI	0.185
VUN	0.169
PAAS	0.126
NA	0.126
POW	0.118
HOD	0.118
VGG	0.118
UFS	0.118
CNR	0.116
BMO	0.115
XWD	0.113
IPL	0.11
SLF	0.108
PPL	0.108
KL	0.108
FTS	0.106
AEM	0.105
BAM.A	0.102
HFU	0.0996
PWF	0.0956
ARX	0.0934
PVG	0.092
BIP.UN	0.0865
OR	0.0848
GIL	0.0847
HXU	0.0846
SSO	0.0843
FM	0.0842
T	0.0822
IMO	0.0815
XQQ	0.0807
ERF	0.0797
ZEB	0.0777
RBA	0.0739
FR	0.0739
FSV	0.0724
GIB.A	0.0677
CPG	0.0655
CCO	0.0594
SW	0.0562
IMG	0.0523
GOOS	0.0521
WCN	0.0455
XEG	0.0417
K	0.0411
G	0.0383
TC	0.0288

ticker	Cramer's V
HQU	0.221
HSU	0.184
VFV	0.223
HVI	0.192
VUN	0.16
PAAS	0.13
NA	0.134
POW	0.137
HOD	0.145
VGG	0.123
UFS	0.108
CNR	0.0913
BMO	0.103
XWD	0.128
IPL	0.122
SLF	0.102
PPL	0.104
KL	0.117
FTS	0.113
AEM	0.0866
BAM.A	0.105
HFU	0.0918
PWF	0.108
ARX	0.0923
PVG	0.0956
BIP.UN	0.0618
OR	0.0844
GIL	0.0852
HXU	0.0793
SSO	0.0922
FM	0.0999
T	0.102
IMO	0.0902
XQQ	0.107
ERF	0.083
ZEB	0.0904
RBA	0.0594
FR	0.0747
FSV	0.0773
GIB.A	0.0647
CPG	0.0717
CCO	0.079
SW	0.0409
IMG	0.0631
GOOS	0.0406
WCN	0.0408
XEG	0.063
K	0.0479
G	0.0478
TC	0.0368

Table 2.4: Summary of chi square test for coarse imbalance and fine price movements order-by-order. Data is taken from June 1-8, 2017. All p-values are zero or very close ($\approx 10^{-46}$ at most) to zero. The left subtable uses data from the entire trading day while the right subtable uses data only from the first hour of the trading day. Tickers are sorted by magnitude of Cramer's V for the whole day.

In the next subsection we break the imbalance down into smaller intervals and looking at the price direction in each interval. We can then apply the same statistical test to that count data as we did in this subsection.

2.4.2 Test 2: Fine Imbalance and Coarse Price Movements

This time we divide the imbalance into bins and condition the sign of the price movement on each imbalance bin. We follow [58] and split the imbalance up into the follow intervals labeled 1, 2, 3, and 4:

Bin 1: $I \in \left[-1, -\frac{1}{3}\right)$

Bin 2: $I \in \left[-\frac{1}{3}, 0\right)$

Bin 3: $I \in \left[0, \frac{1}{3}\right]$

Bin 4: $I \in \left(\frac{1}{3}, 1\right]$

Tables 2.5 and 2.6 show the count data for order-by-order and 5 second intervals, respectively. With these counts we can test how the probability of up and down price movements changes depending on the bin membership of the volume imbalance. As in the previous subsection, we can divide each bin count by total in each bin to produce a distribution for the probability of an up or down movement given the bin. The probabilities are shown in Figure 2.16. One will notice that the total counts in Tables 2.5 and 2.6 are very different than Tables 2.1 and 2.2. This is because we are conditioning on the change in the best ask being nonzero in Tables 2.5 and 2.6 while a significant number of samples in Tables 2.1 and 2.2 correspond to no change in the best ask price.

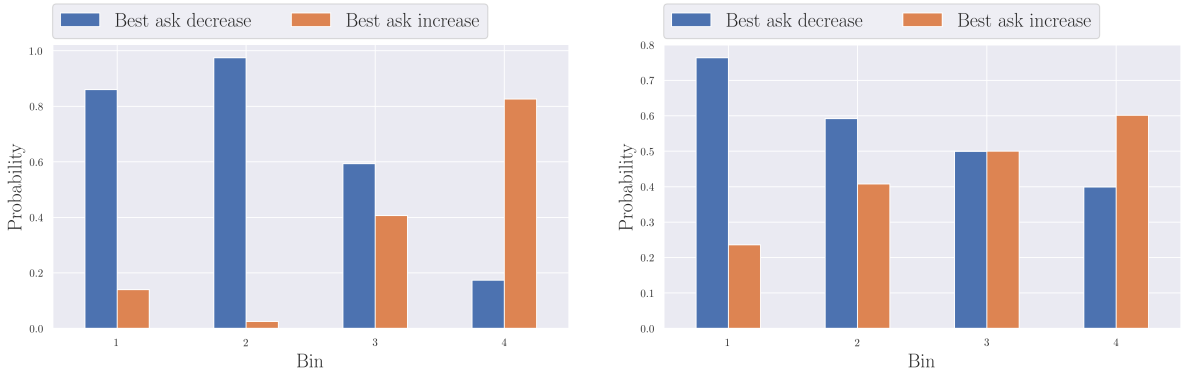
	Bin				
Δp_a	1	2	3	4	Total by Δp_a
$\Delta p_a > 0$	657	58	1015	7259	8989
$\Delta p_a < 0$	4051	2253	1484	1527	9315
Total by Bin	4708	2311	2499	8786	18304

Table 2.5: Count data order-by-order price direction conditioned on imbalance bin. Data from AEM stock on April 17, 2017.

From Figure 2.16 we see order-by-order almost all price movements are down in bins 1 and 2, while almost all price movements are up in bin 4. However, in bin 3 we are still more likely to see a price drop even though the imbalance is positive. We see a similar pattern across all stocks we examined. In Figure 2.17 we are looking at the same plot as

	Bin				
Δp_a	1	2	3	4	Total by Δp_a
$\Delta p_a > 0$	312	1809	3179	1489	6789
$\Delta p_a < 0$	991	2785	3078	978	7832
Total by Bin	1303	4594	6257	2467	14621

Table 2.6: Count data over 5 second intervals for price direction conditioned on imbalance bin. Data from AEM stock on April 17, 2017.



(a) Order-by-order

(b) 5 second intervals

Figure 2.16: Probability of change in best ask price being positive or negative conditioned on the average imbalance being in bins 1, 2, 3, or 4. Data from AEM stock on April 17, 2017.

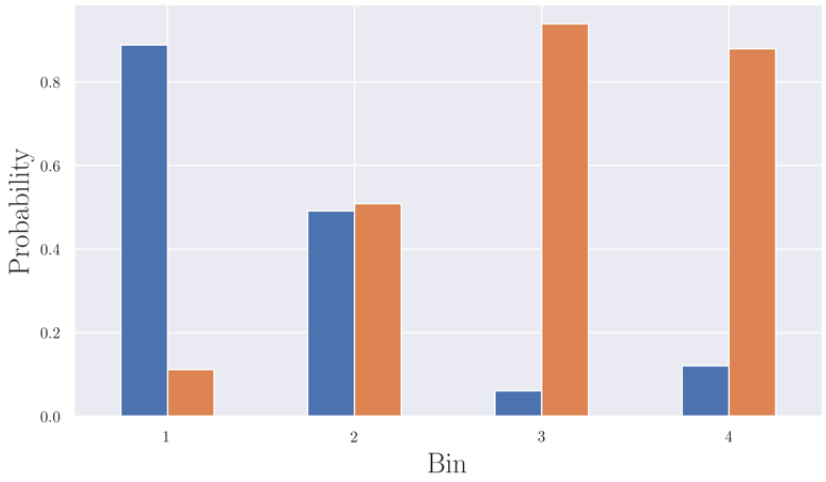


Figure 2.17: Probability of change in best bid price being positive (orange) or negative (blue) conditioned on the average imbalance being in bins 1, 2, 3, or 4. Data from AEM stock on April 17, 2017.

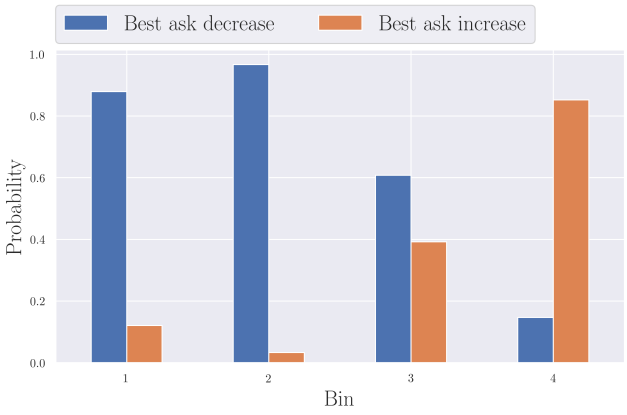
Figure 2.16a except for the change in the best bid price – that is, order-by-order changes in the best bid. Here the pattern has been reversed and we see that there is a slightly

higher probability of a price increase even though the imbalance is negative. Again, we see the same thing across all stocks when looking order-by-order. These two plots would suggest that one needs a much higher (lower) imbalance in order for the best ask (bid) to increase (decrease) than decrease (increase). Gould and Bonart [59] found a similar, but very small, symmetry violation with volume imbalance as well. Focusing on the best ask, this could be because the best ask price can only increase if all shares at the best ask are completely removed. This would be caused by traders placing market orders and we see the best ask will almost only increase if there are few shares left at that price. The best ask can decrease from a limit order placed 1 tick below if the spread is at least 1 tick. It appears that it is simply easier to decrease prices (given the spread is not zero) than to increase them. The same argument applies to changes in the best bid price. This seems to be the case only when looking at price movements order-by-order. As we see in Figure 2.16b, with 5 second intervals, this pattern disappears.

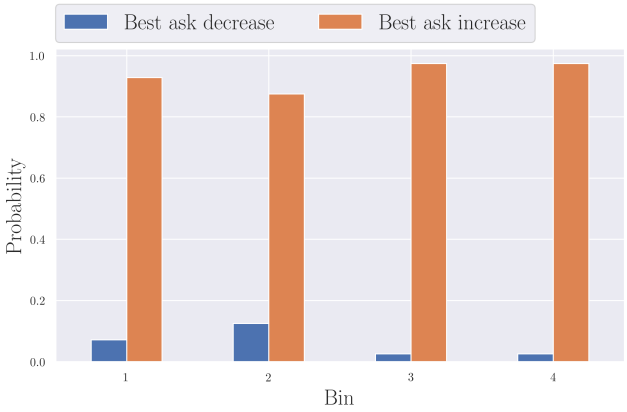
We have a different story when looking at price movements over 5 second intervals, but there is still a skew in the probabilities towards needing a larger imbalance for up movements than one needs a small imbalance for down movements. However, it is not as severe as it was order-by-order. The price does have a steadily increasing probability of moving up as we pass from bin 1 to 4. Likewise we have a steady decrease in the probability of a down movement as we pass from bin 1 to 4.

Before discussing the outcome of the statistical test on Tables 2.5 and 2.6 we take the opportunity to check a previously studied aspect of price movements in limit order books [61]. We take the counts in our two tables, but also condition on what the previous price movement was. The results are converted to probabilities like before and shown in Figures 2.18 and 2.19.

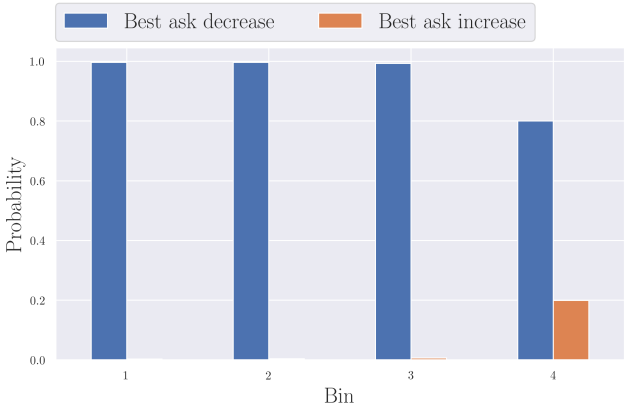
Order-by-order we see that if there was no previous price movement the probabilities are very similar to Figure 2.16a. This is likely because order-by-order there are almost no price movements as shown in Figure 2.14. The interesting result is that if the previous price movement was up, then the next movement is very likely to be down regardless of what the imbalance is. Similarly, if the previous price movement was down it is very likely that the next movement is up regardless of imbalance. This was discussed in [61], but independent of the volume imbalance. Here we see that the reverting aspect of prices happens regardless of the imbalance. When prices move, the next most likely move is in the opposite direction. We do not look at the magnitude of the price movement here though. Even though prices appear to be reverting we could have the best ask increase by 2 cents, but then drop 5 cents. This is not in conflict with Figures 2.18 and 2.19, but is not captured by them either.



(a) No previous price movement

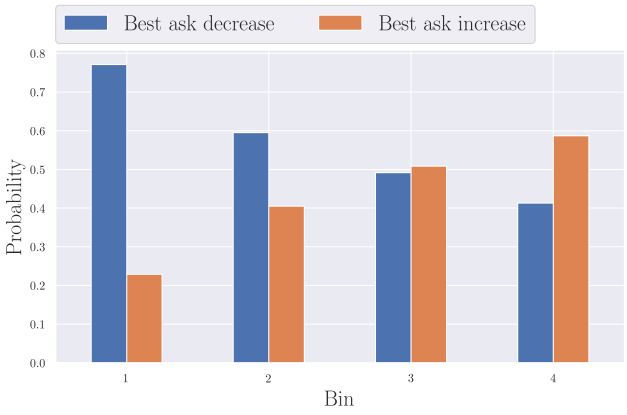


(b) Previous price move down

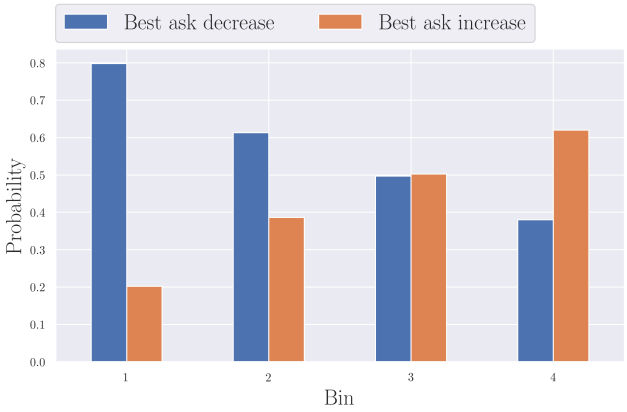


(c) Previous price move up

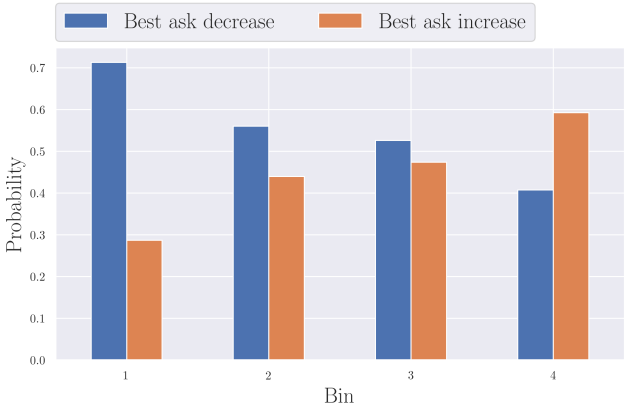
Figure 2.18: Probability of change in best ask price order-by-order being positive or negative conditioned on the average imbalance being in bins 1, 2, 3, or 4. Also conditioned on the last price movement being zero, down, or up. Data from AEM stock on April 17, 2017.



(a) No previous price movement



(b) Previous price move down



(c) Previous price move up

Figure 2.19: Probability of change in best ask price over 5 second intervals being positive or negative conditioned on the average imbalance being in bins 1, 2, 3, or 4. Also conditioned on the last price movement being zero, down, or up. Data from AEM stock on April 17, 2017.

The output of our statistical test is shown in Table 2.7. Like the first test we still get p-values which imply we can easily reject the null hypothesis at any significance level and that the relationship between the imbalance and price movements is positive. The main difference here is the strength of the relationship given by the Cramer's V measure. We get an extremely significant association order-by-order when we break the imbalance into bins compared to when we looked at just the sign of the imbalance in the previous subsection. We also get a stronger association over 5 second intervals, but it is much smaller than order-by-order. This could be because order-by-order we are only looking at the next price movement while over 5 seconds we are seeing an aggregate of many price moves and instantaneous imbalances. Some information is possibly lost along the way in how we choose to aggregate our data, but the positive correlation between the imbalance and price movements still exists. We should also note that the correlation between the change in the best ask price and the imbalance is the same in Tables 2.3 and 2.7 because they use the same sampled data.

	Cramer's V	p-value	Correlation
Order-by-order	0.676	0.0	0.123
5 second intervals	0.200	3.03e-126	0.129

Table 2.7: Summary of chi square test for fine imbalance and coarse price movements. Correlation is taken between change in best ask price and the imbalance.

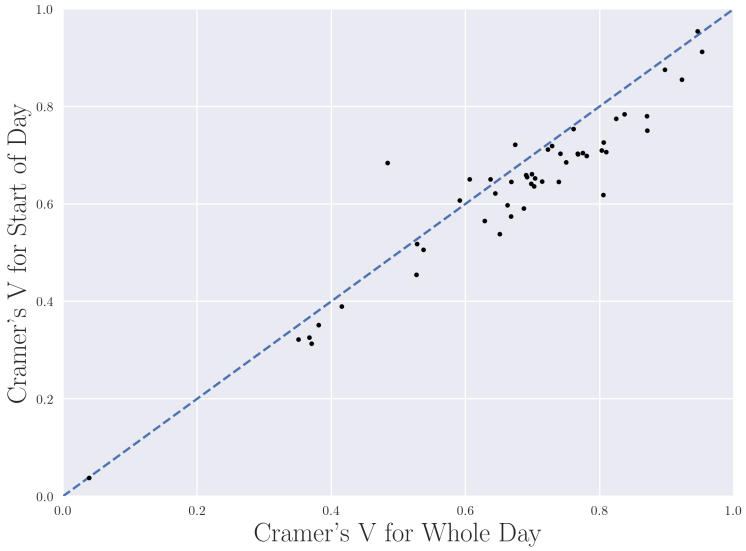


Figure 2.20: Cramer's V for whole day and start of day plotted against each other. Data is from Table 2.8. The dashed line would represent a totally linear relationship between the two statistics.

ticker	Cramer's V
ZEB	0.953
HOD	0.946
XEG	0.922
XWD	0.897
VUN	0.871
HSU	0.87
PWF	0.837
POW	0.824
VFV	0.81
GIL	0.806
FTS	0.805
PPL	0.803
SLF	0.78
XQQ	0.775
IPL	0.767
ARX	0.767
T	0.761
HXU	0.75
IMO	0.741
BAM.A	0.739
CCO	0.729
PVG	0.723
FM	0.714
G	0.703
CPG	0.702
ERF	0.699
HQU	0.698
NA	0.692
SSO	0.69
BMO	0.687
FR	0.674
PAAS	0.668
CNR	0.668
VGG	0.662
GIB.A	0.651
AEM	0.644
IMG	0.637
OR	0.629
HVI	0.606
HFU	0.591
RBA	0.537
KL	0.528
UFS	0.527
K	0.484
FSV	0.415
WCN	0.381
BIP.UN	0.371
GOOS	0.367
SW	0.351
TC	0.0387

ticker	Cramer's V
ZEB	0.912
HOD	0.954
XEG	0.855
XWD	0.875
VUN	0.75
HSU	0.78
PWF	0.784
POW	0.775
VFV	0.706
GIL	0.726
FTS	0.721
PPL	0.71
SLF	0.699
XQQ	0.704
IPL	0.702
ARX	0.703
T	0.754
HXU	0.685
IMO	0.703
BAM.A	0.645
CCO	0.719
PVG	0.712
FM	0.646
G	0.652
CPG	0.636
ERF	0.661
HQU	0.642
NA	0.655
SSO	0.659
BMO	0.591
FR	0.618
PAAS	0.645
CNR	0.574
VGG	0.597
GIB.A	0.538
AEM	0.622
IMG	0.651
OR	0.565
HVI	0.651
HFU	0.607
RBA	0.506
KL	0.517
UFS	0.454
K	0.684
FSV	0.389
WCN	0.352
BIP.UN	0.314
GOOS	0.326
SW	0.322
TC	0.038

Table 2.8: Summary of chi square test for fine imbalance and coarse price movements order-by-order. Data is taken from June 1-8, 2017. All p-values are zero or very close ($\approx 10^{-73}$ at most) to zero. The left subtable uses data from the entire trading day while the right subtable uses data only from the first hour of the trading day. Tickers are sorted by magnitude of Cramer's V for the whole day.

In Table 2.8 we report the same test for a large sample of stocks. As in the previous subsection, we only run the same test on order-by-order data for other stocks as we still need a way of comparing stocks which have very different behaviour over the same time interval. All p-values imply we can reject the null hypothesis at any significance level and we have a very strong association between the binned imbalance and price direction. Figure 2.20 shows a similar relationship as in Figure 2.15 except the Cramer's V during the start of day is usually less than the whole day for this test. This may suggest a weaker association during the start of day between the imbalance and price movements, but the association is still strong and the difference between the Cramer's V in both time periods is small.

2.5 Conclusions

In this chapter we investigated the impact of time on price movements and how the volume imbalance ratio can be used as a predictor of future price movements. We saw that individual stocks have their own time scales over which we see larger and larger price movements. Care must be taken to determine a way to compare all stocks given that they may not have similar behaviour at the time scales. We need a way of fixing the time scale on all stocks so that we can compare any future analysis in an 'apples to apples' way. In chapter 4 we shall explore this comparison between model parameters.

We also saw the problem of dealing with how to aggregate the instantaneous volume imbalance over a given time interval. We gave two ways of doing this - by taking the mean and also weighting by the time between successive orders. The analysis that followed only used the time weighting, but we will come back to this in chapter 4 once we have discussed how to fix the time intervals for each stock.

We used hypothesis testing in two different ways to test how the volume imbalance ratio impacts price movements:

1. How the sign of the imbalance influences specific price movements

2. How the value of the imbalance influences price direction

In both cases we found a strong positive relationship order-by-order and also over 5 second intervals for AEM stock on April 17, 2017. We also ran the same tests over a larger sample of stocks order-by-order and found the same strong relationship in both tests. All tests showed the association between the volume imbalance ratio and price changes was statistically significant, but the Cramer's V allows us to numerically represent the

strength of that association. We will use the Cramer's V to compare models under different parameters once we define a generalized volume imbalance ratio in chapter 3. The generalized form of the volume imbalance ratio that incorporates volumes beyond the best bid and best ask. Improvements in the Cramer's V between models would suggest a stronger association under different choices in how we calculate the volume imbalance.

These results and the existing literature support the use of the volume imbalance, and by extension the volumes in the limit order book, to predict the possible future price movements. In chapter 3 we look at how we incorporate these ideas into a model for determining price changes based on the orders that interact with the limit order book. We can use this model to calculate the costs associated with a spoofer's decision to manipulate the limit order book. The calibration of our price change model is done in chapter 4 along with comparisons between the resulting parameters. In chapter 5 we can use our calibrated models to explore the sensitivity of the limit order book to spoofing and the conditions under which a spoofer decides to manipulate the book.

Chapter 3

Spoofing Cost Model and Generalized Imbalance Ratio

3.1 Introduction

In this chapter we develop the notation and mathematical description necessary to analyze the limit order book. With this notation in place we can build a model for determining the optimal limit order placement to manipulate the best ask price through the volume imbalance ratio. We saw in the previous chapter that the imbalance has a statistically meaningful impact on the distribution of the change in best ask price. The spoofer will place their spoofing limit orders in order to minimize the cost associated with their spoofing strategy. The spoofer's strategy will be made exact in this chapter. The resulting optimal limit order placement will give us a starting point for detecting manipulation - we will know where to look in the limit order book for spoofing orders.

We start this chapter by defining some initial notation before deriving cost functions associated with a trader wishing to purchase a number of shares of stock for three cases:

1. Placing a market order for the shares immediately
2. Delaying their market order to see where the best ask moves then placing a market order
3. Placing a limit order to impact the imbalance which in turn impacts the possible price movements before placing a market order

The third case will necessitate a model for how the imbalance directly changes the distribution of the change in best ask price.

We also saw in the previous chapter that our definition of the volume imbalance ratio was based only on the volume of shares at the best ask and best bid prices. We will generalize this definition to include all prices on both sides of the book with appropriate weights assigned to the volumes. We ultimately want to allow the weights at each depth of the limit order book to be free parameters[1] in our model, but we also explore the use of exponential weights. This is because exponential weights serve as a middle ground between free weights and the classic definition of all weight assigned to the best bid and ask. Using either choice of generalized imbalance will allow us to calibrate a model in which a spoofer can manipulate prices by placing limit orders anywhere in the book - instead of only at the touch.

We will develop a price change model which uses the volume imbalance ratio as the weight in a convex combination of two price change distributions: the price change distributions conditional on I as $I \to \pm 1$. We will use this model to calculate the expected costs of the three decisions listed above to determine what a spoofer would do given a state of the limit order book. The algorithm we use to calibrate our price change model will be discussed in chapter 4 and we implement it in chapter 5 to explore spoofing detection.

3.2 Spoofing Cost Model

3.2.1 Notation and Definitions

We start by assigning notation to the volumes and prices of the limit order book as well as the movements in the best ask price.

Let $\vec{v}_t = [v_{-K}, \ldots, v_K] \in \mathbb{R}^{2K+1}_{\geq 0}$ be the volumes in the limit order book at time t with $v_0 = 0$. We suppress the index t in the components of $\vec{v}_t$ for ease of notation. Then, v_1 is the number of shares sitting at the best ask price p_t^+ at time t and v_{-1} is the number of shares sitting at the best bid price p_t^- at time t. Prices $p_t^\pm$ are in Canadian dollars (CAD). K denotes the number of prices we include from each side of the book so that the volume v_K is located at price $p_t^+ + \frac{K-1}{100}$ (for ask prices). Another way of saying this is that the volume v_K is $K-1$ ticks from the best ask. v_{-K} is then $-K+1$ ticks from the best bid.

Let $\varphi(x_t; I_t) = \mathbb{P}\left[p^+_{t+\Delta t} = p_t^+ + \frac{x_t}{100} | I_t\right]$ denote the probability of the change in best ask price moving $x_t \in \mathbb{Z}$ ticks from time t to time $t + \Delta t$ conditional on the average

[1] We will enforce a constraint where the weight assigned to the best bid and ask is at least as large as any other weight.

imbalance over Δt being I_t.

This is a one period model from time t to $t + \Delta t$ where the spoofer will act at either time, but not in between. We assume that the corresponding price process p_t is right-continuous on a given filtered probability space $(\Omega, \mathcal{F}_t, \mathbb{P})$ where $\mathcal{F}_t$ is the natural filtration of the stochastic processes $\Delta p_t^+ = x_t/100$, Δp_t^-, and limit order book volumes $\vec{v}_t$. However, for our work we will not need the process Δp_t^-. We use the notation that, for example, $p_{t+\Delta t} = p_{t+1}$. The price process is defined as

$$p_{t+1}^+ = p_t^+ + x_t/100\,, \tag{3.2.1}$$

or

$$\Delta p_t^+ = (p_{t+1}^+ - p_t^+) = x_t/100\,, \tag{3.2.2}$$

where the expected value of p_{t+1}^+ given $\mathcal{F}_t$ and I_t is

$$\begin{aligned} E[p_{t+1}^+|\mathcal{F}_t, I_t] &= \sum_i (p_t^+ + i)\mathbb{P}\left[x_t = i|\mathcal{F}_t, I_t\right] \\ &= p_t^+ + \sum_i i\varphi(i; I_t) \end{aligned} \tag{3.2.3}$$

So, for ease of notation, we take $p_t = p^+$ at the start of the period with random variable $p_{t+1} = p^+ + x_t/100$ denoting the best ask price at the end of the period.

Denote the volume of shares on the ask side of the order book N ticks from the best ask price as:

$$V^+(\vec{v}_t, N) = \sum_{i=1}^{N} v_i\,. \tag{3.2.4}$$

Denote the volume of shares on the bid side of the order book N ticks from the best bid price as:

$$V^-(\vec{v}_t, N) = \sum_{i=-N}^{-1} v_i\,. \tag{3.2.5}$$

Define the generalized inverse of $V^\pm$ by

$$U^\pm(\vec{v}_t, H) = \inf\left\{i : V^\pm(\vec{v}_t, i) \geq H\right\} \tag{3.2.6}$$

The financial intuition here is, given a limit order book $\vec{v}_t$, $U^\pm(\vec{v}_t, H)$ gives the limit order book index i necessary to satisfy a market order of at least H shares.

Case 1: Immediate Market Order At time t we could to decide to place a market order based on $\vec{v}_t$ and p_t^+ which would be made immediately after at time t^+. This is represented in Figure 3.1.

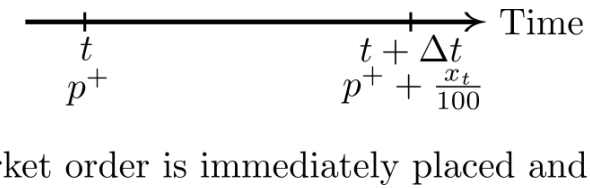


Figure 3.1: We see the limit order book at time t and decide to place our market order. This is sent immediately after at time t^+.

If we place a market order for the H shares immediately there may be fewer than H shares available at the best ask price. We then need to pay a premium for walking the book. We saw this example in Figure 1.8. Let $C_{\text{MO}}(\vec{v}_t, H, p^+)$ denote the cost of placing the market order immediately. We choose to simplify the problem by ignoring the time lag between placement and execution of the market order. It is possible that the price could move against you after placing the order because of other orders in the queue. This problem is explored in [64].

If there are at least H shares available at v_1 then the cost we pay is p^+H. However, if $H > v_1$ we need to walk the book and remove liquidity at increasing prices until we have all H shares. We can satisfy our market order by removing liquidity up to and including index $U^+(\vec{v}_t, H)$. For now we will write $U^+(\vec{v}_t, H) = U$ for clarity. So we pay $(p^+ + \frac{i-1}{100})v_i$ at each price tick $i \in [1, U)$ and purchase the remaining shares $H - v_1 - \cdots - v_{U-1}$ at price $p^+ + \frac{U-1}{100}$. The associated cost would then be:

$$\begin{aligned} C_{\text{MO}}(\vec{v}_t, H, p^+) &= p^+ v_1 + \left(p^+ + \frac{1}{100}\right) v_2 + \left(p^+ + \frac{2}{100}\right) v_3 \cdots + \\ &\qquad \left(p^+ + \frac{U-1}{100}\right)(H - v_1 - \cdots - v_{U-1}) \\ &= \sum_{i=1}^{U-1} \left(p^+ + \frac{i-1}{100}\right) v_i + \left(p^+ + \frac{U-1}{100}\right)\left(H - \sum_{i=1}^{U-1} v_i\right) \end{aligned} \tag{3.2.7}$$

We can expand the sums in equation 3.2.7 and group terms to get

$$\begin{aligned}
C_{\mathrm{MO}}(\vec{v}_t, H, p^+) &= \cancel{p^+ \sum_{i=1}^{U-1} v_i} + \sum_{i=1}^{U-1} \frac{i-1}{100} v_i + p^+ H - \cancel{p^+ \sum_{i=1}^{U-1} v_i} + \\
&\quad \frac{U-1}{100} H - \frac{U-1}{100} \sum_{i=1}^{U-1} v_i \\
&= \frac{1}{100} \sum_{i=1}^{U-1} i v_i - \cancel{\frac{1}{100} \sum_{i=1}^{U-1} v_i} + p^+ H + \frac{U-1}{100} H - \\
&\quad \frac{U}{100} \sum_{i=1}^{U-1} v_i + \cancel{\frac{1}{100} \sum_{i=1}^{U-1} v_i} \\
&= p^+ H + \frac{U-1}{100} H + \frac{1}{100} \sum_{i=1}^{U-1} i v_i - \frac{U}{100} \sum_{i=1}^{U-1} v_i \\
&= \underbrace{p^+ H}_{\text{Market Order}} + \underbrace{G(\vec{v}_t, H)}_{\text{Penalty}}
\end{aligned} \tag{3.2.8}$$

where

$$G(\vec{v}_t, H) = \frac{1}{100} \left[(U^+(\vec{v}_t, H) - 1) H - \sum_{i=1}^{U^+(\vec{v}_t, H)-1} (U^+(\vec{v}_t, H) - i) v_i \right] \tag{3.2.9}$$

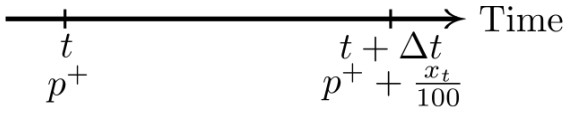
So we have broken $C_{\mathrm{MO}}(\vec{v}_t, H, p^+)$ into two pieces: the cost if we could have bought all H shares at the best ask price p^+ plus a penalty paid for each share purchased beyond the best ask denoted by $G(\vec{v}_t, H)$. Note that the penalty term does not depend on the best ask price p^+.

Since we know $\vec{v}_t$, p^+, and I_t at time t then $C_{\mathrm{MO}}(\vec{v}_t, H, p^+)$ is known at time t and its expected value is simply the value of the function. That is,

$$E\left[C_{\mathrm{MO}}(\vec{v}_t, H, p^+) | \mathcal{F}_t, I_t\right] = C_{\mathrm{MO}}(\vec{v}_t, H, p^+) \tag{3.2.10}$$

Case 2: Delayed Market Order Instead of placing our market order immediately we delay it one time period Δt, so the market order is placed immediately before the end of the period at time $t^- + \Delta t$. This is represented in Figure 3.2.

To find the cost associated with making this decision at time t requires an assumption or approximation for the shape of the order book at time $t + \Delta t$. That is, we do not know what the shape of the book will be at time $t + \Delta t$ and we need to know this to calculate the cost of a market order placed at this time. The simplest modelling assumption would



$t^- + \Delta t$: Market order to be executed at the end of the time interval

Figure 3.2: We see the limit order book at time t and decide to delay our market order to time $t + \Delta t$. This is sent at time $t^- + \Delta t$ just before the end of the period.

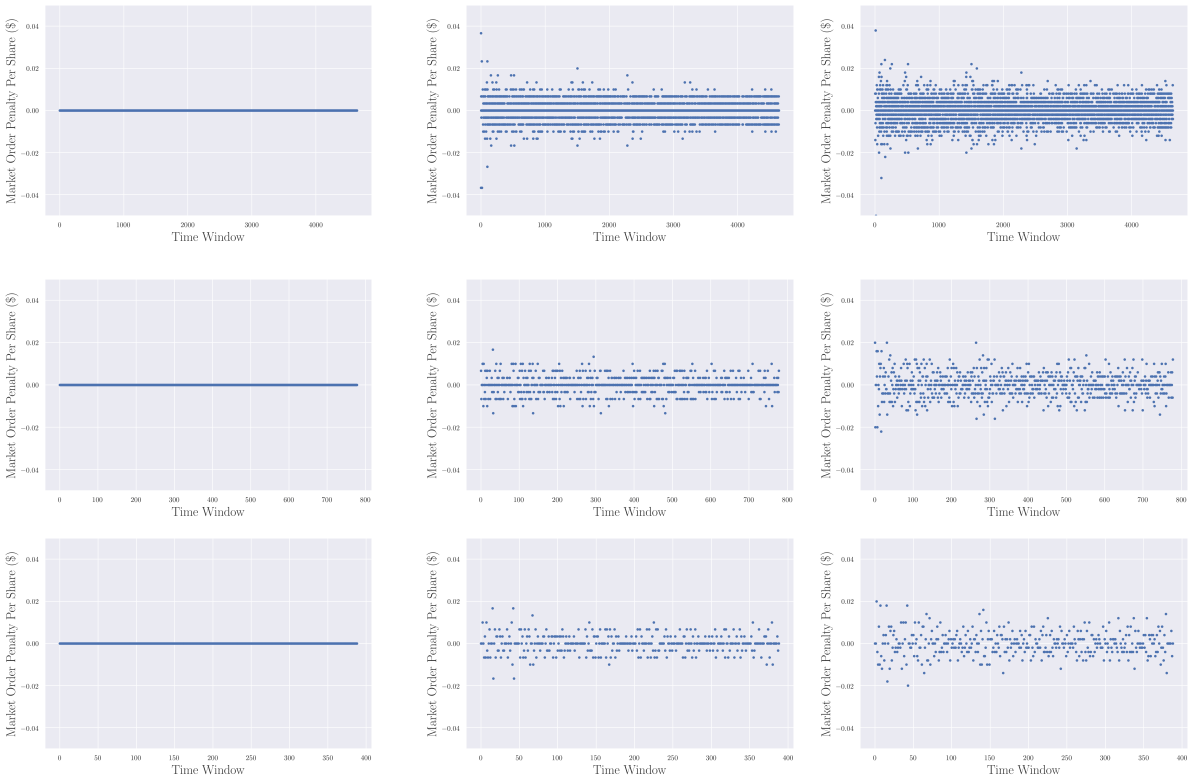
be that the volumes at each price are the same at both times and translate with the price movement x_t. This is unlikely to be true for the entire order book, but if we only intend to spoof to buy a small number of shares H we only need to know how much our penalty $G(\vec{v}_t, H)$ changes over the trading day for different H. If this effect is small we say we can approximate the volumes at the first few ticks in the future by the volumes we see now.

Figure 3.3 shows the impact of purchasing H shares over the entire trading day by the change in cost of walking the book $[G(\vec{v}_t, H) - G(\vec{v}_{t+1}, H)]/H$ between 5, 30, and 60 second time intervals. Note that $G(\vec{v}_t, H)/H$ is a per share cost measure. We see that the difference in the penalty per share is distributed roughly the same over the day with the exception of a few larger penalties at the beginning of the day. This is to be expected before the limit order book fills up. We also see no penalty for small orders of 100 shares which is not surprising since the best ask will always have at least 100 shares.[2] The penalty also remains symmetric about 0 indicating you are just as likely to benefit as be harmed by the assumption that the volume remains the same in the first few ticks. The penalty is almost entirely bounded between ± 0.02\$ per share. We will ignore this for now, but one could set $G(\vec{v}_t, H) \to G(\vec{v}_t, H) \pm \gamma H$ where $\gamma = 0.02$ is an uncertainty factor for our approximation and we could calculate the best and worst case scenarios for the cost function under our assumption. Or we could just limit our analysis to orders of 100 shares to completely avoid worrying about this assumption, as shown in Figure 3.3 (a).

We see similar results for ARX stock on April 17, 2017 in Figure D.8. Here we see a tighter bound even for $H = 1000$ shares, but the main result that the penalty is bounded and symmetric still holds. The price for ARX stock was also about \$ 18 CAD while AEM was about \$ 46 CAD. The penalty per share is roughly that of AEM even though the stock price is over twice as much. We see similar results across other stocks with the penalty being bounded by $\pm$\$0.02 per share.

We denote the cost of a delayed market order by $C_{\text{DMO}}(\vec{v}_t, H, p^+, x_t)$ where the price

[2] 100 shares is the minimum order size for the book.



(a) $H = 100$ (b) $H = 300$ (c) $H = 500$

Figure 3.3: Difference in $G(\vec{v}_t, H)/H$ between different time intervals and H throughout the trading day. The top, middle, and bottom subplots correspond to time intervals of 5, 30, and 60 seconds, respectively. Data taken from AEM stock on April 17, 2017 for the entire trading day. Stock price $\approx$ \$46 CAD.

movement x_t is in ticks. Since we are assuming the limit order book translates with the price and the penalty term $G(\vec{v}_t, H)$ does not depend on p^+ the cost is

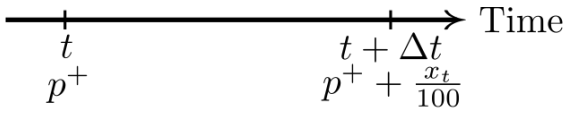
$$\begin{aligned} C_{\text{DMO}}(\vec{v}_t, H, p^+, x_t) &= \left(p^+ + \frac{x_t}{100}\right) H + G(\vec{v}_t, H) \\ &= p^+ H + G(\vec{v}_t, H) + \frac{H}{100} x_t \\ &= C_{\text{MO}}(\vec{v}_t, H, p^+) + \underbrace{\frac{H}{100} x_t}_{\text{Impact of Delay}} \end{aligned} \tag{3.2.11}$$

So when the best ask moves x_t ticks we pick up an extra term which can work for or against us. If $x_t < 0$ we benefit for waiting while we lose if $x_t > 0$. If $x_t = 0$ then case 1 and 2 have the same outcome.

For a time interval Δt, the expected cost for the strategy at time t of delaying our market order to time $t + \Delta t$ is

$$\begin{aligned} E\left[C_{\text{DMO}}(\vec{v}, H, p^+, x_t)|\mathcal{F}_t, I_t\right] &= \sum_i C_{\text{DMO}}(\vec{v}_t, H, p^+, i)\varphi(i; I_t) \\ &= \sum_i \left[C_{\text{MO}}(\vec{v}_t, H, p^+) + \frac{H}{100} i\right] \varphi(i; I_t) \\ &= C_{\text{MO}}(\vec{v}_t, H, p^+) + \frac{H}{100} \sum_i i\, \varphi(i; I_t) \\ &= C_{\text{MO}}(\vec{v}_t, H, p^+) + \frac{H}{100} E\left[x_t|\mathcal{F}_t, I_t\right] \end{aligned} \tag{3.2.12}$$

Case 3: Spoofing With Market Order However, a trader can spoof the order book by placing limit orders at time t^+ on the ask side of the book with a hope of lowering the best ask price and then, at time $t^- + \Delta t$, placing a market order and canceling the previous limit orders. This is represented in Figure 3.4.



t^+ : Spoofing orders are placed
$t^- + \Delta t$: Cancel remaining spoofing orders
$t^- + \Delta t$: Market order to be executed at the end of the time interval

Figure 3.4: We see the limit order book at time t and decide to spoof the book. Our spoofing orders are sent immediately after at time t^+. We then cancel our remaining limit orders and place a market order at time $t^- + \Delta t$.

Denote the volume of the limit orders by $\tilde{v}_t = [\tilde{v}_{-K}, \ldots, \tilde{v}_K] \in \mathbb{R}^{2K+1}_{\geq 0}$ and the change in the best ask price as x_t in ticks. Our limit orders $\tilde{v}$ impact the volume imbalance ratio $I(\vec{v}_t + \tilde{v}_t)$. We suppress the dependency of I_t on $\tilde{v}_t$ now for clarity. We define $\tilde{v}_t$ this way so we can add $\vec{v}_t + \tilde{v}_t$ to get the limit order book plus our spoofing orders, but we are limiting our orders to the ask side of the book based on the analysis in chapter 2.

The trade-off of this strategy is that our limit orders can be executed and we would need to buy more shares to cover our limit orders as in Figure 1.11. We will assume that if the best ask price moves up $x_t > 0$ ticks then all of our spoofing limit orders $\tilde{v}_1, \tilde{v}_2, \tilde{v}_3, \ldots, \tilde{v}_{x_t}$ are executed and we need to cover the shares we sold. To get a more conservative estimate for the cost of spoofing we allow our limit orders at $\tilde{v}_{x_t}$ to be executed if the best ask price moves up x_t ticks, but this may not be the case in reality. This means that we are assuming any limit order we place at $\tilde{v}_{x_t+1}$ and above will not be executed before we can cancel it. The total number of shares we would need to purchase

would be $H + V^+(\tilde{v}_t, x_t)$ as our limit orders are all on the ask side.

Let $C_{\mathrm{S}}(\vec{v}_t, H, p^+, x_t, \tilde{v}_t)$ be the cost of placing limit orders $\tilde{v}_t$ at time t then canceling all limit orders at time $t + \Delta t$ and placing a market order for $H + V^+(\tilde{v}_t, x_t)$ shares after the besk ask price has moved x_t ticks. The cost of the market order would be $C_{\mathrm{DMO}}(\vec{v}_t, H + V^+(\tilde{v}_t, x_t), p^+, x_t)$ and then we get paid $-(p^+ + i)\tilde{v}_{i+1}$, $i \in [0, x_t]$, for each limit order that was executed between 0 and x_t ticks, inclusive, from the best ask price. Putting all this together we get

$$
\begin{aligned}
C_{\mathrm{S}}(\vec{v}_t, H, p^+, x_t, \tilde{v}_t) &= C_{\mathrm{DMO}}(\vec{v}_t, H + V^+(\tilde{v}_t, x_t), p^+, x_t) - \\
&\qquad \sum_{i=0}^{x_t} \left(p^+ + \frac{i}{100} \right) \tilde{v}_{i+1} \\
&= \left(p^+ + \frac{x_t}{100} \right) \left(H + V^+(\tilde{v}_t, x_t) \right) + \\
&\qquad G(\vec{v}_t, H + V^+(\tilde{v}_t, x_t)) - p^+ V^+(\tilde{v}_t, x_t) - \\
&\qquad \frac{1}{100} \sum_{i=0}^{x_t} i \, \tilde{v}_{i+1} \\
&= p^+ H + \cancel{p^+ V^+(\tilde{v}_t, x_t)} + \frac{H}{100} x_t + \\
&\qquad \frac{1}{100} x_t V^+(\tilde{v}_t, x_t) + G(\vec{v}_t, H + V^+(\tilde{v}_t, x_t)) - \\
&\qquad \cancel{p^+ V^+(\tilde{v}_t, x_t)} - \frac{1}{100} \sum_{i=0}^{x_t} i \, \tilde{v}_{i+1} \\
&= p^+ H + G(\vec{v}_t, H + V^+(\tilde{v}_t, x_t)) + \frac{H}{100} x_t + \\
&\qquad \frac{1}{100} \sum_{i=0}^{x_t} (x_t - i) \, \tilde{v}_{i+1} \\
&= p^+ H + G(\vec{v}_t, H + V^+(\tilde{v}_t, x_t)) + \\
&\qquad \frac{H}{100} x_t + \underbrace{C_{\mathrm{LO}}(\tilde{v}_t, x_t)}_{\text{Adjusted Buyback Cost}} ,
\end{aligned}
\tag{3.2.13}
$$

where

$$
C_{\mathrm{LO}}(\tilde{v}_t, x_t) = \frac{1}{100} \sum_{i=0}^{x_t} (x_t - i) \, \tilde{v}_{i+1} \, . \tag{3.2.14}
$$

$C_{\mathrm{LO}}(\tilde{v}_t, x_t)$ is an adjustment to the net cost associated with selling spoofing orders $\tilde{v}_{i+1}$ at price $p^+ + i/100$ then buying them back at price $p^+ + x_t/100$. The function $G(\vec{v}_t, H + V^+(\tilde{v}_t, x_t))$ then takes into account the cost of walking the book for the H

shares the spoofer intends to buy, plus the total executed spoofing orders $V^+(\tilde{v}_t, x_t)$.

For a time interval Δt, the expected cost for the strategy at time t after manipulating $I_t = I(\vec{v}_t + \tilde{v}_t)$ is

$$\begin{aligned} E\left[C_{\mathrm{S}}(\vec{v}_t, H, p^+, x_t, \tilde{v}_t)|\mathcal{F}_t, I_t\right] &= p^+ H + \sum_i G(\vec{v}_t, H + V^+(\tilde{v}_t, i))\varphi(i; I_t) + \\ &\quad \frac{H}{100}\sum_i i\varphi(i; I_t) + \sum_i C_{\mathrm{LO}}(\tilde{v}_t, i)\varphi(i; I_t) \\ &= p^+ H + E\left[G(\vec{v}_t, H + V^+(\tilde{v}_t, x_t))|\mathcal{F}_t, I_t\right] + \\ &\quad \frac{H}{100}E\left[x_t|\mathcal{F}_t, I_t\right] + E\left[C_{\mathrm{LO}}(\tilde{v}_t, x_t)|\mathcal{F}_t, I_t\right] \end{aligned} \tag{3.2.15}$$

Note the cases that if $x_t \leq 0$ or $\tilde{v}_t = \vec{0}$ then $C_{\mathrm{S}}(\vec{v}_t, H, p^+, x_t, \tilde{v}_t) = C_{\mathrm{DMO}}(\vec{v}_t, H, p^+, x_t)$ as $V^+(\tilde{v}_t, x_t) = 0$ and $C_{\mathrm{LO}}(\tilde{v}_t, x_t) = 0$.

3.2.2 When to Spoof?

For our model, the limit order book $\vec{v}_t$ admits a spoofing strategy $\tilde{v}_t$ if the expected cost of spoofing is less than the expected cost of a delayed market order or an immediate market order. That is, using equations 3.2.10 and 3.2.12, we delay our market order if

$$\begin{aligned} & E\left[C_{\mathrm{DMO}}(\vec{v}_t, H, p^+, x_t)|\mathcal{F}_t, I_t\right] < E\left[C_{\mathrm{MO}}(\vec{v}_t, H, p^+)|\mathcal{F}_t, I_t\right] \\ \implies & \cancel{C_{\mathrm{MO}}(\vec{v}_t, H, p^+)} + \frac{H}{100}E\left[x_t|\mathcal{F}_t, I_t\right] < \cancel{C_{\mathrm{MO}}(\vec{v}_t, H, p^+)} \\ \implies & E\left[x_t|\mathcal{F}_t, I_t\right] < 0 \end{aligned} \tag{3.2.16}$$

where the final line is because $H > 0$. The result here is easy enough to understand – we delay our market order if we expect the best ask price to drop.

We adopt the following notation to differentiate between expected values taken with respect to $\varphi(x_t; I(\vec{v}_t))$ and $\varphi(x_t; I(\vec{v}_t + \tilde{v}_t))$.

$$\begin{aligned} E\left[\,\cdot\,\right] &= E\left[\,\cdot\,|\mathcal{F}_t, I(\vec{v}_t)\right] \\ \tilde{E}\left[\,\cdot\,\right] &= E\left[\,\cdot\,|\mathcal{F}_t, I(\vec{v}_t + \tilde{v}_t)\right] \end{aligned} \tag{3.2.17}$$

We should note that both expectation values are taken at time t, when we must make the decision on where to place our spoofing orders $\tilde{v}_t$. With this notation we can write down the condition for spoofing over delaying our market order as

$$\begin{aligned} &\tilde{E}\left[C_{\mathrm{S}}(\vec{v}_t, H, p^+, x_t, \tilde{v}_t)\right] - E\left[C_{\mathrm{DMO}}(\vec{v}_t, H, p^+, x_t)\right] < 0 \\ \implies &\tilde{E}\left[G(\vec{v}_t, H + V^+(\tilde{v}_t, x_t))\right] + \tilde{E}\left[C_{\mathrm{LO}}(\tilde{v}_t, x_t)\right] + \\ &\qquad \frac{H}{100}\left(\tilde{E}\left[x_t\right] - E\left[x_t\right]\right) - G(\vec{v}_t, H) < 0 \end{aligned} \tag{3.2.18}$$

Equation 3.2.18 is not as simple as 3.2.16 since we choose where we place our limit orders $\tilde{v}$. We spoof if it is possible to place our limit orders such that the expected cost of our executed limit orders and subsequent market orders is less than the penalty we take for walking the book with a single market order. We also need to consider the difference in the expected change in the best ask price under both measures. There is a trade off here with $\tilde{v}$: the larger the limit orders the more we can influence the imbalance, but the larger penalty we pay if the price moves against us and those limit orders are executed. We need to find the optimal $\tilde{v}_t$ which balances these opposing forces or pushes the payoff in our favour.

Finally, we can combine equations 3.2.16 and 3.2.18 to get the condition where we would spoof over immediately placing a market order. This gives us

$$\tilde{E}\left[G(\vec{v}_t, H + V^+(\tilde{v}_t, x_t))\right] + \tilde{E}\left[C_{\mathrm{LO}}(\tilde{v}_t, x_t)\right] + \frac{H}{100}\tilde{E}\left[x_t\right] - G(\vec{v}_t, H) < 0 \tag{3.2.19}$$

Equations 3.2.18 and 3.2.19 are very similar except in equation 3.2.19 we gain the entire benefit of the expected change in the best ask price instead of the difference under both measures. This reflects the fact it may be better to spoof over immediately placing a market order, but spoofing may yield no real advantage over simply delaying your market order.

With our spoofing model and equations developed we are left now with determining a model for $\varphi(x_t; I_t)$. We look first at how we can impact I_t through $\tilde{v}_t$. In chapter 2 we used a definition for the volume imbalance ratio that depended only on the volume at the best ask and best bid. We want to be able to impact the distribution of the change in the best ask through the imbalance at all price ticks, hence, we need to generalize our definition of the imbalance ratio to include the entire order book – not just the touch.

3.3 Generalized Imbalance Ratio

We can rewrite equation 2.3.1 for the volume imbalance ratio I_t with our new notation as

$$I_t = I(\vec{v}_t) = \frac{v_{-1} - v_1}{v_{-1} + v_1} \tag{3.3.1}$$

The problem is that we can only impact the imbalance with v_{-1} and v_1, but we want all orders to influence the imbalance. We saw in Figure 1.11 what can happen if a spoofer places their limit orders too close to the touch. Without a generalized definition of the imbalance we can only influence price movements at the touch, so we define the following as the generalized volume imbalance ratio

$$\begin{aligned} I(\vec{v}_t; \vec{w}, K) &= \frac{\sum_{i=1}^{K} w_i v_{-i} - \sum_{i=1}^{K} w_i v_i}{\sum_{i=1}^{K} w_i v_{-i} + \sum_{i=1}^{K} w_i v_i} \\ &= \frac{\sum_{i=1}^{K} w_i (v_{-i} - v_i)}{\sum_{i=1}^{K} w_i (v_{-i} + v_i)}, \end{aligned} \tag{3.3.2}$$

where $K \geq 1$ denotes the number of price increments (ticks) we include on both sides of the book and $\vec{w} = [w_1, \ldots, w_K] \in \mathbb{R}^K_{\geq 0}$ and $\sum_{i=1}^{K} w_i = 1$. The w_i is the weight assigned to the imbalance $v_{-i} - v_i$ at tick i. Dividing the sum of these weighted imbalances by the total weighted volume gives us a volume imbalance ratio between -1 and 1 as before. Equation 3.3.2 then represents a family of definitions that depend on the choice of parameters w and K.

For our work we wish to use free weights for $\vec{w}$ which we calibrate to our data, but we also analyze an intermediate choice for the weights that lies between all weight applied to the touch (classic definition) and fully free weights – exponentially decaying weights. The added benefit of calibrating with the exponential weights before the free weights is that it gives us a numerical check for the free weight results – the free weights should provide calibration results at least as good as the exponential weights. This is because the free weights would be able to reproduce the results of the exponential weights if it were the better choice for the imbalance weights. This statement is made more precise in the next chapter.

The exponentially decaying weights are defined as $w_i = \exp(-(i-1)\alpha)$ for some constant $\alpha \in [0, \infty)$. That is, the significance of each volume imbalance diminishes exponentially as you move deeper into the limit order book with the highest weight assigned to the best ask and best bid. This is a natural extension to the case where all weight is applied to the touch, i.e. $w_i = \delta_{1i}$ and δ_{ij} is the Kronecker delta function. To ensure the free weights also assign the highest weight to the touch we add the constraint $w_1 \geq w_k \quad \forall k \in [1, K]$ for its calibration. Table 3.1 displays all necessary information about the three choices for the imbalance weights.

Weight Name	Definition
Only Touch (Classic)	$w_i = \delta_{1i}$ $\quad \delta_{ij} =$ Kronecker Delta
Exponential	$w_i = \exp(-(i-1)\alpha)$ $\quad \alpha \in [0, \infty)$
Free	w_i subject to $w_1 \geq w_k$ $\quad \forall k \in [1, K]$

Table 3.1: Three choices for imbalance weights $\vec{w}$ used to calibrate our model.

Equipped with a general definition we can now write down a model for the distribution of the change in the best ask price dependent on the volume imbalance ratio $I(\vec{v}; \vec{w}, K)$.

3.4 Price Change Distribution Model

With a general volume imbalance ratio we can describe the model we use for $\varphi(x_t; I_t)$ as influenced by orders placed at the best bid/ask and beyond. Remember from the previous section that the imbalance is now parameterized by the weights w and parameter $K \geq 1$ such that $I = I(\vec{v}_t; \vec{w}, K)$. We note that $\vec{w}$ and K are fixed in the model and do not depend on time. Let $dp \in \mathbb{R}^{2K-1}_{\geq 0}$ be a vector of positive real numbers dp_x, $x \in [-K+1, K-1]$, such that

$$dp = \{dp_{-K+1}, \ldots, dp_0, \ldots, dp_{K-1}\}, \quad \sum_{x=-K+1}^{K-1} dp_x = 1 \quad \text{and} \tag{3.4.1}$$

$$\mu = \sum_{x=-K+1}^{K-1} x\, dp_x \,. \tag{3.4.2}$$

We can then view dp as a probability distribution on the discrete support $x \in [-K+1, K-1]$. The reason we do not take the support $x \in [-K, K]$ is because all volumes v_k, $1 \leq k \leq K$, should be necessary to determine dp_{K-1}. If the best ask were to move up 2 ticks we would expect the volumes up to 2 ticks from the best ask to be important in determining the probability of that movement since all these volumes would need to be depleted for the best ask to move up more than 2 ticks.

From chapter 2 we saw that the distribution of the change in best ask price remains peaked at zero, but the probability skews left when $I \to -1$, and skews right when $I \to 1$. Using equation 3.4.1 we define two distributions dp^+ and dp^- with expected values μ_- and μ_+, respectively. We want $dp \to dp^+$ as $I \to 1$ and $dp \to dp^-$ as $I \to -1$. That is, dp^+ and dp^- are the distributions for the change in the best ask price as I approaches 1 and -1, respectively. The simplest way to do this is to take $\varphi(x; I)$ as the convex combination of dp^+_x and dp^-_x. Our unnormalized model for the distribution of the change

in the best ask price $\varphi(x; I)$ is then

$$\tilde{\varphi}(x; I) = (I + 1)\, dp_x^+ + (1 - I)\, dp_x^- \tag{3.4.3}$$

with normalization constant

$$\begin{aligned} \sum_x \tilde{\varphi}(x; I) &= \sum_x (I + 1)\, dp_x^+ + \sum_x (1 - I)\, dp_x^- \\ &= (I + 1) \sum_x dp_x^+ + (1 - I) \sum_x dp_x^- \\ &= I + 1 + 1 - I \\ &= 2 \end{aligned} \tag{3.4.4}$$

So our normalized model is

$$\varphi(x; I) = \frac{I + 1}{2}\, dp_x^+ + \frac{1 - I}{2}\, dp_x^- \tag{3.4.5}$$

Due to the skewness and symmetry of the volume imbalance, we assume that

$$dp_x^+ = dp_{-x}^- \tag{3.4.6}$$

This guarantees that $\varphi(x; I)$ skews in the direction of the volume imbalance, is symmetric about $I = 0$, and $\mu_+ = -\mu_-$. The last equality is from

$$\mu_+ = \sum_x x\, dp_x^+ = -\sum_x x\, dp_x^- = -\mu_- \tag{3.4.7}$$

Figure 3.5 gives a preview of dp^+, dp^-, and φ, for AEM stock over 5 second intervals. We have dp^+ and dp^- skewing right and left, respectively, as we enforced by equation 3.4.6. Also, φ given $I = 0$ is symmetric about 0 as we enforced by the same equation.

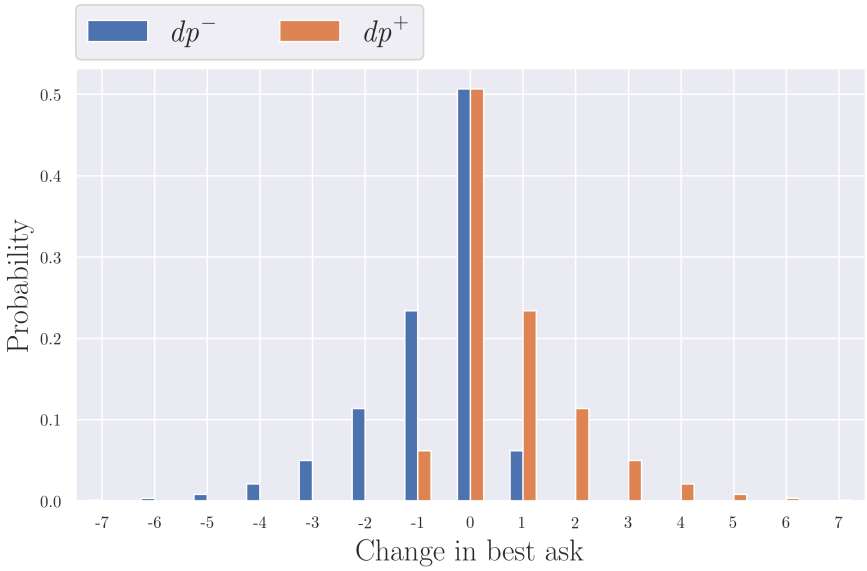
This model would also imply that the probability of the price movements change independent of I. That is,

$$\frac{\partial \varphi(x; I)}{\partial I} = \frac{dp_x^+ - dp_x^-}{2} \tag{3.4.8}$$

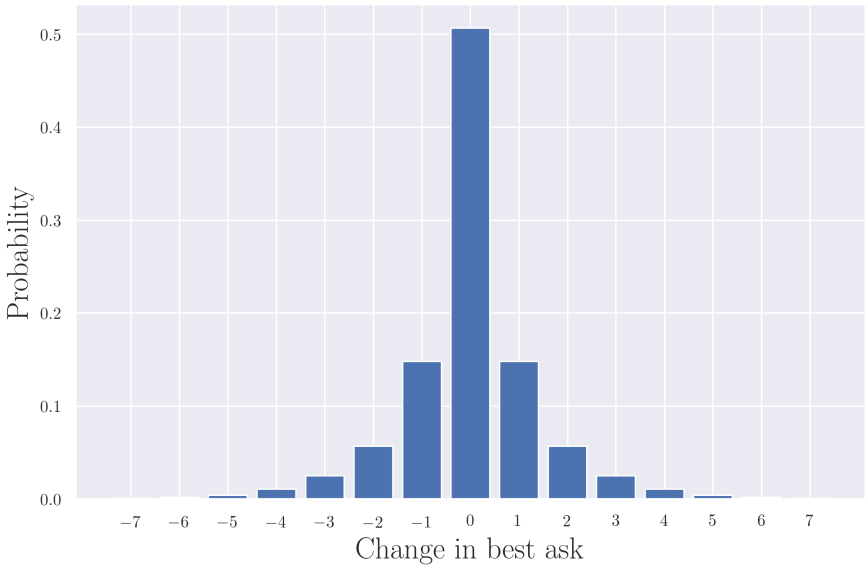
If we wanted to incorporate an asymmetry in the distribution as we change I we could include an extra parameter $\eta \in (0, 1)$ in the unnormalized distribution such that

$$\tilde{\varphi}(x; I) = \eta(I + 1)\, dp_x^+ + (1 - I)\, dp_x^- \tag{3.4.9}$$

This biases $\varphi(x; I)$ one way (in this case, against $I > 0$), so the imbalance does not



(a) dp^+ and dp^-



(b) φ given $I = 0$

Figure 3.5: Preview of dp^+, dp^-, and φ for AEM stock over 5 second intervals. Data used from April 17, 2017 over the entire trading day. We have $\mu_+ = 0.701$, $\sigma_+^2 = 1.42$, $\theta_+ = 2.55$, and $\kappa_+ = 11.96$.

shift the distribution equally in both directions. We found evidence of this bias against positive I for AEM stock in Figure 2.16 which we could incorporate this way. Subplot (b) in Figure 2.16 shows the probability of the price increases less with increasing imbalance than the probability of the price decreases with decreasing imbalance. The placement of η in equation 3.4.9 breaks the symmetry about $I = 0$, but also makes it so that the

expected value of $\varphi(x; I)$ is 0 at a strictly positive imbalance I which is dependent on the value of η. However, we will assume the distributions are symmetric to simplify the model.

Next we will calculate the moments of φ in terms of the moments of dp^+.

3.5 Moments of Distribution Model

In this section we present the moments of our distribution model. These moments will allow us to check that our distribution is satisfying the conditions we want and to provide a better statistical understanding of the model. Each moment is defined as

$$n^{\text{th}} \text{ Moment} = E[(X - E[X])^n] \tag{3.5.1}$$

for random variable X and $n \in \mathbb{N}$. The first and second moments are the mean and variance, respectively. The skewness and kurtosis are defined in terms of the third and fourth moments, respectively. These definitions are

$$\text{Skew}[X] = \frac{E[(X - E[X])^3]}{E[(X - E[X])^2]^{\frac{3}{2}}} \tag{3.5.2}$$

$$\text{Kurt}[X] = \frac{E[(X - E[X])^4]}{E[(X - E[X])^2]^2} \tag{3.5.3}$$

It should be noted that kurtosis is sometimes defined as the 'excess kurtosis' $\text{Kurt}[X] - 3$, but we will use equation 3.5.3 when we refer to the kurtosis. The reason for the use of excess kurtosis is because the kurtosis of the univariate normal distribution is 3, so the excess kurtosis is a measure relative to the normal distribution.

Given the volume imbalance I, the change in best ask price $x \sim \varphi$, and the conditional expected value of x is

$$\begin{aligned} E\left[x \middle| I\right] &= \sum_x x\varphi(x; I) \\ &= \frac{I+1}{2} \sum_x x\, dp_x^+ + \frac{1-I}{2} \sum_x x\, dp_x^- \\ &= \frac{I+1}{2}\mu_+ - \frac{1-I}{2}\mu_- \\ &= I\mu_+ \end{aligned} \tag{3.5.4}$$

In the limit cases where $I = \pm 1$ the mean is $\pm\mu_+$ and smoothly moves between these two values. We also have that the mean is 0 when $I = 0$ as expected – we forced our

distribution to be symmetric with equation 3.4.6.

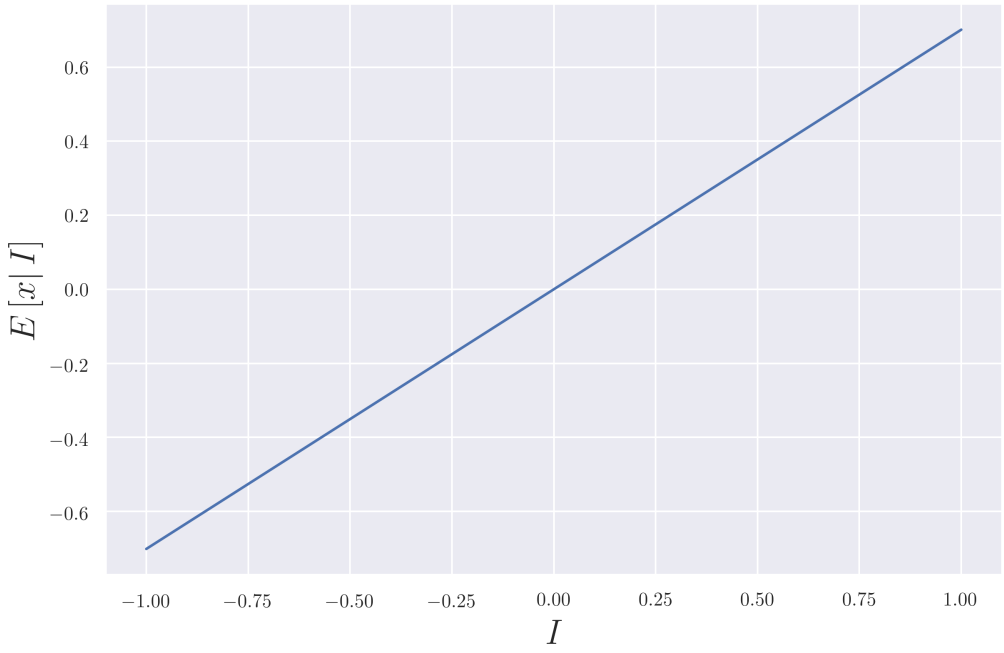


Figure 3.6: $E\left[x\middle|\, I\right]$ where μ_+ is taken from Figure 3.5.

Figure 3.6 presents $E\left[x\middle|\, I\right]$ for dp^+ in Figure 3.5. As in equation 3.5.4 it varies linearly with I, is 0 at $I = 0$, and $\pm\mu_+$ at $I = \pm 1$.

We define the variance σ_+^2 of dp^+ as

$$\begin{aligned} \sigma_+^2 &= \sum_x x^2\, dp_x^+ - \left(\sum_x x\, dp_x^+\right)^2 \\ &= \sum_x x^2\, dp_x^+ - \mu_+^2 \\ \implies \sum_x x^2\, dp_x^+ &= \sigma_+^2 + \mu_+^2 \end{aligned} \tag{3.5.5}$$

Equations 3.4.6 and 3.5.5 imply that $\sigma_+^2 = \sigma_-^2$. The variance of x given I is

$$
\begin{aligned}
\text{Var}[x|I] &= \frac{I+1}{2}\sum_x x^2\, dp_x^+ + \frac{1-I}{2}\sum_x x^2\, dp_x^- - I^2\mu_+^2 \\
&= \frac{I+1}{2}\sum_x x^2\, dp_x^+ + \frac{1-I}{2}\sum_x x^2\, dp_{-x}^+ - I^2\mu_+^2 \\
&= \frac{I+1}{2}\sum_x x^2\, dp_x^+ + \frac{1-I}{2}\sum_x x^2\, dp_x^+ - I^2\mu_+^2 \\
&= \frac{I+1}{2}(\sigma_+^2+\mu_+^2) + \frac{1-I}{2}(\sigma_+^2+\mu_+^2) - I^2\mu_+^2 \\
&= \sigma_+^2 + \mu_+^2 - I^2\mu_+^2 \\
&= \sigma_+^2 + (1-I^2)\mu_+^2
\end{aligned}
\tag{3.5.6}
$$

We find that the variance of x is largest when $I = 0$ and decreases as $I \to \pm 1$. This is because when the imbalance is 0 we have less information about where prices will be moving. This is in contrast to when $I = \pm 1$ where the distribution will be skewed by the extreme imbalance in the book like we saw in chapter 2.

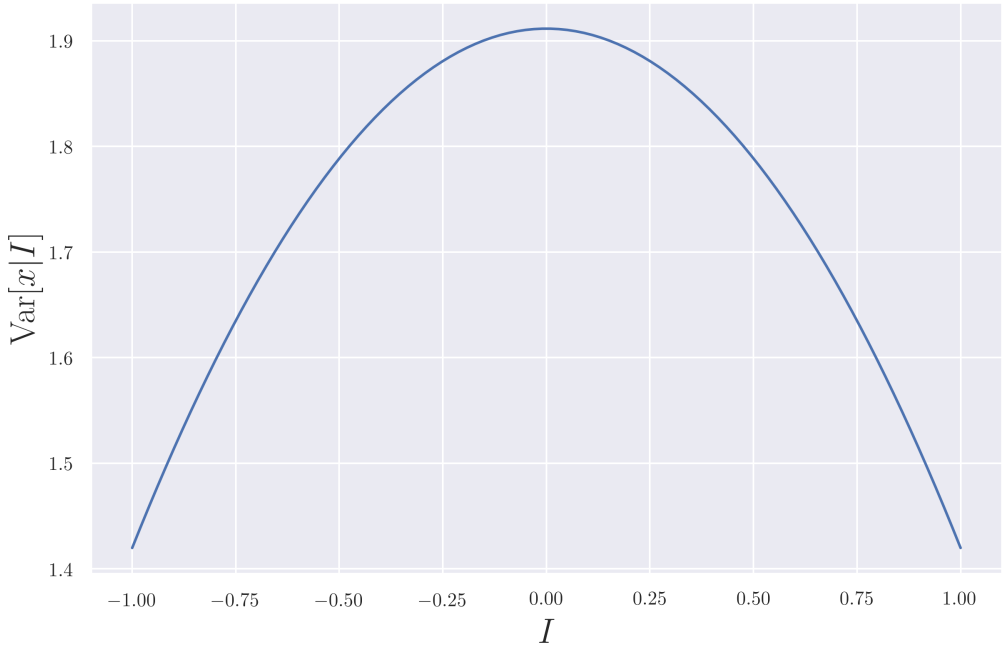


Figure 3.7: $\text{Var}[x|I]$ where μ_+ and σ_+^2 are taken from Figure 3.5.

Figure 3.7 presents $\text{Var}[x|I]$ for dp^+ in Figure 3.5. As in equation 3.5.6 it reaches a maximum at $I = 0$ and a minimum at $I = \pm 1$ where the minimum is σ_+^2.

We now define θ_+ and κ_+ as the third and fourth moments of dp^+, respectively, so we can write down the $\text{Skew}[x|I]$ and $\text{Kurt}[x|I]$. Using the fact that

$$E[(X - E[X])^3] = E[X^3] - 3E[X]E[X^2] + 2E[X]^3 \tag{3.5.7}$$

and

$$E[(X - E[X])^4] = E[X^4] - 4E[X]E[X^3] + 2E[X]^2E[X^2] - 3E[X]^4 \tag{3.5.8}$$

we can write out the third and fourth moments of our distributions. From Equation 3.5.7, the third moment of dp^+ is

$$\begin{aligned} \theta_+ &= \sum_x x^3 dp_x^+ - 3\mu_+ \sum_x x^2 dp_x^+ + 2\mu_+^3 \\ &= \sum_x x^3 dp_x^+ - 3\mu_+(\sigma_+^2 + \mu_+^2) + 2\mu_+^3 \\ \implies \sum_x x^3 dp_x^+ &= \theta_+ + 3\mu_+\sigma_+^2 + \mu_+^3 \end{aligned} \tag{3.5.9}$$

Equations 3.4.6 and 3.5.9 imply that $\theta_+ = -\theta_-$. Equation 3.5.7 then gives the third moment of φ as

$$\begin{aligned} E[(x - I\mu_+)^3|I] &= E[x^3|I] - 3I\mu_+E[x^2|I] + 2I^3\mu_+^3 \\ &= \frac{I+1}{2}\sum_x x^3 dp_x^+ + \frac{1-I}{2}\sum_x x^3 dp_x^- - \\ &\quad 3I\mu_+(\sigma_+^2 + \mu_+^2) + 2I^3\mu_+^3) \\ &= \frac{I+1}{2}(\theta_+ + 3\mu_+\sigma_+^2 + \mu_+^3) + \frac{1-I}{2}(-\theta_+ - 3\mu_+\sigma_+^2 - \mu_+^3) - \\ &\quad 3I\mu_+(\sigma_+^2 + \mu_+^2) + 2I^3\mu_+^3 \\ &= I(\theta_+ + 3\mu_+\sigma_+^2 + \mu_+^3) - 3I\mu_+(\sigma_+^2 + \mu_+^2) + 2I^3\mu_+^3 \\ &= \theta_+ I + 2\mu_+^3 I(1 - I^2) \\ &= I[\theta_+ + 2\mu_+^3(1 - I^2)] \end{aligned} \tag{3.5.10}$$

The skew of φ given I is given by

$$\begin{aligned} \text{Skew}[x|I] &= \frac{E[(x - I\mu_+)^3|I]}{E[(x - I\mu_+)^2|I]^{\frac{3}{2}}} \\ &= I\frac{\theta_+ + 2\mu_+^3(1 - I^2)}{(\sigma_+^2 + (1 - I^2)\mu_+^2)^{\frac{3}{2}}} \end{aligned} \tag{3.5.11}$$

Equation 3.5.11 shows the skewness of φ is 0 when $I = 0$ because we have forced the distribution to be symmetric at $I = 0$. At the limits $I = \pm 1$ we have $\text{Skew}[x|I] = \pm\theta_+/\sigma_+^3$, respectively, which is just the skewness of dp^+ and dp^- as we should expect.

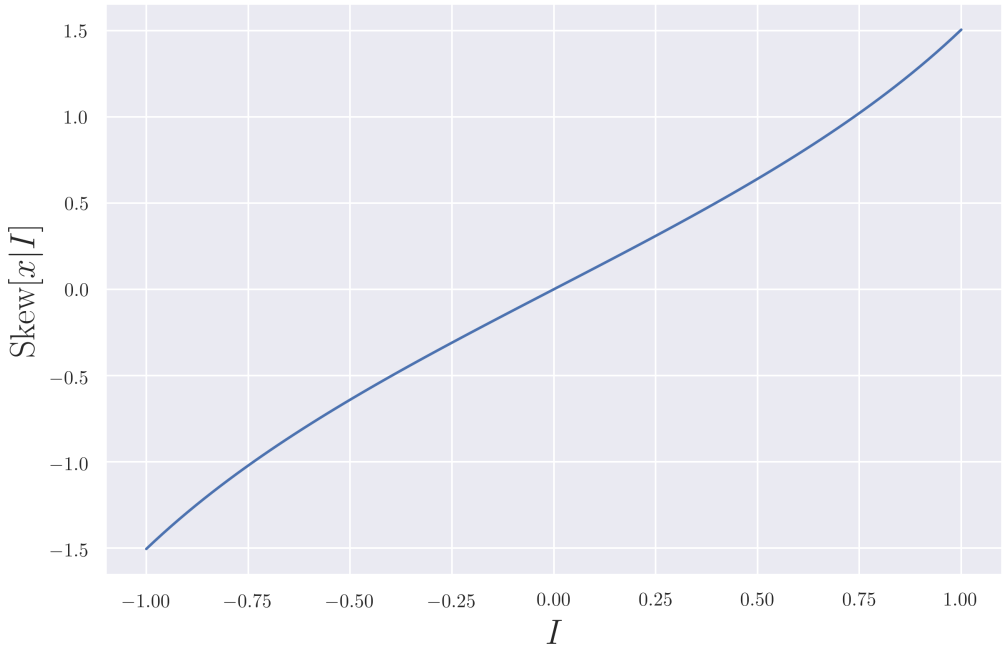


Figure 3.8: Skew$[x|I]$ where μ_+, σ_+^2, and θ_+ are taken from Figure 3.5.

Figure 3.8 presents Skew$[x|I]$ for dp^+ in Figure 3.5. As in equation 3.5.11 the skew is 0 at $I = 0$ and moves to $\pm\theta_+/\sigma_+^3$ at $I = \pm 1$.

From Equation 3.5.8, the fourth moment of dp^+ is

$$\begin{aligned}
\kappa_+ &= \sum_x x^4 dp_x^+ - 4\mu_+ \sum_x x^3 dp_x^+ + 6\mu_+^2 \sum_x x^2 dp_x^+ - 3\mu_+^4 \\
&= \sum_x x^4 dp_x^+ - 4\mu_+(\theta_+ + 3\mu_+\sigma_+^2 + \mu_+^3) + \\
&\quad 6\mu_+^2(\sigma_+^2 + \mu_+^2) - 3\mu_+^4 \\
&= \sum_x x^4 dp_x^+ - 4\mu_+\theta_+ - 6\mu_+^2\sigma_+^2 - \mu_+^4 \\
\implies \sum_x x^4 dp_x^+ &= \kappa_+ + 4\mu_+\theta_+ + 6\mu_+^2\sigma_+^2 + \mu_+^4
\end{aligned} \tag{3.5.12}$$

Equations 3.4.6 and 3.5.12 imply that $\kappa_+ = \kappa_-$. Equation 3.5.8 then gives the fourth moment of φ as

$$\begin{aligned}
E[(x - I\mu_+)^4|I] &= E[x^4|I] - 4\mu_+ E[x^3|I] + 6\mu_+^2 E[x^2|I] - 3\mu_+^4 \\
&= \frac{I+1}{2}\sum_x x^4 dp_x^+ + \frac{1-I}{2}\sum_x x^4 dp_x^- - \\
&\quad 4I^2\mu_+(\theta_+ + 3\mu_+\sigma_+^2 + \mu_+^3) + \\
&\quad 6I^2\mu_+^2(\sigma_+^2 + \mu_+^2) - 3I^4\mu_+^4 \\
&= \kappa_+ + 4\mu_+\theta_+ + 6\mu_+^2\sigma_+^2 + \mu_+^4 - \\
&\quad 4I^2\mu_+(\theta_+ + 3\mu_+\sigma_+^2 + \mu_+^3) + \\
&\quad 6I^2\mu_+^2(\sigma_+^2 + \mu_+^2) - 3I^4\mu_+^4 \\
&= \kappa_+ + 4\mu_+\theta_+(1 - I^2) + 6\mu_+^2\sigma_+^2(1 - I^2) + \\
&\quad \mu_+^4(1 + 3I^2)(1 - I^2) \\
&= \kappa_+ + (1 - I^2)[4\mu_+\theta_+ + 6\mu_+^2\sigma_+^2 + \mu_+^4(1 + 3I^2)]
\end{aligned} \tag{3.5.13}$$

The Kurtosis of φ given I is given by

$$\begin{aligned}
\text{Kurt}[x|I] &= \frac{E[(x - I\mu_+)^4|I]}{E[(x - I\mu_+)^2|I]^2} \\
&= \frac{\kappa_+ + (1 - I^2)[4\mu_+\theta_+ + 6\mu_+^2\sigma_+^2 + \mu_+^4(1 + 3I^2)]}{(\sigma_+^2 + (1 - I^2)\mu_+^2)^2}
\end{aligned} \tag{3.5.14}$$

From equation 3.5.14 we recover the kurtosis of dp^+ and dp^- when $I = \pm 1$, respectively. That is, $\text{Kurt}[x|I] = \kappa_+/\sigma_+^4$ when $I = \pm 1$ as we expect.

Figure 3.9 presents $\text{Kurt}[x|I]$ for dp^+ in Figure 3.5. As in equation 3.5.14, the kurtosis is κ_+/σ_+^4 at $I = \pm 1$ with a local minimum at $I = 0$. Taking the derivative of equation 3.5.14 with respect to I and setting the result to zero allows us to find the extrema of the kurtosis as

$$I = 0 \quad \text{and} \quad I = \pm\sqrt{1 + \frac{\kappa_+\mu_+ - 2\mu_+^3\sigma_+^2 - 3\mu\sigma_+^4 - 2\sigma_+^2\theta_+}{2\mu_+^5 + 6\mu_+^3\sigma_+^2 + 2\mu_+^2\theta_+}} \text{ for } \mu_+ \neq 0 \tag{3.5.15}$$

Equation 3.5.15 matches the extrema in Figure 3.9 which are found at $I = 0, \pm 0.545$.

With equations 3.5.4, 3.5.6, 3.5.11, and 3.5.14, we know how to calculate the measures of our distribution model once we know the moments of dp^+ – as per the examples depicted in Figures 3.6, 3.7, 3.8, and 3.9.

Putting everything together we can now write down the optimization problem to give us a starting point for spoofing detection in the limit order book.

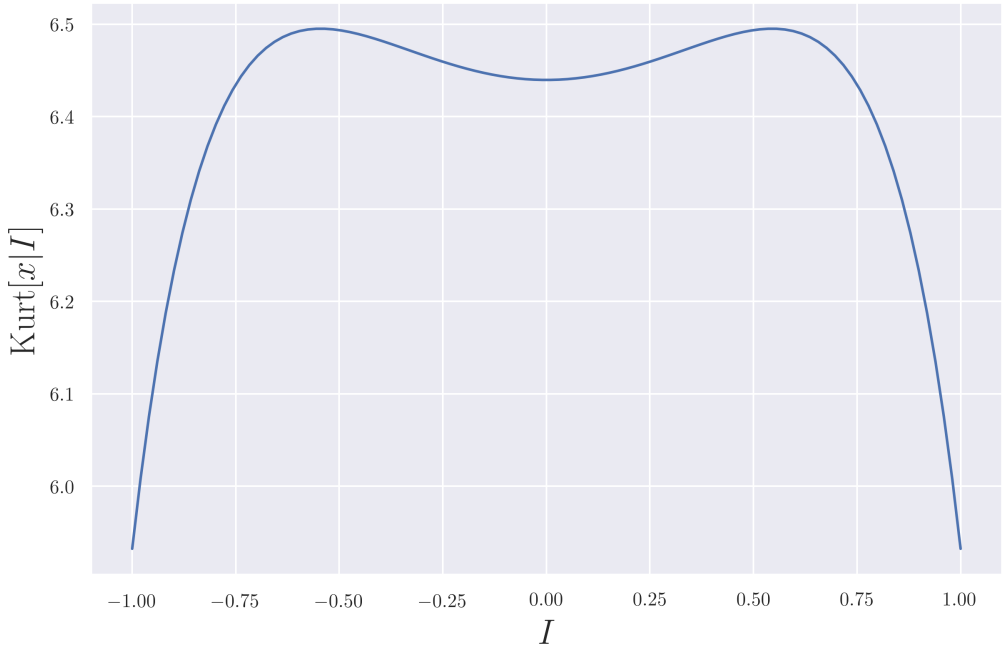


Figure 3.9: Kurt$[x|I]$ where μ_+, σ_+^2, θ_+, and κ_+ are taken from Figure 3.5.

3.6 Optimization Problem

We now have a model which allows us to influence the distribution of the best ask price φ by placing spoofing limit orders $\tilde{v}_t$ in the book $\vec{v}_t$ to alter the volume imbalance ratio $I(\vec{v}_t + \tilde{v}_t; \vec{w}, K)$. This in turn has an impact on the cost functional $C_{\mathrm{S}}(\vec{v}_t, H, p^+, x_t, \tilde{v}_t)$ which we want to minimize. However, we also have conditions given by equations 3.2.16 and 3.2.18 on when to delay our market order and when to spoof, respectively.

Given parameters dp^+, $\vec{w}$, and K, the optimization problem for determining the optimal limit order placement is

$$\min_{\tilde{v}_t} \sum_{x_t} C_{\mathrm{S}}(\vec{v}_t, H, p^+, x_t, \tilde{v}_t)\varphi(x_t; I(\vec{v}_t + \tilde{v}_t; \vec{w}, K)) \tag{3.6.1}$$

We may find a $\tilde{v}_t$ which minimizes equation 3.6.1, but the expected cost may not be smaller than simply placing a market order or delaying your market order – spoofing would not be the optimal strategy. The optimal strategy will depend on the shape of the limit order book as much as it depends on how the price change distribution is influenced by the volume imbalance ratio.

In the next chapter we estimate all of our model parameters so that we can solve equation 3.6.1 over a collection of stocks. The solutions will give us a starting point for

flagging suspicious behaviour in the limit order book.

3.7 Model Summary

In this chapter we developed the notation and definitions necessary for discussing the limit order book as well as the costs associated with market orders, delayed market orders, and spoofing orders in our one period model. We also generalized the definition of the volume imbalance ratio and used this new definition to allow one to influence the distribution of the change in the best ask price by adding volume beyond the touch of the limit order book. We also now have a model for the distribution of the change in best ask price parameterized by the generalized imbalance ratio.

Our model was a convex combination of two distributions which smoothly transition to one another through the volume imbalance ratio. We assume that the total distribution was symmetric about $I = 0$ while providing a starting point for symmetry breaking through equation 3.4.9. We also derived the relevant moments of the model and made checks to show that the results at $I = \pm 1$ were consistent with the definitions of dp^+ and dp^-.

The next step we take in the following chapter is to calibrate our model using data provided by TMX for various stocks. To summarize, our model parameters so far are

- $\vec{w}$: weights in our generalized volume imbalance ratio
- K: depth we take in the limit order book
- dp^+: distribution of the change in best ask price when $I = 1$

Parameters $\vec{w}$ and K tell us how to calculate $I(\vec{v}; \vec{w}, K)$ and dp^+ will give us our price distribution. With these pieces we can solve the optimization problem, given by equation 3.6.1, for optimal placement of limit orders to manipulate the best ask price. This will give us a starting point for flagging possible manipulation in the limit order book as now we can quantify the costs associated with a spoofer's decisions to manipulate the limit order book.

In chapter 4 we discuess how we can compare the results across different stocks. From chapter 2 we saw that different stocks behave differently over different time scales. We start the next chapter by fixing this time scale, so we can compare the results of our calibration between stocks. This allows us to explore the relationships between model parameters and statistics we can draw from activity on the limit order book. We then

return to the statistical tests in chapter 2 to see improvements in our generalized imbalance definitions over using just the touch – increasing Cramer's V under new parameters would suggest an increase in the association between changes in the price and the volume imbalance definition using those same parameters.

Chapter 4

Model Calibration

4.1 Introduction

In this chapter we set up the calibration for the parameters of our price distribution model, but we first need to resolve a problem we discussed back in chapter 2 – what do we do about sampling time? The first thing we will do is outline how to choose the sampling time as a way of comparing the calibration results stock to stock. However, the calibration itself can be done over any sampling time one chooses. We just want an initial set of results which is comparable between stocks so as to investigate relationships between our parameters and statistics we can derive from the limit order book. For example, these statistics could be the spread, time between limit/market/cancellation orders, volume traded, or number of each type of order.

We also define the depth of book we take in our model calibration based on the support of the empirical price change distribution. We then have a method for determining the optimal sampling time Δt and depth K, so we provide an algorithm for calibrating our model to data from the limit order book. With all of our parameters and limit order book statistics we can investigate any relationships we find between them which provide a qualitative way of understanding what each parameter represents for a given limit order book.

We want to calibrate our model using exponential and free imbalance weights $\vec{w}$, but they will require different estimation techniques. The free imbalance weights can be calibrated using maximum likelihood estimation and we will show that a penalized version of maximum likelihood, maximum a posterior estimation, is required for the exponential weights. This is due to the exponential weights being parameterized by an unbounded, positive, real number.

With an optimal sampling time Δt and model parameters w, K, and dp^+, established

for each stock, we can repeat the statistical tests of chapter 2 across the same collection of stocks in Tables 2.4 and 2.8. We present results for the statistical tests using both definitions for calculating the imbalance – equations 3.3.1 and 3.3.2. This will give us a way of comparing if the optimal weights w give an improvement over just using the volumes at the touch. The weights will also tell us which stocks have price movements that have strong association to volume imbalances deep in the limit order book.

We then discuss the goodness of fit between the empirical price change distribution and our calibrated model. We use Kullback–Leibler divergence and probability-probability plots to argue, numerically and visually, that we have an excellent fit for our models. We also return to the discussion of how to appropriately aggregate the instantaneous imbalances into an average imbalance and show that time weighting removes negative correlations between the change in the best ask price and the volume imbalance ratio.

Finally, we present the optimization problem presented to the spoofer which we solve, in detail, in chapter 5.

4.2 Optimal Sampling Time

Stocks are not the same and people do not trade on all stocks at the same frequency. The trading activity of a stock varies during the day and even day to day. The higher the trade volume and frequency the more prices can move in a given time interval. The time needed to observe a particular variation in the distribution of the best ask price is dependent on this sampling time Δt as we saw in chapter 2.

In our calibration we need to specify a sampling time to generate the distribution of a change in the best ask price. Choosing too small a time interval can give us distributions with very small variance, i.e. little movement in the price. We want to be able to compare stocks over time intervals where they show similar movement in their prices. It is a common theme in financial mathematics that variance and time scales are linked. For example, the increment of Brownian motion $dW_t = W_t - W_s$ with $s < t$, which many financial models are built from, is drawn from the normal distribution $\mathcal{N}(0, t-s)$ for time interval $\Delta t = t - s$. The larger the time interval, the larger the variance of the random variable. We can use this idea to fix a time scale by fixing the variance of the distribution of the change in the best ask price.

To do this we can pick a benchmark variance σ_b^2 and then find the sampling time $\Delta t > 0$ which gives an empirical variance $\sigma_{\Delta t}^2$ as close to the benchmark as possible. For each stock we need to find Δt and $\sigma_{\Delta t}^2$ such that

$$\arg\min_{\Delta t>0} |\sigma_b^2 - \sigma_{\Delta t}^2| \tag{4.2.1}$$

It just remains to determine what to choose for our σ_b^2. When we refer to the optimal sampling time we mean optimal with respect to σ_b^2 according to equation 4.2.1. Alternative methods for determining the time scale could be based on the number of seconds to see a set number of trades, number of orders, specific amount of volume entering/exiting the book, etc. For example, Bechler & Ludkovski fix their time scale by the number of seconds to see 20,000 shares traded in the book [51]. We do not take this approach because our goal is to model the distribution in the change in the best ask and fixing the variance guarantees a certain amount of price movement which may be more difficult via an indirect way like trade volume – it will be more difficult to control the price change distribution.

Fixing the variance also gives us the advantage that the shape of the distribution will be heavily determined by the probability that the best ask does not change. Since the distributions are all peaked at 0, if we fix the variance and the $\mathbb{P}[x = 0] \approx 1$ then the probability mass in the tails of the distribution spreads out deep in the support to achieve the fixed variance. By the same argument, as $\mathbb{P}[x = 0]$ decreases the support will decrease with it. The financial intuition for this is that stocks with a best ask that changes very little must increase/decrease multiple ticks when it does move in order to see the same fixed variance as a stock with a best ask which is constantly moving.

Initially, only data for AEM stock on April 17, 2017 was available to us – full access to TMX's data came later. From initial analysis of this data we determined that 5 seconds showed an adequately large support for us to test our model. This is why in chapter 2 our statistical analysis was done over 5 second intervals. The distribution of the change in the best ask price for AEM on this day had a variance of ≈ 2 ticks and the optimal Δt to see a variance of 2 in the change in the best ask price was 5 seconds as well. This will serve as our benchmark variance for determining the optimal sampling time Δt. This result is shown in Figure 4.1.

It is worth repeating that Δt is just the sampling time we are calibrating our model over. We can choose any interval we want, but for the purposes of comparing calibration results we want to pick a sampling time large enough to see a reasonable amount of price movement.

We can also investigate the intraday optimal sampling time for each stock by finding $\sigma_{\Delta t}^2$ for the first hour, the middle 4.5 hours, then the final hour of trading. That is, 9:30 - 10:30, 10:30 - 15:00, and 15:00 - 16:00. We refer to these as the start, mid, and end

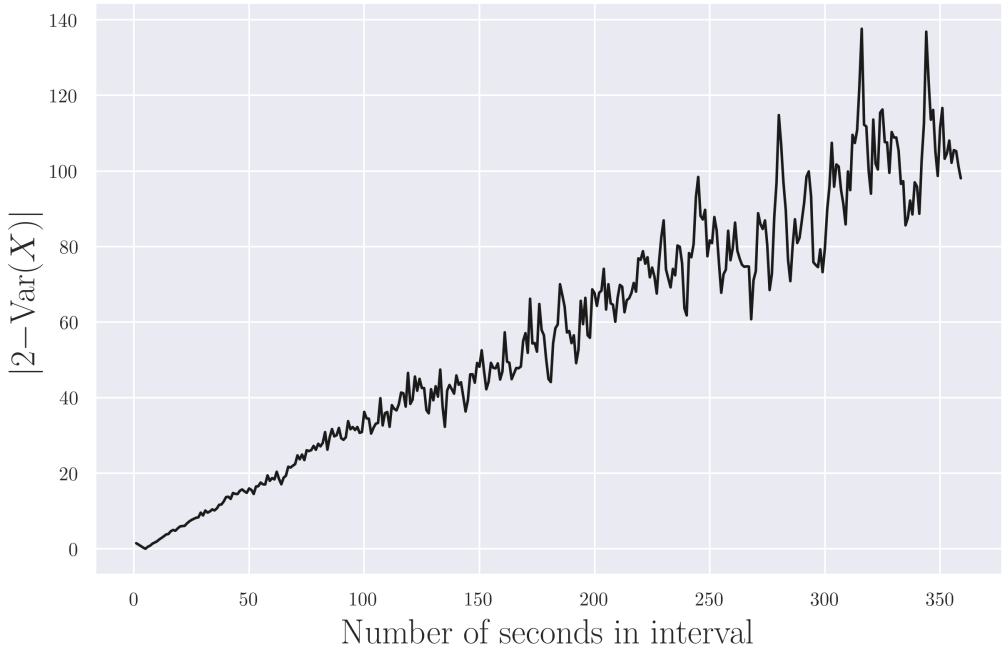


Figure 4.1: Output of equation 4.2.1 for AEM stock on April 17, 2017 with $\sigma_b^2 = 2$. We find the optimal number of seconds is $\Delta t = 5$. Data used is from the entire trading day.

periods, respectively.

4.2.1 Optimal Sampling Time Analysis

We select four stocks from the top 100 most active stocks on the TSX during the year 2017 to discuss the optimal time interval. We take AEM, BMO, PPL, and CPG as these stocks exhibit very different time scales for their price changes – which we will see in this section. The objective is to minimize the absolute value of the difference of 2 and the variance of the distribution calculated over increasing interval lengths. Figures 4.2-4.5 show the results for determining the optimal time interval needed to see a variance of 2 in the distribution of the change in the best ask price.

We see that the optimal sampling time for the first hour of the day is always smaller in our 4 cases than the sampling time over the whole day. This would imply that, at the beginning of the day, price movements happen more frequently and at a larger intensity than they do during the rest of the day. We expect this since the beginning of the trading day is usually the most active time for any stock.

We also see that the general order from smallest to largest interval is the first hour, whole day, mid day, and the last hour can fall above or below the mid day. This is shown, for example, in Figures 4.2 and 4.6 where the last hour falls below the mid day

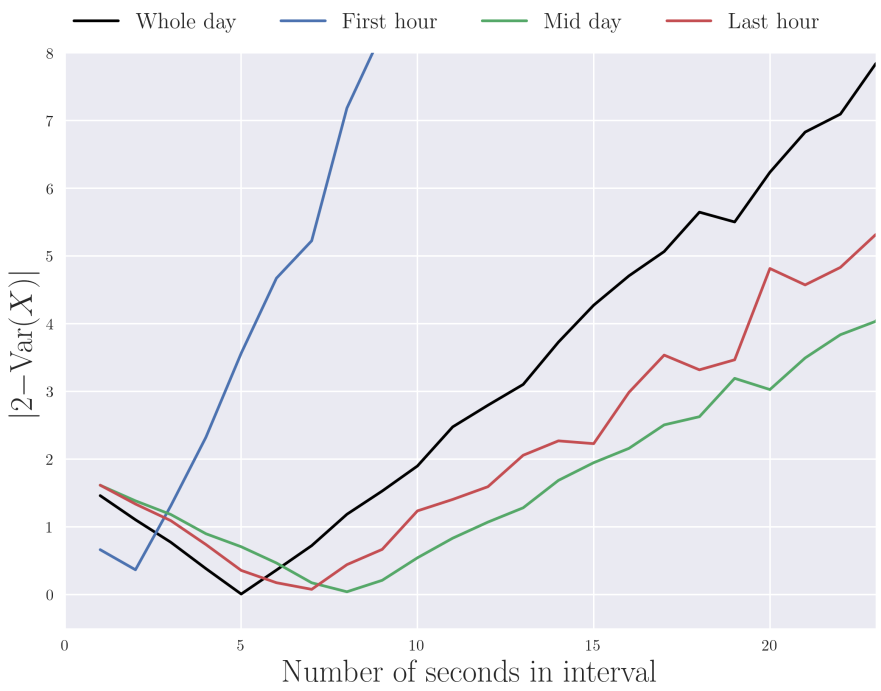


Figure 4.2: Optimal sampling time for AEM stock on June 9, 2017 to see a variance of 2 in the distribution of the change in best ask price. $\Delta t = 5$, 2, 8, and 7 seconds, for the whole day, first hour, mid day, and last hour, respectively.

Figure 4.3: Optimal sampling time for BMO stock on June 9, 2017 to see a variance of 2 in the distribution of the change in best ask price. $\Delta t = 10$, 4, 13, and 7 seconds, for the whole day, first hour, mid day, and last hour, respectively.

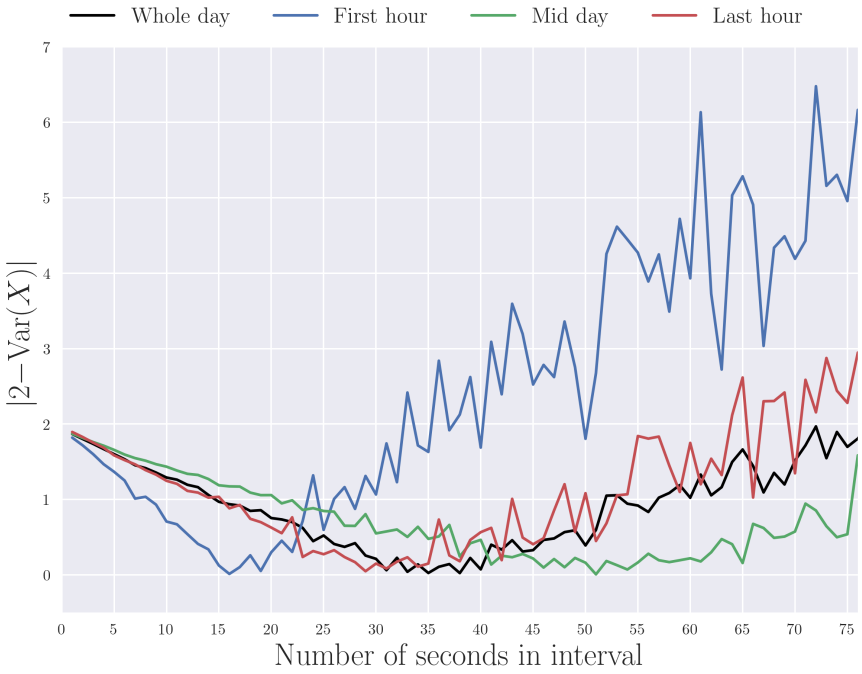


Figure 4.4: Optimal sampling time for PPL stock on June 9, 2017 to see a variance of 2 in the distribution of the change in best ask price. $\Delta t = 38$, 16, 51, and 29 seconds, for the whole day, first hour, mid day, and last hour, respectively.

Figure 4.5: Optimal sampling time for CPG stock on June 9, 2017 to see a variance of 2 in the distribution of the change in best ask price. $\Delta t = 51$, 28, 59, and 68 seconds, for the whole day, first hour, mid day, and last hour, respectively.

and then above the mid day, respectively. However, we do not see the last hour being as active as the first hour of the day. We saw this back in chapter 2 in Figure 2.5 where the distribution over the whole day and the final hour appeared very similar.

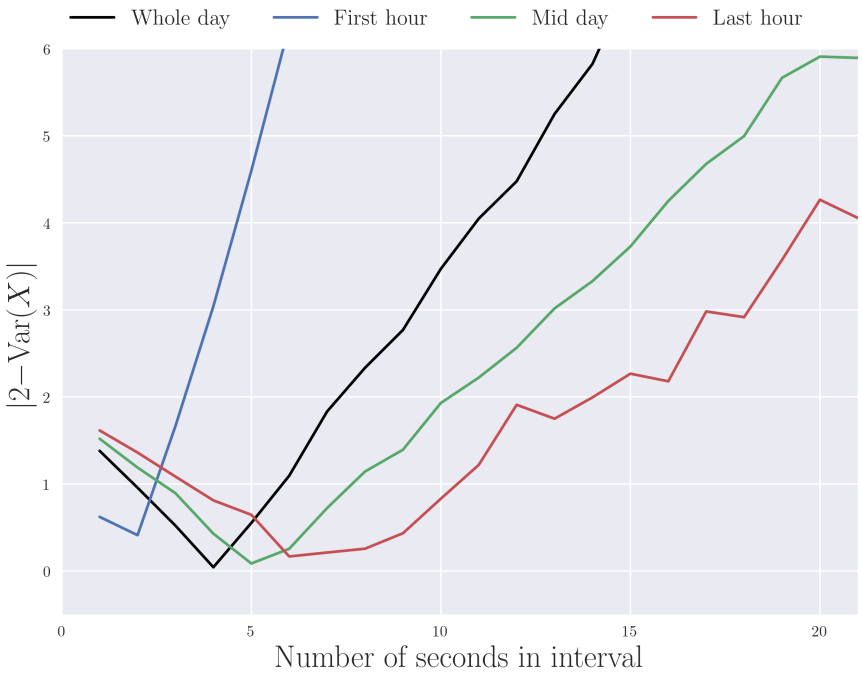


Figure 4.6: Optimal sampling time for AEM stock on June 8, 2017 to see a variance of 2 in the distribution of the change in best ask price. $\Delta t = 4$, 2, 5, and 6 seconds, for the whole day, first hour, mid day, and last hour, respectively.

Now we can investigate if any clusters or patterns emerge within an individual stock or between stocks for the optimal sampling time and any descriptive quantities we can derive from the stock – say the average spread or volume traded, for example. We take the stocks BMO, CNR, HFU, HSU, PAAS, and PPL and calculate the optimal sampling time Δt for each day between May 29, 2017 and August 04, 2017.

In Figure 4.7 we show the relationship between Δt and the average daily spread for the start, mid, and end periods of a day. The spread is largest during the start period of the day and narrows as the day goes on. The optimal time sampling decreases with the increasing spread with the optimal time sampling being smallest during the start period. We see no difference in the optimal sampling time between the mid and end periods, but the spread is generally wider in the mid period. However, we see 6 points as outliers – the two blue dots near a spread of 17 ticks, two red dots between a spread of 4 and 5 ticks, and two green dots between a spread of 5 and 6 ticks. These 6 points are the three periods from May 29, 2017 and July 4, 2017. These two days correspond to the holidays

in the United States for Memorial Day and Independence Day. Also, these two days see significantly fewer orders ($< 50\%$ as many) and generally larger spreads in all periods.

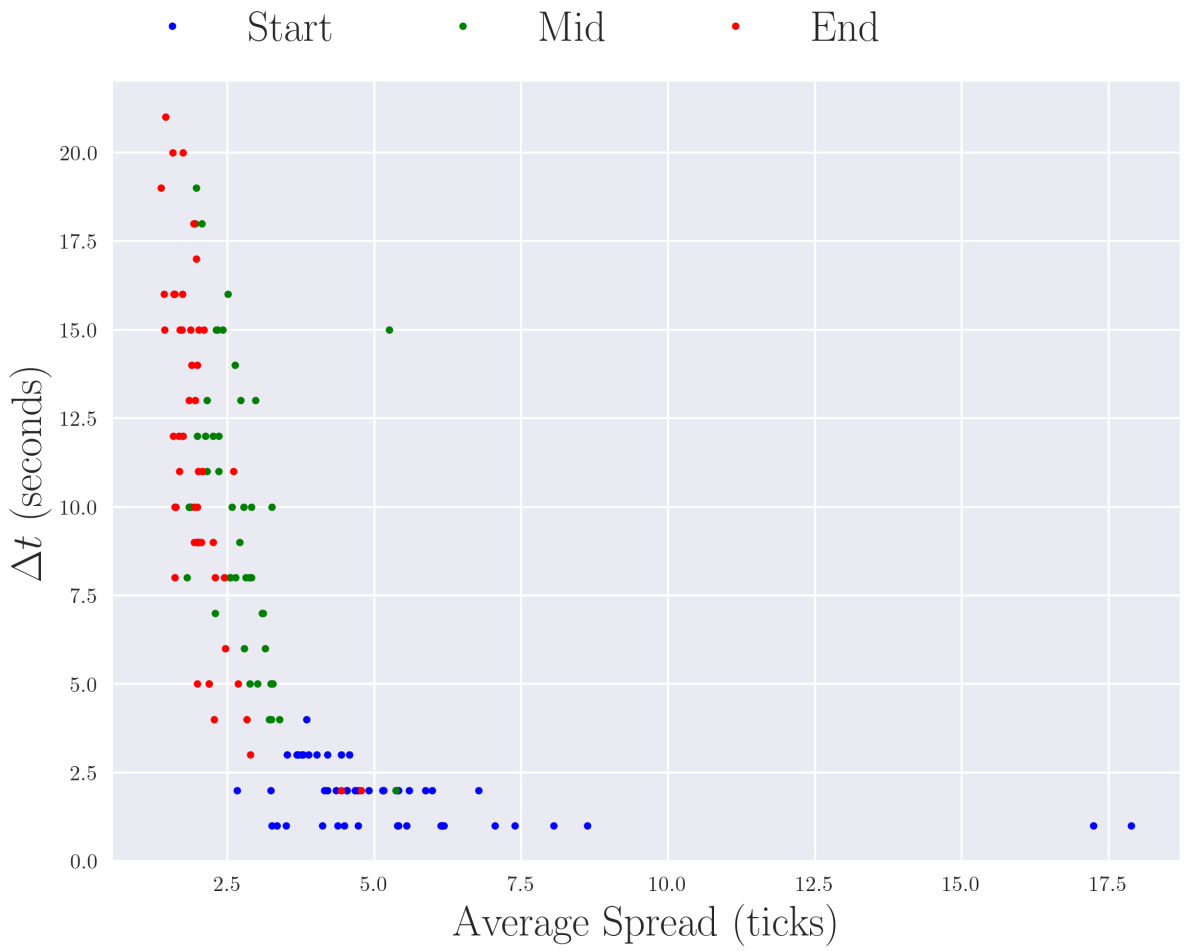


Figure 4.7: Optimal sampling time Δt (seconds) against the average spread (ticks) for CNR stock during start, mid, and end, periods. Each point corresponds to a trading day between May 29, 2017 and August 4, 2017.

Figure 4.8 depicts the relationship between the Δt and the average interarrival time of all orders for each day. We see the optimal sampling time increasing with increasing average interarrival time. The longer the time between orders, the longer it takes to see our variance of 2 in the distribution of the change in best ask price. The average interarrival time is smallest during the start period and then increases into the mid and end periods. We also have 6 more outlier points corresponding to May 29, 2017 and July 4, 2017. In addition, July 3, 2017 was when the TSX was closed for Canada Day. There are fewer orders placed during these days so the time between orders increases accordingly. Interestingly, even in our age of algorithmic trading, we still see a noticeable decrease in

trading activity on days where the American markets are closed. The American traders could still place orders in Canada via computer algorithms without needing to actually place the orders themselves. We see that the same outliers in all stocks we have looked at on those two days which would suggest either American traders do not operate at near the same level on their days off or they do not trust letting a computer place their trades unsupervised.

We also see evidence of why we found little difference between the mid and end periods for CNR in Figures 2.2c and 2.5c. The two periods cluster together with similar ranges in Δt, albeit with the end period having a tighter spread.

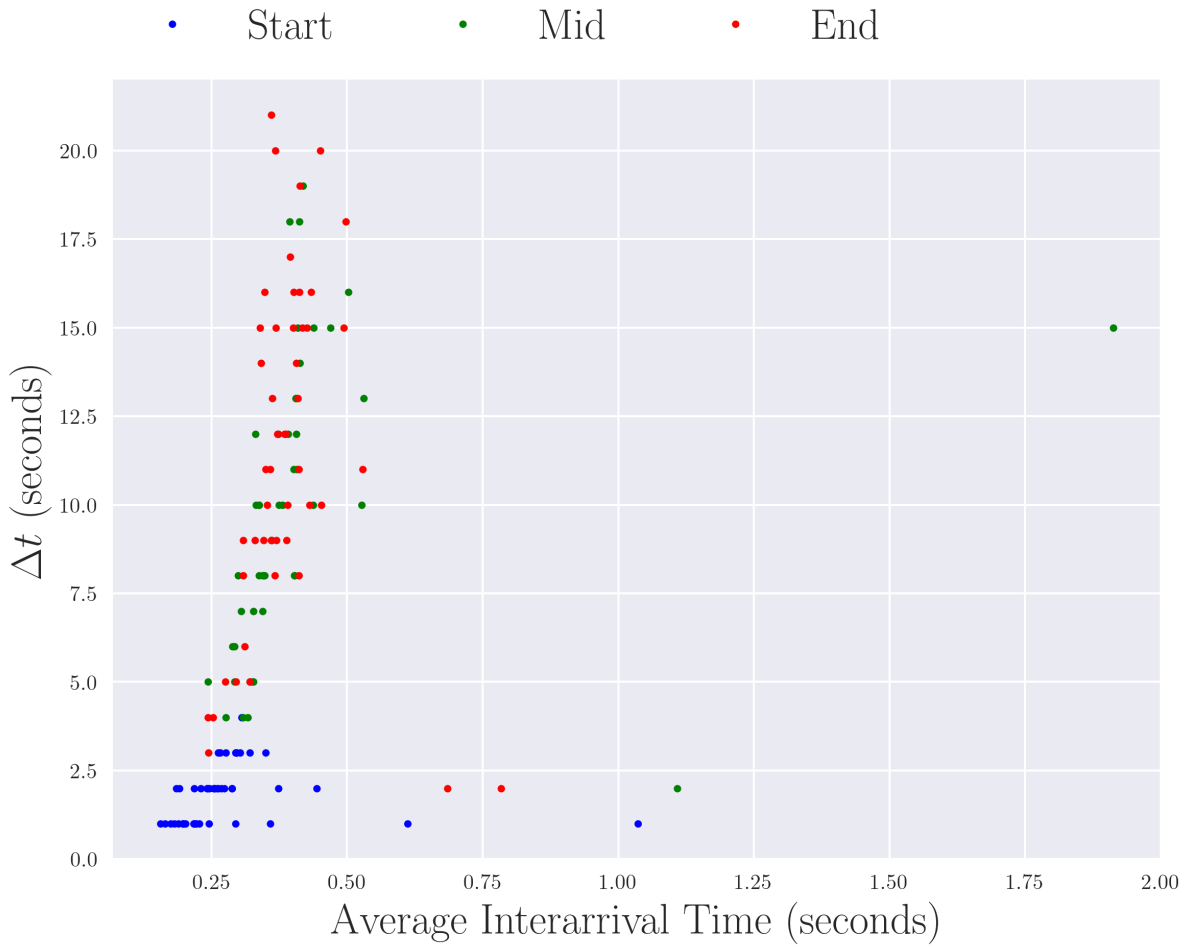


Figure 4.8: Optimal sampling time Δt (seconds) against the average interarrival time (seconds) of all orders for CNR stock during start, mid, and end, periods. All trading days between May 29, 2017 and August 4, 2017.

We also see outlier points for some stocks on days that are not holidays. For example, PPL has data points outside their usual clusters on July 24-26, 2017. This, possibly,

corresponds to the record dividend date for PPL on July 25. The holder of the share on that date receives the dividend payment issued on August 15. July 24 sees significantly slower activity on PPL as holders of the stock are less likely to sell their shares before they become the holder on record for the dividend payment. Then on July 25 and 26 the activity picks up as orders are placed to sell shares once the dividend payment has been locked in.

Similarly, BMO has several unusual points. All of these events were unlikely to have caused this change in trading behaviour on these days, but we were unable to find any other news and there was no dividend record date nearby. June 5, BMO announces it is bringing Android Pay to its Canadian customers with a $\approx 30\%$ drop in order numbers causing an increase in the average interarrival time, but with a normal average spread. June 14, BMO announces increasing US$ prime lending rate from 4.00% to 4.25% which accompanies a significant uptick in activity at the end of the trading day. BMO also has a share repurchase program complete on July 24 and appointed a new Leader for Financial Advisors on July 25. The dates July 24-27 then saw unusual spreads and average interarrival times.

However, there are outliers that we were unable to give possible links to news about the associated stock. For example, PAAS has an unusually low average interarrival time on June 14, 2017, but this date has nothing to do with dividends or exact press releases. So, like the BMO points that are unlikely to have been impacted by the news we did find, there are many points with no clear explanation for what caused them. This may indicate something out of the ordinary happening on these dates. It is not clear that this is caused by manipulation, but the exchange could use this information to scrutinize the trading behaviour of stocks on these days to look for possible irregularities.

Some stocks also had noticeably different behaviour on August 4, 2017. This was the Friday before the Ontario civic holiday long weekend where the TSX would be closed on the Monday. Some of our 6 stocks had an uptick in the number of orders during the day causing a decrease in the average interarrival time, but Δt remained large. This was not the case across all of the 6 stocks though – CNR did not have this issue as we saw in Figures 4.7 and 4.8.

Figures 4.9, 4.10, and 4.11, show the relationship between Δt and the average spread for our 6 stocks on all days except for the days we removed due to their unusual behaviour in at least one of our stocks. The figures are for the start, mid, and end periods, respectively. We see some clear clustering for individual stocks with the general trend being that smaller spreads lead to longer sampling times. This would make sense as smaller spreads mean prices have little room to move when they do. Interestingly, HFU

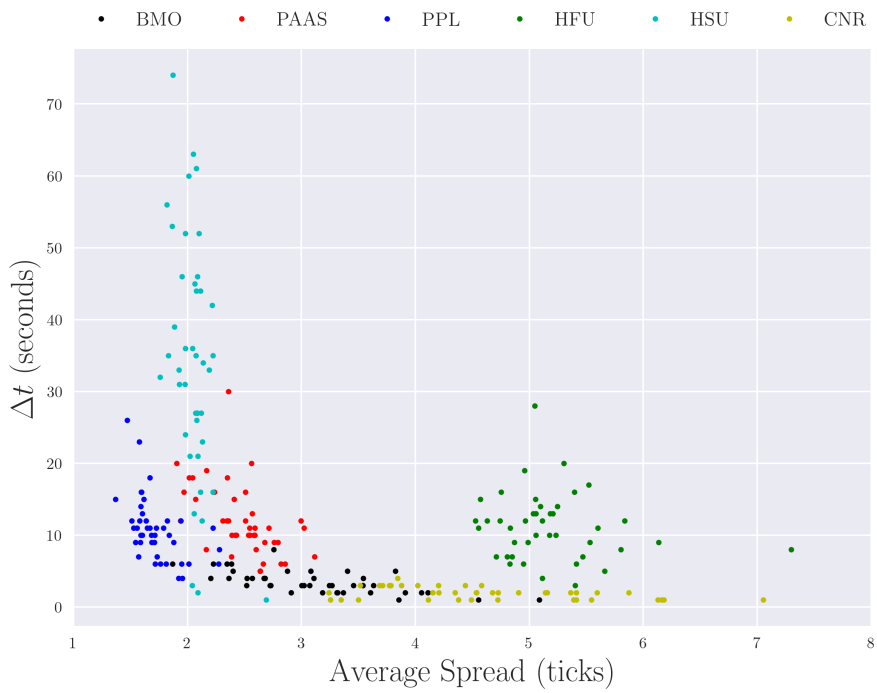


Figure 4.9: Optimal sampling time Δt (seconds) against the average spread (ticks) during the start period of the trading day. Each point corresponds to a trading day between May 29 - August 4, 2017 – excluding May 29, July 4-6, July 24-25, August 4.

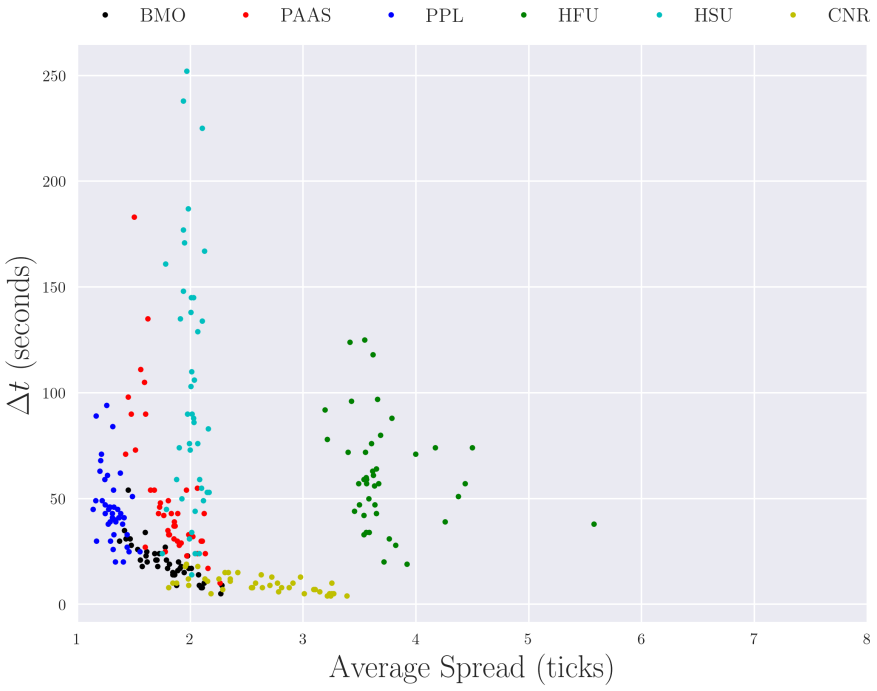


Figure 4.10: Optimal sampling time Δt (seconds) against the average spread (ticks) during the mid period of the trading day. Each point corresponds to a trading day between May 29 - August 4, 2017 – excluding May 29, July 4-6, July 24-25, August 4.

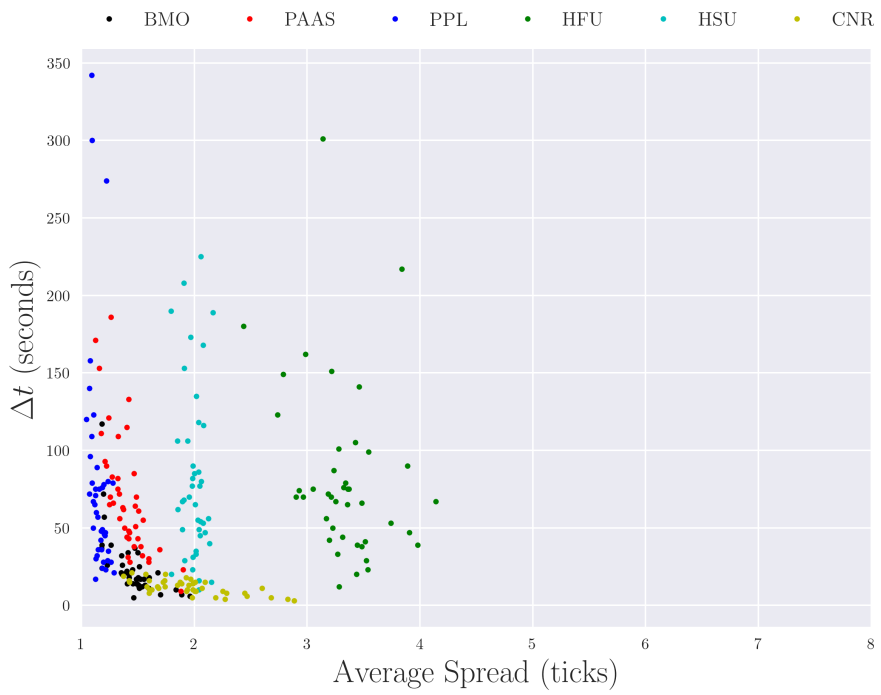


Figure 4.11: Optimal sampling time Δt (seconds) against the average spread (ticks) during the end period of the trading day. Each point corresponds to a trading day between May 29 - August 4, 2017 – excluding May 29, July 4-6, July 24-25, August 4.

and HSU show very different results compared to BMO, PAAS, PPL, and CNR. HSU has a spread of ≈ 2 regardless of time period or day in the 10 week sample, but has no clear relationship with Δt. Similarly, HFU appears to have no relationship between the average spread and Δt, but the average spread decreases over the day like the other stocks. HFU appears to share characteristics of both HSU and our other stocks, whereas HSU is obviously different than the rest.

Figures 4.12, 4.13, and 4.14, show the relationship between Δt and the average interarrival time of all orders for our 6 stocks on all days except for the ones we remove which had unusual behaviour in at least one of our stocks. The figures are for the start, mid, and end periods, respectively. Like the average spread we have clear clustering with the average interarrival time. The general trend is that longer average interarrival times leads to longer sampling times and this relationship becomes stronger as the trading day goes on. During the first period the average interarrival times of all stocks are roughly between 0.2 and 0.5 seconds, regardless of the sampling time. The average interarrival time then increases in the mid and end periods for each stock with the individual stocks still clustering together, but the stock clusters separate from each other.

This surface level look at the optimal sampling time Δt leads to some interesting

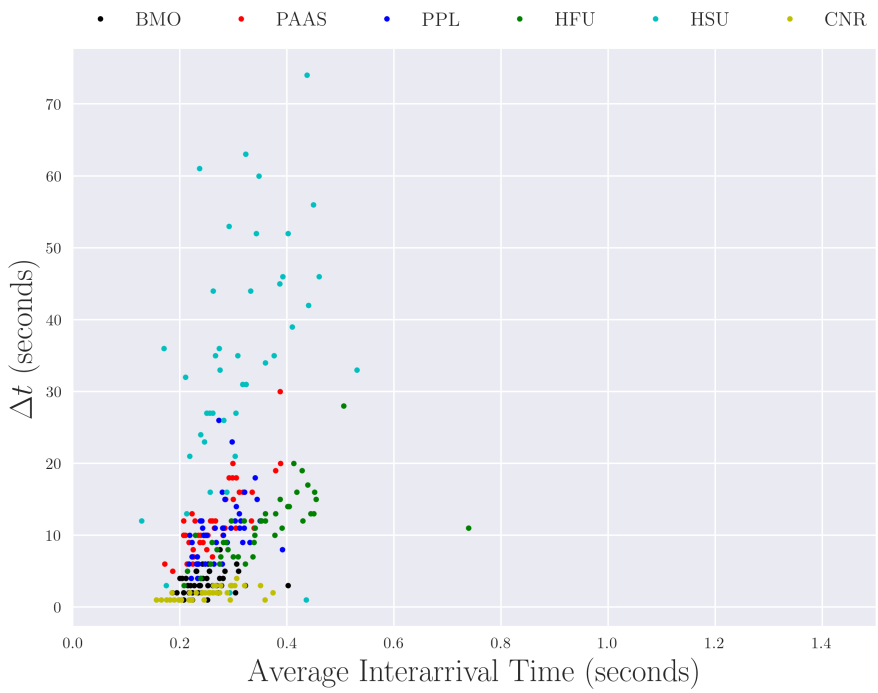


Figure 4.12: Optimal sampling time Δt (seconds) against the average interarrival time (seconds) of all orders during the start period of the trading day. All trading days between May 29 - August 4, 2017 – excluding May 29, July 4-6, July 24-25, August 4.

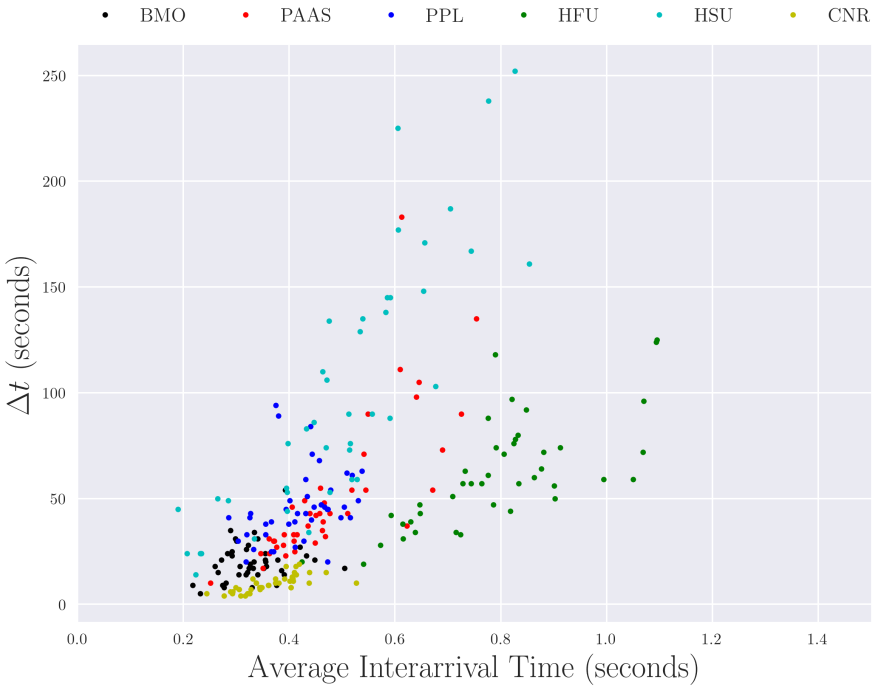


Figure 4.13: Optimal sampling time Δt (seconds) against the average interarrival time (seconds) of all orders during the mid period of the trading day. All trading days between May 29 - August 4, 2017 – excluding May 29, July 4-6, July 24-25, August 4.

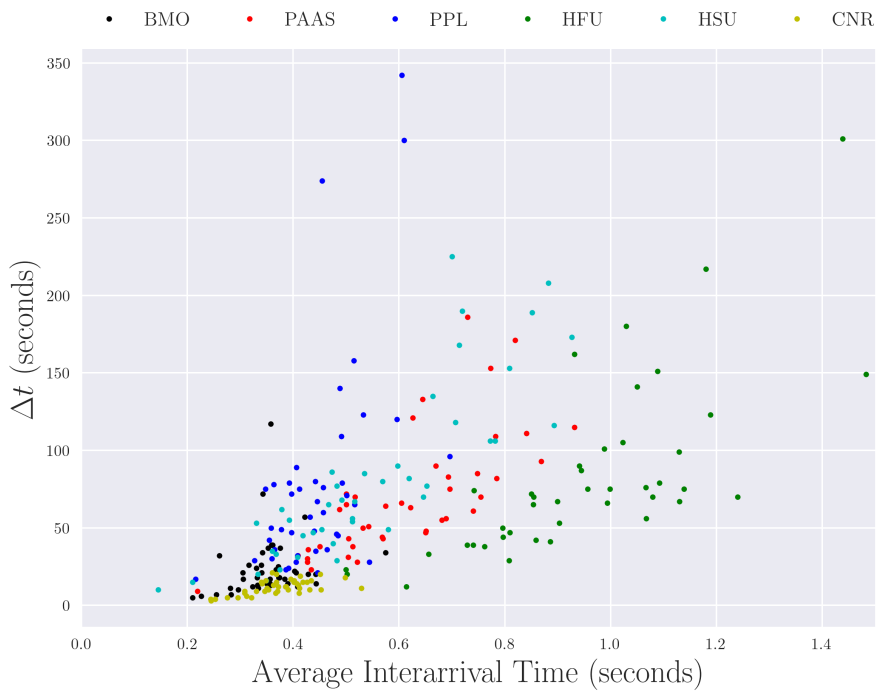


Figure 4.14: Optimal sampling time Δt (seconds) against the average interarrival time (seconds) of all orders during the end period of the trading day. All trading days between May 29 - August 4, 2017 – excluding May 29, July 4-6, July 24-25, August 4.

clustering with the average spread and average interarrival time of orders within and between individual stocks. We can also identify clear outliers in the data which correspond to dividend dates or holidays, but there are also days with unusual results to which we are unable to assign a cause. This could possibly be used to identify days on which the exchange could scrutinize trading behaviour, but we have not seen evidence yet that this is caused by price manipulation. Further investigation would be needed to determine exactly what is causing these days to be so different from any other both within and between stocks.

4.3 Depth of Book

In chapter 3 we defined the parameter K as the depth we take from both sides of the limit order book. So, our limit order book v_k has support in $[-K, K]$ with $v_0 = 0$ and v_1 (v_{-1}) is the volume at the best ask (bid) price. We also made the argument that the price change distribution would then have support in $[-K+1, K-1]$ since the probability of the best ask price moving up one tick will depend on the volumes at the first two ticks. This is because the best ask will move up exactly one tick if only the volume at the first

tick is completely depleted.

From this we can set the depth of the book K after determining the number of ticks in the support of the empirical price change distribution over Δt seconds. Let X denote the empirical price change distribution with $Q_1(X)$ and $Q_2(X)$ as the 0.001% and 99.999% quantiles of X, respectively, defined such that

$$\begin{aligned} \mathbb{P}\left[X \leq Q_1(X)\right] &= 0.001\% \\ \mathbb{P}\left[X \leq Q_2(X)\right] &= 99.999\% \end{aligned} \tag{4.3.1}$$

$Q_1(X)$ and $Q_2(X)$ may not be whole integers so we need to round them to the nearest integer. Then

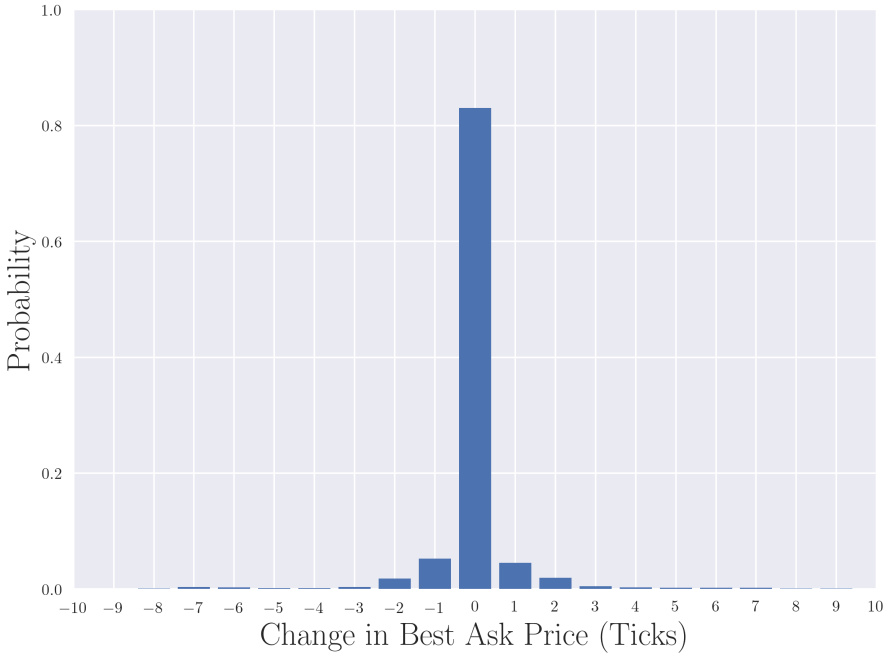
$$K = \max(-Q_1(X), Q_2(X)) + 1 \tag{4.3.2}$$

However, due to the limitations of our AWS cluster, we generated limit order books including only the first 15 prices on either side of the order book. So, if $K > 15$ we take $K = 15$ for calculation purposes. Of the 50 stocks we investigated, only 2 of them had depths larger than 15 after fixing Δt by equation 4.2.1 – FSV and TC.

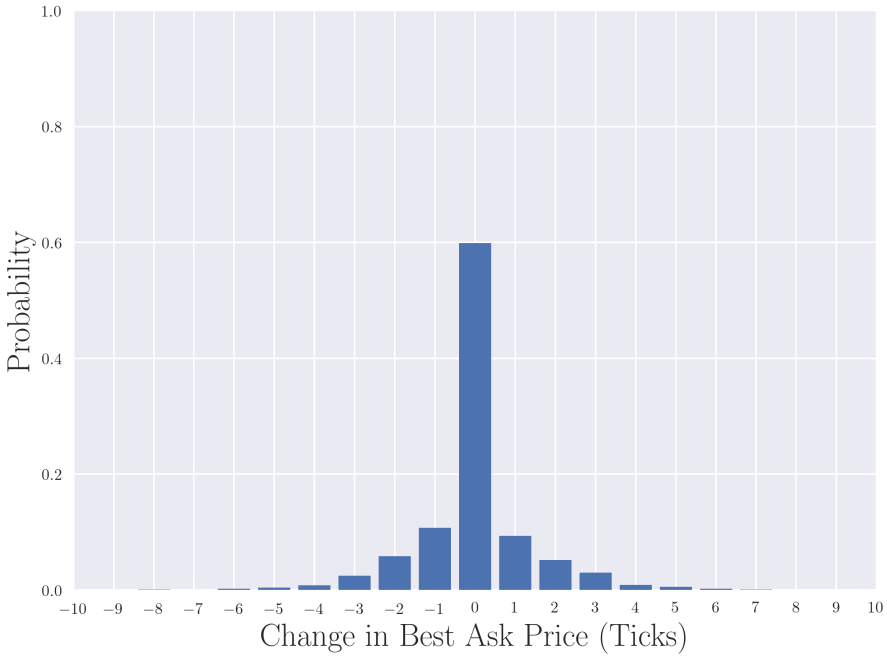
Using equation 4.3.2 we calculate the depth K for stocks TC and AEM from their empirical price change distributions shown in Figure 4.15. Since we have fixed the variance over time interval Δt to be as close to 2 as possible given the data we have that the larger the support of the empirical distribution, the larger the probability that the best ask price does not change over Δt. An example of this is shown in Figure 4.15 where TC has a higher probability of no price movement compared to AEM, which is accompanied by twice the depth. In order for a stock which has a high probability of no price movement to see a variance of 2, the price movements to be large when they actually happen. Like TC, when the price moves it is usually ± 1 or ± 2, but movements of ± 3 to ± 8 ticks occur with almost equal probability to each other. These long tails deep in the support of the price change distribution happen for any of these low movement stocks.

In Figure 4.16 we show the relationship between the depth K and the probability of no change in the best ask price using our collection of stocks from the previous section. As we expect – as the probability of no movement increases to 100%, the depth increases with it since the variance is fixed. The depth then gives us an idea of the frequency in which the stock price moves over its optimal time interval Δt. That is – large depth means higher probability of no movement in the best ask over Δt seconds.

The only clear relationship found between the depth K and other statistics drawn from our data is with the optimal time interval Δt. In Figure 4.17 we see that, in general,



(a) TC, $K = 20$, $\Delta t = 1$



(b) AEM, $K = 10$, $\Delta t = 5$

Figure 4.15: Empirical price change distributions for TC and AEM stocks. Data taken from the week June 1-8, 2017 over the full trading day. From equation 4.3.2 the depth K is 20 and 10 for subplots (a) and (b), respectively.

Figure 4.16: Depth K against the probability of no change in the best ask price for BMO, PAAS, PPL, HFU, HSU, and CNR. Data from all three time periods. Each point corresponds to a trading day between May 29 - August 4, 2017 – excluding May 29, July 4-6, July 24-25, August 4.

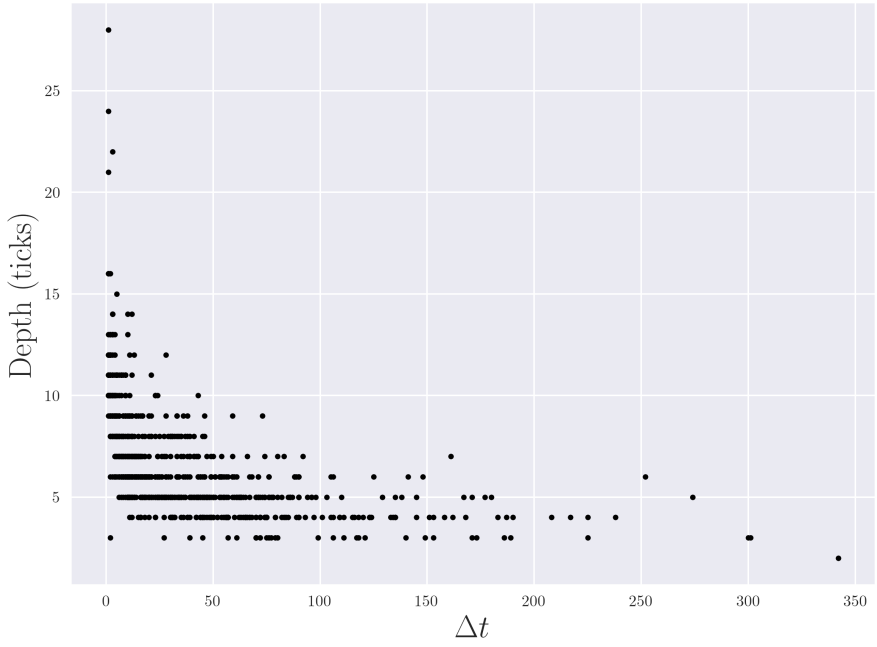


Figure 4.17: Depth K against the optimal time interval Δt for BMO, PAAS, PPL, HFU, HSU, and CNR. Data from all three time periods. Each point corresponds to a trading day between May 29 - August 4, 2017 – excluding May 29, July 4-6, July 24-25, August 4.

price changes which have a variance of 2 over short time scales (less than 5 seconds) also have larger depth – their prices move less frequently, but they move multiple ticks when they do. This is because prices can only move after a new order comes into the book. Order-by-order, the best ask will change with these new incoming orders which is aggregated over Δt seconds. The bigger Δt, the more orders can be included in the time interval to impact the price movements as seen in Figure 4.18.

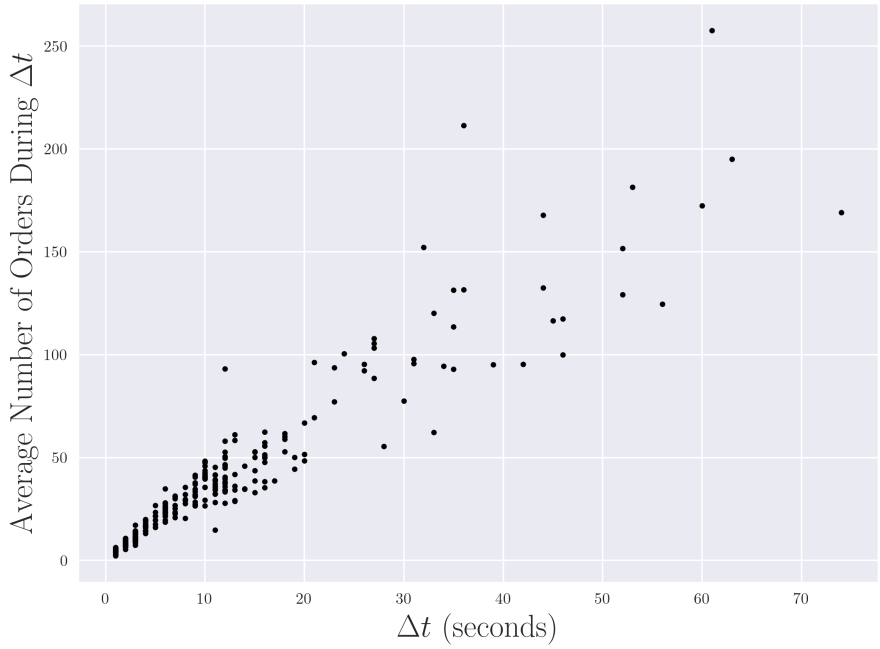


Figure 4.18: Average number of orders during Δt seconds against the optimal time interval Δt for BMO, PAAS, PPL, HFU, HSU, and CNR. Data from the start period. Each point corresponds to a trading day between May 29 - August 4, 2017 – excluding May 29, July 4-6, July 24-25, August 4.

For a small Δt the orders which come into the book would have to have a large impact on the price when they do eventually cause it to move. We see this in Figures 4.19 and 4.18 where a very small Δt comes with a very small number of average orders and a high probability of no change in the best ask price. A high probability of no change in the best ask price means that the price change distribution must have a large support in order to get a variance of 2. In turn, this gives a large depth in the price change distribution.

Likewise, stocks with longer time scales require time for their small price changes to accumulate enough to give a variance of 2 in the distribution of the change in best ask price. These stocks see small price changes order-by-order which add together over Δt to give enough movement to see a variance of 2 in the price change distribution. The longer the time scale, the more orders during Δt and the smaller the probability of no change

Figure 4.19: Average number of orders during Δt seconds against the probability of no change in the best ask price for BMO, PAAS, PPL, HFU, HSU, and CNR. Data from the start period. Each point corresponds to a trading day between May 29 - August 4, 2017 – excluding May 29, July 4-6, July 24-25, August 4.

in the best ask price. In turn, this gives a smaller depth in the price change distribution.

4.3.1 Spread and Price Movement

We also found interesting clustering with the average spread and the probability of no movement in the best ask price. Figure 4.20 shows these clusters for the start period of our six sample stocks. Like the clusters we saw earlier between the spread and Δt, the stocks HFU and HSU are clearly different from the rest. The probability of no movement in the best ask price as no real impact on the spread of HFU or HSU – this is exactly what we saw where the spread for these two stocks is roughly consistent regardless of the other statistics.

However, excluding HFU, we do see the average spread increasing when the best ask price move less and less. This is consistent with what we have seen so far because we have fixed the variance so if a stock has a small probability of price movement, the movement will be large when it happens. Large spreads would facilitate an environment for a stock's best ask price to move multiple ticks at a time when an order does finally cause it to move.

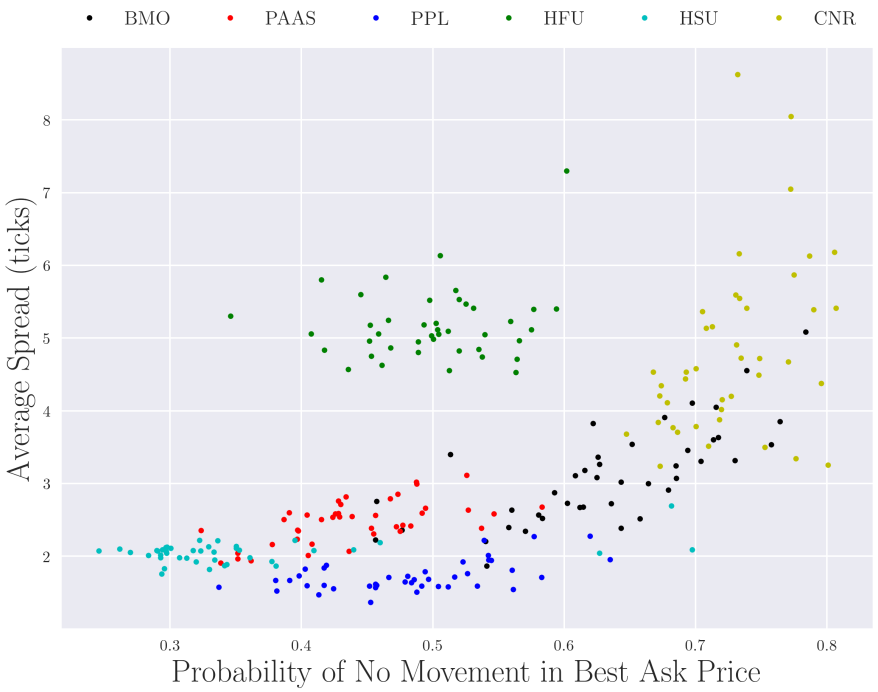


Figure 4.20: Average spread against the probability of no change in the best ask price for BMO, PAAS, PPL, HFU, HSU, and CNR. Data from the start period. Each point corresponds to a trading day between May 29 - August 4, 2017 – excluding May 29, July 4-6, July 24-25, August 4.

4.4 Price and Imbalance Model Calibration

In this section we present two methods for calibrating our price change distribution model to the data – dp^+ and $\vec{w}$. The two methods are maximum likelihood estimation (MLE), and maximum a posteriori estimation (MAP). Details of how these methods are used to fit our model to the data are presented in Appendix B. One can view the MAP estimate as a penalized version of the MLE.

To calibrate our model for dp^+ and $\vec{w}$ we take 20,000 Δt second intervals for dates between June 1-8, 2017. The optimal time interval Δt is calculated from equation 4.2.1. We do the same calibration for the first hour of trading. Sometimes a Δt second interval has only one or no orders in it and we drop such intervals. This happens at most ≈ 400 out of our 20,000 samples. From this data set we can calculate K using equation 4.3.2. The i^{th} Δt second interval then has an associated change in best ask and average imbalance pair denoted by $(x^i, I(\vec{v}^{\,i}; \vec{w}, K))$ for given weights $\vec{w}$ and depth K.

For each stock we then have optimal time interval Δt, depth K, and data pairs $(x^i, I(\vec{v}^{\,i}; \vec{w}, K))$ which we can use to calibrate our model to find dp^+ and $\vec{w}$.

4.4.1 Calibration without Penalty

We first look at maximum likelihood estimation to calibrate our model for exponential and free imbalance weights. Let N denote the number of Δt second intervals and data from the i^{th} interval is denoted by $(x^i, I(\vec{v}^{\,i}; \vec{w}, K))$. We calibrate first using exponential weights as they serve as an intermediate choice between using only the touch and using free weights. Once the model is calibrated using exponential weights we can use the value of the likelihood function to check the results of the free weights – the free weights should produce a likelihood value at least as good as the exponential weights.

We now write down the calibration problem for exponential and free weights using maximum likelihood estimation. For the exponential weights we have that the individual weights $\vec{w}$ are given by

$$w_i = e^{-(i-1)\alpha} \tag{4.4.1}$$

with $\alpha \in [0, \infty)$. The largest weight is assigned to the best bid and best ask volumes. The weights $\vec{w}(\alpha)$ are now entirely parameterized by α. However, we can transform $\alpha \to \bar{\alpha} \in [0, 1)$ to make our calibration more tractable. We make the following transformation

$$\bar{\alpha} = 1 - \frac{1}{1+\alpha}, \quad \alpha = \frac{1}{1-\bar{\alpha}} - 1 \tag{4.4.2}$$

Using maximum likelihood, our calibration problem becomes:

$$\underset{dp^+,\,\bar{\alpha}}{\arg\min}\,\frac{1}{N}\left[-\sum_{i=1}^{N}\log\varphi(x^i;I(\vec{v}^{\,i};\vec{w}\left(\frac{1}{1-\bar{\alpha}}-1\right),K))\right] \tag{4.4.3}$$

with constraints

$$dp_x^+ = dp_{-x}^- \tag{4.4.4}$$

$$\sum_{x=-K+1}^{K-1} dp_x^+ = 1 \tag{4.4.5}$$

$$0 \leq \bar{\alpha} \leq 1 \tag{4.4.6}$$

as maximizing the likelihood is equivalent to minimizing the negative log likelihood. The factor of $1/N$ outside the negative log likelihood provides numerical stability while not affecting the optimization since it is a positive constant.

The calibration problem using maximum likelihood estimation with free weights is then given by

$$\underset{dp^+,\,w}{\arg\min}\,\frac{1}{N}\left[-\sum_{i=1}^{N}\log\varphi(x^i;I(\vec{v}^{\,i};\vec{w},K))\right] \tag{4.4.7}$$

with constraints

$$dp_x^+ = dp_{-x}^- \tag{4.4.8}$$

$$\sum_{x=-K+1}^{K-1} dp_x^+ = 1 \tag{4.4.9}$$

$$\sum_{k=1}^{K} w_k = 1 \tag{4.4.10}$$

$$w_1 \geq w_k \quad \forall k \in [1,K] \tag{4.4.11}$$

The extra constraint in equation 4.4.11 is to ensure that the weight assigned to the best ask/bid is at least as large as any other weight. We want the volume at the best bid/ask to be the most important volume in determining the average imbalance. This way we can also compare our free weight results to the results where we used exponentially

weighting as the exponential weights assign the largest weight to the touch as well. The financial intuition here is that the volumes at the best bid and ask will dictate whether the best ask moves at all, while the volumes at depths beyond the best bid and ask would determine how deep the price moves when it does.

We also found that removing the constraint made the optimization problem unstable and often did not converge to a solution. This may be caused by having too much freedom in the imbalance weights. Adding the constraint solved this issue while also keeping the free weights conceptually similar to the classic and exponential imbalance weight definitions.

Ultimately, we want to use the free weights when exploring spoofing detection, but calibrating with the exponential weights first allows us to numerically check the results of our higher dimensional free weight calibration for consistency.

4.4.2 Calibration with Penalty

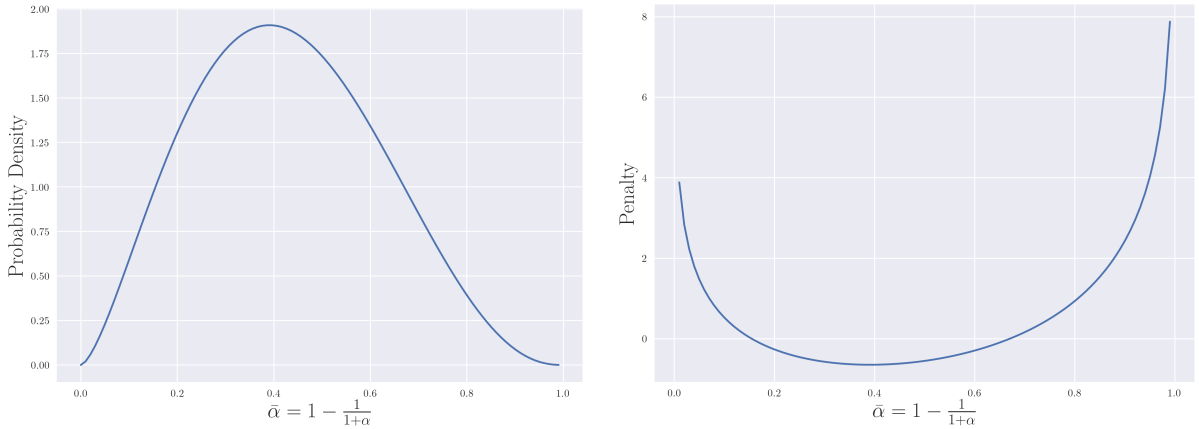
Alternatively, we can take a Bayesian approach and include a penalty function by assuming a prior distribution $P(\bar{\alpha})$ on $\bar{\alpha}$. This is known as maximum a posteriori probability (MAP) estimation and is given by a slight modification to the maximum likelihood estimation. For the exponential weights, an α of 0 would imply that all depths carry equal weight in determining the effect of volume imbalance on price changes - which the literature suggests is not the case [51]. It would also mean a spoofer could place a spoofing limit order so deep in the book it would not be executed while having a large impact on the imbalance. Essentially manipulating the book with no downside. There is also no real difference between an α of 5 or 100 as the exponential weight would decay any contribution beyond the best ask/bid.

That is, we want to penalize α near 0 and as it approaches $+\infty$. After rescaling α as we did in the calibration without penalty we can use a beta distribution $\text{Beta}(\bar{\alpha}; a, b)$ for $P(\bar{\alpha})$ to penalize our bounds at 0 and 1.

Figure 4.21 shows an example penalty function using the beta distribution. We penalize the $\bar{\alpha} = 1$ boundary more harshly because of how we rescaled α as our original α will grow to infinity rapidly for $\bar{\alpha} > 0.8$.

The calibration problem with penalty then becomes

$$\underset{dp^+,\,\bar{\alpha}}{\arg\min}\ \frac{1}{N}\left[-\sum_{i=1}^{N}\log\varphi(x^i; I(\vec{v}^{\,i}; \vec{w}\left(\frac{1}{1-\bar{\alpha}}-1\right), K)) - \log P(\bar{\alpha})\right] \tag{4.4.12}$$



(a) Prior PDF (b) Prior Penalty

Figure 4.21: Subplot (a) is the beta distribution $\text{Beta}(\bar{\alpha}; a, b)$ and subplot (b) is negative log of $\text{Beta}(\bar{\alpha}; a, b)$. We take $a \approx 2.55, b \approx 3.42$ to be the smallest pair where 95% of the probability of $P(\bar{\alpha})$ lies in the interval [0.1,0.8] ($\approx$ [0.11,4.00] in the unscaled α).

with constraints

$$dp_x^+ = dp_{-x}^- \tag{4.4.13}$$

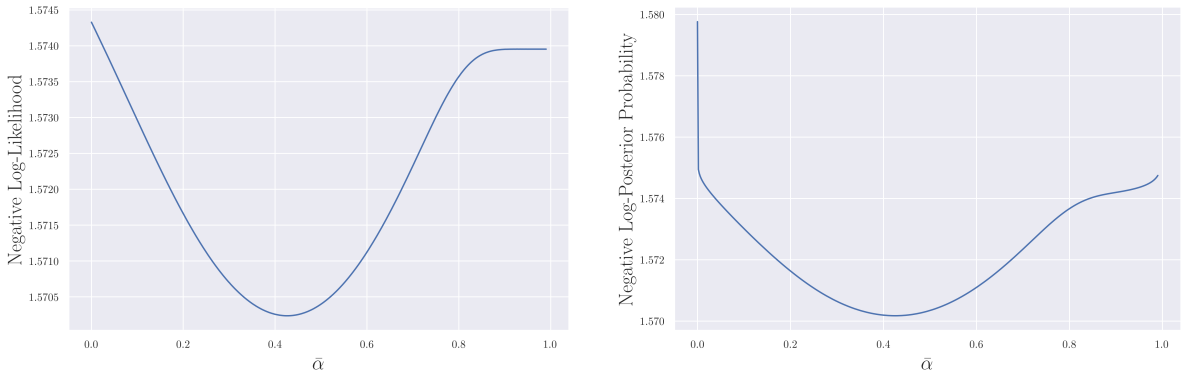
$$\sum_{x=-K+1}^{K-1} dp_x^+ = 1 \tag{4.4.14}$$

$$0 \leq \bar{\alpha} \leq 1 \tag{4.4.15}$$

where $P(\bar{\alpha}) = \text{Beta}(\bar{\alpha}; a, b)$ for suitable constants a and b. As in Figure 4.21, we will take $(a, b) \approx (2.55, 3.42)$ as this is the smallest pair of numbers for which 95% of the probability in $\text{Beta}(\bar{\alpha}; a, b)$ lies in the interval [0.1,0.8]. This way our penalty is almost entirely applied near our boundaries and should have little impact on $\bar{\alpha} \in [0.1, 0.8]$.

Figure 4.22 shows the impact of the MAP estimate over MLE for AEM stock on April 17, 2017 for the entire trading day. There is little difference between the optimal $\bar{\alpha}$ for either method as the minimum value falls in the interval with the smallest penalty, but we have a significant enough penalty to the boundaries of $\bar{\alpha}$.

It should be mentioned that the MAP estimate, unlike the MLE, is not parameterization invariant. This is because the maximum likelihood is a function over the parameter space, while the maximum a posteriori is a probability density over the parameter space. The MAP estimate is the maximum mode of the posterior density which can change with reparameterization as regions of the parameter space can be stretched/contracted



(a) MLE

(b) MAP Estimate

Figure 4.22: MLE and MAP estimate for AEM stock on April 17, 2017 for the entire trading day.

under a non-linear map, but the probability in those regions must be conserved. However, in our calibrations the reparameterization from $\bar{\alpha} \to \alpha$ had minimal impact on location of the optimal α outside of the heavily penalized regions of the parameter space. We also run all calibrations using $\bar{\alpha}$ so we can compare results and any impact on the reparameterization is negligible for our current purposes.

4.5 Statistical Tests for Exponential Weights

Now we can return to the statistical tests we did in chapter 2, but this time we calculate the average imbalance using the weights found from our model calibration. Again, we calculate the p-values and Cramer's V of the test statistic to report the level of association between the average imbalance and changes in the best ask price.

The results for the coarse imbalance and fine price movement test are presented in Table 4.1 where we have included the Cramer's V, $C_{\mathrm{V}}^{\mathrm{Touch}}$ and C_{V}^{α}, obtained by calculating the average imbalance using only the volumes at the touch, and from the optimal weights, respectively. The results in the table are presented for the whole trading day and for the first hour of the trading day. We also include the optimal Δt for each stock.

Table 4.2 presents the results from the fine imbalance and coarse price movement test with the same formatting as Table 4.1.

We aggregate the results in Table 4.1 of the Cramer's V in Figure 4.23. We have plotted $C_{\mathrm{V}}^{\mathrm{Touch}}$ against C_{V}^{α} and compare the results against a relationship where the two values are the same – given by the dashed blue line. We see that there is an overall improvement in the association between the average imbalance and changes in the best

ticker	Δt	α	$C_{\mathrm{V}}^{\mathrm{Touch}}$	C_{V}^{α}
CNR	6	2.16	0.163	0.171
GIL	41	4.05	0.151	0.166
XWD	237	5.48	0.141	0.158
XQQ	126	7.28	0.123	0.158
HXU	38	4.95	0.122	0.136
PPL	22	4.31	0.113	0.113
IMO	21	6.78	0.112	0.122
G	62	0.0202	0.112	0.235
GIB.A	11	1.66	0.109	0.133
FSV	1	0.447	0.108	0.0757
BMO	10	3.39	0.105	0.108
ARX	42	4.31	0.105	0.12
NA	19	4.05	0.105	0.106
T	67	4.61	0.104	0.125
WCN	4	0.559	0.103	0.108
UFS	8	0.419	0.103	0.151
BIP.UN	15	2.69	0.102	0.115
HOD	36	0.00598	0.102	0.246
PAAS	21	1.13	0.101	0.139
FR	74	0.01	0.101	0.167
BAM.A	27	0.608	0.101	0.131
IPL	53	8.02	0.0932	0.109
FM	78	8.18	0.088	0.106
HQU	28	0.525	0.0879	0.182
AEM	5	0.593	0.0875	0.137
RBA	18	1.2	0.0867	0.12
ERF	68	0.01	0.0864	0.201
PWF	46	0.00598	0.0819	0.158
CPG	55	7.42	0.0807	0.0787
FTS	42	5.31	0.08	0.102
K	326	4.94	0.0788	0.0928
ZEB	180	0.01	0.0751	0.207
SLF	18	3.95	0.0723	0.0764
IMG	281	5.74	0.0665	0.0792
KL	78	6.77	0.0629	0.0713
CCO	122	0.01	0.062	0.138
OR	37	0.01	0.0602	0.147
HVI	88	0.012	0.0592	0.0753
TC	1	0.512	0.0556	0.0804
GOOS	2	1.14	0.0555	0.0593
POW	67	7.01	0.0539	0.0658
HSU	58	0.01	0.0514	0.245
SW	5	4.26	0.0505	0.0588
VUN	179	0.399	0.0453	0.238
VFV	112	0.00799	0.0438	0.109
SSO	61	0.00398	0.0304	0.105
HFU	34	0.01	0.0278	0.123
XEG	246	0.00598	0.0245	0.308
VGG	190	0.0202	0.0238	0.181
PVG	45	0.00398	0	0.114

ticker	Δt	α	$C_{\mathrm{V}}^{\mathrm{Touch}}$	C_{V}^{α}
CNR	1	1.46	0.0963	0.108
GIL	10	1.66	0.123	0.155
XWD	109	6.32	0.121	0.149
XQQ	18	6.11	0.175	0.196
HXU	11	6.21	0.243	0.261
PPL	9	3.04	0.106	0.0995
IMO	7	2.74	0.0653	0.0692
G	20	0.01	0.0843	0.245
GIB.A	3	1.08	0.0873	0.108
FSV	1	0.247	0.119	0.0818
BMO	3	1.59	0.086	0.0872
ARX	15	5.73	0.102	0.106
NA	6	4.61	0.07	0.0673
T	13	2.22	0.0969	0.125
WCN	1	0.263	0.0604	0.0906
UFS	2	0.656	0.097	0.117
BIP.UN	5	0.791	0.0727	0.1
HOD	29	0.00799	0.0556	0.32
PAAS	8	1.82	0.107	0.143
FR	24	0.0202	0.0417	0.0862
BAM.A	7	1.13	0.091	0.125
IPL	18	6.89	0.0624	0.075
FM	19	6.48	0.0283	0.0328
HQU	1	2.3	0.132	0.135
AEM	2	0.549	0.112	0.146
RBA	5	0.934	0.0875	0.131
ERF	25	5.51	0.114	0.12
PWF	18	5.92	0.0699	0.0789
CPG	21	7.42	0.0871	0.0862
FTS	13	2.26	0.0872	0.0894
K	99	5.73	0.168	0.185
ZEB	42	0.0161	0.139	0.211
SLF	5	3.59	0.0942	0.0991
IMG	57	0.981	0.0261	0.0786
KL	21	6.77	0.131	0.137
CCO	27	6.22	0.102	0.108
OR	14	0.00598	0.0722	0.0948
HVI	28	0.00799	0.0354	0.104
TC	1	0.878	0.0388	0.0483
GOOS	1	3.59	0.051	0.0685
POW	20	6.11	0.0948	0.109
HSU	19	0.01	0.0414	0.253
SW	1	0.823	0.0558	0.0926
VUN	74	0.226	0.0292	0.332
VFV	45	0.00799	0.1	0.125
SSO	15	3.01	0.0558	0.083
HFU	8	0.01	0.0152	0.179
XEG	75	0.00598	0.0547	0.351
VGG	76	0.354	0.0874	0.182
PVG	15	0.0202	0	0.0254

Table 4.1: Summary of chi square test for coarse imbalance and fine price movements over Δt seconds. Data is taken from June 1-8, 2017. All p-values are zero or very close ($\approx 10^{-46}$ at most) to zero. The left subtable uses data from the entire trading day while the right subtable uses data only from the first hour of the trading day. Tickers are sorted by magnitude of Cramer's V for the whole day.

ticker	Δt	α	$C_{\mathrm{V}}^{\mathrm{Touch}}$	C_{V}^{α}
CNR	6	2.16	0.284	0.286
GIL	41	4.05	0.281	0.277
GIB.A	11	1.66	0.264	0.278
XQQ	126	7.28	0.22	0.221
HXU	38	4.95	0.205	0.205
XWD	237	5.48	0.2	0.201
IMO	21	6.78	0.193	0.193
T	67	4.61	0.191	0.192
BIP.UN	15	2.69	0.19	0.199
RBA	18	1.2	0.188	0.205
G	62	0.0202	0.187	0.278
HQU	28	0.525	0.186	0.212
PPL	22	4.31	0.182	0.184
BMO	10	3.39	0.177	0.178
UFS	8	0.419	0.177	0.245
WCN	4	0.559	0.175	0.224
BAM.A	27	0.608	0.172	0.167
ARX	42	4.31	0.166	0.169
FM	78	8.18	0.165	0.165
FSV	1	0.447	0.164	0.136
AEM	5	0.593	0.163	0.204
NA	19	4.05	0.162	0.161
PAAS	21	1.13	0.153	0.189
IPL	53	8.02	0.15	0.15
HOD	36	0.00598	0.149	0.319
SW	5	4.26	0.147	0.153
GOOS	2	1.14	0.144	0.164
FTS	42	5.31	0.139	0.14
PWF	46	0.00598	0.137	0.178
ERF	68	0.01	0.123	0.236
SLF	18	3.95	0.116	0.12
CCO	122	0.01	0.113	0.166
FR	74	0.01	0.106	0.196
VUN	179	0.399	0.103	0.27
CPG	55	7.42	0.102	0.102
TC	1	0.512	0.0999	0.141
KL	78	6.77	0.0938	0.0935
OR	37	0.01	0.0843	0.207
IMG	281	5.74	0.0808	0.0835
ZEB	180	0.01	0.0787	0.193
HFU	34	0.01	0.0782	0.163
K	326	4.94	0.0728	0.0673
POW	67	7.01	0.0698	0.0697
VFV	112	0.00799	0.0692	0.107
HSU	58	0.01	0.066	0.272
SSO	61	0.00398	0.0656	0.129
HVI	88	0.012	0.0624	0.0552
VGG	190	0.0202	0.0497	0.201
XEG	246	0.00598	0.0286	0.357
PVG	45	0.00398	0.0277	0.0995

ticker	Δt	α	$C_{\mathrm{V}}^{\mathrm{Touch}}$	C_{V}^{α}
CNR	1	1.46	0.29	0.289
GIL	10	1.66	0.256	0.293
GIB.A	3	1.08	0.273	0.287
XQQ	18	6.11	0.32	0.319
HXU	11	6.21	0.43	0.429
XWD	109	6.32	0.173	0.17
IMO	7	2.74	0.132	0.128
T	13	2.22	0.202	0.202
BIP.UN	5	0.791	0.162	0.207
RBA	5	0.934	0.244	0.298
G	20	0.01	0.121	0.336
HQU	1	2.3	0.255	0.252
PPL	9	3.04	0.141	0.135
BMO	3	1.59	0.174	0.177
UFS	2	0.656	0.193	0.234
WCN	1	0.263	0.187	0.253
BAM.A	7	1.13	0.189	0.206
ARX	15	5.73	0.162	0.164
FM	19	6.48	0.0609	0.06
FSV	1	0.247	0.182	0.0998
AEM	2	0.549	0.236	0.257
NA	6	4.61	0.121	0.124
PAAS	8	1.82	0.199	0.217
IPL	18	6.89	0.141	0.141
HOD	29	0.00799	0.0881	0.393
SW	1	0.823	0.202	0.226
GOOS	1	3.59	0.162	0.166
FTS	13	2.26	0.142	0.151
PWF	18	5.92	0.111	0.112
ERF	25	5.51	0.168	0.166
SLF	5	3.59	0.174	0.174
CCO	27	6.22	0.161	0.163
FR	24	0.0202	0.0527	0.0872
VUN	74	0.226	0.0393	0.383
CPG	21	7.42	0.122	0.122
TC	1	0.878	0.103	0.0715
KL	21	6.77	0.235	0.234
OR	14	0.00598	0.124	0.141
IMG	57	0.981	0.0565	0.0351
ZEB	42	0.0161	0.171	0.222
HFU	8	0.01	0.0654	0.241
K	99	5.73	0.239	0.241
POW	20	6.11	0.115	0.114
VFV	45	0.00799	0.108	0.15
HSU	19	0.01	0.0859	0.319
SSO	15	3.01	0.106	0.112
HVI	28	0.00799	0.0399	0.105
VGG	76	0.354	0.124	0.206
XEG	75	0.00598	0.0577	0.421
PVG	15	0.0202	0.0274	0.0351

Table 4.2: Summary of chi square test for fine imbalance and coarse price movements over Δt seconds. Data is taken from June 1-8, 2017. All p-values are zero or very close ($\approx 10^{-46}$ at most) to zero. The left subtable uses data from the entire trading day while the right subtable uses data only from the first hour of the trading day. Tickers are sorted by magnitude of Cramer's V for the whole day.

ask price for this statistical test across all stocks, with the exception of the stock FSV who's points lie well below the dashed blue line.

FSV is a stock with an average spread of over 25 ticks, $\Delta t = 1$ second, and an average interarrival time of ≈ 0.6 seconds. Running the calibration multiple times often yielded different results each time depending on the time intervals we sampled over the week. FSV is a very inactive stock which would require special care in that we likely need more data and sample more points. It may also be that our restriction of looking at stocks over time intervals where they show a variance of 2 ticks in the change in the best ask price is unfair to stocks which move very little or move many ticks at a time when they do move. This was the only stock encountered in our list that exhibited this behaviour, but there are very likely more stocks like this as we look at less active stocks.

Aside from FSV we have the greatest improvement in the Cramer's V with our calibration on stocks with small α values. This makes sense as these are stocks where the association with the imbalance increases as we take volumes deeper in the book. Some stocks have an optimal α which just reproduces the imbalance taking only the touch as these stocks are more likely to have their price dynamics determined by the volumes at the touch.

Similarly, we aggregate the results in Table 4.2 of the Cramer's V in Figure 4.24. We see the same improvement from stocks in each α category as in Figure 4.23 with FSV being the only outlier again for the same reasons stated before.

Both statistical tests show that there are some stocks which gain significant improvement in the association between the average imbalance and changes in the best ask price while others see little to no benefit. This may imply there are some stocks which have their price dynamics influenced more heavily than others by volumes deeper in the book than by volumes only at the touch. This dependency on the depth of book may make these particular stocks more vulnerable to manipulation. Prices could still be manipulated at the touch, but it would be a much riskier strategy in our model.

The take away here is that we see a statistically significant increase in our desired association by capturing information deeper in the limit order book for some stocks. It is also good to see some stocks which gain nothing from orders placed deeper in the book and whose price dynamics are most associated with the touch. It would have been very surprising to see an across the board improvement in all stocks using our model, so we are at least able to say that the existing literature involving only the touch is still very relevant across most of the stocks we presented in Tables 4.1 and 4.2.

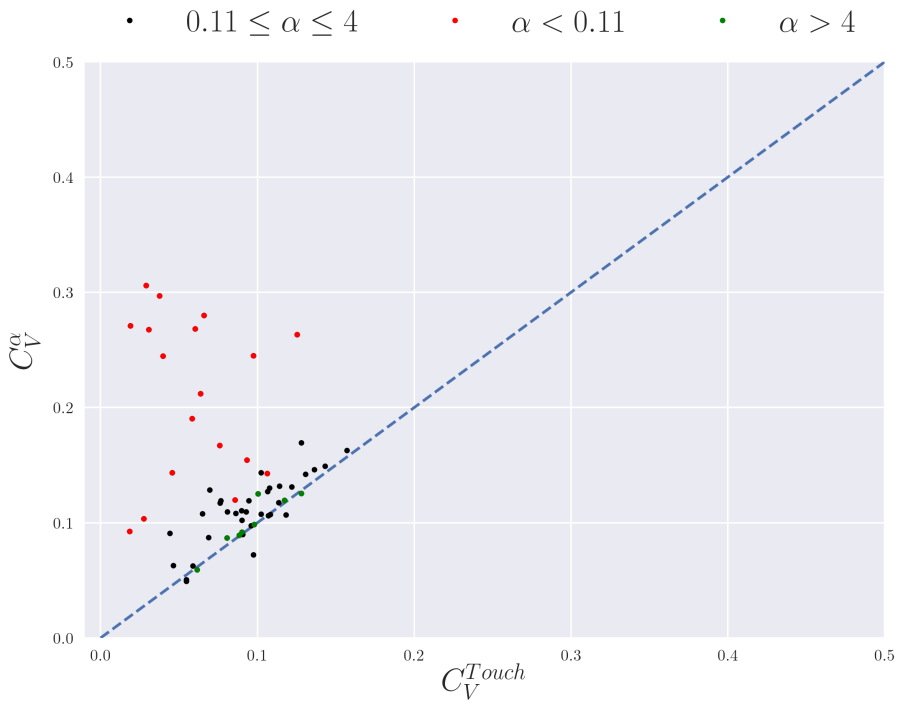


Figure 4.23: $C_{\mathrm{V}}^{\mathrm{Touch}}$ against C_{V}^{α} for chi square test for coarse imbalance and fine price movements. The blue dashed line would be the relationship that both Cramer's V values are the same. We separate the Cramer's V values by the optimal α of each stock. The intervals are chosen based on the regions we assigned our penalties to in Figure 4.21.

4.6 Calculating Average Imbalance Over Δt

Returning to a point we left in chapter 2 – we decided to calculate the average imbalance using time weighting instead of taking a simple mean. Now that we have an algorithm for determining our sampling time Δt we can come back to another reason for choosing the time weight over the simple mean, in addition to the reasons presented back in chapter 2.

When calculating the correlation between the average imbalance and the change in the best ask price using a simple mean for the average imbalance we often found negative correlations between these two variables after calibrating for the optimal α. Regardless of the α picked, we would always see a negative correlation even after resampling from the larger data set. This might not be too surprising if the optimal α was found to be small so as to incorporate volumes deeper in the book when calculating the imbalance, but this was the case even for stocks with large α where only the touch is used in the calculation. No matter the model we should find a positive correlation between the volumes at the touch and the price dynamics because the best ask can never increase without the volume

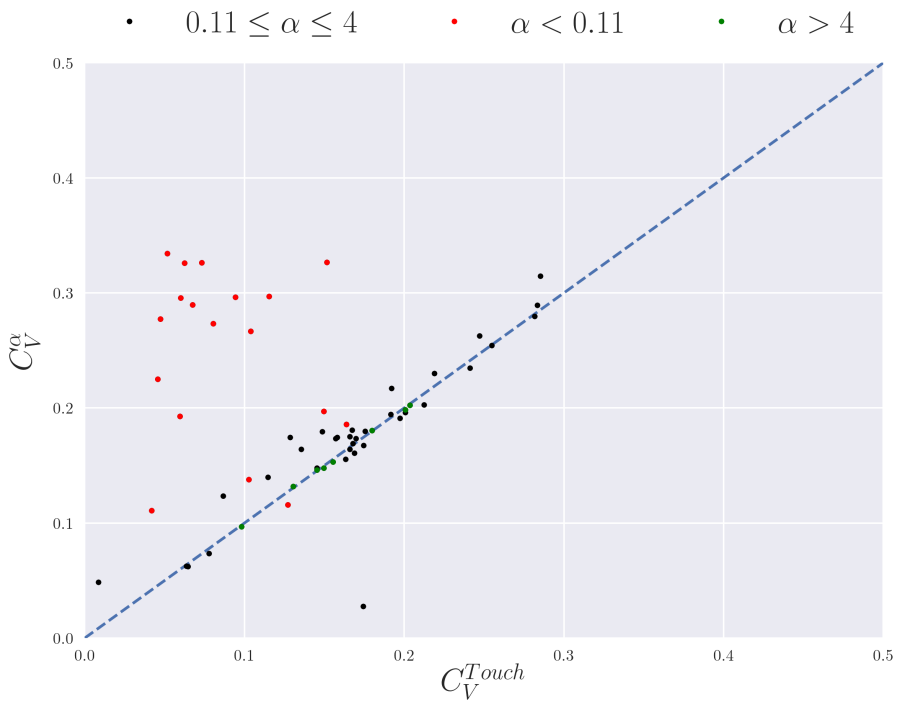


Figure 4.24: $C_{\mathrm{V}}^{\mathrm{Touch}}$ against C_{V}^{α} for chi square test for fine imbalance and coarse price movements. The blue dashed line would be the relationship that both Cramer's V values are the same. We separate the Cramer's V values by the optimal α of each stock. The intervals are chosen based on the regions we assigned our penalties to in Figure 4.21.

at the best ask going to zero so the best ask moves to the next highest price with a non-empty volume. Also we should see prices almost never decreasing unless the best bid volume has been depleted. Of course, if the spread is large enough we could have new limit orders come in below the best ask which causes it to decrease, but even stocks with tight spreads of a 1 tick would appear to have negative correlations between imbalance and price movements.

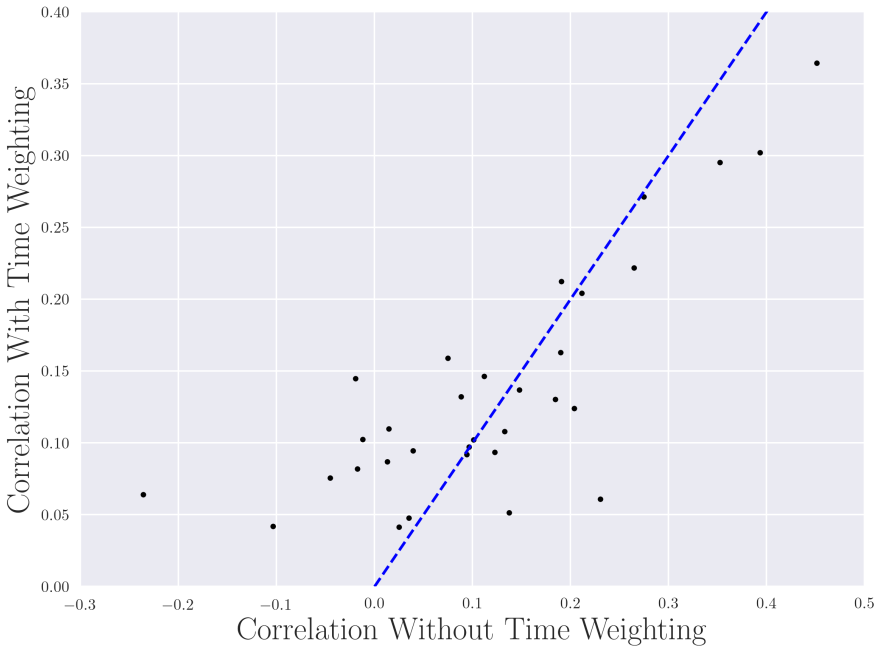


Figure 4.25: Comparison between average imbalance with and without time weighting captured by the correlation between the average imbalance and the change in best ask price. The average imbalance is calculated using the calibrated α. Data is taken from June 1-8, 2017 for the entire trading day across the stocks presented in Table 4.1. The dashed blue line would represent identical correlation using both methods.

In Figure 4.25 we show the two methods for calculating the average imbalance and their impact on the correlations between this average imbalance and price movements. The time weighting approach completely removed the negative correlations we would find after our calibrations. This is not to say that the time weighting always increased correlation. There are stocks that had their correlations decreased, but still positive. Our arguements back in chapter 2 were statistical and financial in nature, but here we can see the impact of one method over the other in finding a predictor of price dynamics which is consistent across all stocks we investigated – that is we no longer found any stocks with strong negative correlations. We would still find some stocks with very small negative correlations (≈ -0.02, for example), but we would argue this is really no different than

a very small positive correlation in our case and that we would find no correlation at all if taking a sufficient number of samples.

4.7 dp^+ Goodness of Fit

We would like to have a goodness of fit for the calibrated dp^+ for each stock when compared to the respective empirical distribution of price movements. We refer to the calibrated distribution dp^+ here because the contraints we apply to our price change model dp made it fully determined by dp^+ and the imbalance weights w_i. After the calibration we have all the model parameters which we can plug into equation 3.4.5 to get our fitted distribution for the change in the best ask price given some average imbalance I. To compare our fitted distribution to the empirical we then need an estimate of I over the entire relavent time period. We simply take the mean of the average imbalances calculated for each Δt time sample as our estimate for the stock's overall average imbalance.

We need a way to compare two discrete distributions – one way involves the Kullback–Leibler (KL) divergence [65] in equation 4.7.1.

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \tag{4.7.1}$$

where $P(x)$ and $Q(x)$ are two discrete distributions over the support $\mathcal{X}$. Typically, $P(x)$ is taken to be the empirical distribution and $Q(x)$ is the fitted distribution. We should note that the KL divergence is not a metric on the space of probability distributions as it is not symmetric and it does not satisfy the triangle inequality. Instead, the KL divergence is a measure of how much information is lost when approximating P by Q. The KL divergence is zero for identical distributions and is unbounded to positive infinity.

Figure 4.26 presents a histogram of the KL divergence for the stocks presented in Tables 4.1 and 4.2. Overall the KL divergence is small across all stocks and provides evidence of a good fit for our model price change distribution.

We can also show the fitted distribution to the empirical distribution as in Figure 4.27. The fitted distributions provide an excellent approximation to the empirical distribution except for at ± 1 and ± 2 ticks. This was not the case across all stocks, but for many the empirical distribution was not symmetric about 0 when not conditioning on the average imbalance. Any discrepancies we found with our fits were likely because of the modelling assumption that the change in the best ask price is a symmetric distribution. Returning to Figures 2.16a and 2.17 we saw that order-by-order the probability of the best ask

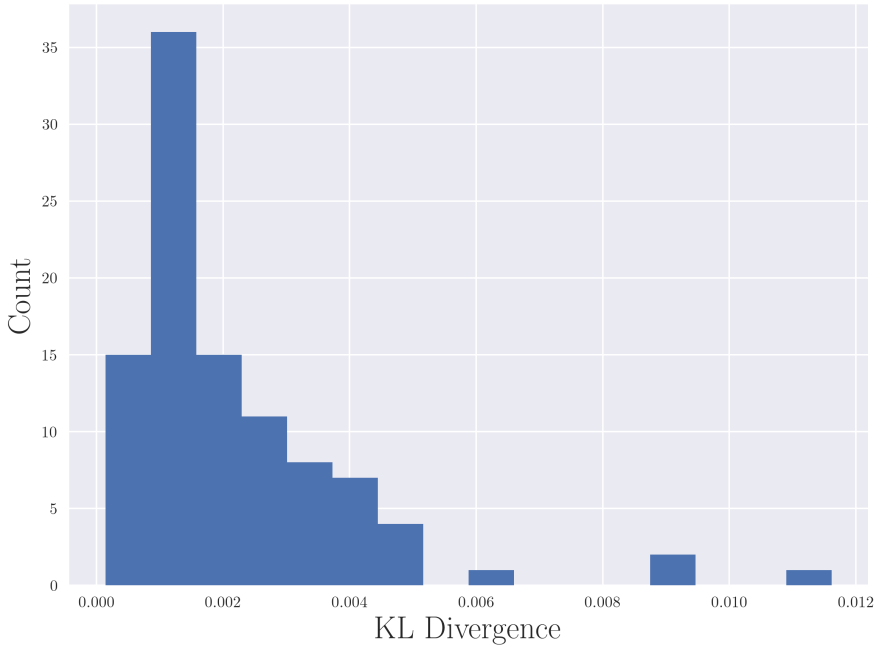


Figure 4.26: Histogram of the KL divergence calculated for each stock after calibration. We are comparing the empirical price change distribution to the fitted distribution with the average imbalance taken to be the mean of the sampled average imbalances.

increasing or decreasing was not symmetric about $I = 0$. This asymmetry extends over Δt as in Figure 2.16.

Figure 4.28 shows the probability-probability (PP) plot of the distributions in Figure 4.27. A PP plot displays the value of the cumulative distribution function at each tick for the empirical and fitted price change distributions against each other. A perfect fit to the data would have all black points lying on the dashed blue line. Again, here we can see the two distributions do not completely match at ticks ± 1 and ± 2. However, we get a very good approximation in the tails. We use AEM here as a representative plot because none of our other fits are any worse than the one we present here.

Future work would be to incorporate the imbalance asymmetry in the model to see if this remedies the problem in our distribution fit, but for now we at least have a fairly accurate model for how the distribution of changes in the best ask price are impacted through the average imbalance. Though we may not have a perfect fit at all ticks, we do have a great fit deep in the tails of the distribution so that the least likely price movements will be properly incorporated in the cost functions we derived in chapter 3. It is important to fit the least likely events to properly capture the profits and losses the spoofer realizes in the worst (large price increase) or best (large price decrease) case scenarios. These events may be unlikely, but a spoofer would want to have them incorporated in their

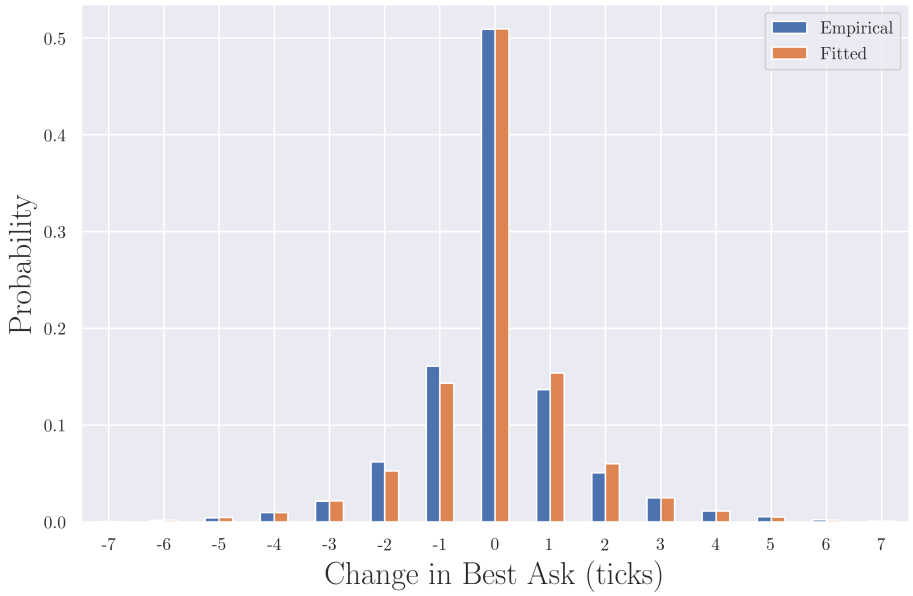


Figure 4.27: Empirical and fitted price change distributions for AEM stock on April 17, 2017 over the entire trading day. The average imbalance is the mean of the sampled average imbalances.

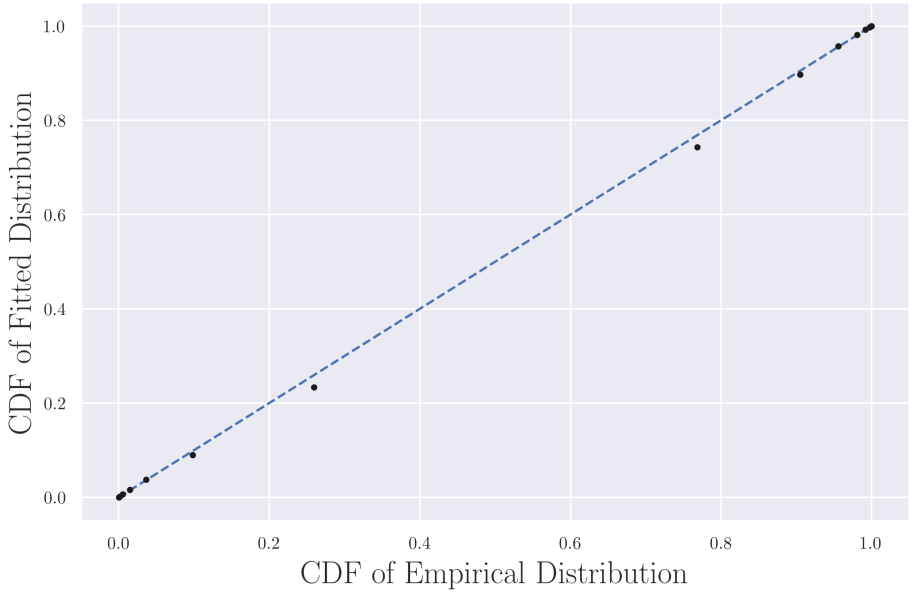


Figure 4.28: Probability-probability plot for the two distributions in Figure 4.27. The CDF of the two distributions is evaluated at each tick in the support and plotted against each other. The points would lie on the dashed blue line if the two distributions were identical.

expected costs to properly account for the risks associated with placing spoofing orders deep in the book. This would allow the spoofer to, in theory, make better decisions.

4.8 Exponential and Free Imbalance Weights

Tables 4.1 and 4.2 show that there are many stocks with $\alpha \approx 0$ even after the penalty. This value of α would imply that all volumes up to depth K carry roughly equal weight when determining how to calculate the average imbalance which has the highest association to changes in the best ask price. It is unlikely all weights would be equally important, but due to our choice of exponentially decreasing weights if a stock has a significant enough correlation to volumes deep in the book then $\alpha \approx 0$ is the only choice in order to capture that volume.

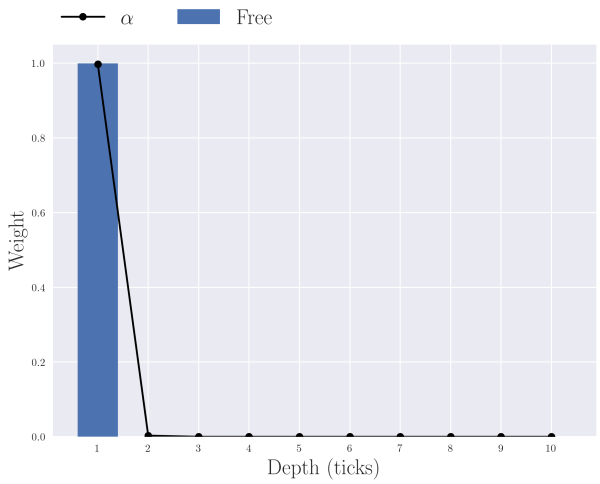
We now present the results for the free weights using the stocks BMO, CNR, HFU, HSU, PAAS, and PPL calibrated weekly between May 29, 2017 and August 04, 2017 for the entire trading day. We will then compare the free and exponential weights from the two calibrations and then return to the two statistical tests using the free weights.

In Figure 4.29 we show free and exponential weights determined by our calibrations for BMO, CNR, and HSU stocks during the week of May 29 - June 2, 2017 over the entire trading day. We use these three examples to show the outcome of the free calibration when we have a large, medium, and small α, respectively. In each case the free weights still pick up the exponentially decaying portion near the touch, but we also see weights deeper in the order book for CNR and HSU. Even with the free weights BMO still only wants the volume at the touch for best determining movements in the best ask price, while CNR and HSU benefit from information several ticks into the book.

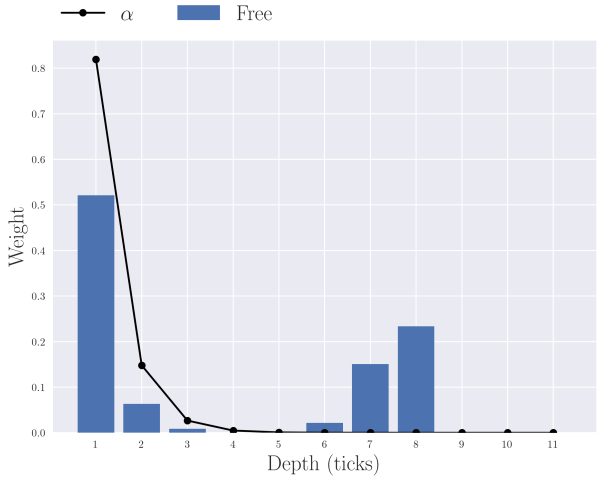
CNR has free weights deep in the book which we cannot capture with the exponential weights without giving more weights to the volumes that have a weaker association to the price dynamics – you would not be able to include the weights at 6, 7, or 8 ticks without giving increasing weight to ticks 4 and 5, for example. However, this is not an issue for HSU as the weights are so strong a few ticks into the book that the unimportant volumes (ticks 2, 7, and 8) are overshadowed by the increased association in the price dynamics we get from including the several volumes that do matter (ticks 3,4,5, and 6).

We see the same patterns emerge in each of the stocks over the 10 week period. We just present these three examples to showcase the three 'regimes' (small, medium, and large α) of α when calibrating with free weights.

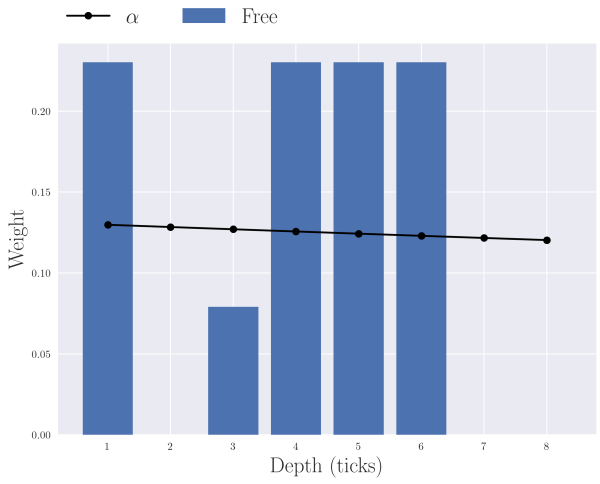
In Figures 4.30 and 4.31 we plot the Cramer's V for our two statistical tests using the free weights as well as the exponential weights. We see that the free weights either give



(a) BMO, $\alpha \approx 5.79$



(b) CNR, $\alpha \approx 1.71$



(c) HSU, $\alpha \approx 0.0108$

Figure 4.29: Free and exponential weights for BMO, CNR, and HSU calibrated using data from May 29 - June 2, 2017 over the entire trading day.

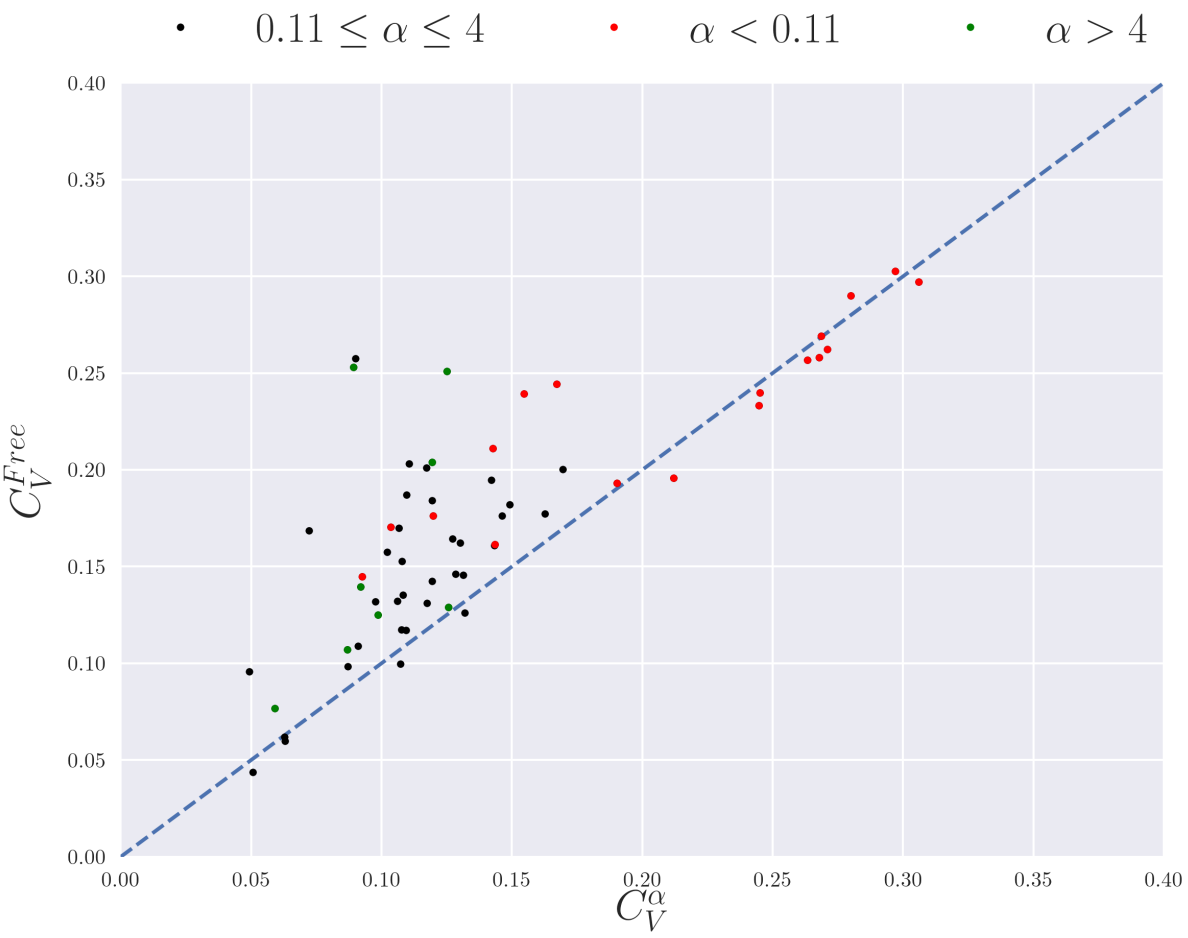


Figure 4.30: $C_{\mathrm{V}}^{\mathrm{Free}}$ against C_{V}^{α} for chi square test for coarse imbalance and fine price movements. The blue dashed line would be the relationship that both Cramer's V values are the same. We separate the Cramer's V values by the optimal α of each stock. The intervals are chosen based on the regions to which we assigned our penalties in Figure 4.21.

back the same results we had originally or they greatly improve the association of the average imbalance and changes in the best ask price. Clearly, the free weights provide a better association as they will be at least as good as the exponential weights while also giving more flexiability in which volumes are incorporated in calculating the average imbalance. Interestingly, we also see that the improvements are not isolated to one α category as even stocks with a large α found improvements by including volumes deeper in the book.

The first test in Figure 4.30 has some stocks with no improvement over the exponential weights as this test is only against the sign of the average imbalance in predicting price movements – the changes to the weights may not have any statistically significant change

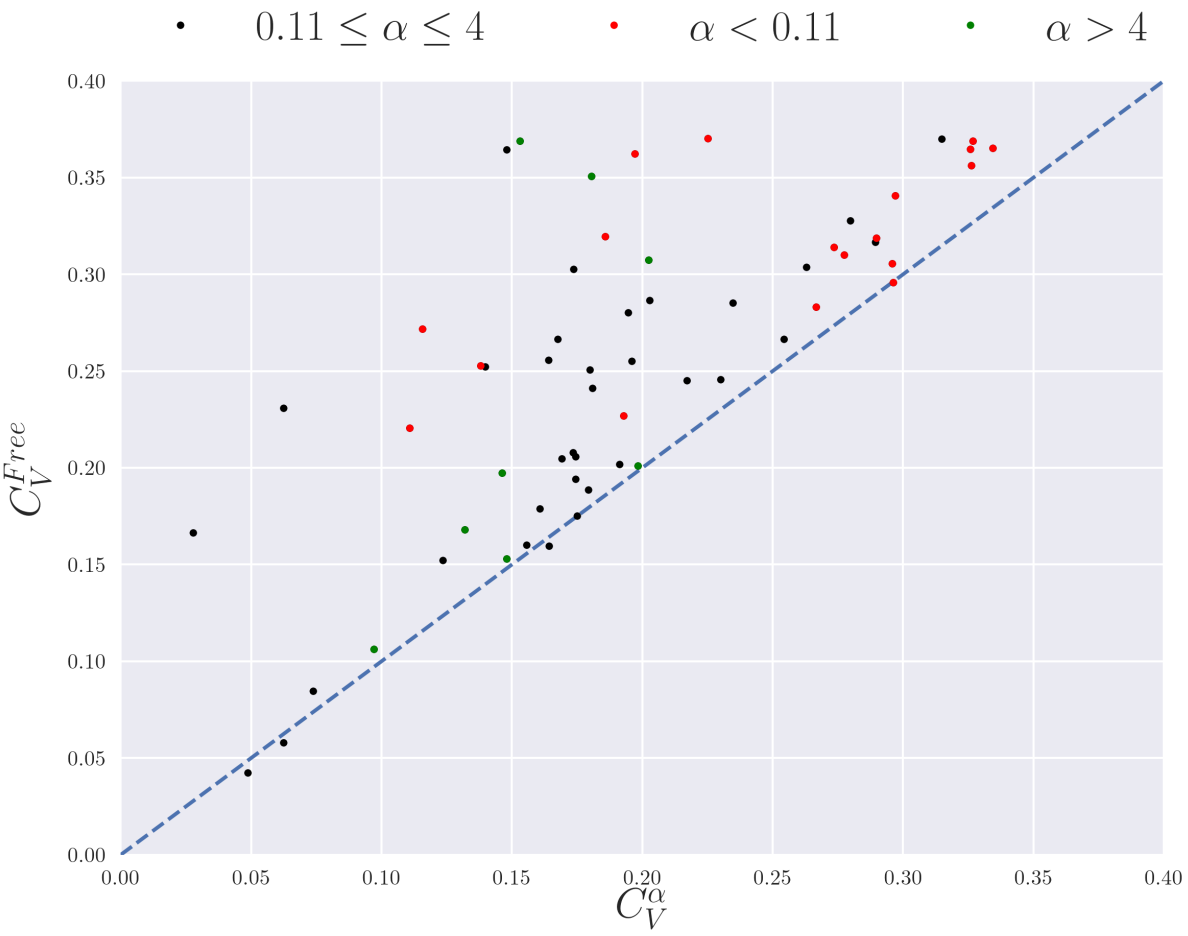


Figure 4.31: $C_{\mathrm{V}}^{\mathrm{Free}}$ against C_{V}^{α} for chi square test for fine imbalance and coarse price movements. The blue dashed line would be the relationship that both Cramer's V values are the same. We separate the Cramer's V values by the optimal α of each stock. The intervals are chosen based on the regions to which we assigned our penalties in Figure 4.21.

to the sign of the imbalance calculated from either weight choice. However, Figure 4.31 shows the second test where we broke the average imbalance into bins. Changing the weights can change the average imbalance enough that our observations are moved from one bin to another. The increased granularity of the imbalance in the second statistical test can then pick up improvements in how the imbalance is calculated when there is no change in sign. This could be why we see such huge improvements in Figure 4.31 over Figure 4.30.

4.9 Conclusions

In this chapter we introduced an algorithm for fitting our model parameters $\vec{w}$, K, and dp^+ to stock data in a given time period over some time interval Δt. In order to compare the results across stocks we decided to fix Δt for each stock based on a benchmark variance in changes in the best ask price. This way we are comparing all stocks over time intervals where they have approximately the same variance. We then found relationships between Δt, the average spread, and average interarrival time of orders which also cluster for each stock during different periods of the day. Increasing spreads and decreasing average interarrival times are associated with decreasing Δt. We also found that we could spot outliers within the data using these parameters and stock statistics which were linked to holidays in Canada and the United States and dividend dates. There were other outliers which we were unable to pin to any specific event associated with the stock.

With a time interval Δt set we could discuss price movements. We then defined the depth of book K based on the support of the distribution of the change in the best ask price. Since we had fixed the variance we had forced a relationship between K and the probability of no change in the best ask price. We then found increasing depth was associated with decreasing Δt which we motivated based on the number of orders a stock receives during Δt and how this related to movements in the stock price. We also found clusters for each stock during different time periods using the average spread and the probability of no movement in the best ask price.

With some meaning assigned to our model parameters we then calibrated our model to stock data using MAP estimation – maximum likelihood estimation with penalty. Here we assumed the weights $\vec{w}$ decayed exponentially as one included more depth in the book and were parameterized by α – the exponential decay constant. With our calibrated model parameters we repeated the two statistical tests we ran in chapter 2 for the average imbalance calulated from α. We found improvements across many stocks over using only the volume at the touch, but mostly for stocks with smaller α. This would be because the smaller the α the more volumes we incorporate deeper in the limit order book.

We also returned to the point of why we used time weighting over taking a simple mean for calulating the average imbalance. We found that we often had negative correlations between the volume imbalance and changes in the best ask price even when using just the volumes at the touch. This was against conventional wisdom of how the limit order book operates and when we used time weighting these negative correlations vanished.

We then provided evidence of goodness of fit for dp^+ based on the Kullback-Leiber divegerence and probability-probability plots. Overall we had excellent fits for the price

change distributions, but there was still a discrepancy between the two distributions. This was likely caused by our assumption that the distribution is symmetric about the average imbalance I as we had evidence of an asymmetry from our statistical tests in chapter 2. Our fits were still very good, but this would be one avenue of improvement in future work.

Since we found many stocks had α values close to zero we also had our calibration with free weights $\vec{w}$ to see if we could find which volumes in the book these stocks had the most association with their price movements. We chose a particular constraint on $\vec{w}$ to be consistent with our exponential weights. The results of the free weights then included the results of the exponential decaying weights, but we also found important volumes beyond the touch that were not included by the original results. We repeated our two statistical tests again using the free weights and found that all stocks either benefited from applying weights deeper in the book or were no worse off. This is further statistical evidence that information about the volumes beyond the touch in the limit order book give us a better association with price dynamics.

In the next chapter we can finally implement our spoofing model from chapter 3 using parameters we generate from our calibration algorithm.

Chapter 5

Spoofing Detection

5.1 Introduction

Now that we have a methodology for calibrating our model we can solve the optimization problem we originally presented in chapter 3. That is, we want to buy H shares and we need to determine if we should immediately place a market order at time t for the H shares or should we delay our market order to $t + \Delta t$. Alternatively, as the spoofer, we could spoof the book at time t and delay our market order to time $t + \Delta t$, cancel our spoofing orders, and lower our cost of purchasing the H shares.

We derived expressions for the expected cost associated with each of these three decisions. We just need to compare which option saves us the most money for purchasing H shares, bearing in mind that the option to spoof the book also involves an optimization problem for picking where to place our spoofing limit orders. We compare two methods of determining the optimal strategy – comparing expected costs, and a hybrid using expected costs and the Sharpe ratio.

In this chapter we determine the optimal decision at each time throughout the day for multiple stocks to see if there are periods throughout the day where the limit order book is susceptible to spoofing, and if so, how much more profitable is spoofing the book? This will depend on the number of shares H we wish to purchase as well as the number of shares $\tilde{V}$ that we spoof with. We show that incorporating risk into the spoofer's decision making process yields better decision clusters which allow us to explore the dependency on the boundary between spoofing and placing market orders based on H and $\tilde{V}$.

We also analyze the optimal spoofing strategy with BMO stock using four example limit order books taken from the data. We show that the spoofer's strategy not only depends on H and $\tilde{V}$, but the initial volume imbalance and the predicted shape of the book when the spoofer needs to make their decision. We argue that properly modelling a

spoofer's prediction of the shape of the order book in the next time period is important to determining how they will act as to better combat them.

5.2 Determining the Optimal Strategy

From chapter 3 we wrote down equations for the cost associated with our three strategies - market order, delayed market order, and spoofing. We then choose the strategy which yields the smallest expected cost. We also round each result to the nearest penny when making our strategy decision. Spoofing may yield the lowest cost between the three strategies, but if it is not lower by at least a penny then there is no practical reason to take the risk in spoofing. We take this more conservative approach to decision making because spoofing may always yield a numerically lower cost than delaying a market order (due to manipulating the imbalance), but if that cost is only lower by some negligible amount then we argue the spoofer cannot justify the risk associated with manipulating the book.

Since one could, in theory, always gain an advantage by spoofing the book with an arbitrarily large number of shares we have to limit our investigation to a finite number. Also, limit orders of 100 shares are the smallest possible order one can place on the book[1] and orders are in increments of 100 shares. This significantly reduces the size of our solution space and we can even solve the optimization problem by brute force in some cases. Given parameters dp^+, $\vec{w}$, and K, the optimal spoofing strategy is obtained by solving the optimization problem

$$\min_{\tilde{v}_t} \sum_{x_t} C_{\mathrm{S}}(\vec{v}_t, H, p^+, x_t, \tilde{v}_t)\varphi(x_t; I(\vec{v}_t + \tilde{v}_t; \vec{w}, K)) \tag{5.2.1}$$

Subject to

$$\sum_{i=-K}^{0} \tilde{v}_i = 0, \quad \sum_{i=1}^{K} \tilde{v}_i \leq \tilde{V}, \quad \tilde{v}_i \geq 0 \text{ and } \tilde{v}_i \in 100\mathbb{N} \ \ \forall i \in [-K, K] \tag{5.2.2}$$

The first condition in equation 5.2.2 is to further reduce the solution space, but by construction we would never spoof the bid side of the book because that would push the volume imbalance against us. The second condition limits the total size of our spoofing orders to $\tilde{V}$. The final condition forces all spoofing orders to be 0 or multiples of 100 shares. The solution to equation 5.2.1 subject to 5.2.2 then gives the optimal spoofing strategy. We, as the spoofer, decide how many shares $\tilde{V}$ we are willing to spoof with.

[1] Orders less than 100 shares go to a separate book called Oddlot. This book has lower liquidity.

We cannot assume that all orders would have happened exactly the same over Δt if we place a spoofing order at the start of the time period, so we cannot calculate the average imbalance in the same way as we did for the calibration. We can only use the information we have at time t. We have to make our optimal spoofing decision based on the assumption that the imbalance we create from our spoofing order remains constant over Δt. There is no other option in the absence of data on how the market would react to specific limit order sizes and placement to model the market dynamics over the time interval based on our choices at time t.[2] This is where a multi-period model would come into play where we could update our strategy at multiple times over the day – which would be a more realistic spoofing strategy. This work is, however, a first step towards a more complicated and complete model to capture the behaviour of a market manipulator.

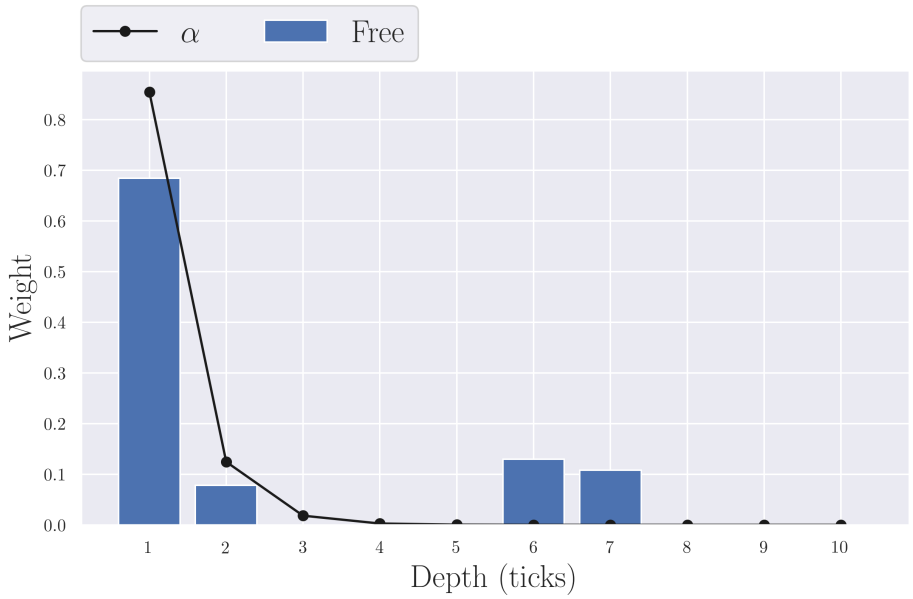


Figure 5.1: Calibrated weights $\vec{w}$ for BMO stock on April 17, 2017 over the entire trading day. $\Delta t = 5$ seconds and $K = 10$.

As an example, Figures 5.1 and 5.2 show the weights $\vec{w}$ and the optimal limit order placement for a specific limit order book configuration for BMO stock on April 17, 2017. Optimal limit order placement is decided using the free imbalance weights. We spoof with up to 500 shares and wish to purchase 1000 shares. From the weights in Figure 5.1 we can easily tell that the volume imbalance heavily favours the best ask to decrease over $\Delta t = 5$ seconds. For purchasing 1000 shares with this limit order book the expected

[2]One would need considerable money and resources to test this type of market sensitivity to limit order placement.

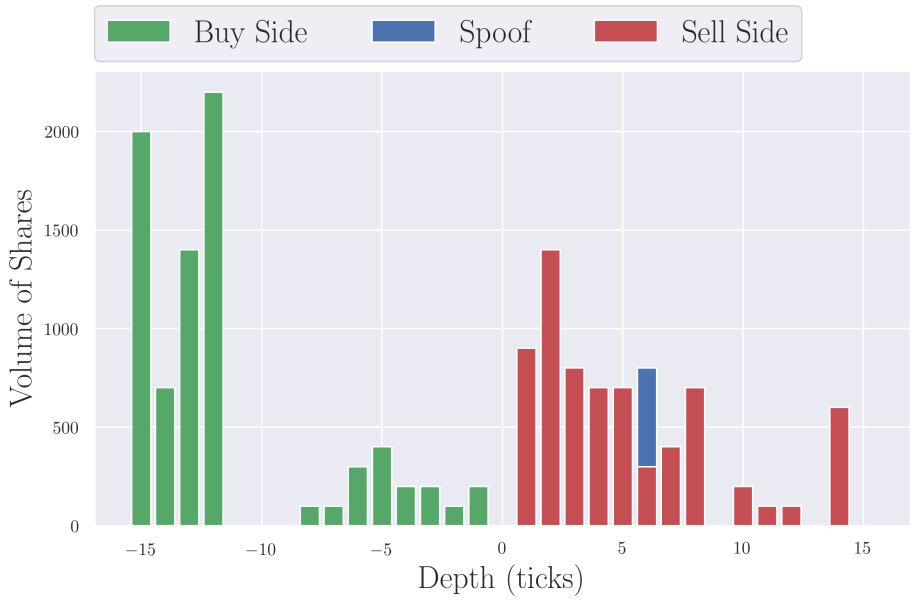


Figure 5.2: Optimal limit order placement for spoofing with example book for BMO stock taken on April 17, 2017. This is a stylized figure where we only show the first 15 prices in the bid and ask sides of the book and ignore the spread. $\tilde{V} = 500$ and we use the free weights for calculating the imbalance.

cost savings are \$2.50 if we spoof and \$2.36 if we delay our market order, implying an improved cost of \$0.14 with spoofing.

We now need to compare this results to our other two available choices – delayed and immediate market orders. Instead of using the expected cost of all three strategies, we can subtract the cost of an immediate market order to see the expected net cost savings. We denote these as μ_{S} and μ_{DMO} for the relative savings of spoofing and a delayed market order, respectively. That is,

$$\begin{aligned} \mu_{\text{S}} &= E\left[\, C_{\text{S}}(\vec{v}_t, H, p^+, x_t, \tilde{v}_t) \,|\mathcal{F}_t, I(\vec{v}_t + \tilde{v}_t)\right] - C_{\text{MO}}(\vec{v}_t, H, p^+) \\ \mu_{\text{DMO}} &= E\left[\, C_{\text{DMO}}(\vec{v}_t, H, p^+, x_t) \,|\mathcal{F}_t, I(\vec{v}_t)\right] - C_{\text{MO}}(\vec{v}_t, H, p^+) \end{aligned} \tag{5.2.3}$$

We can bring $C_{\text{MO}}(\vec{v}_t, H, p^+)$ outside the expected value since it is $\mathcal{F}_t$ measurable. This way we can also compare the savings seen across different limit order book states since the costs are recorded relative to the immediate market order option for each time period.

In Figure 5.3 we show the decision process for picking the optimal strategy based purely on the expected net cost of the three options. We place an immediate market order if we cannot save money by spoofing or delaying our market order. We spoof or delay our market order if we can expect to save money and we choose the option

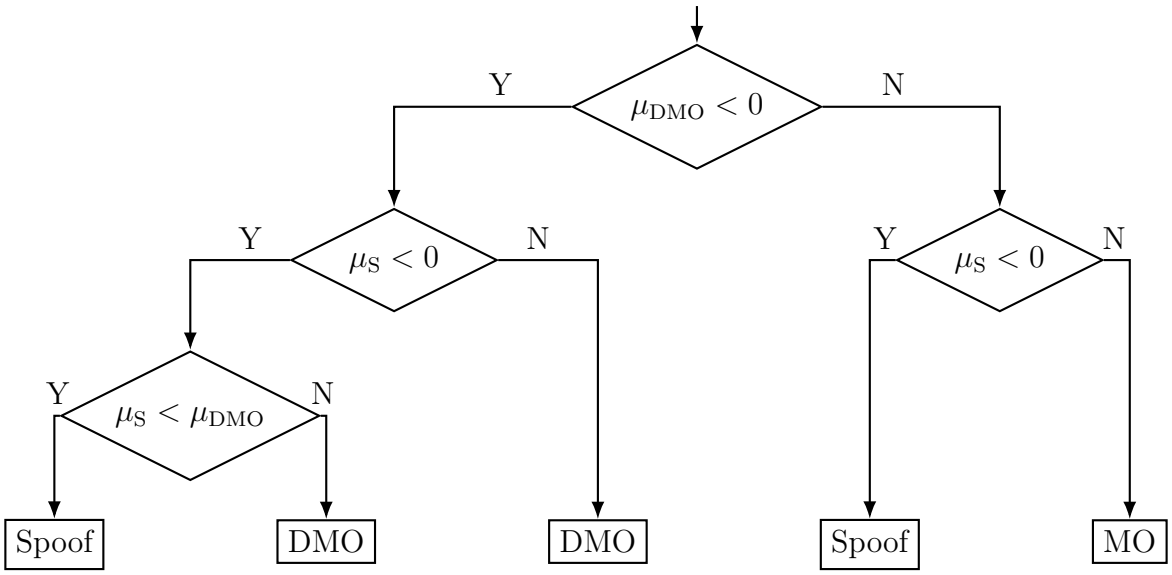


Figure 5.3: Decision tree using only expected costs to determine optimal strategy. The three options: Spoof, DMO, and MO, refer to spoofing, delayed market order, and immediate market order, respectively.

which yields the greatest savings. However, using only the expected net savings in the decision making process also ignores the risk associated with spoofing or delaying our market order. We can incorporate the risk through the Sharpe ratio [66], S, defined for an investment as

$$S = \frac{\mu - r}{\sigma} \tag{5.2.4}$$

where μ is the expected rate of return, r is the risk-free rate, and σ is the standard deviation or volatility. One can think of the Sharpe ratio as the net return per unit of increased risk. To unpack the Sharpe ratio we first subtract the risk-free rate from the expected return of our investment since if $\mu < r$ we would have been better off investing in the risk-free asset (usually taken to be a government bond). We then divide the excess return by the volatility as a stand-in for the risk associated with the investment. This is done to compare the ratio of excess returns to risk for a collection of different investments (stocks, for example) that could make up a portfolio. Ideally, one wants to maximize their return per unit of risk and would invest in assets with the highest Sharpe ratio.

Since we are looking at such small time intervals we can take $r \approx 0$ in our case[3]. The volatility σ for the cost of spoofing or delaying a market order is then given by

[3]A large risk free rate of 5% per year would yield $1.05^{\frac{1}{250\times 6.5\times 60\times 60}} - 1 = 8.34 \times 10^{-7}\%$ per second. Assuming approximately 250 trading days with 6.5 hours of trading per day.

$$\begin{aligned}\sigma_{\mathrm{S}}^2 &= \mathrm{Var}\left[\, C_{\mathrm{S}}(\vec{v}_t, H, p^+, x_t, \tilde{v}_t) - C_{\mathrm{MO}}(\vec{v}_t, H, p^+) \,|\mathcal{F}_t, I(\vec{v}_t + \tilde{v}_t)\right] \\ \sigma_{\mathrm{DMO}}^2 &= \mathrm{Var}\left[\, C_{\mathrm{DMO}}(\vec{v}_t, H, p^+, x_t) - C_{\mathrm{MO}}(\vec{v}_t, H, p^+) \,|\mathcal{F}_t, I(\vec{v}_t)\right]\end{aligned} \tag{5.2.5}$$

and the simple returns from spoofing and the delayed market order relative to the immediate market order are given by

$$\begin{aligned}\text{Spoofing Return} &= E\left[\, \frac{C_{\mathrm{S}}(\vec{v}_t, H, p^+, x_t, \tilde{v}_t) - C_{\mathrm{MO}}(\vec{v}_t, H, p^+)}{C_{\mathrm{MO}}(\vec{v}_t, H, p^+)} \,\middle|\, \mathcal{F}_t, I(\vec{v}_t + \tilde{v}_t)\right] \\ &= \frac{\mu_{\mathrm{S}}}{C_{\mathrm{MO}}(\vec{v}_t, H, p^+)} \\ \text{DMO Return} &= E\left[\, \frac{C_{\mathrm{DMO}}(\vec{v}_t, H, p^+, x_t) - C_{\mathrm{MO}}(\vec{v}_t, H, p^+)}{C_{\mathrm{MO}}(\vec{v}_t, H, p^+)} \,\middle|\, \mathcal{F}_t, I(\vec{v}_t + \tilde{v}_t)\right] \\ &= \frac{\mu_{\mathrm{DMO}}}{C_{\mathrm{MO}}(\vec{v}_t, H, p^+)}\end{aligned} \tag{5.2.6}$$

Similarly, the volatility of each strategy relative to the immediate market order is given by

$$\begin{aligned}\text{Spoofing Volatility} &= \mathrm{Var}\left[\, \frac{C_{\mathrm{S}}(\vec{v}_t, H, p^+, x_t, \tilde{v}_t) - C_{\mathrm{MO}}(\vec{v}_t, H, p^+)}{C_{\mathrm{MO}}(\vec{v}_t, H, p^+)} \,\middle|\, \mathcal{F}_t, I(\vec{v}_t + \tilde{v}_t)\right] \\ &= \frac{\sigma_{\mathrm{S}}^2}{C_{\mathrm{MO}}(\vec{v}_t, H, p^+)^2} \\ \text{DMO Volatility} &= \mathrm{Var}\left[\, \frac{C_{\mathrm{DMO}}(\vec{v}_t, H, p^+, x_t) - C_{\mathrm{MO}}(\vec{v}_t, H, p^+)}{C_{\mathrm{MO}}(\vec{v}_t, H, p^+)} \,\middle|\, \mathcal{F}_t, I(\vec{v}_t + \tilde{v}_t)\right] \\ &= \frac{\sigma_{\mathrm{DMO}}^2}{C_{\mathrm{MO}}(\vec{v}_t, H, p^+)^2}\end{aligned} \tag{5.2.7}$$

then finally putting equations 5.2.6 and 5.2.7 together we get the Sharpe ratio as

$$\begin{aligned}S_{\mathrm{S}} &= \frac{\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}} \\ S_{\mathrm{DMO}} &= \frac{\mu_{\mathrm{DMO}}}{\sigma_{\mathrm{DMO}}}\end{aligned} \tag{5.2.8}$$

The advantage of using the Sharpe ratio is that spoofing may yield a lower expected cost than delaying the market order, but the improved return on the strategy may not be worth the increased risk.

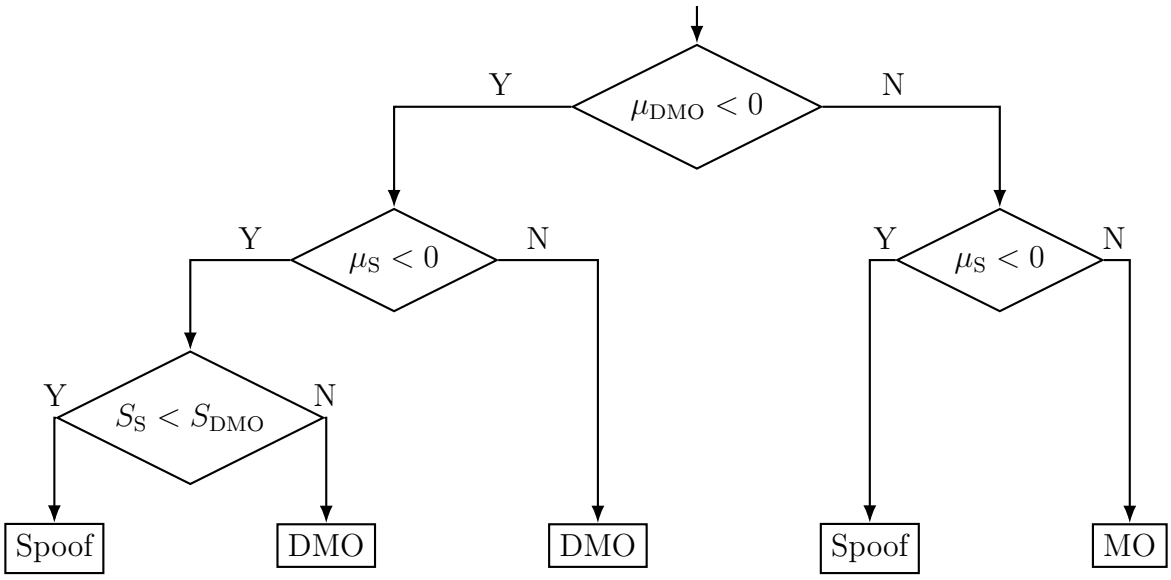


Figure 5.4: Decision tree using expected costs and Sharpe ratio to determine optimal strategy. The three options: Spoof, DMO, and MO, refer to spoofing, delayed market order, and immediate market order, respectively.

In Figure 5.4 we show a slightly different decision making process from Figure 5.3. In this case if we can expect to save money from both spoofing and delaying our market order we choose the option which yields the highest savings per unit of risk – represented by the Sharpe ratio. Spoofing will always be riskier than delaying our market order because our spoofing limit orders have a chance to be executed against us. Taking into account the Sharpe ratio gives us a way to quantify if that increased risk is worth the associated expected savings.

With our optimal strategy criteria established we can investigate how our decisions change over time for a given stock and their possible dependency on the number of shares H we wish to buy and the number of shares $\tilde{V}$ we are willing to spoof with.

5.3 Spoofing Payoff and Optimal Strategy

5.3.1 Spoofing Criteria

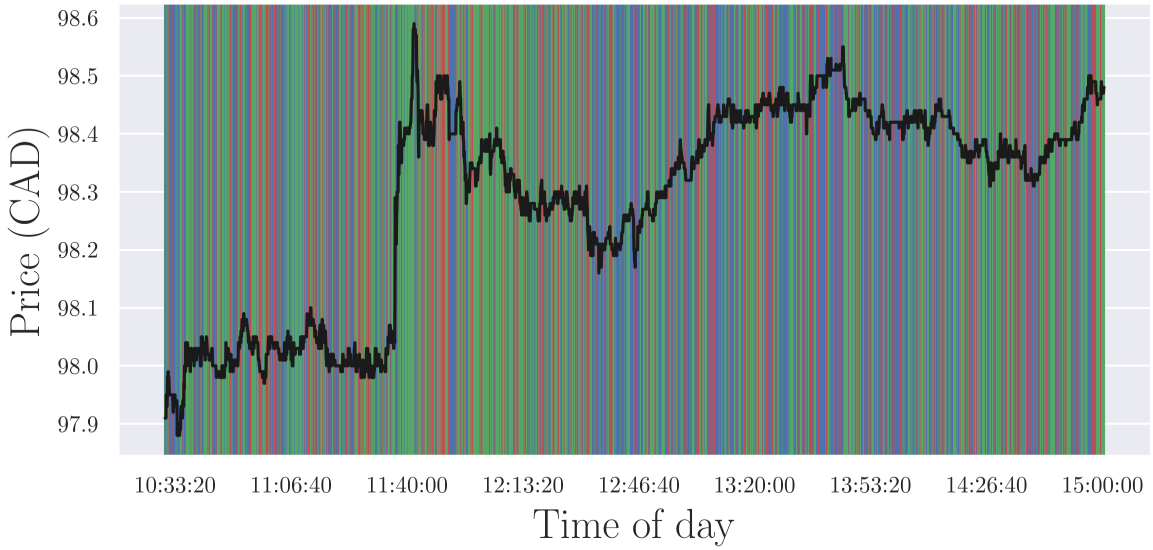
To investigate the difference between our two decision criteria when both $\mu_{\text{S}} < 0$ and $\mu_{\text{DMO}} < 0$ we use the calibration presented in Figure 5.1 for BMO stock on April 17, 2017. The optimal decision for both criteria is presented in Figure 5.5 for the time period 10:30 – 11:30 AM over 5 second intervals. In this example we also use the free weights when calculating the imbalance and we aim to buy $H = 200$ shares and spoof with up to

$\tilde{V} = 500$ shares. Using the Sharpe ratio over just the net-expected cost gives more regions where we would delay our market order instead of spoofing the book. To understand what is happening here we first look at the impact spoofing has on the imbalance at the start of each 5 second time period. The optimal decisions (as color coded in the figures) are being made after rounding μ_{S} and μ_{DMO} to the nearest penny.

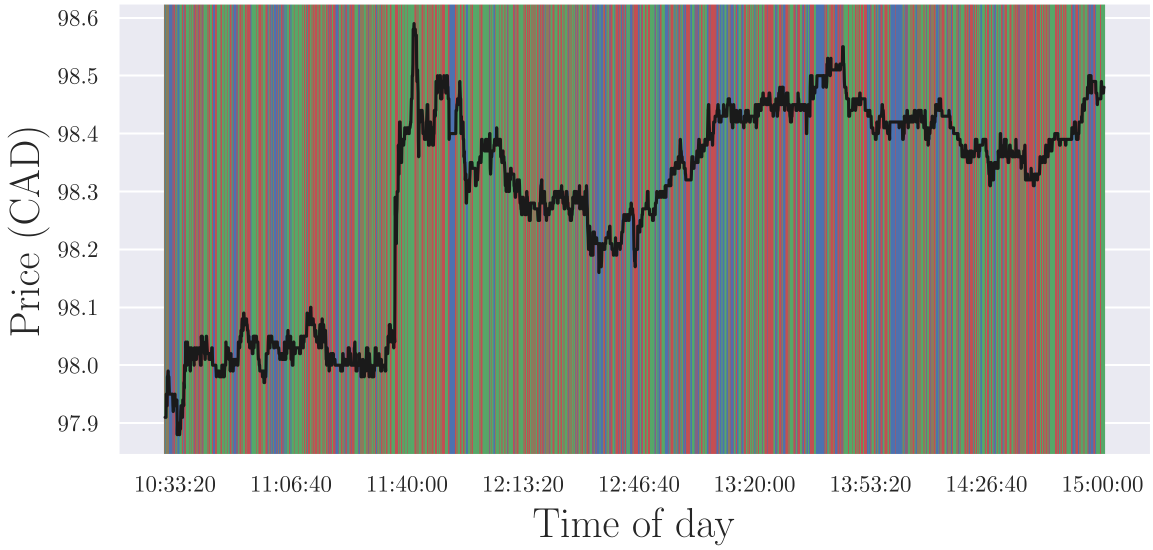
Figure 5.6 shows the pre- and post-spoofing imbalance for each 5 second interval presented in Figure 5.5. We see a clear divide between immediately placing a market order and either spoofing or delaying the market order. We see that we immediately place the market order when the pre-spoofing imbalance is initially positive and cannot be made sufficiently negative with our spoofing order. We then spoof when we can drive the imbalance negative so it is more likely for the best ask to decrease over the next 5 seconds. When using the net-expected cost criteria we delay our market order only if the imbalance is already so negative that we cannot make it better for us with a spoofing order, but when we use the Sharpe ratio there is a new region of points where we delay over spoofing. These points lie in the region where the imbalance is already so negative we can hardly improve it, but we can still get a modest improvement by still spoofing. Why would we delay the market order over spoofing then? We have to see what is happening between the net savings, μ_{S} and μ_{DMO}, and the Sharpe ratios, S_{S} and S_{DMO}.

In Figure 5.7 we compare the two criteria inequalities, $\mu_{\mathrm{S}} - \mu_{\mathrm{DMO}}$ and $S_{\mathrm{S}} - S_{\mathrm{DMO}}$. We see there is a region where using only the expected net savings determines that we should spoof instead of delaying our market order even though the net savings on our spoofing strategy is smaller per unit of volatility than the delayed market order strategy. That is, the net savings with spoofing may be better, but that net savings is not worth the associated increased risk which is captured by the Sharpe ratio.

Then in Figure 5.8 we compare the net savings, $\mu_{\mathrm{S}} - \mu_{\mathrm{DMO}}$, to the spoofing strategy Sharpe ratio S_{S}. We see that when the spoofing strategy has a positive cost we place an immediate market order as delaying the market order or spoofing will more than likely increase our cost of buying H shares. We then spoof the book if S_{S} is negative with a net savings sufficiently greater than the net savings on delaying the market order. We also see in the limit that the $\mu_{\mathrm{S}} - \mu_{\mathrm{DMO}} \to 0$ we delay our market orders instead of spoofing as the net savings is not there to justify spoofing. The Sharpe ratio then captures the fact that the net savings of spoofing over delaying the market order is not worth the associated increased risk. Going back to Figure 5.6 we have that there are times where we can spoof to push the imbalance further negative than it already is, but the increased risk of exposing ourselves to our limit order being executed can offset the small improvement in the expected net savings.



(a) Net-Expected Cost Criteria



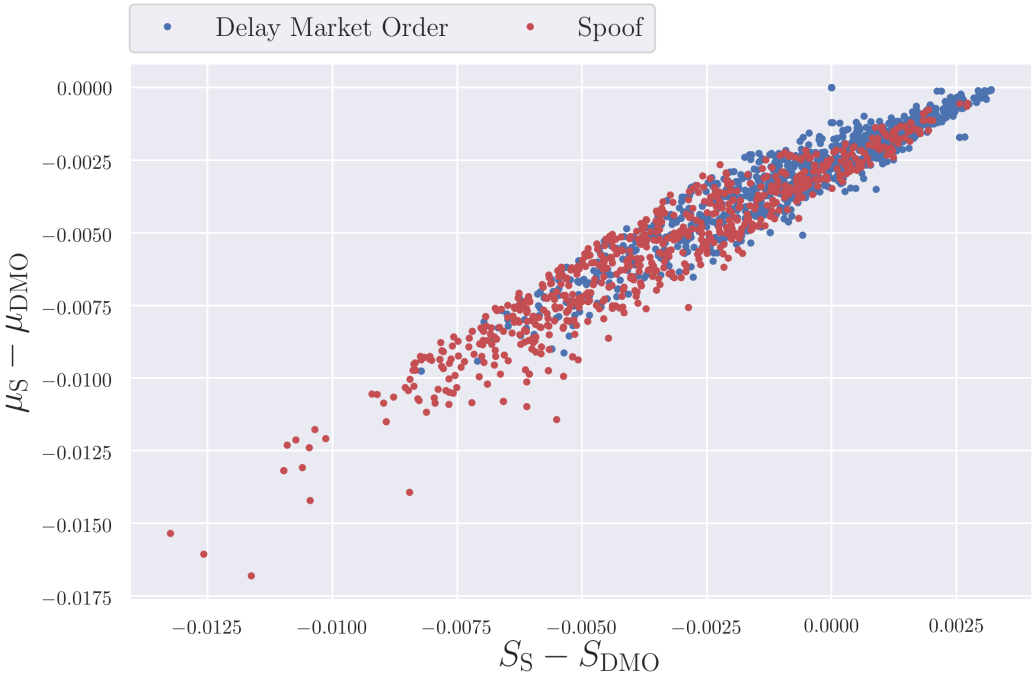
(b) Sharpe Ratio Criteria

Figure 5.5: Optimal strategy over 5 second intervals from 10:30 AM – 3:00 PM on April 17, 2017 for BMO stock using both selection criteria. The black line is the best ask price time series. You place an immediate market order in the green regions, a delayed market order in the blue regions, and spoof in the red regions. $H = 100$ and $\tilde{V} = 200$.

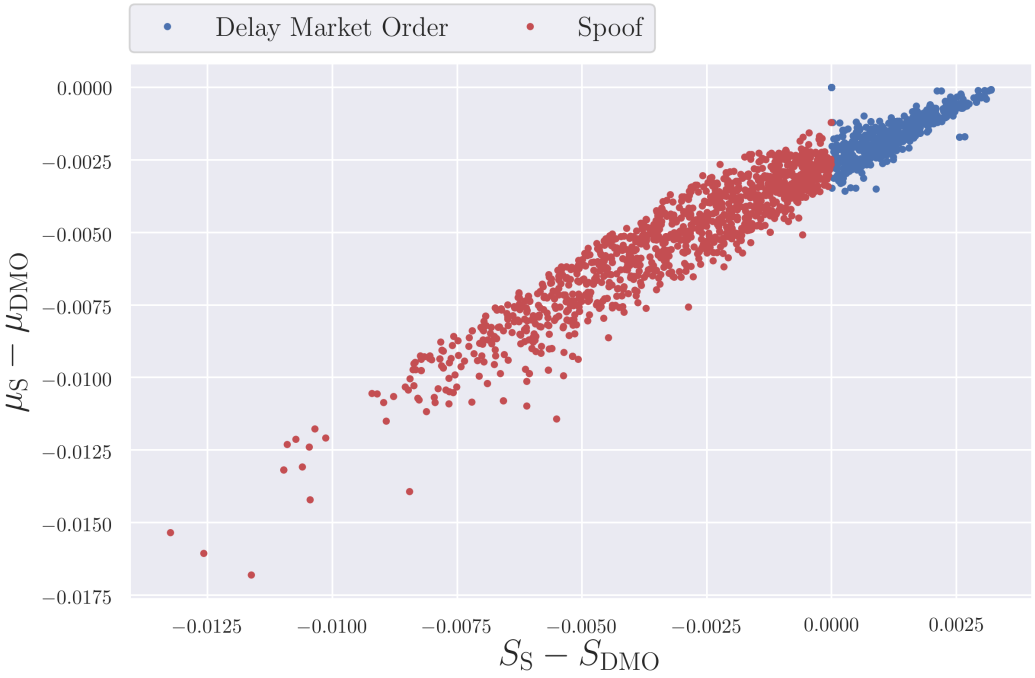
(a) Net-Expected Cost Criteria

(b) Sharpe Ratio Criteria

Figure 5.6: The pre- and post-spoofing imbalance for each 5 second time period in Figure 5.5 labeled by optimal strategy using both selection criteria. The post-spoofing imbalance is determined by the imbalance after the optimal spoofing order – even if spoofing was not the optimal decision.

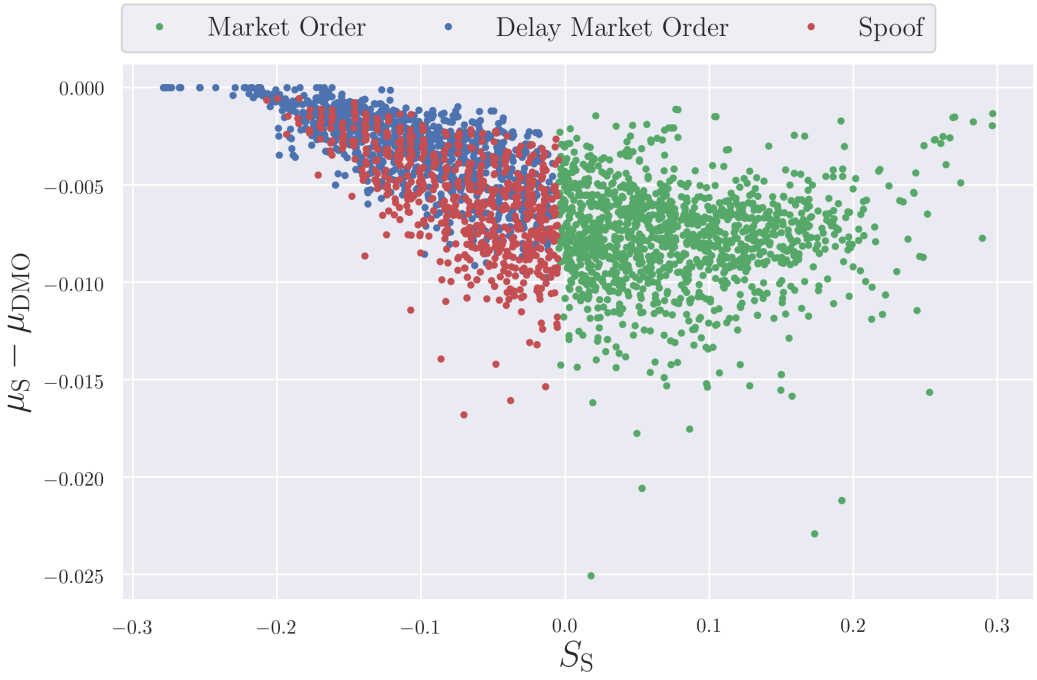


(a) Net-Expected Cost Criteria

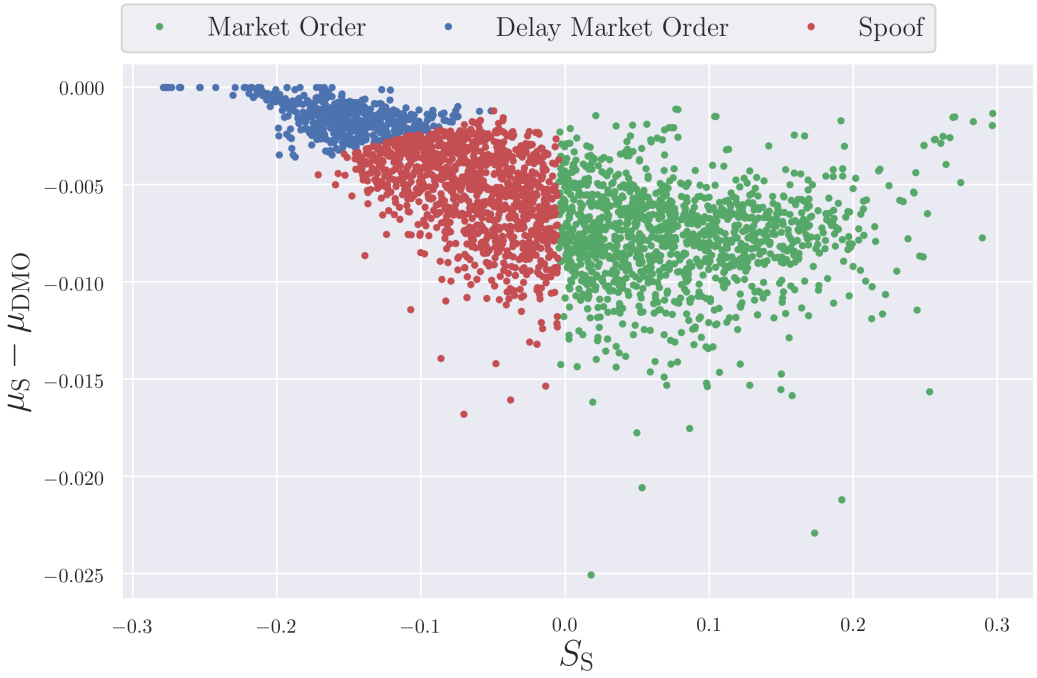


(b) Sharpe Ratio Criteria

Figure 5.7: Comparing net spoofing savings and Sharpe ratio over a delayed market order for each 5 second time period in Figure 5.5 labeled by optimal strategy using both selection criteria.



(a) Net-Expected Cost Criteria



(b) Sharpe Ratio Criteria

Figure 5.8: Comparing net spoofing savings over a delayed market order to spoofing Sharpe ratio for each 5 second time period in Figure 5.5 labeled by optimal strategy using both selection criteria.

Figures 5.9, 5.10, and 5.11 compare the decision clusters with increasing H and $\tilde{V}$. Increasing $\tilde{V}$ gives the spoofer an increased ability to manipulate prices, so the spoofing decision becomes a profitable investment over the delayed market order and optimal decisions switch from the delayed market order to spoofing. In both cases we see that using the Sharpe ratio makes better clusters compared to using expected net savings as it provides a distinct boundary between the two decisions.

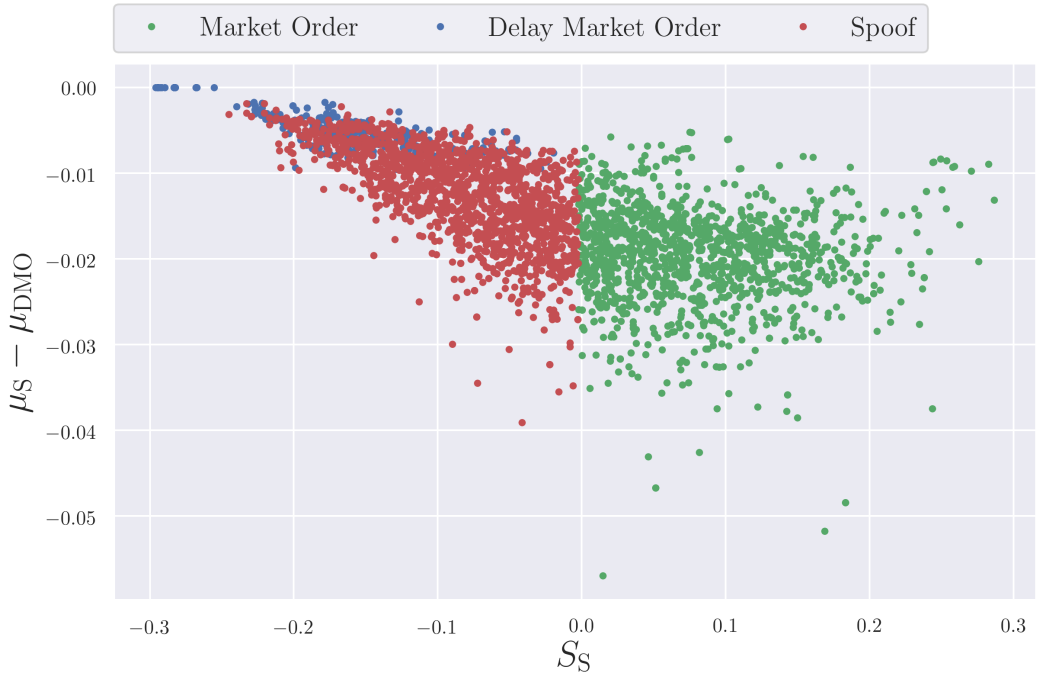
Ultimately, both criteria yield similar results, but using the Sharpe ratio could narrow our search for potential price manipulators to time periods where not only was it profitable to spoof the book, but the potential profit earned was worth the risk of spoofing as well. If we are to catch people attempting to manipulate the limit order book of every stock each day we will need ways to save time and narrow down where to look for when the books were most vulnerable. This was an example of a single stock over one hour, but the purpose was to illustrate how our model could be used in practice.

5.3.2 Decision Boundary, H, and $\tilde{V}$

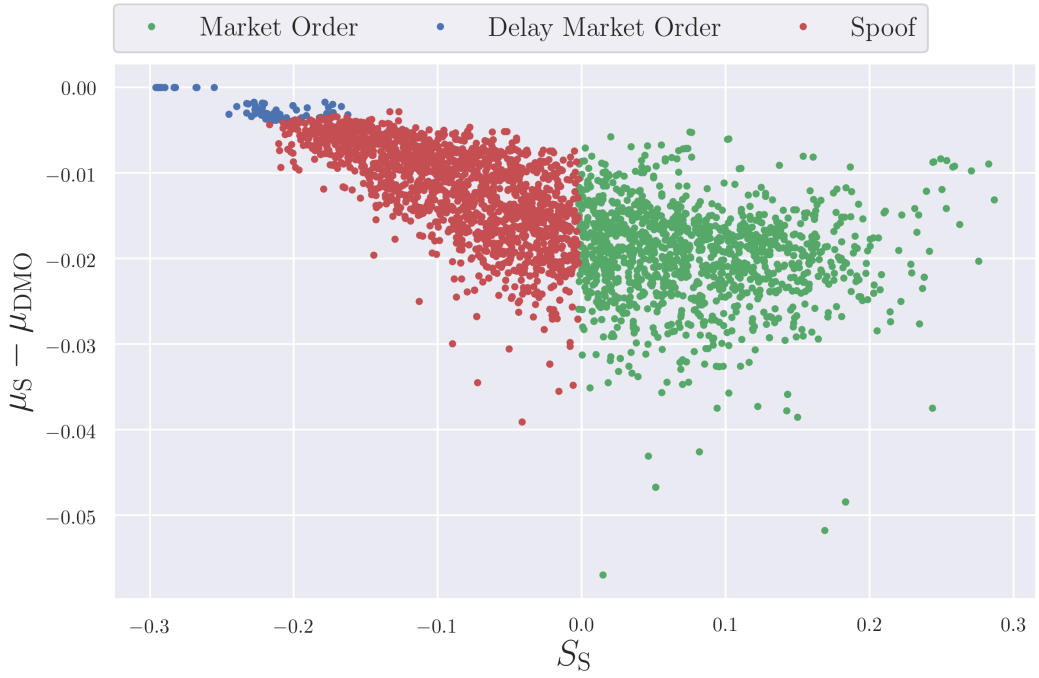
In Figure 5.6 we can see there is a clear decision boundary between the immediate market order choice and either spoofing or delaying the market order. This decision boundary is dependent on the number of shares H we wish to purchase and the number of shares $\tilde{V}$ we are willing to spoof with. We want to know if we can extract some general decision rules based on the initial imbalance of the book, H, and $\tilde{V}$ which a spoofer may use or the exchange may use to narrow their search down for possible price manipulators. These rules will of course vary between stocks and the time of day, but we want to see if some simple 'rules of thumb' can be gained from our model.

We present the results using the free weights for the imbalance as this allows for limit order placement deeper in the book with a larger association between the imbalance and changes in the best ask price. This gives a 'worse case scenario' for the book's vulnerability to spoofing. We also found our model would predict almost no advantage to spoofing when using exponential weights – the risk of spoofing near the touch completely removed the advantages of moving the best ask price in a favourable direction. The free weights allow for the same limit order placement near the touch as the exponential weights, so even if that was the best spoofing order placement it would be captured with the free weights anyway.

We determine the decision boundary using a support vector machine (SVM) with a linear kernel. This will find the linear boundary between the classes, immediate market orders and the spoofing/delayed market orders, which separates both classes on either

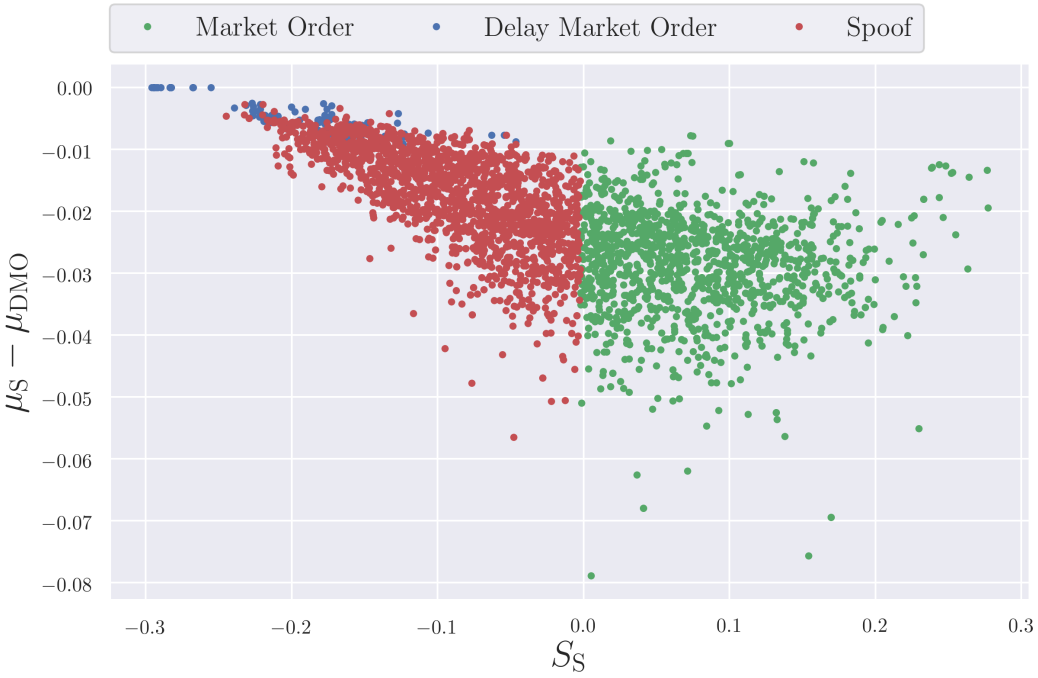


(a) Net-Expected Cost Criteria

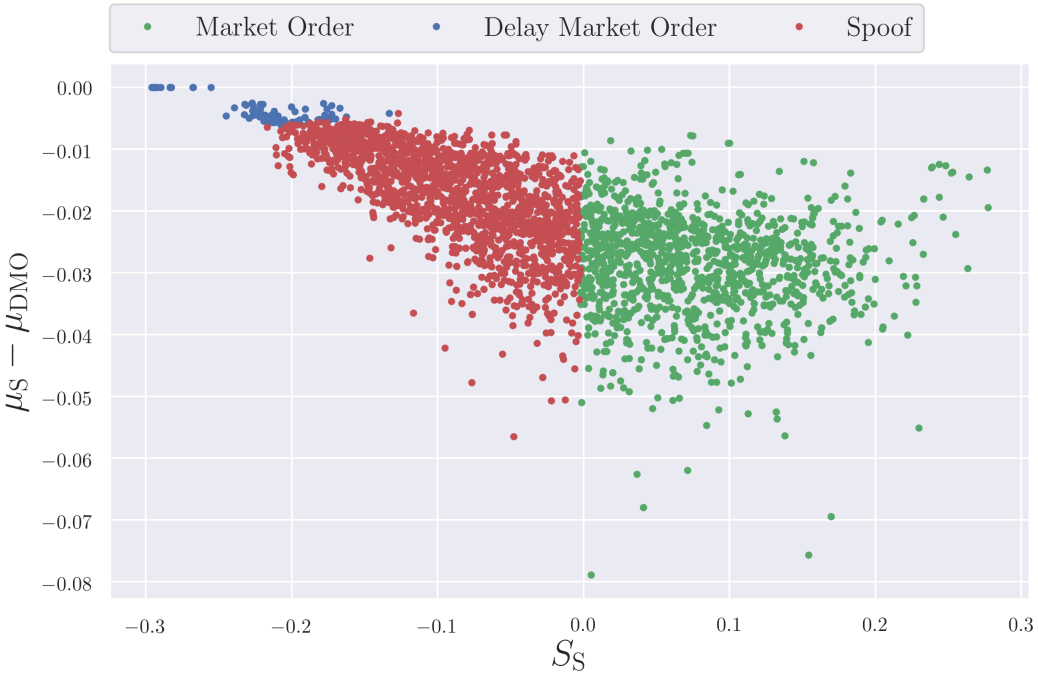


(b) Sharpe Ratio Criteria

Figure 5.9: Comparing net spoofing savings over a delayed market order to spoofing Sharpe ratio for each 5 second time period in Figure 5.5 labeled by optimal strategy using both selection criteria. $H = 200$ and $\tilde{V} = 200$ for this case.

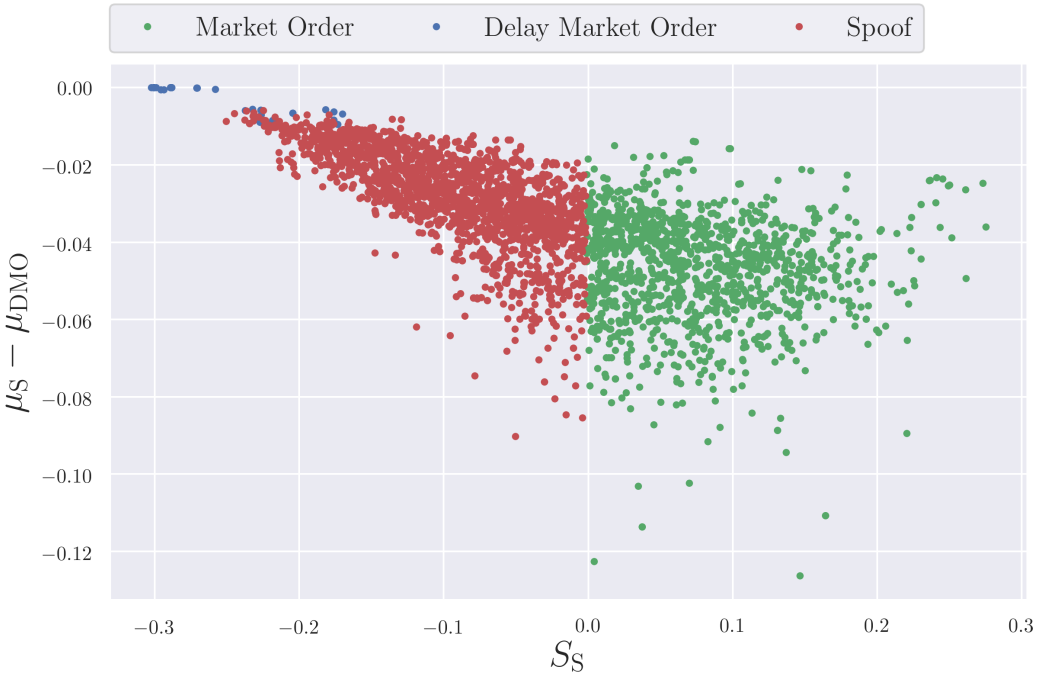


(a) Net-Expected Cost Criteria

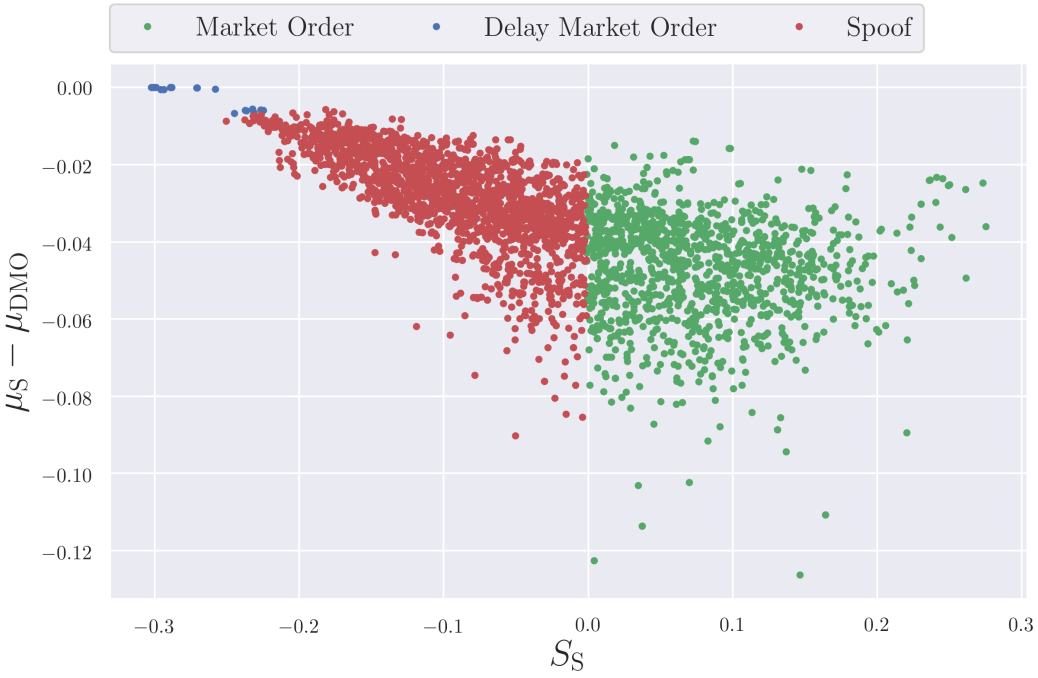


(b) Sharpe Ratio Criteria

Figure 5.10: Comparing net spoofing savings over a delayed market order to spoofing Sharpe ratio for each 5 second time period in Figure 5.5 labeled by optimal strategy using both selection criteria. $H = 200$ and $\tilde{V} = 300$ for this case.



(a) Net-Expected Cost Criteria



(b) Sharpe Ratio Criteria

Figure 5.11: Comparing net spoofing savings over a delayed market order to spoofing Sharpe ratio for each 5 second time period in Figure 5.5 labeled by optimal strategy using both selection criteria. $H = 300$ and $\tilde{V} = 300$ for this case.

side of the boundary and has the largest distance between the boundary and points in both class. We also perform a linear regression for the pre- and post-spoofing imbalance as seen in Figure 5.6 so that the intersection of our SVM boundary and the regression line will approximate the location of the boundary. The intersection point will be referred to as the 'midpoint' of the decision boundary. We can then see how this midpoint moves as we change H and $\tilde{V}$ for different stocks. An example of this process is presented in Figure 5.12 where the SVM boundary is the dashed black line and the regression line is in solid black.

Figure 5.12: Example decision boundary for AEM stock on April 17, 2017 using data from 10:30 AM – 3:00 PM over 5 second intervals. $H = 100$ and $\tilde{V} = 300$. Decision boundary midpoint determined by the intersection of the regression line (solid black) and the support vector (dashed black).

Figure 5.13 presents the results of the boundary midpoint's movement from changing H and $\tilde{V}$ for AEM stock on April 17, 2017 using data from 10:30 AM – 3:00 PM over 5 second intervals. We excluded the first and last hours as there are some 5 second intervals where there is not enough shares available within the first 15 ticks of the best ask price to fulfill a market order of say, 1000 shares, and definitely not enough if the spoofing orders are also executed. This would prevent us from calculating the cost of walking the book when buying that many shares.

The points, which increase to the right, are color coded by the number of shares $\tilde{V}$ with which we are willing to spoof. Within each color the points move up with increasing

H. There are a few things to note in the figure. First, the number of shares we are willing to spoof with for a given H, in general, increases the initial imbalance we are willing to spoof at. This is intuitive as the more shares we spoof with the more impact with can have on the imbalance to push prices in our favour. Second, the more shares we need to purchase increases the spoofer's willingness to manipulate prices even if they cannot move the imbalance much with their spoofing orders – likely because the book is in a state where they cannot push the imbalance negative without a massive spoofing order. This would reflect the spoofer's willingness to take risks for the smallest chance to lower the best ask because they intend to purchase so many shares that any price improvement they get is worth it.

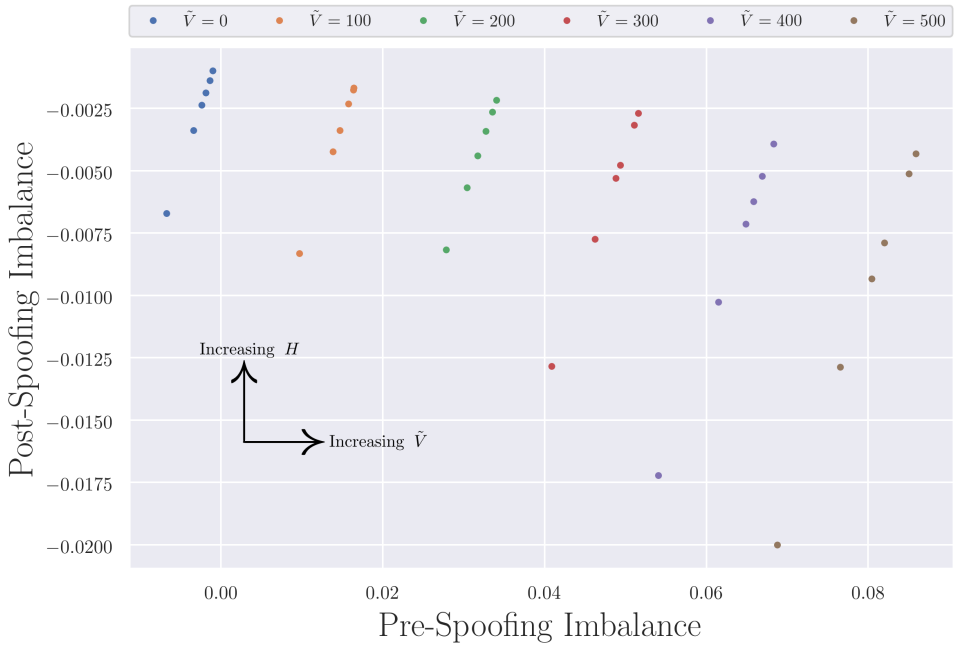


Figure 5.13: Midpoint of decision boundary for changing H and $\tilde{V}$ for AEM stock on April 17, 2017 using data from 10:30 AM – 3:00 PM over 5 second intervals. For a given $\tilde{V}$, points move up the graph with increasing H. $H \in [100, 200, 300, 400, 700, 1000]$.

We also see the same patterns and behaviour in Figures 5.14 and 5.15 for BMO and CNR stock on the same day over the same time period and also over 5 second intervals. We only present these three cases for brevity, but this pattern was consistent across all stocks we investigated. In short, increasing H causes the spoofer to manipulate prices even if they cannot influence the imbalance as much as they would like and increasing $\tilde{V}$ causes the spoofer to manipulate more often because they have more shares to impact the imbalance.

In addition to this, we can look at the dependency of the slope and intercept for the

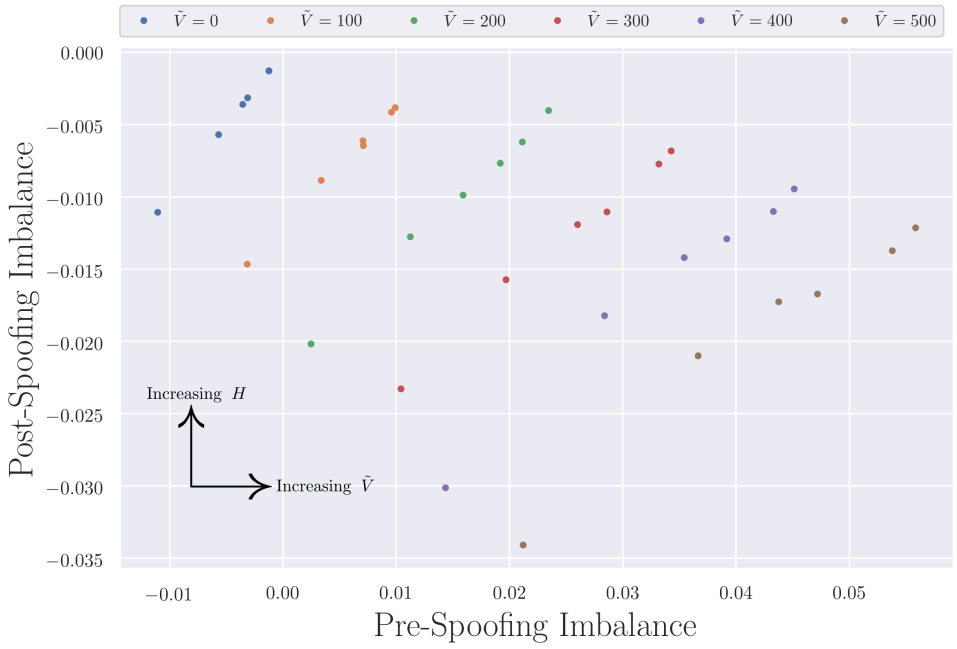


Figure 5.14: Midpoint of decision boundary for changing H and $\tilde{V}$ for BMO stock on April 17, 2017 using data from 10:30 AM – 3:00 PM over 5 second intervals. For a given $\tilde{V}$, points move up the graph with increasing H. $H \in [100, 200, 300, 400, 700, 1000]$.

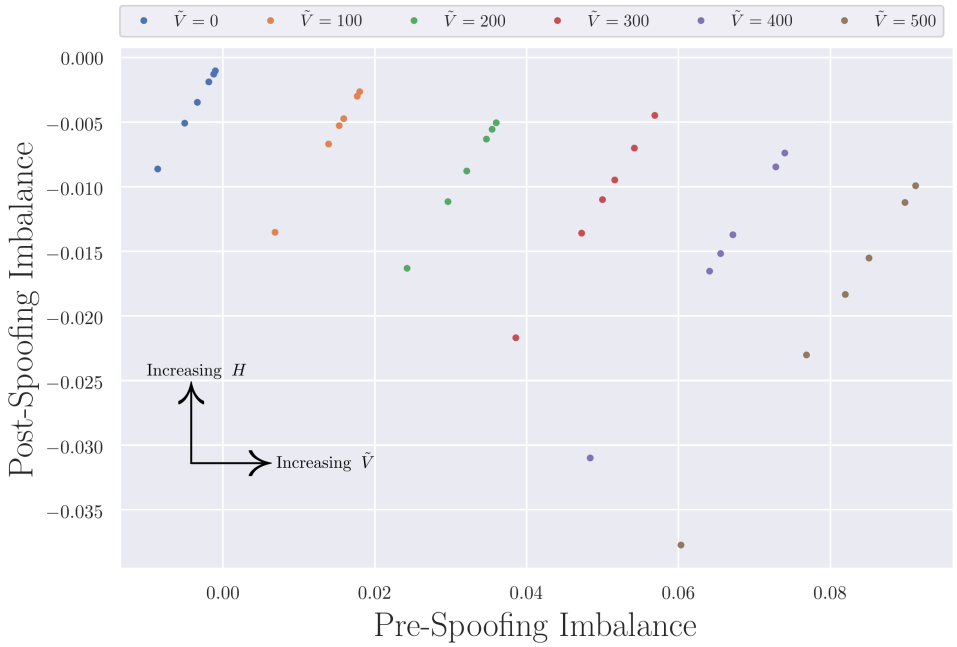


Figure 5.15: Midpoint of decision boundary for changing H and $\tilde{V}$ for CNR stock on April 17, 2017 using data from 10:30 AM – 3:00 PM over 5 second intervals. For a given $\tilde{V}$, points move up the graph with increasing H. $H \in [100, 200, 300, 400, 700, 1000]$.

regression line on H and $\tilde{V}$. In Figure 5.16 we present the exponential and free imbalance weights for AEM and BMO stock to keep in mind how the imbalance is being calculated for these two stocks. The key difference is that BMO has 2 spots beyond the touch which have a strong association to movements in the best ask price – compared to the one spot for AEM stock. AEM also has a significant weight deep in the book compared to the smaller weights for BMO.

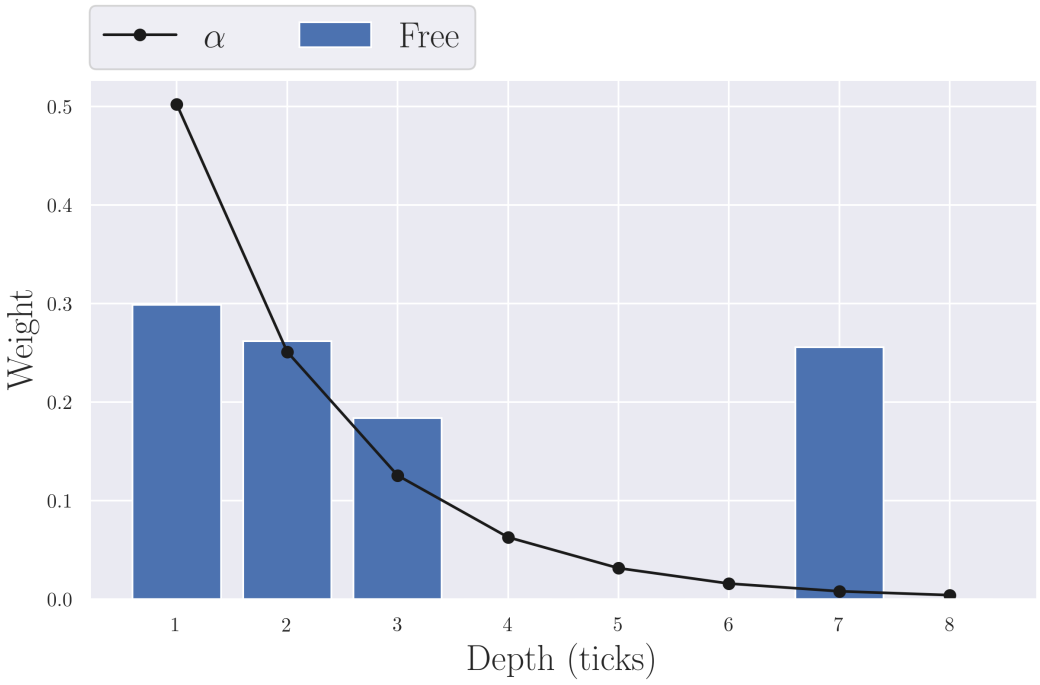
In Figures 5.17 and 5.18 we present the impact of H and $\tilde{V}$ on the slope and intercept of the regression line for AEM and BMO stock. Both AEM and BMO have decreasing slope and intercept with increasing $\tilde{V}$ because increasing $\tilde{V}$ allows for a larger impact on the post-spoofing imbalance. However, we see a difference in the dependency in the slope and intercept on H for BMO, but not for AEM. This is because AEM has only a single tick location deeper in the book where a spoofer can place orders to manipulate prices, but BMO has two. The only way to impact the imbalance through H is by the optimal spoofing strategy defined in equation 5.2.1, so the optimal strategy is changing depending on the number of shares a spoofer wishes to buy.

It is also clear why the optimal strategy is changing – the spoofing orders are being rearranged to give largest impact on the post-spoofing imbalance that is possible. This is interesting because this would, by Figure 5.16, mean that riskier orders are being placed at tick 6 instead of tick 7 to capture that extra weight when calculating the imbalance. Likewise, when H is small the imbalance is smaller than it could be for a given $\tilde{V}$ which implies the spoofing orders are placed at tick 7 to have a smaller chance of being executed. The more shares the spoofer wants to buy, the more risk they are willing to take to cut costs. This is an important insight as it provides evidence the spoofer's strategy will change depending on the number of shares they want to buy and this will need to be considered when employing detection technology to catch spoofers purchasing different amounts of H.

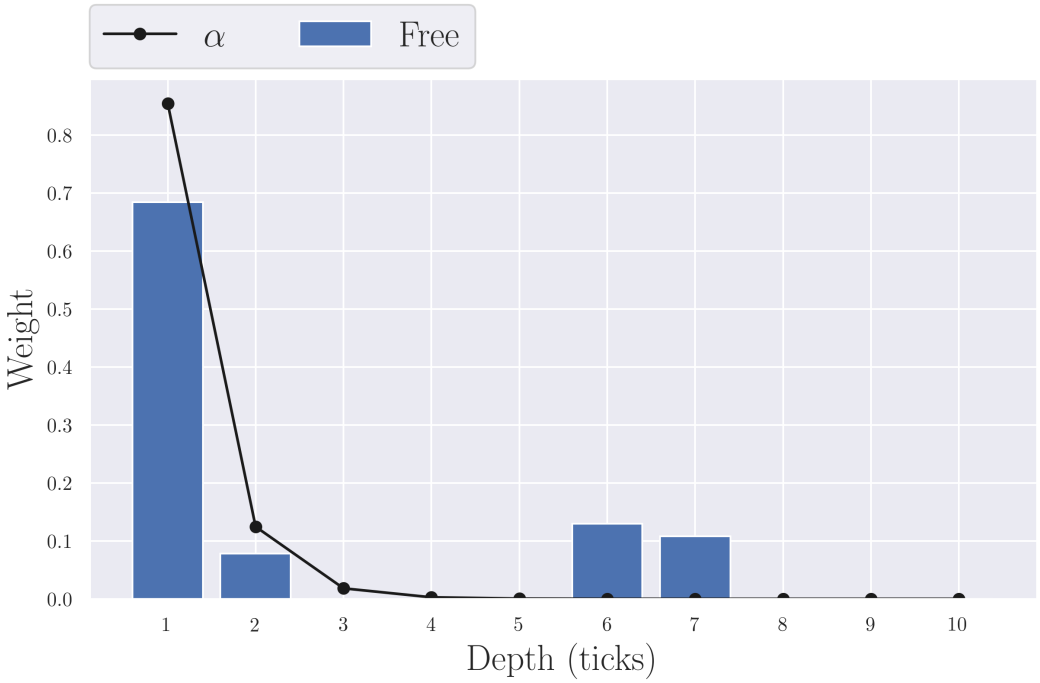
When there are multiple locations at which spoofing can impact the imbalance deeper in the book we see the optimal strategy changing with H. To understand why this is happening we can look deeper into how BMO's optimal spoofing strategy changes with H and $\tilde{V}$.

5.4 BMO Optimal Spoofing Strategy

We found in the previous section that when we have two available locations to spoof and a fixed number of shares to spoof with we are faced with allocating our 'spoofing resources' optimally. This did not occur with AEM as there was only one location deep

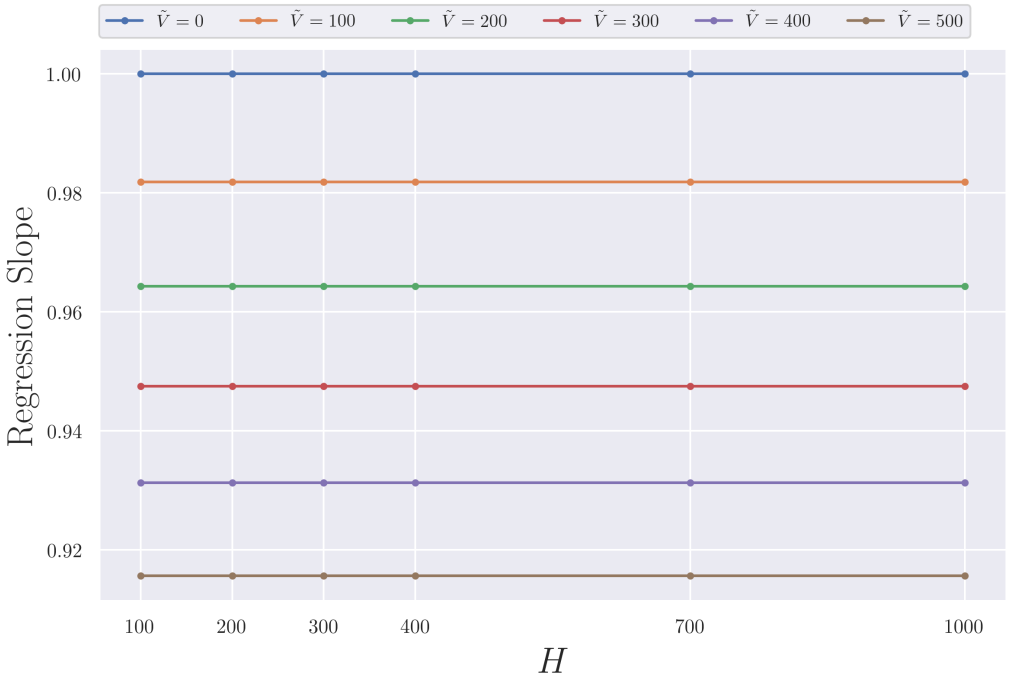


(a) AEM

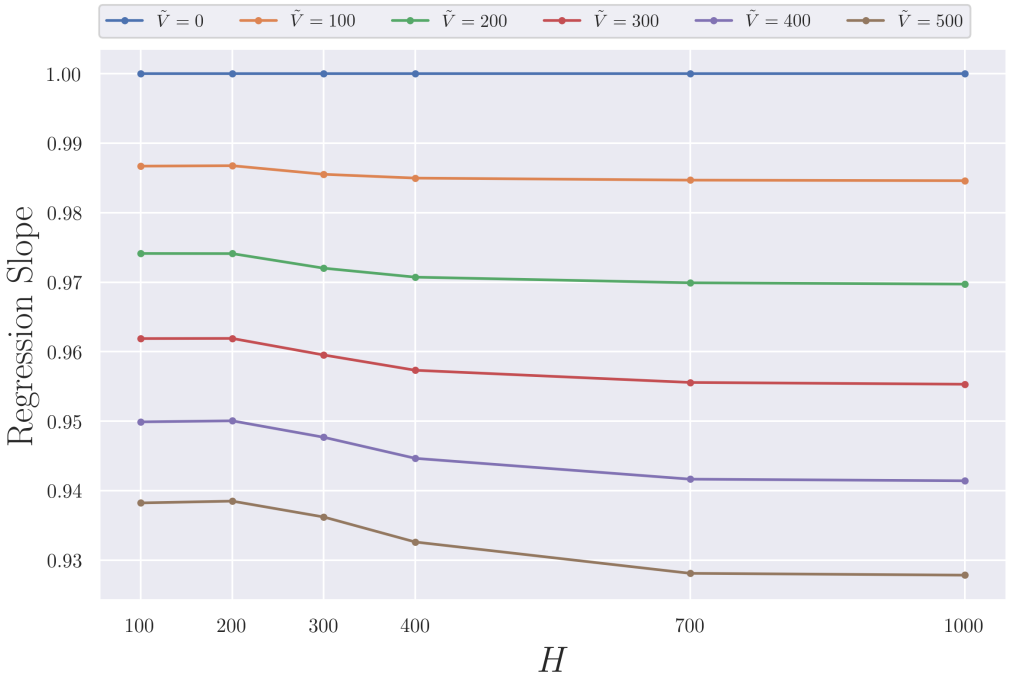


(b) BMO

Figure 5.16: Comparing exponential and free imbalance weights for BMO and AEM on April 17, 2017 using the entire trading day over 5 second intervals.

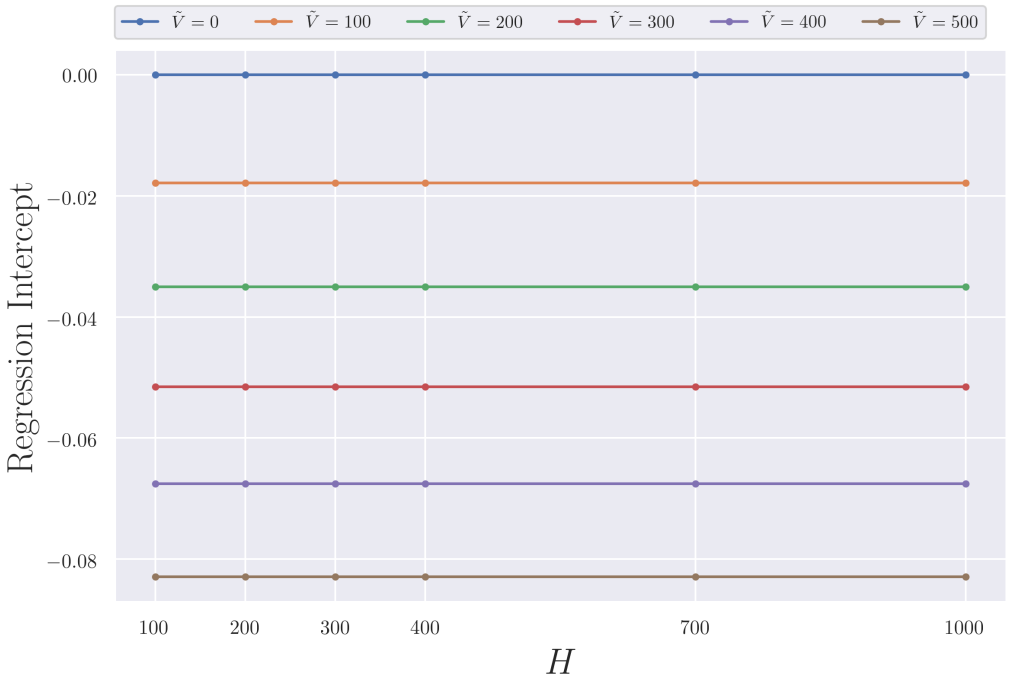


(a) AEM

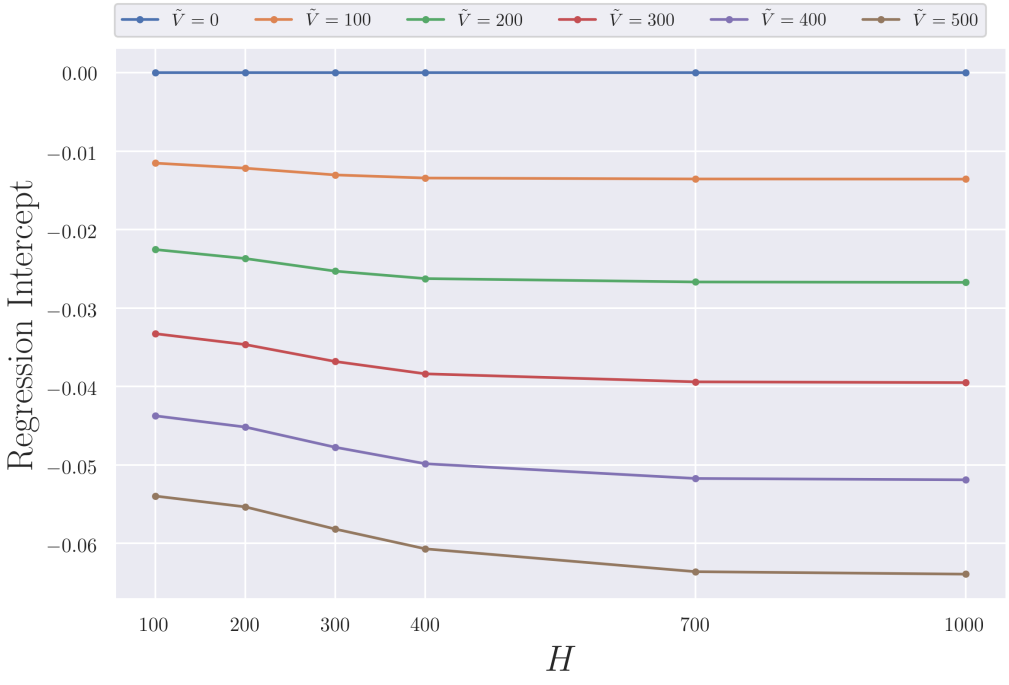


(b) BMO

Figure 5.17: Comparing the dependency of the slope of the regression line on H and $\tilde{V}$ for AEM and BMO stock on April 17, 2017 using the entire trading day over 5 second intervals.



(a) AEM



(b) BMO

Figure 5.18: Comparing the dependency of the intercept of the regression line on H and $\tilde{V}$ for AEM and BMO stock on April 17, 2017 using the entire trading day over 5 second intervals.

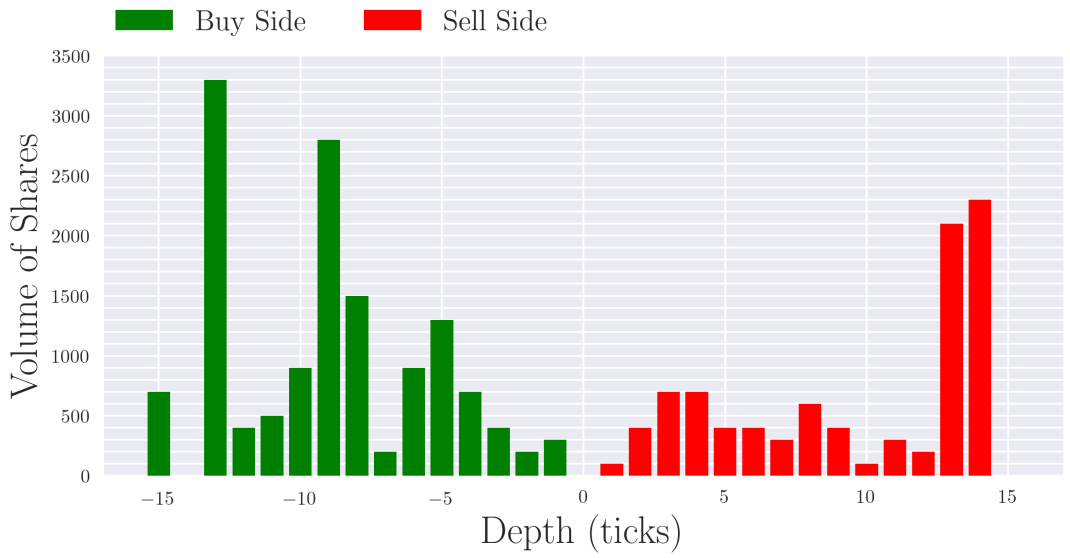
in the book with a nonzero weight. Also, for BMO, in Figure 5.16 (b) there is more weight at tick 6 than tick 7 – because of this, it may be optimal to place spoofing orders at tick 6 to gain the impact of the larger weight. If tick 7 had the larger weight you would always spoof that location since it is less risky and carries a larger impact on the imbalance.

From Figures 5.17 and 5.18 we see both the slope m and intercept b of the regression line $I^{\text{post-spoof}} = mI^{\text{pre-spoof}} + b$ of Figure 5.12 with increasing H for a fixed $\tilde{V}$. This is caused by the spoofing orders being shifted to the larger weight tick 6 to the smaller weight tick 7 – thus giving a smaller impact on I after spoofing. However, the regression slope and intercept are just capturing an aggregate over the entire day of this behaviour, so we can take specific limit order book examples from the day and investigate the optimal spoofing order placement for a given H and $\tilde{V}$ to see how the controls change in each case.

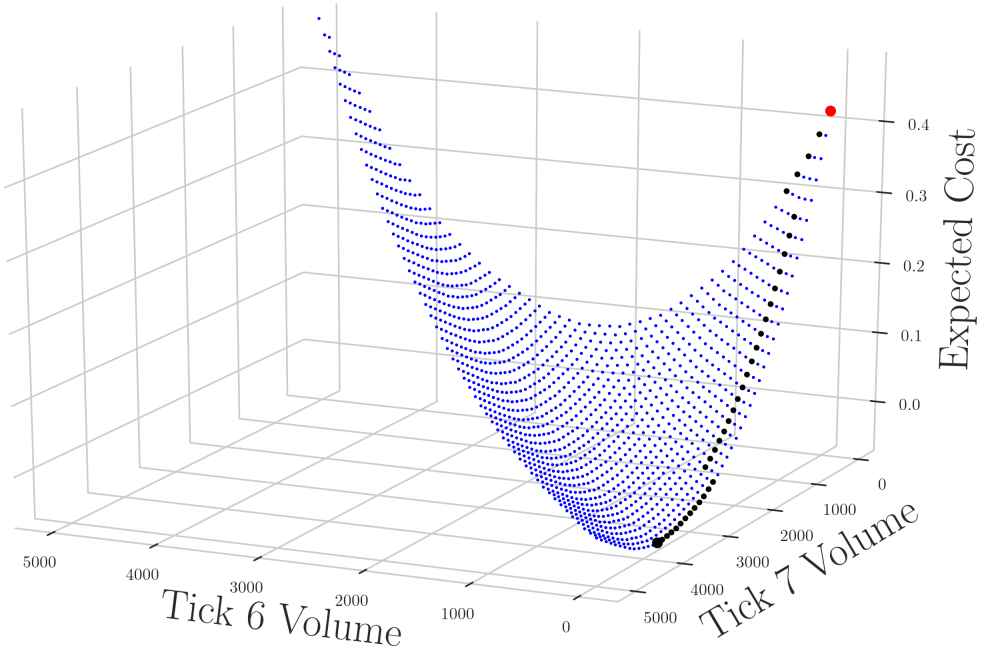
We choose four different limit order books for BMO on April 17, 2017 as shown in subplot (a) of Figures 5.19, 5.20, 5.21, and 5.22 to represent large positive, small positive, small negative, and large negative imbalances, respectively. In subplot (b) we show the objective surface of the expected spoofing cost μ_{S} for different values of $\tilde{v}_6$ and $\tilde{v}_7$ with $H = 300$. We know from the previous optimizations of BMO on April 17, 2017 that those were the only locations where spoofing orders were placed. We also have a large red point at $(\tilde{v}_6, \tilde{v}_7) = (0, 0)$ representing the case where we do not spoof – this would be the value of μ_{DMO}. Moving out from this red point are lines of constant $\tilde{V}$ with the optimal allocation of the spoofing shares for each $\tilde{V}$ being represented by a black point. For example, in Figure 5.19, the first line is $\tilde{V} = 100$ where we can place either 100 shares at tick 6 or 100 shares at tick 7 – these are the two points closest to the red point. In this case placing 100 shares at tick 6 provides the biggest expected cost reduction. For $\tilde{V} = 200$ we have 3 points now – 200 shares at tick 6, 200 shares at tick 7, or 100 shares on each. Here the optimal strategy is again to place all 200 shares on tick 6. This process is repeated until we find the global minimum at $(\tilde{v}_6, \tilde{v}_7) = (0, 3700)$.

We should also note the existence of a global minimum for our four cases and that the spoofer would not just place an infinite number of shares at either tick 6 or tick 7 to drive the imbalance to -1. The fear of their spoofing orders being executed is strong enough that even with infinite resources the spoofer can still lose if they are not allocated properly. Clearly, for these cases, the spoofer realizes diminishing returns when spoofing with an increasing number of shares. This is not a proof of the convexity of the problem, but it suggests that convexity may be provable in some cases.

Interestingly, it is not just the initial imbalance that determines how much the spoofer

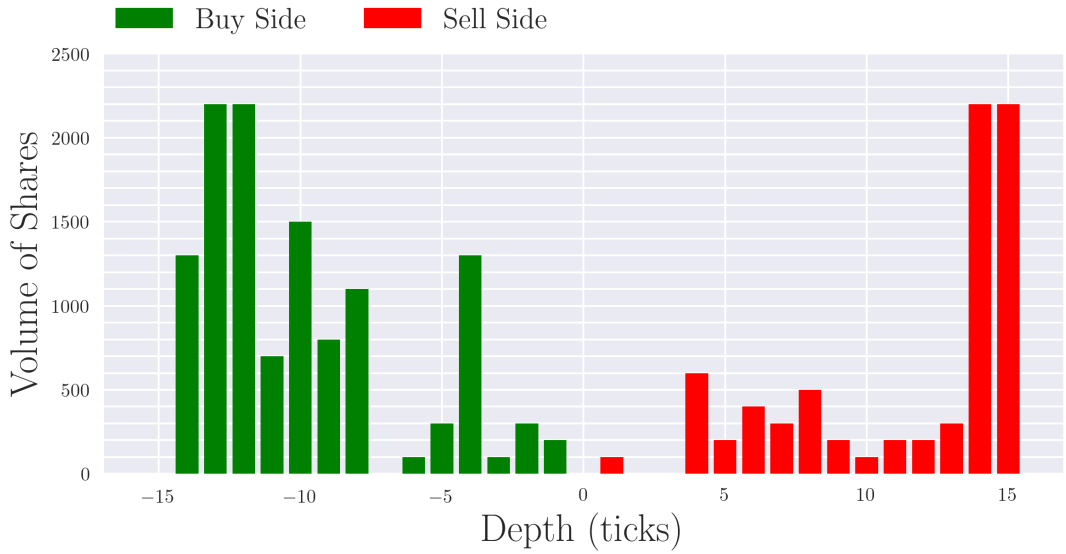


(a) Large positive imbalance ≈ 0.35

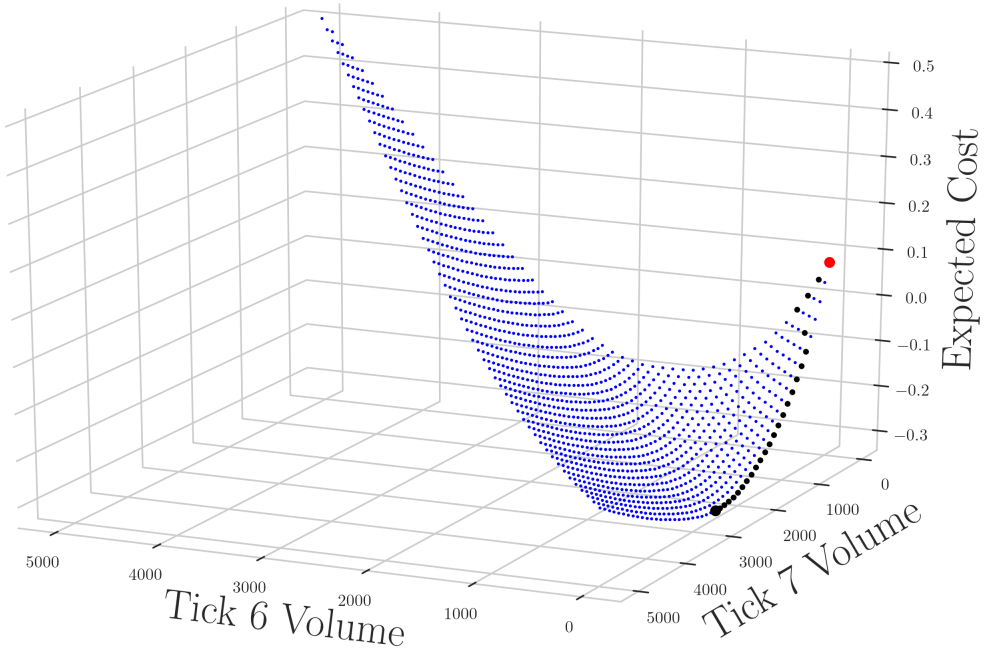


(b) μ_S by strategy for $H = 300$

Figure 5.19: Example limit order book for BMO stock on April 17, 2017 with large positive imbalance. The second plot is the surface μ_S as a function of the spoofing volumes $\tilde{v}_6$ and $\tilde{v}_7$ – the volumes placed 5 and 6 ticks from the best ask, respectively. The large red point represents the case where $\tilde{V} = 0$ which is the expected net savings for a delayed market order. The path of black points leading to the largest point is the optimal strategy for increasing $\tilde{V}$ to the global minimum at $(\tilde{v}_6, \tilde{v}_7) = (0, 3700)$ with $\mu_S = -0.0601$. The imbalance after spoofing is $I \approx -0.238$.

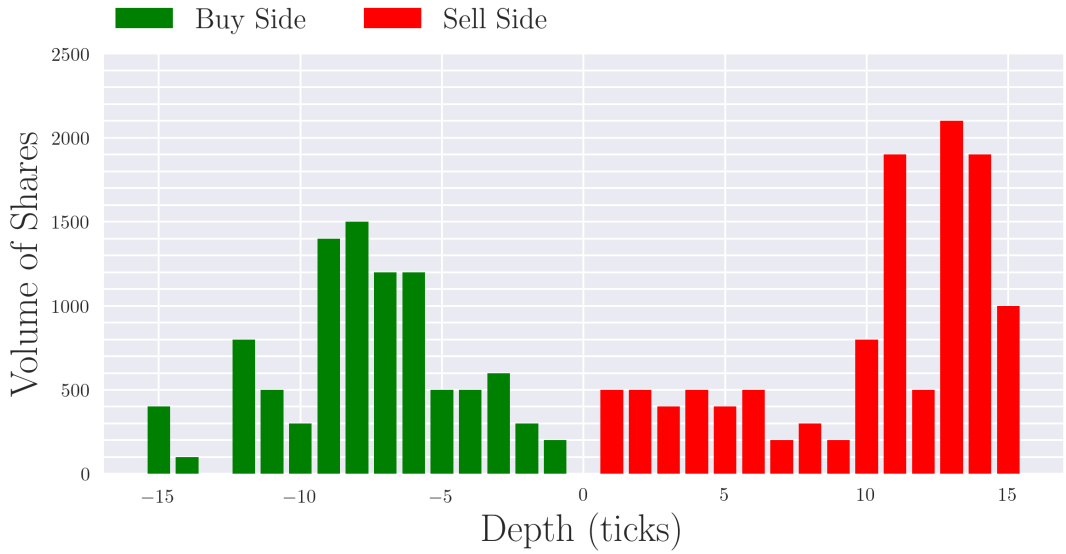


(a) Small positive imbalance ≈ 0.062

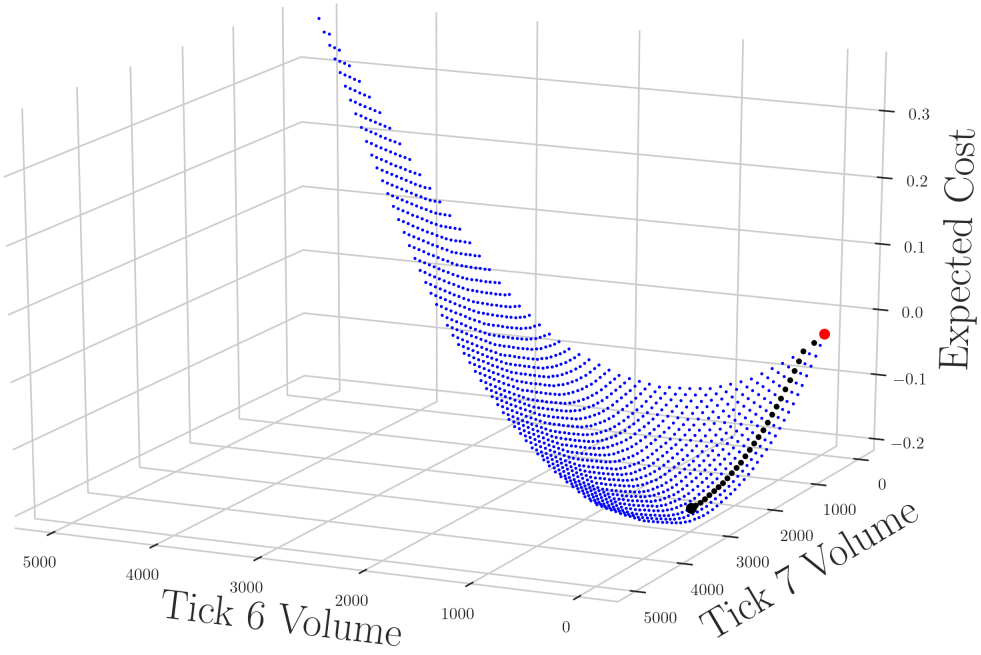


(b) μ_S by strategy for $H = 300$

Figure 5.20: Example limit order book for BMO stock on April 17, 2017 with small positive imbalance. The second plot is the surface μ_S as a function of the spoofing volumes $\tilde{v}_6$ and $\tilde{v}_7$ – the volumes placed 5 and 6 ticks from the best ask, respectively. The large red point represents the case where $\tilde{V} = 0$ which is the expected net savings for a delayed market order. The path of black points leading to the largest point is the optimal strategy for increasing $\tilde{V}$ to the global minimum at $(\tilde{v}_6, \tilde{v}_7) = (0, 2500)$ with $\mu_S = -0.327$. The imbalance after spoofing is $I \approx -0.327$.

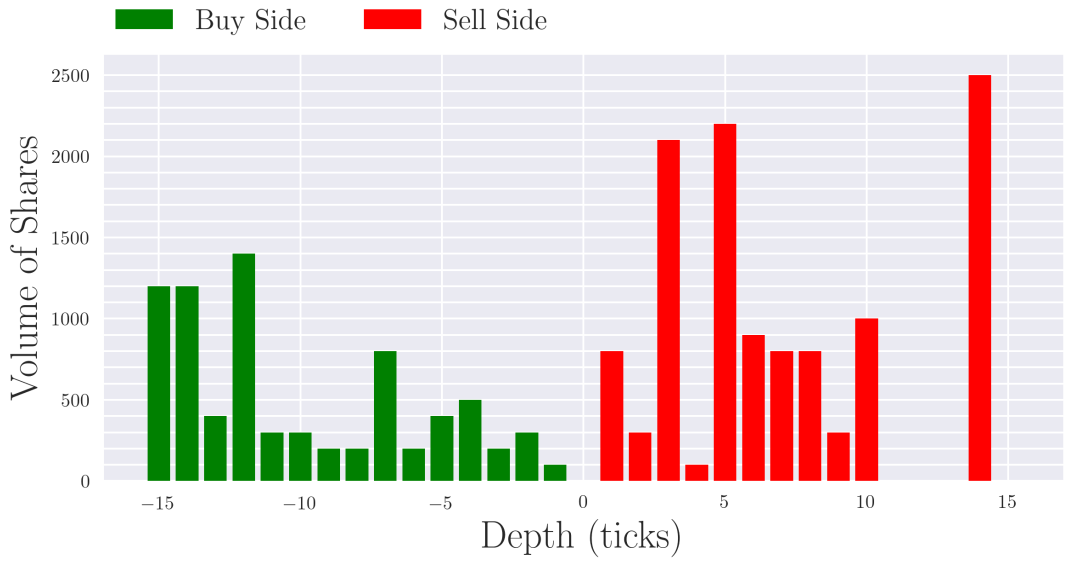


(a) Small negative imbalance ≈ -0.024

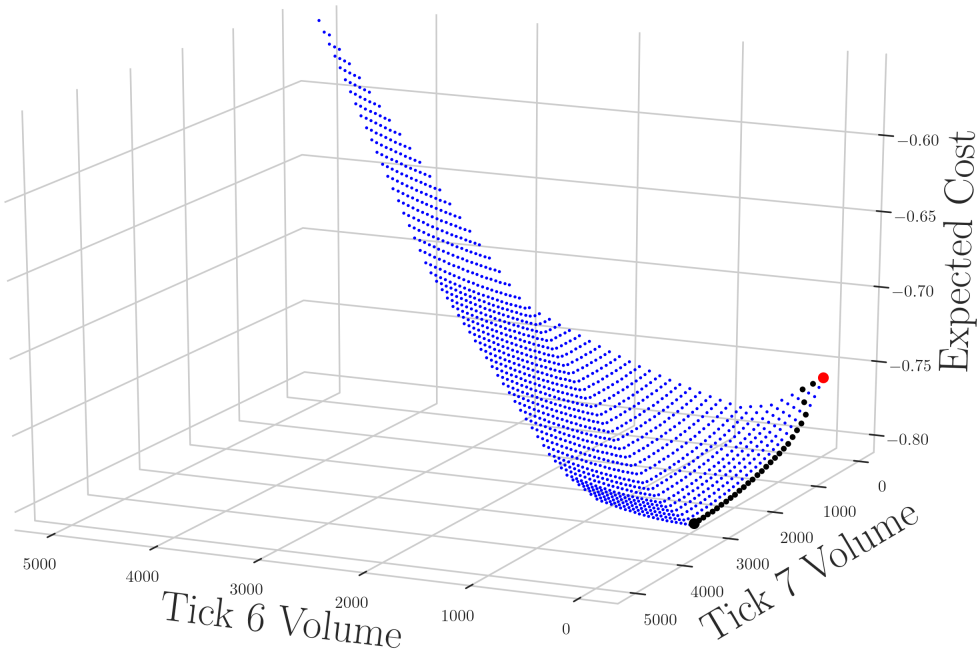


(b) μ_S by strategy for $H = 300$

Figure 5.21: Example limit order book for BMO stock on April 17, 2017 with small negative imbalance. The second plot is the surface μ_S as a function of the spoofing volumes $\tilde{v}_6$ and $\tilde{v}_7$ – the volumes placed 5 and 6 ticks from the best ask, respectively. The large red point represents the case where $\tilde{V} = 0$ which is the expected net savings for a delayed market order. The path of black points leading to the largest point is the optimal strategy for increasing $\tilde{V}$ to the global minimum at $(\tilde{v}_6, \tilde{v}_7) = (200, 2500)$ with $\mu_S = -0.206$. The imbalance after spoofing is $I \approx -0.263$.



(a) Large negative imbalance ≈ -0.58



(b) μ_S by strategy for $H = 300$

Figure 5.22: Example limit order book for BMO stock on April 17, 2017 with large negative imbalance. The second plot is the surface μ_S as a function of the spoofing volumes $\tilde{v}_6$ and $\tilde{v}_7$ – the volumes placed 5 and 6 ticks from the best ask, respectively. The large red point represents the case where $\tilde{V} = 0$ which is the expected net savings for a delayed market order. The path of black points leading to the largest point is the optimal strategy for increasing $\tilde{V}$ to the global minimum at $(\tilde{v}_6, \tilde{v}_7) = (0, 2900)$ with $\mu_S = -0.807$. The imbalance after spoofing is $I \approx -0.684$.

can impact the book. The actual shape of both sides of the book is just as important for determining just how negative the imbalance can be brought through spoofing. For example, in Figure 5.20 the imbalance after spoofing at the global minimum is ≈ -0.327 while in Figure 5.21 the imbalance after spoofing is ≈ -0.263 even though Figure 5.20 has an initial positive imbalance while Figure 5.21 has a small negative imbalance. The initial imbalance does not tell the whole story: the imbalances of some limit order books are more sensitive to spoofing than others simply due to how the volume of shares are spread over both sides of the book.

In each of the four figures we see that the optimal strategy moves from allocating all shares at tick 6 to tick 7 with increasing $\tilde{V}$. In all cases, except Figure 5.21, all shares are eventually placed entirely at tick 7. In Figure 5.21 the spoofer will still place 200 shares at tick 6 even at the global minimum. This is behaviour we see in multiple cases for H and $\tilde{V}$ for each of our four limit order books. The spoofer is willing to take the extra risk with these 200 shares for the increased impact on the imbalance because even if they are executed they would not walk the book according to Figure 5.21 (a) since the 200 executed shares at tick 6 plus the 300 shares we buy is equal to the 500 shares at the best ask. So, the shape of the limit order book near the touch is also important for determining optimal spoofing strategies. We see these results because we have made the modelling assumption that the book would remain constant over Δt, but this is obviously not typically the case. Because of this the spoofer is able to allocate their spoofing shares to locations that are guaranteed they minimize the distance they walk the book. A more accurate model would have the spoofer evaluating expected values not only with respect to changes in the best ask, but also changes to the volumes near the best ask given the current state of the book. However, the modelling required there is much harder.

The impact of the shape of the book on the objective function is even more clear in Figure 5.22 (b) with the obvious changes in slope at specific points. Since the book already has a large negative imbalance there is little improvement from spoofing and changes in the objective function are mostly determined by the costs associated with executed spoofing orders and walking the book to recover those shares while still having to purchase H shares. If we follow the points away from the red point where all spoofing shares are placed at tick 6 we see obvious changes in slope at $\tilde{V} = 500, 800,$ and 2900. Combining with the $H = 300$ we also have to purchase, these three quantities correspond to the 800, 300, and 2100 shares at the first three ticks of the ask side of the limit order book. We do not see such clear changes in the objective function caused by the shape of the book in the other cases because we can impact the imbalance more and benefit from the increased probability of the best ask decreasing – this smooths out the surface and

removes edges we see in Figure 5.22 (b).

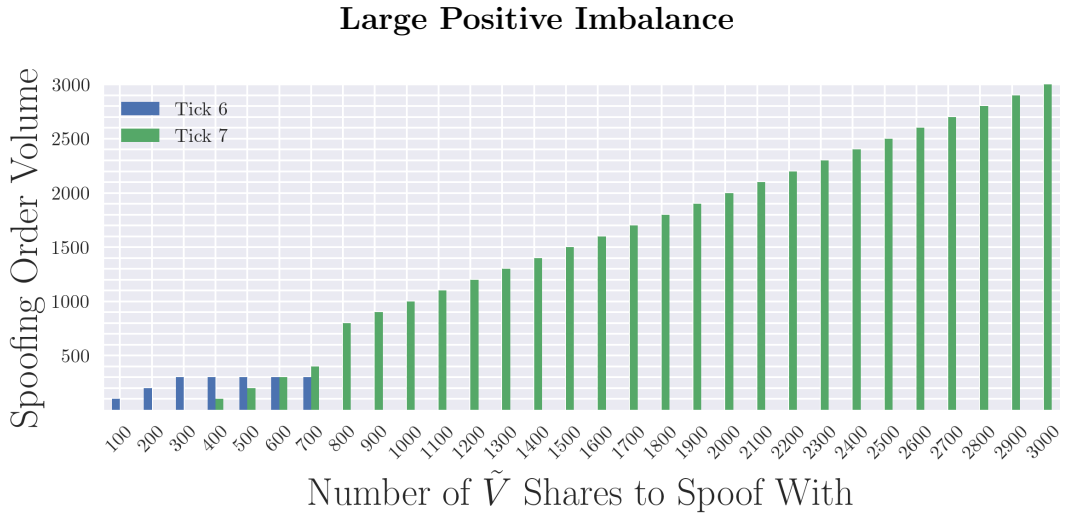
In Figures 5.23, 5.24, 5.25, and 5.26 we present the optimal spoofing strategy for our four limit order books for different values of H and increasing values of $\tilde{V}$. For each value of $\tilde{V}$ we plot the number of shares placed at tick 6 and tick 7 in blue and green, respectively. The more shares the spoofer wishes to buy the more risk they are willing to take with their spoofing orders if $\tilde{V}$ is small since they want to influence the book as best they can with what limited resources they have available, but as the spoofer is willing to spoof with increasing shares they are instead placed at the less risky tick 7. There is then a transition period between these two cases where the spoofer mixes their spoofing orders between the two locations. Then, as seen in Figure 5.21, the spoofer is still willing to leave some shares at tick 6 in the limit $\tilde{V} \to \infty$ based on how deep they would have to walk the book to recover those shares if they were executed.

For example, in Figure 5.23 (b), the spoofer still leaves 100 shares at tick 6. They are already buying 400 shares and, according to Figure 5.19 (a), would already walk the book one tick to the right of the best ask and leave 100 shares at this position. So, the spoofer could have 100 shares executed against them and not have to walk the book to recover them. Again, in Figure 5.23 (c) the spoofer leaves 600 shares at tick 6 because buying 600 shares would cause you to walk the book two ticks from the best ask anyway with 600 shares left at that position.

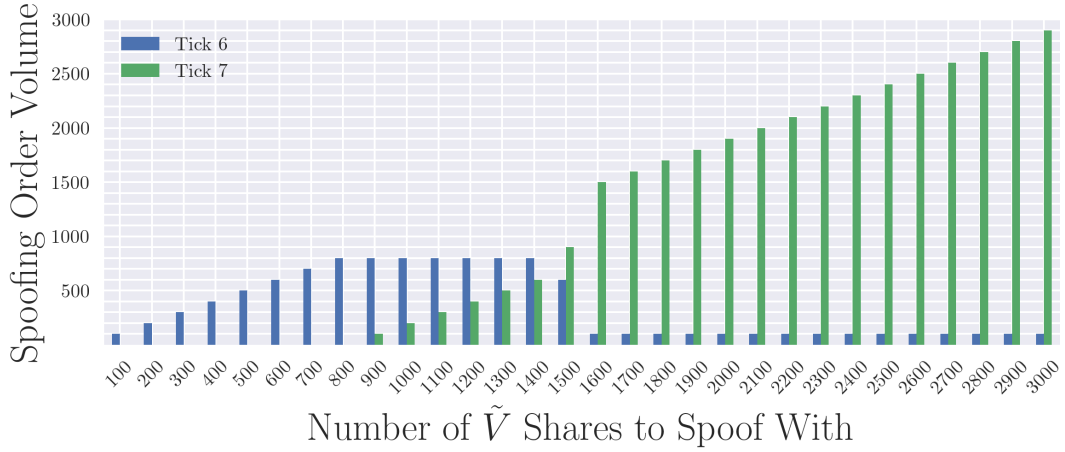
This behaviour is not followed in all cases for H and $\tilde{V}$ for each limit order book snapshot, but for the cases it is we are able to see why the spoofer would still risk some amount of shares being executed. In some cases the spoofer could impact the imbalance enough at tick 7 that the risk of any shares being executed at tick 6 is not worth it – as in Figures 5.24 (b), 5.25 (a), or 5.26 (b) for example. These results further emphasize the importance of properly modelling the shape of the book at the next time period given the current shape.

We have seen that minimize spoofing order execution is important for the strategy of the spoofer. Built into the model is the idea that the spoofer attempts to minimize their risks while maximizing their profits through lower the expected cost of their intended market order. Just like the previous section it may be more paramount to the spoofer to minimize the expected net savings relative to the risk associated with that net savings – minimizing the Sharpe ratio of their spoofing strategy rather than the pure savings itself. In this spirit we can also see how the optimal strategy from Figures 5.19, 5.20, 5.21, and 5.22 changes under a different minimization criteria. So, rather than minimize μ_{S} with respect to $(\tilde{v}_6, \tilde{v}_7)$ we minimize S_{S}.

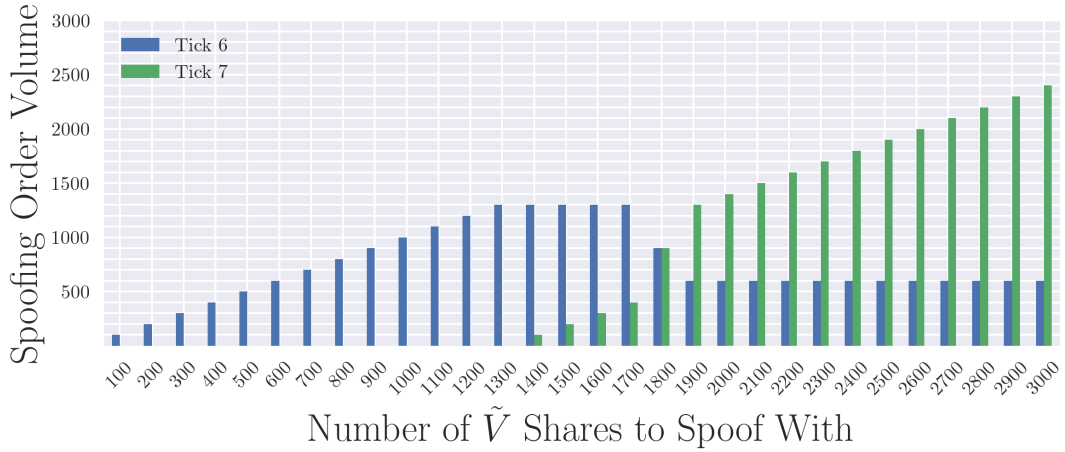
This is similar to the optimization problem in equations 5.2.1 and 5.2.2, but we write



(a) $H = 200$

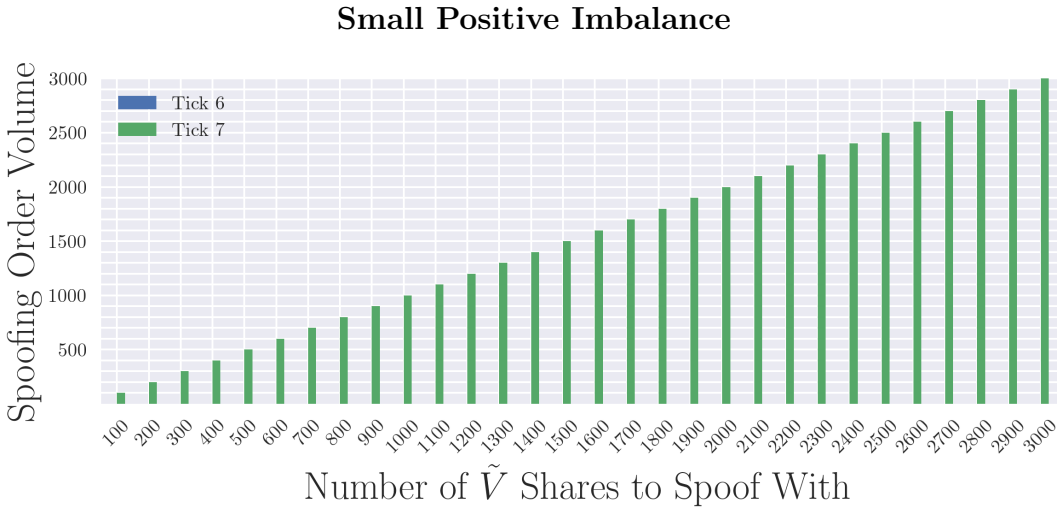


(b) $H = 400$

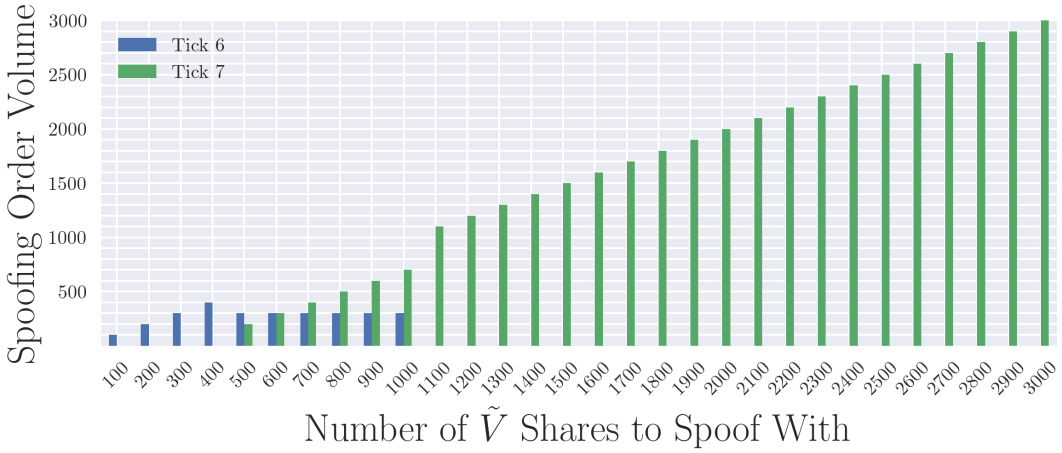


(c) $H = 600$

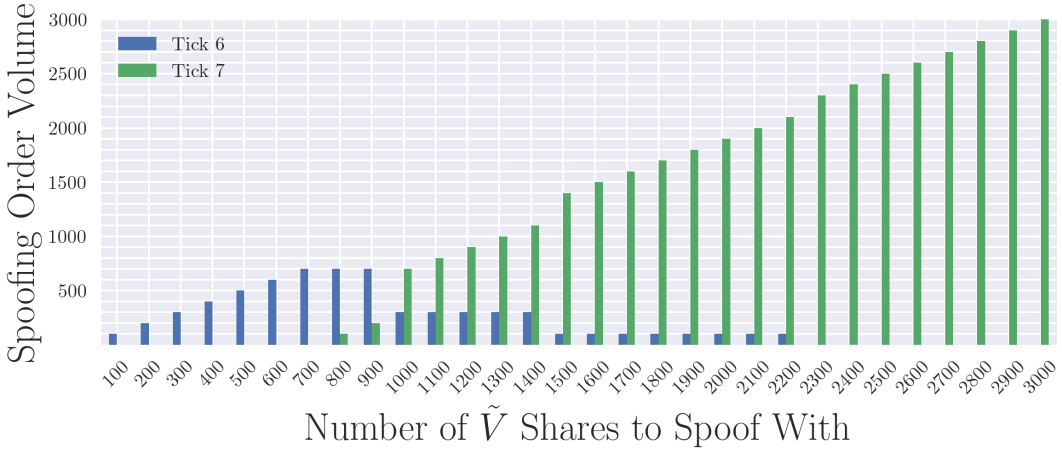
Figure 5.23: Optimal spoofing order placement with changing H and $\tilde{V}$ for large positive imbalance as seen in Figure 5.19.



(a) $H = 200$

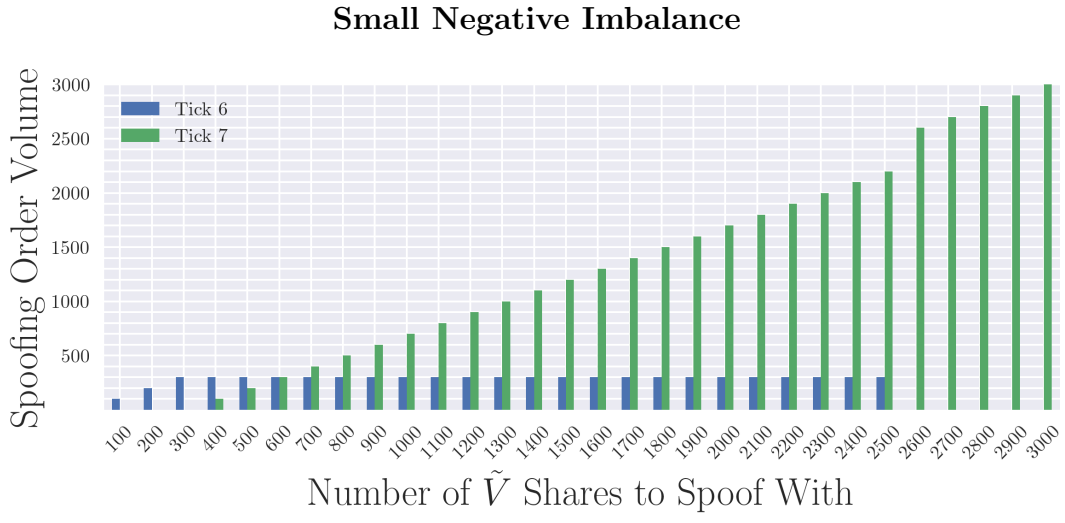


(b) $H = 400$

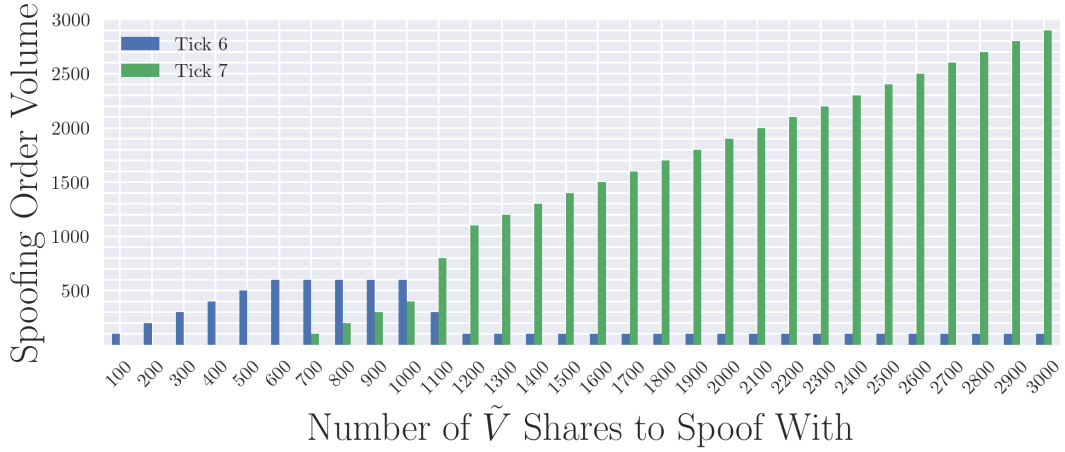


(c) $H = 600$

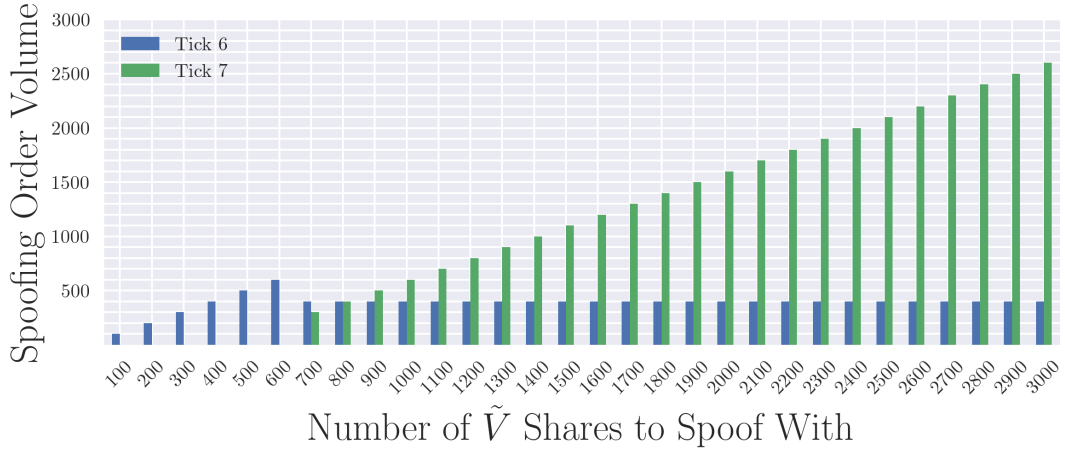
Figure 5.24: Optimal spoofing order placement with changing H and $\tilde{V}$ for small positive imbalance as seen in Figure 5.20.



(a) $H = 200$

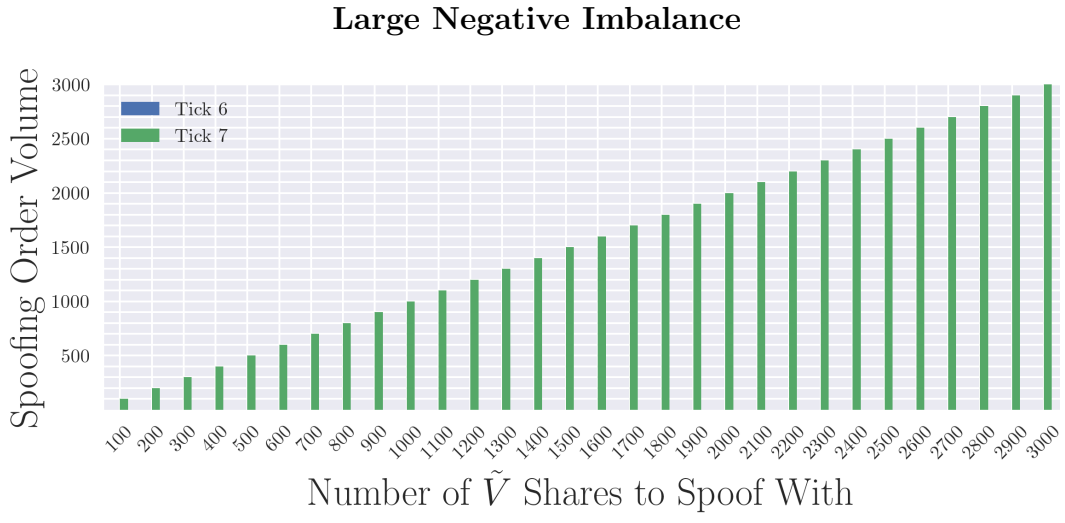


(b) $H = 400$

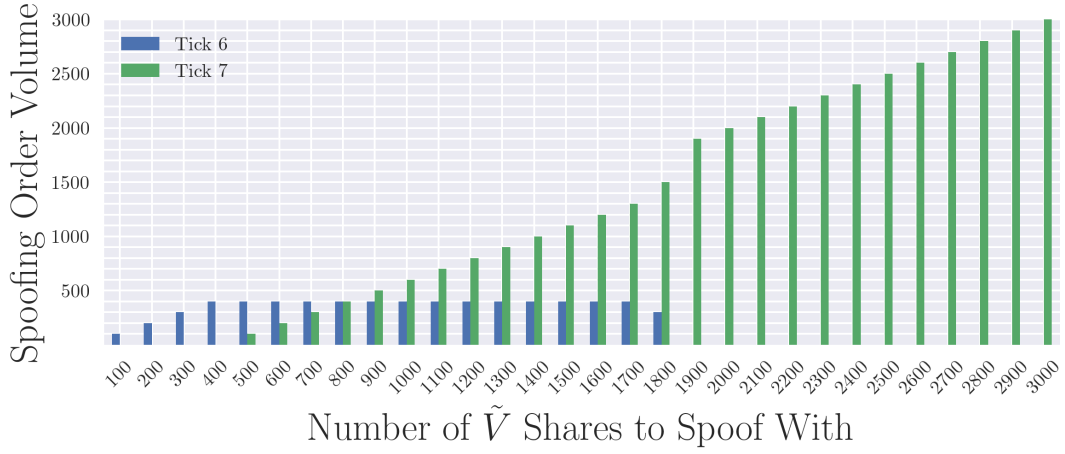


(c) $H = 600$

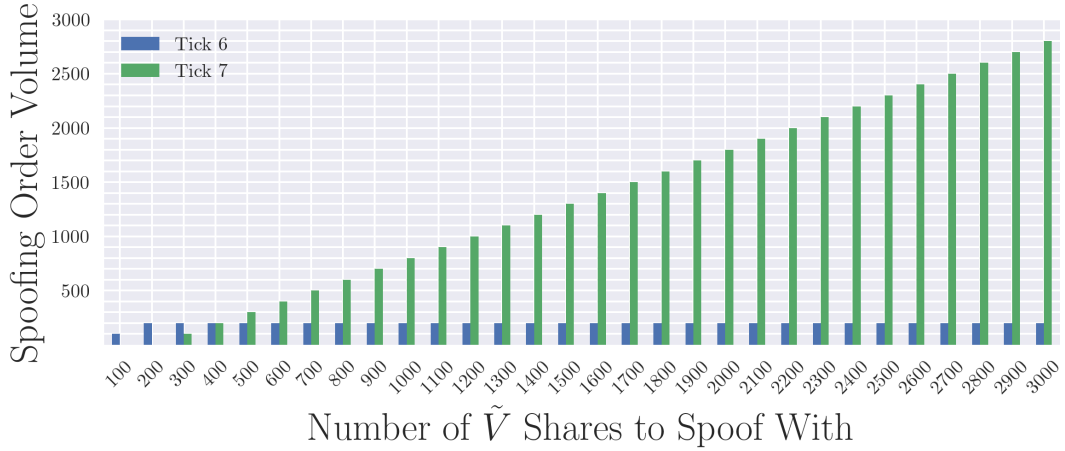
Figure 5.25: Optimal spoofing order placement with changing H and $\tilde{V}$ for small negative imbalance as seen in Figure 5.21.



(a) $H = 200$



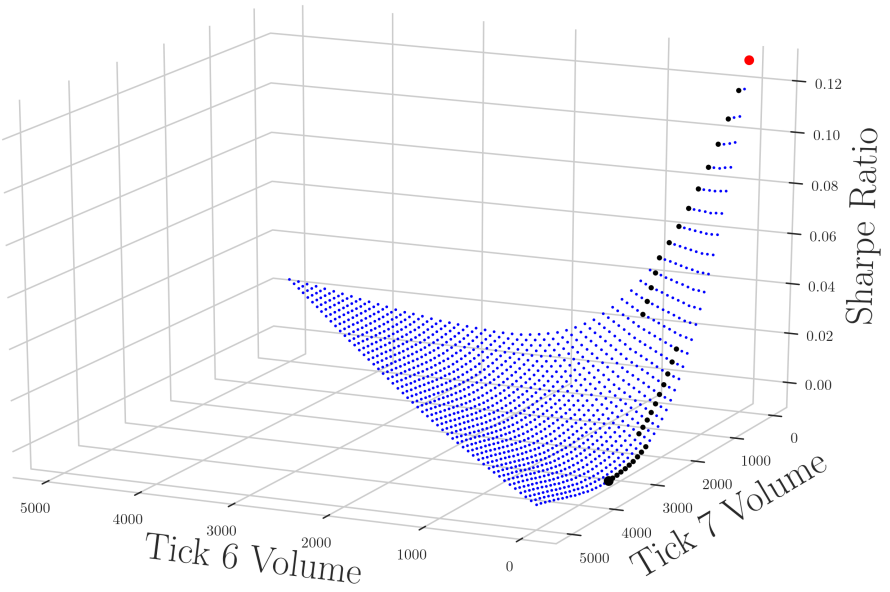
(b) $H = 400$



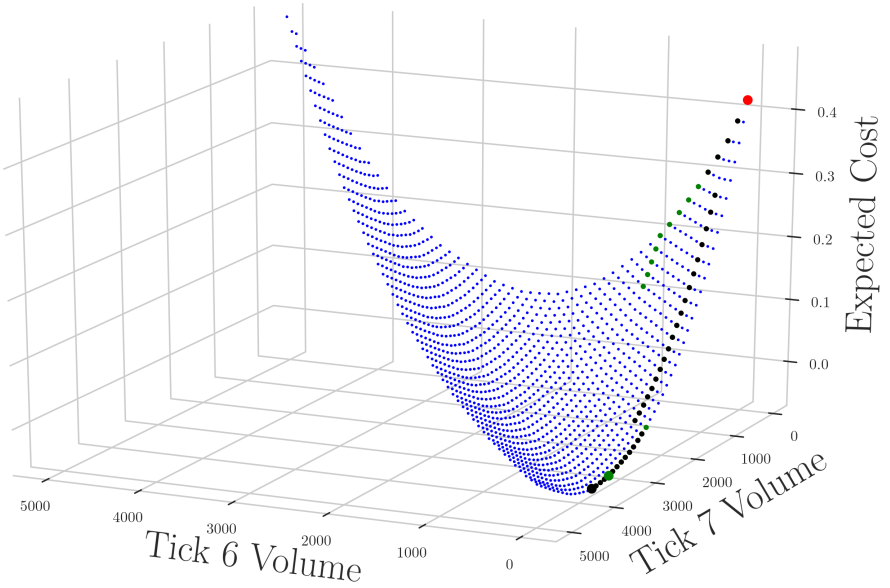
(c) $H = 600$

Figure 5.26: Optimal spoofing order placement with changing H and $\tilde{V}$ for large negative imbalance as seen in Figure 5.22.

Large Positive Imbalance



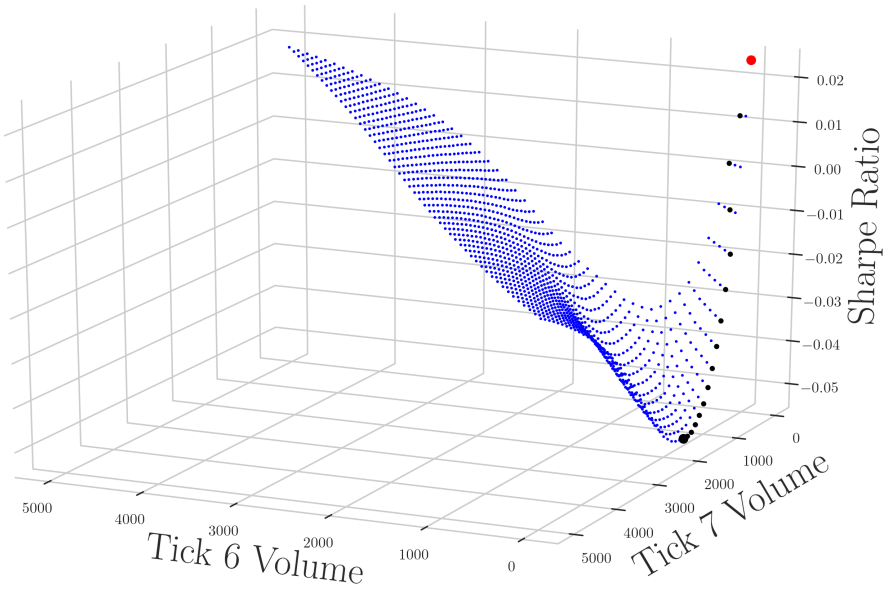
(a) S_{S} by strategy for $H = 300$



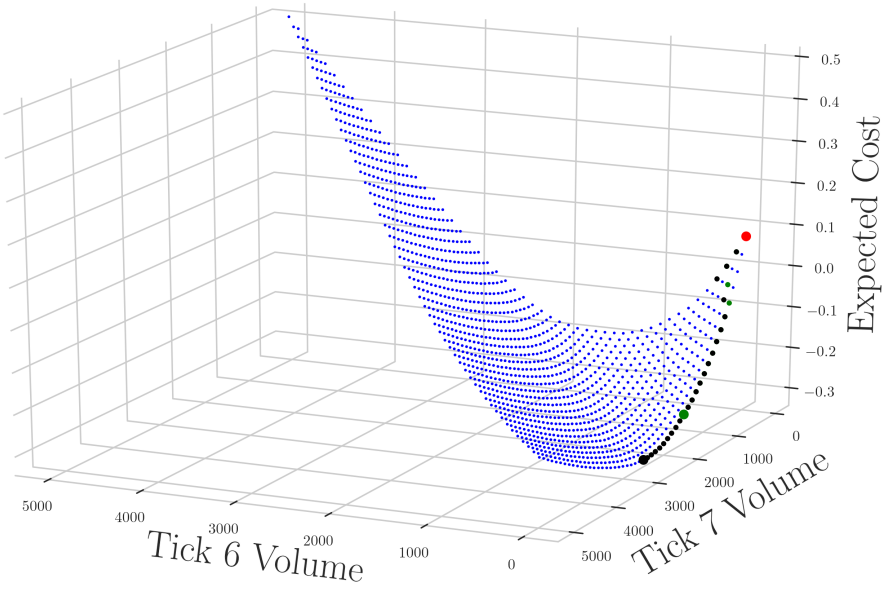
(b) μ_{S} by strategy for $H = 300$

Figure 5.27: Comparison of optimal spoofing strategy for limit order book in Figure 5.19 using Sharpe ratio S_{S} instead of expected net savings μ_{S}. Both strategies are plotted together in subplot (b) with green points representing the Sharpe ratio strategy. The largest green point being the global minimum for the Sharpe ratio at $(\tilde{v}_6, \tilde{v}_7) = (0, 3300)$. If a line of $\tilde{V} =$ constant has only a black point then the two strategies are the same. At the Sharpe ratio minimum we have $\mu_{\mathrm{S}} = -0.0551$, $S_{\mathrm{S}} = -0.00734$, and $I = -0.202$. At the expected cost minimum $S_{\mathrm{S}} = -0.00696$.

Small Positive Imbalance

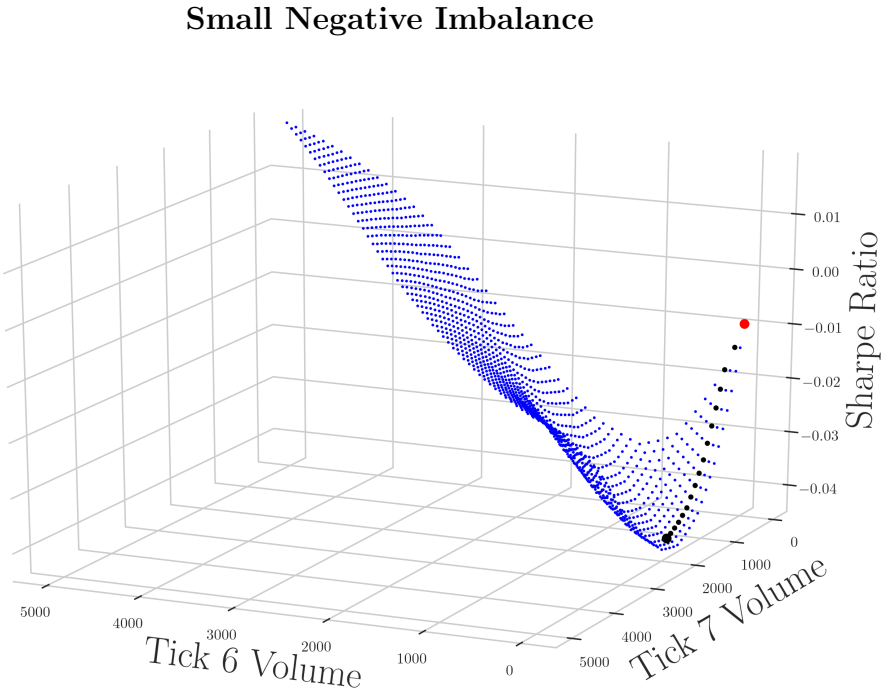


(a) S_{S} by strategy for $H = 300$

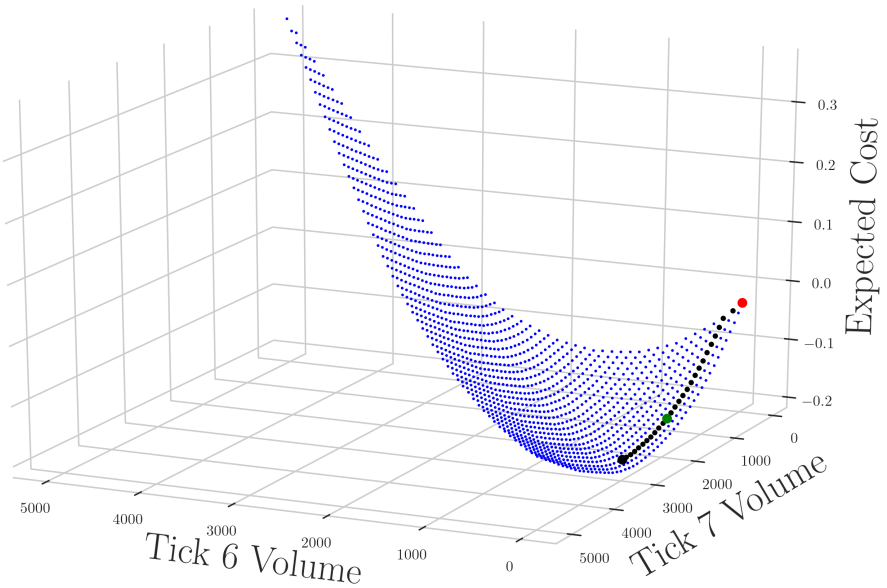


(b) μ_{S} by strategy for $H = 300$

Figure 5.28: Comparison of optimal spoofing strategy for limit order book in Figure 5.20 using Sharpe ratio S_{S} instead of expected net savings μ_{S}. Both strategies are plotted together in subplot (b) with green points representing the Sharpe ratio strategy. The largest green point being the global minimum for the Sharpe ratio at $(\tilde{v}_6, \tilde{v}_7) = (0, 1500)$. If a line of $\tilde{V} =$ constant has only a black point then the two strategies are the same. At the Sharpe ratio minimum we have $\mu_{\mathrm{S}} = -0.272$, $S_{\mathrm{S}} = -0.0539$, and $I = -0.291$. At the expected cost minimum $S_{\mathrm{S}} = -0.0414$.



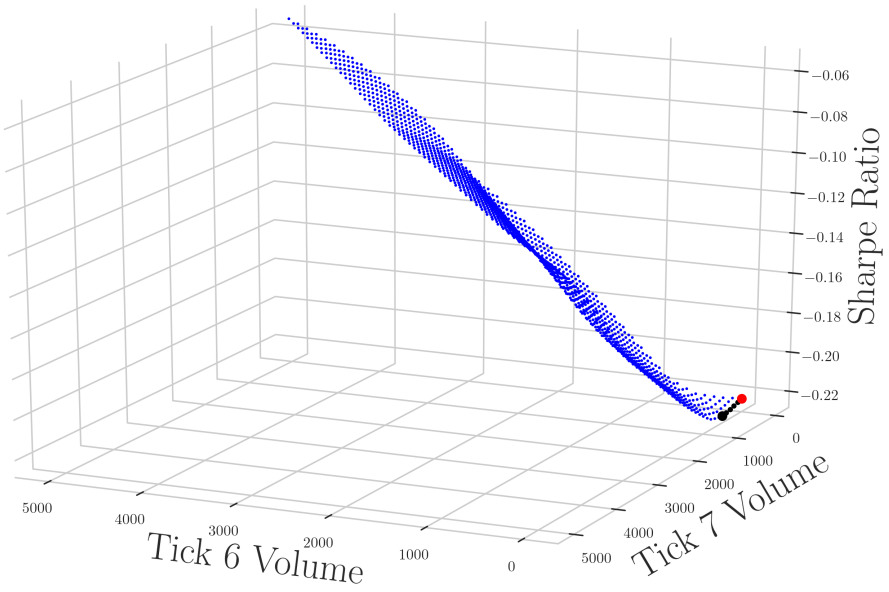
(a) S_{S} by strategy for $H = 300$



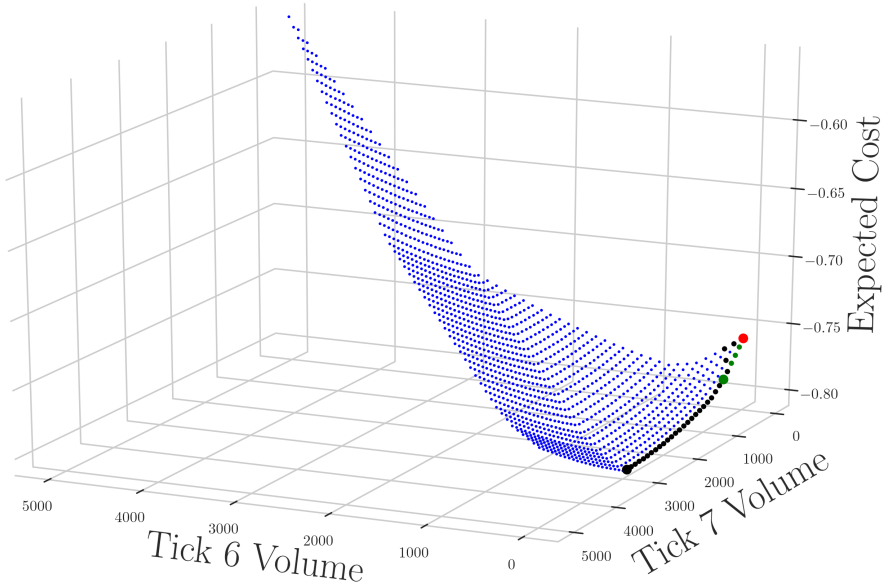
(b) μ_{S} by strategy for $H = 300$

Figure 5.29: Comparison of optimal spoofing strategy for limit order book in Figure 5.21 using Sharpe ratio S_{S} instead of expected net savings μ_{S}. Both strategies are plotted together in subplot (b) with green points representing the Sharpe ratio strategy. The largest green point being the global minimum for the Sharpe ratio at $(\tilde{v}_6, \tilde{v}_7) = (200, 1400)$. If a line of $\tilde{V} =$ constant has only a black point then the two strategies are the same. At the Sharpe ratio minimum we have $\mu_{\mathrm{S}} = -0.179$, $S_{\mathrm{S}} = -0.0437$, and $I = -0.182$. At the expected cost minimum $S_{\mathrm{S}} = -0.0358$.

Large Negative Imbalance



(a) S_{S} by strategy for $H = 300$



(b) μ_{S} by strategy for $H = 300$

Figure 5.30: Comparison of optimal spoofing strategy for limit order book in Figure 5.22 using Sharpe ratio S_{S} instead of expected net savings μ_{S}. Both strategies are plotted together in subplot (b) with green points representing the Sharpe ratio strategy. The largest green point being the global minimum for the Sharpe ratio at $(\tilde{v}_6, \tilde{v}_7) = (0, 500)$. If a line of $\tilde{V} =$ constant has only a black point then the two strategies are the same. At the Sharpe ratio minimum we have $\mu_{\mathrm{S}} = -0.780$, $S_{\mathrm{S}} = -0.225$, and $I = -0.604$. At the expected cost minimum $S_{\mathrm{S}} = -0.170$.

it here for clarity:

$$\min_{\tilde{v}_t} \frac{E\left[\, C_{\mathrm{S}}(\vec{v}_t, H, p^+, x_t, \tilde{v}_t) \,|\mathcal{F}_t, I(\vec{v}_t + \tilde{v}_t)\right] - C_{\mathrm{MO}}(\vec{v}_t, H, p^+)}{\sqrt{\mathrm{Var}\left[\, C_{\mathrm{S}}(\vec{v}_t, H, p^+, x_t, \tilde{v}_t) \,|\mathcal{F}_t, I(\vec{v}_t + \tilde{v}_t)\right] - C_{\mathrm{MO}}(\vec{v}_t, H, p^+)}} \tag{5.4.1}$$

Subject to

$$\sum_{i=-K}^{0} \tilde{v}_i = 0, \quad \sum_{i=1}^{K} \tilde{v}_i \leq \tilde{V}, \quad \tilde{v}_i \geq 0 \text{ and } \tilde{v}_i \in 100\mathbb{N} \quad \forall i \in [-K, K] \tag{5.4.2}$$

Subplot (a) in Figures 5.27, 5.28, 5.29, and 5.30 gives the optimal spoofing strategy according to the Sharpe ratio S_{S}. Subplot (b) plots the optimal strategy according to S_{S} and μ_{S} on the same objective surface μ_{S}. However, the optimal strategy according to the Sharpe ratio is given by green points with the global minimum a large green point. If only a black point is present for a line of constant $\tilde{V}$ then the two strategies coincide.

The optimal strategy according to the Sharpe ratio follows a similar path to the one found from the expected cost, but ultimately the global minimum is always found using fewer spoofing shares. The structure of the Sharpe ratio is also interesting in that as $\tilde{V} \to \infty$ the Sharpe ratio asymptotes to 0 since the variance grows to infinity as you spoof with an increasing number of shares. The other limit, as $\tilde{V} \to 0$, determines how large a cost reduction the spoofer can expect for their risk. If the red point lies too far above zero there may be no global minimum at all or it may lie so far out in the surface the net savings is still too risky (as in Figure 5.27). If it lies too far below zero the optimal strategy will be to do nothing at all (as in Figure 5.30). However, in between these extremes gives the spoofer an opportunity to see real cost savings for their risk.

For example, in Figure 5.27, the imbalance is initially so positive that a spoofer would need to spoof with 3300 shares to see an expected cost reduction on the purchase of only 300 shares. This is a significant risk to take on comparatively small market order, and the Sharpe ratio of less than 1% return per unit of risk reflects this. In Figures 5.28 and 5.29, the imbalance is small enough that the spoofer can find a solid return for their risk using only 1500 and 1600 spoofing shares, respectively. However, in Figure 5.30, the imbalance is already so in favour of the spoofer that spoofing with more than 500 shares only hurts the Sharpe ratio.

In the previous sections we used the expected cost of spoofing the book for determining the optimal strategy, but the Sharpe ratio builds in the actions one would expect from a spoofer attempting to manipulate the book with the smallest amount of risk to their orders. This may lead to further interesting results if we repeated our previous analysis under this new optimization criteria, but the expected net savings at the point which

minimizes the Sharpe ratio is never significantly worse and we would not expect the results of the previous sections to deviate much from what we have presented. The major different between the two methods would be when looking into the book for specific spoofing orders because when using the Sharpe ratio to determine the optimal strategy the spoofer is not willing to risk as many spoofing shares compared to if they are driven purely by achieving the largest possible cost reduction. We would need access to already labeled data sets which include known spoofing orders into order to make any judgment on which method would provide a more accurate strategy to what a spoofer might actually do in reality.

5.5 Conclusions

This chapter is a culmination of the model building and calibration we presented in the previous chapters. Equipped with the tools needed to explore the costs associated with spoofing limit order books we were able to analyze the sensitivity of the book for a given stock throughout the day as well as draw some general rules from the aggregated results.

We started by establishing two criteria for determining whether a spoofer would either immediately place a market order, delay their market order to the next time period, or spoof the book combined with a delayed market order. These two criteria were based on the expected cost reduction and the Sharpe ratio – the latter being used to take into account the risk associated with the execution of the spoofing orders against the spoofer. We found that the Sharpe ratio criteria provided a more consistent decision boundary between these three possible actions as the Sharpe ratio punished spoofing the book with a considerable volume of shares to earn tiny cost reductions of a couple of pennies on market orders of hundreds of shares. This gave better clusters between the decisions to spoof and immediately place a market order. The improvement between these two clusters allowed us to explore the dependency of the boundary between them as a function of H and $\tilde{V}$.

We then used these decision boundaries to determine a relationship between the number of shares H the spoofer wishes to buy, the number of shares $\tilde{V}$ they are willing to spoof with, and the decision boundary between spoofing and immediately placing a market order. We explored this relationship for AEM, BMO, and CNR stocks and found that a spoofer is willing to spoof a more positive imbalanced book with increasing $\tilde{V}$. A spoofer aiming to purchase more shares was also more willing to manipulate the book even if their spoofing orders were less able to make large impacts on the imbalance – a spoofer was willing to take more risks the more shares they needed to purchase.

From this analysis of the decision boundary we found that BMO showed different behaviour from AEM and CNR because there were two locations a spoofer could impact the imbalance deep in the book as opposed to the single location for the other two stocks. The more risky locations carried more weight in the imbalance, so the spoofer needed to properly allocate their spoofing shares to balance risk with return. This balancing act made a spoofer more willing to take risks the more shares they wanted to buy as well as if they had a smaller number of shares to spoof with. The spoofer was less willing to take risks the more shares they had available to manipulate the book with. We also found that the shape of book at the next time period was important for determining the optimal strategy based on the costs associated with walking the book to buy H and recover any spoofing shares executed by movement in the best ask against the spoofer.

Finally, we explored how different the optimal spoofing strategy would be if the Sharpe ratio was used to take into account risk based on how effective this was in determing the spoofer's ultimate strategy between spoofing, delayed market orders, and immediate market orders. We found that a spoofer was less willing to risk a large number of shares using the Sharpe ratio and the strategies themselves reflected more accurately what one would naively expect from a spoofer without attempting to model their behaviour. We do not believe this alternative strategy would drastically change the decision boundary analysis, but would be possibly more effective when attempting to flag limit orders in the book as suspicious. However, this type of work would require an actual labelled data set which included limit orders which were already deemed as spoof orders by some regulator of the exchange.

Chapter 6

Conclusions and Future Work

6.1 Summary of Conclusions

The goal of this project was to develop a model which would allow us to explore the conditions under which a spoofer may attempt to manipulate a limit order book in order to gain insights into identifying potential illegal behaviour for regulators to investigate further. Not only is it difficult to find the figurative 'needle in a haystack', but one also needs evidence to the spoofer's intent to illegally spoof the book. We take the approach of stepping back from individual broker IDs to look at the aggregate behaviour of the book in response to the orders placed, so that we can find a general relationship between the shape of the book and movements in the best ask or bid. Then, a regulator could check to see if an individual trader was attempting to abuse this general relationship through spoofing limit orders. Since we did not have access to individual trader IDs or labeled cases of actual spoofing we were only able to probe the vulnerability of the limit order book to a potential spoofer using our new model.

The spoofer pads the limit order book with limit orders they never intend to execute in an attempt to get other traders to move prices in the spoofer's best interest. The spoofer manipulates the shape of the book itself to gain an advantage, so the natural start for a model would be to use a statistic which captures shape information about the limit order book and is correlated to price movements – the volume imbalance ratio. We provided initial statistical tests across 50 different stocks which support the existing literature that positive/negative price movements accompany positive/negative imbalance ratios.

The usual definition of the volume imbalance ratio involves only volumes of shares at the best ask and best bid of the book, but we want to be able to capture the imbalance deeper into the book. We use a generalized definition of the imbalance which uses weights, w, applied to each share volume in the book. We assume the weights are the same

for the bid and ask side. The first natural extension of this is to allow exponentially decaying weights from the touch, so the volumes at the best bid/ask remain the most important when calculating the imbalance ratio. Then, we can relax the exponentially decaying weight assumption and allow the weights to be free parameters, but still with the constraint that the largest weight is applied to the best ask/bid. The statistical tests we performed on the classical imbalance definition gave consistent results for these two new methods and also provided increased association between changes in the best ask and the imbalance ratio as we move from the classical definition $\rightarrow$ exponentially decaying weights $\rightarrow$ free weights. With a generalized form of the imbalance including information about volumes deeper in the limit order book we set up a model for how changes in the best ask are impacted through the volume imbalance.

With price changes one needs to also talk about the time interval Δt over which the price changed. Aggregating the instantaneous imbalances over Δt into an average imbalance was done by weighting the imbalances by the length of time the book remained in that state. This method removed the negative correlations between price changes and the imbalance ratio that we saw for some stocks when aggregating the instantaneous imbalances with a simple arithmetic mean. We set Δt for all stocks as the Δt that gives the closest variance in the change in the best ask price to 2. This way we can compare them since each stock has its own time scale in which prices move – some move more over smaller time intervals, for example. Our specific choice of how to fix Δt was to allow us to compare results between stocks, but Δt is a free parameter which can be set based on what time scales one wishes to examine the book.

Our model incorporates the imbalance into the distribution of the change in the best ask price by using the imbalance as the weight in a convex combination of two distributions, dp^+ and dp^-, representing the change in the best ask price if $I = 1$, and $I = -1$, respectively. If we know dp^+ we also know dp^- by the assumed symmetry constraints of our model. The other model parameter is the depth of book, K, which we define based on the support of the empirical price change distribution. So, after fixing Δt we have three model parameters dp^+, K, and w.

Using maximum a posteriori (maximum likelihood) estimation for the exponential (free) weight model we calibrated a model for each stock using data from different time periods. Investigating the model parameters which came out of the calibrations gave insights into relationships between Δt, K, the average spread, the average interarrival time of orders, and the probability of no movement in the best ask price. We were also able to visually identify outliers in the data which corresponded to American holidays or days related to financial events. We found a clear dependence of the model parameters

with the time period of the trading day which was consistent with the usual beliefs in trade activity over time. We used a large number of stocks over a small time period to investigate the relationship between price changes and the imbalance. However, we had to select a smaller number of stocks when investigating the model parameters over a longer time period due to the computational time needed to run over a hundred calibrations for each stock. In both cases the calibrations provide excellent fits to the data which we explored through the KL divergence and repeating our statistical tests from earlier in the work for both the exponential and free imbalance weights. We found many stocks with a significantly increased association when weight was given to the depth of book, but also stocks in which the classic imbalance definition was still optimal.

With an analysis of the model parameters completed we could apply our model to the problem of spoofing detection. We determined the optimal strategy a spoofer would take with $\Delta t = 5$ seconds given a limit order book and compare the pay off to immediately making a market order, or waiting to place a market order at the start of the next time period. We found that comparing the Sharpe ratio of the three strategies, rather than the expected value, we got consistent results for three clusters where the spoofer would determine their strategy. We can then use these values to determine what periods of the day spoofer was most optimal in which a regular could use to limit their search for manipulators.

We then tested the decision boundary between an immediate market order and spoofing the book as we changed the number of shares H the spoofer intends to purchase and the number of shares $\tilde{V}$ that the spoofer is willing to risk manipulating the book. As $\tilde{V}$ increases the spoofer is willing to manipulate the book as it is increasingly positively imbalanced while as H increases the spoofer is willing to manipulate even if their optimal spoofing strategy has a smaller impact on the imbalance – the spoofer is willing to take a bigger risk the more shares they need to buy. We noticed, however, that one stock had a different behaviour which was attributed to the optimal imbalance weights. There were two locations in which a spoofer could place limit orders to most impact the imbalance with the location closest to the touch carrying more weight. How the spoofer allocated their shares, for a given H, then depended on the number of shares $\tilde{V}$ they were willing to spoof with. The fewer the shares the spoofer has for manipulation the riskier they are willing to be with them. As they are willing to spoof with more shares they start to move them to the less risky position. The shape of the book played heavily into how the shares were allocated in order to minimize the impact of walking the book in case their spoofing limit orders were executed. We found that there was a global minimum in the cases were investigated for BMO stock which tells us a spoofer would not arbitrarily

spoof with an increasing number of shares – that there is a point where spoofing with more only hurts the spoofer.

Based on the success of using the Sharpe ratio for the optimal decision making process we also checked using the Sharpe ratio to determine the optimal spoofing strategy instead of the expected value of the cost. The Sharpe ratio provided a more conservative strategy when spoofing in that the spoofer was using less shares at the global minimum of the Sharpe ratio. Also, the Sharpe ratio more accurately reflected what one would expect a manipulator to do in a very positively or negatively imbalanced limit order book – we would not expect a spoofer to manipulate with many shares, if at all in these cases. The optimal strategies according to both objective functions could then be used by regulators to look for traders placing orders of similar sizes in similar locations during the vulnerable periods of the trading day for the limit order book on a given stock.

Overall, we completed an analysis and application of a one period model in which the shape of the limit order book determines the distribution of the change in the best ask price. The model itself could be used in other applications, but we focused on testing the sensitivity of the book to this type of manipulation through the volume imbalance ratio as well as determining the optimal strategy a spoofer would take as predicted by our model. This provides a starting point for further investigation into price manipulating behaviour by traders abusing the nature of high frequency trading. In the following section we discuss possible improvements to the model and directions of research to take based on results from this work.

6.2 Future Work

There are significant directions we can take to improve and expand this work, but first we will talk about what we can do with the existing model setup. To start, a further investigation into a proper way of fixing Δt for a given stock based on some external objective function. That is, we would like to find some optimal time interval over which a spoofer is most likely to operate on a given stock rather than simply fixing Δt to whatever we want and then investigating the expected costs over that time interval. From this we could analyze more illiquid stocks or stocks which would require longer time intervals. Based on the results of this work the longer the time interval the larger the support of the price change distribution which would require us to know more about the volumes deeper in the book. We only took the first 15 prices on each side of the book based on memory restrictions, but we may be required to store information significantly further into the book than that to investigate spoofing over larger time intervals. We

could also look into alternative models for weighting the instantaneous imbalances when aggregating them into an average imbalance over Δt. We only looked at taking a simple mean and using time weighting in this work, but there may be better ways to calculate the average imbalance.

We also saw in chapter 2 that the average imbalance does not affect the probability of an up or down movement equally. An average imbalance of -0.2 gave a higher probability of the best ask decreasing than an average imbalance of 0.2 gave for the best ask increasing. However, for simplicity, we built into our model that the change in the best ask was symmetric about a zero imbalance. Further study needs to be done into this behaviour and how best to model it. A starting point would be to introduce a new parameter into the price change distribution model which breaks this symmetry and can be calibrated with the other model parameters. Alternatively, we could drop the symmetry constraint in the distribution so that dp^- is not completely determined by dp^+ and becomes a new distribution we need to find. The downside to this is that we would be introducing significantly more parameters than a single parameter to break the symmetry.

In this work we looked at three different ways of calculating the instantaneous imbalance – using only the touch, exponentially decaying weights, and free weights. Ideally, one would find a distribution (or mixture of distributions) to model the imbalance weights with fewer parameters than the free weights use. Over fitting would be less of a risk if we can reduce the model parameters as much as possible. Also, if we can find an appropriate weight model we could perform cluster analysis based on the weight parameters for collections of stocks or even single stocks over time.

Along with new models for the imbalance weights we saw in chapter 5 that the shape of the book itself is important for determining the strategy costs as they also impact the optimal spoofing strategy. The importance is introduced through the impact of walking the book at the end of the time period which we would need to estimate at the start of the time interval. In our work we assumed the shape of the book remained constant which we found to work for and against the spoofer roughly equally, so if the process was repeated enough times the spoofer would 'win' as often as they 'lose' from the modeling assumption. However, if a spoofer does not manipulate the book regularly they would ideally want a way to model the volume of shares near the touch to prevent themselves from spoofing the book during a period where the costs associated with walking the book would destroy the profits they made from spoofing. If we modeled the volumes at the end of the time interval conditional on the volumes at the start of the time with a distribution then the cost of walking the book would be calculated with an expected value with respect to this distribution.

After discussing ways we can expand on the existing model we can discuss about how we can extend it beyond a simple one period model and introduce the bid side of the book into the problem. We only looked at the problem of spoofing the ask side of the book in order to lower the best ask and buy shares at a discount[1], but some traders may actually be spoofing both sides of the book as part of their strategy. In order for us to model this we would need a joint distribution of the change in the best ask and best bid prices dependent on the average imbalance. This joint distribution would also be constrained by the spread since the best bid cannot be greater than or equal to the best ask. A model of the full joint distribution would give us the ability to investigate any price manipulation strategy either or both sides of the limit order book.

To further give the model realism we would like to expand the implementation of the spoofer's decisions into a multi-period model rather than a single time step. This way the spoofer could update their strategy continuously over the entire trading day until the optimal times to place their market orders. If we want to avoid the problems associated with estimating the cost of walking the book with a delayed market order we could have the spoofer placing multiple 100 share market orders throughout the day as they spoof the book to totally avoid these extra costs. This could be implemented as a dynamic programming problem. Alternatively, one could model the price change joint distribution as a general mixture or hidden Markov model where the regime changes come from changes in the average imbalance. These would be significantly more complicated models, but should provide a more realistic approach to replicating a spoofer's behaviour so a regular would know what to look for.

The ultimate goal of this work was to highlight an application of our price change model to spoofing detection. Since we did not have access to a labeled data set of known examples of spoofing we were only able to do a type of sensitivity analysis or optimal control given the states of the order book. With data sets of a known spoofer operating during the day we would be able to actually develop techniques which should at least be able to detect them. We also only had access to the broker ID for each action taken on the limit order book. There are too many traders operating under any given broker ID to make the claim that a broker's behaviour may be suspicious when a collection of potentially manipulative orders were actually being placed by multiple traders. If we had the trader IDs we could at least look for times during the trading day a trader was acting in a manner according to our spoofing model which could be grounds for a regulator to flag them, but is not concrete evidence of illegal behaviour. A fully labeled data

[1]Spoofing the bid side to increase the best bid and sell shares at a higher was conceptually the same due to the symmetry of the book.

set of known spoofing examples would provide further insights into the model building procedure and allow us to backtest any results of the models.

Bibliography

[1] J. Hull, *Options, Futures and Other Derivatives*. Prentice Hall finance series, Pearson/Prentice Hall, 2009.

[2] L. Bachelier, "Théorie de la spéculation," *Annales Scientifiques de L'Ecole Normale Supérieure*, vol. 17, pp. 21–88, 1900. Reprinted in P. H. Cootner (ed), 1964, The Random Character of Stock Market Prices, Cambridge, Mass. MIT Press.

[3] F. Black and M. S. Scholes, "The pricing of options and corporate liabilities," *Journal of Political Economy*, vol. 81, pp. 637–654, May-June 1973.

[4] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, pp. 77–91, Mar. 1952.

[5] R. C. Merton, "Option pricing when underlying stock returns are discontinuous," *Journal of Financial Economics*, vol. 3, no. 1, pp. 125 – 144, 1976.

[6] F. Allen and D. Gale, "Stock-price manipulation," *The Review of Financial Studies*, vol. 5, no. 3, pp. 503–529, 1992.

[7] R. A. Jarrow, "Market manipulation, bubbles, corners, and short squeezes," *Journal of Financial and Quantitative Analysis*, vol. 27, no. 3, pp. 311–336, 1992.

[8] A. Alfonsi and A. Schied, "Optimal trade execution and absence of price manipulations in limit order book models," *SIAM Journal on Financial Mathematics*, vol. 1, no. 1, pp. 490–522, 2010.

[9] A. Alfonsi and J. I. Acevedo, "Optimal execution and price manipulations in time-varying limit order books," *Applied Mathematical Finance*, vol. 21, no. 3, pp. 201–237, 2014.

[10] F. Allen and G. Gorton, "Stock price manipulation, market microstructure and asymmetric information," *European Economic Review*, 1991.

[11] K. John and R. Narayanan, "Market manipulation and the role of insider trading regulations," *The Journal of Business*, vol. 70, no. 2, pp. 217–247, 1997.

[12] R. K. Aggarwal and G. Wu, "Stock market manipulation-theory and evidence," in *AFA 2004 San Diego Meetings*, 2003.

[13] J. Mei, G. Wu, and C. Zhou, "Behavior based manipulation: theory and prosecution evidence," *Available at SSRN 457880*, 2004.

[14] A. Chakraborty and B. Yılmaz, "Informed manipulation," *Journal of Economic Theory*, vol. 114, no. 1, pp. 132–152, 2004.

[15] G. Jiang, P. G. Mahoney, and J. Mei, "Market manipulation: A comprehensive study of stock pools," *Journal of Financial Economics*, vol. 77, no. 1, pp. 147–170, 2005.

[16] R. Böhme and T. Holz, "The effect of stock spam on financial markets," *Available at SSRN 897431*, 2006.

[17] L. Frieder and J. L. Zittrain, "Spam works: evidence from stock touts and corresponding market activity," *Berkman Center Research Publication*, no. 2006-11, 2007.

[18] M. Hanke and F. Hauser, "On the effects of stock spam e-mails," *Journal of Financial markets*, vol. 11, no. 1, pp. 57–83, 2008.

[19] R. A. Jarrow, "Derivative security markets, market manipulation, and option pricing theory," *Journal of Financial and Quantitative Analysis*, vol. 29, no. 2, pp. 241–261, 1994.

[20] F. Klöck, A. Schied, and Y. Sun, "Price manipulation in a market impact model with dark pool," *Applied Mathematical Finance*, vol. 24, no. 5, pp. 417–450, 2017.

[21] N. F. Johnson, M. Hart, P. M. Hui, and D. Zheng, "Trader dynamics in a model market," *International Journal of Theoretical and Applied Finance*, vol. 3, no. 03, pp. 443–450, 2000.

[22] W. A. Brock, C. H. Hommes, and F. O. Wagener, "Evolutionary dynamics in markets with many trader types," *Journal of Mathematical Economics*, vol. 41, no. 1-2, pp. 7–42, 2005.

[23] P. Casgrain and S. Jaimungal, "Mean-field games with differing beliefs for algorithmic trading," *arXiv preprint arXiv:1810.06101*, 2018.

[24] P. Casgrain and S. Jaimungal, "Mean field games with partial information for algorithmic trading," tech. rep., arXiv. org, 2019.

[25] C. C. Moallemi and K. Yuan, "A model for queue position valuation in a limit order book," *Columbia Business School Research Paper*, no. 17-70, 2016.

[26] I. Muni Toke and N. Yoshida, "Modelling intensities of order flows in a limit order book," *Quantitative Finance*, vol. 17, no. 5, pp. 683–701, 2017.

[27] F. Abergel and A. Jedidi, "A mathematical approach to order book modeling," *International Journal of Theoretical and Applied Finance*, vol. 16, no. 05, p. 1350025, 2013.

[28] R. Martins and D. Hendricks, "The statistical significance of multivariate hawkes processes fitted to limit order book data," *arXiv preprint arXiv:1604.01824*, 2016.

[29] R. Cont and M. S. Mueller, "A stochastic pde model for limit order book dynamics," *arXiv preprint arXiv:1904.03058*, 2019.

[30] A. Alfonsi, A. Fruth, and A. Schied, "Optimal execution strategies in limit order books with general shape functions," *Quantitative Finance*, vol. 10, no. 2, pp. 143–157, 2010.

[31] X. Guo, A. De Larrard, and Z. Ruan, "Optimal placement in a limit order book: an analytical approach," *Mathematics and Financial Economics*, vol. 11, no. 2, pp. 189–213, 2017.

[32] R. Malhotra, "'flash crash' course: what is 'layering?'," *CNBC*, 2015. [Online; accessed 2019-11-21].

[33] R. Pellecchia, "FINRA issues first cross-market report cards covering spoofing and layering," *FINRA*, 2016. [Online; accessed 2019-11-21].

[34] E. J. Lee, K. S. Eom, and K. S. Park, "Microstructure-based manipulation: Strategic behavior and performance of spoofing traders," *Journal of Financial Markets*, vol. 16, no. 2, pp. 227–252, 2013.

[35] Y. Wang, "Strategic spoofing order trading by different types of investors in the futures markets," *Wall Street Journal*, 2015.

[36] A. Harris and M. Leising, "High-speed trader accused of commodity market 'spoofing'," *Bloomberg*, 2014. [Online; accessed 2019-10-28].

[37] T. Polansek, "CFTC tells CME group to work more on 'spoofing' detection," *Reuters*, 2014. [Online; accessed 2019-10-28].

[38] P. Stafford, L. Whipp, and G. Meyer, "US trader found guilty in landmark 'spoofing' case," *CNBC*, 2015. [Online; accessed 2019-10-28].

[39] E. Rosenfeld, "UK trader charged for manipulation contributing to 2010 flash crash," *CNBC*, 2015. [Online; accessed 2019-10-28].

[40] N. Raymond, "Canadian accused by U.S. of high-speed trading scheme pleads guilty," *Reuters*, 2015. [Online; accessed 2019-10-28].

[41] J. Macfarland, "OSC reaches deal in alleged 'spoofing' case," *The Globe and Mail*, 2015. [Online; accessed 2019-10-28].

[42] B. Shecter, "Investment fund K2 accused of manipulative order 'spoofing' by OSC staff," *Financial Post*, 2018. [Online; accessed 2019-10-28].

[43] L. Moyer, "UBS, Deutsche Bank and HSBC to pay millions in spoofing settlement, CFTC says," *CNBC*, 2018. [Online; accessed 2019-10-28].

[44] G. Porter Jr., "Spoofing crackdown nets a guilty plea by ex-Bear Stearns trader," *BNN Bloomberg*, 2019. [Online; accessed 2019-10-28].

[45] K. Benner, "3 from JPMorgan accused in scheme to game precious metals market," *The New York Times*, 2019. [Online; accessed 2019-10-28].

[46] P. Henning, "The problem with prosecuting 'spoofing'," *The New York Times*, 2018. [Online; accessed 2019-11-08].

[47] Y. Cao, Y. Li, S. Coleman, A. Belatreche, and T. M. McGinnity, "Adaptive hidden markov model with anomaly states for price manipulation detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 2, pp. 318–330, 2014.

[48] Y. Cao, Y. Li, S. Coleman, A. Belatreche, and T. M. McGinnity, "Detecting price manipulation in the financial market," *2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, pp. 77–84, 2014.

[49] Á. Cartea, S. Jaimungal, and Y. Wang, "Spoofing and price manipulation in order driven markets," *Available at SSRN 3431139*, 2019.

[50] R. Cont, A. Kukanov, and S. Stoikov, "The price impact of order book events," *Journal of Financial Econometrics*, vol. 12, no. 1, pp. 47–88, 2014.

[51] K. Bechler and M. Ludkovski, "Order flows and limit order book resiliency on the meso-scale," *Market Microstructure and Liquidity*, vol. 3, no. 03n04, p. 1850006, 2017.

[52] K. Xu, M. Gould, and S. Howison, "Multi-level order-flow imbalance in a limit order book," *Available at SSRN 3479741*, 2019.

[53] M. Dixon, "Sequence classification of the limit order book using recurrent neural networks," *Journal of Computational Science*, vol. 24, pp. 277–286, 2018.

[54] P. Nousi, A. Tsantekidis, N. Passalis, A. Ntakaris, J. Kanniainen, A. Tefas, M. Gabbouj, and A. Iosifidis, "Machine learning for forecasting mid-price movements using limit order book data," *IEEE Access*, vol. 7, pp. 64722–64736, 2019.

[55] J. A. Sirignano, "Deep learning for limit order books," *Quantitative Finance*, vol. 19, no. 4, pp. 549–570, 2019.

[56] C. Cao, O. Hansch, and X. Wang, "The information content of an open limit-order book," *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, vol. 29, no. 1, pp. 16–41, 2009.

[57] J. Hellström and O. Simonsen, "Does the open limit order book reveal information about short-run stock price movements?," *Umeå Economic Studies*, no. 687, 2006.

[58] Á. Cartea, R. Donnelly, and S. Jaimungal, "Enhancing trading strategies with order book signals," *Applied Mathematical Finance*, vol. 25, no. 1, pp. 1–35, 2018.

[59] M. D. Gould and J. Bonart, "Queue imbalance as a one-tick-ahead price predictor in a limit order book," *Market Microstructure and Liquidity*, vol. 2, no. 02, p. 1650006, 2016.

[60] L. Kish, *Survey Sampling*. John Wiley & Sons, Inc., 1965.

[61] Á. Cartea, S. Jaimungal, and J. Penalva, *Algorithmic and High-Frequency Trading*. Cambridge University Press, 2015.

[62] H. Cramér, *Mathematical Methods of Statistics*, vol. 43. Princeton University Press, 1999.

[63] W. Bergsma, "A bias-correction for Cramér's v and Tschuprow's t," *Journal of the Korean Statistical Society*, vol. 42, no. 3, pp. 323–328, 2013.

[64] Á. Cartea, S. Jaimungal, and L. Sánchez-Betancourt, "Latency and liquidity risk," *Available at SSRN 3433739*, 2019.

[65] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[66] W. F. Sharpe, "Adjusting for risk in portfolio performance measurement," *The Journal of Portfolio Management*, vol. 1, no. 2, pp. 29–34, 1975.

Appendix A

Broker Behaviour

Individual traders go through brokers in order to place orders on the exchange. The trader ID's are hidden even in the level 2 data and are usually only available to regulatory bodies in the financial industry. However, we can see the aggregate activity of these traders through the broker ID's in the level 2 data. The actions of all active brokers during the market open on AEM stock on April 17, 2017 is shown in Figure A.1. The total number of booked orders by broker ID for the same day and stock ticker is shown in Figure A.2. The broker ID's shown in the figure have been changed from their true values in the dataset in order to keep their identities anonymous.

From Figure A.1 we can see that not all brokers share the same activity on AEM stock on this day. Similar plots can be made for other stocks on any given day of the year. For example, brokers 0 through 8 cancel almost all orders they book on AEM stock. This would suggest algorithmic traders use these brokers to place and quickly cancel limit orders throughout the day while making very few market orders. Brokers 9 through 18 also cancel many of their orders, while allowing some to be executed as well as placing market orders. Brokers 19 through 25 cancel very few orders and mostly book orders and trade. Brokers 9 through 25 are likely to be made up of mixes of algorithmic traders as well as real people placing orders on the exchange.

Figure A.2 gives us a better picture of which brokers are acting as market makers for AEM stock on April 17, 2017. Brokers 0, 2, 3, 4, 5, 6, and 8 are clearly contributed the most limit orders to the order book on this day with broker 4 dominating the rest. It is also interesting to note that these are the same brokers that also cancel almost all orders they place on the order book.

Figures A.3 and A.4 show the total number of shares booked and cancelled on the limit order book by broker ID. Here we can see that the brokers we highlighted from Figure A.2 also place large numbers of shares on the limit order book in addition to

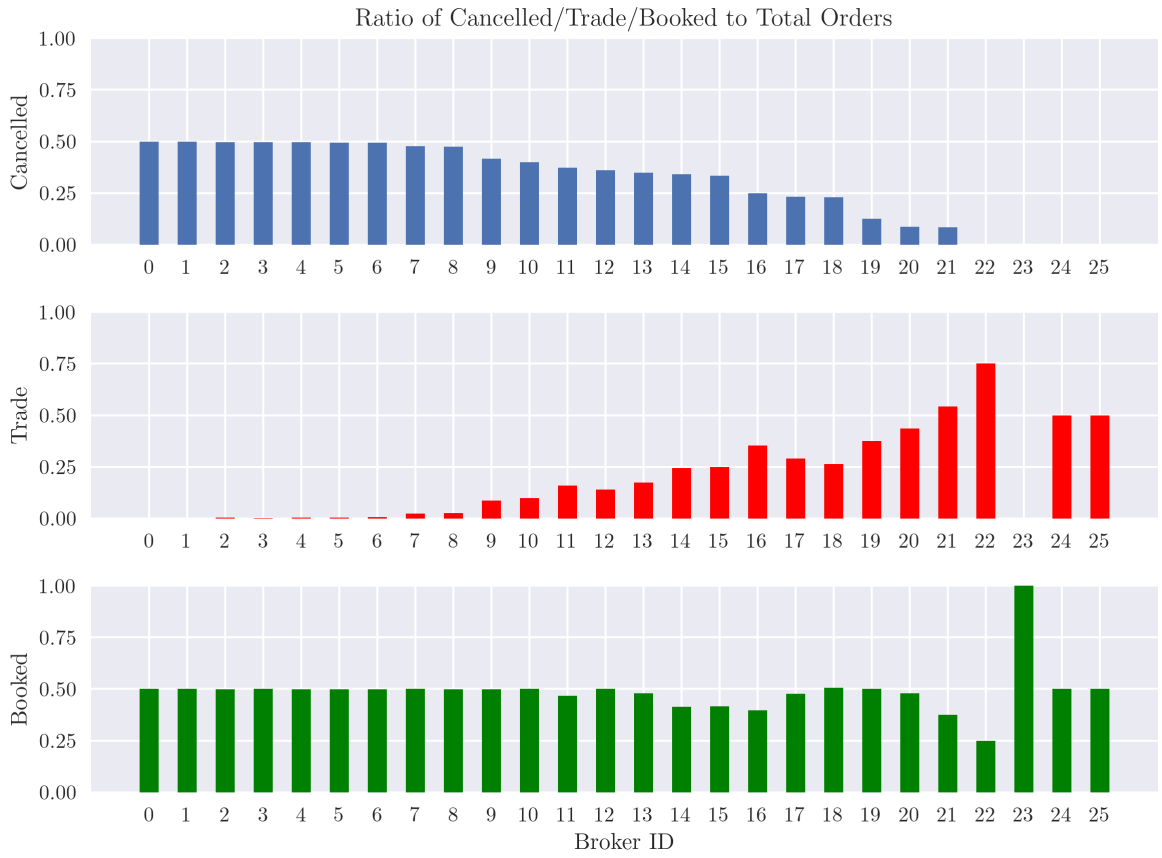


Figure A.1: Ratio of cancelled, trade, and booked orders to total orders for all active brokers on AEM stock on April 17, 2017. The broker ID's displayed above are not the true broker ID's in the dataset.

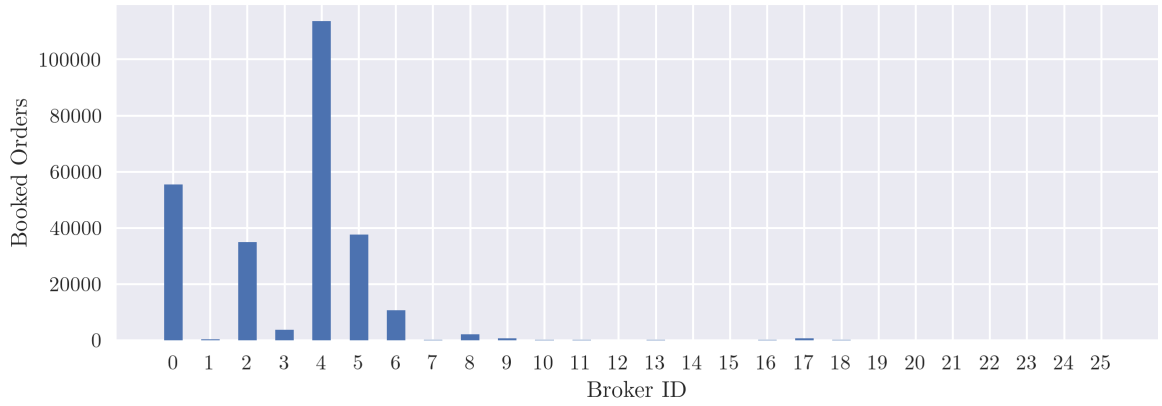


Figure A.2: Total booked orders for all active brokers on AEM stock on April 17, 2017. The broker ID's displayed above are not the true broker ID's in the dataset. Same broker ID's as Figure A.1

placing the most orders. All other brokers contribute relatively little shares to book. We see the same situations across other stock tickers where a small number of brokers are driving the dynamics of the limit order book through large and frequent limit order placements and cancellations.

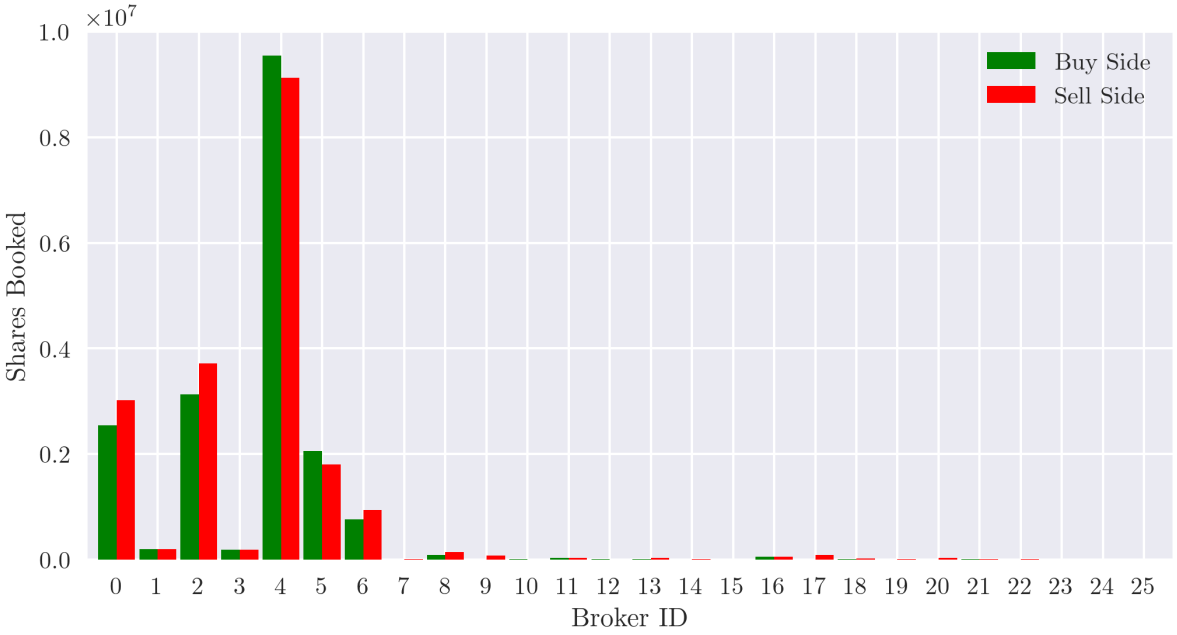


Figure A.3: Total booked shares of AEM stock on April 17, 2017. Green/red for shares booked on buy/sell side of limit order book. Same broker ID's as Figure A.1

Figure A.5 shows the total number of shares traded on the limit order book by broker ID. Even though brokers 0 through 6 make up most of the activity on the book there are still brokers that execute trades throughout the day either by market orders or allowing what few limit orders they have to be matched. The point here is that even though all brokers are initiating trades on the book there are only a few which provide the vast majority of the liquidity to the market.

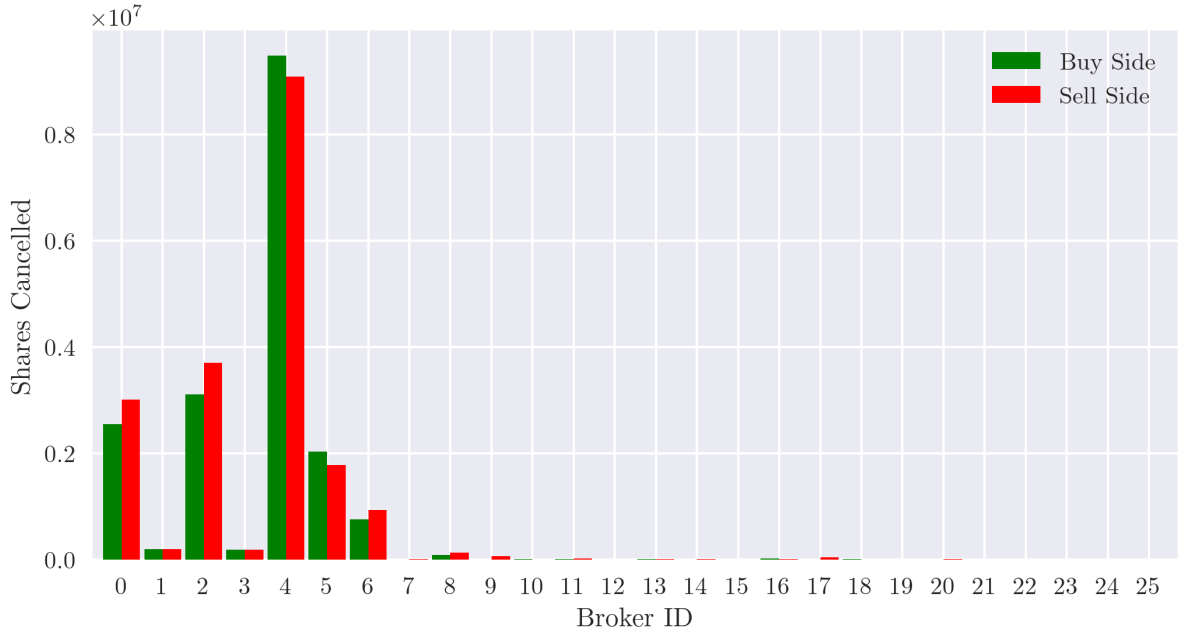


Figure A.4: Total cancelled shares of AEM stock on April 17, 2017. Green/red for shares cancelled on buy/sell side of limit order book. Same broker ID's as Figure A.1

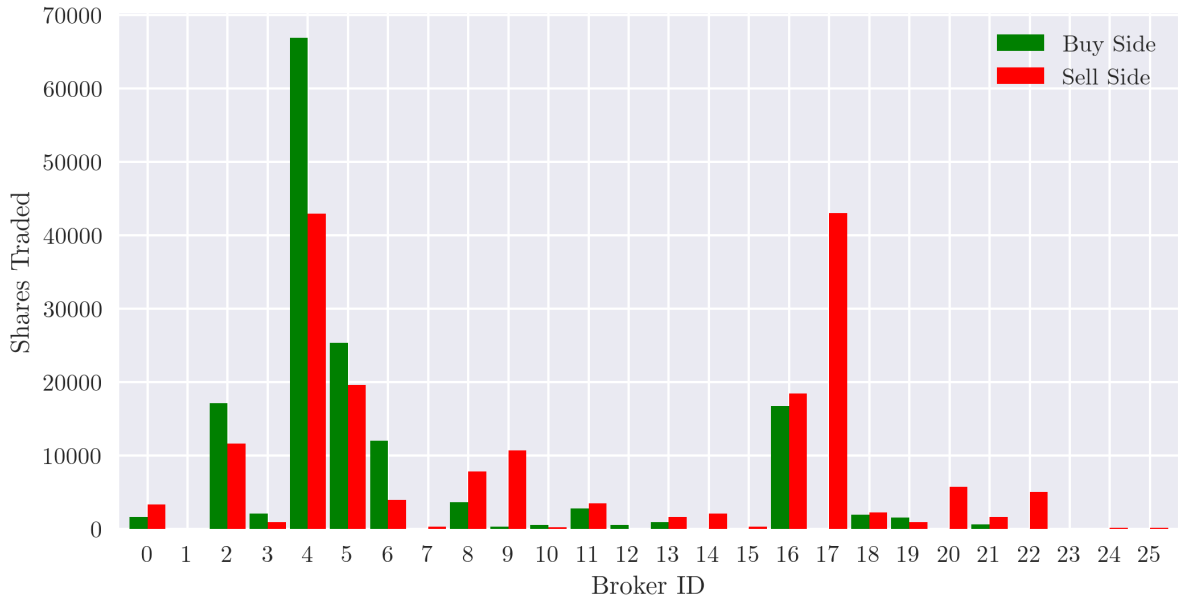


Figure A.5: Total traded shares of AEM stock on April 17, 2017. Green/red for shares traded on buy/sell side of limit order book. Same broker ID's as Figure A.1

Appendix B

Maximum a Posteriori Estimation

Say we want to estimate parameters of a distribution, but we want to incorporate our prior knowledge or belief in how the parameters of the model are distributed. For example, if we had a new baseball player to the major leagues and we wanted to determine the probability of this player hitting a home run given that he has only been up to bat once and struck a home run on his first swing. We have a single data point suggesting, naively, that he has a 100% probability of hitting a home run. However, we know this could not possibly be true since no one has ever had such a home run batting average. Given our limited data set for this new player - how can we incorporate our past knowledge of other player's home run batting averages to estimate the probability for this new player? A technique for doing this is the maximum a posteriori (MAP) estimation.

The maximum likelihood estimator (MLE) for data $X = (x_1, x_2, \ldots, x_n)$ with probability distribution $f(X|\theta)$ parameterized by θ is:

$$\theta^\star = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta|X) = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} f(X|\theta) \tag{B.0.1}$$

where $\mathcal{L}(\theta|X)$ is the likelihood function.

If instead we assume that θ itself is a random variable with prior distribution $g(\theta)$ then we can calculate the posterior distribution $f(\theta|X)$ using Bayes' theorem.

$$f(\theta|X) = \frac{f(X|\theta)g(\theta)}{\int_\Theta f(X|\phi)g(\phi)d\phi} \tag{B.0.2}$$

The maximum a posteriori estimation is then the mode of the posterior distribution $f(\theta|X)$ which gives:

$$\theta^\star = \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} \frac{f(x_i|\theta)g(\theta)}{\int_\Theta f(x_i|\phi)g(\phi)d\phi} \tag{B.0.3}$$

However,

$$\int_\Theta f(x_i|\phi)g(\phi)d\phi \tag{B.0.4}$$

is independent of θ and strictly positive so can be ignored without affecting the maximum of equation B.0.3. Substituting in the definition of $\mathcal{L}(\theta|X)$ we get

$$\theta^\star = \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} f(x_i|\theta)g(\theta) = \arg\max_{\theta\in\Theta} \mathcal{L}(\theta|X)g(\theta) \tag{B.0.5}$$

as the MAP estimation. We can think of this procedure as assuming some initial prior distribution $g(\theta)$ on θ which is updated by data through the likelihood function to assign new probabilities to the θ parameters by the posterior distribution. As we see more data the changes in our beliefs are reflected in the posterior distribution. Again, like MLE, one usually sees this written in the form of minimizing the negative log of the MAP estimator. So, the alternative MAP estimation is:

$$\theta^\star = \arg\min_{\theta\in\Theta} \left[-\left(\sum_{i=1}^{n} \log f(x_i|\theta) \right) - \log g(\theta) \right] \tag{B.0.6}$$

A few things to note are that if we take $g(\theta)$ as the uniform distribution over Θ then the MAP estimation is equivalent to MLE. Also, MAP estimation is the maximum mode of the posterior distribution unlike the MLE which was the maximum of the likelihood function. The MAP estimation is also not invariant under reparameterization since $g(\theta)$ is the prior for random variable θ, so if we map $\theta \to \hat{\theta}$ the prior distribution for $\hat{\theta}$ may not be $g(\hat{\theta})$. We would need to determine the new prior for $\hat{\theta}$ from the Jacobian of the reparameterization.

From equation B.0.6 we can also view the MAP estimator as applying a penalty to the MLE where we penalize values of $\theta \in \Theta$ we deem unlikely from past experience or belief.

Appendix C

Statistical Tests

C.1 Pearson's Chi-Squared Test

Pearson's chi-squared test is a statistical test for sets of categorical data to determine how likely the difference between the sets came from random chance. So named because Pearson's test statistic χ^2 asymptotically approaches the χ^2-distribution.

The test is a way for determining the significance level α in which we can reject the null hypothesis H_0 that the sets of categorical data are independent of each other. The alternative hypothesis H_1 is that the sets of categorical data are not independent. We only use two sets of categorical data in this thesis so we can restrict our case to two data sets: X_1 and X_2. This is usually written as:

$$H_0 : \text{'Variable } X_1 \text{ is independent of variable } X_2\text{'} \tag{C.1.1}$$

$$H_1 : \text{'Variable } X_1 \text{ is not independent of variable } X_2\text{'} \tag{C.1.2}$$

To calculate χ^2 we need to construct a $R \times C$ contingency table like Table 2.5 where R is the number of rows and C is the number of columns. For Table 2.5, $R = 2$ and $C = 4$. The test statistic χ^2 is then calculated by:

$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{C.1.3}$$

where

o_{ij} is the observed cell count in the i^{th} row and j^{th} column of the table (C.1.4)

e_{ij} is the expected cell count in the i^{th} row and j^{th} column of the table (C.1.5)

and

$$e_{ij} = \frac{\left(\sum_{\ell=1}^{R} o_{\ell j}\right)\left(\sum_{k=1}^{C} o_{ik}\right)}{n} \tag{C.1.6}$$

$$n = \sum_{i=1}^{R}\sum_{j=1}^{C} o_{ij} \tag{C.1.7}$$

The χ^2 statistic is then compared to the critical value χ^2_{crit} from the χ^2-distribution with degrees of freedom $dof = (R-1)(C-1)$ and chosen significance level α. If $\chi^2 > \chi^2_{\text{crit}}$, then we say we can reject the null hypothesis at the α significance level.

For example, using Table 2.5, we have o_{ij} and e_{ij} as:

$$o_{ij} = \begin{bmatrix} 657 & 58 & 1015 & 7259 \\ 4051 & 2253 & 1484 & 1527 \end{bmatrix} \tag{C.1.8}$$

$$e_{ij} = \begin{bmatrix} 1184.76 & 1021.88 & 37.71 & 2009.05 \\ 1143.30 & 986.12 & 35.42 & 1938.73 \end{bmatrix} \tag{C.1.9}$$

Then using Equation C.1.3, $\chi^2 = 8355.99$ with $dof = 3$. A significance level $\alpha = 0.999$ would give $\chi^2_{\text{crit}} = 16.266$ so we could easily reject the null hypothesis that the two sets are independent at the 99.9% significance level. A $\chi^2 = 8355.99$ would mean we could reject the null hypothesis at some significance level arbitrarily close to 1.

A p-value for the test can be computed as

$$\text{p-value} = \mathbb{P}\left[X \geq \chi^2 | H_0\right] \tag{C.1.10}$$

where X is distributed by the χ^2-distribution with 3 degrees of free and χ^2 is the test statistic we calculated above. That is, the p-value is the probability of observing a test statistic at least as extreme as χ^2 under the null hypothesis.

To give us an idea of how strong the association between the two data sets is we can calculate the Cramer's V statistic.

C.2 Cramer's V

The Cramer's V is a statistic which can be calculated from the χ^2 statistic that is independent of the sample size and bound between 0 and 1. A Cramer's V statistic of 1 would imply the two data sets are identical, 0 would imply they are totally independent. This gives a measure of association between the data sets used to calculate the test statistic. The advantage here is if we find our tests producing p-values arbitrarily close to zero then we can use the Cramer's V to differentiate between which data sets have the strongest association.

The Cramer's V, C_{V}, is defined as

$$C_{\mathrm{V}} = \sqrt{\frac{\varphi^2}{\min(R-1, C-1)}} \tag{C.2.1}$$

with the Phi coefficient defined as

$$\varphi^2 = \frac{\chi^2}{n} \tag{C.2.2}$$

We can introduce a bias correction to give a more conservative estimate of the association with a small modification to our definitions [63].

$$\tilde{\varphi}^2 = \max\left(0, \varphi^2 - \frac{(R-1)(C-1)}{n-1}\right) \tag{C.2.3}$$

$$\tilde{R} = R - \frac{(R-1)^2}{n-1} \tag{C.2.4}$$

$$\tilde{C} = C - \frac{(C-1)^2}{n-1} \tag{C.2.5}$$

$$\tilde{C}_{\mathrm{V}} = \sqrt{\frac{\tilde{\varphi}^2}{\min(\tilde{R}-1, \tilde{C}-1)}} \tag{C.2.6}$$

From the previous example, we had $\chi^2 = 8355.99$, $n = 18304$, $R = 2$, and $C = 4$. Plugging these into Equation C.2.6 we find the $\tilde{C}_{\mathrm{V}} = 0.676$. From our test we have clear statistical significance between our two data sets and their corresponding association is very strong.

Appendix D

Additional Plots

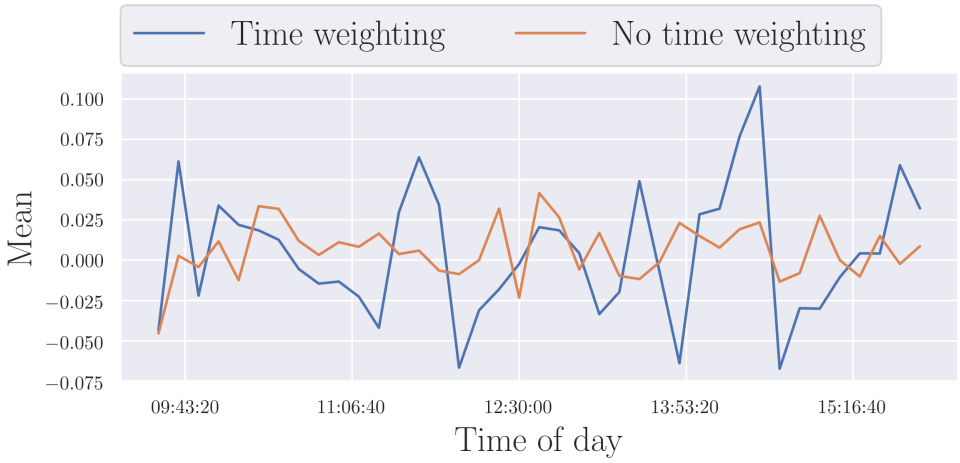


Figure D.1: Mean of time series in Figure 2.7 over 10 minute intervals after replacing the instantaneous imbalance with random draws from normal distribution with mean 0 and variance 1.

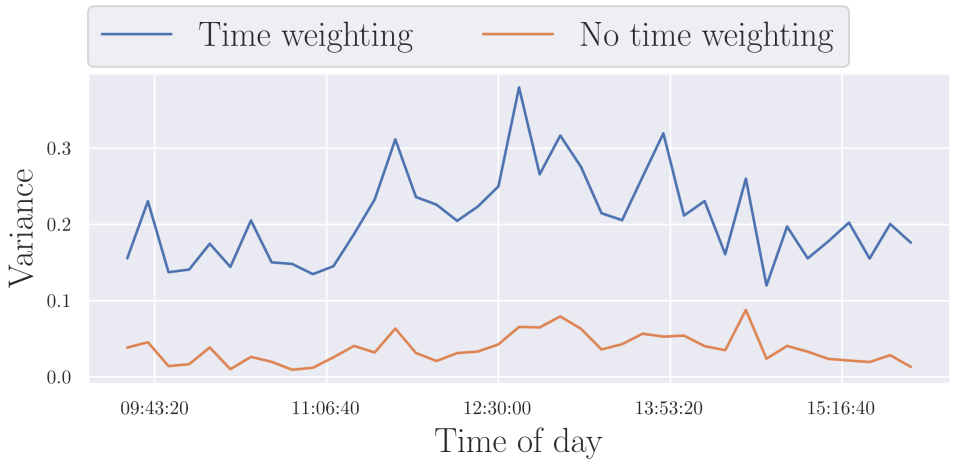


Figure D.2: Variance of time series in Figure 2.7 over 10 minute intervals after replacing the instantaneous imbalance with random draws from normal distribution with mean 0 and variance 1.

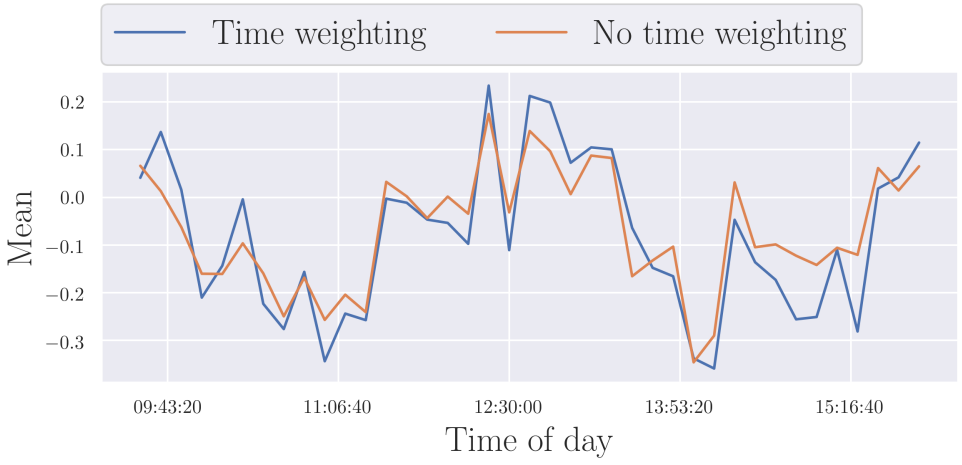


Figure D.3: Mean of average imbalance over 10 minute intervals. Mean varies slightly with both methods having periods being greater than the other. Data from ARX stock on April 17, 2017 for the entire trading day.

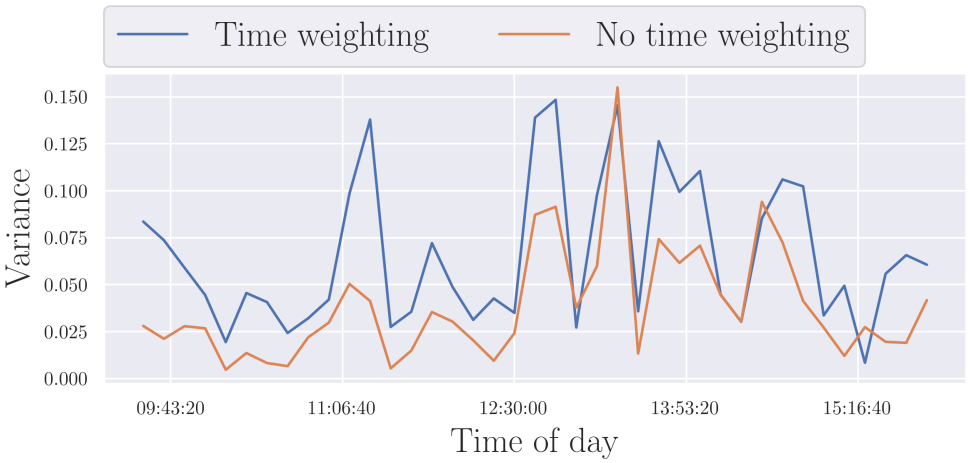


Figure D.4: Variance of average imbalance over 10 minute intervals. Variance is almost always greater with time weighting. Data from ARX stock on April 17, 2017 for the entire trading day.

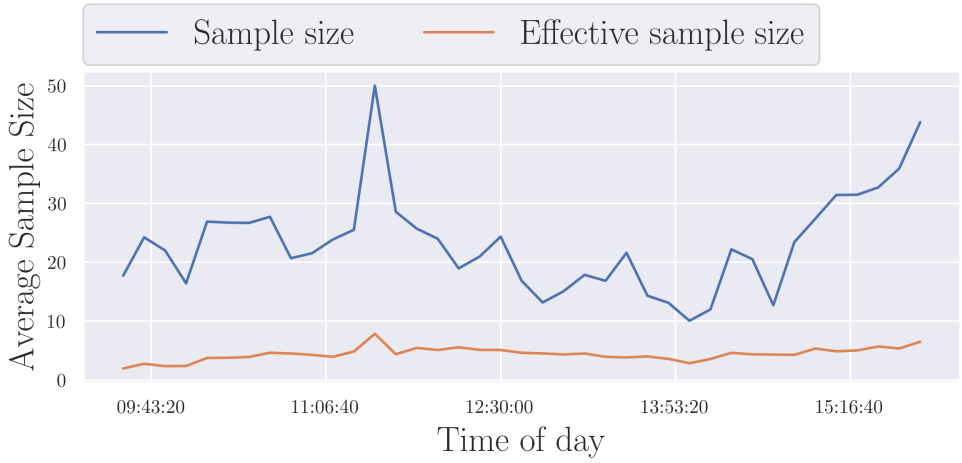
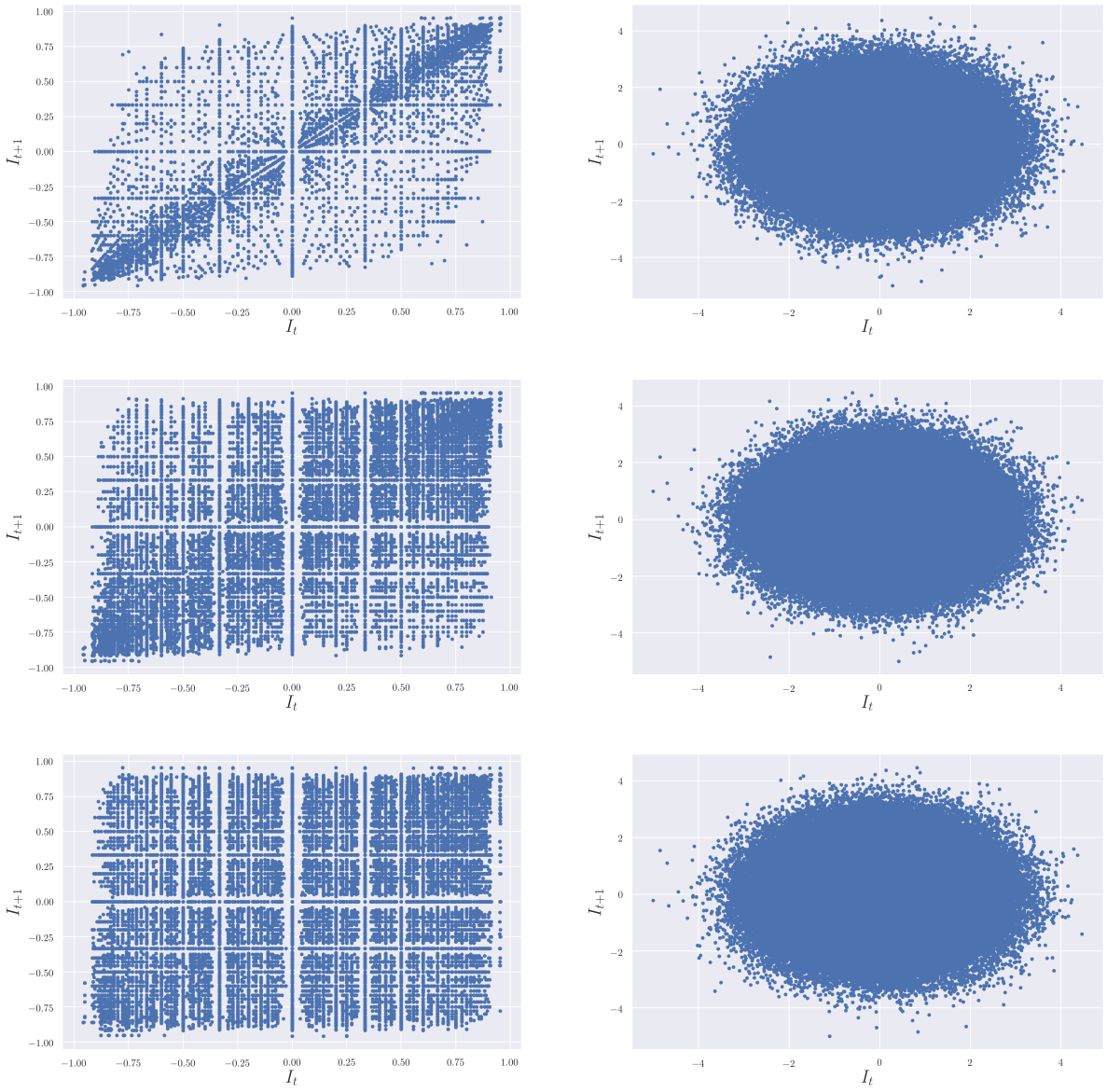


Figure D.5: Average sample size over 10 minute intervals. Effective sample size is also known as Kish's effective sample size. Data from ARX stock on April 17, 2017 for the entire trading day.



(a) Instantaneous Imbalance

(b) Normal Random Draw

Figure D.6: Future value of volume imbalance ratio I_{t+1}, t orders from current imbalance ratio I_t. The top, middle, and bottom subplots are for 1, 10, and 100 orders respectively. The left subplots are the actual instantaneous imbalance ratio and the right subplots are replacing I_t with random draws from $\mathcal{N}(0, 1)$. Data was taken from AEM stock on April 17, 2017 for the entire trading day.

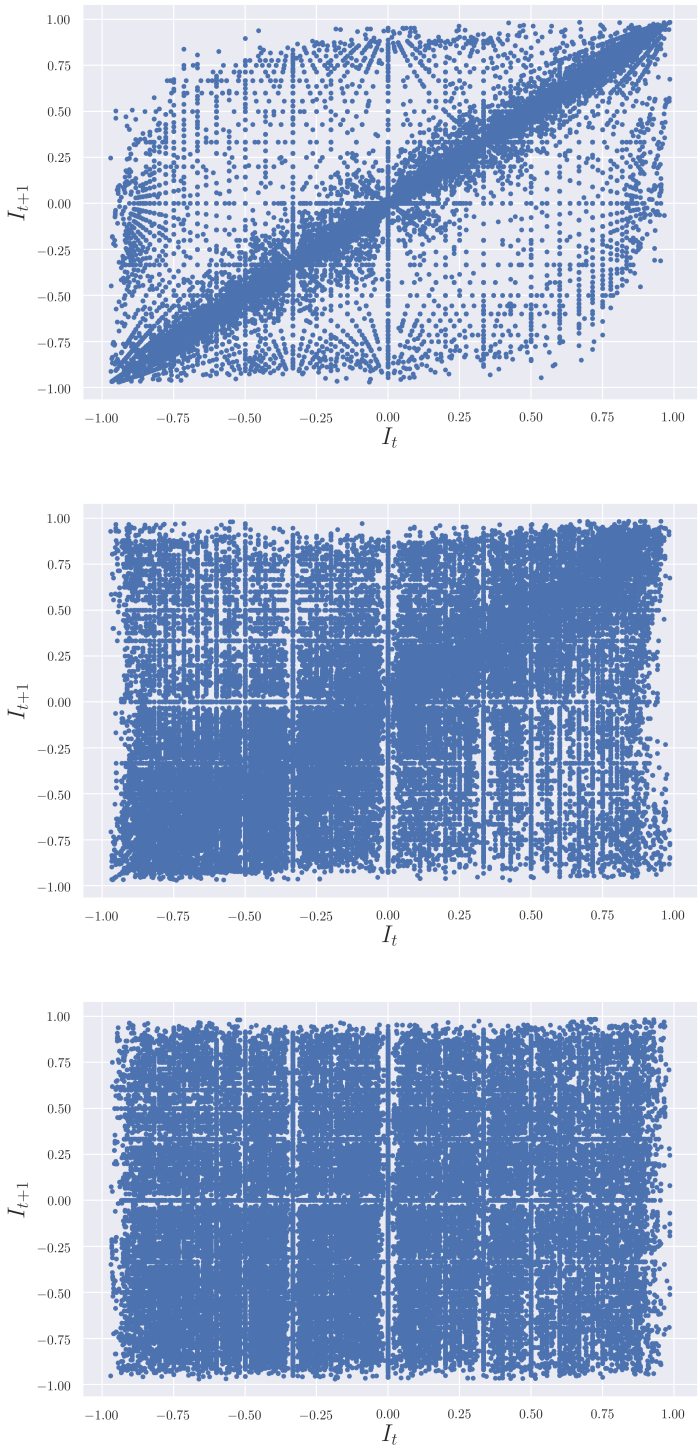


Figure D.7: Future value of volume imbalance ratio I_{t+1}, t orders from current imbalance ratio I_t. The top, middle, and bottom subplots are for 1, 10, and 100 orders respectively. Data was taken from ARX stock on April 17, 2017 for the entire trading day.

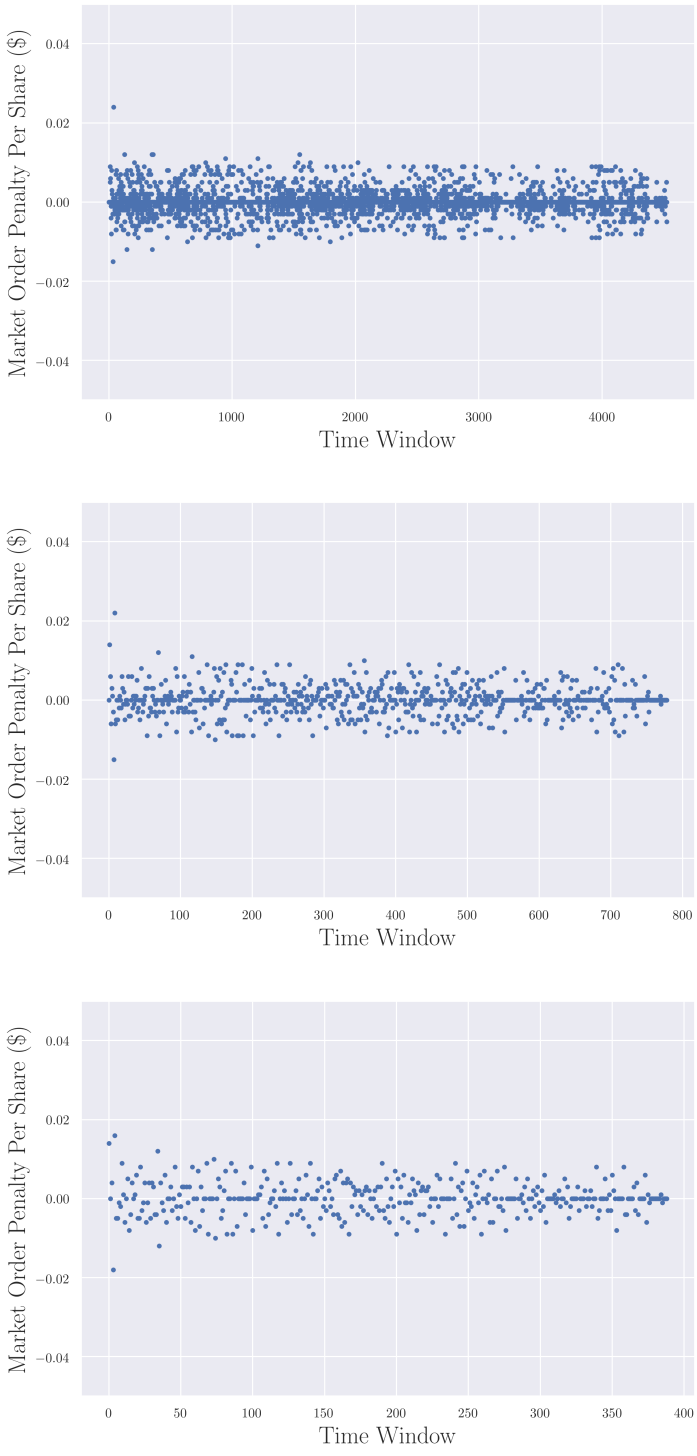


Figure D.8: Difference in $G(\vec{v}_t, H)/H$ between different time intervals and H throughout the trading day where $H = 1000$. The top, middle, and bottom subplots correspond to time intervals of 5, 30, and 60 seconds, respectively. Data taken from ARX stock on April 17, 2017 for the entire trading day. Stock price $\approx$ \$18 CAD.

Curriculum Vitae

Name:	Andrew Day
Post-Secondary Education and Degrees:	Western University London, ON 2013 - 2020 Ph.D. Applied Mathematics University of New Brunswick Fredericton, NB 2011 - 2013 M.Sc. Mathematics Memorial University of Newfoundland St. John's, NL 2006 - 2011 B.Sc. Applied Mathematics and Physics Joint Honours

Honours and Awards:

OGS, 2016 - 2017
NSERC PGS-D, 2013 - 2016

Related Work Experience:

MITACS Internship, TMX Group, 2018 Summer - Fall
AM 2270 Instructor, Western University, 2018
Graduate Teaching Assistant, Western University, 2013 - 2019
Graduate Teaching Assistant, University of New Brunswick, 2011 - 2013

Publications:

A. Buchel, & A. Day, "Universal relaxation in quark-gluon plasma at strong coupling" *Physical Review D* 92.2 (2015): 026009.
A. Day, & I.A. Brown, & S.S. Seahra, "Primordial fluctuations from deformed quantum

algebras" *Journal of Cosmology and Astroparticle Physics* 2014.03 (2014): 005.
J. Gegenberg, A. Day, H. Liu & S.S. Seahra, "An instability of hyperbolic space under the Yang-Mills flow" *Journal of Mathematical Physics* 55.4 (2014): 042501.