

Mathematical Problems in Algorithmic Trading and Financial Regulation



Yixuan Wang

Somerville College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2019

*To my parents,
Yanping Yang and Jianchun Wang.*

*In memory of my grandmother,
Yuehua Xu.*

Acknowledgements

First, I would like to thank my supervisor Prof. Álvaro Cartea for his guidance and encouragement throughout my DPhil. He is not only a great professor and an intelligent supervisor, but also a true friend and a life mentor. It has been a great honour to be his student.

I would like to thank Prof. Sebastian Jaimungal for advices and suggestions of improvements.

I would like to express my deepest gratitude to our crew: *Wei Fang, Dunhong Jin*, and *Zhenru Wang*. We have been more than just friends, but another family. I cannot go through this without your company. Although there were some tears dropped, the past four years have been one of the happiest periods in my life just because of you three. I will certainly miss every moment we spent together, every laugh we had together and every song we sang together. I have learned a lot from you guys and I am sincerely thankful for your love and support.

I would like to thank Vadim Kaushansky for discussions about study and life. I would like to thank Matthieu Mariapragassam, Andrei Cozma and Siyuan Li for help and suggestions as senior DPhil students. A special thanks goes to Yufei Zhang for insightful and helpful discussions. I would like to thank Kaichen Gu, Shiqi Chen and Jing Ren, who are pursuing PhDs in other universities, for sharing thoughts for the experience and encouragement. I would also like to thank Zhen Jiang, Qianqian Pu and Yue Wang for their distant friendships, support and encouragement through my hard times.

Finally, I would like to thank my parents, Yanping Yang and Jianchun Wang, for their consistent support and endless love. I would like to thank other members of my big family: Yongbin Cai, Liping Wang, Yueye Cai, Zhiqiang Yang, Huiya Lu, Mengyi Yang, Bingsheng Yang, Qin Zhang, Guohua Wang and Yuehua Xu, without whom this thesis would have never been written.

Statement of Originality

I hereby declare that this thesis contains no material which has been accepted or is currently being submitted for any other degree, diploma, certificate or other qualifications at the University of Oxford or elsewhere. To the best of my knowledge, this thesis contains no material previously published and precise reference is made when a previously published result is used or discussed.

This thesis includes three papers published. Chapter 2 is published in International Journal of Theoretical and Applied Finance (Cartea and Wang (2020)), Chapter 3 is published in Applied Mathematical Finance (Cartea et al. (2020)), Chapter 4 is published in Quantitative Finance (Cartea and Wang (2019)). Chapter 2 and 4 are co-authored with my supervisor Prof. Álvaro Cartea, Chapter 3 is co-authored with Prof. Sebastian Jaimungal and my supervisor Prof. Álvaro Cartea.

The ideas, development and writing up of the thesis, and the papers included in it, were the principal responsibility of myself, Yixuan Wang, working under the supervision of Prof. Álvaro Cartea.

Abstract

We study algorithmic trading strategies in order driven markets. We make three contributions to the literature. One, we show how a market maker employs information about the momentum in the price of the asset to design liquidity provision strategies. The momentum in the midprice of the asset depends on the arrival of liquidity taking orders and the arrival of news. Buy market orders (MOs) exert a short-lived upward pressure on the midprice and sell MOs exert a downward pressure of the price. We employ high-frequency data to estimate model parameters and show the performance of the market making strategy. Two, we model the trading strategy of an investor who spoofs the limit order book (LOB) to increase the revenue she obtains from selling a position in an asset. The strategy employs, in addition to sell limit orders (LOs) and sell market orders (MOs), a large number of spoof buy LOs to manipulate the volume imbalance of the LOB. Our results show that spoofing considerably increases the revenues from liquidating a position. The spoof strategy employs, on average, fewer sell MOs (than a strategy without spoof LOs) and from executing roundtrip trades that are initiated by buy spoof LOs that are inadvertently filled and subsequently unwound with sell LOs. Spoofing is illegal and difficult to detect. We show that as the financial penalty for spoofing increases, the spoof strategy relies less on spoof LOs. There is a critical point where the gains from spoofing are outweighed by the financial penalty, so it is optimal no not to spoof the LOB. Three, we show how the supply of liquidity in order driven markets is affected if LOs are forced to rest in the LOB for a minimum resting time (MRT) before they can be cancelled. The bid-ask spread increases as the MRT increases because market makers (MMs) increase the depth of their LOs to protect them from being picked off by other traders. The expected profits of the MMs increase when the MRT increases. The intuition is as follows. As the MRT increases, there are two opposing forces at work. (i) The longer the MRT, the more likely the

LOs are to be filled and, on average, shares are sold at a loss. (ii) because the depth of the posted LOs increases, the probability that the LO is picked off by other traders before the end of the MRT decreases. The net effect is that a longer MRT leads to a higher expected profit. We also show that the depth of LOs increases when the volatility of the price of the asset increases. Also, the depth of LOs increases when the arrival rate of market orders increases because it is less likely that LOs will be picked off by the end of the MRT. Finally, our model also makes predictions about the overall liquidity of the market. We show that MMs choose to supply the minimum amount of shares per LO allowed by the exchange because expected profits are maximised when liquidity provided is lowest.

Contents

1	Introduction	1
1.1	Background	1
1.2	Literature	2
1.2.1	Algorithmic trading	2
1.2.2	Effects on market quality	3
1.3	The algorithmic trading problem	4
1.3.1	Almgren-Chriss model	4
1.3.2	Cartea-Jaimungal model	9
1.4	Contributions and outline	13
2	Market making with momentum in prices	15
2.1	Introduction	15
2.2	Model	17
2.2.1	Parameter estimates	22
2.2.2	Optimal strategy	24
2.3	Simulation and performance of strategy	26
2.3.1	Sample paths	26
2.3.2	Distribution of orders traded	28
2.3.3	Tradeoff: mean and standard deviation of PnL	29
2.4	Simulation results with parameters estimated from Nasdaq data	31
2.5	Numerical scheme	32
2.6	Conclusions	35
2.7	Proofs	36
2.7.1	Proof of Theorem 2.2.1	36
2.7.2	Proof of Lemma 2.5.1	38
2.7.3	Proof of Proposition 2.5.2	39

2.7.4	Proof of Theorem 2.5.2	40
3	Spoofing and price manipulation in order driven markets	42
3.1	Introduction	42
3.2	Limit order book and volume imbalance	46
3.2.1	Volume imbalance	46
3.2.2	Market order activity	48
3.2.3	Limit order activity	49
3.2.4	Volumes at best prices and spoofing	49
3.3	The model	51
3.3.1	Liquidating shares without spoofing	52
3.3.2	Liquidating shares with spoof buy LOs	52
3.4	Investor's optimisation problem	54
3.4.1	Trading in lots of shares	58
3.4.2	Optimal trading strategy	59
3.5	Simulation and performance of strategy	60
3.5.1	PnL and price manipulation	61
3.5.2	Tradeoff: mean and standard deviation of PnL	64
3.6	Numerical scheme	66
3.7	Conclusions and Further Research	68
3.8	Proofs	70
3.8.1	Proof of Theorem 3.4.1	70
3.8.2	Comparison principle	72
3.8.3	Proof of Lemma 3.6.1	77
3.8.4	Proof of Proposition 3.6.2	78
3.8.5	Proof of Theorem 3.6.1	80
4	Market making with minimum resting times	83
4.1	Introduction	83
4.2	Literature review	85
4.3	Model I: limit orders of same volume	86
4.3.1	Performance criterion: expected profit	89
4.3.2	Value function	95
4.3.3	Numerical study and simulations	97

4.4	Model II: Limit orders of various volumes	100
4.4.1	Performance criterion: expected profit	102
4.4.2	Value function	105
4.4.3	Numerical study and simulations	107
4.5	Optimal depth	109
4.6	Expected profits and liquidity provision	111
4.7	Conclusions	112
4.8	Proofs	113
4.8.1	Proof of Proposition 4.3.2	113
4.8.2	Proof of Proposition 4.3.3	114
4.8.3	Proof of Theorem 4.3.1	116
4.8.4	Proof of Proposition 4.4.2	118
4.8.5	Proof of Proposition 4.4.3	118
4.8.6	Proof of Proposition 4.4.4	119
5	Conclusions	121
5.1	Summary of contributions	121
5.2	Directions for future research	122
5.2.1	Momentum in prices	122
5.2.2	Spoofing and price manipulation	122
5.2.3	Minimum resting times	123
A	Supplementary Background Material	124
A.1	Hamilton-Jacobi-Bellman Quasi-Variational Inequalities (HJBQVIs)	124
A.2	Viscosity solutions	125
	Bibliography	126

List of Figures

1.1	Optimal trajectories of the Almgren Chriss model for various values of λ . . .	8
1.2	Tradeoff between expected loss and variance of inventory as a function of the risk-aversion parameter λ . The parameter λ decreases from left to right. . .	9
1.3	Optimal trajectories of the Cartea-Jaimungal model with different ϕ 's. . .	12
2.1	Optimal strategy of the market maker. Dark blue: no LOs in the LOB and no MOs. Blue: post sell LO ($l^+ = 1$). Light blue: send sell MO (τ^-). Orange: post buy LO ($l^- = 1$). Green: send buy MO (τ^+). Yellow: post both sell and buy LO ($l^+ = l^- = 1$). Near maturity, there is a green sliver when $q = -1$, a dark blue sliver when $q = 0$, and a light blue sliver when $q = 1$	26
2.2	Surface of function \tilde{h} . The value of \tilde{h} is highest when the value of inventory is expected to appreciate and lowest when the value of inventory is expected to depreciate.	27
2.3	One simulation: sample path of midprice, alpha signal and inventory. . . .	28
2.4	Distributions of the number of filled LOs and executed MOs for three values of the inventory penalty parameter ϕ	29
2.5	Risk-reward profile with $\bar{Q} = 4$, $T = 60$ seconds. From left to right, $\phi = 10^{-3}, 5 \times 10^{-4}, 2 \times 10^{-4}, 10^{-6}$	30
2.6	Risk-reward profile with $\phi = 10^{-6}$, $T = 60$ seconds. From left to right, $\bar{Q} = 1, 2, 4, 5, 7$	30
3.1	Volume imbalance ρ_t of INTC from 10:00:00 to 10:02:00 on October 2 2017. . .	47
3.2	Number of MOs for various levels of volume imbalance ρ_t of INTC. Data: Nasdaq October 2017.	47

3.3	Optimal strategies for the investor. Parameter values as listed above and $\phi_q = 2 \times 10^{-5}, \phi_f = 4.5 \times 10^{-3}, p_{sell} = 0.3, p_{buy} = (0, 0.5, 0.6), V = 2$. Black: investor only posts sell LOs and does not spoof the book. White: investor spoofs the book and also posts sell LOs. Grey: investor executes sell MOs.	60
3.4	Heatmaps of the evolution of the midprice. Parameters: $\phi_q = 10^{-10}, \phi_f = 0, p_{sell} = 0.3, p_{buy} = (0, 0.5, 0.6), V = 2, \mathfrak{N} = 30$	64
3.5	Performance of strategy for ϕ_f in the range $[e^{-6}, e^{-2}]$. Other parameters: $\phi_q = 10^{-10}, p_{sell} = 0.3, p_{buy} = (0, 0.5, 0.6), V = 2, p_{sell} = 0.3$	65
4.1	Expected profit and 2,000 simulations (standard deviation of each simulation is also shown). MMs post LOs of same volume. Parameters: $\kappa = 10^2, M = 1, T = 0.5$ seconds, $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$ and $\lambda = 0.1/\text{second}$	98
4.2	Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}, \lambda = 0.1/\text{second}, T = 0.5$ seconds.	99
4.3	Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}, \lambda = 0.1/\text{second}, M = 1$	99
4.4	Expected profit. Parameters: $T = 0.5$ seconds, $\lambda = 0.1/\text{second}, M = 1$	99
4.5	Expected profit. Parameters: $T = 0.5$ seconds, $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}, M = 1$	99
4.6	Expected profit and 2,000 simulations (standard deviation of each simulation is also shown). The other MMs post LOs of various volumes. $M = 2, T = 0.5$ seconds, $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}, \lambda = 0.1/\text{second}$	108
4.7	Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}, \lambda = 0.1/\text{second}, T = 0.5$ seconds.	108
4.8	Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}, \lambda = 0.1/\text{second}, M = 2$	108
4.9	Expected profit. Parameters: $T = 0.5$ seconds, $\lambda = 0.1/\text{second}, M = 2$	109
4.10	Expected profit. Parameters: $T = 0.5$ seconds, $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}, M = 2$	109
4.11	Optimal delta. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}, \lambda = 0.1/\text{second}, T = 0.5$ seconds.	110
4.12	Optimal delta. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}, \lambda = 0.1/\text{second}, M = 3$	110
4.13	Optimal delta. Parameters: $T = 0.5$ seconds, $\lambda = 0.1/\text{second}, M = 3$	110
4.14	Optimal delta. Parameters: $T = 0.5$ seconds, $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}, M = 3$	110
4.15	Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}, \lambda = 0.1/\text{second}, T = 0.5$ seconds.	112
4.16	Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}, \lambda = 0.1/\text{second}, M = 3$	112

- 4.17 Expected profit. Parameters: $T = 0.5$ seconds, $\lambda = 0.1/\text{second}$, $M = 3$. . . 112
- 4.18 Expected profit. Parameters: $T = 0.5$ seconds, $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $M = 3$. . 112

Chapter 1

Introduction

1.1 Background

Exchanges play a major role in the efficient allocation of resources in capital markets and in the price discovery process by providing a platform where liquidity makers and takers interact. On the supply side, liquidity makers take into account public and private information when deciding how much liquidity to supply in the exchanges. To stay in business, liquidity makers are constantly monitoring and updating the prices and quantities they make to the market to ensure their quotes are up-to-date. In particular, liquidity makers consider a number of factors including: market conditions, arrival of news, market order flow, changes in the provision of liquidity, financial constraints, and inventory risk.

In electronic markets, traders send instructions to the exchange and a matching engine clears the market following a set of trading rules. The instructions are the output of computerised trading algorithms that process market information to make trading decisions and handle inventories.

Most modern equity exchanges are order-driven markets in which market makers (MMs) provide liquidity by posting limit orders (LOs) and takers consume liquidity by executing market orders (MOs). LOs quote prices and quantities that show intention to sell or buy shares and the exchange amalgamates them in the limit order book (LOB). LOs rest in the LOB until they are executed against an incoming market order (MO) or until they are cancelled. Nowadays, the vast majority of trades in financial markets are performed by trading algorithms. Trading algorithms are designed in the backdrop of

financial regulation, trading rules, and the architecture of electronic markets, all of which determine the overall market quality.

1.2 Literature

In this section we provide a brief summary of the literature on algorithmic trading and discuss some of the literature on high-frequency and algorithmic trading that focuses on market quality.

1.2.1 Algorithmic trading

An early paper on algorithmic trading is that of Almgren and Chriss (2001). The authors consider the problem of a trader who employs MOs to liquidate a large amount of shares within a fixed time horizon. The MOs of the trader can adversely affect the price of the shares. In their model setup, price impact is decomposed into temporary price impact and permanent price impact, both of which are assumed to be linear in the speed of trading. The authors propose several risk metrics including mean-variance and mean-VaR (value at risk). Section 1.3 introduces the Almgren-Chriss model in more detail.

There are many extensions to the work of Almgren and Chriss (2001). Obizhaeva and Wang (2013) propose transient price impact to model the resilience of the LOB, and Alfonsi et al. (2010) extend it to the case of a general price impact function. Almgren (2003) extends the price impact function to be nonlinear and Almgren (2012) considers stochastic price impact and stochastic volatility of the midprice. Cartea and Jaimungal (2016b) take into account the order flow from other market participants and Cartea and Jaimungal (2016a) look at the optimal execution of a large order where the agent aims to beat the volume weighted average price (VWAP) benchmark. All of the above employ liquidity taking orders in the execution programme. The work of Cartea and Jaimungal (2015a) considers the optimal execution problem with both limit and market orders while targeting the inventory schedule implied by the time-weighted-average-price (TWAP) or the optimal strategy in Almgren and Chriss (2001).

There is also abundant literature that focuses on optimal market making. One of the first papers to employ stochastic optimal control techniques to solve the problem of a market maker is Ho et al. (1981). The authors consider the problem of optimal quoting,

where the dealer decides the optimal depth of the bid and ask prices, relative to the ‘true’ price of the asset. The dealer assumes there is a linear-decay relation between the depth and the intensity of transactions. Avellaneda and Stoikov (2008) consider a similar problem where the exponential decay relation between the depth of LOs and the intensity of transactions. Both Ho et al. (1981) and Avellaneda and Stoikov (2008) assume the utility function of the market maker to be exponential. Guéant et al. (2013) solves the problem of a market maker that imposes bounds on the inventory holdings. Cartea et al. (2014) consider a more advanced model that takes into account the dependence between MOs, the LOB dynamics, and midprice moves, see also Cartea et al. (2018a). Cartea and Jaimungal (2015c) introduce several risk metrics so the market maker fine-tunes her strategies to trade off inventory risk, which also proxy for capital risk, against expected profits. In Cartea et al. (2017) the authors allow for model ambiguity in the framework of market maker to make the market making model robust to misspecification in the: arrival rate of MOs, fill probability of LOs, and midprice dynamics.

1.2.2 Effects on market quality

There are a number of papers that discuss how algorithmic and high-frequency trading affect the quality of markets. Cartea and Penalva (2012) provide a theoretical model to show that the volatility of prices and the price impact of liquidity trades increase in the presence of ultra-fast traders. Martinez and Rosu (2013) show that high-frequency traders increase volatility of the midprice and volume of trading, and also make markets more efficient. Hoffmann (2014) shows that in a market with slow and fast traders, being fast is valuable because it enables traders to avoid being picked off by slower traders. On the other hand, due to speed disadvantages, slow traders face a relative loss in bargaining power which leads to a reduction in trading and, consequently, a reduction in welfare. In a similar vein, Biais et al. (2015) show that ultra-fast traders can generate profits from trade or adverse selection because they have a relative speed advantage. However, this increase in speed increases adverse selection for all and incentivises other participants to become faster, which might lead to a socially sub-optimal over-investment in technology.

Several empirical papers examine the effect that algorithmic and ultra-fast trading have on market quality by looking at volatility, quoted spreads, effective spreads, and price discovery. An early study in this area, Hendershott et al. (2011), uses NYSE data from 2001

to 2005 to show that algorithmic trading reduces spreads, adverse selection, and trade-related price discovery and that these effects are stronger for large cap stocks. Cartea et al. (2019) employ data from NASDAQ and show that an increase in ultra-fast trading leads to lower liquidity: greater quoted and effective spreads and lower depth posted in the LOB — these effects are economically significant. Their results also hold in periods of unusually high ultra-fast trading (a proxy for quote stuffing) and periods where ultra-fast trading is primarily driven by fleeting orders inside the spread (a proxy for spoofing and competition between liquidity providers). In a different asset class, Chaboud et al. (2014) study the impact of algorithmic trading in the foreign exchange market. One of their key findings is that the presence of more algorithmic trading is associated with lower volatility of the fundamental value of exchange rates.

Finally, Boehmer et al. (2015) employ data from 39 exchanges (excluding US exchanges) for the period 2001 to 2009 to assess the effect of algorithmic trading, proxied by co-location facilities, on market quality. They find that for large (small) capitalization stocks an increase in algorithmic trading activity improves (worsens) liquidity and leads to faster price discovery. More algorithmic trading increases volatility for all stocks but with a larger effect on the volatility of small cap stocks.

1.3 The algorithmic trading problem

In this section we introduce two well-known mathematical frameworks in algorithmic trading. The first framework is Almgren-Chriss model from Almgren and Chriss (2001) and the second one is Cartea-Jaimungal model from Cartea et al. (2015). For both frameworks we introduce the model setup, state the mathematical problem and present the main results.

1.3.1 Almgren-Chriss model

Suppose we hold X units of a security, which we aim to liquidate before time T . We divide T into N intervals of length $\tau = T/N$, and define the discrete times $t_k = k\tau$, for $k = 0, \dots, N$. We define a trading trajectory to be a list x_0, \dots, x_N , where x_k is the number of units that we plan to hold at time t_k . Thus, our initial holding is $x_0 = X$, and full liquidation by time T requires that $x_N = 0$. We also define the corresponding trade

list n_1, \dots, n_N , where $n_k = x_{k-1} - x_k$ is the number of units that we sell between times t_{k-1} and t_k . The quantities x_k and n_k are related by

$$x_k = X - \sum_{j=1}^k n_j = \sum_{j=k+1}^N n_j, \quad k = 0, \dots, N. \quad (1.1)$$

Suppose that the initial price of the security is S_0 . The evolution of the price of the asset is determined by two exogenous factors: volatility and drift, and one endogenous factor: market impact. Volatility and drift are assumed to be the result of market forces that occur randomly and are independent of our trading. As market participants ‘begin to detect’ the volume we are selling (resp. buying) they adjust their bids (resp. offers) downward (resp. upward).

There are two types of market price impact: temporary price impact and permanent price impact. Temporary price impact refers to temporary imbalances in supply and demand caused by our trading leading to temporary price movements away from equilibrium. Permanent price impact means changes in the ‘equilibrium’ price due to our trading, which remain at least for the life of the liquidation strategy.

Permanent price impact. Assume that the price of the asset evolves according to the discrete arithmetic random walk

$$S_k = S_{k-1} + \sigma \tau^{1/2} \xi_k - \tau g\left(\frac{n_k}{\tau}\right), \quad (1.2)$$

for $k = 1, \dots, N$. Here, σ is the volatility of the asset, the random variables ξ_j are i.i.d. with mean zero and unit variance, and $g(v)$ is a function of the average speed of trading $v = n_k/\tau$ during the interval t_{k-1} to t_k .

Temporary price impact. When the trader sells n_k units of security between t_{k-1} and t_k , the execution price may decrease between t_{k-1} and t_k due to exhausting the supply of liquidity. We assume that this effect is short-lived and in particular, liquidity returns after each period and a new equilibrium price is established.

We model this effect by introducing a temporary price impact function $h(v)$, which represents the temporary drop in average price per share caused by trading at average speed v during one time interval. Thus, the price per share received on sale k is

$$\tilde{S}_k = S_{k-1} - h\left(\frac{n_k}{\tau}\right), \quad (1.3)$$

and the effect of temporary price impact $h(v)$ does not appear in the next ‘market’ price S_k .

We define the capture of a trajectory to be the full trading revenue upon completion of all trades:

$$\sum_{k=0}^N n_k \tilde{S}_k = X S_0 + \sum_{k=1}^N \left(\sigma \tau^{1/2} \xi_k - \tau g\left(\frac{n_k}{\tau}\right) \right) x_k - \sum_{k=1}^N n_k h\left(\frac{n_k}{\tau}\right), \quad (1.4)$$

and the shortfall to be $X S_0 - \sum_{k=1}^N n_k \tilde{S}_k$, which is the difference between the initial market value and the capture. We write $E(x)$ for the expected shortfall and $V(x)$ for the variance of the shortfall. Given the price dynamics,

$$E(x) = \sum_{k=1}^N \tau x_k g\left(\frac{n_k}{\tau}\right) + \sum_{k=1}^N n_k h\left(\frac{n_k}{\tau}\right), \quad (1.5)$$

and

$$V(x) = \sigma^2 \sum_{k=1}^N \tau x_k^2. \quad (1.6)$$

Note that (1.6) only holds if each x_k is non-random. The exact forms of the impact functions g and h may be chosen to reflect various properties of price impact. For computational purposes, we assume the permanent and temporary impact functions to be linear in the speed of trading. In particular, we have

$$g(v) = \gamma v, \quad (1.7)$$

and

$$h\left(\frac{n_k}{\tau}\right) = \epsilon \operatorname{sgn}\left(\frac{n_k}{\tau}\right) + \eta \frac{n_k}{\tau}, \quad (1.8)$$

where sgn is the sign function, γ , η and ϵ are positive constants. The parameter ϵ is interpreted as the fixed costs of selling. The expected shortfall becomes

$$E(x) = \frac{1}{2} \gamma X^2 + \epsilon \sum_{k=1}^N |n_k| + \frac{\tilde{\eta}}{\tau} \sum_{k=1}^N n_k^2, \quad (1.9)$$

where $\tilde{\eta} = \eta - \gamma \tau/2$. We observe that E is a strictly convex function if $\tilde{\eta} > 0$. The trader seeks the optimal trajectory by solving the unconstrained problem

$$\min_x \left(E(x) + \lambda V(x) \right), \quad (1.10)$$

where the risk parameter $\lambda > 0$ is the Lagrange multiplier and represents how much we penalise variance relative to expected cost. The trader optimizes over the set of admissible strategies consisting of the strategies with initial holding $x_0 = X$ and terminal holding $x_N = 0$.

The objective function $E + \lambda V$ is strictly convex when $\lambda > 0$, and in this case (1.10) has a unique solution. The optimal solution is a trading trajectory of the form:

$$x_j = \frac{\sinh(\kappa(T - t_j))}{\sinh(\kappa T)} X, \quad j = 0, \dots, N, \quad (1.11)$$

and the associated trade list

$$n_j = \frac{2 \sinh(\frac{1}{2} \kappa \tau)}{\sinh(\kappa T)} \cosh\left(\kappa \left(T - t_{j-\frac{1}{2}}\right)\right) X, \quad j = 0, \dots, N, \quad (1.12)$$

where $t_{j-\frac{1}{2}} = (j - \frac{1}{2}) \tau$, and κ satisfies

$$\frac{2}{\tau^2} \left(\cosh(\kappa \tau) - 1 \right) = \tilde{\kappa}^2, \quad \text{with} \quad \tilde{\kappa}^2 = \frac{\lambda \sigma^2}{\tilde{\eta}} = \frac{\lambda \sigma^2}{\eta \left(1 - \frac{\gamma \tau}{2\eta}\right)}.$$

We use parameters reported in Table 1.1 to provide a numerical example. Figure 1.1 shows optimal trajectories for three values of the risk parameter λ . The blue line is the optimal trajectory for $\lambda = 2 \times 10^{-6}$ in which case the trader is risk averse and wants to liquidate the position in the asset quickly to reduce volatility risk from the midprice, despite the extra transaction costs that arise from selling part of the inventory too quickly at the beginning. The purple line is the optimal strategy when $\lambda = 0$. The trader minimizes the expected shortfall without considering the variance of the shortfall. We call this the naïve strategy. We observe that in these examples when there is no drift in the security traded and the transaction costs are linear in the trading speed, the optimal trajectory is linear in time. Finally, the red line corresponds to $\lambda = -2 \times 10^{-7}$ and the trader is risk-seeker. The trader postpones selling at the beginning and causes higher variance $V(x)$ by liquidating part of the inventory very slowly at the beginning of the trading horizon.

S_0	50 \$/share
X	10^6 shares
T	5 days
N	5
σ	0.95 (\$/share)/day
ϵ	0.0625 \$/share
γ	2.5×10^{-7} \$/share ²
η	2.5×10^{-6} (\$/share)/(share/day)

Table 1.1: Parameters for the Almgren-Chriss model.

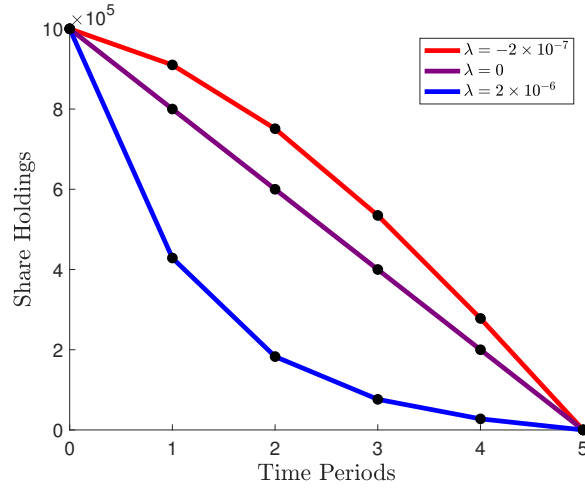


Figure 1.1: Optimal trajectories of the Almgren Chriss model for various values of λ .

The expectation shortfall and variance of the shortfall, given the optimal strategy, are

$$E(X) = \frac{1}{2} \gamma X^2 + \epsilon X + \tilde{\eta} X^2 \frac{\tanh(\frac{1}{2} \kappa \tau) \left(\tau \sinh(2 \kappa T) + 2 T \sinh(\kappa \tau) \right)}{2 \tau^2 \sinh^2(\kappa T)} \quad (1.13)$$

where we recall that $\tilde{\eta} = \eta - \gamma \tau / 2$, and

$$V(X) = \frac{1}{2} \sigma^2 X^2 \frac{\tau \sinh(\kappa T) \cosh(\kappa (T - t)) - T \sinh(\kappa \tau)}{\sinh^2(\kappa T) \sinh(\kappa \tau)}. \quad (1.14)$$

Figure 1.2 shows the tradeoff between expected loss and variance of inventory as a function of the risk-aversion parameter λ — other parameters are as in Table 1.1. The blue solid line is when $\lambda > 0$, the red dash line is when $\lambda < 0$, and λ decreases from left to right. As λ decreases, the variance of the shortfall increases and the expected shortfall first decreases and then increases. Near $\lambda = 0$, the trader can significantly decrease the variance of the shortfall without increasing the expected shortfall much.

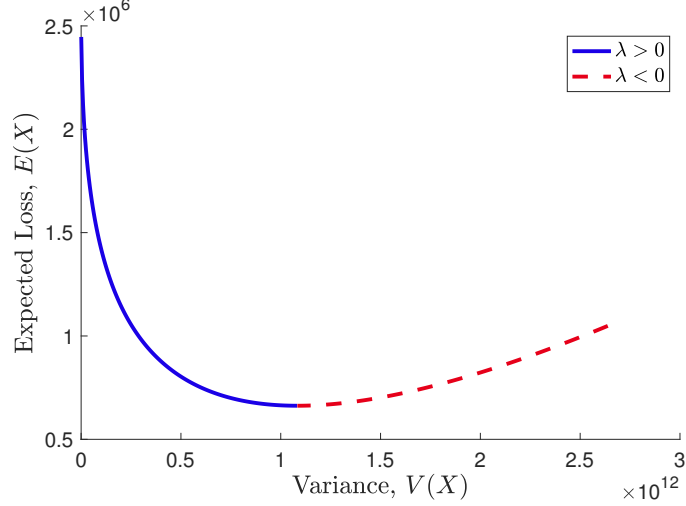


Figure 1.2: Tradeoff between expected loss and variance of inventory as a function of the risk-aversion parameter λ . The parameter λ decreases from left to right.

1.3.2 Cartea-Jaimungal model

The trader wants to liquidate \mathfrak{N} shares between time $t = 0$ and $t = T$, where $T > 0$ is constant. The trader uses MOs to liquidate the shares continuously and we denote by $\nu = (\nu_t)_{(0 \leq t \leq T)}$ the trading rate, i.e., the speed of trading.

Denote by $S^\nu = (S_t^\nu)_{(0 \leq t \leq T)}$ the midprice process of the asset, which satisfies

$$dS_t^\nu = -g(\nu_t) dt + \sigma dW_t, \quad (1.15)$$

where $W = (W_t)_{(0 \leq t \leq T)}$ is a standard Brownian motion, σ is a positive constant that represents the volatility of the midprice, and $g : \mathbb{R} \rightarrow \mathbb{R}$ denotes the permanent price impact that the trader's trading action has on the midprice. When the trader sells the shares using MOs, the execution price is less than the quoted midprice because the MOs walk the limit order book (and because the quoted spread in the LOB is generally positive). Denote by $\tilde{S}^\nu = (\tilde{S}_t^\nu)_{(0 \leq t \leq T)}$ the execution price, which satisfies

$$\tilde{S}_t^\nu = S_t^\nu - f(\nu_t), \quad (1.16)$$

where $f : \mathbb{R} \rightarrow \mathbb{R}$ denotes the temporary price impact.

Denote by $Q^\nu = (Q_t^\nu)_{(0 \leq t \leq T)}$ the trader's inventory, thus

$$dQ_t^\nu = -\nu_t dt, \quad Q_0^\nu = \mathfrak{N}. \quad (1.17)$$

The trader's cash is denoted by $X^\nu = (X_t^\nu)_{(0 \leq t \leq T)}$ the trader's cash and it satisfies

$$dX_t^\nu = \tilde{S}_t^\nu \nu_t dt = (S_t^\nu - f(\nu_t)) \nu_t dt, \quad X_0^\nu = 0. \quad (1.18)$$

Let τ represent the time when the trader completes the liquidation programme before T , thus

$$\tau = \inf \{t \geq 0 \mid Q_t^\nu \leq 0\} \wedge T. \quad (1.19)$$

The trader's the value function is

$$H(t, x, S, q) = \sup_{\nu \in \mathcal{A}} \mathbb{E}_{t,x,S,q} \left[X_\tau^\nu + Q_\tau^\nu (S_\tau^\nu - \alpha Q_\tau^\nu) - \phi \int_t^\tau (Q_u^\nu)^2 du \right], \quad (1.20)$$

where $\mathbb{E}_{t,x,S,q}[\cdot]$ denotes the expectation operator conditioned on $X_{t-}^\nu = x$, $S_{t-}^\nu = S$, $Q_{t-}^\nu = q$, and \mathcal{A} denotes the set of admissible strategies, consisting of \mathcal{F}_t -predictable processes such that $\mathbb{E} \left[\int_0^T |\nu_t| dt \right] < \infty$ \mathbb{P} -a.s..

We interpret each term inside the expectation operator in (1.20). The first term, X_τ^ν , is the final amount of cash that the trader receives from liquidating shares in the period $t \in [0, T]$. The second term, $Q_\tau^\nu (S_\tau^\nu - \alpha Q_\tau^\nu)$, is the revenue from liquidating any remaining inventory at τ using a MO, where the positive constant α is the terminal liquidation penalty parameter. Here we assume the LOB is flat, so the average price per share decreases linearly in the amount of shares remaining at terminal time, Q_τ^ν . In general we can replace αQ_τ^ν with any positive non-decreasing function of the remaining inventory Q_τ^ν to reflect the shape of the LOB, perhaps at the expense of mathematical tractability.

The third term, $\phi \int_t^\tau (Q_u^\nu)^2 du$, is the running penalty on the inventory throughout the trading horizon, where the positive constant ϕ is the running inventory penalty parameter. This penalty term does not affect the investor's revenues but affects the optimal trading speed. When the value of ϕ is large, the penalisation on the inventory level is high so the trader increases the trading rate to liquidate faster.

Here the trader gets penalised for any inventory level that is different from zero along the entire path of the trading time, to reflect the aim of liquidation. One can extend the inventory penalty by replacing $\phi \int_t^\tau (Q_u^\nu)^2 du$ with $\phi \int_t^\tau (Q_u^\nu - \mathbf{q}_u)^2 du$, where $\mathbf{q} = (\mathbf{q}_t)_{0 \leq t \leq T}$ is the investor's target schedule. Instead of penalising any inventory level different from zero,

it penalises any deviations from the investor's target schedule \mathbf{q}_t . See Cartea and Jaimungal (2015a) for the strategy of time-weighted average price (TWAP) and the Almgren-Chriss strategy as target schedule, and Cartea and Jaimungal (2016a) for a percentage of cumulative volume as target schedule.

The third term $\phi \int_t^\tau (Q_u^\nu)^2 du$ can also be motivated by the quadratic variation of the trader's mark-to-market trading value, which is

$$\left\langle \int_0^\cdot Q_u^\nu dS_u^\nu \right\rangle_\tau - \left\langle \int_0^\cdot Q_u^\nu dS_u^\nu \right\rangle_t = \sigma^2 \int_t^\tau (Q_u^\nu)^2 du,$$

where $\langle \cdot \rangle_t$ denotes the quadratic variation operator, see Cartea et al. (2018b). Recall that the Almgren-Chriss model penalises the variance of the terminal wealth. Such a penalty fails to capture the trader's inventory risk throughout the trading horizon. Instead, we use the quadratic variation of the trading value, which is essentially the 'variance (of the diffusion term) per second' of the trader's trading value. Another interpretation of the term $\phi \int_t^\tau (Q_u^\nu)^2 du$ is that the quadratic penalisation on running inventory is equivalent to the ambiguity aversion to the midprice drift, see Cartea et al. (2017).

To solve the optimal control problem (1.20), a dynamic programming principle holds and the value function $H(t, x, S, q)$ satisfies the Hamilton-Jacobi-Bellman (HJB) equation (see, e.g., Pham (2009))

$$\partial_t H + \frac{1}{2} \sigma^2 \partial_{SS} H - \phi q^2 + \sup_\nu \left\{ \nu \left(S - f(\nu) \right) \partial_x H - g(\nu) \partial_S H - \nu \partial_q H \right\} = 0, \quad (1.21)$$

with the terminal condition $H(T, x, S, q) = x + qS - \alpha q^2$.

For computational purposes, we use the simplifying assumptions that the permanent and temporary price impact functions are linear in the trading rate, i.e.,

$$f(\nu) = k\nu \quad \text{and} \quad g(\nu) = b\nu$$

for finite constants $k > 0$ and $b \geq 0$. The form of the performance criterion, together with the linear assumptions of the price impact functions, allows us to obtain a closed-form solution for the trading rate. That is,

$$\nu_t^* = \gamma \frac{\zeta e^{\gamma(T-t)} + e^{\gamma(T-t)}}{\zeta e^{\gamma T} - e^{\gamma T}} \mathfrak{N}, \quad (1.22)$$

and the optimal inventory level is

$$Q_t^{\nu^*} = \frac{\zeta e^{\gamma(T-t)} - e^{\gamma(T-t)}}{\zeta e^{\gamma T} - e^{\gamma T}} \mathfrak{N}, \quad (1.23)$$

where

$$\gamma = \sqrt{\frac{\phi}{k}} \quad \text{and} \quad \zeta = \frac{\alpha - \frac{1}{2}b + \sqrt{k\phi}}{\alpha - \frac{1}{2}b - \sqrt{k\phi}}. \quad (1.24)$$

We use the parameters in Table 1.2 to provide a numerical example. Figure 1.3 shows optimal trajectories for three values of the penalty parameter ϕ . As the value of ϕ increases, the optimal trajectory becomes more convex and the trader chooses to liquidate more shares at the beginning of the trading horizon. For larger values of ϕ the penalisation the running inventory penalty is higher, which incentivises the trader to liquidate faster — note that throughout the trading horizon, the trajectory with a larger value of ϕ is always below the trajectory with smaller value of ϕ at the same point in time.

\mathfrak{N}	T	k	b	α
10^6	5	2×10^{-3}	2×10^{-3}	0.01

Table 1.2: Parameters for the Cartea-Jaimungal model

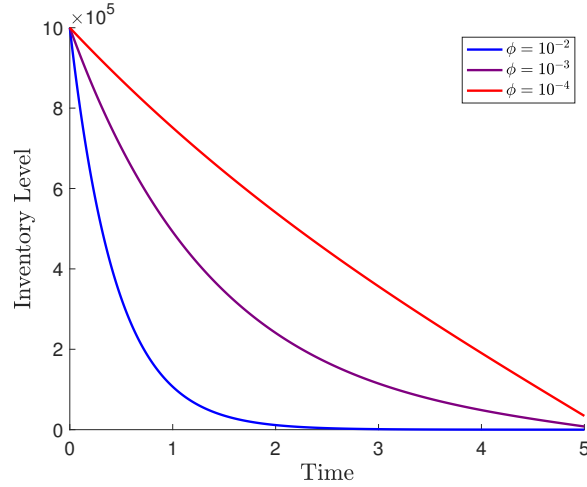


Figure 1.3: Optimal trajectories of the Cartea-Jaimungal model with different ϕ 's.

We also notice that at terminal time, there is inventory left and the trader needs to pay the terminal liquidation penalty which is $\alpha Q_\tau^{\nu^*}$ per share. If we let $\alpha \rightarrow +\infty$, which represents infinity penalty for terminal inventory, we obtain $\zeta \rightarrow 1$ and

$$\lim_{\alpha \rightarrow +\infty} Q_t^{\nu^*} = \lim_{\alpha \rightarrow +\infty} \frac{\zeta e^{\gamma(T-t)} - e^{\gamma(T-t)}}{\zeta e^{\gamma T} - e^{\gamma T}} \mathfrak{N} = \frac{\sinh(\gamma(T-t))}{\sinh(\gamma T)} \mathfrak{N}, \quad (1.25)$$

which has the same form as the optimal trajectory (1.11) as in the Almgren-Chriss model. Recall that as part of the model setup, the Almgren-Chriss model forces the terminal

inventory to be zero. This limiting behaviour shows that the Almgren-Chriss model is a special case of the Cartea-Jaimungal model when the trader requires complete liquidation by the terminal date.

1.4 Contributions and outline

In this thesis we address three mathematical problems arising in algorithmic trading for makers and takers of liquidity in an order driven market.

In Chapter 2 we consider the problem of market making with momentum in prices. We show how a market maker employs information about the momentum in the price of the asset (i.e., alpha signal) to make decisions in her liquidity provision strategy in an order driven electronic market. The market maker makes profits using both limit orders and market orders. We assume the drift of the midprice is affected by the arrival of market orders from other market participants and the market maker herself. We show that the market maker profits not only from earning the spreads by providing liquidity to the market, but also from adjusting her strategy according to the alpha signal, and experiencing the midprice change by taking positions with same direction as the alpha signal.

In Chapter 3 we consider the problem of spoofing in order driven markets. We model the trading strategy of an investor who spoofs the limit order book (LOB) to increase the revenue obtained from selling a position in a security. The strategy employs, in addition to sell limit orders (LOs) and sell market orders (MOs), a large number of spoof buy LOs to manipulate the volume imbalance of the LOB. The arrival rate of buy MOs increases because other traders believe that the spoofed buy-heavy LOB shows the true supply of liquidity and interpret this imbalance as an upward pressure in prices. Our results show that spoofing considerably increases the revenues from liquidating a position. The spoof strategy employs, on average, fewer sell MOs (than a strategy without spoof LOs) and from executing roundtrip trades that are initiated by buy spoof LOs that are ‘inadvertently’ filled and subsequently unwound with sell LOs.

In Chapter 4 we consider the problem of market making with minimum resting times. We show how the supply of liquidity in order driven markets is affected if limit orders (LOs) are forced to rest in the limit order book (LOB) for a minimum resting time (MRT)

before they can be cancelled. The bid-ask spread increases as the MRT increases because market makers (MMs) increase the depth of their LOs to protect them from being picked off by other traders. We also show that the expected profits of the MMs increase when the MRT increases. The intuition is as follows. As the MRT increases, there are two opposing forces at work. One, the longer the MRT, the more likely the LOs are to be filled and, on average, shares are sold at a loss. Two, because the depth of the posted LOs increases, the probability that the LO is picked off by other traders before the end of the MRT decreases. The net effect is that a longer MRT leads to a higher expected profit. We also show that the depth of LOs increases when the volatility of the price of the asset increases. Also, the depth of LOs increases when the arrival rate of market orders increases because it is less likely that LOs will be picked off by the end of the MRT. Finally, our model also makes predictions about the overall liquidity of the market. We show that MMs choose to supply the minimum amount of shares per LO allowed by the exchange because expected profits are maximised when liquidity provided is lowest.

Finally we draw conclusions and state possible future research directions in Chapter 5.

Chapter 2

Market making with momentum in prices

2.1 Introduction

In this chapter we show how a market maker employs information about the momentum in the price of the asset to provide liquidity in an order driven electronic market. The midprice of the asset is modelled by a pure jump process and we refer to the arrival rate of the midprice innovations as the alpha signal. Liquidity taking orders, news, and other information in the marketplace affect the alpha signal. When a buy market order (MO) arrives in the exchange, the alpha signal jumps up and when a sell MO arrives in the exchange, the alpha signal jumps down. The alpha signal also undergoes diffusive shocks that represent other events that affect the limit order book (LOB), in particular the best bid and best ask prices of the asset. When the value of the alpha signal is positive (negative), it is more likely to observe positive (negative) price innovations — the sign of the alpha signal determines the momentum in the midprice of the asset.

The market maker maximises terminal expected wealth, while penalising inventory holdings throughout the life of the trading strategy. We formulate the problem as a stochastic and impulse control problem and derive the corresponding Hamilton-Jacobi-Bellman Quasi-Variational Inequality (HJBQVI), which we solve numerically.

We use Nasdaq high-frequency data to estimate the parameters of the model and run simulations to show the performance of the strategy. When inventory is close to zero and the alpha signal is near zero, the market maker posts both sell and buy limit orders (LOs) — this is optimal because the arrival rates of upward and downward jumps in the

midprice is approximately the same, so there is zero momentum in the midprice. As the value of the alpha signal increases, the market maker posts buy LOs and does not post sell LOs to minimise adverse selection costs, i.e., to avoid being picked off by high-frequency traders. When the value of the alpha signal increases beyond a critical level, the market maker executes buy MOs in anticipation of an imminent increase in the price of the asset. This expected increase in price is large enough to cover the costs that stem from liquidity taking fees (i.e., exchange fees) and from crossing the spread — the behaviour of the strategy is similar when the value of the alpha signal is negative and inventory is close to zero.

As the market maker increases her tolerance to inventory risk, the expected profits that stem from employing the alpha signal increase because the strategy employs more speculative MOs in the direction of the expected change in prices and performs more roundtrip trades with LOs. On the other hand, these speculative directional trades and additional LOs increase the volatility of the inventory, and consequently, the volatility of the profits of the strategy also increase. The optimal strategy trades off expected profits and the volatility of the inventory holdings. We show that there is a range of tolerance to inventory risk where the expected profits of the strategy increase and the volatility of the profits of the strategy hardly changes as the tolerance to inventory risk increases. Finally, if the market maker's tolerance to inventory risk is very low, the strategy benefits very little from the alpha signal and, all else being equal, the expected profits of the strategy are lowest.

There is abundant literature on market making. Ho et al. (1981) consider the problem of optimal quoting, where the market maker decides the optimal depth of the bid and ask prices, relative to the price of the asset. The market maker assumes there is a linear-decay relation between the depth and the intensity of the market orders. Avellaneda and Stoikov (2008) consider a similar problem where they assume an exponential decay relation between the depth of LOs and the arrival intensity of MOs. In a similar vein, Guéant et al. (2013) solve the problem of a market maker who imposes bounds on the inventory holdings. Cartea et al. (2014) consider a model that accounts for the dependence between MOs, LOB dynamics, and midprice innovations, see also Cartea et al. (2018a). The work of Cartea and Jaimungal (2015c) introduces several risk metrics that the market maker employs to trade off inventory risk against expected profits. In Cartea et al. (2017) the authors allow for model ambiguity to make the market making model robust

to misspecification in the: arrival rate of MOs, fill probability of LOs, and midprice dynamics.

The remainder of the chapter is organised as follows. Section 2.2 introduces the model setup, shows how to estimate the parameters of the model, and illustrates some features of the optimal market making strategy. Section 2.3 presents simulations of the optimal strategy and discusses the distributions of the profit and loss (PnL) and of the number of filled LOs and executed MOs. Section 2.5 introduces the numerical scheme and shows convergence results. Finally, Section 2.6 presents our conclusions and we collect proofs in Section 2.7.

2.2 Model

We fix a complete filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \in [0, T]}, \mathbb{P})$, where $\{\mathcal{F}_t\}_{t \in [0, T]}$ is the natural filtration generated by the collection of observable stochastic processes that we define below. The market maker trades over the time window $[0, T]$, where the terminal time T is a positive constant.

Let $S = (S_t)_{t \geq 0}$ denote the midprice of the asset and assume it follows the pure jump process

$$dS_t = \sigma (dJ_t^+ - dJ_t^-), \quad (2.1)$$

where $\sigma > 0$ is the size of the tick in the LOB. Here, J_t^+ and J_t^- are doubly stochastic Poisson processes with intensities μ_t^+ and μ_t^- , respectively, and given by

$$\mu_t^+ = (\alpha_t)_+ + \theta \quad \text{and} \quad \mu_t^- = (\alpha_t)_- + \theta, \quad (2.2)$$

where $\theta > 0$ is the baseline arrival intensity of the midprice innovations.

The process $\alpha = (\alpha_t)_{t \geq 0}$ is the alpha signal and $(\cdot)_+$ and $(\cdot)_-$ are operators that take the absolute value of the positive and negative parts of the argument, so that $\alpha_t = (\alpha_t)_+ - (\alpha_t)_-$.

Other market participants send buy and sell MOs to the electronic exchange to trade the asset. Buy MOs arrive with intensity λ^+ and sell MOs arrive with intensity λ^- . Let $M^{0+} = (M_t^{0+})_{t \geq 0}$ and $M^{0-} = (M_t^{0-})_{t \geq 0}$ denote the counting processes of buy and sell

MOs executed by other market participants – below we provide details of the effect that MOs have on the alpha signal.

The market maker decides when to post buy and sell LOs. The control process $l_t^+ \in \{0, 1\}$ represents the activity of sell LOs and $l_t^- \in \{0, 1\}$ represents the activity of buy LOs. When $l_t^+ = 1$, the market maker posts a sell LO and when $l_t^+ = 0$, the strategy does not post a sell LO; the control for a buy LO is similar. Sell and buy LOs are of volume one and are posted at the best ask and best bid prices $S_t + \Delta$ and $S_t - \Delta$, respectively, where $\Delta \geq 0$ denotes the half-spread. Conditioned on the arrival of a buy (sell) MO from other participants, the sell (buy) LO posted by the market maker is filled with probability p^+ (p^-) $\in (0, 1]$. The processes $N^{+,l} = (N_t^{+,l})_{t \geq 0}$ and $N^{-,l} = (N_t^{-,l})_{t \geq 0}$ count the number of filled sell and buy LOs by the market maker.

The vectors $\boldsymbol{\tau}^+ = (\tau_1^+, \tau_2^+, \tau_3^+ \dots)$ and $\boldsymbol{\tau}^- = (\tau_1^-, \tau_2^-, \tau_3^- \dots)$ represent the times when the market maker sends MOs, where the superscript $+$ denotes buy MOs and the superscript $-$ denotes sell MOs. The market maker sends MOs of volume one. The counting processes for the market maker's buy and sell MOs are denoted by $M^+ = (M_t^+)_{t \geq 0}$ and $M^- = (M_t^-)_{t \geq 0}$ and are given by

$$M_t^+ = \sum_{k=1}^{\infty} \mathbb{1}_{\{\tau_k^+ \leq t\}} \quad \text{and} \quad M_t^- = \sum_{k=1}^{\infty} \mathbb{1}_{\{\tau_k^- \leq t\}}. \quad (2.3)$$

Buy and sell MOs execute at prices $S_t + \Upsilon_{MO}$ and $S_t - \Upsilon_{MO}$, respectively, where the liquidity taking costs are $\Upsilon_{MO} = \Delta + \varepsilon_{MO}$ and $\varepsilon_{MO} > 0$ is the exchange fee. We denote the control of the market maker by $\nu = (l^\pm, \boldsymbol{\tau}^\pm)$.

The assumptions about the volume of the market maker's LOs and MOs can be easily relaxed. For example, the volume of the MOs and LOs could be greater than one, and in particular, we can extend the model to one in which the market maker controls the volume of each LO and each MO she sends to the exchange. For the scope of this chapter, we assume that conditioned on the arrival of a MO, the LOs posted by the market maker are filled with probabilities $p^+ = p^- = 1$.

Finally, the alpha signal follows the dynamics

$$d\alpha_t^\nu = -\kappa \alpha_t^\nu dt + \xi dW_t + \eta^+ (dM_t^{0+} + dM_t^+) - \eta^- (dM_t^{0-} + dM_t^-), \quad \alpha_0 = 0, \quad (2.4)$$

where $W = (W_t)_{t \geq 0}$ is a standard Brownian motion, and $\kappa, \xi, \eta^+, \eta^-$ are positive constants. Observe that the arrival of MOs in the exchange cause the alpha signal to jump

up by η^+ if the MO is a buy order or jump down by η^- if the MO is a sell order. The Brownian component of the alpha signal represents the effect of the arrival of news in the market and other information that liquidity providers employ in their strategies.

The market maker's controlled inventory is denoted by $Q^\nu = (Q_t^\nu)_{t \geq 0}$ with dynamics

$$Q_t^\nu = N_t^{-,l} - N_t^{+,l} + M_t^+ - M_t^- ,$$

and the inventory obeys the bounds $-\bar{Q} \leq Q_t^\nu \leq \bar{Q}$ for some integer $\bar{Q} > 0$. The controlled cash process is denoted by $X^\nu = (X_t^\nu)_{t \geq 0}$ and satisfies

$$dX_t^\nu = (S_t^\nu + \Upsilon_{LO}) dN_t^{+,l} - (S_t^\nu - \Upsilon_{LO}) dN_t^{-,l} - (S_t^\nu + \Upsilon_{MO}) dM_t^+ + (S_t^\nu - \Upsilon_{MO}) dM_t^- .$$

The market maker's optimisation problem is

$$H(t, x, S, \alpha, q) = \sup_{\nu \in \mathcal{A}} \mathbb{E}_{t,x,S,\alpha,q} \left[X_T^\nu + Q_T^\nu (S_T^\nu - \text{sign}(Q_T^\nu) \Upsilon_{MO} - \psi Q_T^\nu) - \phi \int_t^T (Q_s^\nu)^2 ds \right] , \quad (2.5)$$

where $\mathbb{E}_{t,x,S,\alpha,q}[\cdot]$ is the expectation operator conditional on $X_{t-}^\nu = x, S_{t-}^\nu = S, \alpha_{t-}^\nu = \alpha, Q_{t-}^\nu = q$, and \mathcal{A} is the set of admissible strategies which consists of \mathcal{F}_t -stopping times τ^\pm and \mathcal{F}_t -predictable stochastic control processes l^\pm so that $Q_t^\nu \in [-\bar{Q}, \bar{Q}]$. We give a brief interpretation of each term in the market maker's value function. The terminal cash of the market maker is X_T^ν . The term $Q_T^\nu (S_T^\nu - \text{sign}(Q_T^\nu) \Upsilon_{MO} - \psi Q_T^\nu)$ represents the earnings from employing MOs to liquidate any outstanding position at the end of the trading horizon, the parameter ψ is a positive constant and ψQ_T^ν represents the extra costs of walking the LOB. The term $\phi \int_t^T (Q_s^\nu)^2 ds$, with $\phi \geq 0$, is the running inventory penalty. The penalty parameter ϕ and the inventory cap \bar{Q} represent the market maker's tolerance to inventory risk, see Cartea et al. (2015) and Guéant (2016). Another interpretation of the running inventory penalty is that the market maker is ambiguity averse to the drift of the midprice, see Cartea et al. (2017).

By standard results, see Øksendal and Sulem (2007), the value function (2.5) is the unique

viscosity solution of the Hamilton-Jacobi-Bellman quasi-variational inequality (HJBQVI)

$$\begin{aligned}
\max \Bigg\{ & \partial_t H + (\alpha^+ + \theta) \left(H(t, x, S + \sigma, \alpha, q) - H \right) \\
& + (\alpha^- + \theta) \left(H(t, x, S - \sigma, \alpha, q) - H \right) - \kappa \alpha \partial_\alpha H + \frac{1}{2} \xi^2 \partial_{\alpha\alpha} H - \phi q^2 \\
& + \lambda^+ \sup_{l^+ \in \{0,1\}} \left[l^+ \left(H(t, x + (S + \Upsilon_{LO}), S, \alpha + \eta^+, q - 1) - H \right) \right. \\
& \quad \left. + (1 - l^+) \left(H(t, x, S, \alpha + \eta^+, q) - H \right) \right] \\
& + \lambda^- \sup_{l^- \in \{0,1\}} \left[l^- \left(H(t, x - (S - \Upsilon_{LO}), S, \alpha - \eta^-, q + 1) - H \right) \right. \\
& \quad \left. + (1 - l^-) \left(H(t, x, S, \alpha - \eta^-, q) - H \right) \right]; \\
& H(t, x + (S - \Upsilon_{MO}), S, \alpha, q - 1) - H; \\
& H(t, x - (S + \Upsilon_{MO}), S, \alpha, q + 1) - H \Bigg\} = 0, \tag{2.6}
\end{aligned}$$

with terminal condition

$$H(T, x, S, \alpha, q) = x + q \left(S - \text{sign}(q) \Upsilon_{MO} - \psi q \right). \tag{2.7}$$

The stochastic controls, in feedback form, to post LOs are

$$\begin{aligned}
l^+ &= \mathbb{1}_{\{H(t, x + (S + \Upsilon_{LO}), S, \alpha + \eta^+, q - 1) > H(t, x, S, \alpha + \eta^+, q)\}}, & \text{sell LO,} \\
l^- &= \mathbb{1}_{\{H(t, x - (S - \Upsilon_{LO}), S, \alpha - \eta^-, q + 1) > H(t, x, S, \alpha - \eta^-, q)\}}, & \text{buy LO.}
\end{aligned} \tag{2.8}$$

In the expression for the control of sell LOs, the term $H(t, x + (S + \Upsilon_{LO}), S, \alpha + \eta^+, q - 1)$ is the value function when the market maker's sell LO is filled by a buy MO. The term $H(t, x, S, \alpha + \eta^+, q)$ is the value function when a buy MO arrives, but the market maker does not post a sell LO. Similarly, in the expression for the control of buy LOs, the term $H(t, x - (S - \Upsilon_{LO}), S, \alpha - \eta^-, q + 1)$ is the value function when the market maker's buy LO is filled by a sell MO. The term $H(t, x, S, \alpha - \eta^-, q)$ is the value function when a buy MO arrives, but the market maker did not post a buy LO.

Substitute the ansatz $H(t, x, S, \alpha, q) = x + q S + \tilde{h}(t, \alpha, q)$ in (2.6) and write

$$\begin{aligned} \max \left\{ \partial_t \tilde{h} + \alpha \sigma q - \kappa \alpha \partial_\alpha \tilde{h} + \frac{1}{2} \xi^2 \partial_{\alpha\alpha} \tilde{h} - \phi q^2 \right. \\ + \lambda^+ \sup_{l^+ \in \{0,1\}} \left[l^+ \left(\Upsilon_{LO} + \tilde{h}(t, \alpha + \eta^+, q - 1) - \tilde{h} \right) + (1 - l^+) \left(\tilde{h}(t, \alpha + \eta^+, q) - \tilde{h} \right) \right] \\ + \lambda^- \sup_{l^- \in \{0,1\}} \left[l^- \left(\Upsilon_{LO} + \tilde{h}(t, \alpha - \eta^-, q + 1) - \tilde{h} \right) + (1 - l^-) \left(\tilde{h}(t, \alpha - \eta^-, q) - \tilde{h} \right) \right]; \\ \left. - \Upsilon_{MO} + \tilde{h}(t, \alpha, q - 1) - \tilde{h}; \right. \\ \left. - \Upsilon_{MO} + \tilde{h}(t, \alpha, q + 1) - \tilde{h} \right\} = 0, \end{aligned} \quad (2.9)$$

with terminal condition

$$\tilde{h}(T, \alpha, q) = q \left(-\text{sign}(q) \Upsilon_{MO} - \psi q \right). \quad (2.10)$$

The optimal feedback controls to post LOs become

$$l^+ = \mathbb{1}_{\{\Upsilon_{LO} + \tilde{h}(t, \alpha + \eta^+, q - 1) > \tilde{h}(t, \alpha + \eta^+, q)\}} \quad \text{and} \quad l^- = \mathbb{1}_{\{\Upsilon_{LO} + \tilde{h}(t, \alpha - \eta^-, q + 1) > \tilde{h}(t, \alpha - \eta^-, q)\}} \quad (2.11)$$

The three terms of the ansatz represent: the cash accumulated by the strategy up until time t , the mark-to-market value of the inventory, and the function \tilde{h} is the extra value the market maker obtains from making markets optimally over the remaining life of the strategy.

Theorem 2.2.1. (Verification) *Let \tilde{h} be a solution to (2.9) and define a candidate solution $\tilde{H} = x + q S + \tilde{h}(t, \alpha, q)$. Then \tilde{H} equals the value function as defined in (2.5).*

Proof. For a proof see Section 2.7. □

Finally, we note that in our model the jumps in the midprice are always of size one tick. One could also assume that σ in (2.1) is the average number of ticks the midprice jumps when there is a price innovation. Also, one could consider a model where the spread is also stochastic, see Cartea et al. (2018). For simplicity, we assume the spread is constant at one tick, which is the case of the large-stick stocks such as the ones we study here.

2.2.1 Parameter estimates

In this section we employ Nasdaq high-frequency data to estimate some of the model parameters. The data are messages sent to the Nasdaq exchange, with nanosecond time stamps, that indicate the events in the LOB such as arrival and cancellation of LOs, and arrival of MOs. For simplicity, we assume that the volatility parameter ξ in (2.4) is zero.

For the time interval $[0, T]$, let

$$t^+ = \{t_1^+, t_2^+, \dots, t_{m^+}^+\} \quad \text{and} \quad t^- = \{t_1^-, t_2^-, \dots, t_{m^-}^-\} \quad (2.12)$$

denote the times when the midprice jumps, up (+) and down (−), and let the arrays

$$\tau^{0+} = \{\tau_1^{0+}, \tau_2^{0+}, \dots, \tau_{n^+}^{0+}\} \quad \text{and} \quad \tau^{0-} = \{\tau_1^{0-}, \tau_2^{0-}, \dots, \tau_{n^-}^{0-}\} \quad (2.13)$$

represent the arrival times of the MOs: the superscripts 0+ and 0− denote buy and sell orders, respectively.

We also denote by τ^0 the array that combines τ^{0+} , τ^{0-} and $\{0, T\}$, such that τ^0 is an increasing sequence that starts at $\tau_0^0 = 0$ and ends at time $\tau_{n^++n^-+1}^0 = T$. The entries of the array τ^0 are observed, so we write the alpha signal as follows:

$$\alpha_t = \eta^+ \sum_{i=1}^{n^+} \exp\left(-\kappa(t - \tau_i^{0+})\right) \mathbb{1}_{\{t \geq \tau_i^{0+}\}} - \eta^- \sum_{i=1}^{n^-} \exp\left(-\kappa(t - \tau_i^{0-})\right) \mathbb{1}_{\{t \geq \tau_i^{0-}\}} \quad (2.14)$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function.

The log-likelihood of t^\pm , given $\tau^{0\pm}$, is

$$\begin{aligned} \mathcal{L}(\Theta) &= \log \mathbb{P}(t^\pm | \tau^{0\pm}, \Theta) \\ &= \log \mathbb{P}(t^+ | \tau^{0+}, \Theta) + \log \mathbb{P}(t^- | \tau^{0-}, \Theta) \\ &= \log \left[e^{-\int_0^T \mu_s^+ ds} \prod_{i=1}^{m^+} \mu_{t_i^+}^+ \right] + \log \left[e^{-\int_0^T \mu_s^- ds} \prod_{i=1}^{m^-} \mu_{t_i^-}^- \right] \\ &= -\int_0^T \mu_s^+ ds + \sum_{i=1}^{m^+} \log \mu_{t_i^+}^+ - \int_0^T \mu_s^- ds + \sum_{i=1}^{m^-} \log \mu_{t_i^-}^- \\ &= -2\theta T - \int_0^T ((\alpha_s)_+ - (\alpha_s)_-) ds + \sum_{i=1}^{m^+} \log \left[\left(\alpha_{t_i^+}^+ \right)_+ + \theta \right] + \sum_{i=1}^{m^-} \log \left[\left(\alpha_{t_i^-}^- \right)_- + \theta \right], \end{aligned} \quad (2.15)$$

where $\Theta = (\kappa, \eta^+, \eta^-, \theta)$ denotes the set of parameters to estimate.

Use (2.14) to write each term in the log-likelihood. We start with the term α_{t-} , which is given by

$$\alpha_{t-} = \eta^+ \sum_{i=1}^{n^+} e^{-\kappa(t-\tau_i^{0+})} \mathbb{1}_{\{t > \tau_i^{0+}\}} - \eta^- \sum_{i=1}^{n^-} e^{-\kappa(t-\tau_i^{0-})} \mathbb{1}_{\{t > \tau_i^{0-}\}}. \quad (2.16)$$

Now, to write the integrated alpha signal that appears in the last line of (2.15) we note that the process α_t does not change sign in the interval $[\tau_i^0, \tau_{i+1}^0)$ for $i = 0, 1, \dots, n^+ + n^-$, so by the definition of α_t we have

$$\begin{aligned} \int_0^T (\alpha_s)_+ ds &= -\frac{1}{\kappa} \sum_{i=0}^{n^++n^-} \mathbb{1}_{\{\alpha_{\tau_i^0} \geq 0\}} \left[\eta^+ \sum_{j=1}^{n^+} \left(e^{-\kappa(\tau_{i+1}^0 \vee \tau_j^{0+} - \tau_j^{0+})} - e^{-\kappa(\tau_i^0 \vee \tau_j^{0+} - \tau_j^{0+})} \right) \right. \\ &\quad \left. - \eta^- \sum_{j=1}^{n^-} \left(e^{-\kappa(\tau_{i+1}^0 \vee \tau_j^{0-} - \tau_j^{0-})} - e^{-\kappa(\tau_i^0 \vee \tau_j^{0-} - \tau_j^{0-})} \right) \right], \end{aligned} \quad (2.17)$$

and

$$\begin{aligned} \int_0^T (\alpha_s)_- ds &= \frac{1}{\kappa} \sum_{i=0}^{n^++n^-} \mathbb{1}_{\{\alpha_{\tau_i^0} \leq 0\}} \left[\eta^+ \sum_{j=1}^{n^+} \left(e^{-\kappa(\tau_{i+1}^0 \vee \tau_j^{0+} - \tau_j^{0+})} - e^{-\kappa(\tau_i^0 \vee \tau_j^{0+} - \tau_j^{0+})} \right) \right. \\ &\quad \left. - \eta^- \sum_{j=1}^{n^-} \left(e^{-\kappa(\tau_{i+1}^0 \vee \tau_j^{0-} - \tau_j^{0-})} - e^{-\kappa(\tau_i^0 \vee \tau_j^{0-} - \tau_j^{0-})} \right) \right]. \end{aligned} \quad (2.18)$$

Finally, maximise the log-likelihood to obtain the parameter estimates:

$$\hat{\Theta} = \operatorname{argmax} \mathcal{L}(\Theta). \quad (2.19)$$

Table 2.1 reports parameter estimates for 10 stocks traded in Nasdaq – we employ data from 10:00 to 15:30 on 11 February 2019. From Table 2.1 we observe that the size of the jumps in the alpha signal induced by the arrival of MOs is much larger than the baseline of the intensities for midprice jumps, i.e., $\hat{\eta}^\pm \gg \hat{\theta}$. The value of the mean-reverting speed of the alpha process, $\hat{\kappa}$, is at least as large as the size of the jumps in the alpha process, and for some stocks it is much larger the size of the jumps. We recall the half-life of α is $\kappa^{-1} \log 2$, and from Table 2.1 we see that the estimated half-lives are all less than 0.02

seconds for the stocks we study. This shows that the alpha signal decays very quickly, so the upward (downward) momentum of the midprice lasts for a very short period of time. Only market participants (e.g., high-frequency traders and high-frequency market makers) who are able to process information and access the market within milliseconds will be able to take advantage of the alpha signal, see Cartea and Sánchez-Betancourt (2018) for a discussion of latency in electronic markets.

	$\hat{\kappa}$	$\hat{\eta}^+$	$\hat{\eta}^-$	$\hat{\theta}$	$\hat{\lambda}^+$	$\hat{\lambda}^-$
COST	85.669	92.024	78.548	0.446	0.074	0.086
CSCO	310.790	75.263	52.646	0.055	0.086	0.108
EBAY	267.880	30.221	54.676	0.046	0.129	0.054
EXPE	62.425	54.592	49.350	0.342	0.084	0.099
GILD	269.255	118.323	105.776	0.383	0.101	0.105
MSFT	225.456	97.085	84.773	0.575	0.287	0.268
ORCL	355.617	105.237	77.118	0.091	0.072	0.061
PYPL	236.140	134.227	106.058	0.519	0.148	0.156
QCOM	459.756	104.341	93.905	0.146	0.122	0.125
VRTX	46.754	26.092	35.014	0.118	0.022	0.027

Table 2.1: Parameter estimates. Data: Nasdaq, from 10:00 to 15:30, 11 February 2019.

	$\hat{\kappa}$	$\hat{\eta}$	$\hat{\theta}$	$\hat{\lambda}$
COST	83.622	82.638	0.446	0.080
CSCO	309.208	62.119	0.055	0.097
EBAY	270.145	37.881	0.046	0.092
EXPE	66.551	54.665	0.343	0.092
GILD	265.501	109.970	0.383	0.103
MSFT	238.284	95.052	0.576	0.278
ORCL	370.179	95.664	0.091	0.066
PYPL	254.785	126.808	0.521	0.152
QCOM	441.456	95.253	0.146	0.124
VRTX	45.930	30.428	0.118	0.025

Table 2.2: Parameter estimates with constraints of symmetry. Data: Nasdaq, from 10:00 to 15:30, 11 February 2019.

2.2.2 Optimal strategy

In this section we solve the market maker’s HJBQVI numerically, where we employ the parameters in Table 2.3, and discuss the stylised features of the optimal strategy. Note

that for the large-tick stocks we study, the quoted spread is very often one tick, i.e., $2\Delta = 0.01$, which is also the size of the price innovations in our model.

Δ	ε_{MO}	ε_{LO}	σ	θ	η	κ	ξ	λ^+	λ^-	ϕ	ψ	T	\bar{Q}
0.005	0.003	0.002	0.01	0.1	60	200	1	1	1	10^{-6}	0	60	4

Table 2.3: Model parameters.

Figure 2.1 shows the market maker's optimal strategy for the last ten seconds of the trading horizon, i.e., $t \in [50, 60]$, for three inventory levels (far from maturity, the optimal strategy does not vary much as time elapses, so we omit the first 50 seconds). We first focus on the case $q = 0$. When the value of α_t is near zero the strategy posts both sell and buy LOs. As the value of α increases, the strategy posts buy LOs because the midprice of the asset will, on average, increase due to the positive trend. The strategy does not post sell LOs at the best ask price because the LOs would be picked off by high-frequency traders who also trade on a momentum signal. As the value of α increases further, the market maker sends buy MOs in anticipation of a price increase, i.e., due to the strong upward pressure on the midprice of the asset. The optimal strategy is similar when the value of α is negative, and in this particular case the optimal strategy is symmetric with respect to $\alpha = 0$ because the parameters we choose are symmetric, i.e., $\lambda^+ = \lambda^-$ and $q = 0$. Near maturity, the regions of LOs and MOs shrink. We notice there is a sliver near maturity where the optimal strategy does not post LOs or send MOs. In this region, the market maker's position is already zero when it is close to maturity and the expected profit from taking positions to earn the spread are offset by the uncertainty in the midprice and liquidation costs. Near maturity, if the non-zero inventory is not cleared, the market maker needs to pay the spread and fees to unwind inventories with MOs.

If the inventory of the market maker is short, for example $q = -1$, the buy LO region and the buy MO region expand compared with that of the strategy for $q = 0$, and the sell LO and the sell MO regions shrink. The optimal strategy for $q = 1$ is symmetric with the optimal strategy when $q = -1$ because of the symmetry in the parameters.

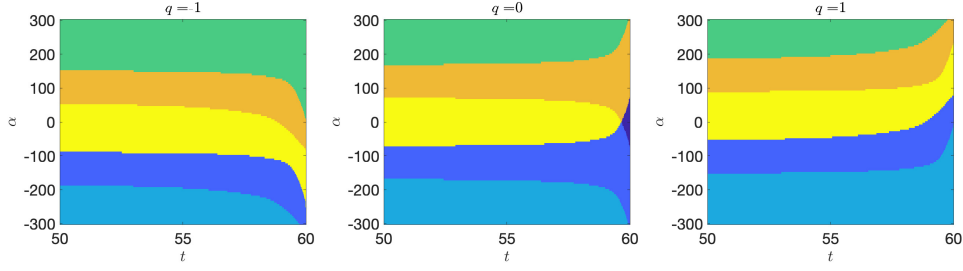


Figure 2.1: Optimal strategy of the market maker. Dark blue: no LOs in the LOB and no MOs. Blue: post sell LO ($l^+ = 1$). Light blue: send sell MO (τ^-). Orange: post buy LO ($l^- = 1$). Green: send buy MO (τ^+). Yellow: post both sell and buy LO ($l^+ = l^- = 1$). Near maturity, there is a green sliver when $q = -1$, a dark blue sliver when $q = 0$, and a light blue sliver when $q = 1$.

We recall that the value function is $H = x + qS + \tilde{h}(t, \alpha, q)$. Figure 2.2 shows the surface of the function \tilde{h} evaluated at $t = 0$. Recall the function \tilde{h} is the extra value the market maker obtains from making markets optimally over the remaining life of the strategy. The value of \tilde{h} is lowest when both the initial inventory is positive and the initial value of alpha signal is negative or when the initial inventory is negative and the alpha signal is positive. Note that if the market maker has a large long position in the asset and its midprice is expected to trend downwards, the value of the inventory is expected to decrease. Similarly, the value of \tilde{h} is highest when the initial inventory is in the direction of the initial alpha signal (i.e., the value of α is positive and q is positive or the value of α is negative and q is negative) so the value of the inventory is expected to appreciate.

2.3 Simulation and performance of strategy

2.3.1 Sample paths

Figure 2.3 shows one simulation of the strategy: midprice sample path and trading activity, alpha signal, and inventory path. We observe that the market maker trades mostly using LOs. When a spike appears in the alpha signal which indicates an imminent change in the midprice, a steep change in the midprice occurs. Also, in response to the spike in the alpha signal, the market maker executes MOs to adjust her inventory. The strategy employs the alpha signal to keep the inventory on target, minimise adverse selection costs, and to execute speculative directional MOs. During most of the trading horizon, the strategy employs either two-sided LOs or one-sided LOs and employs only on MO at

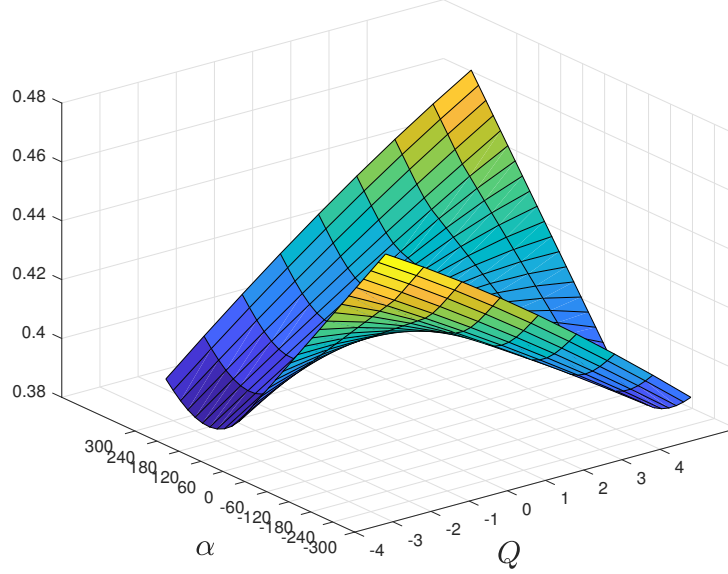
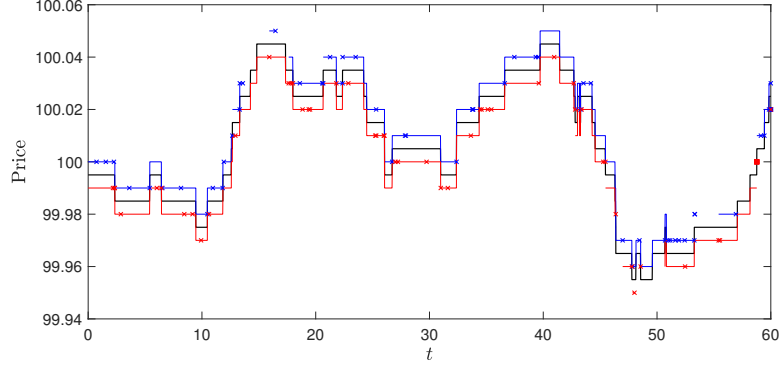
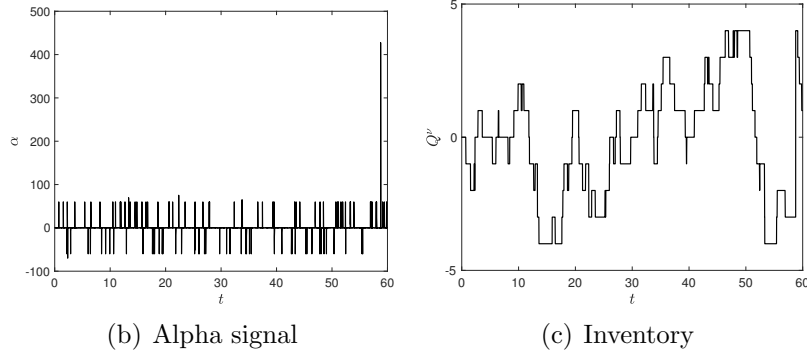


Figure 2.2: Surface of function \tilde{h} . The value of \tilde{h} is highest when the value of inventory is expected to appreciate and lowest when the value of inventory is expected to depreciate.

the end of the trading horizon. When the alpha signal is positive (negative), the strategy tends not to post sell (buy) LOs to avoid adverse selection costs. Near the end of the trading window we observe an upward spike in the alpha signal that triggers a buy MO, which is also required to unwind inventories because the strategy is close to expiry and the value of the inventory position is short, see panel (c). Throughout the trading horizon, the market maker's inventory Q_t^v tends to revert toward zero because the strategy benefits from executing roundtrip trades and accounts for the market maker's tolerance to inventory risk. Notice that panel (b) shows the time of the alpha signal being away from zero is much shorter compared to the time of the alpha signal being practically zero. For the parameters we choose as in Table 2.3, the half life of the alpha signal is approximately 0.02 seconds. This makes the actual implementation of the strategy difficult as the market maker need to react instantaneously, which is challenging in terms of both software and hardware.



(a) Black line: midprice. Blue line: post sell LO. Blue cross: sell LO is filled. Red line: post buy LO. Red cross: buy LO is filled. Blue square: execute sell MO. Red square: execute buy MO.



(b) Alpha signal

(c) Inventory

Figure 2.3: One simulation: sample path of midprice, alpha signal and inventory.

2.3.2 Distribution of orders traded

Figure 2.4 shows the distribution (10,000 simulations) of the orders traded for various values of the inventory penalty parameter ϕ . For example, as the value of the parameter ϕ decreases, both the number of filled LOs and the number of executed MOs increase because the market maker is more willing to take positions with LOs and to employ more MOs to adjust her inventory.

In the figures, light green corresponds to $\phi = 10^{-6}$, red corresponds to $\phi = 5 \times 10^{-4}$, and blue corresponds to $\phi = 10^{-3}$. In panels (c) and (d) the superposition of up to three colours appears in different shades of green. Note that when the inventory penalty parameter ϕ is smallest, there are simulation runs in which the strategy employs more than two sell and two buy MOs, and for larger values of the parameter ϕ the maximum number of sell or buy MOs we observe is two. Clearly, as the market maker becomes more tolerant to bear inventory risk, it is optimal to build larger inventory positions and

employ more speculative MOs to take advantage of the alpha signal and to send MOs to keep the inventory on target.

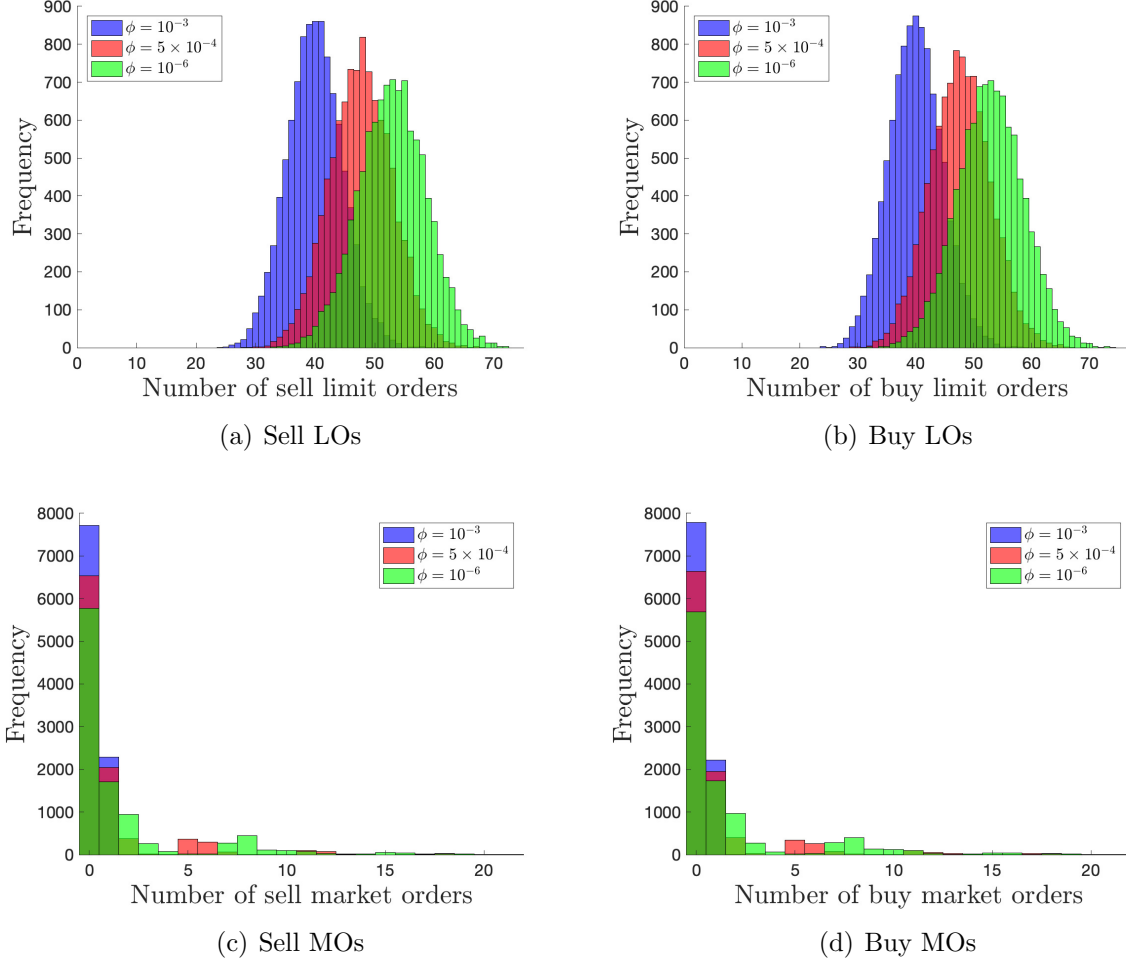


Figure 2.4: Distributions of the number of filled LOs and executed MOs for three values of the inventory penalty parameter ϕ .

2.3.3 Tradeoff: mean and standard deviation of PnL

In this section we examine the mean and the standard deviation of PnL for various values of the running inventory penalty parameter and the cap on the inventory, both of which represent the market maker's willingness to bear inventory risk. The trading horizon is $T = 60$ seconds. We also compare the PnL of a strategy that incorporates the alpha signal with a strategy that assumes $\alpha_t = 0$ throughout the trading horizon.

Figure 2.5 plots the mean of PnL against the and the standard deviation of PnL for a range of values of ϕ and Figure 2.6 plots the mean of PnL against the standard deviation

of PnL for a range of \bar{Q} , which is the cap on inventory. Both figures show the results of the optimal strategy with the alpha signal and without employing the alpha signal, i.e., $\alpha_t = 0$, during the trading horizon. Figure 2.5 shows that both the mean and the standard deviation of the PnL increase when the value of the inventory penalty parameter ϕ decreases. As the inventory penalty becomes less severe, the strategy takes larger positions in the asset (long and short) to benefit from the alpha signal. When the value of the penalty parameter ϕ is high, a slight decrease in the value of ϕ causes an increase in the mean of the PnL without a significant increase in the standard deviation of PnL. Also, the PnL of the optimal strategy is higher than the PnL of the strategy of an investor who assumes $\alpha_t = 0$ for the whole trading horizon. For low values of ϕ , the difference between the two strategies is about 0.02\$ for a trading horizon of 60 seconds.

Figure 2.6 shows that both the mean and the standard deviation of the PnL increase when the value of the cap \bar{Q} increases. A larger allowance of maximum and minimum inventory levels leads to higher expected profits because the strategy can perform, on average, more roundtrip trades and because the strategy can take full advantage of the momentum in the price of the asset. For small values of \bar{Q} , the mean of the PnL increases without a significant increase in the standard deviation of the PnL when \bar{Q} increases. The PnL with the optimal strategy is higher than the PnL of a strategy that assumes $\alpha_t = 0$. For higher values of \bar{Q} , the difference in mean PnLs is about 0.04\$ for a trading horizon of 60 seconds.

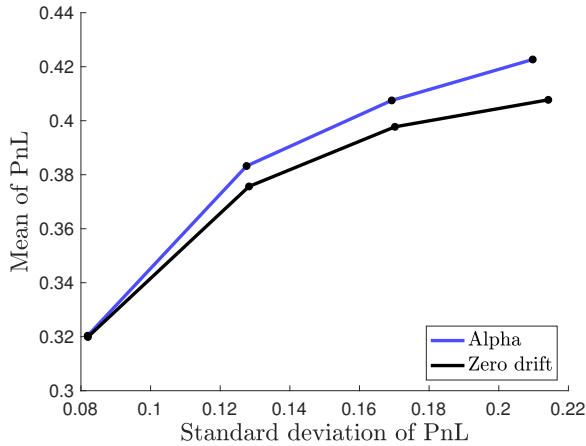


Figure 2.5: Risk-reward profile with $\bar{Q} = 4$, $T = 60$ seconds. From left to right, $\phi = 10^{-3}, 5 \times 10^{-4}, 2 \times 10^{-4}, 10^{-6}$.

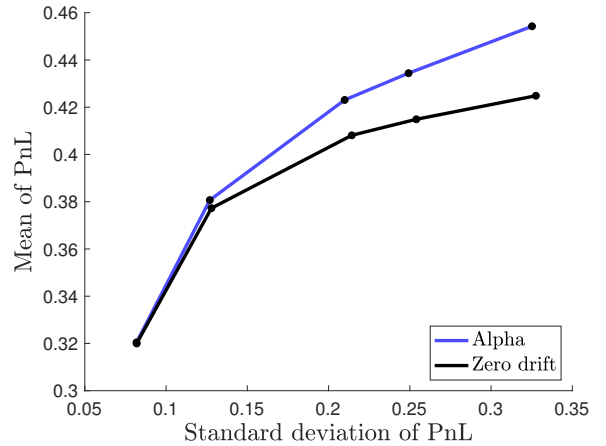


Figure 2.6: Risk-reward profile with $\phi = 10^{-6}$, $T = 60$ seconds. From left to right, $\bar{Q} = 1, 2, 4, 5, 7$.

2.4 Simulation results with parameters estimated from Nasdaq data

We use the parameter estimates of 3 stocks reported in Table 2.1 and assume $\xi = 0$, i.e., there is no diffusion in the alpha signal. We run 10,000 simulations with $T = 300$ seconds and other model parameters as in Table 2.3. Recall that the parameter estimates are obtained from high-frequency Nasdaq data on 11 February 2019 (10:00 to 15:30).

Table 2.4 reports the mean and the standard deviation of the simulated PnL for a range of values of the running inventory penalty parameter ϕ . Observe that as the value of the parameter ϕ increases, both the mean and the standard deviation of the PnL decrease.

Stock	ϕ			
	10^{-6}	3×10^{-5}	10^{-4}	3×10^{-4}
CSCO	0.257393 (0.170465)	0.228978 (0.105016)	0.188893 (0.067854)	0.188898 (0.067851)
EBAY	0.266769 (0.158333)	0.240957 (0.099900)	0.238051 (0.096836)	0.199573 (0.065479)
GILD	0.152533 (0.370212)	0.135067 (0.221638)	0.108451 (0.138233)	0.108458 (0.138176)

Table 2.4: Mean, in \$, of the PnL and standard deviation of PnL in brackets. Parameters as in Table 2.1, $T = 300$ seconds, $\bar{Q} = 4$, and various choices of ϕ .

Table 2.5 reports the mean and the standard deviation of the simulated PnL for a range of \bar{Q} with a fixed value of ϕ . As \bar{Q} increases, both the mean and the standard deviation of PnL increase.

Stock	\bar{Q}			
	2	4	7	10
CSCO	0.228793 (0.105515)	0.257393 (0.170465)	0.267399 (0.239815)	0.267638 (0.250639)
EBAY	0.241087 (0.100206)	0.266769 (0.158333)	0.270478 (0.219711)	0.270970 (0.229488)
GILD	0.135825 (0.222287)	0.152533 (0.370212)	0.175081 (0.538307)	0.175016 (0.538375)

Table 2.5: Mean, in \$, of the PnL and standard deviation of PnL in brackets. Parameters as in Table 2.1, $T = 300$ seconds, $\phi = 10^{-6}$, and various choices of \bar{Q} .

These results coincide with the properties we found in Section 2.3.3.

2.5 Numerical scheme

To solve (2.9) numerically, we assume the signal α lives on a grid with boundaries $-A \leq \alpha \leq A$ for some constant $A \in (0, +\infty)$, and the additional boundary conditions:

$$\partial_{\alpha\alpha}\tilde{h}(t, -A, q) = 0 \quad \text{and} \quad \partial_{\alpha\alpha}\tilde{h}(t, A, q) = 0 \quad (2.20)$$

for all $t \in [0, T]$ and $q \in [-\bar{Q}, \bar{Q}] \cap \mathbb{Z}$. For any $\alpha > A$ we use linear extrapolation from $\tilde{h}(t, A - \delta\alpha, q)$ and $\tilde{h}(t, A, q)$, and similarly for any $\alpha < -A$. For the simulations presented above we used $A = 300$.

To prove the convergence of the numerical scheme we impose bounds on α , which also bound the rate of increase or decrease of the midprice. Assume

$$\mu_t^+ = (\bar{\alpha}_t)_+ + \theta, \quad \mu_t^- = (\bar{\alpha}_t)_- + \theta, \quad (2.21)$$

where

$$\bar{\alpha}_t = \min\{A, \max\{-A, \alpha_t\}\}, \quad (2.22)$$

for some constant $A \in (0, +\infty)$.

We also impose the same bound on the term $-\kappa\alpha\partial_\alpha\tilde{h}$ in (2.9). Thus, we solve the HJBQVI:

$$\begin{aligned} \max \left\{ \partial_t h + \bar{\alpha} \sigma q - \kappa \bar{\alpha} \partial_\alpha h + \frac{1}{2} \xi^2 \partial_{\alpha\alpha} h - \phi q^2 \right. \\ + \lambda^+ \sup_{l^+ \in \{0,1\}} \left[l^+ \left(\Upsilon_{LO} + h(t, \alpha + \eta^+, q - 1) - h \right) + (1 - l^+) \left(h(t, \alpha + \eta^-, q) - h \right) \right] \\ + \lambda^- \sup_{l^- \in \{0,1\}} \left[l^- \left(\Upsilon_{LO} + h(t, \alpha - \eta^-, q + 1) - h \right) + (1 - l^-) \left(h(t, \alpha - \eta^-, q) - h \right) \right]; \\ \left. \begin{aligned} & -\Upsilon_{MO} + h(t, \alpha, q - 1) - h; \\ & -\Upsilon_{MO} + h(t, \alpha, q + 1) - h \end{aligned} \right\} = 0, \quad (2.23) \end{aligned}$$

with terminal condition

$$h(T, \alpha, q) = q \left(-\text{sign}(q) \Upsilon_{MO} - \psi q \right). \quad (2.24)$$

Theorem 2.5.1. (*Comparison principle*) Let h_1 and h_2 be, respectively, a bounded subsolution and a bounded supersolution of (2.23). Suppose that $h_1(T) \leq h_2(T)$. Then $h_1 \leq h_2$.

Proof. See Theorem 2.5.11. of Seydel (2010). \square

Here we describe the numerical scheme we employ to solve (2.23). Let $\mathbb{T}_{\delta t}$ be the uniform grid on $[0, T]$ with step size $\delta t > 0$ and $\mathbb{R}_{\delta\alpha}$ be the uniform grid on \mathbb{R} with step size $\delta\alpha > 0$. We define, for any function $\varphi : [0, T] \times \mathbb{R}_{\delta\alpha} \times ([-\bar{Q}, \bar{Q}] \cap \mathbb{Z}) \rightarrow \mathbb{R}$, the operator

$$\mathcal{S}^{\delta t, \delta\alpha}(t, \alpha, q, \varphi) = \max \left[\mathcal{T}^{\delta t, \delta\alpha}(t, \alpha, q, \varphi), \mathcal{M}^{\delta t, \delta\alpha}(t, \alpha, q, \varphi) \right], \quad (2.25)$$

where

$$\begin{aligned} & \mathcal{T}^{\delta t, \delta\alpha}(t, \alpha, q, \varphi) \\ &= \varphi + \delta t \left\{ \bar{\alpha} \sigma q + \kappa \bar{\alpha}_- \frac{\varphi(t, \alpha + \delta\alpha, q) - \varphi(t, \alpha, q)}{\delta\alpha} - \kappa \bar{\alpha}_+ \frac{\varphi(t, \alpha, q) - \varphi(t, \alpha - \delta\alpha, q)}{\delta\alpha} \right. \\ & \quad + \frac{\xi^2}{2} \frac{\varphi(t, \alpha + \delta\alpha, q) - 2\varphi(t, \alpha, q) + \varphi(t, \alpha - \delta\alpha, q)}{\delta\alpha^2} - \phi q^2 \\ & \quad + \lambda^+ \sup_{l^+ \in \{0,1\}} \left[l^+ \left(\Upsilon_{LO} + \mathcal{I}^+ \varphi(t, \alpha, q-1) - \varphi \right) + (1-l^+) \left(\mathcal{I}^+ \varphi(t, \alpha, q) - \varphi \right) \right] \\ & \quad \left. + \lambda^- \sup_{l^- \in \{0,1\}} \left[l^- \left(\Upsilon_{LO} + \mathcal{I}^- \varphi(t, \alpha, q+1) - \varphi \right) + (1-l^-) \left(\mathcal{I}^- \varphi(t, \alpha, q) - \varphi \right) \right] \right\}, \end{aligned} \quad (2.26)$$

and

$$\mathcal{M}^{\delta t, \delta\alpha}(t, \alpha, q, \varphi) = \max \left\{ \varphi(t, \alpha, q-1) - \Upsilon_{MO}, \quad \varphi(t, \alpha, q+1) - \Upsilon_{MO} \right\}. \quad (2.27)$$

The linear interpolation operators \mathcal{I}^+ and \mathcal{I}^- for the jumps in α are given by

$$\begin{aligned} \mathcal{I}^+ \varphi(t, \alpha, q) &= \varphi \left(t, \alpha + \left\lfloor \frac{\eta^+}{\delta\alpha} \right\rfloor \delta\alpha, q \right) \\ & \quad + \left(\frac{\eta^+}{\delta\alpha} - \left\lfloor \frac{\eta^+}{\delta\alpha} \right\rfloor \right) \left(\varphi \left(t, \alpha + \left\lceil \frac{\eta^+}{\delta\alpha} \right\rceil \delta\alpha, q \right) - \varphi \left(t, \alpha + \left\lfloor \frac{\eta^+}{\delta\alpha} \right\rfloor \delta\alpha, q \right) \right), \end{aligned} \quad (2.28)$$

and

$$\begin{aligned} \mathcal{I}^- \varphi(t, \alpha, q) &= \varphi \left(t, \alpha - \left\lfloor \frac{\eta^-}{\delta\alpha} \right\rfloor \delta\alpha, q \right) \\ & \quad - \left(\frac{\eta^-}{\delta\alpha} - \left\lfloor \frac{\eta^-}{\delta\alpha} \right\rfloor \right) \left(\varphi \left(t, \alpha - \left\lceil \frac{\eta^-}{\delta\alpha} \right\rceil \delta\alpha, q \right) - \varphi \left(t, \alpha - \left\lfloor \frac{\eta^-}{\delta\alpha} \right\rfloor \delta\alpha, q \right) \right). \end{aligned} \quad (2.29)$$

We define the numerical solution $h^{\delta t, \delta \alpha} : \mathbb{T}_{\delta t} \times \mathbb{R}_{\delta \alpha} \times ([-\bar{Q}, \bar{Q}] \cap \mathbb{Z}) \rightarrow \mathbb{R}$ as follows,:

$$\begin{cases} h^{\delta t, \delta \alpha}(T, \alpha, q) &= q \left(-\text{sign}(q) \Upsilon_{MO} - \psi q \right), \\ h^{\delta t, \delta \alpha}(n \delta t, \alpha, q) &= \mathcal{S}^{\delta t, \delta \alpha}(t, \alpha, q, h^{\delta t, \delta \alpha}((n+1) \delta t, \alpha, q)) . \end{cases} \quad (2.30)$$

Next we prove the convergence of (2.30). We first prove monotonicity, stability, and consistency properties (Propositions 2.5.1, 2.5.2 and 2.5.3 respectively) of $\mathcal{S}^{\delta t, \delta \alpha}(t, \alpha, q, \varphi)$. Combined with the comparison principle (Theorem 2.5.1), we prove the convergence, see Barles and Souganidis (1991).

Lemma 2.5.1. *The value function H admits the following bounds:*

$$x + q S - |q| \Upsilon_{MO} - \psi q^2 \leq H \leq x + q S + (\lambda^+ + \lambda^-) (T - t) \Upsilon_{LO} + (T - t) (A + \theta) \sigma \bar{Q}. \quad (2.31)$$

Proof. For a proof see Section 2.7. □

Proposition 2.5.1. *(Monotonicity) For any $\delta t < f(\delta \alpha)$ for some $f : [0, \infty) \rightarrow [0, \infty)$, $\varphi_1, \varphi_2 \in C_b^{1,2}([0, T] \times \mathbb{R} \times ([-\bar{Q}, \bar{Q}] \cap \mathbb{Z}))$ such that $\varphi_1 \leq \varphi_2$, we have $\mathcal{S}^{\delta t, \delta \alpha}(t, \alpha, q, \varphi_1) \leq \mathcal{S}^{\delta t, \delta \alpha}(t, \alpha, q, \varphi_2)$.*

Proof. From expression (2.26), if

$$\delta t < \left(\frac{\kappa A}{\delta \alpha} + \frac{\xi^2}{2 \delta \alpha^2} + \lambda^+ + \lambda^- \right)^{-1}, \quad (2.32)$$

then $\mathcal{T}^{\delta t, \delta \alpha}(t, \alpha, q, \varphi)$ is monotone non-decreasing in φ . The monotonicity of $\mathcal{M}^{\delta t, \delta \alpha}(t, \alpha, q, \varphi)$ is obvious. □

Proposition 2.5.2. *(Stability) For any $\delta t > 0$, there exists a unique solution $h^{\delta t, \delta \alpha}(t, \alpha, q)$ to (2.30). Furthermore, we have the uniform bounds*

$$L(q) \leq h^{\delta t, \delta \alpha}(t, \alpha, q) \leq U(t), \quad (2.33)$$

where

$$U(t) = (T - t) \left[(\lambda^+ + \lambda^-) \Upsilon_{LO} + (A + \theta) \sigma \bar{Q} \right] \quad \text{and} \quad L(q) = -|q| \Upsilon_{MO} - \psi q^2.$$

Proof. For a proof see Section 2.7. □

Proposition 2.5.3. (*Consistency*) For all $(t, \alpha, q) \in [0, T] \times \mathbb{R} \times ([-\bar{Q}, \bar{Q}] \cap \mathbb{Z})$ and $\varphi \in C_b^{1,2}([0, T] \times \mathbb{R} \times ([-\bar{Q}, \bar{Q}] \cap \mathbb{Z}))$, we have

$$\begin{aligned} & \lim_{\substack{(\delta t, \delta \alpha) \rightarrow (0, 0) \\ (t', \alpha') \rightarrow (t, \alpha)}} \frac{1}{\delta t} [\mathcal{T}^{\delta t, \delta \alpha}(t' + \delta t, \alpha', q, \varphi) - \varphi(t', \alpha', q)] \\ &= \partial_t \varphi + \bar{\alpha} \sigma q - \kappa \bar{\alpha} \partial_\alpha \varphi + \frac{1}{2} \xi^2 \partial_{\alpha\alpha} \varphi - \phi q^2 \\ &+ \lambda^+ \sup_{l^+ \in \{0, 1\}} \left[l^+ \left(\Upsilon_{LO} + \varphi(t, \alpha + \eta^+, q - 1) - \varphi \right) + (1 - l^+) \left(\varphi(t, \alpha + \eta^+, q) - \varphi \right) \right] \\ &+ \lambda^- \sup_{l^- \in \{0, 1\}} \left[l^- \left(\Upsilon_{LO} + \varphi(t, \alpha - \eta^-, q + 1) - \varphi \right) + (1 - l^-) \left(\varphi(t, \alpha - \eta^-, q) - \varphi \right) \right]; \end{aligned}$$

and

$$\lim_{\substack{(\delta t, \delta \alpha) \rightarrow (0, 0) \\ (t', \alpha') \rightarrow (t, \alpha)}} \mathcal{M}^{\delta t, \delta \alpha}(t' + \delta t, \alpha', q, \varphi) = \max \left\{ \varphi(t, \alpha, q - 1) - \Upsilon_{MO}, \quad \varphi(t, \alpha, q + 1) - \Upsilon_{MO} \right\}.$$

Proof. The limits converge by directly applying $\varphi \in C_b^{1,2}([0, T] \times \mathbb{R} \times ([-\bar{Q}, \bar{Q}] \cap \mathbb{Z}))$. \square

Theorem 2.5.2. (*Convergence*) $h^{\delta t, \delta \alpha}(t, \alpha, q)$ converges locally uniformly to the unique viscosity solution $h(t, \alpha, q)$ as $(\delta t, \delta \alpha) \rightarrow (0, 0)$, provided $\delta t < f(\delta \alpha)$ for some $f : [0, \infty) \rightarrow [0, \infty)$.

Proof. For a proof see Section 2.7. \square

2.6 Conclusions

In this chapter we derived the optimal strategy for a market maker who employs information of the drift of the asset to place limit orders and to send market orders to maximise expected profits. We refer to the drift of the asset as the alpha signal of the midprice, which follows a mean-reverting jump-diffusion process, where the jumps are caused by the arrival of market orders in the exchange.

We formulated the market maker's problem as a stochastic and impulse control problem. We solved the corresponding HJBQVI numerically and proved the convergence of the numerical scheme. We provided maximum likelihood estimates of the model parameters using Nasdaq LOB data. When the alpha signal is near zero, the market maker tends to

post both sell and buy limit orders (LOs) because the arrival rate of upward jumps in the midprice is approximately equal to the arrival rate of downward jumps in the midprice. As the alpha signal increases or decreases, the market maker posts LOs on one side of the LOB to avoid being picked off on the wrong side of the LOB. Also, when the alpha signal increases (decreases) beyond a critical level, the market maker executes buy (sell) MOs in anticipation of an increase (decrease) in the price of the asset.

The optimal strategy requires almost instantaneous reaction to the change in the alpha signal which has a half life much less than a second. Although nowadays most market participants are equipped with both software and hardware technology, there are still difficulties in the actual implementation of the optimal strategy.

Finally, we also showed that the PnL of a market maker who does not employ the alpha signal in her strategy will be lower than that of a market maker who employs the signal when the running inventory penalty parameter is low. The alpha signal is a high-frequency signal that requires computer power and speed to compute and implement the strategy in real time. Not all market participants have the latency and the hardware to execute these types of strategies, see the recent work Cartea and Sánchez-Betancourt (2018).

The alpha signal developed in this chapter is relevant to all types of trading strategies, and not only to those of market makers. For example, those who execute a large trade over a trading window, will also benefit from incorporating alpha signals in their strategy, see, e.g., Cartea and Jaimungal (2016b).

2.7 Proofs

2.7.1 Proof of Theorem 2.2.1

Proof. We follow Øksendal and Sulem (2007). Let the function $f(q) = -\phi q^2$. We define the operators \mathfrak{L}^ν and \mathfrak{M} such that the HJBQVI (2.9) is represented by the form of

$$\max \left\{ \sup_{c \in \mathcal{U}} \left[\mathfrak{L}^\nu \tilde{h} + f \right], \mathfrak{M} \tilde{h} - \tilde{h} \right\} = 0. \quad (2.34)$$

Let \tilde{h} be the solution to (2.9) and define the candidate solution $\tilde{H} = x + qS + \tilde{h}(t, \alpha, q)$. We want to show that $\tilde{H} = H$.

For any control $\nu = (l^\pm, \tau)$, where τ is the union vector of τ^\pm in increasing order, from Itô's Lemma we have

$$\begin{aligned} & \mathbb{E} \left[\tilde{H}(\tau_n^-, X_{\tau_n}^\nu, S_{\tau_n}^\nu, \alpha_{\tau_n}^\nu, Q_{\tau_n}^\nu) \middle| \mathcal{F}_t \right] - \mathbb{E} \left[\tilde{H}(\tau_{n-1}, X_{\tau_{n-1}}^\nu, S_{\tau_{n-1}}^\nu, \alpha_{\tau_{n-1}}^\nu, Q_{\tau_{n-1}}^\nu) \middle| \mathcal{F}_t \right] \\ &= \mathbb{E} \left[\int_{\tau_{n-1}}^{\tau_n} \mathfrak{L}^\nu \tilde{h}(u, \alpha_u^\nu, Q_u^\nu) du \middle| \mathcal{F}_t \right], \end{aligned}$$

and

$$\begin{aligned} & \tilde{H}(\tau_n, X_{\tau_n}^\nu, S_{\tau_n}^\nu, \alpha_{\tau_n}^\nu, Q_{\tau_n}^\nu) - \tilde{H}(\tau_n^-, X_{\tau_n}^\nu, S_{\tau_n}^\nu, \alpha_{\tau_n}^\nu, Q_{\tau_n}^\nu) \\ &= \mathfrak{M} \tilde{h}(\tau_n^-, \alpha_{\tau_n}^\nu, Q_{\tau_n}^\nu) - \tilde{h}(\tau_n^-, \alpha_{\tau_n}^\nu, Q_{\tau_n}^\nu), \end{aligned}$$

where, with a slight abuse of notation, we only include the jumps from the control. Summing over $[t, T]$, taking an expectation conditional on \mathcal{F}_t and rearranging yields

$$\begin{aligned} \tilde{H}(t, x, S, \alpha, q) &= \mathbb{E} \left[\tilde{H}(T, X_T^\nu, S_T^\nu, \alpha_T^\nu, Q_T^\nu) \middle| \mathcal{F}_t \right] \\ &\quad - \mathbb{E} \left[\int_t^T \mathfrak{L}^\nu \tilde{h}(u, \alpha_u^\nu, Q_u^\nu) du \middle| \mathcal{F}_t \right] \\ &\quad - \mathbb{E} \left[\sum_{\tau_n \leq T} \mathfrak{M} \tilde{h}(\tau_n^-, \alpha_{\tau_n}^\nu, Q_{\tau_n}^\nu) - \tilde{h}(\tau_n^-, \alpha_{\tau_n}^\nu, Q_{\tau_n}^\nu) \middle| \mathcal{F}_t \right] \end{aligned} \tag{2.35}$$

From HJBQVI (2.9) we have

$$\begin{aligned} \tilde{H}(t, x, S, \alpha, q) &\geq \mathbb{E} \left[\tilde{H}(T, X_T^\nu, S_T^\nu, \alpha_T^\nu, Q_T^\nu) - \phi \int_t^T (Q_u^\nu)^2 du \middle| \mathcal{F}_t \right] \\ &= \mathbb{E} \left[X_T^\nu + Q_T^\nu (S_T^\nu - \text{sign}(Q_T^\nu) \Upsilon_{MO} - \psi Q_T^\nu) - \phi \int_t^T (Q_u^\nu)^2 du \middle| \mathcal{F}_t \right]. \end{aligned} \tag{2.36}$$

Since the inequality above holds for any control ν , we have

$$\begin{aligned} \tilde{H}(t, x, S, \alpha, q) &\geq \sup_{\nu \in \mathcal{A}} \mathbb{E}_{t, x, S, \alpha, q} \left[X_T^\nu + Q_T^\nu (S_T^\nu - \text{sign}(Q_T^\nu) \Upsilon_{MO} - \psi Q_T^\nu) - \phi \int_t^T (Q_u^\nu)^2 du \right] \\ &= H(t, x, S, \alpha, q). \end{aligned} \tag{2.37}$$

Now if we use the optimal control ν^* , (2.35) becomes

$$\begin{aligned}
\tilde{H}(t, x, S, \alpha, q) &= \mathbb{E} \left[\tilde{H}(T, X_T^{\nu^*}, S_T^{\nu^*}, \alpha_T^{\nu^*}, Q_T^{\nu^*}) - \phi \int_t^T (Q_u^{\nu^*})^2 du \middle| \mathcal{F}_t \right] \\
&= \mathbb{E} \left[X_T^{\nu^*} + Q_T^{\nu^*} (S_T^{\nu^*} - \text{sign}(Q_T^{\nu^*}) \Upsilon_{MO} - \psi Q_T^{\nu^*}) - \phi \int_t^T (Q_u^{\nu^*})^2 du \middle| \mathcal{F}_t \right] \\
&\leq \sup_{\nu \in \mathcal{A}} \mathbb{E}_{t,x,S,\alpha,q} \left[X_T^\nu + Q_T^\nu (S_T^\nu - \text{sign}(Q_T^\nu) \Upsilon_{MO} - \psi Q_T^\nu) - \phi \int_t^T (Q_u^\nu)^2 du \right] \\
&= H(t, x, S, \alpha, q).
\end{aligned} \tag{2.38}$$

Hence $H = \hat{H}$. □

2.7.2 Proof of Lemma 2.5.1

Proof.

$$\begin{aligned}
&H(t, x, S, \alpha, q) \\
&= \sup_{\nu \in \mathcal{A}} \mathbb{E}_{t,x,S,\alpha,q} \left[X_T^\nu + Q_T^\nu (S_T^\nu - \text{sign}(Q_T^\nu) \Upsilon_{MO} - \psi Q_T^\nu) - \phi \int_t^T (Q_s^\nu)^2 ds \right] \\
&= x + q S \\
&\quad + \sup_{\nu \in \mathcal{A}} \mathbb{E}_{t,x,S,\alpha,q} \left[\int_t^T dX_u^\nu + \int_t^T S_u^\nu dQ_u^\nu + \int_t^T Q_u^\nu dS_u^\nu - |Q_T^\nu| \Upsilon_{MO} - \psi (Q_T^\nu)^2 \right. \\
&\quad \left. - \phi \int_t^T (Q_s^\nu)^2 ds \right] \\
&= x + q S \\
&\quad + \sup_{\nu \in \mathcal{A}} \mathbb{E}_{t,x,S,\alpha,q} \left[\int_t^T (S_u^\nu + \Upsilon_{LO}) dN_u^{+,l} - \int_t^T (S_u^\nu - \Upsilon_{LO}) dN_u^{-,l} \right. \\
&\quad - \int_t^T (S_u^\nu + \Upsilon_{MO}) dM_u^+ + \int_t^T (S_u^\nu - \Upsilon_{MO}) dM_u^- \\
&\quad + \int_t^T S_u^\nu dN_u^{-,l} - \int_t^T S_u^\nu dN_u^{+,l} + \int_t^T S_u^\nu dM_u^+ - \int_t^T S_u^\nu dM_u^- \\
&\quad \left. + \int_t^T Q_u^\nu dS_u^\nu - |Q_T^\nu| \Upsilon_{MO} - \psi (Q_T^\nu)^2 - \phi \int_t^T (Q_s^\nu)^2 ds \right].
\end{aligned}$$

After cancelling terms, we have

$$\begin{aligned}
H(t, x, S, \alpha, q) &= x + q S \\
&+ \sup_{\nu \in \mathcal{A}} \mathbb{E}_{t, x, S, \alpha, q} \left[\Upsilon_{LO} (N_T^{+,l} - N_t^{+,l}) + \Upsilon_{LO} (N_T^{-,l} - N_t^{-,l}) \right. \\
&\quad - \Upsilon_{MO} (M_T^+ - M_t^+) - \Upsilon_{MO} (M_T^- - M_t^-) \\
&\quad \left. + \int_t^T Q_u^\nu dS_u^\nu - |Q_T^\nu| \Upsilon_{MO} - \psi (Q_T^\nu)^2 - \phi \int_t^T (Q_s^\nu)^2 ds \right].
\end{aligned}$$

For the lower bound, we first restrict the set of admissible strategies such that the market maker clears out her positions with MOs immediately. In this case, we have

$$H(t, x, S, \alpha, q) \geq x + q S - |q| \Upsilon_{MO} - \psi q^2.$$

For the upper bound, we first drop all the non-positive terms, and obtain

$$\begin{aligned}
H(t, x, S, \alpha, q) &\leq x + q S + \sup_{\nu \in \mathcal{A}} \mathbb{E}_{t, x, S, \alpha, q} \left[\Upsilon_{LO} (N_T^{+,l} - N_t^{+,l}) + \Upsilon_{LO} (N_T^{-,l} - N_t^{-,l}) + \int_t^T Q_u^\nu dS_u^\nu \right] \\
&\leq x + q S + (\lambda^+ + \lambda^-) (T - t) \Upsilon_{LO} + (T - t) (A + \theta) \sigma \bar{Q}.
\end{aligned}$$

□

2.7.3 Proof of Proposition 2.5.2

Proof. Existence and uniqueness follows immediately from the definition of the explicit scheme (2.30). We first prove the upper bound. We notice that $h^{\delta t, \delta \alpha}(T, \alpha, q) = q (-\text{sign}(q) \Upsilon_{MO} - \psi q) \leq 0 = U(T)$, and

$$\begin{aligned}
\mathcal{T}^{\delta t, \delta \alpha}(t, \alpha, q, U) &= U(t) + \delta t \left\{ \bar{\alpha} \sigma q - \phi q^2 + (\lambda^+ + \lambda^-) \Upsilon_{LO} \right\} \\
&\leq U(t) + \delta t \left\{ (A + \theta) \sigma \bar{Q} + (\lambda^+ + \lambda^-) \Upsilon_{LO} \right\} \\
&= U(t - \delta t),
\end{aligned}$$

and

$$\mathcal{M}^{\delta t, \delta \alpha}(t, \alpha, q, U) = U(t) - \Upsilon_{MO} \leq U(t - \delta t).$$

Thus, we have

$$\mathcal{S}^{\delta t, \delta \alpha}(t, \alpha, q, U) \leq U(t - \delta t).$$

As $h^{\delta t, \delta \alpha}(t - \delta t, \alpha, q) = \mathcal{S}^{\delta t, \delta \alpha}(t, \alpha, q, h^{\delta t, \delta \alpha}(t, \alpha, q))$, we can prove by induction that $h^{\delta t, \delta \alpha}(t, \alpha, q) \leq U(t)$.

The lower bound is attained by the strategy that immediately clears out any position and then does nothing until time T . \square

2.7.4 Proof of Theorem 2.5.2

Proof. Follows Theorem 2.5.1, Propositions 2.5.1, 2.5.2, 2.5.3, and Barles and Souganidis (1991).

We define

$$h_*(t, \alpha, q) = \liminf_{\substack{(\delta t, \delta \alpha) \rightarrow (0, 0) \\ (t', \alpha') \rightarrow (t, \alpha)}} h^{\delta t}(t, \alpha, q) \quad \text{and} \quad h^*(t, \alpha, q) = \limsup_{\substack{(\delta t, \delta \alpha) \rightarrow (0, 0) \\ (t', \alpha') \rightarrow (t, \alpha)}} h^{\delta t}(t, \alpha, q),$$

which are, respectively, lower and upper semi-continuous functions on $[0, T] \times \mathbb{R} \times ([-\bar{Q}, \bar{Q}] \cap \mathbb{Z})$, and inherit the boundedness of $h^{\delta t, \delta \alpha}(t, \alpha, q)$ by stability from Proposition 2.5.2. By definition, we have $h_* \leq h^*$. We claim that h_* and h^* are, respectively, a viscosity supersolution and a viscosity subsolution of (2.23), then by Theorem 2.5.1 (Comparison principle) we have $h^* \leq h_*$ and hence the equality. By symmetry, it suffices to show the viscosity supersolution property of h_* .

Let $(\tilde{t}, \tilde{\alpha}, \tilde{q}) \in [0, T] \times \mathbb{R} \times ([-\bar{Q}, \bar{Q}] \cap \mathbb{Z})$ and $\varphi \in C_b^{1,2}([0, T] \times \mathbb{R} \times ([-\bar{Q}, \bar{Q}] \cap \mathbb{Z}))$ such that $(\tilde{t}, \tilde{\alpha}, \tilde{q})$ attains the strict global minimum of $h_* - \varphi$. Then there exists a sequence $\{(t'_k, \alpha'_k, q'_k)\}_k \in [0, T] \times \mathbb{R} \times ([-\bar{Q}, \bar{Q}] \cap \mathbb{Z})$ and $\{(\delta t_k, \delta \alpha_k)\}_k$ such that

$$\begin{aligned} (t'_k, \alpha'_k, q'_k) &\rightarrow (\tilde{t}, \tilde{\alpha}, \tilde{q}), \\ (\delta t_k, \delta \alpha_k) &\rightarrow (0, 0), \\ h^{\delta t_k, \delta \alpha_k} &\rightarrow h_*(\tilde{t}, \tilde{\alpha}, \tilde{q}), \end{aligned}$$

and (t'_k, α'_k, q'_k) is the global minimizer of $h^{\delta t_k, \delta \alpha_k} - \varphi$.

We restrict $(\delta t_k, \delta \alpha_k)$ to satisfy the condition in Proposition 2.5.1, so that we can apply the monotonicity of $\mathcal{S}^{\delta t, \delta \alpha}$.

Define $\varepsilon_k = (h^{\delta t_k, \delta \alpha_k} - \varphi)(t'_k, \alpha'_k, q'_k)$. Then by the numerical scheme in (2.30) and the monotonicity from Proposition 2.5.1, we have

$$\begin{aligned}
& \varepsilon_k + \varphi(t'_k, \alpha'_k, q'_k) \\
&= h^{\delta t_k, \delta \alpha_k}(t'_k, \alpha'_k, q'_k) \\
&= \mathcal{S}^{\delta t_k, \delta \alpha_k}(t'_k + \delta t_k, \alpha'_k, q'_k, h^{\delta t_k, \delta \alpha_k}) \\
&\geq \mathcal{S}^{\delta t_k, \delta \alpha_k}(t'_k + \delta t_k, \alpha'_k, q'_k, \varphi + \varepsilon_k) \\
&= \mathcal{S}^{\delta t_k, \delta \alpha_k}(t'_k + \delta t_k, \alpha'_k, q'_k, \varphi) + \varepsilon_k \\
&= \max \left\{ \mathcal{T}^{\delta t_k, \delta \alpha_k}(t'_k + \delta t_k, \alpha'_k, q'_k, \varphi), \mathcal{M}^{\delta t_k, \delta \alpha_k}(t'_k + \delta t_k, \alpha'_k, q'_k, \varphi) \right\} + \varepsilon_k.
\end{aligned}$$

After rearranging,

$$\max \left\{ \frac{1}{\delta t_k} \left[\mathcal{T}^{\delta t_k, \delta \alpha_k}(t'_k + \delta t_k, \alpha'_k, q'_k, \varphi) - \varphi(t'_k, \alpha'_k, q'_k) \right], \right. \\
\left. \left[\mathcal{M}^{\delta t_k, \delta \alpha_k}(t'_k + \delta t_k, \alpha'_k, q'_k, \varphi) - \varphi(t'_k, \alpha'_k, q'_k) \right] \right\} \leq 0.$$

Apply the consistency in Proposition 2.5.3 and let $k \rightarrow \infty$, to obtain

$$\begin{aligned}
& \max \left\{ \partial_t \varphi + \bar{\alpha} \sigma q - \kappa \bar{\alpha} \partial_\alpha \varphi + \frac{1}{2} \xi^2 \partial_{\alpha\alpha} \varphi - \phi q^2 \right. \\
& \quad + \lambda^+ \sup_{l^+ \in \{0,1\}} \left[l^+ \left(\Upsilon_{LO} + \varphi(t, \alpha + \eta^+, q - 1) - \varphi \right) + (1 - l^+) \left(\varphi(t, \alpha + \eta^+, q) - \varphi \right) \right] \\
& \quad + \lambda^- \sup_{l^- \in \{0,1\}} \left[l^- \left(\Upsilon_{LO} + \varphi(t, \alpha - \eta^-, q + 1) - \varphi \right) + (1 - l^-) \left(\varphi(t, \alpha - \eta^-, q) - \varphi \right) \right]; \\
& \quad - \Upsilon_{MO} + \varphi(t, \alpha, q - 1) - \varphi; \\
& \quad \left. - \Upsilon_{MO} + \varphi(t, \alpha, q + 1) - \varphi \right\} \leq 0, \\
& \hspace{25em} (2.39)
\end{aligned}$$

which is the viscosity supersolution property, as desired. Therefore, $h^* \leq h_*$, and hence $h^* = h_*$. \square

Chapter 3

Spoofing and price manipulation in order driven markets

3.1 Introduction

In order-driven markets, ‘spoofing’ and ‘layering’ are strategies that provide false information about the demand and supply of an asset. These trading strategies are illegal and profit from market participants who trade on misleading market signals. The Dodd-Frank Act describes spoofing as “bidding or offering with the intent to cancel the bid or offer before execution,” see Dodd-Frank (2010). Similar to spoofing, layering consists in submitting a relatively large number of orders to one side of the limit order book (LOB) to precipitate an unwarranted change in the price of the asset. Layering and spoofing may cause prices to change because the market interprets the one-sided pressure in the LOB as a shift in the balance of the number of investors who wish to purchase or sell the asset, which causes prices to increase (more buyers than sellers) or prices to decline (more sellers than buyers).¹

Spoofing and layering are very difficult to detect. These strategies are camouflaged behind the vast number of updates in the LOB and use sophisticated automated algorithms to avoid detection. Nevertheless, regulators and financial authorities have successfully prosecuted many traders for market abuse strategies and price manipulation as a result of spoofing and layering, e.g., see Agency for the Cooperation of Energy Regulators (2019b).²

¹See <https://www.fca.org.uk/news/statements/statement-regarding-swift-trade-court-appeal-judgment>.

²See also the cases described in the U.S. Commodity Futures Trade Commission (www.cftc.gov) and the UK’s Financial Conduct Authority (www.fca.org.uk).

We show how an investor employs spoof limit orders (LOs) to manipulate the volume imbalance of the LOB to trade at more advantageous prices. We compute volume imbalance of the LOB as the ratio of volume posted at the best bid price minus the volume posted at the best ask price to the sum of the volume at the best bid and best ask prices. We assume that volume imbalance follows a Markov chain to describe three regimes of the LOB: buy-heavy, neutral, and sell-heavy. For a selection of stocks traded in Nasdaq, we show that the arrival rate of buy market orders (MOs) is highest in the buy-heavy regime, second-highest in the neutral regime, and lowest in the sell-heavy regime. The opposite pattern is found for the arrival rate of sell MOs, see Cartea et al. (2018).

The investor’s objective is to maximise the expected profit and loss (PnL) from liquidating a position in shares. The strategy employs LOs and MOs to trade shares, and penalises inventory holdings throughout the trading horizon. Although the objective is to liquidate shares, the investor may employ spoof buy LOs to tilt the volume imbalance in the LOB from sell-heavy and neutral, into the buy-heavy regime — if the LOB is already buy-heavy there is no need to spoof the book. When the book is tilted into the buy-heavy regime, the arrival rate of buy MOs increases and the price of the asset exhibits a positive trend.

In our model, the investor pays a fine if the financial authorities detect spoofing in the LOB and manipulation of prices. The investor’s strategy trades off the gains that originate from spoofing against the expected financial losses due to the fine. As the expected value of the fine increases, the investor relies less on spoofing, and if the expected fine is large enough, it is optimal for the investor not to spoof the LOB because the fine outweighs the benefits from spoofing.

Compared with a no-spoof strategy, when the fine for spoofing is low, we show that the investor’s spoof strategy liquidates (on average) more shares using sell LOs and relies on fewer MOs to sell shares and to stay on target throughout the trading horizon. Sell MOs receive worse prices than sell LOs because the former are executed at the midprice minus half the quoted spread and incur liquidity taking fees, while the latter are filled at the midprice plus half the quoted spread and liquidity making rebates.

In addition to the risk of being prosecuted by financial authorities, spoofers bear the risk that their spoof buy LOs could be filled and there may be little time left to employ sell LOs to unwind the ‘inadvertently’ acquired shares. The probability that a spoof buy LO receives full or partial execution depends on the arrival rate of sell MOs when the LOB is

in the buy-heavy regime and on the volume of the spoof buy LO posted at the best bid price to tilt the book. For example, a sell-heavy book requires more spoof buy LOs than a book in the neutral regime to tilt it into the buy-heavy regime.

We use Nasdaq high-frequency data to estimate the parameters of the model and use simulations to analyse the performance of the optimal liquidation strategy for an investor who spoofs the LOB. In one of the examples we provide, we assume that the investor needs to liquidate 30 lots of shares of Intel Corporation, ticker symbol INTC, over a period of 300 seconds. Each lot consists of 300 shares, which is approximately the average volume the LOs posted at the best bid in the LOB of INTC. If the investor does not spoof the LOB, the strategy employs, on average, 24 lots of sell MOs and 6 lots of (filled) sell LOs to liquidate the position.

In the extreme case where there is no penalty for spoofing the book, and market participants believe that the buy-heavy LOB conveys truthful information about demand and supply of the asset, the spoofing strategy liquidates the initial position with an average of: 15 lots of sell MOs, 24 lots of (filled) sell LOs, and 9 lots of (filled) spoof buy LOs. In this example, the difference between the mean PnL received by the investor when the strategy spoofs the LOB and when it does not spoof the book to liquidate an initial target of 9,000 shares is $\$12.13 \times 300$. This extra revenue, which stems from employing spoof LOs in the strategy, is approximately $1,213 \times 300$ times the quoted spread in INTC.

In general, the PnL of the spoof strategy is higher than that of a no-spoof strategy for two reasons. First, the investor employs fewer MOs to draw the inventory to zero and benefits from roundtrip trades which stem from spoof buy LOs that are ‘inadvertently’ filled and subsequently unwound with sell LOs. Second, the midprice trends upward when the book is buy-heavy, therefore, as time evolves, the spoofer sells the asset at better prices (on average).

The spoofing strategy deviates the price of the asset from its fundamental value. Our simulations show that once the investor finalises the liquidation programme, which lasts five minutes, the mean price of the asset is approximately 52 cents higher than the mean price of the asset in the absence of spoofing. The effects of manipulating the price with spoof LOs are expected to subside, so the price of the asset will return to its fundamental value. In the meantime, however, market participants make trading decisions on distorted

information about the price of the asset, which may also have a knock-on effect on other assets in the marketplace.

The literature on spoofing is scant. Allen and Gale (1992) propose a simple model for trade-based stock-price manipulation from a uninformed speculator. They show that manipulation is profitable. Lee et al. (2013) define a spoof LO order as an order resting in the book at least 6 ticks away from the market price (in our analysis, the investor posts spoof LOs at the best bid) and the volume of the spoof order is at least twice as large as the average volume of the LOs posted the previous day. They use a proprietary data set with account information from the Korea Exchange. They show empirically that the spoof orders create the impression of imbalance in the LOB, which moves the price, and also show that the probability of filling a spoof order is very low. They also show empirically that spoofing achieves substantial extra profits and that spoofing tends to target stocks with: higher volatility of returns, lower market capitalisation, lower price level, and lower managerial transparency. Wang (2015) uses data from the Taiwan Futures Exchange to show that market participants spoof the LOB in the stocks that exhibit high volumes of trading, high volatility, and high prices. The author also shows that spoofing increases the volume of trading, increases the volatility of prices, and increases the quoted spread. In a recent article, the Agency for the Cooperation of Energy Regulators (2019a) investigates spoofing and layering in wholesale energy markets and suggests a list of indicators to identify these illegal activities. There are other studies that look at price manipulation in other contexts, see for example Alfonsi and Acevedo (2014) and Klöck et al. (2017).

Finally, there are studies that employ machine learning algorithms to detect manipulation in markets. Cao et al. (2014) use k-Nearest Neighbour (kNN) and One Class Support Vector Machine (OCSVM) on transformed LOB data to detect price manipulation. They show that most non-stationary features of the data can be removed by transformations based on techniques of time series analysis such as the autoregressive integrated moving average model. Their results indicate that kNN and OCSVM work effectively to detect spoofing and the transformations of the data also improve the power of machine learning algorithms to detect spoofing — the performance of detection is measured in terms of area under Receiver Operating Characteristic (ROC) curve. Cao et al. (2015) use Adaptive Hidden Markov Model with Anomaly States (AHMMAS) to detect manipulation activities in level 2 data from Nasdaq and the London Stock Exchange, and compare the performance of detecting price manipulation with standard machine learning algorithms

such as Gaussian Mixture Models, kNN and OCSVM. They find that AHMMAS performs better in terms of the area under the ROC curve and the F-measure.

The remainder of this chapter proceeds as follows. In Section 3.2 we use market data to show the relationship between volume imbalance and the arrival of MOs and LOs. In Section 3.3 we introduce the model. In Section 3.4 we solve the optimisation problem of the investor and derive the optimal spoofing strategy. In Section 3.5 we employ simulations to illustrate the performance of the optimal strategy. We introduce the numerical scheme and prove the convergence results in Section 3.6. Finally, we conclude in Section 3.7 and collect proofs in Section 3.8.

3.2 Limit order book and volume imbalance

In this section we provide a measure of volume imbalance in the LOB and discuss its relationship with the arrival rates and volumes of MOs and LOs. We employ Nasdaq data for October 2017 to compute arrival rates and average sizes of MOs and LOs for 10 stocks. We remove the first and the last 30 minutes of each trading day to exclude the behaviour of the LOB during the opening and closing auctions of each trading day.

3.2.1 Volume imbalance

We define volume imbalance in the LOB at time t as

$$\rho_t = \frac{V_t^b - V_t^a}{V_t^b + V_t^a} \in (-1, 1), \quad (3.1)$$

where $V_t^b > 0$ and $V_t^a > 0$ are the volumes at time t of LOs posted at the best bid and the best ask, respectively. Volume imbalance is a key quantity because it summarises the willingness of agents to buy or sell assets using LOs. Clearly, when ρ_t is close to 1 there is strong buy pressure and when it is close to -1 there is strong sell pressure. Figure 3.1 shows the path of ρ_t for stock INTC during a window of 2 minutes. Figure 3.2 shows the number of MOs (buy and sell) for different levels of volume imbalance for INTC during October 2017. We observe that as the absolute value of volume imbalance increases, the number of MOs increases.

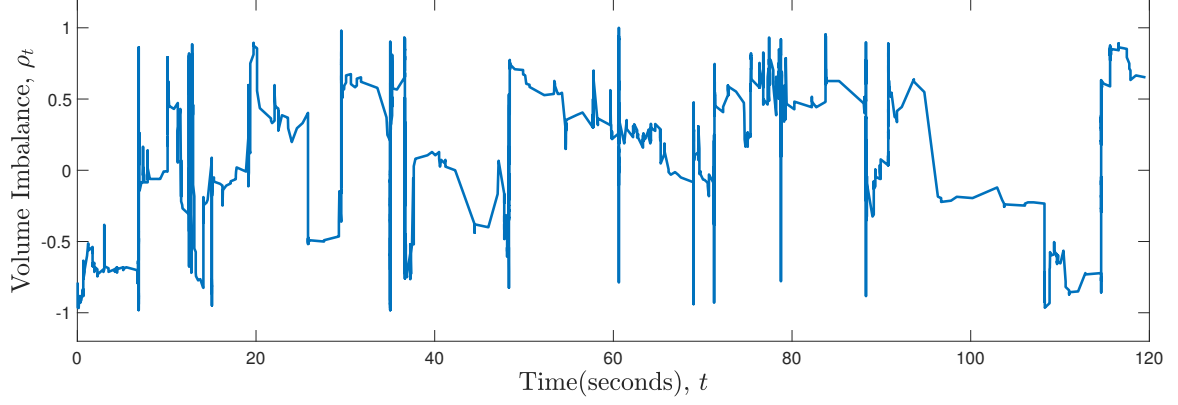


Figure 3.1: Volume imbalance ρ_t of INTC from 10:00:00 to 10:02:00 on October 2 2017.

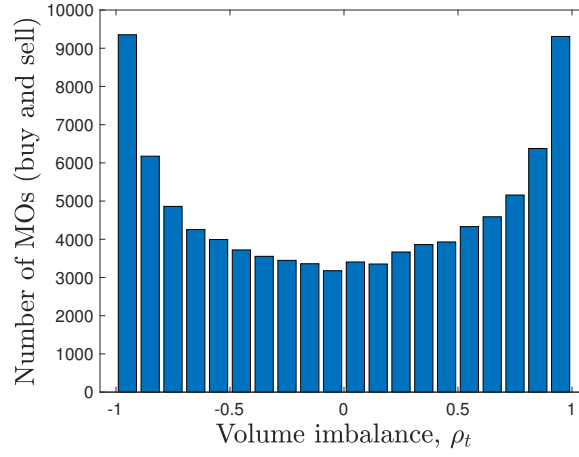


Figure 3.2: Number of MOs for various levels of volume imbalance ρ_t of INTC. Data: Nasdaq October 2017.

Cartea et al. (2018) show that the volume imbalance measure (3.1) helps to predict the rate of incoming MOs and to predict the direction and magnitude of price movements that follow the arrival of a MO. We divide the range of the volume imbalance measure $(-1, 1)$ into three subintervals. We refer to the interval $(1/3, 1)$ as the buy-heavy regime, the interval $[-1/3, 1/3]$ as the neutral regime, and the interval $(-1, -1/3)$ as the sell-heavy regime. When volume imbalance is buy-heavy (sell-heavy) and a MO arrives, there is a high probability that this MO is a buy (sell) order. Furthermore, immediately following a buy (sell) MO, the magnitude and sign of the change in the midprice is, on average, positive (negative) when volume imbalance is buy-heavy (sell-heavy). Table 3.1 shows the arrival rates of changes in the midprice conditioned on the volume imbalance regimes for 10 stocks in October 2017. For most stocks, the rate of midprice increase is highest in the buy-heavy regime and lowest in the sell-heavy regime, which indicates the predictive

power of volume imbalance on price movements. For example, for the stock INTC in the buy-heavy regime, the arrival intensity of positive jumps in prices is 0.112 per second.

	Arrival rates (per second) Midprice increase			Arrival rates (per second) Midprice decrease		
	Buy-heavy	Neutral	Sell-heavy	Buy-heavy	Neutral	Sell-heavy
INTC	0.112	0.407	0.044	0.048	0.412	0.096
AAPL	1.207	1.436	0.609	0.635	1.413	1.133
BIDU	0.543	0.339	0.389	0.417	0.343	0.525
MSFT	0.164	3.203	0.080	0.084	3.207	0.160
AMZN	0.861	0.659	0.607	0.610	0.624	0.793
GOOG	0.508	0.456	0.548	0.489	0.448	0.540
NVDA	1.724	1.302	1.016	1.027	1.081	1.897
CSCO	0.030	0.134	0.014	0.013	0.118	0.032

Table 3.1: Arrival rates of midprice increase/decrease (per second). Data: Nasdaq October 2017, with first and last 30 minutes of each trading day removed.

3.2.2 Market order activity

Table 3.2 shows the arrival rates (per second) of buy and sell MOs in each regime of volume imbalance for 10 stocks in October 2017. For most stocks, the arrival rate of buy MOs is highest in the buy-heavy regime and lowest in the sell-heavy regime. Similarly, for most stocks, the arrival rate of sell MOs is highest in the sell-heavy regime and lowest in the buy-heavy regime. When the LOB is buy-heavy (sell-heavy), the arrival rate of buy (sell) MOs is highest because some traders anticipate an increase (decrease) in prices, see Cartea et al. (2018). So, rather than posting LOs in the book, they send buy (sell) MOs to ensure they receive immediate execution at the best ask (bid) price ‘before’ prices increase (decrease).

	Buy MO arrival rates (per second)			Sell MO arrival rates (per second)		
	Buy-heavy	Neutral	Sell-heavy	Buy-heavy	Neutral	Sell-heavy
INTC	0.270	0.051	0.047	0.040	0.052	0.312
AAPL	0.631	0.225	0.174	0.152	0.219	0.653
BIDU	0.110	0.076	0.135	0.168	0.090	0.124
MSFT	0.402	0.089	0.077	0.066	0.094	0.490
AMZN	0.196	0.144	0.206	0.237	0.147	0.171
GOOG	0.078	0.062	0.115	0.106	0.062	0.079
NVDA	0.299	0.215	0.273	0.277	0.215	0.300
CSCO	0.138	0.021	0.015	0.016	0.019	0.111

Table 3.2: Arrival rates of MOs (per second). Data: Nasdaq October 2017, with first and last 30 minutes of each trading day removed.

Table 3.3 shows the average volume of MOs in each regime of volume imbalance. The average volume of a buy MO in the buy-heavy regime is less than the average volume in the sell-heavy regime. This shows that although liquidity takers increase the rate at which they send buy MOs when volume imbalance is buy-heavy, the average size of the buy MOs is smaller than the average size of the sell MOs when the LOB is neutral or sell-heavy. We find similar results for sell MOs. The last column in the table shows the average daily volume of traded, buy and sell, MOs.

	Buy MO average volume			Sell MO average volume			Avg. daily volume
	Buy-heavy	Neutral	Sell-heavy	Buy-heavy	Neutral	Sell-heavy	
INTC	414	904	734	703	872	392	2,353,814
AAPL	114	218	315	390	231	123	2,363,229
BIDU	47	95	135	138	96	44	416,692
MSFT	232	449	456	480	456	230	2,090,697
AMZN	25	59	100	95	60	24	464,922
GOOG	21	53	68	68	53	19	163,120
NVDA	53	109	242	236	111	50	1,303,358
CSCO	562	927	684	813	1064	539	1,136,837

Table 3.3: Average volume of MOs. Data: Nasdaq October 2017, with first and last 30 minutes of each trading day removed.

3.2.3 Limit order activity

Table 3.4 shows the arrival rates of LOs (per second) in each regime of volume imbalance. For most stocks, the arrival rate of buy LOs is highest in the buy-heavy regime and lowest in the sell-heavy regime. Similarly, for most stocks, the arrival rate of sell LOs is highest in the sell-heavy regime and lowest in the buy-heavy regime. These results are similar to those described above for the arrival rate of MOs, however the arrival rates of LOs are larger than the arrival rates of MOs.

Table 3.5 shows the average size of LOs in each regime of volume imbalance. For each stock, the average LO size remains approximately the same in the three regimes. The average size of the orders can be thought of as one lot of shares when we develop the mathematical framework below in Section 3.3.

3.2.4 Volumes at best prices and spoofing

The previous two subsections showed that volume imbalance conveys important information about market activity of liquidity providers and takers. In most electronic markets it

	Buy LO arrival rates (per second)			Sell LO arrival rates (per second)		
	Buy-heavy	Neutral	Sell-heavy	Buy-heavy	Neutral	Sell-heavy
INTC	5.00	2.50	2.21	1.98	2.48	5.36
AAPL	5.22	4.04	4.37	4.02	4.16	5.97
BIDU	0.43	0.34	0.42	0.40	0.30	0.38
MSFT	5.83	3.07	3.46	2.89	3.12	6.93
AMZN	0.62	0.53	0.47	0.49	0.52	0.54
GOOG	0.44	0.41	0.53	0.43	0.39	0.41
NVDA	2.39	2.35	2.28	1.21	1.20	1.19
CSCO	2.98	1.32	0.97	1.01	1.29	2.71

Table 3.4: Arrival rates of LOs (per second). Data: Nasdaq October 2017, with first and last 30 minutes of each trading day removed.

	Buy LO average volume			Sell LO average volume		
	Buy-heavy	Neutral	Sell-heavy	Buy-heavy	Neutral	Sell-heavy
INTC	320	313	327	347	319	310
AAPL	163	151	150	148	151	167
BIDU	72	75	59	65	80	78
MSFT	140	139	142	137	136	138
AMZN	59	58	47	49	56	55
GOOG	64	60	55	53	56	55
NVDA	56	57	55	79	82	81
CSCO	423	379	419	426	378	410

Table 3.5: Average volume of LOs. Data: Nasdaq October 2017, with first and last 30 minutes of each trading day removed.

is almost costless to cancel or amend LOs, hence a trader can intentionally manipulate the volume imbalance in the LOB by posting LOs on one side of the LOB to take advantage of traders who use volume imbalance as one of the inputs in their trading strategies.

For example, suppose an investor wants to sell stocks. She could sell the stocks by sending MOs to the exchange, which guarantee immediate execution and pay the costs of crossing the spread plus other fees. Alternatively, although illegal, the investor can spoof the LOB. The trading strategy is based on the following two actions. (i) Post sell LOs at the best ask price in the LOB. (ii) Post spoof buy LOs on the best bid. The volumes of the spoof buy LOs must be large enough to tilt, momentarily, the LOB into the buy-heavy regime. Some traders will interpret this shift in regime as a signal of a likely increase in the price of the stock, so they send buy MOs that are likely to be filled by the spoofer's LOs resting on the sell side on the book.

Table 3.6 shows the time-weighted average volumes posted at the best bid and best ask prices in each regime of volume imbalance for the period October 2017 (first and last 30 minutes of each trading day removed). The information in the table provides a rough

idea of the volume of spoof orders required to tilt the LOB. For example, to spoof the bid side of the book of INTC from sell-heavy to buy-heavy the investor must post spoof LOs that add up to approximately 10,154 shares.

	Best bid average volume			Best ask average volume		
	Buy-heavy	Neutral	Sell-heavy	Buy-heavy	Neutral	Sell-heavy
INTC	6,165	3,376	1,513	1,660	3,313	5,833
AAPL	879	436	217	189	438	1,143
BIDU	323	162	47	68	160	271
MSFT	2,598	1,519	735	742	1,473	3,130
AMZN	230	115	26	27	113	225
GOOG	204	116	24	25	114	173
NVDA	535	177	76	81	178	570
CSCO	8,360	4,304	1,948	2,260	4,326	7,129

Table 3.6: Time-weighted average volume posted at the best bid and best ask prices. Data: Nasdaq October 2017, with first and last 30 minutes of each trading day removed.

3.3 The model

In this section we present a model for an investor who wants to liquidate a certain amount of inventory of a traded stock. Although spoofing is illegal, we assume that the investor is willing to spoof the LOB to improve the PnL from liquidating her inventory. We work on a completed filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F})_{t \geq 0}, \mathbb{P})$, where the filtration is the natural filtration generated by the collection of observable stochastic processes that we define below.

The finite state imbalance regime process without spoofing is $Z_t \in \{1, 2, 3\}$, which represents $\{buy-heavy, neutral, sell-heavy\}$ respectively, and we assume that Z is a continuous-time Markov chain with constant generator matrix G .

The midprice of the asset is denoted by $S = (S_t)_{t \geq 0}$ and is given by

$$S_t = S_0 + \sigma (J_t^+ - J_t^-) , \quad (3.2)$$

where $J^\pm = (J_t^\pm)_{t \geq 0}$ are conditionally independent doubly stochastic Poisson processes, with regime-dependent arrival rates $\gamma^\pm(Z_t) > 0$ when there is no spoofing and $\sigma > 0$ denotes the size of the tick in the LOB, which is assumed to be constant. The spread between the best bid and best ask prices is constant at 2Δ and we refer to $\Delta > 0$ as the half-spread. One could consider a model where the spread is also stochastic, see Cartea

et al. (2018). For simplicity, we assume the spread is constant at one tick, which is the case of the large-stick stocks such as the ones we study here.

3.3.1 Liquidating shares without spoofing

The investor wishes to sell $\mathfrak{N} > 0$ shares before or by a fixed time horizon $T > 0$. The investor can post sell LOs at the price $S_t + \Delta$, and wait for buy MOs from other market participants to trade with her LOs resting in the book. Every time a sell LO is filled, the investor receives $S_t + \Delta + \varepsilon_{LO}$, which is the best ask price in the LOB plus the liquidity making rebate $\varepsilon_{LO} > 0$. We denote the total effective half-spread of a LO by $\Upsilon_{LO} := \Delta + \varepsilon_{LO}$. Also, the investor can aggressively cross the spread by submitting sell MOs. We assume there is enough liquidity in the best bid of the LOB to fill the investor's sell MO. The investor's market sell order receives the cash $S - \Delta - \varepsilon_{MO}$ per share, which is the best bid price in the LOB minus the liquidity taking fee $\varepsilon_{MO} \geq 0$. We denote the total effective half-spread of a MO by $\Upsilon_{MO} := \Delta + \varepsilon_{MO}$. The volume of the investor's sell LOs and the volume of the investor's sell MOs are one — e.g., one share or one lot of shares.

The arrival rate of MOs is assumed to be constant in each regime of volume imbalance. We denote by $\lambda^+(Z_t)$ the arrival rate of buy MOs, where $\lambda^+(1) > \lambda^+(2) > \lambda^+(3)$ — see for example Table 3.2. Finally, when a buy MO arrives, the sell LO of the investor is filled with probability p_{sell} , which we assume to be constant.

3.3.2 Liquidating shares with spoof buy LOs

In addition to the strategy described above, the investor can also include spoof orders as part of her trading strategy. That is, the investor employs a combination of: sell LOs at the best ask price, sell MOs, and spoof buy LOs at the best bid price $S_t - \Delta$ and recall that every time the spoof buy LO is filled, the investor receives a liquidity making rebate of ε_{LO} per share. The volume of the spoof buy LOs must be large enough to tilt the book into the buy-heavy regime. This ‘phantom liquidity’ on the buy side of the LOB makes traders believe that volume imbalance is buy-heavy. This entices other market participants to post buy LOs (see Table 3.4) and the more impatient traders will send buy MOs in anticipation of an increase in prices, thus the arrival rate of buy MOs also increases (see Table 3.2). Consequently, the fill rate of the sell LOs posted by the investor

increases, so the investor expects to liquidate her inventory of shares more quickly and at better prices. The sizes of the spoof buy LOs depend on the volume required to tilt the LOB into buy-heavy regime, see Table 3.6.

We denote the arrival rate of buy MOs in a LOB with spoof buy LOs by λ^s . We assume that the rate is constant and is equal to the arrival rate of buy MOs when the book is buy-heavy, i.e., $\lambda^s = \lambda^+(1)$. We note that if volume imbalance without spoofing is already in the buy-heavy regime, the investor does not need to spoof the LOB because the arrival rate of buy MOs is already highest.

When the investor spoofs the buy side of the LOB, the arrival rates of J^\pm (i.e., arrival of price innovations) become $\gamma^\pm(1)$ because the market believes that volume imbalance is buy-heavy. Therefore, market participants change their strategies in anticipation to an increase in the price of the asset. That is, some traders send buy MOs to purchase the asset and other traders reshuffle their LOs in the book. Both strategies are self-fulfilling because they exert an upward pressure in the price of the asset. This is an extreme example where the spoofer misleads market participants and the price of the asset deviates from its ‘true’ price — in an efficient market the price of the asset will eventually return to its true value.

The investor bears the risk that an incoming sell MO from another market participant hits her spoof buy LOs, which is against the investor’s goal of selling shares in the stock. We denote by λ^- the arrival rate of sell MOs in the buy-heavy regime. When a sell MO arrives, V shares in the spoof buy LOs of the investor are filled with probability $p_{buy}(Z_t)$. The value of $p_{buy}(Z_t)$ is determined by the volume of the spoof LOs, which itself depends on the imbalance regime of the LOB immediately before tilting the book with spoof orders. Clearly, the proportion of spoof LOs in the best bid queue is highest when the LOB was sell-heavy immediately before the investor sends the spoof orders. Thus, the probability that a spoof LO is filled is such that $p_{buy}(2) < p_{buy}(3)$, which is generally borne out by the data (recall that regime 3 is when the book is sell-heavy). Finally, the investor mitigates the risk that her spoof buy LOs are filled by setting an upper bound, denoted by \bar{Q} which is greater than \mathfrak{N} , on the inventory she is willing to hold.

3.4 Investor's optimisation problem

The control process $c = (c_t)_{t \geq 0}$ takes values in $\{0, 1\}$, where $c_t = 1$ denotes that the investor spoofs the LOB and $c_t = 0$ represents no spoofing. The investor cannot spoof the book when her inventory is greater than $\bar{Q} - V$ because it might increase beyond the cap \bar{Q} if the spoof LOs are filled (recall that the investor imposes the cap \bar{Q} on the inventory holdings and V represents the number of shares that every time the spoof buy LO is filled by. The investor does not need to spoof the LOB when volume imbalance is already buy-heavy, because the arrival rate of buy MOs is highest. We denote by \mathcal{U} the set of admissible strategies c , consisting of \mathcal{F}_t -predictable control process $c \in \{0, 1\}$, such that the inventory does not exceed the upper bound \bar{Q} , and $c_t = 0$ when $Z_t = 1$ because the book is already in the buy-heavy regime.

We represent by $N^{\pm, c} = (N_t^{\pm, c})_{t \geq 0}$ the counting processes that keep track of the number of filled sell (+) and buy (−) LOs posted by the investor.

The investor also controls the times at which she sends sell MOs. These times are given by the impulse control $\tau = (\tau_1, \tau_2, \tau_3, \dots)$ and $M_t = \sum_{k=1}^{\infty} \mathbb{1}_{\{\tau_k \leq t\}}$ represents the counting process of the investor's sell MOs. We denote by \mathcal{V} the set of admissible strategies τ , which consists of increasing \mathcal{F}_t -stopping times less than T .

Denote by $Q^{c, \tau} = (Q_t^{c, \tau})_{t \geq 0}$ the investor's controlled inventory, which satisfies

$$dQ_t^{c, \tau} = dN_t^{-, c} - dN_t^{+, c} - dM_t, \quad (3.3)$$

and $X^{c, \tau} = (X_t^{c, \tau})_{t \geq 0}$ represents the cash process with dynamics

$$dX_t^{c, \tau} = (S_t^c + \Upsilon_{LO}) dN_t^{+, c} - (S_t^c - \Upsilon_{LO}) dN_t^{-, c} + (S_t^c - \Upsilon_{MO}) dM_t. \quad (3.4)$$

The time $\tau_S = \inf\{t : Q_t^{c, \tau} = 0\}$ is when the investor sells all her shares and stops trading.

The investor's value function is

$$H(t, x, S, q, Z) = \sup_{(c_t, \tau) \in \mathcal{A}} \mathbb{E}_{t, x, S, q, Z} \left[X_{\tau_S \wedge T}^{c, \tau} + Q_{\tau_S \wedge T}^{c, \tau} (S_{\tau_S \wedge T}^c - \Upsilon_{MO} - \alpha Q_{\tau_S \wedge T}^{c, \tau}) - \phi_q \int_t^{\tau_S \wedge T} (Q_u^{c, \tau})^2 du - \phi_f \int_t^{\tau_S \wedge T} c_u du \right], \quad (3.5)$$

where $\mathbb{E}_{t, x, S, q, Z}[\cdot]$ is the expectation operator conditional on $X_{t-}^{c, \tau} = x, S_{t-}^c = S, Q_{t-}^{c, \tau} = q$, and on the regime Z of the volume imbalance. The set of admissible strategies is

represented by $\mathcal{A} = \mathcal{U} \times \mathcal{V}$, which consists of \mathcal{F}_t -stopping times and \mathcal{F}_t -predictable control process $c \in \{0, 1\}$, such that the inventory does not exceed the upper bound \bar{Q} , and $c_t = 0$ when $Z_t = 1$ because the book is already in the buy-heavy regime.

We give a brief interpretation of the terms in the value function. The term $X_{\tau_S \wedge T}^{c, \tau}$ is the terminal cash held by the investor. The term $Q_{\tau_S \wedge T}^{c, \tau}(S_{\tau_S \wedge T}^c - \Upsilon_{MO} - \alpha Q_{\tau_S \wedge T}^{c, \tau})$ is the revenue from liquidating any remaining inventory using sell MOs at the terminal time. Here, $\alpha Q_{\tau_S \wedge T}^{c, \tau}$ represents the cost of walking the LOB and the terminal inventory penalty parameter is a non-negative constant. The term, $\phi_q \int_t^{\tau_S \wedge T} (Q_u^{c, \tau})^2 du$, is a running inventory penalty, where the non-negative constant ϕ_q is the running inventory penalty parameter, see Cartea and Jaimungal (2015c), Cartea et al. (2015), Guéant (2016). The last term, $\phi_f \int_t^{\tau_S \wedge T} c_u du$, represents the fine imposed by financial authorities if they detect spoofing and manipulation of prices, where the parameter $\phi_f \geq 0$ determines the severity of the fine.

By standard results, see Øksendal and Sulem (2007), the value function (3.5) is the unique viscosity solution of the Hamilton-Jacobi-Bellman quasi-variational inequality (HJBQVI)

$$\begin{aligned}
& \max \left\{ \partial_t H - \phi_q q^2 \right. \\
& \quad + \sup_{c \in \mathcal{U}} \left[c \left(-\phi_f + \gamma^+(1) \left(H(t, x, S + \sigma, q, Z) - H(t, x, S, q, Z) \right) \right. \right. \\
& \quad \quad \quad + \gamma^-(1) \left(H(t, x, S - \sigma, q, Z) - H(t, x, S, q, Z) \right) \\
& \quad \quad \quad + \lambda^s p_{sell} \left(H(t, x + (S + \Upsilon_{LO}), S, q - 1, Z) - H \right) \\
& \quad \quad \quad \left. \left. + \lambda^- p_{buy}(Z) \left(H(t, x - V(S - \Upsilon_{LO}), S, q + V, Z) - H \right) \right) \right. \\
& \quad \quad \quad + (1 - c) \left(\gamma^+(Z) \left(H(t, x, S + \sigma, q, Z) - H(t, x, S, q, Z) \right) \right. \\
& \quad \quad \quad \quad + \gamma^-(Z) \left(H(t, x, S - \sigma, q, Z) - H(t, x, S, q, Z) \right) \\
& \quad \quad \quad \left. \left. + \lambda^+(Z) p_{sell} \left(H(t, x + (S + \Upsilon_{LO}), S, q - 1, Z) - H \right) \right) \right] \\
& \quad \quad \quad + \sum_K (H(t, x, S, q, K) - H(t, x, S, q, Z)) G_{Z, K}, \\
& \quad \quad \quad \left. H(t, x + (S - \Upsilon_{MO}), S, q - 1, Z) - H \right\} = 0,
\end{aligned} \tag{3.6}$$

with terminal and boundary conditions

$$H(T, x, S, q, Z) = x + q(S - \Upsilon_{MO} - \alpha q) \quad \text{and} \quad H(t, x, S, 0, Z) = x,$$

respectively. For simplicity, we suppress the arguments of $H(t, x, S, q, Z)$ and write H ; we do the same for other functions we define later.

If $q < \bar{Q}$ and

$$\begin{aligned} & -\phi_f + \gamma^+(1) \left(H(t, x, S + \sigma, q, Z) - H \right) + \gamma^-(1) \left(H(t, x, S - \sigma, q, Z) - H \right) \\ & + \lambda^s p_{sell} \left(H(t, x + (S + \Upsilon_{LO}), S, q - 1, Z) - H \right) \end{aligned} \quad (3.7)$$

$$\begin{aligned} & + \lambda^- p_{buy}(Z) \left(H(t, x - V(S - \Upsilon_{LO}), S, q + V, Z) - H \right) \\ & > \gamma^+(Z) \left(H(t, x, S + \sigma, q, Z) - H \right) + \gamma^-(Z) \left(H(t, x, S - \sigma, q, Z) - H \right) \\ & + \lambda^+(Z) p_{sell} \left(H(t, x + (S + \Upsilon_{LO}), S, q - 1, Z) - H \right), \end{aligned} \quad (3.8)$$

then the optimal control process in feedback form is

$$c^*(t, x, S, q, Z) = 1. \quad (3.9)$$

Otherwise, $c^*(t, x, S, q, Z) = 0$.

The left-hand side of the inequality above shows the expected change in the value function when the investor rests a sell LO in the book and spoofs the buy side of the book. Specifically, (3.7) consists of five terms. The first term represents the fine imposed by financial authorities if they detect spoofing and manipulation of prices. The second and the third terms represent the change in the value function when the midprice moves. The intensities γ^\pm are those in the buy-heavy region because the investor spoofs the LOB into the buy-heavy regime. The fourth term represents the change in the value function when a sell LO is filled (with probability $\lambda^s p_{sell}$). The fifth term is the change in the value function when the investor's spoof buy LOs are filled (with probability $\lambda^- p_{buy}(Z)$). The right-hand side of the inequality, i.e., (3.8), consists of three parts. The first two terms represent the change in the value function when the midprice moves without spoofing. The third term is the change in the value function when the investor does not spoof the buy side of the book, so the sell LOs are filled with probability $\lambda^+(Z) p_{sell}$.

The optimal stopping times τ^* when the investor submits sell MOs satisfies

$$\tau_k^* = \inf \left\{ t > \tau_{k-1}^*, H(t, x + (S - \Upsilon_{MO}), S, q - 1, Z) - H = 0 \right\}, \quad \text{for } k > 1, \quad (3.10)$$

and

$$\tau_1^* = \inf \left\{ t > 0, H(t, x + (S - \Upsilon_{MO}), S, q - 1, Z) - H = 0 \right\}, \quad (3.11)$$

i.e., the investor executes a MO when the change in the value function is zero.

Substitute ansatz $H(t, x, S, q, Z) = x + qS + \tilde{h}(t, q, Z)$ in (3.6) to obtain

$$\begin{aligned} \max \left\{ \partial_t \tilde{h} - \phi_q q^2 \right. \\ + \sup_{c \in \mathcal{U}} \left[c \left(-\phi_f + (\gamma^+(1) - \gamma^-(1)) \sigma q + \lambda^s p_{sell} \left(\Upsilon_{LO} + \tilde{h}(t, q - 1, Z) - \tilde{h} \right) \right. \right. \\ \left. \left. + \lambda^- p_{buy}(Z) \left(V \Upsilon_{LO} + \tilde{h}(t, q + V, Z) - \tilde{h} \right) \right) \right. \\ \left. + (1 - c) \left((\gamma^+(Z) - \gamma^-(Z)) \sigma q + \lambda^+(Z) p_{sell} \left(\Upsilon_{LO} + \tilde{h}(t, q - 1, Z) - \tilde{h} \right) \right) \right] \\ + \sum_{K \neq Z} \left(\tilde{h}(t, q, K) - \tilde{h}(t, q, Z) \right) G_{Z,K}, \\ \left. - \Upsilon_{MO} + \tilde{h}(t, q - 1, Z) - \tilde{h} \right\} = 0, \end{aligned} \quad (3.12)$$

with terminal and boundary conditions

$$\tilde{h}(T, q, Z) = q(-\Upsilon_{MO} - \alpha q) \quad \text{and} \quad \tilde{h}(t, 0, Z) = 0,$$

respectively.

Now, if $q < \bar{Q}$ and

$$\begin{aligned} & -\phi_f + (\gamma^+(1) - \gamma^-(1)) \sigma q \\ & + \lambda^s p_{sell} \left(\Upsilon_{LO} + \tilde{h}(t, q - 1, Z) - \tilde{h} \right) + \lambda^- p_{buy}(Z) \left(V \Upsilon_{LO} + \tilde{h}(t, q + V, Z) - \tilde{h} \right) \\ & > (\gamma^+(Z) - \gamma^-(Z)) \sigma q + \lambda^+(Z) p_{sell} \left(\Upsilon_{LO} + \tilde{h}(t, q - 1, Z) - \tilde{h} \right), \end{aligned}$$

the feedback control becomes

$$c^*(t, q, Z) = 1. \quad (3.13)$$

Otherwise, $c^*(t, q, Z) = 0$.

The investor submits sell MOs at the stopping times τ^* such that

$$\tau_k^* = \inf \left\{ t > \tau_{k-1}^*, h(t, q - 1, Z) - h = \Upsilon_{MO} \right\}, \quad \text{for } k > 1, \quad (3.14)$$

and

$$\tau_1^* = \inf \left\{ t > 0, h(t, q-1, Z) - h = \Upsilon_{MO} \right\}. \quad (3.15)$$

Theorem 3.4.1. (*Verification*) Let \tilde{h} be a solution to (3.12) and define a candidate solution $\tilde{H} = x + qS + \tilde{h}(t, q, Z)$. Then \tilde{H} equals the value function as defined in (3.5).

Proof. For a proof please see Section 3.8. □

3.4.1 Trading in lots of shares

In our model setup, we assume the volume of sell LOs and buy MOs is one share and the volume that the spoof buy LOs is filled by is V shares. It is straightforward to scale the size of the investor's orders by a factor $\psi > 0$ and show that the value function of the investor scales linearly in ψ if the running penalty and terminal inventory penalty parameters also scale with ψ . First, note that if the investor's buy MO and sell LO are of size ψ , then the cash she obtains per filled LO is $\hat{S}^c + \hat{\Upsilon}_{LO}$, where $\hat{S}^c = \psi S^c$ and $\hat{\Upsilon}_{LO} = \psi \Upsilon_{LO}$. Similarly, the price received for a sell MO of size ψ is $\hat{S}^c - \hat{\Upsilon}_{MO}$, where $\hat{\Upsilon}_{MO} = \psi \Upsilon_{MO}$.

Therefore, when the investor trades in lots of size ψ , her value function is

$$\begin{aligned} \hat{H}(t, \hat{x}, \hat{S}, q, Z) = & \sup_{(c_t, \tau) \in \mathcal{A}} \mathbb{E}_{t, \hat{x}, \hat{S}, q, Z} \left[\hat{X}_{\tau_S \wedge T}^{c, \tau} + Q_{\tau_S \wedge T}^{c, \tau} \left(\hat{S}_{\tau_S \wedge T}^c - \hat{\Upsilon}_{MO} - \hat{\alpha} Q_{\tau_S \wedge T}^{c, \tau} \right) \right. \\ & \left. - \hat{\phi}_q \int_t^{\tau_S \wedge T} (Q_u^{c, \tau})^2 du - \hat{\phi}_f \int_t^{\tau_S \wedge T} c_u du \right], \end{aligned}$$

where $\hat{X}^{c, \tau}$ denotes the cash obtained from trading in lots of size ψ , and (with a slight abuse of notation) the units of inventory $Q^{c, \tau}$ and of the volume V (lots of shares filled when a spoof buy LO is hit), and the counting processes for filled LOs and executed MOs are in lots of shares of size ψ .

If we assume $\hat{\phi}_q = \psi \phi_q$, $\hat{\phi}_f = \psi \phi_f$, and $\hat{\alpha} = \psi \alpha$ it is easy to see that $\hat{H}(t, \hat{x}, \hat{S}, q, Z) = \psi H(t, x, S, q, Z)$. Thus, the optimal strategies of the original problem and the scaled problem are the same, and for each path of the strategy, terminal cash is $\hat{X}_{\tau_S \wedge T}^{c, \tau} = \psi X_{\tau_S \wedge T}^{c, \tau}$.

3.4.2 Optimal trading strategy

We use Nasdaq data to estimate model parameters for the stock INTC in October 2017 — some parameter estimates are in Section 3.2. As in the empirical analysis above, we remove the first and last 30 minutes of each day to exclude the behaviour of the LOB during the opening and closing auctions of each trading day. We use the methodology in Cartea et al. (2018) to calculate the maximum likelihood estimates of the parameters. Recall that λ^- is the arrival rate of sell MOs when the volume imbalance is buy-heavy, $\lambda^+(Z_t)$ is the arrival rate of buy MOs when the investor does not spoof the LOB, λ^s is the arrival rate of buy MOs when the investor spoofs the book, i.e., the investor tilts the book with buy LOs into the buy-heavy, and we assumed that $\lambda^s = \lambda^+(1)$. The estimate of the arrival rates (per second) are:

$$\hat{\lambda}^- = 0.0395, \quad \hat{\lambda}^s = \hat{\lambda}^+(1) = 0.2698, \quad \hat{\lambda}^+(2) = 0.0514, \quad \hat{\lambda}^+(3) = 0.0469.$$

To ensure the model does not contain any long term speculation based on a non-zero drift in the midprice, we impose the symmetry constraints, as in Cartea et al. (2018), on the arrival rates of the innovations in the midprice of the asset and the generator matrix of the volume imbalance regime process Z . The estimates of the arrival rates of the innovations in the midprice of the asset (per second) and the generator matrix are, respectively, given by

$$\hat{\gamma}^+ = (0.1985, 0.8195, 0.0961), \quad \hat{\gamma}^- = (0.0961, 0.8195, 0.1985),$$

and

$$\hat{G} = \begin{bmatrix} -0.8644 & 0.7642 & 0.1002 \\ 0.3634 & -0.7268 & 0.3634 \\ 0.1002 & 0.7642 & -0.8644 \end{bmatrix}.$$

Other model parameters are (see e.g., Cartea and Jaimungal (2015b) and Foucault (2012)):

$$\begin{aligned} \Delta &= 0.005, \quad \varepsilon_{MO} = 0.003, \quad \varepsilon_{LO} = 0.002, \quad \alpha = 0, \\ \mathfrak{N} &= 30 \text{ shares}, \quad \bar{Q} = 40 \text{ shares}, \quad T = 300 \text{ seconds}. \end{aligned}$$

Figure 3.3 shows the optimal spoof strategy as a function of the inventory and the time remaining to expiry. The black area, denoted by LO, represents the region where the investor **only** posts sell LOs and does not spoof the book. The white area, denoted by

Spoof, represents the region where the investor spoofs the book and also posts sell LOs, and the grey area, denoted by MO, represents the region where the investor executes sell MOs.

When market participants believe volume imbalance genuinely reflects demand and supply of the asset, and there is a potential fine for spoofing (i.e., $\phi_f = 4.5 \times 10^{-3}$), the optimal strategy for the agent is to spoof the LOB when the book is sell-heavy and neutral; the exception in these regimes is when the strategy is near expiry or the inventory is very high. The agent does not spoof when the inventory level is very low because the risk of a fine outweighs the benefit of spoofing. Recall that the investor does not spoof when volume imbalance is already buy-heavy ($Z = 1$) because the arrival rate of buy MOs is highest and the price of the asset is expected to increase.

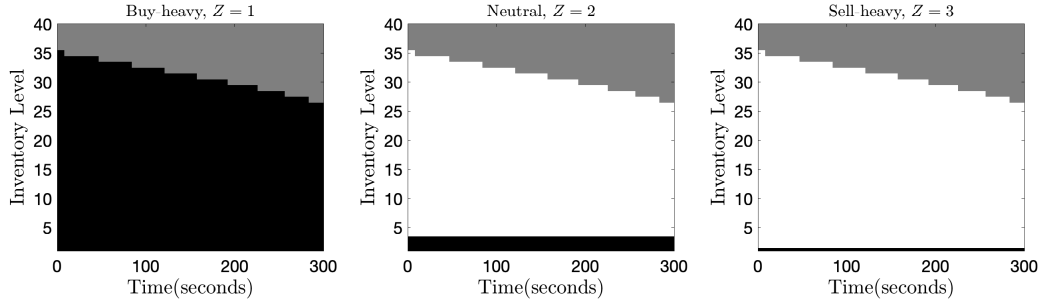


Figure 3.3: Optimal strategies for the investor. Parameter values as listed above and $\phi_q = 2 \times 10^{-5}$, $\phi_f = 4.5 \times 10^{-3}$, $p_{sell} = 0.3$, $p_{buy} = (0, 0.5, 0.6)$, $V = 2$. Black: investor **only** posts sell LOs and does not spoof the book. White: investor spoofs the book and also posts sell LOs. Grey: investor executes sell MOs.

3.5 Simulation and performance of strategy

In this section we employ simulations to analyse the performance of the optimal spoofing strategy. The initial midprice is $S_0 = 99.995$ and the tick size is $\sigma = 0.01$, and the best bid and best ask prices are 99.99 and 100.00, respectively. We assume that market participants believe the information conveyed by the LOB, so the spoofer benefits from an increase in the arrival rate of sell MOs and from price manipulation — recall that in the buy-heavy the midprice of the asset increases on average.

3.5.1 PnL and price manipulation

We run 10,000 simulations to show the performance of the optimal strategy as a function of the fine parameter ϕ_f . At time T the investor's PnL is

$$\text{PnL} = X_T + Q_T (S_T - \Upsilon_{MO} - \alpha Q_T), \quad (3.16)$$

where the units of inventory, LOs, and MOs is one lot of one share. As discussed in subsection 3.4.1, one can rescale the size of the LOs and MOs and the results derived here for lots of one share are easily re-interpreted. For example, the investor could trade lots of 300 shares in INTC, which is approximately the average size of the LOs posted on the book of INTC, see Table 3.5 — below we return to this particular example of lots of shares.

Table 3.7 shows statistics for the number of filled LOs and executed MOs when $\phi_q = 10^{-5}$, $p_{buy} = (0, 0.5, 0.6)$, and $V = 2$. The top panel of the table reports the case where the investor does not spoof the LOB. The other three panels assume that the fine parameter ϕ_f takes on the values $\{4 \times 10^{-2}, 2.7 \times 10^{-2}, 0\}$ respectively. When the fine parameter ϕ_f is 4×10^{-2} the expected penalty from spoofing is so high that is optimal for the investor not to spoof the LOB — compare the results of this case with those of no-spoof to observe that they are the same.

On the other hand, when the fine is zero, the results in the bottom panel show that the benefits from spoofing are considerable. On average, the investor sells more shares using LOs and fewer shares using MOs than the strategy without spoof LOs. The strategy fills an average of 9.5 spoof buy LOs, which are unwound (on average) with sell LOs. Although the spoof strategy bears the risk of fills on the ‘wrong’ side of the LOB, these inadvertently filled buy LOs help to increase the average PnL of the strategy. The mean PnL the investor obtains when there is no fine from spoofing is \$12.13 (which is approximately \$0.40 per share) greater than the mean PnL of the strategy without spoofing or when the fine parameter ϕ_f is 4×10^{-2} . This difference is approximately 1,213 times the quoted spread of INTC in the Nasdaq exchange. This example is extreme because the fine for spoofing is zero, so the investor spoofs the LOB for most of the trading horizon.

The PnL of the spoofing strategy is higher because: (i) the investor employs fewer MOs to draw the inventory to zero and also receives revenue from roundtrip trades, i.e., earns the spread (between the best ask and best bid prices) and the liquidity making rebate, when

filled spoof buy LOs are unwound with sell LOs. (ii) The midprice trends upward when the book is buy-heavy. Thus, as time evolves, the spoofer obtains, on average, better prices for selling the asset.

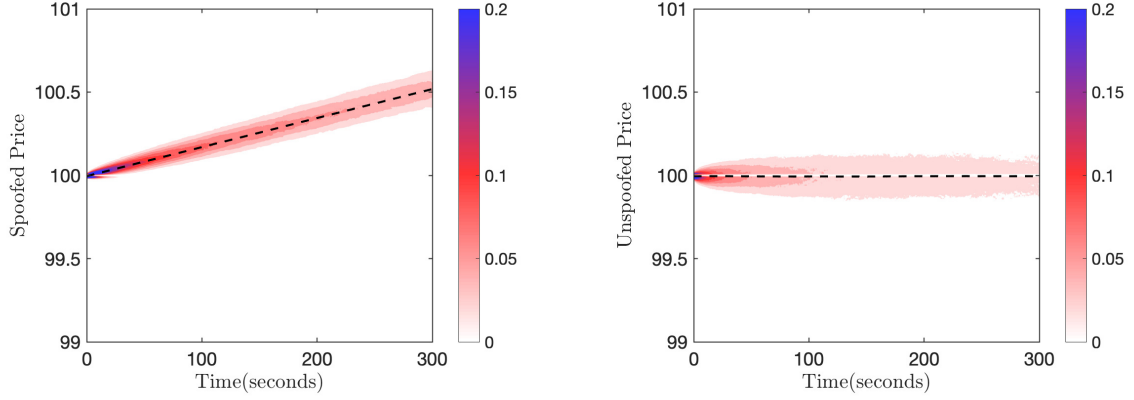
Observe that the standard deviation of the PnL of the strategy with spoof LOs is higher than the standard deviation of the PnL when the investor does not spoof the book. We explain the intuition of this result. First, with our choice of inventory penalty parameter ϕ_q , the no-spoof strategy liquidates the inventory, on average, with more sell MOs than sell LOs. On the other hand, the spoof strategy relies, on average, on more LOs and fewer MOs — thus, the PnL of the spoof strategy exhibits more variability than the PnL of the no-spoof strategy. Second, the liquidation programme with spoofing takes longer (mean time: 284.243 seconds in the case of zero fine) than the time it takes to liquidate the position (mean time: 190.035 seconds) without spoofing because the spoof strategy employs few MOs and many LOs. The execution time is the time when the inventory level first hits zero if it is before $T = 300$ seconds, or 300 seconds if there is still inventory remaining at T .

Finally, note that the standard deviation of the PnL of the spoof case with $\phi_f = 2.7 \times 10^{-2}$ is higher than the standard deviation of the PnL of a spoof strategy with zero fine. This result is explained by the volatility of the midprice, which is affected by the spoof strategy. In the buy-heavy regime the volatility of the midprice is lower than the volatility of the midprice when the LOB could be in any volume imbalance regime. When there is no fine for spoofing the book, the strategy tilts the book into the buy-heavy regime for most of the trading horizon, thus the volatility of the midprice is lower when $\phi_f = 0$; we return to this below when we discuss price manipulation.

		Quantiles							PnL	
		mean	std	0.01	0.25	0.50	0.75	0.99	mean	std
No Spoof	filled sell LOs	5.897	0.754	3	6	6	6	8	2,999.688	0.561
	executed sell MOs	24.103	0.754	22	24	24	24	27		
Spoof		Quantiles							PnL	
ϕ_f		mean	std	0.01	0.25	0.50	0.75	0.99	mean	std
4×10^{-2}	filled sell LOs	5.897	0.754	3	6	6	6	8	2,999.688	0.561
	executed sell MOs	24.103	0.754	22	24	24	24	27		
	filled buy LOs	0	0	0	0	0	0	0		
2.7×10^{-2}	filled sell LOs	15.654	2.924	10	14	15	17	24	3,006.759	4.153
	executed sell MOs	18.823	3.458	13	16	18	21	30		
	filled buy LOs	4.477	4.476	0	2	4	6	20		
0	filled sell LOs	24.290	4.771	14	21	24	27	36	3,011.818	2.717
	executed sell MOs	15.262	6.481	0	11	15	20	31		
	filled buy LOs	9.553	4.411	0	6	10	12	22		

Table 3.7: Each order is for one share of INTC. Parameters: $\phi_q = 10^{-5}$, $p_{sell} = 0.3$, $p_{buy} = (0, 0.5, 0.6)$, $V = 2$, $\mathfrak{N} = 30$. Mean execution time in seconds: i) No Spoofing: 190.035; ii) Spoofing: $\phi_f = 4 \times 10^{-2}$: 190.035, $\phi_f = 2.7 \times 10^{-2}$: 284.243, $\phi_f = 0$: 299.638.

Figure 3.4 shows heatmaps of the evolution of the price of the asset when the investor spoofs and does not spoof the LOB. Clearly, when there is spoofing, and market participants trust that volume imbalance is informative about the demand and supply of the asset, the midprice will exhibit an upward trend (see left panel of figure). Thus, the mean midprice at maturity is highest when the investor spoofs the LOB because the drift of the midprice is positive in this buy-heavy regime. Also, when the book is spoofed into the buy-heavy regime for a large proportion of the trading horizon, the midprice will exhibit lower volatility than when the LOB could be in any volume imbalance regime as is the case with no spoofing.



(a) $\text{Mean}(S_T) = 100.518$ and $\text{Std}(S_T) = 0.081$

(b) $\text{Mean}(S_T) = 99.995$ and $\text{Std}(S_T) = 0.172$

Figure 3.4: Heatmaps of the evolution of the midprice. Parameters: $\phi_q = 10^{-10}$, $\phi_f = 0$, $p_{\text{sell}} = 0.3$, $p_{\text{buy}} = (0, 0.5, 0.6)$, $V = 2$, $\mathfrak{N} = 30$.

Finally, for the examples discussed in Table 3.7, if we assume each lot consists of 300 shares, which is approximately the average volume of a LO posted at the best bid in INTC's book (see Table 3.5), the improvement in the mean PnL with spoofing compared with no spoofing is approximately $1,213 \times 300$ times the quoted spread in INTC. That is, the investor liquidates 9,000 shares (with scaled penalty parameters $\hat{\alpha} = 300\alpha$, $\hat{\phi}_q = 300\phi_q$, and $\hat{\phi}_f = 300\phi_f$) and MOs and LOs are for lots of 300 shares instead of lots of one share.

3.5.2 Tradeoff: mean and standard deviation of PnL

In this section we discuss the mean and the standard deviation of various performance measures that employ the following inputs: the PnL obtained by the investor from selling shares with and without spoofing and a market benchmark.

We introduce three measures of relative performance. The first is

$$\text{rPnL}_1 = \text{PnL} - \mathfrak{N}(S_0 - \Upsilon_{MO}), \quad (3.17)$$

where PnL is given by (3.16).

The two other measures of relative performance are:

$$\text{rPnL}_i = \frac{\text{PnL} - \text{benchmark}_i}{\text{benchmark}_i} \times 10^4, \quad i = 2, 3, \quad (3.18)$$

where $\text{benchmark}_2 = \mathfrak{N}(S_0 - \Upsilon_{MO})$ and $\text{benchmark}_3 = \text{PnL}^{TWAP}$. Here, PnL^{TWAP} denotes the revenue obtained from a time-weighted-average-price strategy known as TWAP, i.e., the investor liquidates shares at the constant speed \mathfrak{N}/T . The measure of performance (3.18) represents the extra PnL of the spoof strategy relative to a benchmark in basis points.

We employ 10,000 simulations to compute the mean and the standard deviation of rPnL_i for $i = 1, 2, 3$. The fine parameter ϕ_f takes values in the interval $[e^{-6}, e^{-2}]$, which is approximately $[2.5 \times 10^{-3}, 1.3 \times 10^{-1}]$. Other model parameters are $\phi_q = 10^{-10}$, $p_{sell} = 0.3$, $p_{buy} = (0, 0.5, 0.6)$, $V = 2$, $p_{sell} = 0.3$.

Figure 3.5 depicts $\text{Mean}(\text{rPnL}_i)/\text{Std}(\text{rPnL}_i)$, where the x -axis is in units of $\log \phi_f$. As expected, the difference between the performance of the spoof and that of the no-spoof strategies is highest when the fine for spoofing is smallest. As the fine for spoofing increases, the strategy relies on fewer spoof buy LOs because the expected fine outweighs the benefits from spoofing.

From panels (b) and (c) of Figure 3.5, we notice that for the TWAP benchmark, the difference in $\text{Mean}(\text{rPnL})/\text{Std}(\text{rPnL})$ between spoof and no-spoof, is smaller than when the benchmark is $\mathfrak{N}(S_0 - \Upsilon_{MO})$. This shows, for our choice of parameters, that TWAP outperforms $\mathfrak{N}(S_0 - \Upsilon_{MO})$ and therefore the increase in $\text{Mean}(\text{rPnL})/\text{Std}(\text{rPnL})$ (panel (c)) is less significant than the case with $\mathfrak{N}(S_0 - \Upsilon_{MO})$ as benchmark (panel (b)).

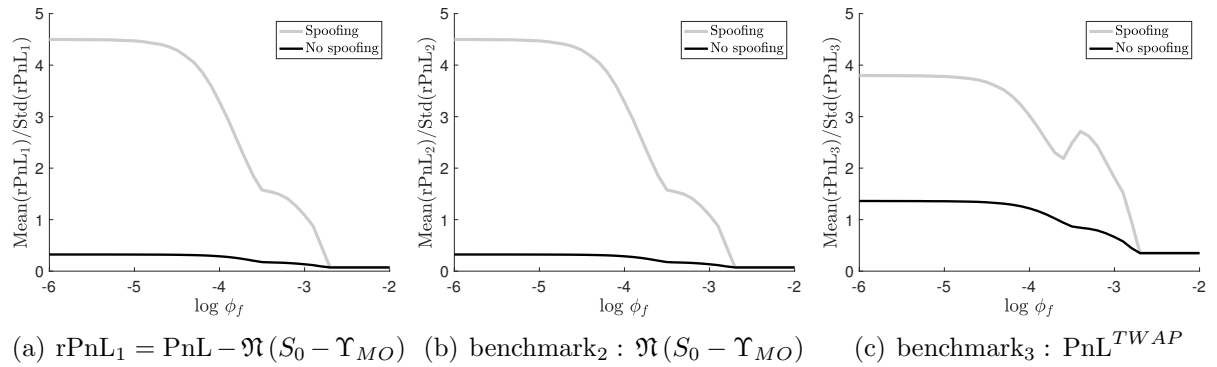


Figure 3.5: Performance of strategy for ϕ_f in the range $[e^{-6}, e^{-2}]$. Other parameters: $\phi_q = 10^{-10}$, $p_{sell} = 0.3$, $p_{buy} = (0, 0.5, 0.6)$, $V = 2$, $p_{sell} = 0.3$.

3.6 Numerical scheme

In this section we describe the numerical scheme we employ to solve (3.12). Let $\mathbb{T}_{\delta t}$ be the uniform grid on $[0, T]$ with step size $\delta t > 0$. For any function $\varphi : [0, T] \times ([0, \bar{Q}] \cap \mathbb{Z}) \times \{1, 2, 3\} \rightarrow \mathbb{R}$ define the operator

$$\mathcal{S}^{\delta t}(t, q, Z, \varphi) = \max [\mathcal{T}^{\delta t}(t, q, Z, \varphi), \mathcal{M}^{\delta t}(t, q, Z, \varphi)] , \quad (3.19)$$

where

$$\begin{aligned} \mathcal{T}^{\delta t}(t, q, Z, \varphi) &= \varphi + \delta t \left[-\phi_q q^2 \right. \\ &\quad + \sup_{c \in \mathcal{U}} \left[c \left(-\phi_f + (\gamma^+(1) - \gamma^-(1)) \sigma q + \lambda^s p_{sell} \left(\Upsilon_{LO} + \varphi(t, q-1, Z) - (1 + \kappa) \varphi \right) \right. \right. \\ &\quad \left. \left. + \lambda^- p_{buy}(Z) \left(V \Upsilon_{LO} + \varphi(t, q+V, Z) - (1 + \kappa) \varphi \right) \right) \right. \\ &\quad \left. \left. + (1-c) \left((\gamma^+(Z) - \gamma^-(Z)) \sigma q + \lambda^+(Z) p_{sell} \left(\Upsilon_{LO} + \varphi(t, q-1, Z) - (1 + \kappa) \varphi \right) \right) \right] \right. \\ &\quad \left. + \sum_{K \neq Z} (\varphi(t, q, K) - (1 + \kappa) \varphi(t, q, Z)) G_{Z,K} \right] , \end{aligned} \quad (3.20)$$

and

$$\mathcal{M}^{\delta t}(t, q, Z, \varphi) = \frac{\varphi(t, q-1, Z) - \Upsilon_{MO}}{1 + \kappa} , \quad (3.21)$$

where $\kappa \downarrow 0$ is a robustness parameter.

Define the numerical solution $h^{\delta t} : \mathbb{T}_{\delta t} \times ([0, \bar{Q}] \cap \mathbb{Z}) \times \{1, 2, 3\} \rightarrow \mathbb{R}$ as follows:

$$\begin{cases} h^{\delta t}(T, q, Z) &= -q (\Upsilon_{MO} + \alpha q) , \\ h^{\delta t}(t, 0, Z) &= 0 , \\ h^{\delta t}(k \delta t, q, Z) &= \mathcal{S}^{\delta t}(t, q, Z, h^{\delta t}((k+1) \delta t, q, Z)) . \end{cases} \quad (3.22)$$

We employ the explicit scheme backwards in time from T .

Next we prove the convergence of (3.22). We first prove monotonicity, stability, and consistency properties (Propositions 3.6.1, 3.6.2, and 3.6.3 respectively) of $\mathcal{S}^{\delta t}(t, q, Z, \varphi)$. Combined with the comparison principle (Corollary 3.8.1), we prove the convergence by following Barles and Souganidis (1991).

Lemma 3.6.1. *The value function H obeys the following bounds:*

$$\begin{aligned} x + q S - \max_i \gamma^-(i) \sigma \bar{Q} (T - t) - \Upsilon_{MO} \bar{Q} - \alpha \bar{Q}^2 - \phi_q \bar{Q}^2 (T - t) - \phi_f (T - t) \\ \leq H \leq \\ x + q S + (T - t) \left[\Upsilon_{LO} \left(\lambda^- V \max_i p_{buy}(i) + \max_i \lambda^+(i) p_{sell} \right) + \max_i \gamma^+(i) \sigma \bar{Q} \right]. \end{aligned} \quad (3.23)$$

Proof. For a proof see Section 3.8. \square

Proposition 3.6.1. *(Monotonicity) For any δt less than some constant, $\varphi_1, \varphi_2 \in C_b^1([0, T] \times ([0, \bar{Q}] \cap \mathbb{Z}) \times \{1, 2, 3\})$ such that $\varphi_1 \leq \varphi_2$, we have $\mathcal{S}^{\delta t}(t, q, Z, \varphi_1) \leq \mathcal{S}^{\delta t}(t, q, Z, \varphi_2)$.*

Proof. From expression (3.20) we observe that $\mathcal{T}^{\delta t}(t, q, Z, \varphi)$ is monotone non-decreasing in φ , given

$$\delta t < \left[(1 + \kappa) \left(\lambda^s p_{sell} + \lambda^- V \max_i p_{buy}(i) + \max_i \lambda^+(i) p_{sell} - G_{Z,Z} \right) \right]^{-1}, \quad (3.24)$$

and monotonicity of $\mathcal{M}^{\delta t}$ is obvious. \square

Proposition 3.6.2. *(Stability) For any $\delta t > 0$, there exists a unique solution $h^{\delta t}(t, q, Z)$ to (3.22). Furthermore, we have the uniform bounds*

$$L(t, q) \leq h^{\delta t}(t, q, Z) \leq U(t), \quad (3.25)$$

where

$$U(t) = (T - t) \left[\Upsilon_{LO} \left(\lambda^- V \max_i p_{buy}(i) + \max_i \lambda^+(i) p_{sell} \right) + \max_i \gamma^+(i) \sigma \bar{Q} \right]$$

and

$$L(t, q) = - \max_i \gamma^-(i) \sigma \bar{Q} (T - t) - \Upsilon_{MO} q - \alpha \bar{Q}^2 - \phi_q \bar{Q}^2 (T - t) - \phi_f (T - t).$$

Proof. For a proof see Section 3.8. \square

Proposition 3.6.3. *(Consistency) For all $(t, q, Z) \in [0, T] \times ([0, \bar{Q}] \cap \mathbb{Z}) \times \{1, 2, 3\}$ and*

$\varphi \in C_b^1\left([0, T] \times ([0, \bar{Q}] \cap \mathbb{Z}) \times \{1, 2, 3\}\right)$, we have

$$\begin{aligned}
& \lim_{\substack{\delta t \rightarrow 0 \\ t' \rightarrow t}} \frac{1}{\delta t} [\mathcal{T}^{\delta t}(t' + \delta t, q, Z, \varphi) - \varphi(t', q, Z)] \\
&= \partial_t \varphi - \phi_q q^2 \\
&\quad + \sup_{c \in \mathcal{U}} \left[c \left(-\phi_f + (\gamma^+(1) - \gamma^-(1)) \sigma q + \lambda^s p_{\text{sell}} \left(\Upsilon_{LO} + \varphi(t, q-1, Z) - (1 + \kappa) \varphi \right) \right. \right. \\
&\quad \left. \left. + \lambda^- p_{\text{buy}}(Z) \left(V \Upsilon_{LO} + \varphi(t, q+V, Z) - (1 + \kappa) \varphi \right) \right) \right. \\
&\quad \left. + (1 - c) \left(+ (\gamma^+(Z) - \gamma^-(Z)) \sigma q \right. \right. \\
&\quad \left. \left. + \lambda^+(Z) p_{\text{sell}} \left(\Upsilon_{LO} + \varphi(t, q-1, Z) - (1 + \kappa) \varphi \right) \right) \right] \\
&\quad + \sum_{K \neq Z} (\varphi(t, q, K) - (1 + \kappa) \varphi(t, q, Z)) G_{Z,K},
\end{aligned}$$

and

$$\lim_{\substack{\delta t \rightarrow 0 \\ t' \rightarrow t}} \mathcal{M}^{\delta t}(t' + \delta t, q, Z, \varphi) = \frac{\varphi(t, q-1, Z) - \Upsilon_{MO}}{1 + \kappa}.$$

Proof. The limits converge by directly applying $\varphi \in C_b^1\left([0, T] \times ([0, \bar{Q}] \cap \mathbb{Z}) \times \{1, 2, 3\}\right)$. \square

Theorem 3.6.1. (*Convergence*) The function $h^{\delta t}(t, q, Z)$ converges locally uniformly to the unique viscosity solution $h(t, q, Z)$ as $\delta t \rightarrow 0$.

Proof. Follows Corollary from 3.8.1 (see the Appendix), Propositions 3.6.1, 3.6.2, 3.6.3, and Barles and Souganidis (1991). See Section 3.8 for more details. \square

3.7 Conclusions and Further Research

In this chapter we derived the optimal trading strategy for an investor who employs spoof LOs to improve the rate and prices at which she sells shares with limit orders in an order-driven electronic market. The spoof strategy is illegal and benefits from misleading other market participants about the supply and demand of an asset.

We focused on a strategy where the investor may choose to tilt the volume imbalance of the LOB by posting LOs on one side of the book. The strategy trades off the benefits from spoofing and the potential fine the investor may receive from the financial authorities. When the expected fine is low, we found that spoofing the book considerably increases the financial performance of an execution strategy. The financial improvement stems from: (i) executing more shares using limit orders (instead of market orders that cross the spread and pay fees), (ii) employing sell limit orders to unwind shares that were ‘inadvertently’ purchased with spoof buy limit orders, and (iii) increasing the drift of the midprice so inventory appreciates and shares are sold at higher prices.

We also showed that spoofing deviates the price of the asset from its fundamental value. This deviation is highest when market participants believe the information conveyed by the LOB and the fine for spoofing is zero. Our simulations showed that at the end of the trading horizon, which lasts five minutes, the mean price of the asset is 52 cents higher than the mean price of the asset in the absence of spoofing. As the penalty for spoofing increases, the investor relies less on spoof LOs, so the manipulation of the price of the asset is less effective and the PnL of the spoofing strategy decreases.

The framework we developed here can be employed to develop other trading strategies that rely on spoof orders to improve their financial performance. For example, a market making strategy that employs sell and buy spoof LOs to open and close positions, respectively. There are other ways to spoof the LOB that could be considered within the framework developed in this chapter. We provide three examples:

- Phantom liquidity inside the spread. The investor wishes to sell shares. The strategy consists of sending spoof buy LOs inside the spread (i.e., improve the bid price) to entice other liquidity providers to join the queue at the improved bid price. As soon as other traders send LOs at the new best bid price, the spoofer cancels her spoof buy LOs and sends sell MOs. A similar strategy is used when the investor wishes to purchase shares.
- Cross-spoofing. This is identical to the strategy developed in this chapter, only that two (or more) investors agree to spoof the market to avoid being detected by financial authorities. In the strategy, one investor(s) spoofs the LOB and the other investor(s) liquidates the position in the shares.

- Layering. The strategy consists of posting several large LOs at different prices on one side of the book. The goal is to move the price because other market participants interpret the one sided pressure in the LOB as a signal of a price move and trade in anticipation of expected change in price.

To the best of our knowledge this is the first work that provides a mathematical framework to develop optimal spoofing strategies. Strategies that layer or spoof the book are detrimental to the integrity of markets. Our framework can be employed to understand patterns of spoofing and to develop data techniques (e.g., machine learning) to identify strategies that are detrimental to the market.

3.8 Proofs

3.8.1 Proof of Theorem 3.4.1

Proof. We follow Øksendal and Sulem (2007). Let the function $f^c(q) = -\phi_q q^2 - \phi_c c$. We define the operators \mathfrak{L}^c and \mathfrak{M} such that the HJBQVI (3.12) is represented by the form of

$$\max \left\{ \sup_{c \in \mathcal{U}} \left[\mathfrak{L}^c \tilde{h} + f^c \right], \mathfrak{M} \tilde{h} - \tilde{h} \right\} = 0. \quad (3.26)$$

Let \tilde{h} be the solution to (3.12) and define the candidate solution $\tilde{H} = x + qS + \tilde{h}(t, q, Z)$. We want to show that $\tilde{H} = H$.

For any control (c, τ) , from Itô's Lemma we have

$$\begin{aligned} & \mathbb{E} \left[\tilde{H}(\tau_n^-, X_{\tau_n^-}^{c, \tau}, S_{\tau_n^-}^c, Q_{\tau_n^-}^{c, \tau}, Z_{\tau_n^-}^{c, \tau}) \middle| \mathcal{F}_t \right] - \mathbb{E} \left[\tilde{H}(\tau_{n-1}^-, X_{\tau_{n-1}^-}^{c, \tau}, S_{\tau_{n-1}^-}^c, Q_{\tau_{n-1}^-}^{c, \tau}, Z_{\tau_{n-1}^-}^{c, \tau}) \middle| \mathcal{F}_t \right] \\ &= \mathbb{E} \left[\int_{\tau_{n-1}}^{\tau_n} \mathfrak{L}^c \hat{h}(u, Q_u^{c, \tau}, Z_u^{c, \tau}) du \middle| \mathcal{F}_t \right], \end{aligned}$$

and

$$\begin{aligned} & \tilde{H}(\tau_n, X_{\tau_n}^{c, \tau}, S_{\tau_n}^c, Q_{\tau_n}^{c, \tau}, Z_{\tau_n}^{c, \tau}) - \tilde{H}(\tau_n^-, X_{\tau_n^-}^{c, \tau}, S_{\tau_n^-}^c, Q_{\tau_n^-}^{c, \tau}, Z_{\tau_n^-}^{c, \tau}) \\ &= \mathfrak{M} \tilde{h}(\tau_n^-, Q_{\tau_n^-}^{c, \tau}, Z_{\tau_n^-}^{c, \tau}) - \tilde{h}(\tau_n^-, Q_{\tau_n^-}^{c, \tau}, Z_{\tau_n^-}^{c, \tau}), \end{aligned}$$

where, with a slight abuse of notation, we only include the jumps from the control. Summing over $[t, \tau_S \wedge T]$, taking an expectation conditional on \mathcal{F}_t and rearranging yields

$$\begin{aligned}
\tilde{H}(t, x, S, q, Z) &= \mathbb{E} \left[\tilde{H}(\tau_S \wedge T, X_{\tau_S \wedge T}^{c, \tau}, S_{\tau_S \wedge T}^c, Q_{\tau_S \wedge T}^{c, \tau}, Z_{\tau_S \wedge T}^{c, \tau}) \middle| \mathcal{F}_t \right] \\
&\quad - \mathbb{E} \left[\int_t^{\tau_S \wedge T} \mathfrak{L}^c \tilde{h}(u, Q_u^{c, \tau}, Z_u^{c, \tau}) du \middle| \mathcal{F}_t \right] \\
&\quad - \mathbb{E} \left[\sum_{\tau_n \leq \tau_S \wedge T} \mathfrak{M} \tilde{h}(\tau_n^-, Q_{\tau_n^-}^{c, \tau}, Z_{\tau_n^-}^{c, \tau}) - \tilde{h}(\tau_n^-, Q_{\tau_n^-}^{c, \tau}, Z_{\tau_n^-}^{c, \tau}) \middle| \mathcal{F}_t \right]
\end{aligned} \tag{3.27}$$

From HJBQVI (3.12) we have

$$\begin{aligned}
\tilde{H}(t, x, S, q, Z) &\geq \mathbb{E} \left[\tilde{H}(\tau_S \wedge T, X_{\tau_S \wedge T}^{c, \tau}, S_{\tau_S \wedge T}^c, Q_{\tau_S \wedge T}^{c, \tau}, Z_{\tau_S \wedge T}^{c, \tau}) \right. \\
&\quad \left. - \phi_q \int_t^{\tau_S \wedge T} (Q_u^{c, \tau})^2 du - \phi_f \int_t^{\tau_S \wedge T} c_u du \middle| \mathcal{F}_t \right] \\
&= \mathbb{E} \left[X_{\tau_S \wedge T}^{c, \tau} + Q_{\tau_S \wedge T}^{c, \tau} (S_{\tau_S \wedge T}^c - \Upsilon_{MO} - \alpha Q_{\tau_S \wedge T}^{c, \tau}) \right. \\
&\quad \left. - \phi_q \int_t^{\tau_S \wedge T} (Q_u^{c, \tau})^2 du - \phi_f \int_t^{\tau_S \wedge T} c_u du \middle| \mathcal{F}_t \right].
\end{aligned} \tag{3.28}$$

Since the inequality above holds for any control (c, τ) , we have

$$\begin{aligned}
\tilde{H}(t, x, S, q, Z) &\geq \sup_{(c_t, \tau) \in \mathcal{A}} \mathbb{E}_{t, x, S, q, Z} \left[X_{\tau_S \wedge T}^{c, \tau} + Q_{\tau_S \wedge T}^{c, \tau} (S_{\tau_S \wedge T}^c - \Upsilon_{MO} - \alpha Q_{\tau_S \wedge T}^{c, \tau}) \right. \\
&\quad \left. - \phi_q \int_t^{\tau_S \wedge T} (Q_u^{c, \tau})^2 du - \phi_f \int_t^{\tau_S \wedge T} c_u du \right] \\
&= H(t, x, S, q, Z).
\end{aligned} \tag{3.29}$$

Now if we use the optimal control (c^*, τ^*) , (3.27) becomes

$$\begin{aligned}
\tilde{H}(t, x, S, q, Z) &= \mathbb{E} \left[\tilde{H}(\tau_S \wedge T, X_{\tau_S \wedge T}^{c^*, \tau^*}, S_{\tau_S \wedge T}^{c^*}, Q_{\tau_S \wedge T}^{c^*, \tau^*}, Z_{\tau_S \wedge T}^{c^*, \tau^*}) \right. \\
&\quad \left. - \phi_q \int_t^{\tau_S \wedge T} (Q_u^{c^*, \tau^*})^2 du - \phi_f \int_t^{\tau_S \wedge T} c_u^* du \middle| \mathcal{F}_t \right] \\
&= \mathbb{E} \left[X_{\tau_S \wedge T}^{c^*, \tau^*} + Q_{\tau_S \wedge T}^{c^*, \tau^*} (S_{\tau_S \wedge T}^{c^*} - \Upsilon_{MO} - \alpha Q_{\tau_S \wedge T}^{c^*, \tau^*}) \right. \\
&\quad \left. - \phi_q \int_t^{\tau_S \wedge T} (Q_u^{c^*, \tau^*})^2 du - \phi_f \int_t^{\tau_S \wedge T} c_u^* du \middle| \mathcal{F}_t \right] \\
&\leq \sup_{(c_t, \tau) \in \mathcal{A}} \mathbb{E}_{t, x, S, q, Z} \left[X_{\tau_S \wedge T}^{c, \tau} + Q_{\tau_S \wedge T}^{c, \tau} (S_{\tau_S \wedge T}^c - \Upsilon_{MO} - \alpha Q_{\tau_S \wedge T}^{c, \tau}) \right. \\
&\quad \left. - \phi_q \int_t^{\tau_S \wedge T} (Q_u^{c, \tau})^2 du - \phi_f \int_t^{\tau_S \wedge T} c_u du \right] \\
&= H(t, x, S, q, Z).
\end{aligned} \tag{3.30}$$

Hence $H = \hat{H}$. □

3.8.2 Comparison principle

In this section we prove the comparison principle of (3.12), which we employ later in the proof of convergence of the numerical scheme. We introduce a regularised version of (3.12) and prove the comparison principle of this regularised version. The proof of convergence of the numerical scheme of the unregularised version is beyond the scope of this work.

The regularised version of HJBQVI (3.12) is

$$\begin{aligned}
& \max \left\{ \partial_t h - \phi_q q^2 \right. \\
& + \sup_{c \in \mathcal{U}} \left[c \left(-\phi_f + (\gamma^+(1) - \gamma^-(1)) \sigma q + \lambda^s p_{sell} \left(\Upsilon_{LO} + h(t, q-1, Z) - (1+\kappa) h \right) \right. \right. \\
& \quad \left. \left. + \lambda^- p_{buy}(Z) \left(V \Upsilon_{LO} + h(t, q+V, Z) - (1+\kappa) h \right) \right) \right. \\
& \quad \left. + (1-c) \left((\gamma^+(Z) - \gamma^-(Z)) \sigma q \right. \right. \\
& \quad \left. \left. + \lambda^+(Z) p_{sell} \left(\Upsilon_{LO} + h(t, q-1, Z) - (1+\kappa) h \right) \right) \right] \\
& \quad \left. + \sum_{K \neq Z} (h(t, q, K) - (1+\kappa) h(t, q, Z)) G_{Z,K}, \right. \\
& \quad \left. - \Upsilon_{MO} + h(t, q-1, Z) - (1+\kappa) h \right\} = 0,
\end{aligned} \tag{3.31}$$

where $\kappa \downarrow 0$ is a robustness parameter, with terminal and boundary conditions

$$h(T, q, Z) = q(-\Upsilon_{MO} - \alpha q) \quad h(t, 0, Z) = 0,$$

respectively.

For notational convenience we make the following changes of variables:

$$s = T - t, \quad g_{q,Z}(s) = -h(t, q, Z), \quad d_{q,Z}(s) = c(t, q, Z),$$

and define \mathcal{U}_d to be the set of admissible strategies in the new variables, and let

$$F_{q,Z}(r, u, p) = \max\{P_{q,Z}(r, u, p), O_{q,Z}(r, u, p)\}, \tag{3.32}$$

where

$$\begin{aligned}
P_{q,Z}(r, u, p) = & p - \phi_q q^2 \\
& + \sup_{d_{q,Z}(r) \in \mathcal{U}_d} \left[d_{q,Z}(r) \left(-\phi_f + (\gamma^+(1) - \gamma^-(1)) \sigma q \right. \right. \\
& \quad + \lambda^s p_{sell} (\Upsilon_{LO} - u_{q-1,Z} + (1 + \kappa) u_{q,Z}) \\
& \quad + \lambda^- p_{buy}(Z) (V \Upsilon_{LO} - u_{q+V,Z} + (1 + \kappa) u_{q,Z}) \Big) \\
& \quad + (1 - d_{q,Z}(r)) \lambda^+(Z) \left((\gamma^+(Z) - \gamma^-(Z)) \sigma q \right. \\
& \quad \quad \left. \left. + p_{sell} (\Upsilon_{LO} - u_{q-1,Z} + (1 + \kappa) u_{q,Z}) \right) \right] \\
& - \sum_{K \neq Z} (u_{q,K} - (1 + \kappa) u_{q,Z}) G_{Z,K},
\end{aligned}$$

and

$$O_{q,Z}(r, u, p) = (1 + \kappa) u_{q,Z} - (\Upsilon_{MO} + u_{q-1,Z}).$$

Then, the HJBQVI becomes

$$F_{q,Z}(s, g, \partial_s g) = 0, \quad (3.33)$$

with terminal and boundary conditions

$$g_{q,Z}(0) = q (\Upsilon_{MO} + \alpha q), \quad g_{0,Z}(t) = 0, \quad (3.34)$$

respectively.

Proposition 3.8.1. (A1) Let $u = (u_{q,Z})$ and $v = (v_{q,Z})$. Suppose that $u_{q^*,Z^*} - v_{q^*,Z^*} = \max_{q,Z} \{u_{q,Z} - v_{q,Z}\}$, then there exists $c_0 > 0$, where

$$c_0 = \kappa \min \left\{ \lambda^s p_{sell} + \lambda^- p_{buy}(Z^*), \lambda^+(Z^*) p_{sell}, -G_{Z^*,Z^*}, 1 \right\},$$

such that

$$F_{q^*,Z^*}(s, u, p) - F_{q^*,Z^*}(s, v, p) \geq c_0 (u_{q^*,Z^*} - v_{q^*,Z^*}). \quad (3.35)$$

(A2) There exists $\beta > 1$ and a continuous function $\omega : \mathbb{R}_+ \rightarrow \mathbb{R}_+$, $\omega(0) = 0$, such that

$$F_{q,Z}(s, u, \beta(s' - s)) - F_{q,Z}(s', u, \beta(s' - s)) \leq \omega(\beta(s' - s)^2 + \beta^{-1}) \quad (3.36)$$

for all q, Z .

Proof. (A1) The first required inequality is stated as follows:

$$\max \{P_{q^*,Z^*}(s, u, p), O_{q^*,Z^*}(s, u, p)\} - \max \{P_{q^*,Z^*}(s, v, p), O_{q^*,Z^*}(s, v, p)\} \geq c_0 (u_{q^*,Z^*} - v_{q^*,Z^*}).$$

We only need to prove the following two inequalities:

$$\begin{aligned} P_{q^*,Z^*}(s, u, p) - P_{q^*,Z^*}(s, v, p) &\geq c_0(u_{q^*,Z^*} - v_{q^*,Z^*}), \\ O_{q^*,Z^*}(s, u, p) - O_{q^*,Z^*}(s, v, p) &\geq c_0(u_{q^*,Z^*} - v_{q^*,Z^*}). \end{aligned}$$

The other two situations are converted to the two inequalities above. If we obtain different terms from the maximum, then we have

$$\begin{aligned} O_{q^*,Z^*}(s, u, p) - P_{q^*,Z^*}(s, v, p) &\geq P_{q^*,Z^*}(s, u, p) - P_{q^*,Z^*}(s, v, p), \\ P_{q^*,Z^*}(s, u, p) - O_{q^*,Z^*}(s, v, p) &\geq O_{q^*,Z^*}(s, u, p) - O_{q^*,Z^*}(s, v, p). \end{aligned}$$

For the first inequality,

$$\begin{aligned} &P_{q^*,Z^*}(s, u, p) - P_{q^*,Z^*}(s, v, p) \\ &= \max \left[-\phi_f + (\gamma^+(1) - \gamma^-(1)) \sigma q + \lambda^s p_{sell} (\Upsilon_{LO} - u_{q^*-1,Z^*} + (1 + \kappa) u_{q^*,Z^*}) \right. \\ &\quad \left. + \lambda^- p_{buy}(Z^*) (V \Upsilon_{LO} - u_{q^*+V,Z^*} + (1 + \kappa) u_{q^*,Z^*}), \right. \\ &\quad \left. (\gamma^+(Z) - \gamma^-(Z)) \sigma q + \lambda^+(Z^*) p_{sell} (\Upsilon_{LO} - u_{q^*-1,Z^*} + (1 + \kappa) u_{q^*,Z^*}) \right] \\ &\quad - \max \left[-\phi_f + (\gamma^+(1) - \gamma^-(1)) \sigma q + \lambda^s p_{sell} (\Upsilon_{LO} - v_{q^*-1,Z^*} + (1 + \kappa) v_{q^*,Z^*}) \right. \\ &\quad \left. + \lambda^- p_{buy}(Z^*) (V \Upsilon_{LO} - v_{q^*+V,Z^*} + (1 + \kappa) v_{q^*,Z^*}), \right. \\ &\quad \left. (\gamma^+(Z) - \gamma^-(Z)) \sigma q + \lambda^+(Z^*) p_{sell} (\Upsilon_{LO} - v_{q^*-1,Z^*} + (1 + \kappa) v_{q^*,Z^*}) \right] \\ &\quad + \sum_{K \neq Z^*} ((1 + \kappa) (u_{q^*,Z^*} - v_{q^*,Z^*}) - (u_{q^*,K} - v_{q^*,K})) G_{Z^*,K}. \end{aligned}$$

For the first two maxima we use the same technique as above and only need to prove that

$$\begin{aligned} &\lambda^s p_{sell} (\Upsilon_{LO} - u_{q^*-1,Z^*} + (1 + \kappa) u_{q^*,Z^*}) + \lambda^- p_{buy}(Z^*) (V \Upsilon_{LO} - u_{q^*+V,Z^*} + (1 + \kappa) u_{q^*,Z^*}) \\ &\quad - \lambda^s p_{sell} (\Upsilon_{LO} - v_{q^*-1,Z^*} + (1 + \kappa) v_{q^*,Z^*}) - \lambda^- p_{buy}(Z^*) (V \Upsilon_{LO} - v_{q^*+V,Z^*} + (1 + \kappa) v_{q^*,Z^*}) \\ &\geq c_0 (u_{q^*,Z^*} - v_{q^*,Z^*}), \end{aligned} \tag{3.37}$$

and

$$\begin{aligned} & \lambda^+(Z^*) p_{sell}(\Upsilon_{LO} - u_{q^*-1, Z^*} + (1 + \kappa) u_{q^*, Z^*}) - \lambda^+(Z^*) p_{sell}(\Upsilon_{LO} - v_{q^*-1, Z^*} + (1 + \kappa) v_{q^*, Z^*}) \\ & \geq c_0 (u_{q^*, Z^*} - v_{q^*, Z^*}). \end{aligned} \quad (3.38)$$

To prove inequality (3.37), we write

$$\begin{aligned} & \lambda^s p_{sell}(\Upsilon_{LO} - u_{q^*-1, Z^*} + (1 + \kappa) u_{q^*, Z^*}) + \lambda^- p_{buy}(Z^*)(V \Upsilon_{LO} - u_{q^*+V, Z^*} + (1 + \kappa) u_{q^*, Z^*}) \\ & \quad - \lambda^s p_{sell}(\Upsilon_{LO} - v_{q^*-1, Z^*} + (1 + \kappa) v_{q^*, Z^*}) - \lambda^- p_{buy}(Z^*)(V \Upsilon_{LO} - v_{q^*+V, Z^*} + (1 + \kappa) v_{q^*, Z^*}) \\ & = \lambda^s p_{sell} \left((1 + \kappa) (u_{q^*, Z^*} - v_{q^*, Z^*}) - (u_{q^*-1, Z^*} - v_{q^*-1, Z^*}) \right) \\ & \quad + \lambda^- p_{buy}(Z^*) \left((1 + \kappa) (u_{q^*, Z^*} - u_{q^*+V, Z^*}) - (u_{q^*-1, Z^*} - v_{q^*+V, Z^*}) \right) \\ & \geq \kappa (\lambda^s p_{sell} + \lambda^- p_{buy}(Z^*)) (u_{q^*, Z^*} - u_{q^*+V, Z^*}) \\ & \geq c_0 (u_{q^*, Z^*} - v_{q^*, Z^*}), \end{aligned}$$

by the definition of (q^*, Z^*) in (A1).

To prove inequality (3.38), we write

$$\begin{aligned} & \lambda^+(Z^*) p_{sell}(\Upsilon_{LO} - u_{q^*-1, Z^*} + (1 + \kappa) u_{q^*, Z^*}) - \lambda^+(Z^*) p_{sell}(\Upsilon_{LO} - v_{q^*-1, Z^*} + (1 + \kappa) v_{q^*, Z^*}) \\ & = \lambda^+(Z^*) p_{sell} \left((1 + \kappa) (u_{q^*, Z^*} - v_{q^*, Z^*}) - (u_{q^*-1, Z^*} - v_{q^*-1, Z^*}) \right) \\ & \geq \kappa \lambda^+(Z^*) p_{sell} (u_{q^*, Z^*} - v_{q^*, Z^*}) \\ & \geq c_0 (u_{q^*, Z^*} - v_{q^*, Z^*}), \end{aligned}$$

by the definition of (q^*, Z^*) in (A1).

We also have

$$\begin{aligned} \sum_{K \neq Z^*} ((1 + \kappa) (u_{q^*, Z^*} - v_{q^*, Z^*}) - (u_{q^*, K} - v_{q^*, K})) G_{Z^*, K} & \geq \kappa \sum_{K \neq Z^*} (u_{q^*, Z^*} - v_{q^*, Z^*}) G_{Z^*, K} \\ & = \kappa (-G_{Z^*, Z^*}) (u_{q^*, Z^*} - v_{q^*, Z^*}) \\ & \geq c_0 (u_{q^*, Z^*} - v_{q^*, Z^*}), \end{aligned}$$

by the definition of (q^*, Z^*) in (A1). Thus $P_{q^*, Z^*}(s, u, p) - P_{q^*, Z^*}(s, v, p) \geq c_0 (u_{q^*, Z^*} - v_{q^*, Z^*})$.

$$\begin{aligned}
& O_{q^*, Z^*}(s, u, p) - O_{q^*, Z^*}(s, v, p) \\
&= \left((1 + \kappa) u_{q^*, Z^*} - (\Upsilon_{MO} + u_{q^*-1, Z^*}) \right) - \left((1 + \kappa) v_{q^*, Z^*} - (\Upsilon_{MO} + v_{q^*-1, Z^*}) \right) \\
&= (1 + \kappa) (u_{q^*, Z^*} - v_{q^*, Z^*}) - (u_{q^*-1, Z^*} - v_{q^*-1, Z^*}) \\
&\geq \kappa (u_{q^*, Z^*} - v_{q^*, Z^*}) \\
&\geq c_0 (u_{q^*, Z^*} - v_{q^*, Z^*}),
\end{aligned}$$

by the definition of (q^*, Z^*) in (A1). Hence, we obtain

$$F_{q^*, Z^*}(s, u, p) - F_{q^*, Z^*}(s, v, p) \geq c_0 (u_{q^*, Z^*} - v_{q^*, Z^*}).$$

(A2) Inequality (3.36) is satisfied because

$$F_{q, Z}(s, u, \beta(s' - s)) - F_{q, Z}(s', u, \beta(s' - s)) = 0.$$

□

Theorem 3.8.1. (*Comparison Principle*) Assume F is continuous. Let g_1 and g_2 be, respectively, a bounded subsolution and a bounded supersolution of (3.33). Suppose that $g_1(0) \leq g_2(0)$. Then $g_1 \leq g_2$.

Proof. See Theorem 4.7 in Ishii and Koike (1991). □

Corollary 3.8.1. Equation (3.31) admits a comparison principle.

Proof. Directly from Theorem 3.8.1 by change of variables. □

3.8.3 Proof of Lemma 3.6.1

Proof.

$$\begin{aligned}
& H(t, x, S, q, Z) \\
&= \sup_{(c_t, \tau) \in \mathcal{A}} \mathbb{E}_{t, x, S, q, Z} \left[X_{\tau_S \wedge T}^{c, \tau} + Q_{\tau_S \wedge T}^{c, \tau} (S_{\tau_S \wedge T}^c - \Upsilon_{MO} - \alpha Q_{\tau_S \wedge T}^{c, \tau}) \right. \\
&\quad \left. - \phi_q \int_t^{\tau_S \wedge T} (Q_u^{c, \tau})^2 du - \phi_f \int_t^{\tau_S \wedge T} c_u du \right] \\
&= x + q(S - \Upsilon_{MO}) \\
&\quad + \sup_{(c_t, \tau) \in \mathcal{A}} \mathbb{E}_{t, x, S, q, Z} \left[\int_t^{\tau_S \wedge T} dX_u^{c, \tau} + \int_t^{\tau_S \wedge T} S_u^c dQ_u^{c, \tau} - \Upsilon_{MO}(Q_{\tau_S \wedge T}^{c, \tau} - q) + \int_t^{\tau_S \wedge T} Q_u^{c, \tau} dS_u^c \right. \\
&\quad \left. - \alpha (Q_{\tau_S \wedge T}^{c, \tau})^2 - \phi_q \int_t^{\tau_S \wedge T} (Q_u^{c, \tau})^2 du - \phi_f \int_t^{\tau_S \wedge T} c_u du \right] \\
&= x + q(S - \Upsilon_{MO}) \\
&\quad + \sup_{(c_t, \tau) \in \mathcal{A}} \mathbb{E}_{t, x, S, q, Z} \left[\int_t^{\tau_S \wedge T} (S_u^c + \Upsilon_{LO}) dN_u^{+, c} - \int_t^{\tau_S \wedge T} (S_u^c - \Upsilon_{LO}) dN_u^{-, c} + \int_t^{\tau_S \wedge T} (S_u^c - \Upsilon_{MO}) dM_u \right. \\
&\quad + \int_t^{\tau_S \wedge T} S_u^c dN_u^{-, c} - \int_t^{\tau_S \wedge T} S_u^c dN_u^{+, c} - \int_t^{\tau_S \wedge T} S_u^c dM_u - \Upsilon_{MO}(Q_{\tau_S \wedge T}^{c, \tau} - q) \\
&\quad \left. + \int_t^{\tau_S \wedge T} Q_u^{c, \tau} dS_u - \alpha (Q_{\tau_S \wedge T}^{c, \tau})^2 - \phi_q \int_t^{\tau_S \wedge T} (Q_u^{c, \tau})^2 du - \phi_f \int_t^{\tau_S \wedge T} c_u du \right].
\end{aligned}$$

After cancelling terms, we have

$$\begin{aligned}
& H(t, x, S, q, Z) \\
&= x + qS \\
&\quad + \sup_{(c_t, \tau) \in \mathcal{A}} \mathbb{E}_{t, x, S, q, Z} \left[\Upsilon_{LO} \left(\int_t^{\tau_S \wedge T} dN_u^{+, c} + \int_t^{\tau_S \wedge T} dN_u^{-, c} \right) + \int_t^{\tau_S \wedge T} Q_u^{c, \tau} dS_u^c \right. \\
&\quad \left. - \Upsilon_{MO} (Q_{\tau_S \wedge T}^{c, \tau} + M_{\tau_S \wedge T} - M_t) - \alpha (Q_{\tau_S \wedge T}^{c, \tau})^2 - \phi_q \int_t^{\tau_S \wedge T} (Q_u^{c, \tau})^2 du - \phi_f \int_t^{\tau_S \wedge T} c_u du \right].
\end{aligned}$$

So far we have not introduced any enlargement in the value function, and we give a brief interpretation of each term. The first two integrals, which are non-negative, represent the profit of posting the LOs and earning the spread. The terms

$$-\Upsilon_{MO} (Q_{\tau_S \wedge T}^{c, \tau} + M_{\tau_S \wedge T} - M_t) - \alpha (Q_{\tau_S \wedge T}^{c, \tau})^2 - \phi_q \int_t^{\tau_S \wedge T} (Q_u^{c, \tau})^2 du - \phi_f \int_t^{\tau_S \wedge T} c_u du$$

represent the cost of sending MOs, the terminal and running penalties on inventory of the value function, and the fine of spoofing respectively.

For the lower bound, we first restrict the set of admissible strategies such that the investor does not submit any MOs, which is clearly admissible and hence has performance criteria less than H . Also, we drop the three non-negative integrals, and since c , Q and τ_S are bounded, we have

$$\begin{aligned}
& H(t, x, S, q, Z) \\
& \geq x + q S \\
& \quad + \sup_{(c_t, \tau) \in \mathcal{A}} \mathbb{E}_{t, x, S, q, Z} \left[\Upsilon_{LO} \left(\int_t^{\tau_S \wedge T} dN_u^{+, c} + \int_t^{\tau_S \wedge T} dN_u^{-, c} \right) + \int_t^{\tau_S \wedge T} Q_u^{c, \tau} dS_u^c \right. \\
& \quad \left. - \Upsilon_{MO} Q_{\tau_S \wedge T}^{c, \tau} - \alpha (Q_{\tau_S \wedge T}^{c, \tau})^2 - \phi_q \int_t^{\tau_S \wedge T} (Q_u^{c, \tau})^2 du - \phi_f \int_t^{\tau_S \wedge T} c_u du \right] \\
& \geq x + q S - \max_i \gamma^-(i) \sigma \bar{Q} (T - t) - \Upsilon_{MO} \bar{Q} - \alpha \bar{Q}^2 - \phi_q \bar{Q}^2 (T - t) - \phi_f (T - t).
\end{aligned}$$

For the upper bound, we first drop all the non-positive terms, and by enlarging $\tau_S \wedge T$ and c_u , we obtain

$$\begin{aligned}
& H(t, x, S, q, Z) \\
& \leq x + q S \\
& \quad + \sup_{(c_t, \tau) \in \mathcal{A}} \mathbb{E}_{t, x, S, q, Z} \left[\Upsilon_{LO} \left(\int_t^{\tau_S \wedge T} dN_u^{+, c} + \int_t^{\tau_S \wedge T} dN_u^{-, c} \right) + \int_t^{\tau_S \wedge T} Q_u^{c, \tau} dS_u^c \right] \\
& \leq x + q S + (T - t) \left[\Upsilon_{LO} \left(\lambda^- V \max_i p_{buy}(i) + \max_i \lambda^+(i) p_{sell} \right) + \max_i \gamma^+(i) \sigma \bar{Q} \right].
\end{aligned}$$

□

3.8.4 Proof of Proposition 3.6.2

Proof. Existence and uniqueness follow immediately from the definition of the explicit scheme (3.22). We first prove the upper bound. We notice that $h^{\delta t}(T, q, Z) = -q(\Upsilon_{MO} +$

$\alpha q) \leq U(T) = 0$, and

$$\begin{aligned}
\mathcal{T}^{\delta t}(t, q, Z, U) &= U(t) + \delta t \left[-\phi_q q^2 \right. \\
&\quad + \sup_{c \in \mathcal{U}} \left[c \left(-\phi_f + (\gamma^+(1) - \gamma^-(1)) \sigma q + \lambda^s p_{sell} (\Upsilon_{LO} - \kappa U(t)) \right. \right. \\
&\quad \left. \left. + \lambda^- p_{buy}(Z) (V \Upsilon_{LO} - \kappa U(t)) \right) \right. \\
&\quad \left. + (1 - c) \left((\gamma^+(Z) - \gamma^-(Z)) \sigma q \right. \right. \\
&\quad \left. \left. + \lambda^+(Z) p_{sell} (\Upsilon_{LO} - \kappa U(t)) \right) \right] \\
&\quad \left. + \sum_{K \neq Z} -\kappa U(t) G_{Z,K} \right] \\
&\leq U(t) + \delta t \left[\Upsilon_{LO} \left(\lambda^- V \max_i p_{buy}(i) + \max_i \lambda^+(i) p_{sell} \right) + \max_i \gamma^+(i) \sigma \bar{Q} \right] \\
&= U(t - \delta t).
\end{aligned}$$

Together with

$$\mathcal{M}^{\delta t}(t, q, Z, U) = \frac{U(t) - \Upsilon_{MO}}{1 + \kappa} \leq U(t) \leq U(t - \delta t),$$

we have

$$\mathcal{S}^{\delta t}(t, q, Z, U) \leq U(t - \delta t).$$

As $h^{\delta t}(t - \delta t, q, Z) = \mathcal{S}^{\delta t}(t, q, Z, h^{\delta t}(t, q, Z))$, we prove by induction that $h^{\delta t}(t, q, Z) \leq U(t)$.

Similarly for the lower bound, we have $h^{\delta t}(T, q, Z) = -q(\Upsilon_{MO} + \alpha q) \geq L(T, q) =$

$-\Upsilon_{MO} q - \alpha \bar{Q}^2$ because $q \leq \bar{Q}$, and

$$\begin{aligned}
& \mathcal{T}^{\delta t}(t, q, Z, L) \\
= & L(t, q) + \delta t \left[-\phi_q q^2 \right. \\
& + \sup_{c \in \mathcal{U}} \left[c \left(-\phi_f + (\gamma^+(1) - \gamma^-(1)) \sigma q + \lambda^s p_{sell}(\Upsilon_{LO} + \Upsilon_{MO} - \kappa L(t, q)) \right. \right. \\
& \quad \left. \left. + \lambda^- p_{buy}(Z) (V \Upsilon_{LO} - V \Upsilon_{MO} - \kappa L(t, q)) \right) \right. \\
& \quad \left. + (1 - c) \left((\gamma^+(Z) - \gamma^-(Z)) \sigma q \right. \right. \\
& \quad \left. \left. + \lambda^+(Z) p_{sell}(\Upsilon_{LO} + \Upsilon_{MO} - \kappa L(t, q)) \right) \right] \\
& \quad \left. + \sum_{K \neq Z} -\kappa L(t, q) G_{Z,K} \right] \\
\geq & L(t, q) - \delta t \left(\max_i \gamma^-(i) \sigma q + \phi_q q^2 + \phi_f \right) \\
\geq & L(t, q) - \delta t \left(\max_i \gamma^-(i) \sigma \bar{Q} + \phi_q \bar{Q}^2 + \phi_f \right) = L(t - \delta t, q).
\end{aligned}$$

Together with

$$\begin{aligned}
\mathcal{M}^{\delta t}(t, q, Z, L) &= \frac{L(t, q)}{1 + \kappa} \\
&= \frac{L(t - \delta t, q) + \delta t \left(\max_i \gamma^-(i) \sigma \bar{Q} + \phi_q \bar{Q}^2 + \phi_f \right)}{1 + \kappa} \\
&\geq \frac{L(t - \delta t, q) + \kappa L(t - \delta t, q)}{1 + \kappa} \\
&= L(t - \delta t, q),
\end{aligned}$$

we have

$$\mathcal{S}^{\delta t}(t, q, Z, L) \geq L(t - \delta t, q).$$

As $h^{\delta t}(t - \delta t, q, Z) = \mathcal{S}^{\delta t}(t, q, Z, h^{\delta t}(t, q, Z))$, we prove by induction that $h^{\delta t}(t, q, Z) \geq L(t, q)$.

□

3.8.5 Proof of Theorem 3.6.1

Proof. The proof follows from Corollary 3.8.1, Proposition 3.6.1, 3.6.2, 3.6.3, and Barles and Souganidis (1991).

We define

$$h_*(t, q, Z) = \liminf_{\substack{\delta t \rightarrow 0 \\ t' \rightarrow t}} h^{\delta t}(t, q, Z) \quad \text{and} \quad h^*(t, q, Z) = \limsup_{\substack{\delta t \rightarrow 0 \\ t' \rightarrow t}} h^{\delta t}(t, q, Z),$$

which are, respectively, lower and upper semi-continuous functions on $[0, T] \times ([0, \bar{Q}] \cap \mathbb{Z}) \times \{1, 2, 3\}$, and inherit the boundedness of $\{h^{\delta t}(t, q, Z)\}$ by stability from Proposition 3.6.2. By definition, we have $h_* \leq h^*$. We claim that h_* and h^* are, respectively, a viscosity supersolution and a viscosity subsolution of (3.31) and (3.32), then by Corollary 3.8.1 (Comparison Principle) we have $h^* \leq h_*$ and hence the equality. By symmetry, it suffices to show the viscosity supersolution property of h_* .

Let $(\bar{t}, \bar{q}, \bar{Z}) \in [0, T) \times ([0, \bar{Q}] \cap \mathbb{Z}) \times \{1, 2, 3\}$ and $\varphi \in C_b^1([0, T] \times ([0, \bar{Q}] \cap \mathbb{Z}) \times \{1, 2, 3\})$ such that $(\bar{t}, \bar{q}, \bar{Z})$ attains the strict global minimum of $h_* - \varphi$. Then there exists a sequence $\{(t'_k, q'_k, Z'_k)\}_k \in [0, T) \times ([0, \bar{Q}] \cap \mathbb{Z}) \times \{1, 2, 3\}$ and $\{\delta t_k\}_k$ such that

$$\begin{aligned} (t'_k, q'_k, Z'_k) &\rightarrow (\bar{t}, \bar{q}, \bar{Z}), \\ \delta t_k &\rightarrow 0, \\ h^{\delta t_k} &\rightarrow h_*(\bar{t}, \bar{q}, \bar{Z}), \end{aligned}$$

and (t'_k, q'_k, Z'_k) is the global minimizer of $h^{\delta t_k} - \varphi$.

Here we restrict δt_k to satisfy the condition in Proposition 3.6.1, so that we apply the monotonicity of $\mathcal{S}^{\delta t}$.

We define $\varepsilon_k = (h^{\delta t_k} - \varphi)(t'_k, q'_k, Z'_k)$. Then by the numerical scheme in (3.22) and the monotonicity from Proposition 3.6.1, we have

$$\begin{aligned} \varepsilon_k + \varphi(t'_k, q'_k, Z'_k) &= h^{\delta t_k}(t'_k, q'_k, Z'_k) \\ &= \mathcal{S}^{\delta t_k}(t'_k + \delta t_k, q'_k, Z'_k, h^{\delta t_k}) \\ &\geq \mathcal{S}^{\delta t_k}(t'_k + \delta t_k, q'_k, Z'_k, \varphi + \varepsilon_k) \\ &= \mathcal{S}^{\delta t_k}(t'_k + \delta t_k, q'_k, Z'_k, \varphi) + \varepsilon_k \\ &= \max \{ \mathcal{T}^{\delta t_k}(t'_k + \delta t_k, q'_k, Z'_k, \varphi), \mathcal{M}^{\delta t_k}(t'_k + \delta t_k, q'_k, Z'_k, \varphi) \} + \varepsilon_k. \end{aligned}$$

After rearranging, we have

$$\begin{aligned} \max \left\{ \frac{1}{\delta t_k} [\mathcal{T}^{\delta t_k}(t'_k + \delta t_k, q'_k, Z'_k, \varphi) - \varphi(t'_k, q'_k, Z'_k)], \right. \\ \left. (1 + \kappa) [\mathcal{M}^{\delta t_k}(t'_k + \delta t_k, q'_k, Z'_k, \varphi) - \varphi(t'_k, q'_k, Z'_k)] \right\} \leq 0. \end{aligned}$$

We apply the consistency in Proposition 3.6.3 and let $k \rightarrow \infty$, and obtain

$$\begin{aligned}
& \max \left\{ \partial_t \varphi - \phi_q q^2 \right. \\
& + \sup_{c \in \mathcal{U}} \left[c \left(-\phi_f + (\gamma^+(1) - \gamma^-(1)) \sigma q + \lambda^s p_{sell} \left(\Upsilon_{LO} + \varphi(t, q-1, Z) - (1 + \kappa) \varphi \right) \right. \right. \\
& \quad \left. \left. + \lambda^- p_{buy}(Z) \left(V \Upsilon_{LO} + \varphi(t, q+V, Z) - (1 + \kappa) \varphi \right) \right) \right. \\
& \left. + (1-c) \left((\gamma^+(Z) - \gamma^-(Z)) \sigma q + \lambda^+(Z) p_{sell} \left(\Upsilon_{LO} + \varphi(t, q-1, Z) - (1 + \kappa) \varphi \right) \right) \right] \\
& \quad + \sum_{K \neq Z} (\varphi(t, q, K) - (1 + \kappa) \varphi(t, q, Z)) G_{Z,K}, \\
& \quad \left. \varphi(t, q-1, Z) - \Upsilon_{MO} - (1 + \kappa) \varphi \right\} \leq 0,
\end{aligned}$$

which is the viscosity supersolution property as desired. Therefore we have $h^* \leq h_*$, and hence $h^* = h_*$.

□

Chapter 4

Market making with minimum resting times

4.1 Introduction

With the advent of computerised trading and electronic exchanges, the speed at which market participants process information and make trading decisions has increased dramatically, and so has the number of messages sent by traders to the exchanges. These messages are instructions sent by computerised trading algorithms that make decisions to execute MOs, manage inventories, and to post, amend, and cancel LOs. Essential to the price discovery process of stocks, and to the timely dissemination of new information impounded in equity prices, is the ability of liquidity providers to cancel stale LOs and re-post orders in the exchange's book. Orders resting in the LOB are options given to liquidity takers, some of which have the relative speed advantage to process information and snipe stale LOs. Thus, liquidity makers are exposed to being picked off if they do not update their quotes quickly.

While the total number of messages sent to exchanges has surged over the last decade, there has been a disproportionate increase in the number of messages dedicated to cancelling LOs. Stakeholders, market observers, and regulators have asked how this steep increase in the number of cancelled LOs affects the quality of markets. One may interpret the cancellation of LOs as beneficial to the market because liquidity providers update their views and refresh their quotes. This guarantees the market displays the supply of liquidity at the most up-to-date prices, i.e., prices are efficient. On the other hand, it is not clear if there is intention to trade when the vast majority of LOs are cancelled,

and some of them are cancelled over a time window so short that is nearly impossible for market takers to execute against those orders.

Van Ness et al. (2015) study monthly rates of cancellations in over 25 exchanges in the US. They define the cancellation rate as the number of shares cancelled divided by the number of shares posted for each month in the period 2001 to 2010. They show that the rate of cancellations during 2001 is between 30% and 40% and by 2010 it increased to levels above 90%. Moreover, the authors show that for flagship exchanges such as NYSE, NASDAQ, and BATS, the cancellation rates rose from around 70% in 2006 to between 93% and 94% in 2010. The authors conclude that cancellation activity is detrimental to market quality.

Cartea et al. (2019) employ messages sent to NASDAQ to build a measure of ultra-fast activity. This measure records how many LOs are posted and subsequently cancelled within 100 milliseconds. The authors show that an increase in ultra-fast activity leads to lower liquidity: greater quoted and effective spreads, and lower depth posted in the LOB.

Although there is no conclusive evidence on the overall effect that high levels of cancellation have on the quality of markets, financial regulators have discussed possible rules to curb the number of LOs that are cancelled. In the ‘Review of MiFID’ Commission et al. (2010), the European Commission suggests forcing LOs to rest in the LOB for a minimum period before being cancelled, see also Farmer and Skouras (2012). The objective is to slow down activity from traders who post fleeting or short-lived LOs, so that liquidity provision is more stable. An alternative proposal by the European Commission suggested that the ratio of LOs to executed transactions (i.e., filled LOs) for individual market participants should not exceed a pre-specified level.

In this chapter we show how MMs adjust their LOs when the exchange enforces a minimum resting time (MRT) on the LOs before they can be cancelled. We assume that the price of the asset follows an arithmetic Brownian motion. MMs are profit maximisers and decide the depth of the LOs in the book. LOs cannot be cancelled before the compulsory MRT, so this affects the optimal depth of the LOs and the amount of shares that the MMs are willing to supply to the market.

Our findings shed light into the regulatory discussion of the effect of MRTs on the quality of order driven markets. We show that: (i) the depth (relative to the price of the asset) of the LOs in the book increases as the MRT increases, (ii) everything else being equal,

the larger the volume of the LOB, the deeper in the book orders are posted; (iii) the optimal depth of the LOs increases when volatility increases because the probability that LOs become stale and are picked off increases in volatility; (iv) the LOs of MMs supply the minimum amount of shares required by the exchange per LO.

Finally, we also show that the expected profits of MMs increase as the MRT increases. The intuition behind the result is as follows. When the MRT increases, and the depth of posted LO increases, there are two opposing forces at work. (i) The longer the MRT, the more likely the LOs are to be filled and, on average, shares are sold at a loss. (ii) The chance that all posted volume is picked off by other traders before the end of the MRT decreases as the depth of the posted LO increases. The net effect is that longer MRTs lead to higher expected profits. This is due to our model setup and in particular is caused by the MM considering the lifetime of a LO as the time horizon.

The remainder of this chapter proceeds as follows. In Section 4.2 we provide a review of the literature of MRT. In Section 4.3 we introduce the mathematical model assuming all MMs are identical, and we derive an asymptotic integral expression for the expected profit of the MM. Section 4.4 extends the model so MMs can post LOs of any positive integer volume. In Section 4.5 we investigate the optimal depth that the MM chooses. In Section 4.6 we look at the expected profit faced by the MM when she chooses the depth and quantity of shares to post in the LO. We draw conclusions in Section 4.7 and collect proofs Section 4.8.

4.2 Literature review

The extant literature that studies the effects of imposing MRTs in order driven markets is scant and mostly based on simulation platforms. In this vein, Brewer et al. (2013) show that longer MRTs creates liquidity and reduces the volatility of prices because LOs are forced to stay in the book. The authors also show that MRTs reduce the effect of a very large order causing the flash crash (i.e., sharp decrease in the price of the asset followed by a quick recovery in the price), and make the recovery faster.

Hayes et al. (2012) use an agent based model (ABM) and simulations to examine the impact of MRTs on the E-Mini S&P 500 market. They show that MRTs decrease price volatility and MRTs improve market liquidity in several measures, but the changes are

statistically insignificant. They show that MRT tightens the bid-ask spread by a marginal amount, however they assume that market participants trade in the same manner before and after the MRT rule is implemented. In our work the depth and volume of the LOs of the MM depend on the MRT and we find that the bid-ask spread increases because MRTs cause an increase in the depth of the LOs.

Leal and Napoletano (2017) use an ABM in which the interactions between low- and high-frequency traders can generate flash crashes. Their results indicate that MRTs can dampen market volatility and reduce the incidence of flash crashes, but at the same time MRTs increase the time it takes prices to recover from extreme market conditions. In their model setup, the strategies employed by high-frequency traders can lead to wide bid-ask spreads during a flash crash without MRT. When an MRT is implemented, the wider spreads caused by the flash crash persist for a longer period of time in the LOB. In our model, volatility is exogenous but we show that if volatility increases, then the bid-ask spread increases.

Ait-Sahalia and Sağlam (2017) propose a theoretical model to derive an optimal quoting policy with MRTs. They assume that MRTs are exponentially distributed random variables with expected value of 500 milliseconds. Their results are opposite to those in our findings. They show that the expected profits of MMs decrease when the expected MRT increases, while we show that the expected profits of MMs increase when the MRT increases (note that MRTs in our model are not random). They also show that after implementing the MRT, MMs are more sensitive to market volatility than in the absence of MRTs — they find that the spread is low when volatility is low, and high when volatility is high. This result is in line with ours, we find that spreads increase when the volatility of prices increases.

4.3 Model I: limit orders of same volume

In this section we present the model of the MM. First, we derive the expressions for the profit the MM receives when i) all the volume posted in her LO is filled before reaching the MRT, and ii) not all her volume posted in the LO is filled before the end of the MRT. Second, we specify the fill ratio probability of the LOs. Finally we specify a model for the dynamics of the fundamental stock price and compute the expected profits of the MM.

We denote by $S = (S_t)_{t \geq 0}$ the fundamental price of the stock. At time t the MM posts a buy and a sell LO of volume M in the LOB. We assume that the LOs have the same depth, so the ask and bid prices posted by the MM are

$$S_t^a = S_t + \delta/2, \quad S_t^b = S_t - \delta/2, \quad (4.1)$$

respectively, where $\delta > 0$.

We assume there are other MMs following the same strategy as that of the MM, i.e., they post an LO of volume M on each side of the LOB at depth $\delta/2$ (in Section 4.4 we present Model II to examine the case where MMs may post LOs of any integer volume). For simplicity we assume that the depths on both sides are the same and the volume of the LOs are the same. The model can be easily extended so the MM posts a limit sell order of volume M_a and a limit buy order of volume M_b , and the orders are posted at depths δ_a and δ_b respectively.

After the MM posts a limit sell order of volume M at time $t = 0$ and price $S_0 + \delta/2$, other MMs are continuously posting new bid and ask LOs with depth $\delta/2$ relative to the up-to-date fundamental price, i.e., bid at $S_t - \delta/2$ and ask at $S_t + \delta/2$.

All market participants know the fundamental price S_t , but can only trade the stock via the exchange by either posting LOs or by executing MOs. Thus, trades at the price S_t only occur if the quoted depth in the LOB is zero.

In the subsequent analysis we focus only on the limit sell orders of the MM. This simplifies and streamlines the exposition of the model and results. By symmetry, the LOs of the MM on the buy side are the mirror image of the sell LOs, so it is straightforward to extend the results when the buy and sell LOs of the MM are considered.

In our model we denote the length of the MRT by $T \geq 0$ and measure it in seconds. Farmer and Skouras (2012) employ an MRT of 1 second to perform economic impact assessments. Government proposals suggest MRTs of 350 milliseconds and 500 milliseconds.

When the MM posts a limit sell order at time $t = 0$ she will not be able to cancel it before T , so faces the risk of the order becoming stale. During the interval $[0, T]$, the sell LO of the MM posted at $t = 0$ may be partially or fully filled by incoming buy MOs and by limit buy orders of other market participants. Recall that during this time window, other MMs are continuously posting buy LOs at the bid prices $S_t - \delta/2$, $t \in [0, T]$, so if

the fundamental price increases by δ the limit sell orders of the MM will be filled by buy LOs resting in the LOB.

Liquidity takers execute buy MOs that may be filled by the limit sell order posted by the MM. We denote by $\lambda^+ = (\lambda_t^+)_{t \geq 0}$ the fill rate of the sell LOs of the MM posted at $t = 0$, and we denote by N_t^+ the number of shares that the MM has sold by time t and the stochastic intensity of this counting process is λ_t^+ .

We denote by

$$\tau_a = \inf \{t \geq 0, S_t - S_0 = \delta\} \quad (4.2)$$

the first hitting time that the fundamental price has increased by δ .

If the fundamental price has not increased by δ before the MRT expires, the MM cancels any remaining unfilled part of the LO. Note that the MM's strategy is to post LOs at depth $\delta/2$ and only if $S_T = S_0$ would the agent prefer not to cancel the volume left in the LO, but the probability that $S_T = S_0$ is zero, so the MM cancels any remaining volume in the stale LO almost surely.

Furthermore, we denote by

$$\tau = \tau_a \wedge T \quad (4.3)$$

the time when the limit sell order posted by the MM is either completely filled or cancelled. Here the function $\cdot \wedge \cdot$ yields the minimum of the two arguments.

When $\tau = \tau_a$, the fundamental price has increased by δ before T . Until τ_a , the volume of shares the sell LO has filled by incoming buy MOs is $\min(N_{\tau_a}^+, M)$. Now, because at this hitting time there will be limit buy orders posted by other MMs at exactly the same price of the sell LO posted by the MM at $t = 0$, the remaining quantity $\max(M - N_{\tau_a}^+, 0)$ of shares is matched by those new LOs. In other words, if $\tau = \tau_a$, the whole limit sell order of volume M posted by the MM is filled before T .

The other case is when $\tau = T$, that is $S_u - S_0 < \delta$ for all $u \leq T$. Then the MM's limit sell order has filled $\min(N_T^+, M)$ shares, and she cancels any unfilled part.

Throughout the chapter, we work in the completed filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$, where the filtration is generated by the duplet (S_t, N_t^+) , which we define below.

4.3.1 Performance criterion: expected profit

We assume the MM marks-to-market the inventory at time τ and we break down the calculation of the expected profits of the MM in two components. One, we consider the case when $\tau = \tau_a$, which is when the fundamental price increases to levels where all volume posted by the LOs is consumed before the MM is able to cancel stale LOs. Two, when $\tau = T$ and the MM cancels any remaining liquidity posted at time $t = 0$.

When $\tau = \tau_a$, the MM marks-to-market her inventory at the fundamental price S_{τ_a} , which is equal to $S_0 + \delta$ (by the definition of τ_a), and denotes this value by

$$\begin{aligned}\Pi_1 &= \left[\left(S_0 + \frac{\delta}{2} \right) M - S_{\tau_a} M \right] \mathbb{1}_{\{\tau=\tau_a\}} \\ &= -\frac{\delta}{2} M \mathbb{1}_{\{\tau=\tau_a\}},\end{aligned}$$

where $\mathbb{1}$ is the indicator function.

When $\tau = T$, the mark-to-market value of the inventory is

$$\Pi_2 = \left[\min(N_T^+, M) \left(S_0 + \frac{\delta}{2} \right) - \min(N_T^+, M) S_T \right] \mathbb{1}_{\{\tau=T\}}. \quad (4.4)$$

The first term on the right-hand side of (4.4) denotes the cash obtained by the MM from selling shares at the price $S_0 + \delta/2$ per share during the interval $[0, T]$. The second term is the time $\tau = T$ mark-to-market value of the shares sold.

Hence, the mark-to-market value of the inventory is

$$\Pi = \Pi_1 + \Pi_2, \quad (4.5)$$

and the MM maximises the expected profit by solving

$$\max_{\delta \geq 0} \mathbb{E}[\Pi]. \quad (4.6)$$

Here we assume that the performance criterion employed by the MM is the expected profit obtained from posting LOs and that the MM does not penalise the variance of profits or accounts for inventory risk, see e.g., Cartea et al. (2015), Cartea et al. (2015). Our choice captures the essence of how MRTs affect the optimal posting of MMs without adding mathematical complexity to the model. Also, note that the timescale of the MRT is

seconds, so employing more realistic, yet mathematically more challenging, performance criteria will not add further insights. For example, we could assume that the performance criterion of the MM is expected utility of wealth or a mean-variance approach, both of which are employed in the extant literature by many authors, see Cheridito and Sepin (2014), Lorenz and Almgren (2011), Guéant (2015), Schied et al. (2010), Donnelly and Gan (2018).

Until now we have not made any assumptions about the dynamics of the stock price or the fill probabilities. This makes our framework versatile and the choice of price dynamics and model of fill probabilities can be adapted to the objective of the MM. Below we provide specific choices and we find a closed-form expression for $\mathbb{E}[\Pi_1]$. We cannot solve $\mathbb{E}[\Pi_2]$ in closed-form, so we employ a Feynman-Kac formula to derive the associated partial differential equation (PDE) and solve it using perturbation methods to obtain an asymptotic closed-form solution.

Fill rate probability. The fill rate of the LOs is given by

$$\lambda_t^+ = \lambda e^{-\kappa(S_0^a - S_t)}, \quad \text{for } S_0^a - S_t < \delta. \quad (4.7)$$

Here $\kappa > 0$ is the exponential rate of decay of the fill rate and $\lambda > 0$ is a reference fill rate, which we discuss below. Recall that the quantity of sell volume posted by the MM is M .

When $S_0^a - S_t < \delta$, the fill rate decays exponentially with respect to the depth of the limit sell order placed in the book at time 0. The choice of exponential decay, which captures qualitative properties of the fill rate, while keeping some mathematical tractability, is widely used in the literature. Cartea et al. (2014) and Guéant (2017) discuss the general conditions of the fill rate as a function of the depth of the LO, and look at two specific examples: exponential decay and power decay. The reference fill rate λ denotes the fill rate of the LO when the current fundamental price S_t is equal to the price of the limit order S_0^a , i.e., the particular case when $S_0^a = S_t$, so $\lambda_t^+ = \lambda$.

The LOs of the MM are stale in the interval $t \in (0, T]$ when $S_t \neq S_0$. During times when the fundamental price S_t is above the ask price S_0^a posted by the MM, the fill rate of the LO is greater than the reference rate, i.e., $\lambda_t^+ > \lambda$. This captures the increasing intensity of incoming MOs because fast traders snipe stale LOs.

Furthermore, LOs that do not rest at the best prices in the LOB have a much smaller fill rate than those posted at the best prices. Cartea et al. (2018) use data of eight stocks from a full month of trading in NASDAQ (January, 2014) and show that an MO walks beyond the best quote with probability between 0.001 and 0.09. This illustrates the order of the decay rate κ ; we return to this point below.

Dynamics of the fundamental price. The fundamental price of the stock follows an arithmetic Brownian motion:

$$S_t = S_0 + X_t = S_0 + \sigma W_t, \quad (4.8)$$

where $\sigma > 0$ is a constant volatility parameter.

We use data for 6 stocks traded in NASDAQ over 21 trading days in January 2014. The data are recorded at a millisecond frequency, and we use these data to illustrate the order of magnitude of parameters we employ in the numerical study. For each stock, we use data from 10am to 3pm of each trading day to avoid the open and close auctions. We estimate the parameter σ by calculating the volatility of the fundamental price (per second) and we use the arrival rate of buy MOs to estimate the reference fill rate λ (per second). Table 4.1 shows the mean of all daily estimates (and the standard deviation of the mean).

Symbol	$\hat{\lambda}_{MO}$	$std(\hat{\lambda}_{MO})$	$\hat{\sigma}^2$	$\sqrt{\hat{\sigma}^2}$	$std(\hat{\sigma}^2)$
EBAY	0.173684	0.083055	0.000102	0.010099	0.000055
INTC	0.085932	0.029271	0.000007	0.002645	0.000005
ORCL	0.072359	0.021808	0.000040	0.006324	0.000020
NTAP	0.069191	0.030184	0.000151	0.012288	0.000077
AMAT	0.033098	0.011564	0.000007	0.002645	0.000006
FMER	0.027130	0.007663	0.000140	0.011832	0.000067

Table 4.1: Estimates of σ^2 (per second) and λ (per second), with corresponding standard deviations.

In the table, the volatility of prices has the order of magnitude 10^{-2} . For the numerical experiments later we choose the rate of decay of the fill rate of the LOs to be $\kappa = 100$. This makes the decay of the fill probability consistent with the findings in Cartea et al. (2018) mentioned earlier.

Expected Profit. The expected profit from posting LOs results from computing

$$\mathbb{E}[\Pi_1] = \mathbb{E}\left[-\frac{\delta}{2} M \mathbb{1}_{\{\tau=\tau_a\}}\right]$$

and

$$\mathbb{E} [\Pi_2] = \mathbb{E} \left[\min(N_T^+, M) \left(-X_T + \frac{\delta}{2} \right) \mathbb{1}_{\{\tau=T\}} \right],$$

which are the expectations of the first and second terms in the right-hand side of (4.5). The first expectation is straightforward to calculate and is shown in the following proposition.

Proposition 4.3.1. *The expectation of Π_1 is given by*

$$\mathbb{E} [\Pi_1] = -\delta M \left(1 - \Phi \left(\frac{\delta}{\sigma \sqrt{T}} \right) \right), \quad (4.9)$$

where $\Phi(\cdot)$ denotes the cumulative density function of the standard normal distribution.

Proof.

$$\begin{aligned} \mathbb{E} [\Pi_1] &= \mathbb{E} \left[-\frac{\delta}{2} M \mathbb{1}_{\{\tau=\tau_a\}} \right] \\ &= -\frac{\delta}{2} M \mathbb{P}(\tau = \tau_a) \\ &= -\frac{\delta}{2} M \mathbb{P} \left(\max_{0 < t < T} \sigma W_t > \delta \right) \\ &= -\delta M \mathbb{P}(\sigma W_T > \delta) \\ &= -\delta M \left(1 - \Phi \left(\frac{\delta}{\sigma \sqrt{T}} \right) \right). \end{aligned}$$

□

To calculate the expectation of Π_2 we proceed as follows. Let

$$g(t, x, q) = \mathbb{E} \left[\min(N_T^+, M) \left(-X_T + \frac{\delta}{2} \right) \mathbb{1}_{\{\tau=T\}} \middle| X_t = x, N_t^+ = q \right].$$

Then, by the Feynman-Kac formula, see Proposition 12.5 in Cont and Tankov (2004), the function g satisfies the partial differential equation (PDE)

$$\partial_t g + \frac{1}{2} \sigma^2 \partial_{xx} g + \lambda e^{-\kappa(\frac{\delta}{2}-x)} [g(t, x, q+1) - g(t, x, q)] = 0, \quad (4.10)$$

with terminal and boundary conditions

$$g(T, x, q) = \min(q, M) \left(-x + \frac{\delta}{2} \right), \quad g(t, \delta, q) = 0, \quad x < \delta, \quad t \in [0, T]. \quad (4.11)$$

We are not able to find a closed-form solution for (4.10), but instead we seek approximate solutions by an asymptotic expansion. To asymptotically expand the solution, we first apply nondimensionalization to (4.10). We introduce a typical scale of δ , $\delta_0 = 0.01$, and apply the change of variables,

$$t' = \frac{\sigma^2 t}{\delta_0^2}, \quad \lambda' = \frac{\delta_0^2 \lambda}{\sigma^2}, \quad \delta' = \frac{\delta}{\delta_0}, \quad x' = \frac{x}{\delta_0}, \quad \kappa' = \delta_0 \kappa, \quad (4.12)$$

and

$$g'(t', x', q) = \frac{g(t, x, q)}{\delta_0}. \quad (4.13)$$

(4.10) then becomes

$$\partial_{t'} g' + \frac{1}{2} \partial_{x'x'} g' + \lambda' e^{-\kappa' \left(\frac{\delta'}{2} - x' \right)} [g'(t', x', q+1) - g'(t', x', q)] = 0, \quad (4.14)$$

with terminal and boundary conditions

$$g'(T', x', q) = \min(q, M) \left(-x' + \frac{\delta'}{2} \right), \quad g'(t', \delta', q) = 0, \quad x' < \delta', \quad t' \in [0, T'], \quad (4.15)$$

where $T' = \frac{\sigma^2 T}{\delta_0^2}$.

From Table 4.1 we see for liquid stocks traded in NASDAQ, the order of magnitude of the parameter λ is 0.1 per second and the order of magnitude of σ^2 is 10^{-4} . Hence λ' , which is dimensionless, has the order of magnitude 0.1. We asymptotically expand the function g' in λ' , so that

$$g'(t', x', q) = g'_0(t', x', q) + \lambda' g'_1(t', x', q) + \dots. \quad (4.16)$$

We substitute the expression for $g(t, x, q)$ given in (4.16) into the PDE (4.10) to obtain

$$\begin{aligned} 0 = & \partial_{t'} g'_0 + \frac{1}{2} \partial_{x'x'} g'_0 \\ & + \lambda' \partial_{t'} g'_1 + \frac{\lambda'}{2} \partial_{x'x'} g'_1 + \lambda' e^{-\kappa' \left(\frac{\delta'}{2} - x' \right)} [g'_0(q+1) - g'_0(q)] + \dots. \end{aligned}$$

Equate the terms of order 1 and λ' to zero and obtain the PDEs satisfied by g'_0 and g'_1 .

The first PDE is

$$\partial_{t'} g'_0 + \frac{1}{2} \partial_{x'x'} g'_0 = 0, \quad (4.17)$$

with terminal and boundary conditions

$$g'_0(T', x', q) = \min(q, M) \left(-x' + \frac{\delta'}{2} \right), \quad g'_0(t', \delta', q) = 0, \quad x' < \delta', \quad t' \in [0, T']. \quad (4.18)$$

The second PDE is

$$\partial_{t'} g'_1 + \frac{1}{2} \partial_{x'x'} g'_1 + e^{-\kappa' \left(\frac{\delta'}{2} - x' \right)} [g'_0(t', x', q+1) - g'_0(t', x', q)] = 0, \quad (4.19)$$

with terminal and boundary conditions

$$g'_1(T', x', q) = 0, \quad g'_1(t', \delta', q) = 0, \quad x' < \delta', \quad t' \in [0, T']. \quad (4.20)$$

We absorb the terminal and boundary conditions in the function g'_0 because they do not include the parameter λ' . Also, our calculations omit the dependence on q (treat it as a constant) because there is no differential term in q . The following two propositions provide expressions for the functions g'_0 and g'_1 .

Proposition 4.3.2. *Let the function $g'_0(t', x', q)$ satisfy (4.17) with boundary and terminal conditions (4.18), then*

$$g'_0(t', x', q) = \min(q, M) \left(-x' + \frac{3\delta'}{2} - \delta' \Phi \left(\frac{\delta' - x'}{\sqrt{T' - t'}} \right) \right). \quad (4.21)$$

Proof. For a proof see Section 4.8. □

Proposition 4.3.3. *Let $g'_1(t', x', q)$ satisfy (4.19) with terminal and boundary conditions*

(4.20), then

$$\begin{aligned}
g'_1 &= D(q, M) \exp\left(\frac{3\kappa'\delta'}{2} - \kappa'x'\right) \\
&\times \int_0^{\tilde{T}'} \exp\left(\frac{1}{2}\kappa'^2\tilde{s}\right) \left\{ -\sqrt{\frac{\tilde{s}}{2\pi}} \exp\left(-\frac{(x' - \delta' - \tilde{s}\kappa')^2}{2\tilde{s}}\right) \right. \\
&\quad - \left(x' - \tilde{s}\kappa' - \frac{\delta'}{2}\right) \Phi\left(\frac{x' - \delta' - \tilde{s}\kappa'}{\sqrt{\tilde{s}}}\right) \\
&\quad \left. + \delta' \Phi\left(\frac{x' - \delta' - \tilde{s}\kappa'}{\sqrt{\tilde{s}}}, \frac{x' - \delta' - \tilde{s}\kappa'}{\sqrt{\tilde{T}'}}; \sqrt{\frac{\tilde{s}}{\tilde{T}'}}\right) \right\} ds \\
&+ D(q, M) \exp\left(-\frac{\kappa'\delta'}{2} + \kappa'x'\right) \\
&\times \int_0^{\tilde{T}'} \exp\left(\frac{1}{2}\kappa'^2\tilde{s}\right) \left\{ \sqrt{\frac{\tilde{s}}{2\pi}} \exp\left(-\frac{(x' - \delta' + \tilde{s}\kappa')^2}{2\tilde{s}}\right) \right. \\
&\quad - \left(x' + \tilde{s}\kappa' - \frac{3\delta'}{2}\right) \Phi\left(-\frac{x' - \delta' + \tilde{s}\kappa'}{\sqrt{\tilde{s}}}\right) \\
&\quad \left. - \delta' \Phi\left(-\frac{x' - \delta' + \tilde{s}\kappa'}{\sqrt{\tilde{s}}}, -\frac{x' - \delta' + \tilde{s}\kappa'}{\sqrt{\tilde{T}'}}; \sqrt{\frac{\tilde{s}}{\tilde{T}'}}\right) \right\} ds,
\end{aligned}$$

where $\tilde{s} = T' - t' - s$, $\tilde{T}' = T' - t'$, $D(q, M) = \min(q + 1, M) - \min(q, M)$, and $\Phi(x, y; \rho)$ is the cumulative distribution function of the standard bivariate normal distribution with correlation ρ .

Proof. For a proof see Section 4.8. □

Now we show the accuracy of the asymptotic expansion.

Theorem 4.3.1. Accuracy of asymptotic expansion. *When λ' is small,*

$$g'_e(t', x', q) := g'(t', x', q) - g'_0(t', x', q) - \lambda' g'_1(t', x', q) = o(\lambda').$$

Proof. For a proof see Section 4.8. □

4.3.2 Value function

From now on we work with the approximation of the expected profit function of the MM.

Hence, we define

$$G_1(\delta; M, T, \sigma) = -\delta M \left(1 - \Phi\left(\frac{\delta}{\sigma\sqrt{T}}\right)\right). \quad (4.22)$$

For the second expectation (4.4), we change the variables back to the dimensional version and define

$$g_0(t, x, q) = \delta_0 g'_0(t', x', q), \quad g_1(t, x, q) = \delta_0 g'_1(t', x', q).$$

With Proposition 4.3.2 and 4.3.3, we define

$$\begin{aligned} G_2 &= g_0(0, 0, 0; \delta, M, T, \sigma) + \lambda g_1(0, 0, 0; \delta, M, T, \sigma) \\ &= \lambda g_1(0, 0, 0; \delta, M, T, \sigma) \\ &= \lambda \exp\left(\frac{3\kappa\delta}{2}\right) \int_0^T \exp\left(\frac{1}{2}\sigma^2\kappa^2\tilde{s}\right) \left\{ -\sqrt{\frac{\sigma^2\tilde{s}}{2\pi}} \exp\left(-\frac{(\delta + \sigma^2\tilde{s}\kappa)^2}{2\sigma^2\tilde{s}}\right) \right. \\ &\quad \left. + \left(\sigma^2\tilde{s}\kappa + \frac{\delta}{2}\right) \Phi\left(\frac{-\delta - \sigma^2\tilde{s}\kappa}{\sigma\sqrt{\tilde{s}}}\right) \right. \\ &\quad \left. + \delta \Phi\left(\frac{-\delta - \sigma^2\tilde{s}\kappa}{\sigma\sqrt{\tilde{s}}}, \frac{-\delta - \sigma^2\tilde{s}\kappa}{\sigma\sqrt{T}}; \sqrt{\frac{\tilde{s}}{T}}\right) \right\} ds \\ &\quad + \lambda \exp\left(-\frac{\kappa\delta}{2}\right) \int_0^T \exp\left(\frac{1}{2}\sigma^2\kappa^2\tilde{s}\right) \left\{ \sqrt{\frac{\sigma^2\tilde{s}}{2\pi}} \exp\left(-\frac{(\delta - \sigma^2\tilde{s}\kappa)^2}{2\sigma^2\tilde{s}}\right) \right. \\ &\quad \left. - \left(\sigma^2\tilde{s}\kappa - \frac{3\delta}{2}\right) \Phi\left(\frac{\delta - \sigma^2\tilde{s}\kappa}{\sigma\sqrt{\tilde{s}}}\right) \right. \\ &\quad \left. - \delta \Phi\left(\frac{\delta - \sigma^2\tilde{s}\kappa}{\sigma\sqrt{\tilde{s}}}, \frac{\delta - \sigma^2\tilde{s}\kappa}{\sigma\sqrt{T}}; \sqrt{\frac{\tilde{s}}{T}}\right) \right\} ds, \end{aligned}$$

where $\tilde{s} = T - s$.

Therefore, the value function is given by

$$G(\delta; M, T, \lambda, \sigma) = G_1(\delta; M, T, \sigma) + G_2(\delta; M, T, \lambda, \sigma). \quad (4.23)$$

In the sequel, we refer to G as the expected profit $\mathbb{E}[\Pi]$. Note that the integral to compute G is over a bounded interval and the integrand is continuous and bounded in s , thus one can use any standard numerical integration method to compute the integral.

The following proposition shows that the expected profit G is decreasing in the volume of the LO, hence the MM will obtain a higher expected profit by posting an LO with less volume. If all MMs are identical, then the optimal quantity to post is $M = 1$ for all MMs. This shows that the MMs will provide less liquidity to protect themselves from loss in expected profit arising from LOs becoming stale due to the MRT.

Proposition 4.3.4. *For $M \geq 1$,*

$$G(\delta; M, T, \lambda, \sigma) \geq G(\delta; M + 1, T, \lambda, \sigma). \quad (4.24)$$

Proof. Directly from the definition of the function G . □

Finally, the next proposition shows that the value function is bounded and attains a maximum.

Proposition 4.3.5.

$$G(0) = 0, \quad \lim_{\delta \rightarrow +\infty} G(\delta) = 0, \quad (4.25)$$

and $G(\delta)$ is bounded and attains a maximum.

Proof. Directly from the definition of G and because G is continuous in δ . □

Note that (4.25) shows that if the MM posts the LO at the fundamental price, then the expected profit is zero. Also, if the depth of the LO is arbitrarily large, then the expected profit is zero because the LO will never be filled.

Finally, we write the MM's optimisation problem posed in (4.6) as

$$\delta^* = \operatorname{argmax}_{\delta \in [0, +\infty)} G(\delta). \quad (4.26)$$

So far we cannot prove the uniqueness of the maximum, but for the range of the parameters we employ in the numerical study below, the maximum is unique. Therefore we assume δ^* is well defined and we obtain it via numerical optimisation.

4.3.3 Numerical study and simulations

Figure 4.1 plots the expected profit G (red solid line) and the mean profit of 2,000 simulations with standard errors. We observe that the value function obtained using (4.23) and the expected profits obtained from simulations coincide — this lends strong support to the accuracy of the asymptotic expansion used to approximate the expected profit function.

The figure shows that the expected profit is zero when the depth of the LOs is zero because sell LOs posted by the MM are immediately matched with other buy LOs, resulting in zero PnL for the MM. Moreover, there is a range of δ where the depth of the LO is so

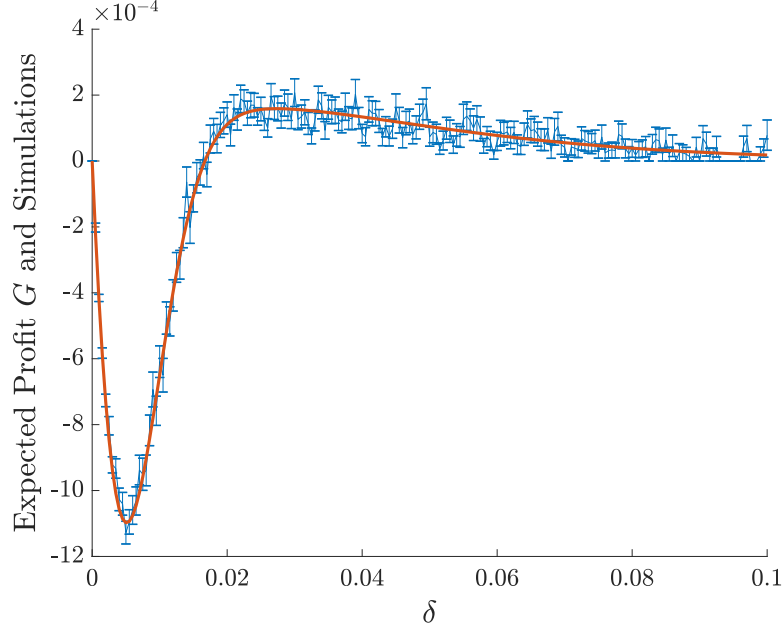


Figure 4.1: Expected profit and 2,000 simulations (standard deviation of each simulation is also shown). MMs post LOs of same volume. Parameters: $\kappa = 10^2$, $M = 1$, $T = 0.5$ seconds, $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$ and $\lambda = 0.1/\text{second}$.

small that the MM incurs expected losses. These losses arise because the probability of being matched at a loss before the MRT expires is high.

As the depth of the LO increases, the expected profit increases and becomes positive. In other words, when the MM posts deeper in the book, the probability that all volume in the stale LO is sniped by MOs and filled by other buy LOs decreases. Finally, from the figure we see there is a value of δ that maximises the expected profit — in the sequel we denote this value by δ^* .

As we increase the value of δ further, the expected profit starts to decrease and converges to zero as δ goes to infinity. This is because the fill rate decreases as the depth of the LO increases, so the LOs of the MM are hardly ever filled, see Proposition 4.3.5.

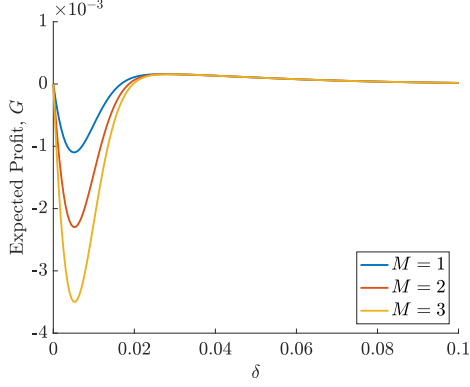


Figure 4.2: Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1/\text{second}$, $T = 0.5$ seconds.

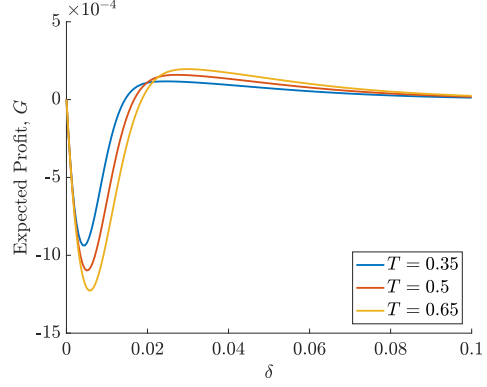


Figure 4.3: Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1/\text{second}$, $M = 1$.

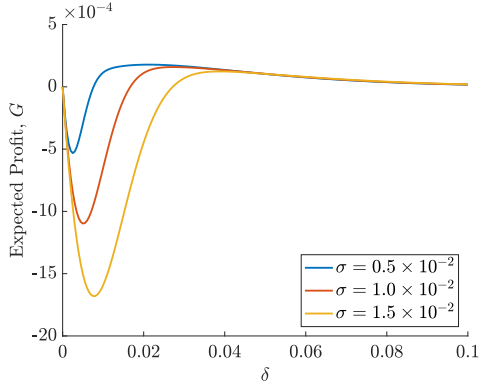


Figure 4.4: Expected profit. Parameters: $T = 0.5$ seconds, $\lambda = 0.1/\text{second}$, $M = 1$.

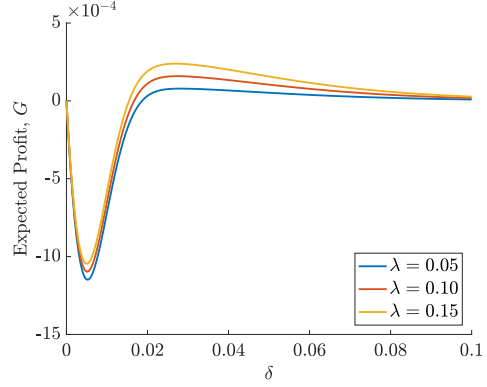


Figure 4.5: Expected profit. Parameters: $T = 0.5$ seconds, $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $M = 1$.

Figures 4.2 to 4.5 show the expected profit for various parameter values. For example, Figure 4.2 shows that the expected profit decreases as the volume of the LO posted by the MM increases, which agrees with Proposition 4.3.4. Clearly, as the MM posts LOs of higher volumes, the exposure to being picked off is higher.

Figure 4.3 shows that the range of δ for which the expected profit is negative is wider as the MRT increases. The intuition for this result is as follows. For small δ and longer MRT, the postings of the MM are more exposed to losses that stem from stale quotes and it is often the case that all volume in the LO is filled at stale prices before the end of the MRT.

The expected profit of the MM decreases as volatility increases, see Figure 4.4. Everything else being equal, when volatility is high, the fundamental price will fluctuate more and it

is more likely to observe an increase in the fundamental price to levels where all volume is traded at a loss.

Finally, Figure 4.5 shows the effect of the reference fill rate λ has on the expected profit of the MM. As the value of the parameter λ increases, the chances of being picked off diminish and when picked off, the losses are also smaller because most of the volume of the LO is filled by incoming buy MOs.

4.4 Model II: Limit orders of various volumes

In the setup discussed above we assumed that all MMs are identical and, in particular, we assumed that all MMs post LOs of the same volume M and therefore at the same depth $\delta^*(M)/2$; this notation stresses that the optimal depth $\delta^*(M)/2$ is a function of the volume of shares posted in the LO.

In this section we extend the Model I, so that MMs may post LOs of any positive integer volumes and we also discuss how they choose the optimal volume. The MMs are identical in all other aspects. Thus, MMs with LOs of same volume will post the LOs at the same depth. We denote the depth, for a given volume (not necessarily the profit maximising depth) by $\hat{\delta}(M)/2$. For clarity we employ the hat notation when we refer to the model in which MMs choose the volume of the LOs.

As above, we derive an expression to approximate the expected profit of the MM as a function of $\hat{\delta}$. The approximation to the expected profits is given by the value function $\widehat{G}(\hat{\delta})$ and we obtain the optimal depth of LOs, for a fixed volume M , by solving

$$\hat{\delta}^*(M) = \operatorname{argmax}_{\hat{\delta} \in [0, +\infty)} \widehat{G}(\hat{\delta}; M). \quad (4.27)$$

This is the analogue of (4.26) and our notation in (4.27) emphasises that MMs choose the volume of the LO.

As above in Model I, we consider the case of limit sell orders. We recall the sequence of events the MM faces. After the MM posts a limit sell order of volume M at price $S_0 + \hat{\delta}(M)/2$, she cannot cancel it until the MRT expires. There will be instances in which the fundamental price has increased by an amount such that any remaining volume of the limit sell order of the MM is matched with the buy LOs posted by other MMs.

In Model I, all MMs post LOs of same volume at the same depth $\delta/2$, so the increase in price required for the limit sell order to be matched (i.e., filled by other buy LOs) is δ . However, if other MMs post LOs with different volumes at corresponding optimal depths, the price increase is $\hat{\delta}(M)/2$ plus the smallest posted depth of all other LOs. Thus, the ordering of the optimal posted depth is important to determine when the LO of the MM is completely picked off by the limit buy orders of other MMs.

We make the following conjecture.

Conjecture 4.4.1. *Optimal depth for LOs with volume M . The best offers in the LOB are of volume $M = 1$, i.e.,*

$$\hat{\delta}^*(M) > \hat{\delta}^*(1), \quad \text{for } M > 1. \quad (4.28)$$

Based on Conjecture 4.4.1, the MMs whose LOs are of volume $M = 1$ will choose the optimal depth of their LOs as in the model described above in Section 4.3. We stress that the important step in computing the expected profits of the MMs is the quantity by which the fundamental price has to increase, so that the limit buy orders of other MMs fill any outstanding volume of previously posted limit sell orders. Thus, we have the following proposition.

Proposition 4.4.1. *Optimal depth for LOs with volume $M = 1$. The optimal depth for LOs with volume $M = 1$ in a market with LOs of all positive integer volumes is the same as that obtained in a market where all LOs are of volume $M = 1$, i.e.,*

$$\hat{\delta}^*(1) = \delta^*(1), \quad (4.29)$$

where $\delta^*(1)$ is derived above, see (4.26).

Proof. From Conjecture 4.4.1 and the setup of the two models. □

For simplicity of notation, from now on δ_1^* is short-hand notation for $\delta^*(1)$.

Therefore, sell LOs of volume $M > 1$ posted at depth $\hat{\delta}(M)/2$ are matched with buy LOs of volume 1 when the fundamental price increases (relative to S_0 when the LO was posted) by the amount $\hat{\delta}(M)/2 + \delta_1^*/2$.

For $M > 1$, the first hitting time for the LO to be matched when $\hat{\delta}(M) > \delta_1^*$, $M > 1$ is given by

$$\hat{\tau}_a = \inf \left\{ t \geq 0, S_t - S_0 = \hat{\delta}(M)/2 + \delta_1^*/2 \right\}. \quad (4.30)$$

Based on Conjecture 4.4.1, we expect to attain the maximum expected profit at $\hat{\delta}(M) > \delta_1^*$ when $M > 1$.

We denote by

$$\hat{\tau} = \hat{\tau}_a \wedge T \quad (4.31)$$

the time when the sell LO posted by the MM is either completely filled or cancelled. Also, as above in Model I, we assume that the MM marks her inventory to market at time $\hat{\tau}$, which is also the reference time employed to calculate the expected profit.

When $\hat{\tau} = \hat{\tau}_a$, the fundamental price has gone up by $\hat{\delta}(M)/2 + \delta_1^*/2$ before the time elapsed since posting the LO hits the MRT. By the time $\hat{\tau}_a$, the volume of the sell LO filled is $\min(N_{\hat{\tau}_a}^+, M)$ shares. At this hitting time there will be buy LOs posted by other MMs at exactly the same price as the sell LO posted by the MM at $t = 0$, the remaining quantity $\max(M - N_{\hat{\tau}_a}^+, 0)$ of the LO posted is matched by those new LOs. In other words, the whole limit sell order of volume M is filled before T if $\hat{\tau} = \hat{\tau}_a$.

The other case is when $\hat{\tau} = T$, that is $S_u - S_0 < \hat{\delta}(M)/2 + \delta_1^*/2$ for $u \leq T$. The limit sell order of the MM fills $\min(N_T^+, M)$ shares and she cancels any unfilled volume left in the LO.

The next sections follow the same steps as those in Model I. We show how to compute the expected profits of the MM and how the optimal depth of the LO depends on various parameters of the model.

4.4.1 Performance criterion: expected profit

When $\hat{\tau} = \hat{\tau}_a$, the mark-to-market value of the inventory at the fundamental price $S_{\hat{\tau}_a} = S_0 + \hat{\delta}(M)/2 + \delta_1^*/2$ is

$$\hat{\Pi}_1 = \left[M \left(-\frac{\delta_1^*}{2} \right) \right] \mathbb{1}_{\{\hat{\tau}=\hat{\tau}_a\}}, \quad (4.32)$$

and when $\hat{\tau} = T$, the mark-to-market value of the inventory is

$$\hat{\Pi}_2 = \left[\min(N_T^+, M) \left(-X_T + \frac{\hat{\delta}(M)}{2} \right) \right] \mathbb{1}_{\{\hat{\tau}=T\}}. \quad (4.33)$$

The MM maximises the expected profit by solving

$$\max_{\hat{\delta}} \mathbb{E} \left[\hat{\Pi}_1 + \hat{\Pi}_2 \right]. \quad (4.34)$$

As above, to compute $\mathbb{E} \left[\widehat{\Pi}_2 \right]$ we employ a Feynman-Kac formula to derive the associated PDE and solve it using perturbation methods to obtain an asymptotic solution.

Proposition 4.4.2.

$$\mathbb{E} \left[\widehat{\Pi}_1 \right] = -M \delta_1^* \left(1 - \Phi \left(\frac{\hat{\delta}(M) + \delta_1^*}{2 \sigma \sqrt{T}} \right) \right). \quad (4.35)$$

Recall that $\Phi(\cdot)$ denotes the cumulative density function of a standard normal random variable.

Proof. Similar to Proposition 4.3.1. For a proof see Section 4.8. \square

To calculate $\mathbb{E} \left[\widehat{\Pi}_2 \right]$ we proceed as follows. Let

$$\hat{g}(t, x, q) = \mathbb{E} \left[\min(N_T^+, M) \left(-X_T + \frac{\hat{\delta}}{2} \right) \mathbb{1}_{\{\tilde{\tau}=T\}} \middle| X_t = x, N_t^+ = q \right].$$

Then, by the Feynman-Kac formula, \hat{g} satisfies the PDE

$$\partial_t \hat{g} + \frac{1}{2} \sigma^2 \partial_{xx} \hat{g} + \lambda e^{-\kappa \left(\frac{\hat{\delta}}{2} - x \right)} [\hat{g}(t, x, q+1) - \hat{g}(t, x, q)] = 0, \quad (4.36)$$

with boundary and terminal conditions

$$\hat{g}(T, x, q) = \min(q, M) \left(-x + \frac{\hat{\delta}}{2} \right), \quad \hat{g}(t, \tilde{\delta}, q) = 0, \quad x < \tilde{\delta}, \quad t \in [0, T], \quad (4.37)$$

where

$$\tilde{\delta} = \frac{\hat{\delta}(M) + \delta_1^*}{2}.$$

We introduce the same scale δ_0 as we did for (4.10) and apply the same change of variables,

$$t' = \frac{\sigma^2 t}{\delta_0^2}, \quad \lambda' = \frac{\delta_0^2 \lambda}{\sigma^2}, \quad \hat{\delta}' = \frac{\hat{\delta}}{\delta_0}, \quad \delta_1^{*'} = \frac{\delta_1^*}{\delta_0}, \quad \tilde{\delta}' = \frac{\tilde{\delta}}{\delta_0}, \quad x' = \frac{x}{\delta_0}, \quad \kappa' = \delta_0 \kappa, \quad (4.38)$$

and

$$\hat{g}'(t', x', q) = \frac{g(t, x, q)}{\delta_0}. \quad (4.39)$$

As in (4.16), we asymptotically expand g in λ , so

$$\hat{g}'(t', x', q) = \hat{g}'_0(t', x', q) + \lambda' \hat{g}'_1(t', x', q) + \dots. \quad (4.40)$$

We substitute the above expansion of \hat{g}' in PDE (4.36), and by equating to zero the terms of order 1 and the terms of order λ' we obtain PDEs for \hat{g}'_0 and \hat{g}'_1 . We absorb the boundary and terminal conditions in \hat{g}'_0 because they do not include the parameter λ' . The first PDE is

$$\partial_{t'} \hat{g}'_0 + \frac{1}{2} \partial_{x'x'} \hat{g}'_0 = 0, \quad (4.41)$$

with terminal and boundary conditions

$$\hat{g}'_0(T', x', q) = \min(q, M) \left(-x' + \frac{\hat{\delta}'}{2} \right), \quad \hat{g}_0(t', \tilde{\delta}', q) = 0, \quad x' < \tilde{\delta}', \quad t' \in [0, T'], \quad (4.42)$$

where $T' = \frac{\sigma^2 T}{\delta_0^2}$, and the second PDE is

$$\partial_{t'} \hat{g}'_1 + \frac{1}{2} \partial_{x'x'} \hat{g}'_1 + e^{-\kappa' \left(\frac{\delta'}{2} - x' \right)} [\hat{g}'_0(t', x', q+1) - \hat{g}'_0(t', x', q)] = 0, \quad (4.43)$$

with terminal and boundary conditions

$$\hat{g}'_1(T', x', q) = 0, \quad \hat{g}'_1(t', \tilde{\delta}', q) = 0, \quad x' < \tilde{\delta}', \quad t' \in [0, T']. \quad (4.44)$$

In the following calculations we omit the dependence of q (treat it as a constant) because there is no differential term in q .

Proposition 4.4.3. *Let \hat{g}'_0 satisfy (4.41) with terminal and boundary conditions (4.42), then*

$$\hat{g}'_0(t', x', q) = \min(q, M) \left(-x' + \frac{\hat{\delta}'(M)}{2} + \delta_1^{*'} - \delta_1^{*'} \Phi \left(\frac{\tilde{\delta}' - x'}{\sqrt{T' - t'}} \right) \right). \quad (4.45)$$

Proof. Similar to Proposition 4.3.2. For a proof see Section 4.8. □

Proposition 4.4.4. *Let \hat{g}'_1 satisfy (4.43) with terminal and boundary conditions in (4.44),*

then

$$\begin{aligned}
\hat{g}'_1 &= D(q, M) \exp\left(\frac{\kappa' \delta_1^{*'}}{2} - \kappa'(x' - \tilde{\delta}') \\
&\times \int_0^{\tilde{T}'} \exp\left(\frac{1}{2} \kappa'^2 \tilde{s}\right) \left\{ -\sqrt{\frac{\tilde{s}}{2\pi}} \exp\left(-\frac{(x' - \tilde{\delta}' - \tilde{s} \kappa')^2}{2\tilde{s}}\right) \right. \\
&\quad - \left(x' - \tilde{\delta}' - \tilde{s} \kappa' + \frac{\delta_1^{*'}}{2}\right) \Phi\left(\frac{x' - \tilde{\delta}' - \tilde{s} \kappa'}{\sqrt{\tilde{s}}}\right) \\
&\quad \left. + \delta_1^{*'} \Phi\left(\frac{x' - \tilde{\delta}' - \tilde{s} \kappa'}{\sqrt{\tilde{s}}}, \frac{x' - \tilde{\delta}' - \tilde{s} \kappa'}{\sqrt{\tilde{T}'}}; \sqrt{\frac{\tilde{s}}{\tilde{T}'}}\right) \right\} ds \\
&+ D(q, M) \exp\left(\frac{\kappa' \delta_1^{*'}}{2} + \kappa'(x' - \tilde{\delta}') \\
&\times \int_0^{\tilde{T}'} \exp\left(\frac{1}{2} \kappa'^2 \tilde{s}\right) \left\{ \sqrt{\frac{\tilde{s}}{2\pi}} \exp\left(-\frac{(x' - \tilde{\delta}' + \tilde{s} \kappa')^2}{2\tilde{s}}\right) \right. \\
&\quad - \left(x' - \tilde{\delta}' + \tilde{s} \kappa' - \frac{\delta_1^{*'}}{2}\right) \Phi\left(-\frac{x' - \tilde{\delta}' + \tilde{s} \kappa'}{\sqrt{\tilde{s}}}\right) \\
&\quad \left. - \delta_1^{*'} \Phi\left(-\frac{x' - \tilde{\delta}' + \tilde{s} \kappa'}{\sqrt{\tilde{s}}}, -\frac{x' - \tilde{\delta}' + \tilde{s} \kappa'}{\sqrt{\tilde{T}'}}; \sqrt{\frac{\tilde{s}}{\tilde{T}'}}\right) \right\} ds,
\end{aligned}$$

where $\tilde{T}' = T' - t'$ and $\tilde{s} = T' - t' - s$.

Proof. Similar to Proposition 4.3.3. For a proof see Section 4.8. \square

4.4.2 Value function

For $M \geq 2$ we define

$$\widehat{G}_1(\hat{\delta}; M, T, \sigma) = -M \delta_1^* \left(1 - \Phi\left(\frac{\hat{\delta} + \delta_1^*}{2\sigma\sqrt{T}}\right)\right). \quad (4.46)$$

For the expectation (4.33) we change the variables back to the dimensional version and define

$$\hat{g}_0(t, x, q) = \delta_0 \hat{g}'_0(t', x', q), \quad \hat{g}_1(t, x, q) = \delta_0 \hat{g}'_1(t', x', q).$$

With Proposition 4.4.3 and 4.4.4 we define

$$\begin{aligned}
\widehat{G}_2 &= \hat{g}_0(0, 0, 0; \delta, M, T, \sigma) + \lambda \hat{g}_1(0, 0, 0; \delta, M, T, \sigma) \\
&= \lambda \hat{g}_1(0, 0, 0; \delta, M, T, \sigma) \\
&= \lambda \exp\left(\frac{\kappa \delta_1^*}{2} + \kappa \tilde{\delta} + \frac{1}{2} \kappa^2 \sigma^2 T\right) \\
&\quad \times \int_0^T \exp\left(-\frac{1}{2} \sigma^2 \kappa^2 s\right) \left\{ -\sqrt{\frac{\sigma^2 \tilde{s}}{2\pi}} \exp\left(-\frac{(-\tilde{\delta} - \sigma^2 \tilde{s} \kappa)^2}{2 \sigma^2 \tilde{s}}\right) \right. \\
&\quad \left. - \left(-\tilde{\delta} - \sigma^2 \tilde{s} \kappa + \frac{\delta_1^*}{2}\right) \Phi\left(\frac{-\tilde{\delta} - \sigma^2 \tilde{s} \kappa}{\sigma \sqrt{\tilde{s}}}\right) \right. \\
&\quad \left. + \delta_1^* \Phi\left(\frac{-\tilde{\delta} - \sigma^2 \tilde{s} \kappa}{\sigma \sqrt{\tilde{s}}}, \frac{-\tilde{\delta} - \sigma^2 \tilde{s} \kappa}{\sigma \sqrt{T}}; \sqrt{\frac{\tilde{s}}{T}}\right) \right\} ds \\
&\quad + \lambda \exp\left(\frac{\kappa \delta_1^*}{2} - \kappa \tilde{\delta} + \frac{1}{2} \kappa^2 \sigma^2 T\right) \\
&\quad \times \int_0^T \exp\left(-\frac{1}{2} \sigma^2 \kappa^2 s\right) \left\{ \sqrt{\frac{\sigma^2 \tilde{s}}{2\pi}} \exp\left(-\frac{(-\tilde{\delta} + \sigma^2 \tilde{s} \kappa)^2}{2 \sigma^2 \tilde{s}}\right) \right. \\
&\quad \left. - \left(-\tilde{\delta} + \sigma^2 \tilde{s} \kappa - \frac{\delta_1^*}{2}\right) \Phi\left(-\frac{-\tilde{\delta} + \sigma^2 \tilde{s} \kappa}{\sigma \sqrt{\tilde{s}}}\right) \right. \\
&\quad \left. - \delta_1^* \Phi\left(-\frac{-\tilde{\delta} + \sigma^2 \tilde{s} \kappa}{\sigma \sqrt{\tilde{s}}}, -\frac{-\tilde{\delta} + \sigma^2 \tilde{s} \kappa}{\sigma \sqrt{T}}; \sqrt{\frac{\tilde{s}}{T}}\right) \right\} ds,
\end{aligned}$$

where $\tilde{s} = T - t$.

Therefore the value function is given by

$$\widehat{G}(\hat{\delta}; M, T, \lambda, \sigma) = \widehat{G}_1(\hat{\delta}; M, T, \sigma) + \widehat{G}_2(\hat{\delta}; M, T, \lambda, \sigma), \quad (4.47)$$

which is an approximation to the expected mark-to-market inventory $\mathbb{E}[\widehat{\Pi}]$ and is the value function the MM maximises by choosing the optimal $\hat{\delta}$. From now on we refer to \widehat{G} as the expected profit $\mathbb{E}[\widehat{\Pi}]$.

The proposition below shows that \widehat{G} is decreasing in M , so the expected profit is higher when the MM posts a LO with less volume for $M > 2$. Therefore, provided there are other MMs who post LOs with volume 1, the MM prefers to post a LO with volume $M = 2$ instead of volume $M > 2$. We cannot compare theoretically with the case of $M = 1$, because the value function is different, but from the numerical results below, we show that all MMs choose to post LOs with volume $M = 1$.

Proposition 4.4.5. *For $M \geq 2$,*

$$\widehat{G}(\hat{\delta}; M, T, \lambda, \sigma) \geq \widehat{G}(\hat{\delta}; M + 1, T, \lambda, \sigma). \quad (4.48)$$

Proof. Directly from the definition of \widehat{G} . □

The following proposition shows properties of the expected profit.

Proposition 4.4.6.

$$\widehat{G}(0) \text{ is bounded, } \lim_{\hat{\delta} \rightarrow +\infty} \widehat{G}(\hat{\delta}) = 0, \quad (4.49)$$

and $\widehat{G}(\delta)$ is bounded.

Proof. Directly from the definition of \widehat{G} and because \widehat{G} is continuous in $\hat{\delta}$. □

So far we cannot prove the uniqueness of the maximum, but for the range of parameters we employ in the numerical study below, the maximum is unique. Thus,

$$\hat{\delta}^* = \operatorname{argmax}_{\hat{\delta} \in [\delta_1^*, +\infty)} \widehat{G}(\hat{\delta}). \quad (4.50)$$

4.4.3 Numerical study and simulations

Figure 4.6 plots the expected profit \widehat{G} (red solid line) and the mean profit of 2,000 simulations with standard errors. Figures 4.7 to 4.10 show the expected profit faced by the MM for a range of values of the model parameters. The interpretation of the figures is similar to that of Figures 4.2 to 4.5.

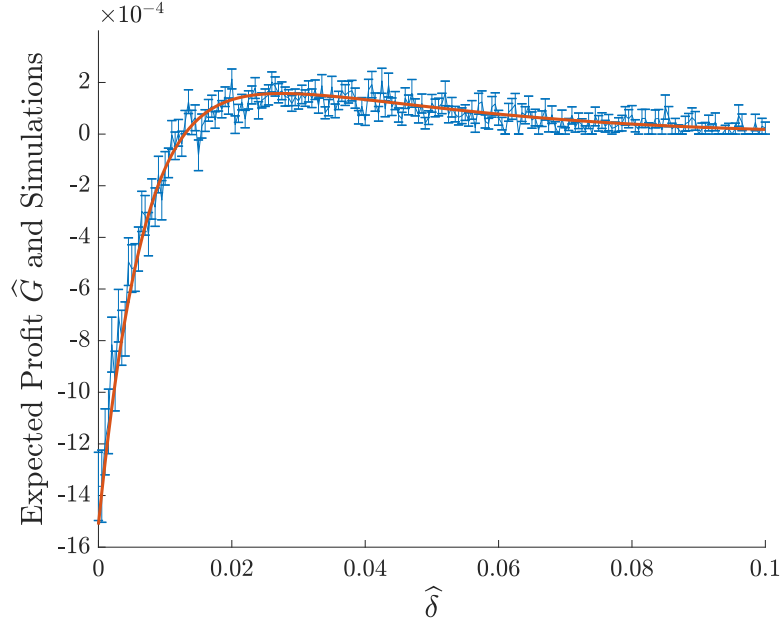


Figure 4.6: Expected profit and 2,000 simulations (standard deviation of each simulation is also shown). The other MMs post LOs of various volumes. $M = 2$, $T = 0.5$ seconds, $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1/\text{second}$.

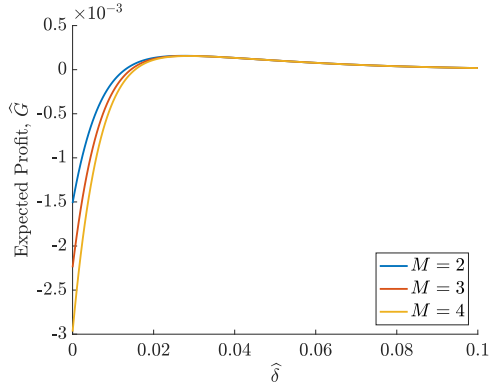


Figure 4.7: Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1/\text{second}$, $T = 0.5$ seconds.

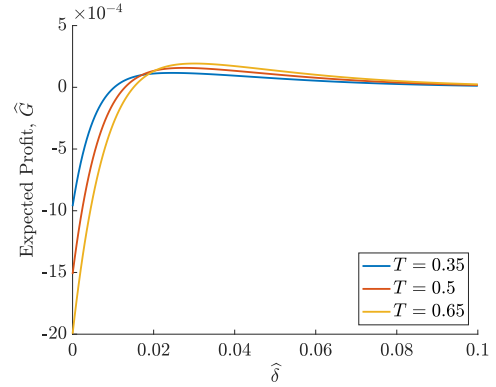


Figure 4.8: Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1/\text{second}$, $M = 2$.

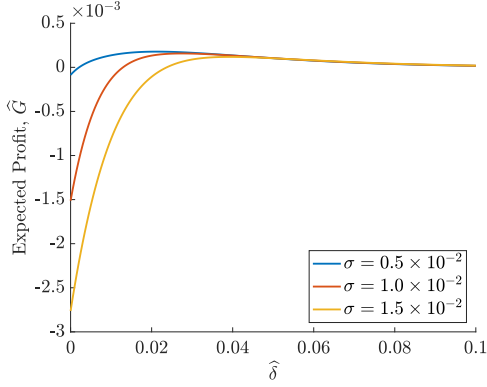


Figure 4.9: Expected profit. Parameters: $T = 0.5$ seconds, $\lambda = 0.1/\text{second}$, $M = 2$.

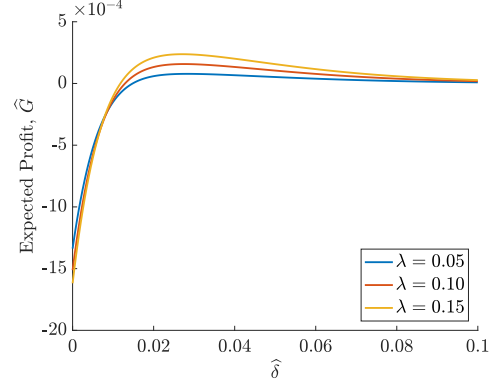


Figure 4.10: Expected profit. Parameters: $T = 0.5$ seconds, $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $M = 2$.

4.5 Optimal depth

In this section we discuss the depths that result from the strategy developed in Section 4.3, where we assume that the LOs posted by MMs are always of the same volume, i.e., Model I, and those that result from the strategy developed in Section 4.4, where we assume that MMs post LOs of any positive integer volume, i.e., Model II.

Figure 4.11 shows the optimal depth (y -axis) of a LO with volume M (x -axis), other parameters fixed at $\sigma = 10^{-2}$, $\lambda = 0.1$ per second, $T = 0.5$ seconds. The dots (resp. stars) correspond to depths of the LOs in Model I (resp. Model II). In both models the depth of the LO increases as the volume posted in the LO increases. Recall that in our notation the depth of the LO is $\delta/2$, which is the ‘distance’ from the fundamental price of the traded asset to the price at which the MM is willing to trade with a LO. The intuition for this result is as follows. The larger the volume of the LO, the higher the expected losses due to stale quotes being picked off. Thus, the MM posts deeper in the book to decrease the probability of being filled and to protect herself from loss-leading stale quotes. We also see that for small values of the volume, the optimal depth in Model I is larger than that resulting from Model II. As M increases, the ordering in the magnitude of the optimal depths for the two models is reversed. Finally, the figure also shows that for $M = 1$, the optimal depth in Models I and II coincides.

Figure 4.12 shows that the optimal depth of the LOs increases as the MRT increases (other parameters fixed at $\sigma = 10^{-2}$, $\lambda = 0.1$ per second, $M = 3$). As the MRT increases, it

is more likely that sell LOs become stale, because they are forced to rest in the book, and other traders pick off these orders. Thus, the MM posts the LOs deeper in the book to decrease the probability of fills to protect herself from financial losses. In Figure 4.12 volume is held fixed at $M = 3$ and we see that the optimal depth in Model I is larger than that in Model II.

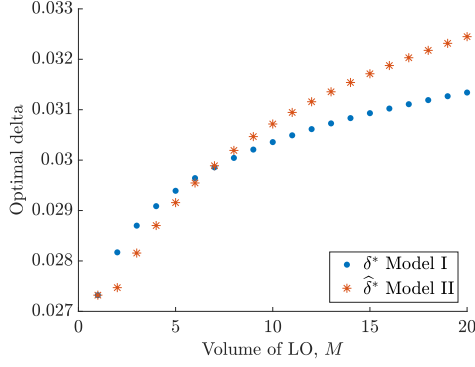


Figure 4.11: Optimal delta. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1/\text{second}$, $T = 0.5$ seconds.

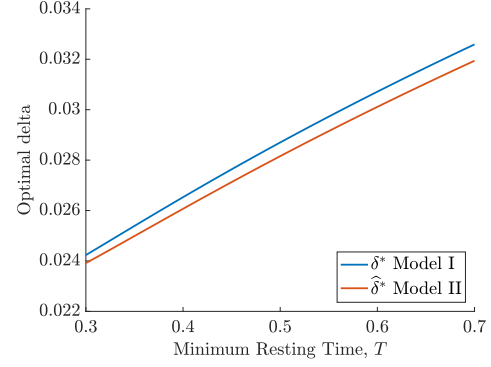


Figure 4.12: Optimal delta. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1/\text{second}$, $M = 3$.

Figure 4.13 shows that the optimal depth increases as volatility increases. As volatility increases the chance of LOs being matched also increases. Thus, the MM posts the LO deeper in the book to mitigate losses.

Figure 4.14 shows that the optimal depth decreases as the reference fill rate increases. This is because there are, on average, more incoming MOs that fill the LO, so the probability of being picked off by limit buy order decreases.

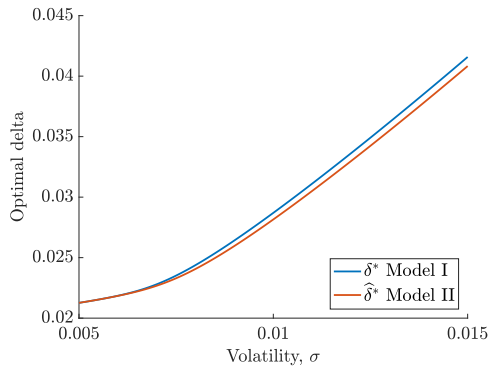


Figure 4.13: Optimal delta. Parameters: $T = 0.5$ seconds, $\lambda = 0.1/\text{second}$, $M = 3$.

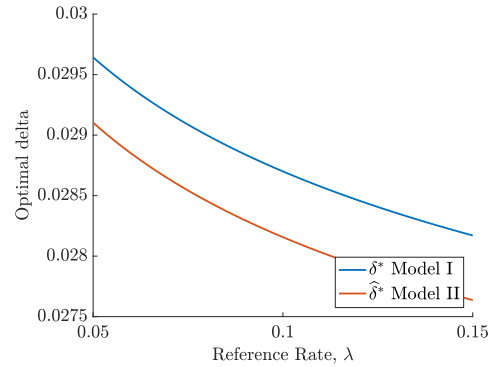


Figure 4.14: Optimal delta. Parameters: $T = 0.5$ seconds, $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $M = 3$.

4.6 Expected profits and liquidity provision

Figures 4.15 to 4.18 show the expected profits obtained by the MM for different values of: posted volume, MRT, volatility, and reference fill rate.

Figure 4.15 shows that the expected profits decrease as the volume of LO increases when the MM employs the optimal posting strategy. Recall that for Model I we showed that the expected profit is decreasing in M (see Proposition 4.3.4) and numerical results show that (for the set of parameters we employ) the expected profits in Model II decrease in M for $M > 1$ (see Proposition 4.4.5). This indicates that for both Models I and II the optimal amount of shares supplied in each LO will be the smallest amount required by the exchange. That is, when the exchange enforces an MRT, each MM will provide the minimum amount of liquidity.

Interestingly, Figure 4.16 shows that the longer the MRT, the higher the expected profit achieved by the MM when she chooses δ optimally. The intuition is as follows. As the MRT increases, and the depth of posted LO increases, there are two opposing forces at work. One, the longer the MRT, the more likely the LO is to become stale and to be picked off by liquidity takers, thus resulting in a loss. Two, as the depth of the posted LO increases, the chance that all volume is filled, before the end of the MRT, decreases. Thus, for the range of parameters we employ, we see that the loss-leading trades that result from the enforcement of longer MRTs are offset by wider depths of the LOs resting in the LOB. In addition, as the LO rests in the book for longer, more MOs arrive (on average) to fill the LO. Hence the expected profit of the MM increases as the MRTs increases. This is due to our model setup and in particular is caused by the MM considering the lifetime of a LO as the time horizon. We expect the result to be different if we choose different time horizons, for example a whole trading day where the MM can post many more than just one LO.

Figure 4.17 shows that the expected profits decrease with volatility of the fundamental price. This shows that even if the MMs choose the depth optimally, they will profit less when the stock is volatile, everything else being equal, because the probability of LOs being matched (by other LOs resting on the other side of the LOB) increases with volatility.

Finally, Figure 4.18 shows that the expected profits increase with the reference fill rate λ , see (4.7). This shows that the MMs will benefit from increasing the intensity of the arrival of MOs because, everything else being equal, the chances of the volume in the LOs being filled at stale quotes are lower if MO activity increases.

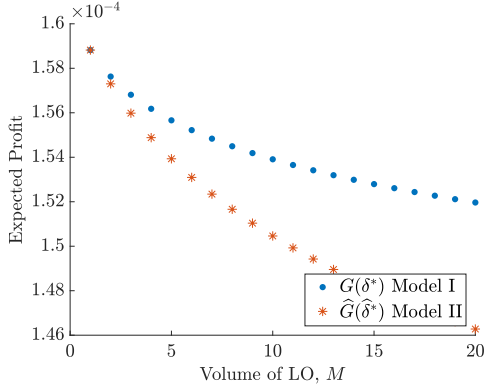


Figure 4.15: Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1/\text{second}$, $T = 0.5$ seconds.

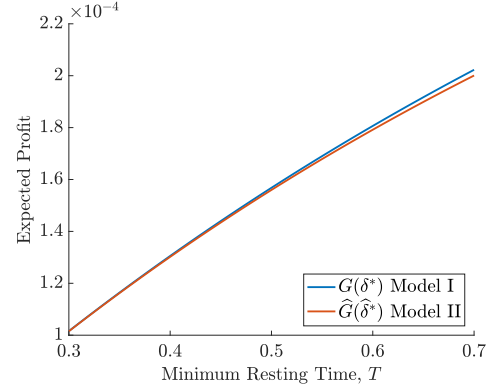


Figure 4.16: Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1/\text{second}$, $M = 3$

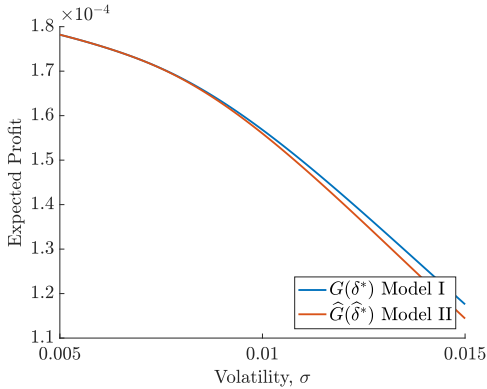


Figure 4.17: Expected profit. Parameters: $T = 0.5$ seconds, $\lambda = 0.1/\text{second}$, $M = 3$.

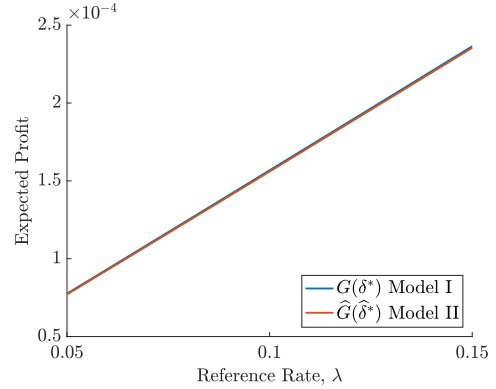


Figure 4.18: Expected profit. Parameters: $T = 0.5$ seconds, $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $M = 3$.

4.7 Conclusions

We developed a mathematical model of market making with the rule of minimum resting time (MRT). We derived an asymptotic integral expression for the expected profit of a market maker (MM) who chooses the depth of the limit orders (LOs) to maximise expected profits. We computed the optimal depth of the LOs posted by the MM and calculated the expected profits of the MM for various parameters of the model.

We showed that the depth of the LOs posted by the MMs increases as the MRT increases. The increase in the depth reduces the loss-leading trades from stale LOs that are picked off by other market participants. We also showed that the optimal depth of the LO increases when the volume of the LO or the volatility of the fundamental price increases, and when the arrival rate of market orders (MOs) decreases.

One of our most important findings is that when MMs choose the volume of the LOs they post, they supply the minimum amount of shares per LO allowed by the exchange. This is optimal for each MM but it is detrimental to the quality of the market. It is optimal for the MMs because expected profits are maximised when liquidity provided is lowest. This result suggests that implementing MRTs will decrease market liquidity provided by liquidity makers, while the objective of the regulators is to improve the quality of liquidity provision.

4.8 Proofs

4.8.1 Proof of Proposition 4.3.2

Proof. We make the standard change of variables,

$$y' = x' - \delta', \tilde{t}' = T' - t', h_0(\tilde{t}', y') = g'_0(t', x'),$$

and the PDE becomes

$$\partial_{\tilde{t}'} h_0 = \frac{1}{2} \partial_{y' y'} h_0. \quad (4.51)$$

The terminal and boundary conditions become

$$h_0(0, y') = \min(q, M) \left(-y' - \frac{\delta'}{2} \right), \quad h_0(\tilde{t}', 0) = 0, \quad y' < 0, \quad \tilde{t}' \in [0, T'].$$

PDE (4.51) is the heat equation with heat kernel

$$K(x' - y', t') = \frac{1}{\sqrt{2\pi t'}} e^{-\frac{(x' - y')^2}{2t'}}.$$

To apply the technique of the fundamental solution to the heat equation, we make an odd expansion of the initial condition:

$$u_0(y') = \begin{cases} \min(q, M) \left(-y' - \frac{\delta'}{2} \right), & y' < 0, \\ 0, & y' = 0, \\ \min(q, M) \left(-y' + \frac{\delta'}{2} \right), & y' > 0. \end{cases}$$

Then the solution of (4.51) is given by

$$\begin{aligned}
h_0(\tilde{t}', y') &= \int_{-\infty}^{\infty} K(y' - z, \tilde{t}') u_0(z) dz \\
&= \min(q, M) \left(\int_0^{\infty} K(y' - z, \tilde{t}') \left(-z + \frac{\delta'}{2}\right) dz \right. \\
&\quad \left. + \int_{-\infty}^0 K(y' - z, \tilde{t}') \left(-z - \frac{\delta'}{2}\right) dz \right) \\
&= \min(q, M) \left(- \int_0^{\infty} \frac{z}{\sqrt{2\pi\tilde{t}'}} e^{-\frac{(y'-z)^2}{2\tilde{t}'}} dz + \frac{\delta'}{2} \int_0^{\infty} \frac{1}{\sqrt{2\pi\tilde{t}'}} e^{-\frac{(y'-z)^2}{2\tilde{t}'}} dz \right. \\
&\quad \left. - \int_{-\infty}^0 \frac{z}{\sqrt{2\pi\tilde{t}'}} e^{-\frac{(y'-z)^2}{2\tilde{t}'}} dz - \frac{\delta'}{2} \int_{-\infty}^0 \frac{1}{\sqrt{2\pi\tilde{t}'}} e^{-\frac{(y'-z)^2}{2\tilde{t}'}} dz \right) \\
&= \min(q, M) \left(-y' + \frac{\delta'}{2} - \delta' \Phi\left(\frac{-y'}{\sqrt{\tilde{t}'}}\right) \right),
\end{aligned}$$

and by changing variables back to the original coordinates we obtain the desired result. \square

4.8.2 Proof of Proposition 4.3.3

Proof. We proceed as above and let $y' = x' - \delta'$, $\tilde{t}' = T' - t'$, $h_1(\tilde{t}', y') = g'_1(t', x')$. Then PDE (4.19) becomes

$$\partial_{\tilde{t}'} h_1 = \frac{1}{2} \partial_{y' y'} h_1 + f(\tilde{t}', y'), \quad (4.52)$$

where

$$f(\tilde{t}', y') = D(q, M) e^{\kappa' \left(\frac{\delta'}{2} + y'\right)} \left[-y' + \frac{\delta'}{2} - \delta' \Phi\left(\frac{-y'}{\sqrt{\tilde{t}'}}\right) \right],$$

and $D(q, M) = \min(q+1, M) - \min(q, M)$. The terminal and boundary conditions become

$$h_1(0, y') = 0, \quad h_1(\tilde{t}', 0) = 0, \quad y' < 0, \quad \tilde{t}' \in [0, T'].$$

We make an odd expansion on f and apply the corresponding heat kernel, thus

$$\begin{aligned}
& h_1(\tilde{t}', y') \\
&= \int_0^{\tilde{t}'} \int_{-\infty}^{\infty} K(y' - z, \tilde{t}' - s) f_0(s, z) dz ds \\
&= \int_0^{\tilde{t}'} \int_0^{\infty} K(y' - z, \tilde{t}' - s) (-f(s, -z)) dz ds + \int_0^{\tilde{t}'} \int_{-\infty}^0 K(y' - z, \tilde{t}' - s) f(s, z) dz ds \\
&= D(q, M) \int_0^{\tilde{t}'} \int_0^{\infty} K(y' - z, \tilde{t}' - s) \left[-e^{\kappa' \left(\frac{\delta'}{2} - z \right)} \left(z + \frac{\delta'}{2} - \delta' \Phi \left(\frac{z}{\sqrt{\tilde{t}'}} \right) \right) \right] dz ds \\
&\quad + D(q, M) \int_0^{\tilde{t}'} \int_{-\infty}^0 K(y' - z, \tilde{t}' - s) \left[e^{\kappa' \left(\frac{\delta'}{2} + z \right)} \left(-z + \frac{\delta'}{2} - \delta' \Phi \left(\frac{-z}{\sqrt{\tilde{t}'}} \right) \right) \right] dz ds \\
&= D(q, M) e^{\frac{\kappa' \delta'}{2}} \int_0^{\tilde{t}'} \int_0^{\infty} K(y' - z, \tilde{t}' - s) e^{-\kappa' z} \left[-z - \frac{\delta'}{2} + \delta' \Phi \left(\frac{z}{\sqrt{\tilde{t}'}} \right) \right] dz ds \\
&\quad + D(q, M) e^{\frac{\kappa' \delta'}{2}} \int_0^{\tilde{t}'} \int_{-\infty}^0 K(y' - z, \tilde{t}' - s) e^{\kappa' z} \left[-z + \frac{\delta'}{2} - \delta' \Phi \left(\frac{-z}{\sqrt{\tilde{t}'}} \right) \right] dz ds \\
&= D(q, M) \exp \left(\frac{\kappa' \delta'}{2} - \kappa' y' + \frac{1}{2} \kappa'^2 \tilde{t}' \right) \\
&\quad \times \int_0^{\tilde{t}'} \exp \left(-\frac{1}{2} \kappa'^2 s \right) \left\{ -\sqrt{\frac{\tilde{t}' - s}{2\pi}} \exp \left(-\frac{(y' - (\tilde{t}' - s) \kappa')^2}{2(\tilde{t}' - s)} \right) \right. \\
&\quad \quad - \left(y' - (\tilde{t}' - s) \kappa' + \frac{\delta'}{2} \right) \Phi \left(\frac{y' - (\tilde{t}' - s) \kappa'}{\sqrt{\tilde{t}' - s}} \right) \\
&\quad \quad \left. + \delta' \Phi \left(\frac{y' - (\tilde{t}' - s) \kappa'}{\sqrt{\tilde{t}' - s}}, \frac{y' - (\tilde{t}' - s) \kappa'}{\sqrt{\tilde{t}'}}; \sqrt{\frac{\tilde{t}' - s}{\tilde{t}'}} \right) \right\} ds, \\
&\quad + D(q, M) \exp \left(\frac{\kappa' \delta'}{2} + \kappa' y' + \frac{1}{2} \kappa'^2 \tilde{t}' \right) \\
&\quad \times \int_0^{\tilde{t}'} \exp \left(-\frac{1}{2} \kappa'^2 s \right) \left\{ \sqrt{\frac{(\tilde{t}' - s)}{2\pi}} \exp \left(-\frac{(y' + (\tilde{t}' - s) \kappa')^2}{2(\tilde{t}' - s)} \right) \right. \\
&\quad \quad - \left(y' + (\tilde{t}' - s) \kappa' - \frac{\delta'}{2} \right) \Phi \left(-\frac{y' + (\tilde{t}' - s) \kappa'}{\sqrt{\tilde{t}' - s}} \right) \\
&\quad \quad \left. - \delta' \Phi \left(-\frac{y' + (\tilde{t}' - s) \kappa'}{\sqrt{\tilde{t}' - s}}, -\frac{y' + (\tilde{t}' - s) \kappa'}{\sqrt{\tilde{t}'}}; \sqrt{\frac{\tilde{t}' - s}{\tilde{t}'}} \right) \right\} ds,
\end{aligned}$$

and by changing variables back to the original coordinates we obtain the desired result. \square

4.8.3 Proof of Theorem 4.3.1

Proof. First we define the infinitesimal generator \mathcal{L} , acting on a sufficiently differentiable function φ , as

$$\mathcal{L}\varphi = \partial_{t'}\varphi + \frac{1}{2}\partial_{x'x'}\varphi + \lambda' e^{-\kappa'\left(\frac{\delta'}{2}-x'\right)} [\varphi(t', x', q+1) - \varphi(t', x', q)] .$$

We define the stochastic processes $(t', X', N^{+'})$ such that the infinitesimal generator of $(t', X', N^{+'})$ is \mathcal{L} . We apply the infinitesimal generator \mathcal{L} on g'_e to obtain

$$\begin{aligned} \mathcal{L}g'_e &= \partial_{t'}g'_e + \frac{1}{2}\partial_{x'x'}g'_e + \lambda' e^{-\kappa'\left(\frac{\delta'}{2}-x'\right)} [g'_e(q+1) - g'_e(q)] \\ &= \partial_{t'}g' + \frac{1}{2}\partial_{x'x'}g' + \lambda' e^{-\kappa'\left(\frac{\delta'}{2}-x'\right)} [g'(q+1) - g'(q)] \\ &\quad - \partial_{t'}g'_0 - \frac{1}{2}\partial_{x'x'}g'_0 - \lambda' e^{-\kappa'\left(\frac{\delta'}{2}-x'\right)} [g'_0(q+1) - g'_0(q)] \\ &\quad - \lambda' \partial_{t'}g'_1 - \frac{\lambda'}{2}\partial_{x'x'}g'_1 - \lambda'^2 e^{-\kappa'\left(\frac{\delta'}{2}-x'\right)} [g'_1(q+1) - g'_1(q)] \\ &= -\lambda'^2 e^{-\kappa'\left(\frac{\delta'}{2}-x'\right)} [g'_1(q+1) - g'_1(q)] . \end{aligned}$$

Thus, we write

$$\begin{aligned} g'_e(t', x', q) &= \mathbb{E}_{t', X'_{t'}, N_{t'}^{+'}} \left[g'_e(T', X'_{T'}, N_{T'}^{+'}) - \int_{t'}^{T'} \mathcal{L}g'_e(s, X'_s, N_s^{+'}) ds \right] \\ &= \lambda'^2 \mathbb{E}_{t', X'_{t'}, N_{t'}^{+'}} \left[\int_{t'}^{T'} e^{-\kappa'\left(\frac{\delta'}{2}-X'_s\right)} \left(g'_1(s, X'_s, N_s^{+'} + 1) - g'_1(s, X'_s, N_s^{+'}) \right) ds \right] . \end{aligned}$$

The expectation above is bounded:

$$\begin{aligned} &\left| \mathbb{E}_{t', X'_{t'}, N_{t'}^{+'}} \left[\int_{t'}^{T'} e^{-\kappa'\left(\frac{\delta'}{2}-X'_s\right)} \left(g'_1(s, X'_s, N_s^{+'} + 1) - g'_1(s, X'_s, N_s^{+'}) \right) ds \right] \right| \\ &\leq \mathbb{E}_{t', X'_{t'}, N_{t'}^{+'}} \left[\int_{t'}^{T'} e^{-\kappa'\left(\frac{\delta'}{2}-X'_s\right)} \left| g'_1(s, X'_s, N_s^{+'} + 1) - g'_1(s, X'_s, N_s^{+'}) \right| ds \right] . \end{aligned}$$

To bound the integrand, we have

$$\begin{aligned}
& |g'_1(t', x', q+1) - g'_1(t', x', q)| \\
& \leq |D(q+1, M) - D(q, M)| \exp\left(\frac{3\kappa'\delta'}{2} - \kappa'x' + \frac{1}{2}\kappa'^2(T' - t')\right) \\
& \quad \times \int_0^{T'-t'} \exp\left(-\frac{1}{2}\kappa'^2 s\right) \left\{ \sqrt{\frac{(T' - t' - s)}{2\pi}} \exp\left(-\frac{(x' - \delta' - (T' - t' - s)\kappa')^2}{2(T' - t' - s)}\right) \right. \\
& \quad \left. + \left(|x'| + (T' - t' - s)\kappa' + \frac{\delta'}{2}\right) \Phi\left(\frac{x' - \delta' - (T' - t' - s)\kappa'}{\sqrt{T' - t' - s}}\right) \right. \\
& \quad \left. + \delta' \Phi\left(\frac{x' - \delta' - (T' - t' - s)\kappa'}{\sqrt{T' - t' - s}}, \frac{x' - \delta' - (T' - t' - s)\kappa'}{\sqrt{T' - t'}}; \sqrt{\frac{T' - t' - s}{T' - t'}}\right) \right\} ds \\
& \quad + |D(q+1, M) - D(q, M)| \exp\left(-\frac{\kappa'\delta'}{2} + \kappa'x' + \frac{1}{2}\kappa'^2(T' - t')\right) \\
& \quad \times \int_0^{T'-t'} \exp\left(-\frac{1}{2}\kappa'^2 s\right) \left\{ \sqrt{\frac{(T' - t' - s)}{2\pi}} \exp\left(-\frac{(x' - \delta' + (T' - t' - s)\kappa')^2}{2(T' - t' - s)}\right) \right. \\
& \quad \left. + \left(|x'| + (T' - t' - s)\kappa' + \frac{3\delta'}{2}\right) \Phi\left(-\frac{x' - \delta' + (T' - t' - s)\kappa'}{\sqrt{T' - t' - s}}\right) \right. \\
& \quad \left. + \delta' \Phi\left(-\frac{x' - \delta' + (T' - t' - s)\kappa'}{\sqrt{T' - t' - s}}, -\frac{x' - \delta' + (T' - t' - s)\kappa'}{\sqrt{T' - t'}}; \sqrt{\frac{T' - t' - s}{T' - t'}}\right) \right\} ds \\
& \leq 2 \exp\left(\frac{3\kappa'\delta'}{2} - \kappa'x' + \frac{1}{2}\kappa'^2(T' - t')\right) \\
& \quad \times \int_0^{T'-t'} \sqrt{\frac{(T' - t' - s)}{2\pi}} + \left(|x'| + (T' - t' - s)\kappa' + \frac{\delta'}{2}\right) + \delta' ds \\
& \quad + 2 \exp\left(-\frac{\kappa'\delta'}{2} + \kappa'x' + \frac{1}{2}\kappa'^2(T' - t')\right) \\
& \quad \times \int_0^{T'-t'} \sqrt{\frac{(T' - t' - s)}{2\pi}} + \left(|x'| + (T' - t' - s)\kappa' + \frac{3\delta'}{2}\right) + \delta' ds \\
& = 2 \exp\left(\frac{3\kappa'\delta'}{2} - \kappa'x' + \frac{1}{2}\kappa'^2(T' - t')\right) \left(\left(|x'| + \frac{3\delta'}{2}\right)(T' - t') + \frac{1}{3}\sqrt{\frac{2}{\pi}}(T' - t')^{\frac{3}{2}} + \frac{\kappa'}{2}(T' - t')^2 \right) \\
& \quad + 2 \exp\left(-\frac{\kappa'\delta'}{2} + \kappa'x' + \frac{1}{2}\kappa'^2(T' - t')\right) \left(\left(|x'| + \frac{5\delta'}{2}\right)(T' - t') + \frac{1}{3}\sqrt{\frac{2}{\pi}}(T' - t')^{\frac{3}{2}} + \frac{\kappa'}{2}(T' - t')^2 \right),
\end{aligned}$$

and we recall that

$$D(q, M) = \min(q+1, M) - \min(q, M).$$

The expectations $\mathbb{E}[e^{|X'_s|}]$ and $\mathbb{E}[|X'_s| e^{|X'_s|}]$ are bounded because X'_s is normal. Hence, the expectation $\mathbb{E}[|g'_1(s, X'_s, N_s^{++} + 1) - g'_1(s, X'_s, N_s^{++})|]$ is bounded, and we obtain the desired result. \square

4.8.4 Proof of Proposition 4.4.2

Proof.

$$\begin{aligned}
\mathbb{E} \left[\widehat{\Pi}_1 \right] &= \mathbb{E} \left[M \left(-\frac{\delta_1^*}{2} \right) \mathbb{1}_{\{\widehat{\tau}=\widehat{\tau}_a\}} \right] \\
&= -\frac{M \delta_1^*}{2} \mathbb{P}(\tau = \widehat{\tau}_a) \\
&= -\frac{M \delta_1^*}{2} \mathbb{P} \left(\max_{0 \leq t \leq T} \sigma W_t > \widehat{\delta}(M)/2 + \delta_1^*/2 \right) \\
&= -M \delta_1^* \mathbb{P} \left(\sigma W_T > \widehat{\delta}(M)/2 + \delta_1^*/2 \right) \\
&= -M \delta_1^* \left(1 - \Phi \left(\frac{\widehat{\delta}(M) + \delta_1^*}{2 \sigma \sqrt{T}} \right) \right).
\end{aligned}$$

□

4.8.5 Proof of Proposition 4.4.3

Proof. We make the standard change of variables,

$$y' = x' - \widetilde{\delta}', \tilde{t}' = T' - t', \hat{h}_0(\tilde{t}', y') = \hat{g}'_0(t', x')'$$

so that the PDE (4.41) becomes

$$\partial_{\tilde{t}'} \hat{h}_0 = \frac{1}{2} \partial_{y' y'} \hat{h}_0, \quad (4.53)$$

and the terminal and boundary conditions become

$$\hat{h}_0(0, y') = \min(q, M) \left(-y' - \frac{\delta_1^{*'}}{2} \right), \quad \hat{h}_0(\tilde{t}', 0) = 0, \quad y' < 0, \quad \tilde{t}' \in [0, T'].$$

PDE (4.53) is the heat equation with heat kernel

$$K(x' - y', t') = \frac{1}{\sqrt{2\pi t'}} e^{-\frac{(x' - y')^2}{2t'}}$$

and initial condition

$$\widehat{u}_0(y') = \begin{cases} \min(q, M) \left(-y' - \frac{\delta_1^{*'}}{2} \right), & y' < 0, \\ 0, & y' = 0, \\ \min(q, M) \left(-y' + \frac{\delta_1^{*'}}{2} \right), & y' > 0. \end{cases}$$

Then,

$$\begin{aligned}
\hat{h}_0(\tilde{t}', y') &= \int_{-\infty}^{\infty} K(y' - z, \tilde{t}') \hat{u}_0(z) dz \\
&= \min(q, M) \left(\int_0^{\infty} K(y' - z, \tilde{t}') \left(-z + \frac{\delta_1^{*'}}{2}\right) dz \right. \\
&\quad \left. + \int_{-\infty}^0 K(y' - z, \tilde{t}') \left(-z - \frac{\delta_1^{*'}}{2}\right) dz \right) \\
&= \min(q, M) \left(- \int_0^{\infty} \frac{z}{\sqrt{2\pi\tilde{t}'}} e^{-\frac{(y'-z)^2}{2\tilde{t}'}} dz + \frac{\delta_1^{*'}}{2} \int_0^{\infty} \frac{1}{\sqrt{2\pi\tilde{t}'}} e^{-\frac{(y'-z)^2}{2\tilde{t}'}} dz \right. \\
&\quad \left. - \int_{-\infty}^0 \frac{z}{\sqrt{2\pi\tilde{t}'}} e^{-\frac{(y'-z)^2}{2\tilde{t}'}} dz - \frac{\delta_1^{*'}}{2} \int_{-\infty}^0 \frac{1}{\sqrt{2\pi\tilde{t}'}} e^{-\frac{(y'-z)^2}{2\tilde{t}'}} dz \right) \\
&= \min(q, M) \left(-y' + \frac{\delta_1^{*'}}{2} - \delta_1^{*'} \Phi\left(\frac{-y'}{\sqrt{\tilde{t}'}}\right) \right),
\end{aligned}$$

and by changing variables back to the original coordinates we obtain the desired result. \square

4.8.6 Proof of Proposition 4.4.4

Proof. We make the standard change of variables

$$y' = x' - \tilde{\delta}', \quad \tilde{t}' = T' - t', \quad \hat{h}_1(\tilde{t}', y') = \hat{g}_1'(t', x').$$

Then

$$\partial_{\tilde{t}'} \hat{h}_1 = \frac{1}{2} \partial_{y'y'} \hat{h}_1 + \hat{f}(\tilde{t}', y'), \quad (4.54)$$

where

$$\hat{f}(\tilde{t}', y') = D(q, M) e^{\kappa' \left(\frac{\delta_1^{*'}}{2} + y'\right)} \left[-y' + \frac{\delta_1^{*'}}{2} - \delta_1^{*'} \Phi\left(\frac{-y'}{\sqrt{\tilde{t}'}}\right) \right],$$

where $D(q, M) = \min(q + 1, M) - \min(q, M)$. The terminal and boundary conditions become

$$\hat{h}_1(0, y') = 0, \quad \hat{h}_1(\tilde{t}', 0) = 0, \quad y' < 0, \quad \tilde{t}' \in [0, T'].$$

We make an odd expansion on \hat{f} to \hat{f}_0 and apply the heat kernel to obtain:

$$\begin{aligned}
& \hat{h}_1(\tilde{t}, y) \\
&= \int_0^{\tilde{t}'} \int_{-\infty}^{\infty} K(y' - z, \tilde{t}' - s) \hat{f}_0(s, z) dz ds \\
&= \int_0^{\tilde{t}'} \int_0^{\infty} K(y' - z, \tilde{t}' - s) \left(-\hat{f}(s, -z) \right) dz ds + \int_0^{\tilde{t}'} \int_{-\infty}^0 K(y' - z, \tilde{t}' - s) \hat{f}(s, z) dz ds \\
&= D(q, M) \int_0^{\tilde{t}'} \int_0^{\infty} K(y' - z, \tilde{t}' - s) \left[-e^{\kappa' \left(\frac{\delta_1^{*'}}{2} - z \right)} \left(z + \frac{\delta_1^{*'}}{2} - \delta_1^{*'} \Phi \left(\frac{z}{\sqrt{\tilde{t}'}} \right) \right) \right] dz ds \\
&\quad + D(q, M) \int_0^{\tilde{t}'} \int_{-\infty}^0 K(y' - z, \tilde{t}' - s) \left[e^{\kappa' \left(\frac{\delta_1^{*'}}{2} + z \right)} \left(-z + \frac{\delta_1^{*'}}{2} - \delta_1^{*'} \Phi \left(\frac{-z}{\sqrt{\tilde{t}'}} \right) \right) \right] dz ds \\
&= D(q, M) e^{\frac{\kappa' \delta_1^{*'}}{2}} \int_0^{\tilde{t}'} \int_0^{\infty} K(y' - z, \tilde{t}' - s) e^{-\kappa' z} \left[-z - \frac{\delta_1^{*'}}{2} + \delta_1^{*'} \Phi \left(\frac{z}{\sqrt{\tilde{t}'}} \right) \right] dz ds \\
&\quad + D(q, M) e^{\frac{\kappa' \delta_1^{*'}}{2}} \int_0^{\tilde{t}'} \int_{-\infty}^0 K(y' - z, \tilde{t}' - s) e^{\kappa' z} \left[-z + \frac{\delta_1^{*'}}{2} - \delta_1^{*'} \Phi \left(\frac{-z}{\sqrt{\tilde{t}'}} \right) \right] dz ds \\
&= D(q, M) \exp \left(\frac{\kappa' \delta_1^{*'}}{2} - \kappa' y' + \frac{1}{2} \kappa'^2 \tilde{t}' \right) \\
&\quad \times \int_0^{\tilde{t}'} \exp \left(-\frac{1}{2} \kappa'^2 s \right) \left\{ -\sqrt{\frac{\tilde{t}' - s}{2\pi}} \exp \left(-\frac{(y' - (\tilde{t}' - s) \kappa')^2}{2(\tilde{t}' - s)} \right) \right. \\
&\quad \quad - \left(y' - (\tilde{t}' - s) \kappa' + \frac{\delta_1^{*'}}{2} \right) \Phi \left(\frac{y' - (\tilde{t}' - s) \kappa'}{\sqrt{\tilde{t}' - s}} \right) \\
&\quad \quad \left. + \delta_1^{*'} \Phi \left(\frac{y' - (\tilde{t}' - s) \kappa'}{\sqrt{\tilde{t}' - s}}, \frac{y' - (\tilde{t}' - s) \kappa'}{\sqrt{\tilde{t}'}}; \sqrt{\frac{\tilde{t}' - s}{\tilde{t}'}} \right) \right\} ds, \\
&+ D(q, M) \exp \left(\frac{\kappa' \delta_1^{*'}}{2} + \kappa' y' + \frac{1}{2} \kappa'^2 \tilde{t}' \right) \\
&\quad \times \int_0^{\tilde{t}'} \exp \left(-\frac{1}{2} \kappa'^2 s \right) \left\{ \sqrt{\frac{\tilde{t}' - s}{2\pi}} \exp \left(-\frac{(y' + (\tilde{t}' - s) \kappa')^2}{2(\tilde{t}' - s)} \right) \right. \\
&\quad \quad - \left(y' + (\tilde{t}' - s) \kappa' - \frac{\delta_1^{*'}}{2} \right) \Phi \left(-\frac{y' + (\tilde{t}' - s) \kappa'}{\sqrt{\tilde{t}' - s}} \right) \\
&\quad \quad \left. - \delta_1^{*'} \Phi \left(-\frac{y' + (\tilde{t}' - s) \kappa'}{\sqrt{\tilde{t}' - s}}, -\frac{y' + (\tilde{t}' - s) \kappa'}{\sqrt{\tilde{t}'}}; \sqrt{\frac{\tilde{t}' - s}{\tilde{t}'}} \right) \right\} ds,
\end{aligned}$$

and by changing variables back to the original coordinates we obtain the desired result. \square

Chapter 5

Conclusions

5.1 Summary of contributions

In this thesis we looked at mathematical problems that arise in algorithmic trading in order driven markets. In Chapter 2 we showed how a market maker employs information about the momentum (alpha signal) in the price of the asset and solve a stochastic and impulse control problem to find the optimal liquidity provision strategy. In Chapter 3 we looked at the problem of an investor who wants to spoof the LOB while liquidation, to earn extra revenue and make the liquidation faster. We show that the investor does not only benefit from using more LOs and less MOs, but also from manipulating the midprice to earn extra revenue through liquidation. As the penalty for spoofing increases, the strategy relies less on spoof LOs. There is a critical point where the penalty outweighs the benefits from spoofing so it is optimal not to spoof the LOB. In Chapter 4 we looked at market making with the rule of minimum resting times, which has been proposed by regulators to prevent liquidity providers from spoofing or from providing fleeting liquidity. One of our most important findings is that when MMs choose the volume of the LOs they post, they supply the minimum amount of shares per LO allowed by the exchange. This is optimal for each MM but it is detrimental to the quality of the market. It is optimal for the MMs because expected profits are maximised when liquidity provided is lowest. This result suggests that implementing MRTs will decrease market liquidity provided by liquidity makers, while the objective of the regulators is to improve the quality of liquidity provision.

5.2 Directions for future research

In this section we provide some possible directions for future research.

5.2.1 Momentum in prices

- An extension to the model presented in Chapter 2 is to model directly the LOB instead of modelling intensities of the jumps in the midprice. Another extension is to let the investor control the volume of the LOs and MOs.

5.2.2 Spoofing and price manipulation

The framework we developed in Chapter 3 can be employed to develop other trading strategies that rely on spoof orders to improve their financial performance of the strategy. For example, a market making strategy that employs sell and buy spoof LOs to open and close positions, respectively. There are other ways to spoof the LOB that could be considered within the framework developed in this work. We provide three examples:

- Phantom liquidity inside the spread. The investor wishes to sell shares. The strategy consists of sending spoof buy LOs inside the spread (i.e., improve the bid price) to entice other liquidity providers to join the queue at the improved bid price. As soon as other traders send LOs at the new best bid price, the spoofer cancels her spoof buy LOs and sends sell MOs. A similar strategy is used when the investor wishes to purchase shares.
- Cross-spoofing. This is identical to the strategy developed in this work, only that two (or more) investors agree to spoof the market to avoid being detected by financial authorities. In the strategy, one investor(s) spoofs the LOB and the other investor(s) liquidates the position in the shares.
- Layering. The strategy consists of posting several large LOs at different prices on one side of the book. The goal is to move the price because other market participants interpret the one sided pressure in the LOB as a signal of a price move and trade in anticipation of expected change in price.

5.2.3 Minimum resting times

- In our model presented in Chapter 4, the fill rate of LOs captures stylised facts we observe in the market (NASDAQ) and strikes the right balance between empirical fill rates and mathematical tractability. Future work is needed to have a model of fill rates that better reflects the shape and dynamics of the LOB, see for example Cartea et al. (2014) and Guéant (2017) who discuss desirable conditions of fill rates as a function of the depth of the LOs and shape of the LOB.
- Examine how certainty of execution prices is affected by MRTs. The work of Cartea and Sánchez-Betancourt (2018) discusses how latency affects the efficacy of liquidity taking strategies because liquidity in the LOB may change between the time liquidity takers decide to trade and the time the orders reach the exchange. It is not clear how MRTs will affect the efficacy of MOs from slow traders because fast traders will snipe stale quotes, some of which were the target of the slow traders.
- Examine how price impact of liquidity taking orders is affected by MRTs. In Chapter 4 we find that liquidity will deteriorate when the exchange enforces MRTs. This will have an effect on price impact costs and how optimal trading strategies are designed.

Appendix A

Supplementary Background Material

A.1 Hamilton-Jacobi-Bellman Quasi-Variational Inequalities (HJBQVIs)

In this section we state the Hamilton-Jacobi-Bellman Quasi-Variational Inequalities (HJBQVIs) for a combined stochastic control and impulse control problem. For details please refer to Øksendal and Sulem (2007).

Suppose the state $Y(t) \in \mathbb{R}^k$ of the system we consider is a jump diffusion, and we are free at any state $y \in \mathbb{R}^k$ to choose a Markov control $u(y) \in U$, where U is a given closed convex set in \mathbb{R}^p . Let \mathcal{U} be a set of such Markov controls. For any (admissible) Markov control u , denote the generator of Y_t by $\mathcal{L}^{(u)}$.

Suppose that at any time t and any state y we are free to intervene and give the system an impulse $\zeta \in \mathcal{Z} \subset \mathbb{R}^p$, where \mathcal{Z} is a given of admissible impulse values. Suppose the result of the impulse ζ when the state is y is that the state jumps immediately from $Y(t^-) = y$ to $Y(t) = \Gamma(y, \zeta) \in \mathbb{R}^k$, where $\Gamma : \mathbb{R}^k \times \mathcal{Z} \rightarrow \mathbb{R}^k$ is a given function. An impulse control for this system is a double sequence

$$v = (\tau_1, \tau_2, \dots, \tau_j, \dots; \zeta_1, \zeta_2, \dots, \zeta_j, \dots)_{j \leq M}, \quad M \leq \infty,$$

where $0 \leq \tau_1 \leq \tau_2 \leq \dots$ are \mathcal{F}_t -stopping times (the intervention times) and ζ_1, ζ_2, \dots are the corresponding \mathcal{F}_{τ_j} -measurable impulses at these times.

For a given v , the impulse control affects the controlled process Y_t at the impulse times via

$$Y_{\tau_j}^{(u,v)} = \Gamma\left(Y_{\tau_j^-}^{(u,v)} + \Delta_N Y_{\tau_j}^{(u,v)}, \zeta_j\right),$$

where $\Delta_N Y_{\tau_j}^{(u,v)}$ is the jump of $Y^{(u,v)}$ stemming from the jump component of the jump diffusion only. Suppose the profit of making an intervention with impulse $\zeta \in \mathcal{Z}$ when the state is y is $K(y, \zeta)$, where $K : \mathcal{S} \times \mathcal{Z} \rightarrow \mathbb{R}$ is a continuous function.

Let $\mathcal{S} \subset \mathbb{R}^k$ be a fixed open set (the solvency region), and define

$$\tau_{\mathcal{S}} = \inf \{t \geq 0 : Y(t) \notin \mathcal{S}\}. \quad (\text{A.1})$$

The performance criterion is given by

$$\begin{aligned} J^{(u,v)}(y) = \mathbb{E}_{t,y} \left[\int_0^{\tau_{\mathcal{S}}} f(Y^{(u,v)}(t), u(t)) dt + g(Y^{(u,v)}(\tau_{\mathcal{S}})) \mathbb{1}_{\{\tau_{\mathcal{S}} < \infty\}} \right. \\ \left. + \sum_{\tau_j \leq \tau_{\mathcal{S}}} K(Y_{\tau_j-}^{(u,v)} + \Delta_N Y_{\tau_j}^{(u,v)}, \zeta_j) \right], \end{aligned} \quad (\text{A.2})$$

where $f : \mathcal{S} \rightarrow \mathbb{R}$ and $g : \mathcal{S} \rightarrow \mathbb{R}$ are continuous functions. f , g and K should also satisfy some integrability conditions.

We now define the value function as

$$\Phi(y) = \sup_{(u,v)} J^{(u,v)}(y). \quad (\text{A.3})$$

Through a dynamic programming argument, the HJBQVI to the value function A.3 is then

$$\max \left[\sup_u \left\{ \mathcal{L}^{(u)} \Phi + f(y, u) \right\}; \sup_v \left\{ \Phi(\Gamma(y, \zeta)) + K(y, \zeta) \right\} - \Phi \right] = 0, \quad y \in \mathcal{S}, \quad (\text{A.4})$$

with boundary condition

$$\Phi = g, \quad y \in \partial \mathcal{S}. \quad (\text{A.5})$$

A.2 Viscosity solutions

In most cases, the value function Φ defined by (A.3) need not be C^1 , and (A.4) is not well defined if we interpret the equation in the classical sense. However, if we interpret (A.4) in the weak sense of viscosity then Φ does indeed solve the equation. To this end, we define the notion of viscosity solutions.

Let $h \in C(\bar{\mathcal{S}})$.

- We say h is a viscosity subsolution of (A.4) and (A.5), if (A.5) holds and for every $\varphi \in C^2(\mathbb{R}^k)$ and every $y_0 \in \mathcal{S}$ such that $\varphi \geq h$ on \mathcal{S} and $\varphi(y_0) = h(y_0)$ we have

$$\max \left[\sup_u \left\{ \mathcal{L}^{(u)} \varphi + f(y, u) \right\}; \sup_v \left\{ h(\Gamma(y, \zeta)) + K(y, \zeta) \right\} - h \right] \geq 0. \quad (\text{A.6})$$

- We say h is a viscosity supersolution of (A.4) and (A.5), if (A.5) holds and for every $\varphi \in C^2(\mathbb{R}^k)$ and every $y_0 \in \mathcal{S}$ such that $\varphi \leq h$ on \mathcal{S} and $\varphi(y_0) = h(y_0)$ we have

$$\max \left[\sup_u \left\{ \mathcal{L}^{(u)} \varphi + f(y, u) \right\}; \sup_v \left\{ h(\Gamma(y, \zeta)) + K(y, \zeta) \right\} - h \right] \leq 0. \quad (\text{A.7})$$

- We say h is a viscosity solution of (A.4) and (A.5), if h is both a viscosity subsolution and a viscosity supersolution of (A.4) and (A.5).

Bibliography

Agency for the Cooperation of Energy Regulators (2019a). Layering and spoofin in continuous wholesale energy markets.

Agency for the Cooperation of Energy Regulators (2019b). Remit quarterly.

Ait-Sahalia, Y. and M. Sağlam (2017). High frequency market making: Implications for liquidity. SSRN.

Alfonsi, A. and J. I. Acevedo (2014). Optimal execution and price manipulations in time-varying limit order books. Applied Mathematical Finance 21(3), 201–237.

Alfonsi, A., A. Fruth, and A. Schied (2010). Optimal execution strategies in limit order books with general shape functions. Quantitative Finance 10(2), 143–157.

Allen, F. and D. Gale (1992). Stock-price manipulation. The Review of Financial Studies 5(3), 503–529.

Almgren, R. (2012). Optimal trading with stochastic liquidity and volatility. SIAM Journal on Financial Mathematics 3(1), 163–181.

Almgren, R. and N. Chriss (2001). Optimal execution of portfolio transactions. Journal of Risk 3, 5–40.

Almgren, R. F. (2003). Optimal execution with nonlinear impact functions and trading-enhanced risk. Applied Mathematical Finance 10(1), 1–18.

Avellaneda, M. and S. Stoikov (2008). High-frequency trading in a limit order book. Quantitative Finance 8(3), 217–224.

Barles, G. and P. E. Souganidis (1991). Convergence of approximation schemes for fully nonlinear second order equations. Asymptotic Analysis 4(3), 271–283.

- Biais, B., T. Foucault, and S. Moinas (2015). Equilibrium fast trading. Journal of Financial Economics 116(2), 292–313.
- Boehmer, E., K. Fong, and J. Wu (2015). International evidence on algorithmic trading. SSRN.
- Brewer, P., J. Cvitanic, and C. R. Plott (2013). Market microstructure design and flash crashes: a simulation approach. Journal of Applied Economics 16(2), 223–250.
- Cao, Y., Y. Li, S. Coleman, A. Belatreche, and T. M. McGinnity (2014). Detecting price manipulation in the financial market. In Computational Intelligence for Financial Engineering & Economics (CIFEr), 2104 IEEE Conference on, pp. 77–84. IEEE.
- Cao, Y., Y. Li, S. A. Coleman, A. Belatreche, T. M. McGinnity, et al. (2015). Adaptive hidden markov model with anomaly states for price manipulation detection. IEEE Trans. Neural Netw. Learning Syst. 26(2), 318–330.
- Cartea, Á., R. Donnelly, and S. Jaimungal (2017). Algorithmic trading with model uncertainty. SIAM Journal on Financial Mathematics 8(1), 635–671.
- Cartea, Á., R. Donnelly, and S. Jaimungal (2018). Enhancing trading strategies with order book signals. Applied Mathematical Finance, 1–35.
- Cartea, Á. and S. Jaimungal (2015a). Optimal execution with limit and market orders. Quantitative Finance 15(8), 1279–1291.
- Cartea, Á. and S. Jaimungal (2015b). Optimal execution with limit and market orders. Quantitative Finance 15(8), 1279–1291.
- Cartea, Á. and S. Jaimungal (2015c). Risk metrics and fine tuning of high-frequency trading strategies. Mathematical Finance 25(3), 576–611.
- Cartea, Á. and S. Jaimungal (2016a). A closed-form execution strategy to target volume weighted average price. SIAM Journal on Financial Mathematics 7(1), 760–785.
- Cartea, Á. and S. Jaimungal (2016b). Incorporating order-flow into optimal execution. Mathematics and Financial Economics 10(3), 339–364.
- Cartea, Á., S. Jaimungal, and J. Penalva (2015). Algorithmic and high-frequency trading. Cambridge University Press.

- Cartea, Á., S. Jaimungal, and J. Ricci (2014). Buy low, sell high: A high frequency trading perspective. SIAM Journal on Financial Mathematics 5(1), 415–444.
- Cartea, Á., S. Jaimungal, and J. Ricci (2018a). Algorithmic trading, stochastic control, and mutually exciting processes. SIAM Review 60(3), 673–703.
- Cartea, Á., S. Jaimungal, and J. Ricci (2018b). Trading strategies within the edges of no-arbitrage. International Journal of Theoretical and Applied Finance 21(03), 1850025.
- Cartea, Á., S. Jaimungal, and J. Walton (2015). Foreign exchange markets with last look. Mathematics and Financial Economics, 1–30.
- Cartea, Á., S. Jaimungal, and Y. Wang (2020). Spoofing and price manipulation in order-driven markets. Applied Mathematical Finance, 1–32.
- Cartea, Á., R. Payne, J. Penalva, and M. Tapia (2019). Ultra-fast activity and intraday market quality. Journal of Banking & Finance 99, 157–181.
- Cartea, Á. and J. Penalva (2012). Where is the value in high frequency trading? Quarterly Journal of Finance 2(3), 1–46.
- Cartea, Á. and L. Sánchez-Betancourt (2018). The shadow price of latency: Improving intraday fill ratios in foreign exchange markets. SSRN.
- Cartea, Á. and Y. Wang (2019). Market making with minimum resting times. Quantitative Finance 19(6), 903–920.
- Cartea, Á. and Y. Wang (2020). Market making with alpha signals. International Journal of Theoretical and Applied Finance, 2050016.
- Chaboud, A. P., B. Chiquoine, E. Hjalmarsson, and C. Vega (2014). Rise of the machines: Algorithmic trading in the foreign exchange market. The Journal of Finance 69(5), 2045–2084.
- Cheridito, P. and T. Sepin (2014). Optimal trade execution under stochastic volatility and liquidity. Applied Mathematical Finance 21(4), 342–362.
- Commission, E. et al. (2010). Review of the markets in financial instruments directive (MiFID). Public Consultation Document, EU Commission, Brussels 8.

- Cont, R. and P. Tankov (2004). Financial modelling with jump processes. Chapman and Hall.
- Dodd-Frank (2010). Public Law 111-203, reform, Dodd-Frank Wall Street and Act, consumer protection. US Statutes at Large 124, 1376.
- Donnelly, R. and L. Gan (2018). Optimal decisions in a time priority queue. Applied Mathematical Finance 25(2), 107–147.
- Farmer, J. D. and S. Skouras (2012). Minimum resting times and transaction-to-order ratios: review of amendment 2.3. f and question 20. Foresight, Government Office For Science.
- Foucault, T. (2012). Pricing liquidity in electronic markets. Foresight Driver Review (DR18) 15(8), 1–26.
- Guéant, O. (2015). Optimal execution and block trade pricing: A general framework. Applied Mathematical Finance 22(4), 336–365.
- Guéant, O. (2016). The financial mathematics of market liquidity: From optimal execution to market making, Volume 33. CRC Press.
- Guéant, O. (2017). Optimal market making. Applied Mathematical Finance 24(2), 112–154.
- Guéant, O., C.-A. Lehalle, and J. Fernandez-Tapia (2013). Dealing with the inventory risk: A solution to the market making problem. Mathematics and Financial Economics 7(4), 477–507.
- Hayes, R., M. Paddrik, A. Todd, S. Yang, P. Beling, and W. Scherer (2012). Agent based model of the e-mini future: Application for policy making. In Proceedings of the Winter Simulation Conference, pp. 111. Winter Simulation Conference.
- Hendershott, T., C. M. Jones, and A. J. Menkveld (2011). Does algorithmic trading improve liquidity? The Journal of Finance 66(1), 1–33.
- Ho, T., H. R. Stoll, et al. (1981). Optimal dealer pricing under transactions and return uncertainty. Journal of Financial Economics 9(1), 47–73.

- Hoffmann, P. (2014). A dynamic limit order market with fast and slow traders. Journal of Financial Economics 113(1), 156 – 169.
- Ishii, H. and S. Koike (1991). Viscosity solutions for monotone systems of second-order elliptic PDEs. Communications in Partial Differential Equations 16(6-7), 1095–1128.
- Klöck, F., A. Schied, and Y. Sun (2017). Price manipulation in a market impact model with dark pool. Applied Mathematical Finance 24(5), 417–450.
- Leal, S. J. and M. Napoletano (2017). Market stability vs. market resilience: Regulatory policies experiments in an agent-based model with low-and high-frequency trading. Journal of Economic Behavior & Organization.
- Lee, E. J., K. S. Eom, and K. S. Park (2013). Microstructure-based manipulation: Strategic behavior and performance of spoofing traders. Journal of Financial Markets 16(2), 227–252.
- Lorenz, J. and R. Almgren (2011). Meanvariance optimal adaptive execution. Applied Mathematical Finance 18(5), 395–422.
- Martinez, V. H. and I. Rosu (2013). High frequency traders, news and volatility. In AFA 2013 San Diego Meetings Paper.
- Obizhaeva, A. A. and J. Wang (2013). Optimal trading strategy and supply/demand dynamics. Journal of Financial Markets 16(1), 1–32.
- Øksendal, B. and A. Sulem (2007). Applied Stochastic Control of Jump Diffusions. Springer Science & Business Media.
- Pham, H. (2009). Continuous-time Stochastic Control and Optimization with Financial Applications, Volume 61. Springer Science & Business Media.
- Schied, A., T. Schöneborn, and M. Tehranchi (2010). Optimal basket liquidation for CARA investors is deterministic. Applied Mathematical Finance 17(6), 471–489.
- Seydel, R. C. (2010). Impulse control for jump-diffusions: Viscosity solutions of quasi-variational inequalities and applications in bank risk management. Ph. D. thesis, Verlag nicht ermittelbar.

Van Ness, B. F., R. A. Van Ness, and E. D. Watson (2015). Canceling liquidity. Journal of Financial Research 38(1), 3–33.

Wang, Y. (2015). Strategic spoofing order trading by different types of investors in the futures markets. Wall Street Journal.