Market Making with Minimum Resting Times
Forthcoming: Quantitative Finance

Álvaro Cartea[a], Yixuan Wang[b]

[a] *Mathematical Institute, University of Oxford, Oxford, UK*
Oxford-Man Institute of Quantitative Finance, Oxford, UK

[b] *Mathematical Institute, University of Oxford, Oxford, UK*

Abstract

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.

Keywords: Market making; Minimum resting times; High-frequency trading; Market quality; Regulation

1. Introduction

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.

Email addresses: `Alvaro.Cartea@maths.ox.ac.uk` (Álvaro Cartea), `yixuan.wang@maths.ox.ac.uk` (Yixuan Wang)

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.

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.

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. (2016) 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 paper 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.

The remainder of this paper proceeds as follows. In Section 2 we provide a review of the literature. In Section 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 extends the model so MMs can post LOs of any positive integer volume. In Section 5 we investigate the optimal depth that the MM chooses. In Section 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 7 and collect proofs in the Appendix.

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 paper 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.

At a more general level, there a number of papers that discuss how algorithmic and high-frequency trading affect the quality of markets. Theoretical papers have looked at the effect of speed advantages on market quality and their welfare implications. Cartea and Penalva (2012) provide a theoretical model to show that in the presence of ultra-fast traders both the volatility of prices and the price impact of liquidity trades increase. Martinez and Rosu (2013) show that high-frequency traders increase volatility and volume, 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. (2016) 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.

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

$$S_t^a = S_t + \delta/2\,, \quad S_t^b = S_t - \delta/2\,, \tag{1}$$

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 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 $\lambda^+ = \left(\lambda_t^+\right)_{t \geq 0}$ 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

$$\tau_a = \inf \{t \geq 0, \, S_t - S_0 = \delta\} \tag{2}$$

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

$$\tau = \tau_a \wedge T \tag{3}$$

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 $\max(N_{\tau_a}^+, M)$. 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 $\max\left(M - N^+_{\tau_a}, 0\right)$ 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 $\max(N^+_T,\, M)$ shares, and she cancels any unfilled part.

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

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

$$\Pi_2 = \left[\min(N^+_T,\, M)\, \left(S_0 + \frac{\delta}{2}\right) - \min(N^+_T,\, M)\, S_T\right] \mathbb{1}_{\{\tau=T\}}\,. \tag{4}$$

The first term on the right-hand side of (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

$$\Pi = \Pi_1 + \Pi_2\,, \tag{5}$$

and the MM maximises the expected profit by solving

$$\max_{\delta\geq 0} \mathbb{E}\left[\Pi\right]\,. \tag{6}$$

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

$$\lambda_t^+ = \lambda\, e^{-\kappa\,(S_0^a - S_t)}\,, \qquad \text{for} \qquad S_0^a - S_t < \delta\,. \tag{7}$$

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:

$$S_t = S_0 + X_t = S_0 + \sigma\, W_t\,, \tag{8}$$

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

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

$$\mathbb{E}\left[\Pi_1\right] = \mathbb{E}\left[-\frac{\delta}{2}\, M\, \mathbb{1}_{\{\tau=\tau_a\}}\right] \qquad \text{and} \qquad \mathbb{E}\left[\Pi_2\right] = \mathbb{E}\left[\min(N_T^+,\, M)\, \left(-X_T + \frac{\delta}{2}\right)\, \mathbb{1}_{\{\tau=T\}}\right]\,,$$

which are the expectations of the first and second terms in the right-hand side of (5). The first expectation is straightforward to calculate and is shown in the following proposition.

Proposition 1. *The expectation of* Π_1 *is given by*

$$\mathbb{E}\left[\Pi_1\right] = -\delta\, M\, \left(1 - \Phi\left(\frac{\delta}{\sigma\,\sqrt{T}}\right)\right)\,, \tag{9}$$

where $\Phi(\,\cdot\,)$ *denotes the cumulative density function of a standard Normal random variable.*

Proof.

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

■

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

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

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

$$\partial_t g + \frac{1}{2}\,\sigma^2\, \partial_{xx} g + \lambda\, e^{-\kappa\,\left(\frac{\delta}{2}-x\right)}\, \left[g(t,x,q+1) - g(t,x,q)\right] = 0\,, \tag{10}$$

with terminal and boundary conditions

$$g(T,x,q) = \min(q,\, M)\left(-x + \frac{\delta}{2}\right), \quad g(t,\delta,q) = 0\,, \quad x < \delta\,, \quad t \in [0,T]\,. \tag{11}$$

We are not able to find a closed-form solution for (10), but for small values of λ we obtain approximate solutions by an asymptotic expansion. For liquid stocks traded in NASDAQ, the order of magnitude of the parameter λ is less than 1 per second – below we choose parameter values to study the strategy of the MM, these parameters are provided in Table 1. We conjecture that we can asymptotically expand the function g in λ, so that

$$g(t,x,q) = g_0(t,x,q) + \lambda\, g_1(t,x,q) + \cdots\,. \tag{12}$$

We give an intuitive justification of this conjecture and below, in Theorem 1, we show the accuracy of the expansion. The expectation $\mathbb{E}\left[\Pi_2\right]$ is a multiple of $\min(q,M)$ and the order of q is approximately λT. Policy proposals have suggested MRTs of around 500 milliseconds, so we expect the order of magnitude of q to be less than M. Hence, $\min(q,M)$ is approximately q.

We substitute the expression for $g(t,x,q)$ given in (12) into the PDE (10) to obtain

$$0 = \partial_t g_0 + \frac{1}{2}\,\sigma^2\,\partial_{xx} g_0 + \lambda\,\partial_t g_1 + \frac{\lambda}{2}\,\sigma^2\,\partial_{xx} g_1 + \lambda\, e^{-\kappa\left(\frac{\delta}{2}-x\right)}\left[g_0(q+1) - g_0(q)\right] + \cdots\,.$$

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

$$\partial_t g_0 + \frac{1}{2}\,\sigma^2\,\partial_{xx} g_0 = 0\,, \tag{13}$$

with terminal and boundary conditions

$$g_0(T,x,q) = \min(q,\, M)\left(-x + \frac{\delta}{2}\right), \quad g_0(t,\delta,q) = 0\,, \quad x < \delta\,, \quad t \in [0,T]\,. \tag{14}$$

The second PDE is

$$\partial_t g_1 + \frac{1}{2}\,\sigma^2\,\partial_{xx} g_1 + e^{-\kappa\left(\frac{\delta}{2}-x\right)}\left[g_0(t,x,q+1) - g_0(t,x,q)\right] = 0\,, \tag{15}$$

with terminal and boundary conditions

$$g_1(T,x,q) = 0\,, \quad g_1(t,\delta,q) = 0\,, \quad x < \delta\,, \quad t \in [0,T]\,. \tag{16}$$

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 2. *Let the function g_0 satisfy* (13) *with boundary and terminal conditions (14), then*

$$g_0(t,x,q) = \min(q,\, M)\, \left(-x + \frac{3\,\delta}{2} - \delta\,\Phi\left(\frac{\delta - x}{\sigma\,\sqrt{T-t}}\right)\right)\,. \tag{17}$$

Proof. For a proof see the Appendix. ■

Proposition 3. *Let $g_1(t,x,q)$ satisfy (15) with terminal and boundary conditions (16), then*

$$\begin{aligned}
g_1 \;=\;& D(q,M)\,\exp\left(\frac{3\,\kappa\,\delta}{2} - \kappa\,x\right)\int_0^{\tilde{T}} \exp\left(\frac{1}{2}\,\sigma^2\,\kappa^2\,\tilde{s}\right)\left\{-\sqrt{\frac{\sigma^2\,\tilde{s}}{2\,\pi}}\,\exp\left(-\frac{(x-\delta-\sigma^2\,\tilde{s}\,\kappa)^2}{2\,\sigma^2\,\tilde{s}}\right)\right.\\
&\left.-\left(x - \sigma^2\,\tilde{s}\,\kappa - \frac{\delta}{2}\right)\Phi\left(\frac{x-\delta-\sigma^2\,\tilde{s}\,\kappa}{\sigma\,\sqrt{\tilde{s}}}\right) + \delta\,\Phi\left(\frac{x-\delta-\sigma^2\,\tilde{s}\,\kappa}{\sigma\,\sqrt{\tilde{s}}},\,\frac{x-\delta-\sigma^2\,\tilde{s}\,\kappa}{\sigma\,\sqrt{\tilde{T}}};\,\sqrt{\frac{\tilde{s}}{\tilde{T}}}\right)\right\}\mathrm{d}s\\
&+D(q,M)\,\exp\left(-\frac{\kappa\,\delta}{2} + \kappa\,x\right)\int_0^{\tilde{T}} \exp\left(\frac{1}{2}\,\sigma^2\,\kappa^2\,\tilde{s}\right)\left\{\sqrt{\frac{\sigma^2\,\tilde{s}}{2\,\pi}}\,\exp\left(-\frac{(x-\delta+\sigma^2\,\tilde{s}\,\kappa)^2}{2\,\sigma^2\,\tilde{s}}\right)\right.\\
&\left.-\left(x + \sigma^2\,\tilde{s}\,\kappa - \frac{3\,\delta}{2}\right)\Phi\left(-\frac{x-\delta+\sigma^2\,\tilde{s}\,\kappa}{\sigma\,\sqrt{\tilde{s}}}\right) - \delta\,\Phi\left(-\frac{x-\delta+\sigma^2\,\tilde{s}\,\kappa}{\sigma\,\sqrt{\tilde{s}}},\,-\frac{x-\delta+\sigma^2\,\tilde{s}\,\kappa}{\sigma\,\sqrt{\tilde{T}}};\,\sqrt{\frac{\tilde{s}}{\tilde{T}}}\right)\right\}\mathrm{d}s\,,
\end{aligned}$$

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 bivariate Normal cumulative distribution function with correlation ρ.

Proof. For a proof see the Appendix. ■

Now we show the accuracy of the asymptotic expansion. First we define the infinitesimal generator $\mathcal{L}$, acting on a sufficiently differentiable function f, as

$$\mathcal{L}f = \partial_t f + \frac{1}{2}\,\sigma^2\,\partial_{xx} f + \lambda\, e^{-\kappa\,\left(\frac{\delta}{2}-x\right)}\,\left[f(t,x,q+1) - f(t,x,q)\right]\,.$$

Theorem 1. ***Accuracy of asymptotic expansion.***

$$g_e(t,x,q) := g(t,x,q) - g_0(t,x,q) - \lambda\, g_1(t,x,q) = o(\lambda)\,.$$

Proof. For a proof see the Appendix. ■

3.2. Value function

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

$$G_1(\delta\,;\,M\,,T\,,\sigma) = -\,\delta\,M\,\left(1 - \Phi\left(\frac{\delta}{\sigma\,\sqrt{T}}\right)\right)\,, \tag{18}$$

and

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

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

Therefore, the value function is given by

$$G(\delta\,;\, M\,, T\,, \lambda\,, \sigma) = G_1(\delta\,;\, M\,, T\,, \sigma)\; +\; G_2(\delta\,;\, M\,, T\,, \lambda\,, \sigma)\,. \tag{19}$$

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 the risk arising from LOs becoming stale due to the MRT.

Proposition 4. *For $M \geq 1$,*

$$G(\delta\,;\, M\,, T\,, \lambda\,, \sigma) \geq G(\delta\,;\, M+1\,, T\,, \lambda\,, \sigma)\,. \tag{20}$$

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 5.

$$G(0) = 0\,, \qquad \lim_{\delta \to +\infty} G(\delta) = 0\,, \tag{21}$$

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 (21) 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 (6) as

$$\delta^* = \underset{\delta\in[0,+\infty)}{\operatorname{argmax}} G(\delta)\,. \tag{22}$$

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.

3.3. Numerical study and simulations

We use data for 9 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 1 shows the mean of all daily estimates (and the standard deviation of the mean).

Table 1: Estimates of σ^2 (per second) and λ (per second), with corresponding standard deviations.

Symbol	$\widehat{\lambda}_{MO}$	$std(\widehat{\lambda}_{MO})$	$\widehat{\sigma}^2$	$\sqrt{\widehat{\sigma}^2}$	$std(\widehat{\sigma}^2)$
AAPL	0.253707	0.101395	0.026901	0.164015	0.012417
EBAY	0.173684	0.083055	0.000102	0.010099	0.000055
INTC	0.085932	0.029271	0.000007	0.002645	0.000005
GOOG	0.083900	0.032641	0.261801	0.511664	0.284895
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
FARO	0.004936	0.003560	0.036585	0.191272	0.020275

In the table, the volatility of prices is in the range $10^{-3}/\text{second}^{\frac{1}{2}}$ to $10^{-1}/\text{second}^{\frac{1}{2}}$. Thus, in our numerical study, we choose the order of magnitude of the volatility parameter σ to be 10^{-2} and choose the rate of decay of the fill rate of the LOs to be $\kappa = 100$.

Figure 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 (19) 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.

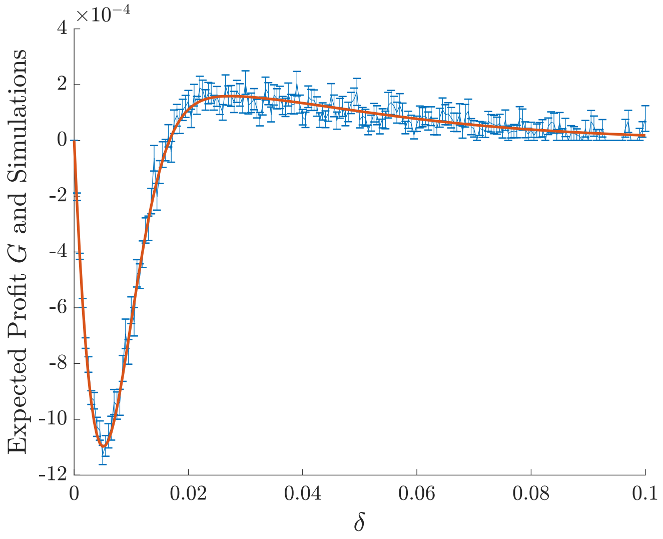

Figure 1: Expected profit and 2,000 simulations (standard deviation of each simulation is also shown). MMs post LOs of same volume. Parameters: $\kappa = 10^2$, $M = 1$, $T = 0.5$ seconds, $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$ and $\lambda = 0.1/\text{second}$.

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 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 5.

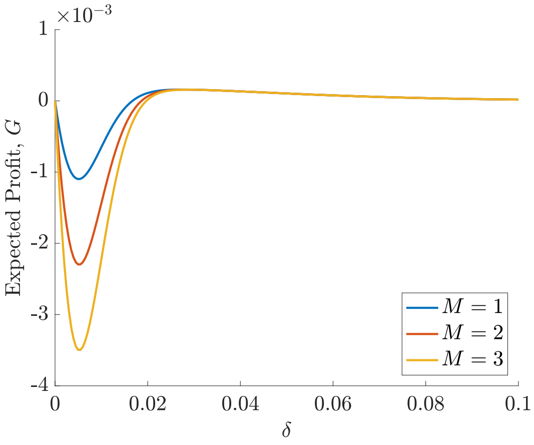

Figure 2: Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1/\text{second}$, $T = 0.5$ seconds.

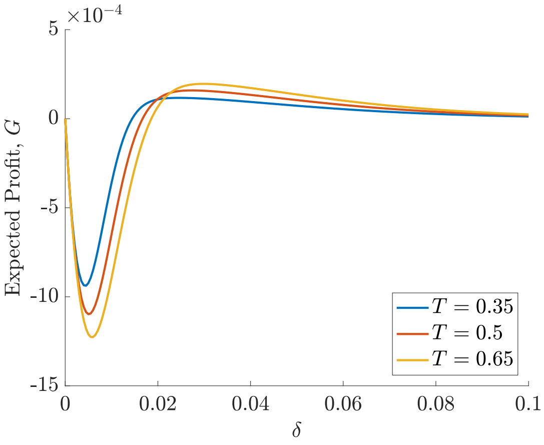

Figure 3: Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1/\text{second}$, $M = 1$.

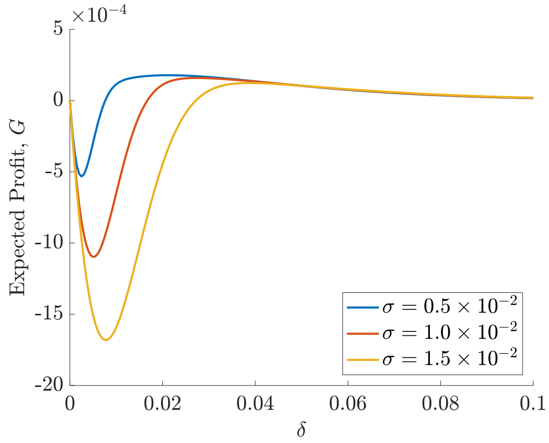

Figure 4: Expected profit. Parameters: $T = 0.5$ seconds, $\lambda = 0.1$/second, $M = 1$.

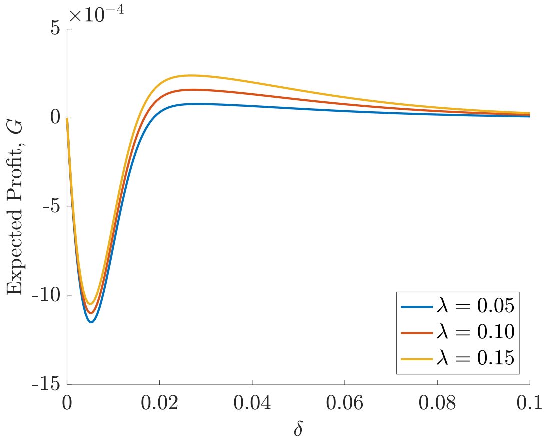

Figure 5: Expected profit. Parameters: $T = 0.5$ seconds , $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $M = 1$.

Figures 2 to 5 show the expected profit for various parameter values. For example, Figure 2 shows that the expected profit decreases as the volume of the LO posted by the MM increases, which agrees with Proposition 4. Clearly, as the MM posts LOs of higher volumes, the exposure to being picked off is higher.

Figure 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. 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 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. 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

$$\hat{\delta}^*(M) = \underset{\hat{\delta}\in[0,+\infty)}{\operatorname{argmax}}\, \widehat{G}(\hat{\delta}; M)\,. \tag{23}$$

This is the analogue of (22) and our notation in (23) 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 1. ***Optimal depth for LOs with volume*** M***.*** *The best offers in the LOB are of volume* $M = 1$*, i.e.,*

$$\hat{\delta}^*(M) > \hat{\delta}^*(1)\,, \quad \textit{for} \quad M > 1\,. \tag{24}$$

Based on Conjecture 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 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 6. ***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.,*

$$\hat{\delta}^*(1) = \delta^*(1)\,, \tag{25}$$

where $\delta^*(1)$ *is derived above, see* (22).

Proof. From Conjecture 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

$$\widehat{\tau}_a = \inf\left\{ t \geq 0,\, S_t - S_0 = \hat{\delta}(M)/2 + \delta_1^*/2 \right\} . \tag{26}$$

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

We denote by

$$\widehat{\tau} = \widehat{\tau}_a \wedge T \tag{27}$$

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 $\widehat{\tau}$, which is also the reference time employed to calculate the expected profit.

When $\widehat{\tau} = \widehat{\tau}_a$, 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 $\widehat{\tau}_a$, the volume of the sell LO filled is $\max(N^+_{\widehat{\tau}_a}, M)$ 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 $\max(M - N^+_{\widehat{\tau}_a}, 0)$ 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 $\widehat{\tau} = \widehat{\tau}_a$.

The other case is when $\widehat{\tau} = T$, 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 $\max(N^+_T, M)$ 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.

4.1. Performance criterion: expected profit

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

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

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

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

The MM maximises the expected profit by solving

$$\max_{\hat{\delta}} \mathbb{E}\left[\widehat{\Pi}_1 + \widehat{\Pi}_2 \right] \,. \tag{28}$$

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 7.

$$\mathbb{E}\left[\widehat{\Pi}_1\right] = -M\,\delta_1^* \left(1 - \Phi\left(\frac{\hat{\delta}(M) + \delta_1^*}{2\,\sigma\,\sqrt{T}} \right) \right) \,. \tag{29}$$

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

Proof. Similar to Proposition 1. For a proof see the Appendix. ■

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

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

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

$$\partial_t \widehat{g} + \frac{1}{2}\,\sigma^2\,\partial_{xx}\widehat{g} + \lambda\, e^{-\kappa\left(\frac{\hat{\delta}}{2}-x\right)}\,\left[\widehat{g}(t,x,q+1) - \widehat{g}(t,x,q)\right] = 0\,, \tag{30}$$

with boundary and terminal conditions

$$\widehat{g}(T,x,q) = \min(q,\, M)\left(-x + \frac{\hat{\delta}}{2} \right)\,, \quad \widehat{g}(t,\widetilde{\delta},q) = 0\,, \quad x < \widetilde{\delta}\,, \quad t \in [0,T]\,, \tag{31}$$

where

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

We conjecture that we can asymptotically expand g in λ, so

$$\widehat{g}(t,x,q) = \widehat{g}_0(t,x,q) + \lambda\,\widehat{g}_1(t,x,q) + \cdots\,.$$

We substitute the above expansion of $\widehat{g}$ in PDE (30), and by equating to zero the terms of order 1 and the terms of order λ we obtain PDEs for $\widehat{g}_0$ and $\widehat{g}_1$. We absorb the boundary and terminal conditions in $\widehat{g}_0$ because they do not include the parameter λ. The first PDE is

$$\partial_t \widehat{g}_0 + \frac{1}{2}\,\sigma^2\,\partial_{xx}\widehat{g}_0 = 0\,, \tag{32}$$

with terminal and boundary conditions

$$\widehat{g}_0(T,x,q) = \min(q,\,M)\left(-x + \frac{\hat{\delta}}{2}\right)\,, \quad \widehat{g}_0(t,\widetilde{\delta},q) = 0\,, \quad x < \widetilde{\delta}\,, \quad t \in [0,T]\,, \tag{33}$$

and the second PDE is

$$\partial_t \widehat{g}_1 + \frac{1}{2}\,\sigma^2\,\partial_{xx}\widehat{g}_1 + e^{-\kappa\left(\frac{\hat{\delta}}{2}-x\right)}\,\left[\widehat{g}_0(t,x,q+1) - \widehat{g}_0(t,x,q)\right] = 0\,, \tag{34}$$

with terminal and boundary conditions

$$\widehat{g}_1(T,x,q) = 0\,, \quad \widehat{g}_1(t,\widetilde{\delta},q) = 0\,, \quad x < \widetilde{\delta}\,, \quad t \in [0,T]\,. \tag{35}$$

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

Proposition 8. *Let $\widehat{g}_0$ satisfy* (32) *with terminal and boundary conditions* (33)*, then*

$$\widehat{g}_0(t,x,q) = \min(q,\,M)\left(-x + \frac{\hat{\delta}(M)}{2} + \delta_1^* - \delta_1^*\,\Phi\left(\frac{\widetilde{\delta}-x}{\sigma\sqrt{T-t}}\right)\right)\,. \tag{36}$$

Proof. Similar to Proposition 2. For a proof see the Appendix. ■

Proposition 9. *Let $\widehat{g}_1$ satisfy (34) with terminal and boundary conditions in (35), then*

$$\begin{aligned}
\widehat{g}_1 \;=\;& D(q,M)\,\exp\left(\frac{\kappa\,\delta_1^*}{2} - \kappa(x-\widetilde{\delta})\right)\int_0^{\tilde{T}} \exp\left(\frac{1}{2}\,\sigma^2\,\kappa^2\,\tilde{s}\right)\left\{-\sqrt{\frac{\sigma^2\,\tilde{s}}{2\,\pi}}\,\exp\left(-\frac{(x-\widetilde{\delta}-\sigma^2\,\tilde{s}\,\kappa)^2}{2\,\sigma^2\,\tilde{s}}\right)\right.\\
&\left. -\left(x-\widetilde{\delta}-\sigma^2\,\tilde{s}\,\kappa + \frac{\delta_1^*}{2}\right)\,\Phi\left(\frac{x-\widetilde{\delta}-\sigma^2\,\tilde{s}\,\kappa}{\sigma\sqrt{\tilde{s}}}\right) + \delta_1^*\,\Phi\left(\frac{x-\widetilde{\delta}-\sigma^2\,\tilde{s}\,\kappa}{\sigma\sqrt{\tilde{s}}}, \frac{x-\widetilde{\delta}-\sigma^2\,\tilde{s}\,\kappa}{\sigma\sqrt{\tilde{T}}}; \sqrt{\frac{\tilde{s}}{\tilde{T}}}\right)\right\}\mathrm{d}s\\
&+D(q,M)\,\exp\left(\frac{\kappa\,\delta_1^*}{2} + \kappa(x-\widetilde{\delta})\right)\int_0^{\tilde{T}} \exp\left(\frac{1}{2}\,\sigma^2\,\kappa^2\tilde{s}\right)\left\{\sqrt{\frac{\sigma^2\,\tilde{s}}{2\,\pi}}\,\exp\left(-\frac{(x-\widetilde{\delta}+\sigma^2\,\tilde{s}\,\kappa)^2}{2\,\sigma^2\,\tilde{s}}\right)\right.\\
&\left. -\left(x-\widetilde{\delta}+\sigma^2\,\tilde{s}\,\kappa - \frac{\delta_1^*}{2}\right)\,\Phi\left(-\frac{x-\widetilde{\delta}+\sigma^2\,\tilde{s}\,\kappa}{\sigma\sqrt{\tilde{s}}}\right) - \delta_1^*\,\Phi\left(-\frac{x-\widetilde{\delta}+\sigma^2\,\tilde{s}\,\kappa}{\sigma\sqrt{\tilde{s}}}, -\frac{x-\widetilde{\delta}+\sigma^2\,\tilde{s}\,\kappa}{\sigma\sqrt{\tilde{T}}}; \sqrt{\frac{\tilde{s}}{\tilde{T}}}\right)\right\}\mathrm{d}s\,,
\end{aligned}$$

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

Proof. Similar to Proposition 3. For a proof see the Appendix. ■

The proof of the accuracy of the asymptotic expansion is the same as that in Theorem 1, note that δ_1^*

is finite.

4.2. Value function

For $M \geq 2$ we define

$$\widehat{G}_1(\hat{\delta}\,;\,M\,,T\,,\sigma) = -M\,\delta_1^*\left(1-\Phi\left(\frac{\hat{\delta}+\delta_1^*}{2\,\sigma\,\sqrt{T}}\right)\right)\,, \tag{37}$$

and

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

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

Therefore the value function is given by

$$\widehat{G}(\hat{\delta}\,;\,M\,,T\,,\lambda\,,\sigma) = \widehat{G}_1(\hat{\delta}\,;\,M\,,T\,,\sigma)\;+\;\widehat{G}_2(\hat{\delta}\,;\,M\,,T\,,\lambda\,,\sigma)\,, \tag{38}$$

which is an approximation to the expected mark-to-market inventory $\mathbb{E}\left[\widehat{\Pi}\right]$ 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 $\mathbb{E}\left[\widehat{\Pi}\right]$, see Figure 6.

The proposition below shows that $\widehat{G}$ is decreasing in M, so the expected profit is higher when the MM posts an LO with less volume for $M > 2$. Therefore, provided there are other MMs who post LOs with volume 1, the MM prefers to post an 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 10. *For $M \geq 2$,*

$$\widehat{G}(\hat{\delta}\,;\,M\,,T\,,\lambda\,,\sigma) \geq \widehat{G}(\hat{\delta}\,;\,M+1\,,T\,,\lambda\,,\sigma)\,. \tag{39}$$

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

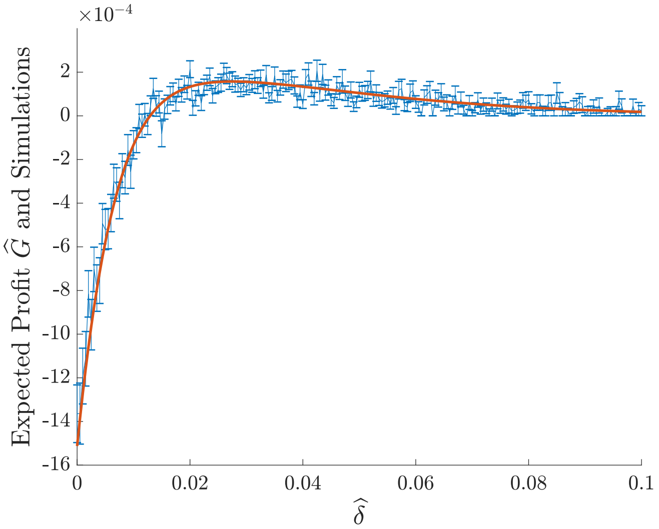

Figure 6: Expected profit and 2,000 simulations (standard deviation of each simulation is also shown). The other MMs post LOs of various volumes. $M = 2$, , $T = 0.5$ seconds, $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1/\text{second}$.

The following proposition shows properties of the expected profit.

Proposition 11.

$$\widehat{G}(0) \textit{ is bounded}, \quad \lim_{\hat{\delta}\to+\infty} \widehat{G}(\hat{\delta}) = 0\,, \tag{40}$$

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,

$$\hat{\delta}^* = \underset{\hat{\delta}\in[\delta_1^*,+\infty)}{\text{argmax}}\ \widehat{G}(\hat{\delta})\,. \tag{41}$$

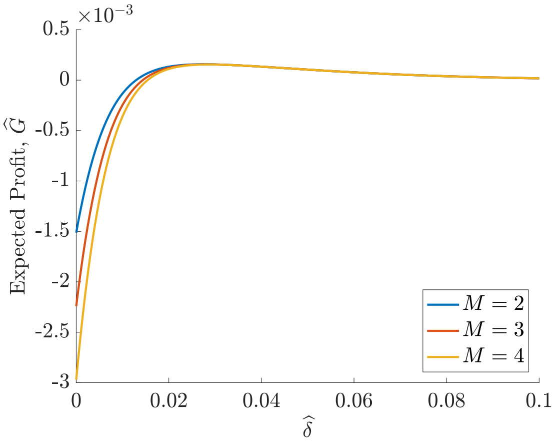

Figure 7: Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1$second, $T = 0.5$ seconds.

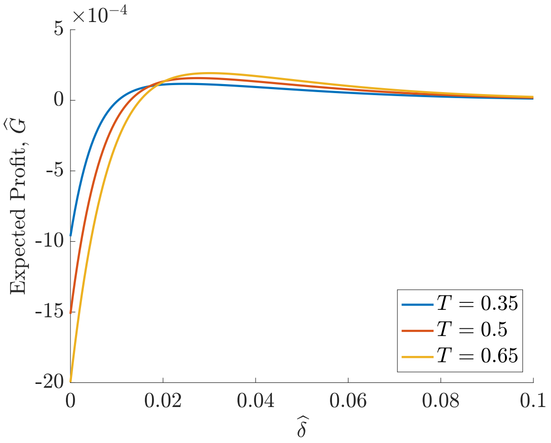

Figure 8: Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1$/second, $M = 2$.

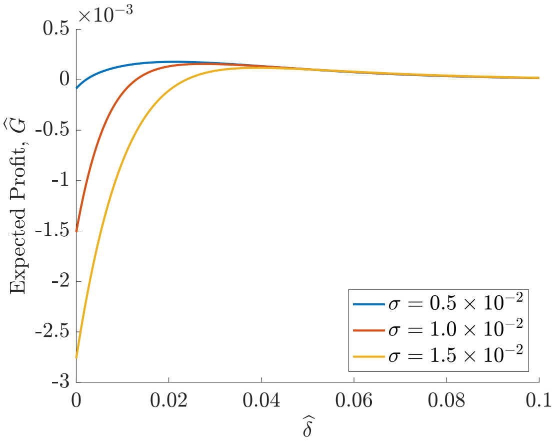

Figure 9: Expected profit. Parameters: $T = 0.5$ seconds, $\lambda = 0.1$/second, $M = 2$.

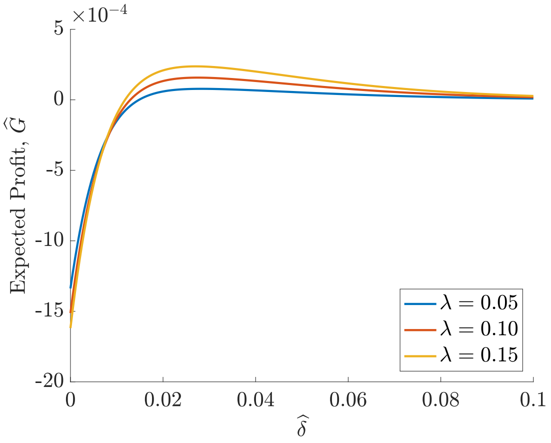

Figure 10: Expected profit. Parameters: $T = 0.5$ seconds, $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $M = 2$.

Figures 7 to 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 2 to 5.

5. Optimal depth

In this section we discuss the depths that result from the strategy developed in Section 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, where we assume that MMs post LOs of any positive integer volume, i.e., Model II.

Figure 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 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 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.

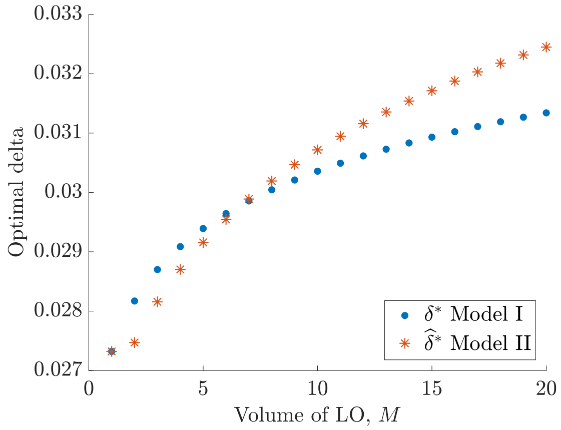

Figure 11: Optimal delta. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1/\text{second}$, $T = 0.5$ seconds.

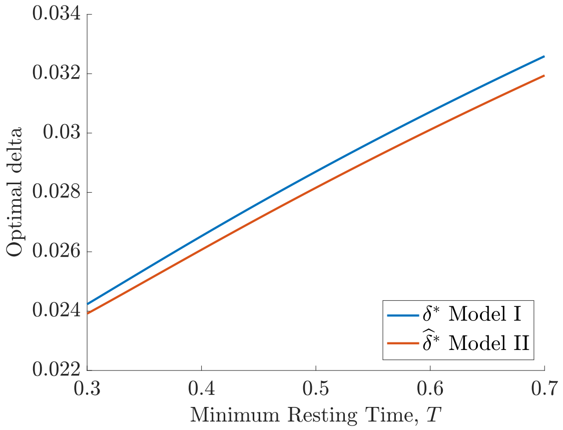

Figure 12: Optimal delta. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1/\text{second}$, $M = 3$.

Figure 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 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.

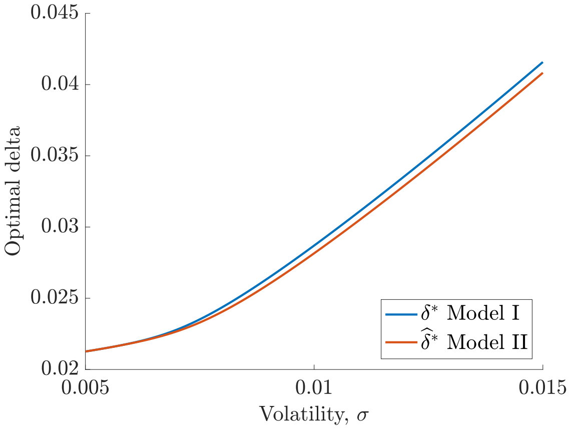

Figure 13: Optimal delta. Parameters: $T = 0.5$ seconds, $\lambda = 0.1$/second, $M = 3$.

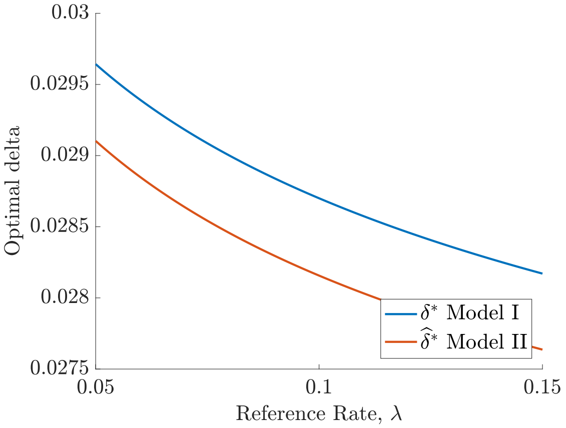

Figure 14: Optimal delta. Parameters: $T = 0.5$ seconds, $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $M = 3$.

6. Expected profits and liquidity provision

Figures 15 to 18 show the expected profits obtained by the MM for different values of: posted volume, MRT, volatility, and reference fill rate.

Figure 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) 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 10). 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 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.

Figure 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 18 shows that the expected profits increase with the reference fill rate λ, see (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.

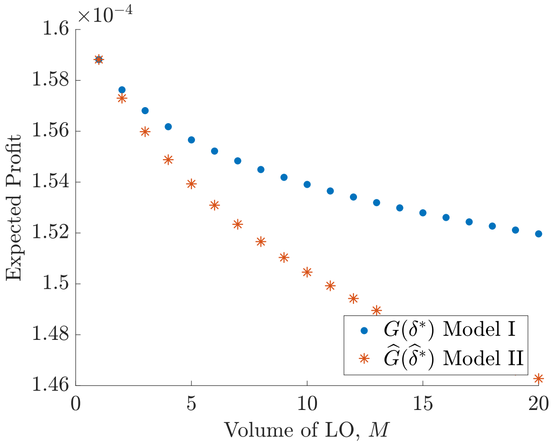

Figure 15: Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1/\text{second}$, $T = 0.5$ seconds.

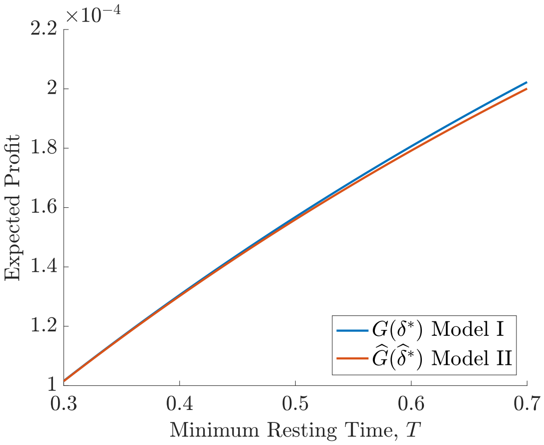

Figure 16: Expected profit. Parameters: $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $\lambda = 0.1/\text{second}$, $M = 3$

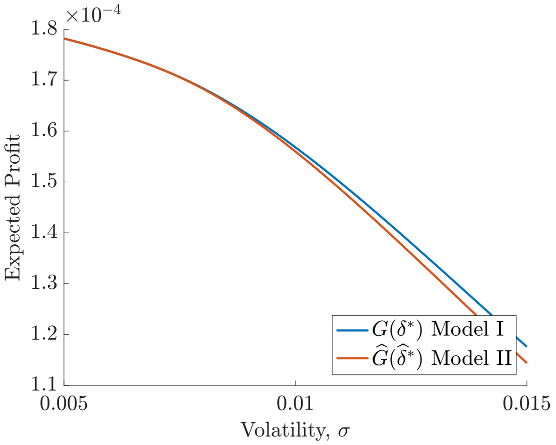

Figure 17: Expected profit. Parameters: $T = 0.5$ seconds, $\lambda = 0.1/\text{second}$, $M = 3$.

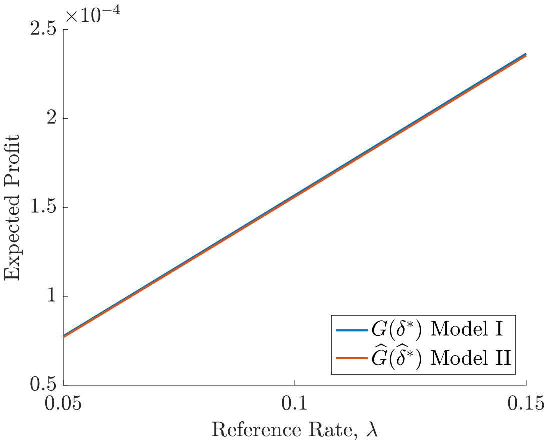

Figure 18: Expected profit. Parameters: $T = 0.5$ seconds, $\sigma = 10^{-2}/\text{second}^{\frac{1}{2}}$, $M = 3$.

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.

We propose several directions for future research.

(i) In our model, 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.

(ii) 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 trader.

(iii) Examine how price impact of liquidity taking orders is affected by MRTs. In this paper 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.

(iv) Finally, a point not discussed in the paper, but of great importance to all stakeholders, is whether MRTs will make spoofing and layering activities less viable. In order-driven markets, spoofing and layering are strategies that provide false information about the demand and supply of an asset, see Cartea et al. (2018). These trading strategies are illegal and profit from market participants who trade on misleading market signals. Spoofing and layering are based on posting and very quickly cancelling LOs. The effect of MRTs on these illegal strategies is not clear and more work is required in this direction.

References

Ait-Sahalia, Y. and M. Sağlam (2017). High frequency market making: Implications for liquidity. SSRN.

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.

Cartea, Á., R. Donnelly, and S. Jaimungal (2018). Enhancing trading strategies with order book signals. Applied Mathematical Finance, 1–35.

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. Walton (2015). Foreign exchange markets with last look. Mathematics and Financial Economics, 1–30.

Cartea, Á., S. Jaimungal, and Y. Wang (2018). Spoofing and manipulation in order driven markets. SSRN.

Cartea, Á., R. Payne, J. Penalva, and M. Tapia (2016). Ultra-fast activity and intraday market quality. SSRN.

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.

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.

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.

Guéant, O. (2015). Optimal execution and block trade pricing: A general framework. Applied Mathematical Finance 22(4), 336–365.

Guéant, O. (2017). Optimal market making. Applied Mathematical Finance 24(2), 112–154.

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.

Hoffmann, P. (2014). A dynamic limit order market with fast and slow traders. Journal of Financial Economics 113(1), 156 – 169.

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.

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.

Schied, A., T. Schöneborn, and M. Tehranchi (2010). Optimal basket liquidation for CARA investors is deterministic. Applied Mathematical Finance 17(6), 471–489.

Van Ness, B. F., R. A. Van Ness, and E. D. Watson (2015). Canceling liquidity. Journal of Financial Research 38(1), 3–33.

A. Proofs

A.1. Proof of Proposition 2

Proof. We make the standard change of variables, $y = x - \delta$, $\tilde{t} = T - t$, $h_0(\tilde{t}, y) = g_0(t, x)$, so the PDE becomes

$$\partial_{\tilde{t}} h_0 = \frac{1}{2}\,\sigma^2\,\partial_{yy} h_0\,, \tag{42}$$

and 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 (42) is the heat equation with heat kernel

$$K(x-y,t) = \frac{1}{\sqrt{2\,\pi\,\sigma^2\,t}}\,e^{-\frac{(x-y)^2}{2\,\sigma^2\,t}}\,.$$

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 (42) is given by

$$\begin{aligned} h_0(\tilde{t},y) &= \int_{-\infty}^{\infty} K\left(y-z,\tilde{t}\right)\,u_0(z)\,\mathrm{d}z\\ &= \min(q,\,M)\,\left(\int_0^{\infty} K\left(y-z,\tilde{t}\right)\left(-z+\frac{\delta}{2}\right)\,\mathrm{d}z + \int_{-\infty}^{0} K\left(y-z,\tilde{t}\right)\left(-z-\frac{\delta}{2}\right)\,\mathrm{d}z\right)\\ &= \min(q,\,M)\,\left(-\int_0^{\infty}\frac{z}{\sqrt{2\,\pi\,\sigma^2\,\tilde{t}}}\,e^{-\frac{(y-z)^2}{2\,\sigma^2\,\tilde{t}}}\,\mathrm{d}z + \frac{\delta}{2}\int_0^{\infty}\frac{1}{\sqrt{2\,\pi\,\sigma^2\,\tilde{t}}}\,e^{-\frac{(y-z)^2}{2\,\sigma^2\,\tilde{t}}}\,\mathrm{d}z\right.\\ &\qquad\left. -\int_{-\infty}^{0}\frac{z}{\sqrt{2\,\pi\,\sigma^2\,\tilde{t}}}\,e^{-\frac{(y-z)^2}{2\,\sigma^2\,\tilde{t}}}\,\mathrm{d}z - \frac{\delta}{2}\int_{-\infty}^{0}\frac{1}{\sqrt{2\,\pi\,\sigma^2\,\tilde{t}}}\,e^{-\frac{(y-z)^2}{2\,\sigma^2\,\tilde{t}}}\,\mathrm{d}z\right)\\ &= \min(q,\,M)\,\left(-y+\frac{\delta}{2}-\delta\,\Phi\left(\frac{-y}{\sigma\,\sqrt{\tilde{t}}}\right)\right)\,, \end{aligned}$$

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

A.2. Proof of Proposition 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 (15) becomes

$$\partial_{\tilde{t}} h_1 = \frac{1}{2}\,\sigma^2\,\partial_{yy} h_1 + f(\tilde{t}, y)\,, \tag{43}$$

where

$$f(\tilde{t}, y) = D(q, M)\, e^{\kappa\,\left(\frac{\delta}{2}+y\right)} \left[-y + \frac{\delta}{2} - \delta\,\Phi\left(\frac{-y}{\sigma\sqrt{\tilde{t}}}\right)\right]\,,$$

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)\, \mathrm{d}z\, \mathrm{d}s \\
=& \int_0^{\tilde{t}} \int_0^{\infty} K(y - z, \tilde{t} - s)\, (-f(s, -z))\, \mathrm{d}z\, \mathrm{d}s + \int_0^{\tilde{t}} \int_{-\infty}^{0} K(y - z, \tilde{t} - s)\, f(s, z)\, \mathrm{d}z\, \mathrm{d}s \\
=& 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}{\sigma\sqrt{\tilde{t}}}\right)\right)\right] \mathrm{d}z\, \mathrm{d}s \\
&+ 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}{\sigma\sqrt{\tilde{t}}}\right)\right)\right] \mathrm{d}z\, \mathrm{d}s \\
=& 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}{\sigma\sqrt{\tilde{t}}}\right)\right] \mathrm{d}z\, \mathrm{d}s \\
&+ 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}{\sigma\sqrt{\tilde{t}}}\right)\right] \mathrm{d}z\, \mathrm{d}s \\
=& D(q, M) \exp\left(\frac{\kappa\,\delta}{2} - \kappa\,y + \frac{1}{2}\,\kappa^2\,\sigma^2\,\tilde{t}\right) \int_0^{\tilde{t}} \exp\left(-\frac{1}{2}\,\sigma^2\,\kappa^2 s\right) \left\{-\sqrt{\frac{\sigma^2\,(\tilde{t} - s)}{2\,\pi}} \exp\left(-\frac{(y - \sigma^2\,(\tilde{t} - s)\,\kappa)^2}{2\,\sigma^2\,(\tilde{t} - s)}\right)\right. \\
&\left. - \left(y - \sigma^2\,(\tilde{t} - s)\,\kappa + \frac{\delta}{2}\right) \Phi\left(\frac{y - \sigma^2\,(\tilde{t} - s)\,\kappa}{\sigma\,\sqrt{\tilde{t} - s}}\right) + \delta\,\Phi\left(\frac{y - \sigma^2\,(\tilde{t} - s)\,\kappa}{\sigma\,\sqrt{\tilde{t} - s}}, \frac{y - \sigma^2\,(\tilde{t} - s)\,\kappa}{\sigma\,\sqrt{\tilde{t}}}; \sqrt{\frac{\tilde{t} - s}{\tilde{t}}}\right)\right\} \mathrm{d}s \\
&+ D(q, M) \exp\left(\frac{\kappa\,\delta}{2} + \kappa\,y + \frac{1}{2}\,\kappa^2\,\sigma^2\,\tilde{t}\right) \int_0^{\tilde{t}} \exp\left(-\frac{1}{2}\,\sigma^2\,\kappa^2 s\right) \left\{\sqrt{\frac{\sigma^2\,(\tilde{t} - s)}{2\,\pi}} \exp\left(-\frac{(y + \sigma^2\,(\tilde{t} - s)\,\kappa)^2}{2\,\sigma^2\,(\tilde{t} - s)}\right)\right. \\
&\left. - \left(y + \sigma^2\,(\tilde{t} - s)\,\kappa - \frac{\delta}{2}\right) \Phi\left(-\frac{y + \sigma^2\,(\tilde{t} - s)\,\kappa}{\sigma\,\sqrt{\tilde{t} - s}}\right) - \delta\,\Phi\left(-\frac{y + \sigma^2\,(\tilde{t} - s)\,\kappa}{\sigma\,\sqrt{\tilde{t} - s}}, -\frac{y + \sigma^2\,(\tilde{t} - s)\,\kappa}{\sigma\,\sqrt{\tilde{t}}}; \sqrt{\frac{\tilde{t} - s}{\tilde{t}}}\right)\right\} \mathrm{d}s\,,
\end{aligned} \tag{44}$$

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

A.3. Proof of Theorem 1

Proof. We apply the infinitesimal generator $\mathcal{L}$ on g_e to obtain

$$\begin{aligned}\mathcal{L}g_e =&\, \partial_t g_e + \frac{1}{2}\,\sigma^2\,\partial_{xx}g_e + \lambda\,e^{-\kappa\left(\frac{\delta}{2}-x\right)}\,[g_e(q+1)-g_e(q)]\\ =&\, \partial_t g + \frac{1}{2}\,\sigma^2\,\partial_{xx}g + \lambda\,e^{-\kappa\left(\frac{\delta}{2}-x\right)}\,[g(q+1)-g(q)]\\ &-\partial_t g_0 - \frac{1}{2}\,\sigma^2\,\partial_{xx}g_0 - \lambda\,e^{-\kappa\left(\frac{\delta}{2}-x\right)}\,[g_0(q+1)-g_0(q)]\\ &-\lambda\,\partial_t g_1 - \frac{\lambda}{2}\,\sigma^2\,\partial_{xx}g_1 - \lambda^2\,e^{-\kappa\left(\frac{\delta}{2}-x\right)}\,[g_1(q+1)-g_1(q)]\\ =&\, -\lambda^2\,e^{-\kappa\left(\frac{\delta}{2}-x\right)}\,[g_1(q+1)-g_1(q)]\;.\end{aligned}$$

Thus, we write

$$\begin{aligned}g_e(t,x,q) =&\, \mathbb{E}_{t,X_t,N_t^+}\left[g_e(T,X_T,N_T^+) - \int_t^T \mathcal{L}g_e(s,X_s,N_s^+)\,\mathrm{d}s\right]\\ =&\, \lambda^2\,\mathbb{E}_{t,X_t,N_t^+}\left[\int_t^T e^{-\kappa\left(\frac{\delta}{2}-X_s\right)}\,\left(g_1(s,X_s,N_s^++1)-g_1(s,X_s,N_s^+)\right)\,\mathrm{d}s\right]\;.\end{aligned}$$

The expectation above is bounded:

$$\begin{aligned}&\left|\mathbb{E}_{t,X_t,N_t^+}\left[\int_t^T e^{-\kappa\left(\frac{\delta}{2}-X_s\right)}\,\left(g_1(s,X_s,N_s^++1)-g_1(s,X_s,N_s^+)\right)\,\mathrm{d}s\right]\right|\\ &\qquad\leq\, \mathbb{E}_{t,X_t,N_t^+}\left[\int_t^T e^{-\kappa\left(\frac{\delta}{2}-X_s\right)}\,\left|g_1(s,X_s,N_s^++1)-g_1(s,X_s,N_s^+)\right|\mathrm{d}s\right]\;.\end{aligned}$$

To bound the integrand, we have

$$
\begin{aligned}
&|g_1(t,x,q+1) - g_1(t,x,q)| \\
\leq\; & |D(q+1,M) - D(q,M)| \exp\left(\frac{3\,\kappa\,\delta}{2} - \kappa\,x + \frac{1}{2}\,\kappa^2\,\sigma^2\,(T-t)\right) \\
&\times \int_0^{T-t} \exp\left(-\frac{1}{2}\,\sigma^2\,\kappa^2\,s\right)\left\{\sqrt{\frac{\sigma^2\,(T-t-s)}{2\,\pi}}\exp\left(-\frac{(x-\delta-\sigma^2\,(T-t-s)\,\kappa)^2}{2\,\sigma^2\,(T-t-s)}\right)\right. \\
&+ \left(|x| + \sigma^2\,(T-t-s)\,\kappa + \frac{\delta}{2}\right)\Phi\left(\frac{x-\delta-\sigma^2\,(T-t-s)\,\kappa}{\sigma\,\sqrt{T-t-s}}\right) \\
&\left.+ \,\delta\,\Phi\left(\frac{x-\delta-\sigma^2\,(T-t-s)\,\kappa}{\sigma\,\sqrt{T-t-s}}, \frac{x-\delta-\sigma^2\,(T-t-s)\,\kappa}{\sigma\,\sqrt{T-t}}; \sqrt{\frac{T-t-s}{T-t}}\right)\right\}\mathrm{d}s \\
&+ |D(q+1,M) - D(q,M)| \exp\left(-\frac{\kappa\,\delta}{2} + \kappa\,x + \frac{1}{2}\,\kappa^2\,\sigma^2\,(T-t)\right) \\
&\times \int_0^{T-t} \exp\left(-\frac{1}{2}\,\sigma^2\,\kappa^2\,s\right)\left\{\sqrt{\frac{\sigma^2\,(T-t-s)}{2\,\pi}}\exp\left(-\frac{(x-\delta+\sigma^2\,(T-t-s)\,\kappa)^2}{2\,\sigma^2\;(T-t-s)}\right)\right. \\
&+ \left(|x| + \sigma^2\,(T-t-s)\,\kappa + \frac{3\,\delta}{2}\right)\Phi\left(-\frac{x-\delta+\sigma^2\,(T-t-s)\,\kappa}{\sigma\,\sqrt{T-t-s}}\right) \\
&\left.+ \,\delta\,\Phi\left(-\frac{x-\delta+\sigma^2\,(T-t-s)\,\kappa}{\sigma\,\sqrt{T-t-s}}, -\frac{x-\delta+\sigma^2\,(T-t-s)\,\kappa}{\sigma\,\sqrt{T-t}}; \sqrt{\frac{T-t-s}{T-t}}\right)\right\}\mathrm{d}s
\end{aligned}
$$

$$
\begin{aligned}
\leq\, & 2\exp\left(\frac{3\,\kappa\,\delta}{2} - \kappa\,x + \frac{1}{2}\,\kappa^2\,\sigma^2\,(T-t)\right)\int_0^{T-t}\sqrt{\frac{\sigma^2\,(T-t-s)}{2\,\pi}} + \left(|x| + \sigma^2\,(T-t-s)\,\kappa + \frac{\delta}{2}\right) + \delta\,\mathrm{d}s \\
&+ 2\exp\left(-\frac{\kappa\,\delta}{2} + \kappa\,x + \frac{1}{2}\,\kappa^2\,\sigma^2\,(T-t)\right)\int_0^{T-t}\sqrt{\frac{\sigma^2\,(T-t-s)}{2\,\pi}} + \left(|x| + \sigma^2\,(T-t-s)\,\kappa + \frac{3\,\delta}{2}\right) + \delta\,\mathrm{d}s
\end{aligned}
$$

$$
\begin{aligned}
=\, & 2\exp\left(\frac{3\,\kappa\,\delta}{2} - \kappa\,x + \frac{1}{2}\,\kappa^2\,\sigma^2\,(T-t)\right)\left(\left(|x| + \frac{3\,\delta}{2}\right)(T-t) + \frac{2}{3}\sqrt{\frac{\sigma^2}{2\,\pi}}\,(T-t)^{\frac{3}{2}} + \frac{\sigma^2\,\kappa}{2}\,(T-t)^2\right) \\
&+ 2\exp\left(-\frac{\kappa\,\delta}{2} + \kappa\,x + \frac{1}{2}\,\kappa^2\,\sigma^2\,(T-t)\right)\left(\left(|x| + \frac{5\,\delta}{2}\right)(T-t) + \frac{2}{3}\sqrt{\frac{\sigma^2}{2\,\pi}}\,(T-t)^{\frac{3}{2}} + \frac{\sigma^2\,\kappa}{2}\,(T-t)^2\right),
\end{aligned}
$$

and we recall that

$$D(q,M) = \min(q+1,M) - \min(q,M)\,.$$

The expectations $\mathbb{E}\left[e^{|X_s|}\right]$ and $\mathbb{E}\left[|X_s|\,e^{|X_s|}\right]$ are bounded because X_s is normal. Hence, the expectation $\mathbb{E}\left[|g_1(s,X_s,N_s^+ + 1) - g_1(s,X_s,N_s^+)|\right]$ is bounded, and we obtain the desired result. ■

A.4. Proof of Proposition 7

Proof.

$$\begin{aligned}\mathbb{E}\left[\widehat{\Pi}_1\right] =& \mathbb{E}\left[M\,\left(-\frac{\delta_1^*}{2}\right)\,\mathbb{1}_{\{\widehat{\tau}=\widehat{\tau}_a\}}\right]\\ =& -\frac{M\,\delta_1^*}{2}\,\mathbb{P}\left(\tau=\widehat{\tau}_a\right)\\ =& -\frac{M\,\delta_1^*}{2}\,\mathbb{P}\left(\max_{0<t<T}\sigma\,W_t > \hat{\delta}(M)/2+\delta_1^*/2\right)\\ =& -M\,\delta_1^*\,\mathbb{P}\left(\sigma\,W_T > \hat{\delta}(M)/2+\delta_1^*/2\right)\\ =& -M\,\delta_1^*\,\left(1-\Phi\left(\frac{\hat{\delta}(M)+\delta_1^*}{2\,\sigma\,\sqrt{T}}\right)\right)\,.\end{aligned}$$

■

A.5. Proof of Proposition 8

Proof. We make the standard change of variables, $y = x-\widetilde{\delta}, \tilde{t} = T-t, \widehat{h}_0(\tilde{t},y) = \widehat{g}_0(t,x)$, so that the PDE

$$\partial_{\tilde{t}}\widehat{h}_0 = \frac{1}{2}\,\sigma^2\,\partial_{yy}\widehat{h}_0\,, \tag{45}$$

and the terminal and boundary conditions become

$$\widehat{h}_0(0,y) = \min(q,\,M)\,\left(-y-\frac{\delta_1^*}{2}\right)\,,\quad \widehat{h}_0(\tilde{t},0)=0\,,\quad y<0\,,\quad \tilde{t}\in[0,T]\,.$$

PDE (45) is the heat equation with heat kernel

$$K(x-y,t) = \frac{1}{\sqrt{2\,\pi\,\sigma^2\,t}}\,e^{-\frac{(x-y)^2}{2\,\sigma^2\,t}}$$

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,

$$
\begin{aligned}
\widehat{h}_0(\tilde{t},y) =& \int_{-\infty}^{\infty} K\left(y-z,\tilde{t}\right)\widehat{u}_0(z)\,\mathrm{d}z \\
=& \min(q,\,M)\,\left(\int_0^{\infty} K\left(y-z,\tilde{t}\right)\left(-z+\frac{\delta_1^*}{2}\right)\,\mathrm{d}z + \int_{-\infty}^{0} K\left(y-z,\tilde{t}\right)\left(-z-\frac{\delta_1^*}{2}\right)\,\mathrm{d}z\right) \\
=& \min(q,\,M)\,\left(-\int_0^{\infty}\frac{z}{\sqrt{2\,\pi\,\sigma^2\,\tilde{t}}}\,e^{-\frac{(y-z)^2}{2\,\sigma^2\,\tilde{t}}}\,\mathrm{d}z + \frac{\delta_1^*}{2}\int_0^{\infty}\frac{1}{\sqrt{2\,\pi\,\sigma^2\,\tilde{t}}}\,e^{-\frac{(y-z)^2}{2\,\sigma^2\,\tilde{t}}}\,\mathrm{d}z\right. \\
&\qquad\left.-\int_{-\infty}^{0}\frac{z}{\sqrt{2\,\pi\,\sigma^2\,\tilde{t}}}\,e^{-\frac{(y-z)^2}{2\,\sigma^2\,\tilde{t}}}\,\mathrm{d}z - \frac{\delta_1^*}{2}\int_{-\infty}^{0}\frac{1}{\sqrt{2\,\pi\,\sigma^2\,\tilde{t}}}\,e^{-\frac{(y-z)^2}{2\,\sigma^2\,\tilde{t}}}\,\mathrm{d}z\right) \\
=& \min(q,\,M)\,\left(-y+\frac{\delta_1^*}{2}-\delta_1^*\,\Phi\left(\frac{-y}{\sigma\,\sqrt{\tilde{t}}}\right)\right)\,,
\end{aligned}
$$

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

A.6. Proof of Proposition 9

Proof. We make the standard change of variables $y = x - \widetilde{\delta}$, $\tilde{t} = T - t$, $\widehat{h}_1(\tilde{t}, y) = \widehat{g}_1(t, x)$. Then

$$
\partial_{\tilde{t}}\widehat{h}_1 = \frac{1}{2}\,\sigma^2\,\partial_{yy}\widehat{h}_1 + \widehat{f}(\tilde{t}, y)\,, \tag{46}
$$

where

$$
\widehat{f}(\tilde{t}, y) = D(q, M)\,e^{\kappa\,\left(\frac{\delta_1^*}{2}+y\right)}\left[-y+\frac{\delta_1^*}{2}-\delta_1^*\,\Phi\left(\frac{-y}{\sigma\,\sqrt{\tilde{t}}}\right)\right]\,,
$$

and

$$
D(q, M) = \min(q+1, M) - \min(q, M)\,.
$$

The terminal and boundary conditions become

$$
\widehat{h}_1(0, y) = 0\,,\quad \widehat{h}_1(\tilde{t}, 0) = 0\,,\quad y < 0\,,\quad \tilde{t}\in[0, T]\,.
$$

We make an odd expansion on $\widehat{f}$ to $\widehat{f}_0$ and apply the heat kernel to obtain:

$$
\begin{aligned}
\widehat{h}_1(\tilde{t}, y) =& \int_0^{\tilde{t}} \int_{-\infty}^{\infty} K(y-z, \tilde{t}-s)\, \widehat{f}_0(s,z)\, \mathrm{d}z\, \mathrm{d}s \\
=& \int_0^{\tilde{t}} \int_0^{\infty} K(y-z, \tilde{t}-s)\, \left(-\widehat{f}(s,-z)\right) \mathrm{d}z\, \mathrm{d}s + \int_0^{\tilde{t}} \int_{-\infty}^{0} K(y-z, \tilde{t}-s)\, \widehat{f}(s,z)\, \mathrm{d}z\, \mathrm{d}s \\
=& 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}{\sigma\sqrt{\tilde{t}}}\right)\right)\right] \mathrm{d}z\, \mathrm{d}s \\
&+ 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}{\sigma\sqrt{\tilde{t}}}\right)\right)\right] \mathrm{d}z\, \mathrm{d}s \\
=& 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}{\sigma\sqrt{\tilde{t}}}\right)\right] \mathrm{d}z\, \mathrm{d}s \\
&+ 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}{\sigma\sqrt{\tilde{t}}}\right)\right] \mathrm{d}z\, \mathrm{d}s \\
=& D(q,M) \exp\left(\frac{\kappa\,\delta_1^*}{2} - \kappa\, y + \frac{1}{2}\,\kappa^2\,\sigma^2\,\tilde{t}\right) \\
&\times \int_0^{\tilde{t}} \exp\left(-\frac{1}{2}\,\sigma^2\,\kappa^2\, s\right) \left\{ -\sqrt{\frac{\sigma^2\,(\tilde{t}-s)}{2\,\pi}} \exp\left(-\frac{(y-\sigma^2\,(\tilde{t}-s)\,\kappa)^2}{2\,\sigma^2\,(\tilde{t}-s)}\right) \right. \\
&- \left(y - \sigma^2\,(\tilde{t}-s)\,\kappa + \frac{\delta_1^*}{2}\right) \Phi\left(\frac{y-\sigma^2\,(\tilde{t}-s)\,\kappa}{\sigma\sqrt{\tilde{t}-s}}\right) \\
&\left. + \;\delta_1^*\, \Phi\left(\frac{y-\sigma^2\,(\tilde{t}-s)\,\kappa}{\sigma\sqrt{\tilde{t}-s}}, \frac{y-\sigma^2\,(\tilde{t}-s)\,\kappa}{\sigma\sqrt{\tilde{t}}}; \sqrt{\frac{\tilde{t}-s}{\tilde{t}}}\right) \right\} \mathrm{d}s \\
&+ D(q,M) \exp\left(\frac{\kappa\,\delta_1^*}{2} + \kappa\, y + \frac{1}{2}\,\kappa^2\,\sigma^2\,\tilde{t}\right) \cdot \\
& \int_0^{\tilde{t}} \exp\left(-\frac{1}{2}\,\sigma^2\,\kappa^2\, s\right) \left\{ \sqrt{\frac{\sigma^2\,(\tilde{t}-s)}{2\,\pi}} \exp\left(-\frac{(y+\sigma^2\,(\tilde{t}-s)\,\kappa)^2}{2\,\sigma^2\,(\tilde{t}-s)}\right) \right. \\
&- \left(y + \sigma^2\,(\tilde{t}-s)\,\kappa - \frac{\delta_1^*}{2}\right) \Phi\left(-\frac{y+\sigma^2\,(\tilde{t}-s)\,\kappa}{\sigma\sqrt{\tilde{t}-s}}\right) \\
&\left. - \;\delta_1^*\, \Phi\left(-\frac{y+\sigma^2\,(\tilde{t}-s)\,\kappa}{\sigma\sqrt{\tilde{t}-s}}, -\frac{y+\sigma^2\,(\tilde{t}-s)\,\kappa}{\sigma\sqrt{\tilde{t}}}; \sqrt{\frac{\tilde{t}-s}{\tilde{t}}}\right) \right\} \mathrm{d}s\,,
\end{aligned}
$$

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