Optimal Strategies of High Frequency Traders

JIANGMIN XU*

Job Market Paper

ABSTRACT

This paper develops a continuous-time model of the optimal strategies of high-frequency traders (HFTs) to rationalize their pinging activities. Pinging, or the most aggressive fleeting orders, is defined as limit orders submitted inside the bid-ask spread that are cancelled shortly thereafter. The current worry is that HFTs utilize their speed advantage to ping inside the spread to manipulate the market. In contrast, the HFT in my model uses pinging to control inventory or to chase short-term price momentum without any learning or manipulative motives. I use historical message data to reconstruct limit order books, and characterize the HFT's optimal strategies under the viscosity solution to my model. Implications on pinging activities from the model are then gauged against data. The result confirms that pinging is not necessarily manipulative and is rationalizable as part of the dynamic trading strategies of HFTs.

A RECENT AND ONGOING heated debate concerns high-frequency traders and high-frequency trading activities (HFT stands for high-frequency trader/trading thereafter). The interest in this subject has grown significantly after the Flash Crash, because HFTs appear as a black-box mystery to the general public as well as to the academic world.[1] One type of HFT activity that has attracted a great deal of attention due to HFTs' speed advantage is the so-called pinging activity. Pinging, or the most aggressive fleeting order activity, is defined as the submission of

*Bendheim Center for Finance, Princeton University. I am deeply indebted to Yacine Aït-Sahalia for his invaluable advice, frequent encouragement and numerous insights. I am also very grateful to Harrison Hong for his many suggestions. I also thank Valentin Haddad, Christopher Sims, David Sraer, Wei Xiong, as well as seminar participants at the Princeton Civitas Foundation finance seminar and the Princeton-QUT-SMU conference for helpful comments.

[1]For a comprehensive review of the Flash Crash, see the study by Kirilenko et. al. (2011). See also Easley et. al. (2011)'s examination on the same subject.

Electronic copy available at: http://ssrn.com/abstract=2382378

limit orders inside the bid-ask spread that get cancelled very shortly.[2] These activities occur in the scale of seconds or milliseconds with extremely low latency, which is "the hallmark of proprietary trading by HFTs" (Hasbrouck and Saar (2013)).[3]

Regulators have expressed concerns over such pinging activities, the main one being that pinging used by HFTs might be manipulative. As pointed out in a concept release of the Securities and Exchange Commission (SEC), HFTs could use pinging orders to detect and learn about hidden orders inside the spread. Hidden orders are limit orders completely non-observable to other market participants. They have become increasingly popular in the past 5 to 10 years and are allowed by many stock exchanges around the world nowadays.[4] This learning would enable HFTs to ascertain the existence of potential large trading interest in the market. Consequently, they would be able to trade ahead and capture a price movement in the direction of the large trading interest.[5]

This paper aims to rationalize the pinging activity levels observable in the data through a theoretical setup without manipulative elements. It develops a continuous-time model of the optimal trading strategies of HFTs absent any learning or strategic feedback effects. The model exploits two well-known forces in the existing literature: inventory control (e.g. Holl and Stoll (1981)) and trend chasing (e.g. Hirschey (2013)).

To achieve my purpose, I first incorporate the existence of hidden orders inside the bid-ask spread into a continuous-time model along the lines of Ho and Stoll (1981). As a result, the HFT in my model would ping for hidden orders inside the spread as a cheaper way to control his inventory compared to using market orders. This would produce pinging orders that execute against hidden orders. However, with inventory control as the sole motive for pinging, it is not possible to obtain a large number of *cancelled* pinging orders at the same time. The reason is that when the HFT uses pinging to control his inventory, he intends to fill his pinging orders and is not incentivized to cancel many of them.

In order to resolve this problem, I then introduce short-term price momentum into the model as a channel for cancelled pinging orders to occur. The momentum is modelled through the predictability of depth imbalance on the direction of price movements. When the HFT sees large

[2]Existing studies (e.g. Hasbrouck and Saar (2009)) have used the term "fleeting orders" to denote submissions and subsequent quick cancellations of limit orders in general. Therefore, to be consistent with the literature, I use the term "pinging" throughout this paper to denote the most aggressive fleeting order activities, i.e. fleeting orders that take place inside the bid-ask spread.

[3]To give a illustration of the speed with which pinging takes place, Hautsch and Huang (2011) find that the median cancellation time is below one second for limit orders submitted inside the spread on NASDAQ.

[4]According to Hautsch and Huang (2012), on average over 20% of trades on NASDAQ are executed against hidden orders in October 2010.

[5]SEC concept release on equity market structure, January 21, 2010.

Electronic copy available at: http://ssrn.com/abstract=2382378

depth imbalance and anticipates a likely directional price move, he could use pinging orders as directional bets to chase the price momentum. Moreover, if there is a subsequent large change in depth imbalance, the HFT would cancel his pinging orders and adjust his strategy according to the variation in momentum. Therefore, the model will now give rise to pinging as well as cancellation due to the HFT's momentum-chasing behaviors.

The model is solved numerically due to its complexity, and I use historical order book message feed data from NASDAQ to reconstruct limit order books and estimate model parameters.[6] The optimal HFT strategies are characterized based on the viscosity solution to my model and the parameter estimates. There are two main findings. First, for stocks whose order books have high depths with relatively stable movements and whose spreads tend to be narrow, approximately 20% of the HFT's optimal strategies are attributable to pinging. On the other hand, for stocks that tend to have low order-book depths, volatile order-book movements and wide spreads, pinging accounts for nearly 50% of the HFT's optimal strategies. These pinging percentages from the model are proven to match most of the observable pinging activity levels from the data. Thus it implies that the pinging activities occurred in reality can be mostly rationalized by my model without any learning or manipulative components. Second, I demonstrate that the mechanism of momentum chasing plays a much more important role than that of inventory control in rationalizing pinging activities for low-depth and wide-spread stocks. In contrast, the two mechanisms carry around similar weights in rationalizing pinging activities for high-depth and narrow-spread stocks. Furthermore, both mechanisms are shown to be necessary for the model to rationalize the pinging occurrences found in the data.

Beyond pinging rationalization, my model also generates a couple of additional auxiliary predictions. They concern about the directions of pinging activities and the frequencies of cancelled pinging activities with regard to the depth imbalance of order books. These predictions are also found to be largely consistent with the data, which further suggests that pinging is rationalizable as part of the dynamic trading strategies of HFTs.

In what follows, I build a continuous-time, partial equilibrium model that captures a wide range of HFT strategies and explore their empirical content and implications. I begin in Section I by discussing the relations between my model and the existing literature. Section II lays out the structure of my model. The model's equilibrium and my numerical solution are presented in Section III. Section IV is devoted to parameter estimations. I then discuss the main findings of this paper in Section V, and draw out the model's auxiliary predictions and evaluate them on empirical data. Finally, Section VI concludes.

[6]I am grateful to NASDAQ for providing me with the access to their historical TotalView-ITCH real-time limit order book message feed database.

Electronic copy available at: http://ssrn.com/abstract=2382378

I. Related Literature

There is a long line of empirical studies (see, e.g., Brogaard et. al. (2013a), Brogaard et. al. (2013b), Hagströmer and Nordén (2013), Hasbrouck and Saar (2013), Hendershott and Riordan (2013), Hendershott et. al. (2011) and Menkveld (2013)) that have tried to understand the effects of high-frequency and algorithmic trading activities on market quality. This line of research is economically important as HFT firms account for over 50% of all US equity trading volume in 2012.[7]

On one hand, the majority of the empirical studies show that, on balance, HFTs are beneficial for market quality. For instance, Brogaard et. al. (2013b) find that HFTs enhance price discovery and market efficiency on NASDAQ, with prices reflecting information more quickly. Hasbrouck and Saar (2013) also show that increased HFTs' low-latency activities are associated with lower posted and effective spreads, lower short-term volatility and increased market depth. Additionally, Hendershott et. al. (2011) illustrate that algorithmic tradings improve liquidity and make quotes more informative for NYSE stocks.

On the other hand, there are empirical analyses demonstrating that some of the HFT strategies are speculative/anticipatory in nature. For example, Hirschey (2013) finds that HFTs on NASDAQ tend to anticipate future order flow and trade ahead of it, through aggressively taking liquidity from the market. Moreover, Kirilenko et. al. (2011) show that HFTs exacerbated market volatility during the Flash Crash by trading in the direction of the downward price spiral. In addition, Baron et. al. (2012) find that aggressive, liquidity-taking HFTs earn short term profits at the expense of other market participants.

This paper focuses on the optimal trading strategies of HFTs and their pinging activities in particular. It builds on the work of Holl and Stoll (1981) – HS81, Avellaneda and Stoikov (2008) – AS08, and Guilbaud and Pham (2013) – GP13. HS81 first introduced, in a continuous-time partial equilibrium model, a market maker who optimally chooses the bid and ask prices of his limit orders to maximize his expected utility of (terminal) wealth. The market maker uses limit orders only, so that inventory can only be managed through limit orders. AS08 recast HS81 in a modern HFT setting by introducing the trading environment for HFTs not present in HS81 – the limit order book. However, an HFT in AS08 is still a pure market maker, since he uses only limit orders to maximize his utility and control his inventory.

GP13 brought in market orders to the framework of HS81 and AS08. In their model, an HFT can trade via both limit orders and market orders. The HFT can post limit orders at the best quotes or improve the quotes by one tick. Otherwise he can use market orders instead

[7] Times Topics: High-Frequency Trading, The New York Times, December 20, 2012.

to change his inventory instantaneously. The purpose of introducing market orders as another control mechanism is to better address the execution risk and the inventory risk faced by HFTs when they use limit orders only.

My model extends the continuous-time configuration of GP13 for HFTs and makes two contributions. Firstly, hidden orders are introduced into the limit order book, which are not present in either AS08 or GP13. Hence the HFT can ping for hidden liquidity when his limit orders improve the best bid/ask prices and are submitted inside the spread. This captures the idea of pinging strategies as identified in Hasbrouck and Saar (2009)'s empirical study. Secondly, I model the order book's depth imbalance at the best quotes as a stochastic process that has an effect on the movement of the mid-prices and on the existence of hidden orders. Thus the HFT can use the depth imbalance as a (imperfect) signal to anticipate the likely price movement. Therefore, the HFT will utilize pinging or market orders to conduct directional/momentum tradings when necessary. This captures the idea of anticipatory strategies as identified in Hirschey (2013), which is not modelled explicitly in GP13.

Due to market orders being impulse (jump) controls in continuous time, the value function of the HFT in my model is not necessarily smooth and differentiable everywhere. Consequently, I will apply the viscosity solution technique to my model, which is similar to the one used in GP13.[8] The viscosity solution concept, originally introduced by Crandall and Lions (1983), is a generalization of the classical solution concept to Hamilton-Jacobi-Bellman equations (or partial differential equations in general) where value functions might not be everywhere differentiable. I will discuss the setup of my model and its solution in detail in the next two sections.

II. The Model

The economy is defined in a continuous time, finite horizon $\mathcal{T} = [0, T]$, with a single risky stock that can be traded in a limit order book (LOB). There is a high-frequency trader (HFT) who trades this stock using either limit orders or market orders. The LOB has a tick size δ, so that prices of all orders are in multiples of δ. Characteristics of the LOB, such as spread and mid-price, evolve stochastically and are exogenous to the HFT.[9] The formalization of my model is further specified as follows.

[8] HS81 and AS08 consider only limit orders that are regular (continuous) controls in continuous time. Therefore, the value functions of HFTs in their models are smooth and differentiable at every point, so that they have standard Hamilton-Jacobi-Bellman equations and classical solution techniques apply.

[9] A limit order of size q at price p is an order to buy or sell q units of the asset at the specified price p; its execution occurs only when it meets a counterpart market order. A market order of size q is an order to buy or sell q units of the asset being traded at the lowest (for buy) or the highest (for sell) available price in the market; its execution is immediate. Given an asset, the best bid (resp. ask) price is the highest (resp. lowest)

A. Processes of Limit Order Book Characteristics

To introduce the main characteristics of the LOB, I will fix a filtered probability space $(\Omega, \mathcal{O}, \mathbb{O}, \mathbb{P})$, $\mathbb{O} = (\mathcal{O}_t)_{t\geq 0}$ satisfying the usual conditions. Therefore, all stochastic processes and random variables are defined on $(\Omega, \mathcal{O}, \mathbb{O}, \mathbb{P})$.

To start with, the depth imbalance F at the best quotes is defined to be the difference of the log of the size of the depth at the best bid price from that at the best ask price, so that $F > (<)0$ means bid (ask) side imbalance. It follows an Ornstein-Uhlenbeck process with mean zero:

$$dF_t = -\alpha_F \, F_t \, dt + \sigma_F \, dW_t \,, \tag{2.1}$$

where α_F measures the speed of mean-reversion and σ_F is a constant volatility parameter.

Next, let

$$S = \{S_t\}_{t\geq 0} \tag{2.2}$$

denote the bid-ask spread of the LOB. It follows a continuous time Markov chain on the state space $\mathbb{S} = \{\delta, 2\delta, 3\delta\}$, with a constant jump intensity λ^S representing the times when orders of participants in the market affect the spread. The probability transition matrix of S is denoted by $\rho = (\rho_{ij})_{1\leq i,j\leq 3}$, with $\rho_{ii} = 0$.

Furthermore, the mid-price P of the stock is assumed to evolve according to a pure jump process:

$$dP_t = dJ_{1t} + dJ_{2t} \,. \tag{2.3}$$

The first component, J_{1t}, has a constant jump intensity λ_1^J, and jump sizes equal to $\delta/2$ with probability $\psi_1(F_t)$ and $-\delta/2$ with probability $1 - \psi_1(F_t)$; while the second component, J_{2t}, has a constant jump intensity λ_2^J, and jump sizes equal to δ with probability $\psi_2(F_t)$ and $-\delta$ with probability $1 - \psi_2(F_t)$, where the functions $\psi_i : \mathbb{R} \mapsto [0, 1]$ are assumed to have the form

$$\psi_i(u) = 1/(1 + \exp(-\beta_i \, u)), \quad \text{for } i = 1, 2 \,,$$

with β_i being positive constants. In addition, J_1 and J_2 are independent.[10] I allow the depth imbalance F to have an impact on the directions of mid-price jumps, since HFTs mostly seek information from the LOB itself to forecast price movements in the millisecond environment.

price among all active buy (resp. to sell) limit orders in the order book. The spread is the difference between the best ask price and the best bid price, which is strictly positive. The mid-price is the mid-point between the best bid and the best ask price. For a more detailed explanation of limit order book variables, please refer to the non-technical survey of Gould et. al. (2010).

[10] The best bid and best ask prices are thus $P_t - S_t/2$ and $P_t + S_t/2$ respectively.

Depth imbalances capture liquidity pressures within the LOB. Therefore, it is informative about at which side of the book the observable depth is likely to become depleted first, resulting in a change in the mid-price. Hence it is a natural signal (albeit imperfect) for the HFTs to infer the direction of future price movements, based on the current state of the LOB.

The influence of the depth imbalance F on the directions of the mid-price jumps is interpreted as the short-term price momentum in the model. If the parameter σ_F is large so that F is volatile, it leads to more frequent appearances of stronger momentum (signals). Thus the HFT would be more likely to engage in trend-chasing actions. This effect is amplified if the jump intensities λ_1^J and λ_2^J of the mid-price are higher, i.e. there are more realizations of price momentum.

Besides these standard LOB characteristic variables, orders with limited pre-trade transparency have become increasingly popular on electronic trading platforms recently. Major US stock exchanges, such as NASDAQ, NYSE and BATS, permit submissions of hidden limit orders into their LOBs.[11] The price, size and location of these orders are completely concealed from other market participants, and they can be placed inside of the observable bid-ask spread without affecting the visible best bid and ask quotes. This interesting mechanism of hidden orders has created a vast number of order activities in markets as HFTs try to "ping" for hidden liquidity inside of the spread by posting aggressive "fleeting orders" that are cancelled a few instants later if not executed.[12]

Consequently, to account for this important phenomenon, I will specify the existence of hidden orders in the following fashion. If the spread S_t equals 2δ, the probability of hidden bid orders sitting at the mid-price P_t is $\varphi_1(F_t)$ and the probability of hidden ask orders is $\pi_1 - \varphi_1(F_t)$. Similarly, if $S_t = 3\delta$, the probabilities of hidden bid orders at $P_t - \delta/2$ and $P + \delta/2$ are $\varphi_1(F_t)$ and $\varphi_2(F_t)$ respectively, so that those of hidden ask orders at $P_t - \delta/2$ and $P_t + \delta/2$ are $\pi_1 - \varphi_1(F_t)$ and $\pi_2 - \varphi_2(F_t)$. Since existing literature suggests that the probability of hidden orders inside the spread is positively correlated with the own-side depth imbalance,[13] I let the functions φ_i take the form of

$$\varphi_i(u) = \pi_i/(1 + \exp(-\kappa\, u)), \quad \text{for } i = 1, 2\,,$$

where κ and π_i are positive constant and $\pi_i < 1$. When π_1 and π_2 are large, hidden orders are

[11] Hautsch and Huang (2012) give a detailed empirical analysis of hidden orders on Nasdaq stocks. They find that, during the month of October 2010, on average, 20.1% of all trades are executed against hidden orders. However, only a small proportion of the hidden depth get executed, which implies that the share of undetected hidden depth is much greater.

[12] This pinging phenomenon was first documented in Hasbrouck and Saar (2009).

[13] Buti and Rindi (2013) provide a theoretical model that gives rise to this positive correlation, and Hautsch and Huang (2012) offer an empirical confirmation.

more likely to be found within the spread. Thus the HFT would have more opportunities to take advantage of these hidden orders through the use of pinging orders.

B. The HFT's Trading Strategy

At any time t, the HFT in my model can submit limit buy and sell orders specifying the prices that he is willing to pay and receive, but they will be executed only when incoming market orders fill his limit orders. The quantity of the HFT's limit orders is fixed at one lot (100 shares).[14] On the other hand, instead of limit orders, the HFT can send out market buy or sell orders for immediate execution. The market orders will cross the spread, i.e. trading at the best ask (resp. best bid) price on the opposite side, and are thus less price-favorable.

B.1. Make Strategy.

When using limit orders to make the market, HFT can place his quotes at the best available bid and ask, hence joining the existing queues at these prices.[15] This strategy simply amounts to traditional market making, which means that the HFT tries to passively capture the spread by posting limit orders at the best available quotes.

Furthermore, the HFT can also "ping" inside the spread by improving either the best bid or ask price by one tick whenever the spread S_t is greater than δ, i.e. submitting a buy (sell) limit order at $P_t - S_t/2 + \delta$ ($P_t + S_t/2 - \delta$). Such a pinging strategy of putting limit orders inside the spread is commonly used by HFTs in practice to capture the market order flow at the best quotes, because of the price-time priority associated with these limit orders. More importantly, they are used to ping for hidden orders inside the spread, as identified by Hasbrouck and Saar (2009). Nevertheless, despite being able to obtain faster executions, these pinging limit orders do receive worse execution prices than their queuing counterpart. Hence this is a trade-off faced by the HFT when he contemplates whether to improve the spread or not.

Because the submission, update or cancellation of limit orders entails no cost, it is natural to model the make strategy of the HFT as a continuous-time predictable control process:

$$\theta_t^{mk} = \{\theta_t^{mk,b}, \theta_t^{mk,a}\}, t \geq 0,$$

[14]The reason that I fix the quantity of limit orders to be one lot is motivated by a common finding that appears in a vast number of empirical studies, whether they examine HFT activities (e.g. Hasbrouck and Saar (2013)) or analyze limit order books (e.g. Hautsch and Huang (2012)). The finding is that the average size of limit orders nowadays is just slightly bigger than 100 shares.

[15]I define market making to be the strategy such that the HFT submit limit orders simultaneously at both sides of the LOB. Quoting at one side of the book only is not allowed, since it does not satisfy the standard definition of market making.

where $\theta_t^{mk,b} \in \{0,1\}$ and $\theta_t^{mk,a} \in \{0,1\}$. b, a stand for the bid and the ask side respectively. The predictable processes $\theta_t^{mk,b}$ and $\theta_t^{mk,b}$, with values equal to 0 or 1, represent the possible make regimes: 0 indicates that the limit order is joining the queue at the best price, whereas 1 indicates improving the best price by δ. Note that if the spread is at its minimum δ, $\theta_t^{mk,b}$ and $\theta_t^{mk,b}$ both can only take the value of 0, since improving the best bid/ask will simply be considered as posting a market buy/sell order instead.

B.2. Take Strategy.

Instead of limit orders, the HFT may also employ market orders (take strategy) to obtain instant execution. However, unlike limit orders, market orders take liquidity from the LOB and are subject to transaction fees. As a result, the costly nature of market orders implies that, if the take strategy is performed continuously, the HFT will go bankrupt in finite time. Hence I shall model the HFT's take strategy as an impulse control in continuous time:

$$\theta^{tk} = \{\tau_n,\, \zeta_n\}_{n\in\mathbb{N}}\,.$$

Here, $\{\tau_n\}$ is an increasing sequence of stopping times denoting the moments that the HFT uses market orders. $\zeta_n \in [-\zeta_{max}, \zeta_{max}]\backslash\{0\}$ are $\mathcal{F}_{\tau_n}$-measurable random variables that represent the number of shares (in lot size) purchased if $\zeta_n > 0$ or sold if $\zeta_n < 0$, at these stopping times. I confine the size of market orders to be less than or equal to some small constant ζ_{max}, so that the HFT remains small relative to the market and his market orders do not eat through the LOB.[16]

C. Order Execution Processes

Limit orders of the HFT, when joining the queue at the best bid and ask, will be executed only if counterpart market orders arrive in the next instant and their sizes are large enough to fill the HFT's limit orders completely. I model the arrivals of exogenous buy and sell market orders by two independent Poisson processes, M^b and M^a, with intensities given by λ^b and λ^a respectively. In addition, when a buy (resp. sell) market order arrives, the HFT's sell (resp. buy) limit order in the queue will be filled with a probability given by the fill-rate function $h(F_t)$ (resp. $h(-F_t)$), where

[16]Eating through the book means the size of an market order is larger than and thus exhausts the available depth at the best price(s), resulting in a price impact and a widening of the spread.

$$h(u) = 1/(1 + \exp(\varsigma_0 + \varsigma_1 u)),\ \varsigma_0,\ \varsigma_1 \text{ are positive constants.}$$

This fill-rate function depicts the idea that the probability of execution of a limit order becomes higher/lower if the own-side queue is relatively shorter/longer, which illustrates the time priority structure of the LOB.

However, if an limit order jumps the queue instead, it will receive an instantaneous execution as long as there is a matching hidden order on the opposite side that resides inside the spread. Otherwise, the limit order will be fully matched if a counterpart market order arrives at the next moment. Limit orders placed inside the spread by the HFT are not subject to the fill rate function, since they have price priority compared to those limit orders at the prevailing best quotes.

Turning to the HFT's market orders, they will execute immediately by hitting either inside-spread hidden orders or limit orders at the best quote on the opposite side. Moreover, there is a per share fee of ϵ associated with each market order that is sent by the HFT. The fee is fixed and paid directly to the market exchange operating the LOB. Thus the HFT cannot claim it back by any means.

D. Comments on the Model

The basic layout of the model resembles Guilband and Pham (2013) and is easy to justify given the goals of the paper. However, there are two important aspects of the model that are new compared to the current literature on high-frequency trading strategies and are essential for rationalizing HFTs' pinging activities without manipulation or learning. First, the opportunity of using limit orders to grasp hidden liquidity has not been properly addressed in existing continuous-time models of HFTs, where the limit order strategy only fulfills the role of the HFT as a traditional market maker or liquidity provider. In my model, limit orders can also be utilized by the HFT as a "pinging" strategy, because of the existence of inside-spread hidden orders. Compared to conventional market making, pinging is certainly more aggressive as it seeks a faster execution at a less favorable price. However, it is not as aggressive as market taking since it does not cross the spread. Given the large number of pinging activities already identified in empirical work, for example, Hasbrouck and Saar (2009, 2013), it is essential to include hidden orders and pinging strategies in a theoretical model of HFT behaviors.

Another aspect that has not been fully dealt with in the continuous-time literature concerns the prospect of directional or momentum tradings by the HFT. In a standard setup, the HFT would only use market orders to reduce its inventory whenever it gets out of control. However, in

my model, the HFT would also implement market orders as a momentum strategy to aggressively take liquidity from the market and pursue directional trading when the signal of depth imbalance on price movement is strong enough. There are empirical studies that have identified some most profitable HFTs as predominantly liquidity takers on the market (e.g. Baron et. al. (2012)) and have recognized the anticipatory/directional trading behaviors of HFTs (e.g. Hirschey (2013)). Therefore, it is crucial to incorporate momentum/directional trading into any theoretical models on HFT strategies.[17]

When both hidden orders and price momentum exist in the model, the HFT has two motives to carry out pinging strategies. One is to control inventory. The HFT could ping inside the spread to hit hidden orders when he needs to unwind his inventory, which is cheaper than using market orders. Nonetheless, if hidden orders existed with high probabilities π_1 and π_2 and inventory control was the HFT's only motive for pinging, the model would generate a large number of pinging orders filled by hidden orders. There would not be many cancelled pinging orders at the same time, contradicting with the numerous ones observed in reality.

The existence of price momentum, i.e. the predictability of depth imbalance on mid-price jumps, are then necessary to bring about cancelled pinging activities. This is because now the other motive for the HFT to ping is to chase price trends. The HFT could employ pinging orders to establish directional positions when anticipating likely price movements. However, if the volatility parameter σ_F is large so that depth imbalance (momentum signal) varies turbulently, it would induce the HFT to cancel his pinging orders frequently and change his strategies based on the swings in momentum. This is escalated if the mid-price jump intensities λ_1^J and λ_2^J are also large. Large intensities prompt more trend-chasing pinging activities from the HFT since his directional bets have higher chances to materialize, yet the HFT often needs to cancel his pinging orders due to volatile changes in price momentum.

As a consequence of the possibility of pinging and momentum trading, the HFT in my model should not be considered exclusively as a turbo-charged market maker in the traditional sense, whereby it provides liquidity to the market with passive limit orders on the one side, and on the other side, liquidity-taking through market orders is only necessitated by inventory-control requirement.

[17]Since market orders are costly, if they are used solely to control inventory when it is really necessary, it would be difficult to reconcile this with the findings that market taking can be (very) profitable.

III. Equilibrium and Solution Method

A. The Objective of the HFT

In order to derive an equilibrium of my model, I begin by stating the cash holding and the inventory processes of the HFT, and defining his optimization problem under stochastic evolutions of the LOB as laid out in the previous section.

Let X and Y denote the cash holdings and the inventory held by the HFT respectively. If a market making strategy θ_t^{mk} is used at t, the cash holding X and the inventory Y evolve according to:

$$dY_t = d\widetilde{M}_t^a - d\widetilde{M}_t^b \tag{3.1}$$

$$dX_t = -\Big(P_t - \frac{S_t}{2} + \delta\theta_t^{mk,b}\Big)d\widetilde{M}_t^a + \Big(P_t + \frac{S_t}{2} - \delta\theta_t^{mk,a}\Big)d\widetilde{M}_t^b\,, \tag{3.2}$$

where

$$\begin{aligned} d\widetilde{M}_t^a &= \theta_t^{mk,b}\Big(B_t^a dt + (1-B_t^a)dM_t^a\Big) + (1-\theta_t^{mk,b})h(F_t)dM_t^a \\ d\widetilde{M}_t^b &= \theta_t^{mk,a}\Big(B_t^b dt + (1-B_t^b)dM_t^b\Big) + (1-\theta_t^{mk,a})h(-F_t)dM_t^b\,. \end{aligned}$$

Here, B_t^a (resp. B_t^b) is an indicator that equals one if ask (resp. bid) hidden orders exist inside the spread. B_t^a and B_t^b have respective distributions $\eta^a(F_t, S_t)$ and $\eta^b(F_t, S_t)$. η^a, η^b match the existence probabilities of ask/bid hidden orders that sit at the best bid/ask price plus δ, given F and S. Therefore, when the HFT's buy (resp. sell) limit order makes the market at the prevailing best bid (resp. ask), its inventory increases (resp. decreases) by one lot if sell (resp. buy) market orders arrive at the next instant *and fill* the limit order of the HFT. Alternatively, if the HFT's buy (resp. sell) limit order is pinging inside the spread, its inventory increases (resp. decreases) by one lot whenever the limit order hits an opposite-side hidden order. Otherwise, it rises (resp. falls) by one lot if there is an arrival of sell (resp. buy) market orders within the next instant. Cash holdings thus increases (resp. decreases) by an amount equal to the quoted price of the sell (resp. buy) limit order multiplied by the order's uncertain execution state.

On the other hand, the dynamics of X and Y jump at t if the HFT exercises a take strategy θ_t^{tk} instead:

$$Y_t = Y_{t-} + \zeta_t\,, \tag{3.3}$$

$$X_t = X_{t-} - \Big[\,\zeta_t P_t + |\zeta_t|\Big(\frac{S_t}{2} - H_t + \epsilon\Big)\,\Big] \tag{3.4}$$

where ϵ is the fixed fee per share paid to the market exchange, H_t is an integer random variable with a probability mass function $G(\cdot \,|\, S_t, F_t)$ that takes the value

$$H_t = \begin{cases} 0 & \text{if } S_t = \delta \\ \delta & \text{if } S_t > \delta \text{ and the market order hits a hidden order at } P_t + sign(\zeta_t)(S_t/2 - \delta) \\ 2\delta & \text{if } S_t > 2\delta \text{ and the market order hits a hidden order at } P_t + sign(\zeta_t)(S_t/2 - 2\delta)\,, \end{cases}$$

and the probability distribution of H_t matches the existence probabilities of inside-spread hidden orders given S_t and F_t. As a result, when the HFT submits a market order, its inventory jumps up or down at t by the size of the order (since take strategies are impluse control). Moreover, its cash holdings changes by the value of the order $\zeta_t P_t$ plus the cost associated with the market order. The cost consists of the uncertain part due to crossing the spread $(S_t/2 - H_t)$ and the constant, fixed transaction fee ϵ.[18]

Given the processes of cash holdings X and inventory Y, the objective of the HFT is the following. He wants to maximize over the finite horizon $[0, T]$ the profit (cash earnings) from his trades in the LOB, while at the same time keeping his inventory at bay. In addition, the HFT has to liquidate all his inventory at the terminal date T. Hence the HFT's optimization problem is given by

$$\max_{\{\theta^{mk},\, \theta^{tk}\}} \; \mathbb{E}_0 \left[X_T - \gamma \int_0^T Y_{t-}^2 d[P,P]_t \right], \quad s.t. \;\; Y_T = 0, \tag{3.5}$$

where the maximization is taken over all admissible strategies Θ. The integral term $\gamma \int_0^T Y_{t-}^2 d[P,P]_t$ is a quadratic-variation penalization term for holding a nonzero inventory in the risky stock, where $\gamma > 0$ is a penalization parameter and $[P,P]_t$ denotes the quadratic variation of the mid-price P.

Let me rewrite the above optimization problem in a more straightforward formulation where the terminal constraint $Y_T = 0$ is removed. To this end, I introduce the function

$$Q(x, y, p, f, s) = x + py - |y| \Big(\frac{s}{2} - H + \epsilon \Big)\,,$$

which represents the total cash obtained after an immediate liquidation of the inventory y via a market order, given the cash holdings x, the mid-price p, the depth imbalance f and the spread

[18] In essence, the term $-H_t$ measures the reduction in the cost of market orders for the HFT given the possible existence of hidden orders inside the spread when $S_t > \delta$, i.e. it does not necessarily pay the full spread-crossing cost of $S_t/2$ (relative to the mid-price).

s.[19] I can now reformulate the problem (3.5) equivalently as

$$\max_{\{\theta^{mk},\,\theta^{tk}\}} \mathbb{E}_0\left[X_T + P_T Y_T - |Y_T|\Big(\frac{S_T}{2} - H_T + \epsilon\Big) - \gamma\int_0^T Y_{t-}^2 d[P,P]_t\right]. \tag{3.6}$$

The proof for the equivalence of the two formulations is shown in the appendix.

Lemma 1. (3.5) *and* (3.6) *are equivalent.*

Having defined the objective, the value function of problem (3.6) for the HFT is then:

$$V(t,x,y,p,f,s) = \sup_{\{\theta^{mk},\,\theta^{tk}\}\in\Theta} \mathbb{E}_t\left[X_T + P_T Y_T - |Y_T|\Big(\frac{S_T}{2} - H_T + \epsilon\Big) - \gamma\int_t^T Y_{u-}^2 d[P,P]_u\right], \tag{3.7}$$

for $t \in [0,T]$, $(x,y,p,f,s) \in \mathbb{R}^2 \times \mathcal{P} \times \mathbb{R} \times \mathbb{S}$. Here, given $\{\theta^{mk},\, \theta^{tk}\} \in \Theta$, $\mathbb{E}_t$ stands for the expectation operator under which the solution (X,Y,P,F,S) to the processes (2.1)-(2.3) and (3.1)-(3.4), with initial state $(X_{t-},Y_{t-},P_{t-},F_{t-},S_{t-}) = (x,y,p,f,s)$, is taken.

Before giving a definition of the equilibrium, I states the following lemma (with its proof shown in the appendix), which provides some bounds on the value function (3.7) and demonstrates that the value function is finite and locally bounded.

Lemma 2. *There exists some constant C_0 and C_1 such that for all $(t,x,y,p,f,s) \in [0,T] \times \mathbb{R}^2 \times \mathcal{P} \times \mathbb{R} \times \mathbb{S}$,*

$$Q(x,y,p,f,s) \le V(t,x,y,p,f,s) \le x + py + C_1 + C_0\,.$$

Both of the lower and the upper bound have a intuitive financial interpretation. The lower bound indicates the value of the particular strategy that eliminates all the current non-zero inventory through a market order, and then waits by doing nothing until the time reaches T. The upper bound is made of three terms. The first term, $x + py$, is the marked-to-market value of the portfolio at mid-price; the second constant C_1 denotes the upper boundary on profit from the fictitious market-making strategy that participates in every trade but with zero cost of controlling inventory; and the constant C_0 at last represents a bound on profit for any directional frictionless market-taking strategy on a virtual asset that is always priced at the mid-price. This lemma is useful later on when I derive the solution to the model's equilibrium.

[19] H is the same integer random variable defined on the previous page.

B. Definition of Equilibrium and Dynamic Programming Equations

The problem of (3.7) is a mixed regular/impulse stochastic control problem in a jump-diffusion continuous time model, to which the methods of dynamic programming naturally lends itself. In order to characterize the equilibrium and the associated dynamic programming equations, I need to introduce two mathematical operators as follows. For any admissible strategy $\theta^{mk} = \{\theta^{mk,b}, \theta^{mk,a}\}$, I will define the second-order non-local operator $\mathcal{L}$:

$$\begin{aligned}\mathcal{L} \circ V(t,x,y,p,f,s) = (\mathcal{L}^P + \mathcal{L}^F + \mathcal{L}^S) \circ V(t,x,y,p,f,s) \\ + g^a(f,s,\theta_t^{mk,b}) \cdot V\big(t,\, x - (p - s/2 + \delta\theta_t^{mk,b}),\, y+1,\, p,\, f,\, s\big) \\ + g^b(f,s,\theta_t^{mk,a}) \cdot V\big(t,\, x + (p + s/2 - \delta\theta_t^{mk,a}),\, y-1,\, p,\, f,\, s\big)\,,\end{aligned}$$

where $\mathcal{L}^P$, $\mathcal{L}^F$, $\mathcal{L}^S$ in the first term are the infinitesimal generators of the processes of the mid-price P, the depth imbalance F and the spread S respectively, and the next two terms denote the non-local operator induced by the (expected) jumps of the cash process X and inventory process Y when the HFT applies an instantaneous make strategy θ_t^{mk} at date t. In addition,

$$\begin{aligned}g^a(f,s,\theta_t^{mk,b}) &= \theta_t^{mk,b}\big(\eta^a(f,s) + (1-\eta^a(f,s))\lambda^a\big) + (1-\theta_t^{mk,b})\lambda^a h(f) \\ g^b(f,s,\theta_t^{mk,a}) &= \theta_t^{mk,a}\big(\eta^b(f,s) + (1-\eta^b(f,s))\lambda^b\big) + (1-\theta_t^{mk,a})\lambda^b h(-f),\end{aligned}$$

which correspond to the expected rate of execution for the HFT's bid and ask limit order respectively.[20]

Besides the operator $\mathcal{L}$, the impulse control operator $\mathcal{M}$ for an admissible take strategy θ^{tk} shall be given by

$$\mathcal{M} \circ V(t,x,y,p,f,s) = \sup_{\zeta \in [-\zeta_{max}, \zeta_{max}]} \int_H V\big(t,\, x - \zeta p - |\zeta|(s/2 - H + \epsilon),\, y+\zeta,\, p,\, f,\, s\big)\, dG(H \,|\, s, f)\,,$$

generated by the jumps in X and Y due to θ_t^{tk} being used at t.

With $\mathcal{L}$ and $\mathcal{M}$ defined, the dynamic programming equation associated with the value function (3.7) is the Hamilton-Jacobi-Bellman quasi-variational inequality (HJB-QVI):

$$\max\left\{\frac{\partial V}{\partial t} + \sup_{\{\theta^{mk}\}}\{\mathcal{L} \circ V\} - \gamma y^2 \frac{\mathbb{E}_t d[P,P]_t}{dt}\,,\;\; \mathcal{M} \circ V - V\right\} = 0,\ \text{on } [0,T) \tag{3.8}$$

[20] The two terms, g^a and g^b, are basically the arrival rates of the "modified" market orders $\widetilde{M}_t^a$ and $\widetilde{M}_t^b$ that have explained under (3.2).

together with the terminal condition

$$V(T,x,y,p,f,s) = x + py - |y|(s/2+\epsilon) + |y| \int_H H\, dG(H \mid s,f)\ . \tag{3.9}$$

There is an explicit expression for the HJB-QVI (3.8) shown in the appendix. In particular, it writes out the full expressions for the infinitesimal generators ($\mathcal{L}^P$, $\mathcal{L}^F$, $\mathcal{L}^S$) and the value of $\mathbb{E}_t[P,P]_t$.

Lemma 3. *The HJB-QVI* (3.8) *admits an explicit expression.*

Having (3.8) and (3.9) at hand, the equilibrium concept of my model is presented in the following definition.

Definition 1. *In the above continuous-time economy where the HFT trades a risky stock in a LOB that is governed by the stochastic processes laid out in Section 2.A-2.C, the partial equilibrium is defined by a value function $v(t;\bullet)$ and policy functions (strategies) $\{\theta^{mk}, \theta^{tk}\}_t$, $t \in [0,T]$, such that*

(a) The policies solve the HFT's maximization problem (3.6)*;*

(b) Given the policy functions, the value function v solves the HJB-QVI (3.8) *for $t \in [0,T)$ with the terminal condition* (3.9) *at T.*

The next proposition demonstrates that a solution to the partial equilibrium exists and is unique. The proof is relegated to the appendix.

Proposition 1. *There is a unique solution to the partial equilibrium of the model. In particular, the value function defined in* (3.7) *is the unique* ***viscosity solution*** *to* (3.8) *and* (3.9).[21]

I shall devote the next subsection to providing a numerical solution to the equilibrium of my model, i.e. the value function (3.7) that solves (3.8) and (3.9) and the corresponding optimal trading strategies of the HFT. However, prior to proceeding further, I present a lemma below that simplifies the value function (3.7) and reduces its dimensionality.

Lemma 4. *(Proof in the appendix) The value function* (3.7) *can be decomposed as $V(t,x,y,p,f,s) = x+py+\nu(t,y,f,s)$. Moreover, the reduced-form value function ν satisfies the*

[21]The viscosity solution concept is a generalization of the classical solution concept to a partial differential equation. For a classic reference on viscosity solutions, see Crandall et. al. (1992) or Fleming and Soner (2005).

quasi-variational inequality and the terminal condition shown below, which are simplified from (3.8)-(3.9) *after decomposing* V*:*

$$\max\Bigg[\frac{\partial\nu}{\partial t}+y\frac{\mathbb{E}_t dP_t}{dt}+\mathcal{L}^{\mathcal{F}}\circ\nu+\mathcal{L}^{\mathcal{S}}\circ\nu-\gamma y^2\frac{\mathbb{E}_t d[P,P]_t}{dt}+$$
$$\sup_{\theta^{mk}}\Big\{g^a(f,s,\theta_t^{mk,b})\cdot\big(\nu(t,y+1,f,s)-\nu(t,y,f,s)+s/2-\delta\theta_t^{mk,b}\big)+$$
$$g^b(f,s,\theta_t^{mk,a})\cdot\big(\nu(t,y-1,f,s)-\nu(t,y,f,s)+s/2-\delta\theta_t^{mk,a}\big)\Big\},$$
$$\sup_{\zeta}\Big\{\nu(t,y+\zeta,f,s)-|\zeta|(s/2+\epsilon)+|y|\int_H H\,dG(H\mid s,f)\Big\}-\nu\Bigg]=0,\quad on\ [0,T) \tag{3.10}$$

with terminal condition:

$$\nu(T,y,f,s)=-|y|(s/2+\epsilon)+|y|\int_H H\,dG(H\mid s,f)\,. \tag{3.11}$$

Before digging into the numerical solution, I would like to provide some economic rationale behind the functions that the pinging strategy serves in the HFT's maximization problem. On one hand, hidden orders exists inside the spread with probability η^a or η^b and the HFT can use pinging orders ($\theta_t^{mk,b}=1$ or $\theta_t^{mk,b}=1$) to control his inventory by trying to hit those hidden orders. This is cheaper than using market orders as market orders cross the spread and pay the transaction fee, which amounts to a cost of $S_t/2+\epsilon$. However, if inventory control is the only motive for the HFT to utilize the pinging strategy, the HFT will always try to execute but not cancel his pinging orders. Thus the model could only produce pinging that executes against hidden orders without many cancellations.

This is why we need short-term price momentum on the other hand, i.e. the effect of depth imbalance F on $\mathbb{E}_t dP_t$ through the functions φ_1 and φ_2. With the presence of momentum, the HFT will also employ the pinging strategy to chase the price trend before it is gone. More importantly, the HFT will cancel his pinging orders when there is an abrupt change in imbalance F. If the momentum becomes too strong to wait, i.e. F becomes very large in absolute value, the HFT will cancel the pinging orders and place his directional bets via market orders. And if the momentum weakens substantially or even reverses, the HFT will also cancel the pinging orders since he does not want to be adversely hit. Hence such trend-chasing behaviors of the HFT enable the model to produce pinging activities as well as cancellations at the same time.

C. Numerical Solution

In this part, I focus on the numerical solution to the value function V of (3.7), and the associated policy functions (optimal strategies). In particular, since $V(t,x,y,p,f,s) = x + py + \nu(t,y,f,s)$, I will provide a backward, finite-difference scheme that solves the quasi-variational inequality (3.10) and (3.11), which completely characterizes the reduced-form value function ν. The numerical method is based on the finite-difference scheme developed by Chen and Forsyth (2008), as well as the scheme used in Guilbaud and Pham (2013).

To begin with, I consider a time discretization on the interval $[0,T]$ with time step $\Delta_T = T/N_T$ and a regular time grid

$$\mathbb{T}_{N_T} = \{t_k = k\Delta_T, k = 0, \ldots, N_T\}.$$

Secondly, I need to discretize and localize the state spaces for Y and F on two finite regular grids, with bounds M_Y, M_F, and step sizes $\Delta_Y = M_Y/N_Y \equiv 1$, $\Delta_F = M_F/N_F$ respectively, where $N_Y, N_F \in \mathbb{N}$, so that

$$\mathbb{Y}_{N_Y} = \{y_i = i\Delta_Y, i = -N_Y, \ldots, N_Y\}, \;\; \mathbb{F}_{N_F} = \{f_j = j\Delta_F, j = -N_F, \ldots, N_F\}.$$

Next, I define two finite-difference matrices, D_1 and D_2, for calculating first and second order derivatives against F on the F-grid $\mathbb{F}_{N_F}$, where D_2 uses central difference and D_1 uses forward difference when $f_j < 0$ and backward difference when $f_j \geq 0$:[22]

$$D_2\nu(t,y,f_j,s) = \frac{\nu(t,y,f_{j+1},s) - 2\nu(t,y,f_j,s) + \nu(t,y,f_{j-1},s)}{(\Delta_F)^2}$$

$$D_1\nu(t,y,f_j,s) = \begin{cases} \frac{\nu(t,y,f_{j+1},s) - \nu(t,y,f_j,s)}{\Delta_F} & \text{if } \; f_j < 0 \\ \frac{\nu(t,y,f_j,s) - \nu(t,y,f_{j-1},s)}{\Delta_F} & \text{if } \; f_j \geq 0 \end{cases}$$

I now state the main part of the numerical scheme. To this end, I introduce the explicit-implicit operator for the time-space discretization of the quasi-variational inequality (3.10), that is, for any $(t,p,f,s) \in [0,T] \times \mathcal{P} \times \mathbb{R} \times \mathbb{S}$ and any real-valued function $\phi :\mapsto \phi(t,y,f,s)$, I define

$$\mathcal{A}(t,y,f,s,\phi) = \max\left\{\widetilde{\mathcal{L}}(t,y,f,s,\phi) \;,\; \widetilde{\mathcal{M}} \circ \widetilde{\mathcal{L}}(t,y,f,s,\phi)\right\},$$

[22] The two matrices D_1 and D_2 are defined in a similar fashion to the finite difference space derivatives in Section 5.1 of Cont and Voltchkova (2005). The reason that I switch from forward to backward difference in D_1 when f_j becomes greater than 0 is detailed over there.

where[23]

$$\widetilde{\mathcal{L}}(t,y,:,s,\phi) = \Big(I_{N_F\times N_F} - \Delta_T\sigma_F^2 D_2 - \Delta_T\alpha_F(\mathbb{F}_{N_F}\mathbf{1}'_{N_F})..D_1\Big)^{-1}\times \\ \Bigg(\phi(t,y,:,s) + \Delta_T y\frac{\mathbb{E}_t dP_t}{dt} + \Delta_T\mathcal{L}^S\big(\phi(t,y,:,s)\big) - \Delta_T\gamma y^2\frac{\mathbb{E}_t d[P,P]_t}{dt} + \\ \Delta_T\sup_{\theta^{mk}}\Big\{g^a(:,s,\theta_t^{mk,b})..\big(\phi(t,y+1,:,s) - \phi(t,y,:,s) + \frac{s}{2} - \delta\theta^{mk,b}\big) + \\ g^b(:,s,\theta_t^{mk,a})..\big(\phi(t,y-1,:,s) - \phi(t,y,:,s) + \frac{s}{2} - \delta\theta^{mk,a}\big)\Big\}\Bigg), \tag{3.12}$$

and

$$\widetilde{\mathcal{M}}\circ\widetilde{\mathcal{L}}(t,y,f,s,\phi) = \sup_{|\zeta|\le\zeta_{max}}\Big\{\widetilde{\mathcal{L}}(t,y+\zeta,f,s,\phi) - |\zeta|(\frac{s}{2}+\epsilon) + |y|\int_H H\,dG(H\,|\,s,f)\Big\}. \tag{3.13}$$

When inventory y is on the boundary of $\mathbb{Y}_{N_Y}$, i.e. $y=-M_Y$ or $y=M_Y$, make and take strategies are confined to the buy side or the sell side only, so that y does not go off its grid. Then, I approximate the solution ν to (3.10)-(3.11) by the numerical solution ϖ on $\mathbb{T}_{N_T}\times\mathbb{Y}_{N_Y}\times\mathbb{F}_{N_F}\times S$ to the backward explicit-implicit finite difference scheme:

$$\varpi(T,y,f,s) = -|y|(\frac{s}{2}+\epsilon) + |y|\int_H H\,dG(H\,|\,s,f) \tag{3.14}$$

$$\varpi(t_k,y,f,s) = \mathcal{A}(t_{k+1},y,f,s,\varpi),\quad k=N_T-1,N_T-2,\ldots,0\,, \tag{3.15}$$

where (3.14) and (3.15) approximate (3.11) and (3.10) respectively.

The complete solution algorithm for the value function and the policy functions is summarized in the backward induction steps below:

1. At the terminal date $t_{N_T}=T$: for each combination of (y,f,s), make $\varpi(T,y,f,s) = -|y|(\frac{s}{2}+\epsilon)+|y|\int_H H\,dG(H\,|\,s,f)$.

2. (*Backward Induction*) From time step t_{k+1} to t_k where k runs from N_T-1 back to 0, for each combination of (y,f,s):

▷ Calculate $\widetilde{\mathcal{L}}(t_{k+1},y,f,s,\varpi)$ from (3.12) and obtain $\theta^{mk,*}$.

[23]In the operator $\widetilde{\mathcal{L}}$, : denotes the column vector of $\mathbb{F}_{N_F}$, $I_{N_F\times N_F}$ is an N_F by N_F identity matrix, $\frac{\mathbb{E}_t dP_t}{dt}$, $\frac{\mathbb{E}_t d[P,P]_t}{dt}$ are vectors evaluated on $\mathbb{F}_{N_F}$, $\mathbf{1}'_{N_F}$ is an $N_F\times 1$ vector of 1s, and .. denotes element-by-element product for vectors and matrices. $\widetilde{\mathcal{L}}$ is expressed as a vector on the grid $\mathbb{F}_{N_F}$ because of the implicit time step that I used when approximating the generator $\mathcal{L}^F$.

▷ Calculate $\widetilde{\mathcal{M}} \circ \widetilde{\mathcal{L}}(t_{k+1}, y, f, s, \varpi)$ from (3.13) and obtain $\theta^{tk,*}$.

▷ If $\widetilde{\mathcal{L}}(t_{k+1}, y, f, s, \varpi) \geq \widetilde{\mathcal{M}} \circ \widetilde{\mathcal{L}}(t_{k+1}, y, f, s, \varpi)$, set $\varpi(t_k, y, f, s) = \widetilde{\mathcal{L}}(t_{k+1}, y, f, s, \varpi)$ and the policy at t_k is thus $\theta^{mk,*}$, given (y, f, s).

▷ Otherwise, let $\varpi(t_k, y, f, s) = \widetilde{\mathcal{M}} \circ \widetilde{\mathcal{L}}(t_{k+1}, y, f, s, \varpi)$, and $\theta^{tk,*}$ is taken to be the policy at t_k, given (y, f, s).

Finally, I state the convergence theorem of my numerical solution ϖ to the reduced-form value function ν (and hence the convergence of discretized policy functions), with its proof left to the appendix.

Proposition 2. *The solution ϖ to the numerical scheme (3.14)-(3.15) and the corresponding discretized policies converge locally uniformly, respectively, to the reduced-form value function ν and the optimal strategies on $[0,T]\times\mathbb{R}\times\mathbb{R}$ as $(\Delta_T, \Delta_Y, M_Y, \Delta_F, M_F) \to (0, 0, \infty, 0, \infty)$, $\forall s \in \mathcal{S}$.*

IV. Estimation

An important aspect of my paper is to examine the optimal trading strategies of the HFT given exogenous evolutions of the LOB. In order to quantify the implications of my model based on the numerical solution, I need to obtain values for the parameters that govern the stochastic processes of the LOB characteristics, and this section concerns the estimation of these parameters. The table below summarizes the parameters to be estimated.

Table 1: Parameters of order book characteristics

Parameters	Explanation	Estimation Procedure
$\lambda^S,\ \rho$	jump intensity and transition matrix of S	Non-parametric estimators
$\alpha_F,\ \sigma_F$	mean-reversion and volatility parameters of F	Maximum likelihood
$\lambda^J_{1,2}$	jump intensities of P	Non-parametric estimators
$\beta_{1,2}$	prob. distribution parameters of directions of P jumps	Logistic regressions
$\kappa,\ \pi_{1,2}$	prob. distribution parameters of hidden orders	Logistic regressions
$\varsigma_{0,1}$	Parameters of limit order fill rates	Logistic regressions
$\lambda^{M,a},\ \lambda^{M,b}$	market order arrival intensities	Non-parametric estimators

A. Data

I use Nasdaq TotalView-ITCH 4.0 limit order book message feed data on three types of stocks listed on Nasdaq during the month of June 2012 (21 trading days). The three types consist of stocks with narrow spreads and high order-book depths, stocks with medium spread and depth levels, and stocks with wide spreads and low order-book depths. I focus on three representative stocks, with one from each of the types: INTC (Intel, narrow spread and high depth), QCOM (Qualcomm, medium spread and medium depth), and AMZN (Amazon, wide spread and low depth).

The TotalView data include all real-time Nasdaq limit order book messages of these stocks, stamped to millisecond precision. The complete message feeds allow me to reconstruct the whole limit order books and their complete evolutions for the three stocks. Since the limit order book characteristics of my model concern only the first level of the book, I track the evolutions at the top of the book for estimation purposes. As in common practices, I use data between 9:45am and 15:45pm to avoid certain erratic market movements.[24]

B. Estimation

I shall employ standard nonparametric estimators for all the intensity parameters as well as the transition matrix ρ.[25] The parameters of the depth imbalance F are then estimated using maximum likelihood, since the transition density of an Ornstein-Uhlenbeck process is known in closed form[26]. In addition, the parameters governing various probabilities are estimated by logistic regressions, as these distribution functions all have logistic forms. I conduct the estimations for all trading days in June 2012, and then calculate the averages of these daily estimates as my final estimated parameter values to be fed into my numerical solutions. The mean values of my daily parameter estimates are presented in Table 2, with Newey-West heteroskedasticity and autocorrelation consistent (HAC) standard errors in parentheses.

There are three aspects of the estimation results that are worth being pointed out. Firstly, compared to Amazon, Intel and Qualcomm have a less volatile imbalance process F as well as a higher tendency to stay at lower-spread positions. Secondly, due to its lower book depth, Amazon's mid-price jumps are more often, as measured by both λ_1^J and λ_2^J. Thirdly, the estimates of hidden order parameters, market order arrival rates and limit order fill rates are more alike for those three stocks. We will see in the next section that these dissimilarities and

[24]Please refer to the appendix for a complete description of the limit order book data.

[25]For a standard reference, see, for example, Karr (1991).

[26]For a reference, see Aït-Sahalia (1999) and Aït-Sahalia and Mykland (2003).

Table 2: Order Book Parameter Estimates

		INTC	QCOM	AMZN
Spread	λ^S	0.161/s	0.312/s	0.578/s
		(0.010)	(0.017)	(0.027)
	ρ	$\begin{pmatrix} 0 & 1 & 0 \\ 0.99 & 0 & 0.01 \\ 0 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0.97 & 0.03 \\ 0.95 & 0 & 0.05 \\ 0.08 & 0.92 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0.743 & 0.257 \\ 0.222 & 0 & 0.778 \\ 0.138 & 0.862 & 0 \end{pmatrix}$
Imbalance	α_F	0.308	0.547	0.734
		(0.016)	(0.027)	(0.035)
	σ_F	0.777	1.429	2.336
		(0.029)	(0.053)	(0.074)
Mid-price	λ_1^J	0.161/s	0.307/s	0.522/s
		(0.010)	(0.018)	(0.027)
	β_1	2.744	2.651	2.610
		(0.077)	(0.076)	(0.075)
	λ_2^J	0.052/s	0.075/s	0.121/s
		(0.003)	(0.005)	(0.007)
	β_2	4.766	2.921	1.881
		(0.244)	(0.257)	(0.129)
Hidden Order	κ	1.196	1.036	0.992
		(0.034)	(0.030)	(0.021)
	π_1	0.238	0.233	0.230
		(0.010)	(0.010)	(0.009)
	π_2	0.125	0.117	0.114
		(0.006)	(0.006)	(0.006)
MO arrival	$\lambda^{M,a}$	0.110/s	0.130/s	0.181/s
		(0.006)	(0.006)	(0.012)
	$\lambda^{M,b}$	0.110/s	0.130/s	0.179/s
		(0.004)	(0.006)	(0.010)
Fill rate	ς_0	1.320	1.454	1.648
		(0.039)	(0.055)	(0.066)
	ς_1	0.399	0.462	0.673
		(0.019)	(0.022)	(0.031)

Note: The table shows the average values of the daily parameter estimates in the month of June 2012, using limit order book data reconstructed from Nasdaq TotalView-ITCH 4.0 real-time message feeds. Newey-West HAC standard errors are in parentheses. The standard errors of the transition matrix estimates are not shown since they are close to zero. Intensity parameters are all measured at per second frequency. In addition, I normalize the tick size δ to be \$0.1 for Amazon, since on average, the spread and limit order prices of Amazon tend to be in multiples of 0.1 instead of the minimum tick size \$0.01 that is used for Intel and Microsoft. For definitions of all the parameters, please refer to Table 1.

similarities in parameter estimates will lead to a much different optimal HFT strategy profile for Amazon as opposed to Intel or Qualcomm.

Besides the estimated parameters, the fixed parameters in Table 3 are also used in my numerical quantification and simulation study of the optimal HFT strategies.

Table 3: Fixed Parameters

		Description	Value
Discrt/loc parameters	T	Time Length in seconds	3600
	Δ_T	Size of time step in seconds	0.5
	M_Y	Inventory grid bound (in lot)	30
	Δ_Y	Inventory grid step size (in lot)	1
	M_F	Depth imbalance grid bound	10
	Δ_F	Depth imbalance grid step size	0.01
Model constants	δ	Tick size	0.01
	ϵ	Per share fee	0.003
	γ	Inventory penalization	2
	ζ_{max}	Max market order size (in lot)	10
Backtest parameters	N_{MC}	Number of MC simulation paths	10000
	X_0	Initial cash	0
	Y_0	Initial inventory	0
	P_0	Initial mid-price of stock	10

Note: I choose the per share fee of market orders to be \$0.003, which corresponds to the same transaction fee stated by Nasdaq for its stocks. Furthermore, the tick size δ is set to be \$0.01, which is the tick size for all Nasdaq stocks that have prices above \$1. However, I make the tick size \$0.1 for Amazon, for the same reason indicated in the note under Table 2.

V. Computation and Simulation Results

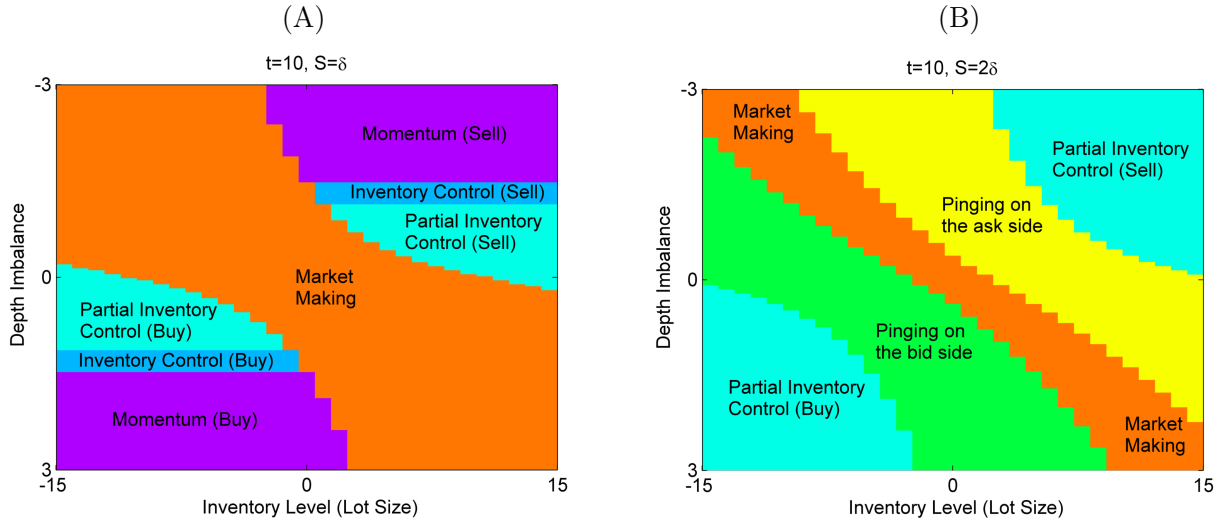
In this section, I provide numerical results obtained with the optimal HFT strategy computed via the implementation of my numerical scheme (3.14)-(3.15), for the three stocks – Intel (INTC), Qualcomm (QCOM) and Amazon (AMZN). I use as inputs the estimated parameter values shown in Table 2, together with the fixed parameters listed in Table 3.

A. Optimal Strategy Profiles

As seen from the reduced-form value function ν, the optimal HFT strategy depends on time t, inventory Y, depth imbalance F and spread level S. Therefore, I will characterize the strategy

as a function of inventory and depth imbalance, for spread equal to δ and 2δ, near $t = 0$ and $t = T$. The strategy profiles are mostly time invariant if time t is not very close to the terminal date T. In addition, the optimal strategy profiles for Qualcomm are not shown since they are in between the ones for Intel and the ones for Amazon.[27]

Figure 1: Optimal Strategy, INTC, $t = 10$, spread $= \delta$ (left) / 2δ (right)



I will start with **Figure 1**, which illustrates the optimal HFT strategies for Intel at $t = 10$, with inventory and depth imbalance level shown on the horizontal and the vertical axis respectively. Consider Figure 1(A) on the left first, where spread equals δ. The orange-colored central region denotes market making through submitting a limit order at both the best bid and ask prices. The two blue regions stand for inventory management; buy (resp. sell) indicates using buy (resp. sell) market orders to increase (resp. decrease) inventory towards zero. Partial inventory control occurs when inventories are only partially unwound, whereas inventory control represents complete liquidation so that inventory jumps precisely back to zero.[28] Furthermore, momentum represents utilizing market orders to change inventory and set up a directional position; (buy) and (sell) denote establishing positive and negative inventory holding positions respectively.[29] Finally, there is no pinging since it is not possible when the spread is

[27]The reason is that the key parameter estimates $(\lambda^S, \alpha_F, \sigma_F, \lambda_1^J, \lambda_2^J)$ of Qualcomm are between those of Intel and Amazon, as pointed out in Table 2.

[28]For instance, if the inventory is -5 lots, the strategy will specify purchasing z lots with $z < 5$ in the region of partial inventory control (buy), yet it will specify buying exactly 5 lots in the region of inventory control (buy) to make inventory become zero.

[29]For example, under momentum (buy), if the inventory is -5 lots, the strategy will dictate a purchase of z

at its minimum δ, and the majority of the graph is represented by traditional market-making / inventory-control behaviors.

The economic intuition behind Figure 1(A) is explained as follows. The situation where depth imbalance is less than zero will be focused on, since a similar but symmetrically reversed exposition can be applied to the opposite case where depth imbalance is greater than zero. To begin with, when depth imbalance is mildly positive, the probability of a positive mid-price jump is somewhat higher than that of a negative one. If the HFT has a negative inventory, he will face a inventory risk as the price would move against his holdings. Thus he will reduce such risk via a market order. Due to the cost of market-taking, the HFT either partially or completely disposes his inventory depending on the amount of the risk he has, i.e. how positive the depth imbalance is.[30] On the contrary, if the HFT holds a positive inventory, he would enjoy a possible gain from a positive mid-price jump and choose to make the market as a result. Since market taking is expensive and the signal effect from depth imbalance on the mid-price movement is not strong enough, the HFT would not use market orders to accumulate additional positive inventory, i.e. the expected return is not large enough.

Furthermore, when depth imbalance becomes considerably more positive, the probability of a positive mid-price jump is much higher than that of a negative one. If the HFT has a negative inventory, he faces a substantial inventory risk, yet also a clear opportunity to chase the upward price momentum. Anticipating the likely price increase, the HFT aggressively takes the liquidity from the market through buy market orders to establish a directional (positive) inventory position, which gives rise to the momentum (buy) region in the graph. However, the HFT stops chasing the price momentum if his inventory is rather positive. As the mid-price jump intensities of Intel are not very large, being too aggressive and obtaining too much positive inventory would result in inventory risk on the opposite side if the mid-price does not jump up soon enough.[31]

Next, consider Figure 1(B) on the right, where spread equals 2δ. The green and yellow regions on each side of market making represent pinging strategies, which the majority of the graph consists of. Here, pinging on the bid (resp. ask) side denotes submitting a buy (resp. sell) limit order inside spread at the best bid plus δ (resp. best ask minus δ), while letting the sell

lots with $z > 5$, and if the inventory is 1 lot, the strategy will dictate a purchase of z lots with $z \geq 1$.

[30]Note that market-making here would not achieve this risk-reduction purpose, for two reasons. First, market-making implies that the HFT would still post a sell limit order at the best ask. Second, the effect of a positive depth imbalance on the fill-rate function means that a buy limit order of the HFT has a smaller chance of being filled. Consequently, instead of decreasing, the two together exacerbate the inventory risk faced by the HFT.

[31]The depth imbalance is mean-reverting towards zero. Consequently, the anticipated price jump would become less likely if it does not occur very soon.

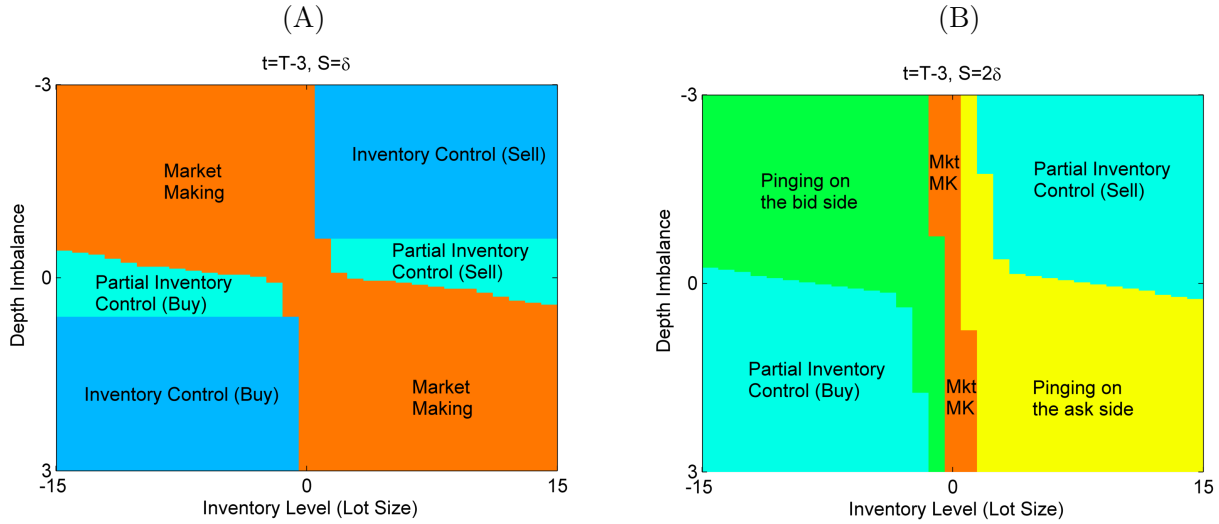
(resp. buy) limit order joining the queue at the best ask (resp. best bid). To understand the intuition behind Figure 1(B) and compare it to the case of Figure 1(A), I will again concentrate on the scenario in which the depth imbalance is less than zero. A symmetrically reversed explanation can be constructed similarly for the opposite scenario.

When depth imbalance is modestly greater than zero, if the HFT's inventory level is quite negative, he will remove the inventory risk through market orders. However, the inventory control is only partial since complete inventory controls are too costly under $S = 2\delta$ as opposed to $S = \delta$. Alternatively, if the HFT's inventory level is closer to zero, he will instead use a pinging strategy from the bid side to reduce his inventory risk, for three reasons. Firstly, despite execution uncertainty, pinging is free and the risk is smaller with inventory level not far from zero, which decreases the HFT's desire for immediacy. Secondly, a buy pinging order might hit a sell hidden order inside the spread. Thirdly, a positive depth imbalance implies a lower fill rate for normal buy limit orders, so that it is optimal for the HFT to jump the queue. On the other hand, if the HFT holds a significant level of positive inventory, he is confronted with the inventory risk due to the possibility of no upward jump in mid-price. Hence he will ping on the ask side, which has a high chance of hitting a bid-side hidden order inside the spread (as the depth imbalance is positive) and entails no cost compared to using market orders.

When the positivity of depth imbalance becomes sizable, the HFT would reduce his inventory risk more aggressively through market orders if his inventory is below zero. However, provided that the inventory is positive but small, the HFT would instead pursue price momentum by pinging on the bid side, which increases the execution probability of his buy limit order. Moreover, chasing the momentum via market-taking is suboptimal in this case since the expected gain is less that the cost of using market orders ($S = 2\delta$). As a result, depending on the configuration of depth imbalance and inventory, pinging strategies would serve two different functionalities: unwinding inventory or pursuing price momentum.

After that, let us examine **Figure 2**, which shows the optimal HFT strategies for Intel at $T-3$. Consider Figure 2(A) on the left first, with spread equal to δ. Since time t is close to the terminal date, inventory management becomes a large concern for the HFT as he must liquidate all positions at T. Consequently, if the HFT is holding negative (resp. positive) inventory and the mid-price is more likely to jump up (resp. down) because of positive (resp. negative) depth imbalance, he will aggressively unload his inventory through market orders to reduce the risk associated with such mismatch of inventory against depth imbalance. This is the reason why approximately half of the graph is made of inventory controls. Conversely, when the HFT's inventory corresponds to the likely price movement, he would still want to seize the expected gains from carrying a directional position. Hence the HFT would make the market instead if

Figure 2: Optimal Strategy, INTC, $t = T - 3$, spread $= \delta$ (left) / 2δ (right)



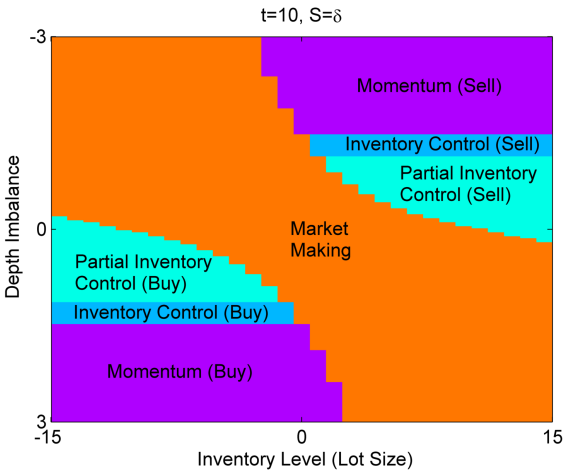
his inventory is positive (resp. negative) and the depth imbalance is the opposite. In this case, the HFT will not find it optimal to employ an aggressive momentum strategy to establish more directional positions in contrast to the situation of $t = 10$, since inventory control is the primary worry when time approaches T.

Consider Figure 2(B) next. Similar to Figure 2(A), inventory management accounts for a large part of the optimal strategy profile, but as in Figure 1(B), the inventory control is partial due to the costly nature of market orders under $S = 2\delta$. The main difference from Figure 1(B) is that pinging strategies now have one sole objective: reducing inventory risk. The HFT would only ping on the ask (resp. bid) side if his inventory and the depth imbalance are both positive (resp. negative). This is because pinging orders are free and have a larger probability of executing against an opposite-side hidden orders inside the spread compared to queuing limit orders.

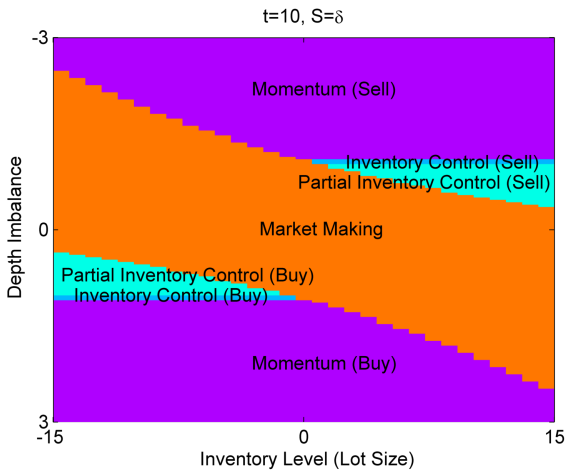
Let us now turn our attention to **Figure 3**, which compares the optimal strategy for Amazon (right) to that for Intel (left) in the case of time $t = 10$ and spread $S = \delta$. It is clear that the major difference occurred to Amazon is that the HFT takes liquidity a lot more often to carry out momentum strategies. The HFT is being more aggressive here since Amazon has much higher mid-price jump intensities as shown in Table 2. Therefore, the HFT is anticipating an upward (resp. downward) directional price movement with a much larger likelihood if depth imbalance swings to the positive (resp. negative) region, which leads to the HFT building a matching directional position forcefully via market orders. The expected reward of a price jump

Figure 3: Optimal Strategy, spread $= \delta$, $t = 10$, INTC (left) v.s. AMZN (right)

(A) Intel



(B) Amazon



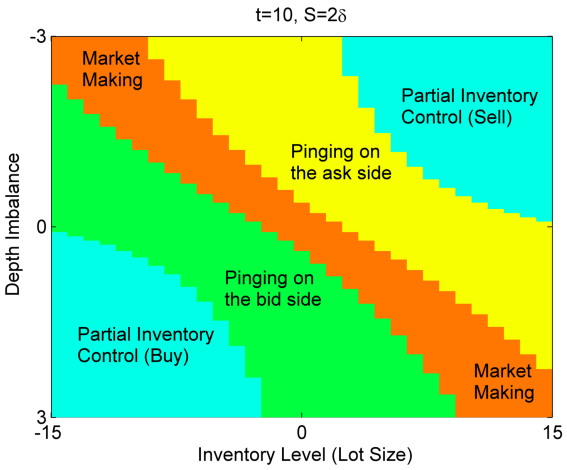
in the short term simply outweighs the cost of aggressive market-taking strategies.

Then, let us look at **Figure 4** that contrasts the optimal strategy for Amazon (right) to that for Intel (left) in the case of time $t = 10$ and spread $S = 2\delta$. There are two main changes occurred to pinging strategies employed for Amazon. First, there is no pinging on the bid (resp. ask) side when the HFT's invetory is negative (resp. positive) and the depth imbalance is also moderately negative (resp. positive). Second, pinging on the bid (resp. ask) side erodes everything when both the HFT's invetory and the depth imbalance are above (resp. below) zero. These imply that when the HFT holds a directional position that matches the likely course of a mid-price change, pinging's only aim is to chase the short-term price momentum. The reason is similar to the one given for **Figure 3**, and it is because Amazon has much higher mid-price jump intensities so that the HFT can afford to be more aggressive. Nevertheless, it is not optimal for the HFT to deploy momentum strategies via market-taking as it is too expensive and the price jump is already in favor of the HFT's inventory. However, if the depth imbalance is extremely high, and the HFT carries inventory that is against the likely directional move of price, we observe that the HFT would actually use momentum strategies in the case of Amazon (purple areas). The large cost of market orders is absorbed by the reduction in inventory risk and the almost certain, short-term benefit from the anticipated price jump.

Finally, **Figure 5** and **Figure 6** are discussed together, which demonstrate the optimal HFT strategies for Amazon (right) towards the terminal date T and compare them to those for Intel (left). From these two figures, we notice that the optimal strategies for Amazon closely

Figure 4: Optimal Strategy, spread $= 2\delta$, $t = 10$, INTC (left) v.s. AMZN (right)

(A) Intel



(B) Amazon

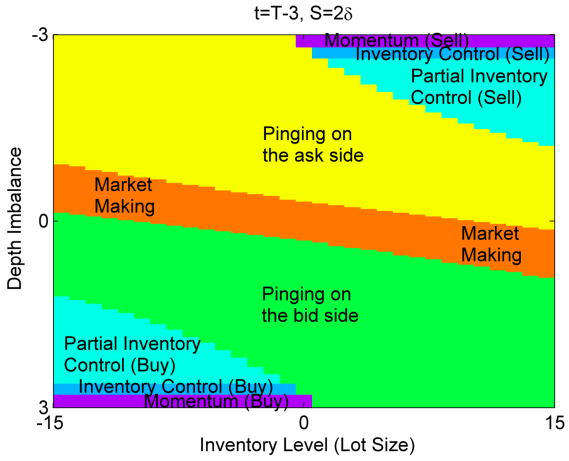
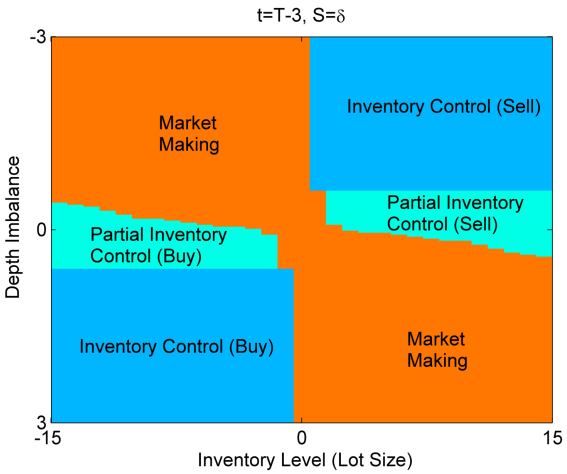


Figure 5: Optimal Strategy, spread $= \delta$, $t = T - 3$, INTC (left) v.s. AMZN (right)

(A) Intel



(B) Amazon

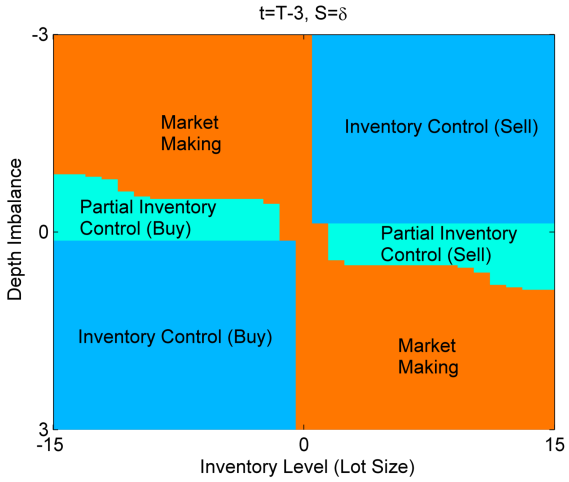
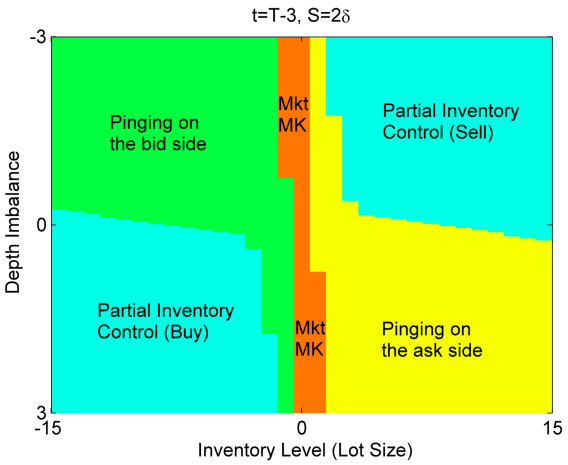
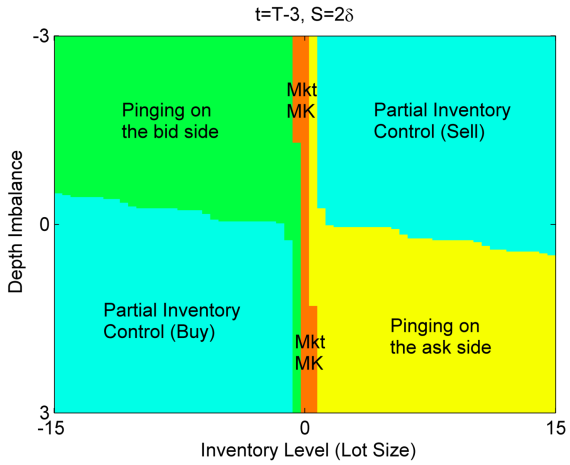


Figure 6: Optimal Strategy, spread $= 2\delta$, $t = T - 3$, INTC (left) v.s. AMZN (right)

(A) Intel



(B) Amazon



resemble those for Intel, despite the differences seen in **Figure 3** and **Figure 4** when $t << T$. Once more, this is the result of inventory management concern being a dominant force as time draws closer to the end, so that establishing directional positions to pursue short-term price momentum is no longer one of the objectives of the HFT.

B. Percentage Attributions of Optimal Strategies

In this part, I conduct Monte Carlo simulation studies to quantify the properties of the HFT's optimal strategies on the three stocks: Intel, Qualcomm and Amazon. The number of Monte Carlo runs is set to be $N_{MC} = 10000$, and I use a standard Euler scheme to simulate the paths of the state variables (P, S, F, X, Y) as well as the exogenous market order arrivals (M^b, M^a).

Table 4: Percentage Breakdown of the Optimal HFT Strategies

	INTC	QCOM	AMZN
Market-making/Inventory-control	70.50%	58.30%	31.80%
Pinging	19.70%	28.10%	46.70%
Momentum (via market orders)	9.80%	13.60%	21.50%

Note: Please refer to Footnote 32 below on how these percentage breakdowns are obtained.

Table 4 summarizes the (Monte Carlo average) percentage attributions of the optimal HFT to three types of activity: traditional market-making/inventory-control, pinging, and momentum/directional trading via market orders.[32]

There are two main implications that we can learn from Table 4. **Firstly**, for stocks with narrow spreads on average and abundant order book depths such as Intel, the HFT behaves like a market maker in the traditional sense. He is providing liquidity to the market most of the time. As a result, pinging activities account for only 20% of the optimal strategies and approximately 70% of the strategies are in the realm of liquidity provision (market-making) and inventory management. It corresponds to what we saw in the profiles of the optimal strategies for Intel early on, as Intel's spread level is often equal to δ. **Secondly**, for stocks like Amazon with wide spreads on average and an order book that has low depths and volatile movements, the HFT looks less like a market maker. Instead, he acts more like a short-term profit/momentum chaser, since pinging and momentum trading together account for about 70% of the strategies for Amazon. In particular, pinging constitutes nearly 50% of the optimal strategies. This also matches with the profiles of the optimal strategies for Amazon, where momentum trading and pinging constitute the majority of the strategy profiles under $S = \delta$ and $S = 2\delta$ respectively. Pinging can be considered as a strategy that demands liquidity from the market when its objective is to build directional inventory position. Hence for stocks with wide spreads and low book depths, the HFT is mainly a market taker and quite often he is trying to take liquidity in order to bet on the directional moves of price.[33]

Next, I compared the pinging percentages obtained from the model (as shown in Table 4) to the pinging percentages observable from the data to see how much pinging in the data can be rationalized by the model. To calculate pinging percentages in the data, I first compute the number of cancelled pinging activities for different stocks in a similar fashion to Hasbrouck and Saar (2009). This is defined as the number of limit orders submitted inside the spread and then cancelled in less than 2 seconds. Second, I need to compute the number of pinging orders that execute against hidden orders inside the spread. This is somewhat problematic since such pinging orders are identical to market orders that hit hidden orders. In order to deal with this

[32]The percentage breakdowns are calculated as follows. For each one of the simulations, I compute the number of choices attributable to each type of the activity/strategy chosen optimally under the numerical solution given the simulated state variables. Then I divide these numbers by the total number of activities, which equals $T/\Delta_T = 7200$, to arrive at the corresponding percentage numbers. Finally, I average these percentage numbers across all 10000 simulations to obtain the numbers shown in Table 4.

[33]For each of the three stock types, I checked 10 stocks, including the representative ones (Intel, Qualcomm, Amazon) being focused here. The percentage attributions are similar, as well as the optimal strategy profiles, i.e. the HFT is more of a market maker (resp. taker) and the pinging percentage is lower (resp. higher) if the stock under consideration has higher (resp. lower) order book depths and narrower (resp. wider) spreads.

issue, I treat non-consecutive orders executing on hidden orders as pinging orders. Therefore, the number of total pinging activities are calculated as the sum of the number of cancelled pinging activities and the number of pinging orders hitting hidden orders. I then divide the number of total pinging activities by the total number of order book activities to arrive at the pinging percentages obtainable from the data:

$$\text{Pinging \% in the data} = \frac{\text{Number of total pinging activities}}{\text{Total number of order book activities}}.$$

Finally, the model's pinging percentages are gauged against those from the data in Table 5.

Table 5: Model's Pinging v.s. Data's Pinging

	INTC	QCOM	AMZN
Pinging % observed in data	23%	39%	70%
Pinging % produced by model	20%	30%	50%
% of data's pinging captured by model	85%	75%	70%

Note: Pinging percentages produced by the model are approximations to the corresponding numbers shown in Table 4.

As clearly seen from Table 5, the pinging percentages produced by the model match quite closely to the percentages from the data. Moreover, at least over 70% of the pinging observable from the data can be captured by the model, and the number is higher for stocks with high depths and low spreads (like Intel). Hence the result indicates that most of the pinging activities observed in the data can be rationalized by the model with the two mechanisms of inventory control and trend chasing.

Table 6: Further Breakdown on Pinging

	Pinging's % in Strategies		
	INTC	QCOM	AMZN
No hidden orders	8%	16%	32%
No price momentum	12%	14%	18%

Table 6 further breaks down the roles that the mechanisms of inventory control and momentum chasing play in rationalizing pinging activities in the model. There are two noticeable features from Table 6. Firstly, for stocks with high depths and narrow spreads such as Intel,

inventory control and momentum chasing contribute comparably to the rationalization of pinging. This can be seen from the similar resulting pinging percentages given by the model if either of the mechanism is shut down. Secondly, for stocks with low depths and wide spreads such as Amazon, momentum chasing carries more weight than inventory control in the rationalization of pinging. This is because the pinging percentage produced by the model decreases by a larger amount when the channel of price momentum is shut down than when the channel of hidden order is.

Additionally, Table 6 also implies that both inventory control and trend chasing are necessary for the purpose of pinging rationalization. Less than 50% of the pinging in the data is rationalized by the model if either mechanism is turned off, suggesting that both mechanisms are indispensable.

C. Auxiliary Predictions

Besides pinging rationalization, the model also yields a couple of other interesting auxiliary predictions regarding pinging activities with respect to depth imbalances. They are presented in this subsection.

The first auxiliary prediction of the model is related to the directions of pinging activities. The model implies that if the HFT sees positive (resp. negative) depth imbalance and hence positive (resp. negative) price momentum more often, he is more likely to pining from the buy (resp. sell) side and take positive (resp. negative) directional bets due to trend chasing motives. Consequently, it yields the following prediction:

Prediction 1. *There is more pinging from the buy (resp. sell) side if depth imbalance is more frequently positive (resp. negative).*

Now I check this prediction against data. To do this, I divide each trading day into 30-second intervals.[34] Next, I calculate the number of buy and sell pinging activities in each of these intervals in a similar manner to the computation of the number of pinging activities shown in Section V(B). In addition, I also compute the durations (measured in seconds) of positive and negative imbalance in each interval. Then I use the following regression to measure and test the effect of imbalance durations on pinging activity directions:

$$PO_t = \alpha + \beta DoI_t + x_t'\gamma + \epsilon_t.$$

[34]The results are similar if I use 15-second or 1-minute intervals instead of 30-second intervals.

PO_t denotes the number of buy or sell pinging activities and DoI_t the duration (measured in seconds) of positive or negative depth imbalance in interval t. x_t stands for other control variables which include DoI_{t-1}, average spread and volatility of imbalance in interval t, and ϵ_t is an error term.

The regression is performed separately for buy pinging activities on positive imbalance and sell pinging activities on negative imbalance, and across all trading days in June 2012. The final parameter estimates are calculated in the same way as in Table 2, i.e. time series averages of daily parameter estimates. Overall, I find that the parameter β is statistically significant for all three types of stocks, with a value around 0.28 on average. This shows that the first auxiliary prediction is confirmed in the data.

The second auxiliary prediction of the model concerns the influence of depth imbalance volatilities on the frequencies of cancelled pinging activities. The model implies that If momentum strengthens (imbalance widens) by a large amount, contemporaneously the HFT would cancel his pinging orders and use market orders instead to chase momentum. And if momentum weakens by a large amount or even reverses (imbalance reduces by a large amount or reverses), contemporaneously the HFT would also cancel his pinging orders as his pinging orders risk being adversely hit. This implication thus yields the following prediction:

Prediction 2. *If depth imbalance is more volatile, there should be more cancelled pinging activities occurring at the same time.*

I employ a similar procedure as before to check this prediction, i.e. I divide each trading day into 30-second intervals and compute the number of cancelled pinging activities in each interval. Then I use the following regression to measure and test the effect of imbalance volatilities on cancelled pinging orders:

$$CPO_t = \alpha + \beta VoD_t + x_t'\gamma + \epsilon_t.$$

CPO_t denotes the number of cancelled pinging orders and VoD_t the log of depth imbalance volatility in interval t. x_t again stands for other control variables which include VoD_{t-1}, average spread and number of market order arrivals in interval t, and ϵ_t is an error term.

The regression is performed across all trading days in June 2012. The final parameter estimates are also calculated as the time series averages of daily parameter estimates. Overall, I find that the parameter β is statistically significant for all three types of stocks as well. Nevertheless, the magnitude of β is much higher for stocks with high depths and narrow spreads (with a value around 3.3) compared to stock with low depths and wide spreads (with a value less than 1). Therefore, the result suggests that the second auxiliary prediction is by and large confirmed in the data too.

VI. Conclusion

In this paper, I build a continuous-time, partial equilibrium model on the optimal strategies of HFTs without any learning or manipulative ingredients to rationalize pinging activities observed in the data. The model improves on the works of Holl and Stoll (1981) as well as Guilbaud and Pham (2013) by introducing hidden orders inside the bid-ask spread and short-term price momentum. The HFT then uses pinging orders inside the spread to either control inventory or chase price trends. I demonstrate that for stocks with high order book depths and narrow spreads, pinging accounts for 20% of the optimal strategies in the model, whereas this number goes up to 50% for stocks with low order book depths and wide spreads. I then compare these pinging percentages from the model to their corresponding counterparts in the data, and find that over 70% of the pinging activities in the data are captured by the model. The result thus suggests that most of the pinging in reality can be rationalized by my model. Furthermore, I show that for low-depth and wide-spread stocks, the majority of the pinging in the model is due to the momentum-chasing motive. However, for high-depth and low-spread stocks, the inventory control motive would play a similar role to the momentum-chasing motive in rationalizing pinging activities. In addition, I also develop a couple of other auxiliary predictions based on the model's implications. They are both assessed on and found to be consistent with the data in general. Therefore, my model gives the overall message that pinging activities do not necessarily have to be manipulative and can be mostly rationalized as part of the standard dynamic trading strategies of HFTs.

Appendix

A. Data Description

In the first part of the appendix, I will give a detailed description of the limit order book data used in this paper. I utilize the Nasdaq TotalView-ITCH 4.0 database, which includes all real-time messages of limit order submissions, cancellations, executions and hidden order executions for every trading day since 7:00am EST when Nasdaq's electronic limit order book (LOB) system starts accepting incoming limit orders, stamped to millisecond precision. The system is initialized by an empty order book where all overnight limit orders are resubmitted automatically at the beginning of each day. The TotalView-ITCH data record a unique identification for any limit order and I can identify the attribute of a limit order (cancelled, executed or neither) by tracking it through its order ID. Furthermore, trades are identified via the records of limit orders

and hidden order executions. Since the trading direction of limit orders and hidden orders is recorded, I can exactly identify whether a trade is buyer-initiated or seller-initiated. Hence the LOB at any time can be reconstructed (up to millisecond precision) through continuously updating the book system according to all reported messages, which exactly represents the historical real-time-disseminated order book states of Nasdaq.

Furthermore, I consolidate all trade transactions into one single trade from a market order if they are logged at the same time-stamp in the data and have the same initiation types. In addition, I use the reconstructed LOB data between 9:45am and 15:45pm only, in order to avoid erratic effects that are likely to occur at market opening or closure.

B. Proofs for Various Lemmas and Propositions

The second part of the appendix is devoted to the proofs for several lemmas and propositions, which are omitted in the main paper.

Proof of Lemma 1: On one hand, I can show that the maximal value of (3.5) is smaller than that of (3.6). This is because, for any admissible strategy such that $Y_T = 0$, we immediately obtain $Q(X_T, Y_T, P_T, F_T, S_T) = X_T$. On the other hand, given an arbitrary admissible strategy θ and its associated state variable processes (X, Y, P, F, S), I can consider an alternative strategy $\tilde{\theta}$, coinciding with θ up to time T and employing an market order that liquidates all the inventory Y_T at the terminal date T. The associated processes of the state variables $(\tilde{X}, \tilde{Y}, P, F, S)$ under $\tilde{\theta}$ satisfy $(\tilde{X}_t, \tilde{Y}_t, P_t, F_t, S_t) = (X_t, Y_t, P_t, F_t, S_t)$ for all $t < T$, and $\tilde{X}_T = Q(X_T, Y_T, P_T, F_T, S_T)$, $\tilde{Y}_T = 0$. Therefore, this shows that the maximal value of (3.6) is smaller than that of (3.5), hence (3.5) is equivalent to (3.6).

Proof of Lemma 2: The lower bound is obvious, since $V(t, x, y, p, f, s) = Q(x, y, p, f, s)$ by adopting the simple strategy that sets $\theta_t^{tk} = y$ and terminates the problem immediately at time t. On the other hand, we have

$$\begin{aligned} V &= (x + py) + \sup_{\theta} \mathbb{E}_t \left[(X_T - x) + (P_T Y_T - py) - |Y_T| \Big(\frac{S_T}{2} - H_T + \epsilon \Big) - \gamma \int_t^T Y_{u-}^2 d[P, P]_u \right] \\ &\leq (x + py) + \sup_{\theta} \mathbb{E}_t \left[(X_T - x) + (P_T Y_T - py) - |Y_T| \Big(\frac{S_T}{2} - H_T + \epsilon \Big) \right] . \end{aligned}$$

Since all the jump intensities $(\lambda^S, \lambda_{1,2}^J, \lambda^a, \lambda^b)$ are finite constants, the term

$$\sup_{\theta} \mathbb{E}_t \left[(X_T - x) + (P_T Y_T - py) - |Y_T| \Big(\frac{S_T}{2} - H_T + \epsilon \Big) \right]$$

cannot be greater than the finite, maximum profit achievable through a combination of a market-making strategy that participates in every trade when market orders arrive (an upper bound on $X_T - x$, denoted by C_0) and a directional, frictionless market-taking strategy that bets on the mid-price jumps (an upper bound on $(P_T Y_T - py) - |Y_T|\Big(\frac{S_T}{2} - H_T + \epsilon\Big)$, denoted by C_1).

Proof of Lemma 3: As stated under (3.9) in the main paper, only the infinitesimal generators ($\mathcal{L}^P$, $\mathcal{L}^F$, $\mathcal{L}^S$) for the value function V and the value of $\mathbb{E}_t d[P,P]_t$ require explicit expressions. They are given below:

$$\begin{aligned}
\mathbb{E}_t d[P,P]_t &= \left(\frac{1}{4}\lambda_1^J + \lambda_2^J\right) dt \\
\mathcal{L}^P \circ V(t,x,y,p,f,s) &= \Big(V(t,x,y,p+\delta/2,f,s)\psi_1(f) + V(t,x,y,p-\delta/2,f,s)(1-\psi_1(f))\Big)\lambda_1^J dt \\
&\quad + \Big(V(t,x,y,p+\delta,f,s)\psi_2(f) + V(t,x,y,p-\delta,f,s)(1-\psi_2(f))\Big)\lambda_2^J dt \\
\mathcal{L}^F \circ V(t,x,y,p,f,s) &= V_f(\alpha_F f)dt + \frac{1}{2}V_{ff}\sigma_F^2 dt \\
\mathcal{L}^S \circ V(t,x,y,p,f,s) &= \Big(\sum_{j=1}^{3}\rho_{ij}[V(t,x,y,p,f,j\delta) - V(t,x,y,p,f,i\delta)]\Big)\lambda^S dt\,,
\end{aligned}$$

where V_f and V_{ff} are the first and second order partial derivatives of V against the state variable F. In addition, $\mathbb{E}_t dP_t = \Big(\lambda_1^J \frac{\delta}{2}(2\psi_1(F_t) - 1) + \lambda_2^J \delta(2\psi_2(F_t) - 1)\Big)dt$.

Proof of Proposition 1: To prove Proposition 1, I need to show that the value function V in (3.7) is the unique viscosity solution to (3.8) and (3.9). Since I have established the necessary growth (boundness) conditions on V that are shown in Lemma 2, the proposition – the existence and the uniqueness of the viscosity solution V (a.k.a. the value function) – is then a direct application of standard arguments and results from stochastic control theory, e.g. Seydel (2009a,b), or Øksendal and Sulem (2007), Chapter 9.

Proof of Lemma 4: It is clear that the quasi-variational inequality (3.10) and the terminal condition (3.11) are simplified versions of (3.8) and (3.9), if I can decompose the value function $V(t,x,y,p,f,s)$ as $x + py + \nu(t,y,f,s)$. The required decomposition can be established by first noting that the mid-price P has constant jump intensities and jump size distributions depending only on the state variable F, and then extending to my scenario the argument for a simpler case shown in Guilbaud and Pham (2013).

Proof of Proposition 2: I shall prove Proposition 2 by first establishing three properties

for my numerical scheme and then applying a power theorem from Barles and Souganidis (1991).

Lemma. *(Monotonicity)*
For any $\Delta_T > 0$ such that $\Delta_T \leq \left(\lambda^S + (\pi_1 + (1-\pi_1)\lambda^a) + (\pi_1 + (1-\pi_1)\lambda^b)\right)^{-1}$, the operator $\mathcal{A}$ defined in (d) of the numerical scheme is non-decreasing in ϕ, i.e.

$$\text{if } \phi < \tilde{\phi}, \text{ then } \mathcal{A}(t,y,f,s,\phi) \leq \mathcal{A}(t,y,f,s,\tilde{\phi}), \ \forall t,\ y,\ s \text{ and } f\,.$$

Proof. From the expression in (e) of the numerical scheme, it is clear that $g^a(f,s,\theta_t^{mk,b}) < \pi_1 + (1-\pi_1)\lambda^a,\ \forall f,\ s,\ \theta_t^{mk,b}$, and $g^b(f,s,\theta_t^{mk,a}) < \pi_1 + (1-\pi_1)\lambda^b,\ \forall f,\ s,\ \theta_t^{mk,a}$. Thus $1 - \Delta_T\lambda^S - \Delta_T g^a(f,s,\theta_t^{mk,b}) - \Delta_T g^b(f,s,\theta_t^{mk,a}) > 0$ as long as $\Delta_T \leq \left(\lambda^S + (\pi_1 + (1-\pi_1)\lambda^a) + (\pi_1 + (1-\pi_1)\lambda^b)\right)^{-1} < \left(\lambda^S + g^a(f,s,\theta_t^{mk,b}) + g^b(f,s,\theta_t^{mk,a})\right)^{-1}$, which implies that $\widetilde{\mathcal{L}}(t,y,f,s,\phi)$ is monotone in ϕ (as the sum of coefficients in front of ϕ is positive) and so is $\mathcal{A}(t,y,f,s,\phi)$.

Lemma. *(Stability)*
For any $\Delta_T, \Delta_Y, M_Y, \Delta_F, M_F > 0$, there exists a unique solution ϖ to (3.14)-(3.15), and the sequence $\{\varpi\}$ is uniformly bounded.

Proof. By the definition of the backward scheme (4.1)-(4.2), the solution ϖ exists and is unique. The uniform bound follows directly from the growth condition (lower and upper bounds) on the reduced-form value function ν (a modification on Lemma 2, i.e. the bounds on V).

Lemma. *(Consistency)*
The scheme (3.14)-(3.15) is consistent in the sense that, for all $(t,y,f) \in [0,T)\times\mathbb{R}\times\mathbb{R}$ and any smooth test function ϕ, as $(\Delta_T, \Delta_Y, M_Y, \Delta_F, M_F) \to (0,0,\infty,0,\infty)$, and $(t',y',f') \to (t,y,f)$, we have

$$\begin{aligned}
&\lim \frac{1}{\Delta_T}[\widetilde{\mathcal{L}}(t'+\Delta_T, y', f', s, \phi) - \phi(t', y', f', s)] \\
&= \frac{\partial\phi}{\partial t} + y\frac{\mathbb{E}_t dP_t}{dt} + \mathcal{L}^{\mathcal{F}}\circ\phi + \mathcal{L}^{\mathcal{S}}\circ\phi - \gamma y^2 \frac{\mathbb{E}_t d[P,P]_t}{dt} + \\
&\qquad \sup_{\theta^{mk}} \Big\{ g^a(f,s,\theta_t^{mk,b}) \cdot \big(\phi(t,y+1,f,s) - \phi(t,y,f,s) + s/2 - \delta\theta_t^{mk,b}\big) + \\
&\qquad\qquad g^b(f,s,\theta_t^{mk,a}) \cdot \big(\phi(t,y-1,f,s) - \phi(t,y,f,s) + s/2 - \delta\theta_t^{mk,a}\big) \Big\}
\end{aligned}$$

and

$$\lim \widetilde{\mathcal{M}} \circ \widetilde{\mathcal{L}}(t', y', f', s, \phi) = \sup_{\zeta} \left\{ \phi(t, y + \zeta, f, s) - |\zeta|(s/2 + \epsilon) + |y| \int_H H \, dG(H \mid s, f) \right\}$$

Proof. This follows from the result established in Section 6.1.2 of Chen and Forsyth (2008).

Proposition. *(Convergence) The solution ϖ to the numerical scheme (3.14)-(3.15) and the corresponding discretized policies converge locally uniformly, respectively, to the reduced-form value function ν and the optimal strategies on $[0, T] \times \mathbb{R} \times \mathbb{R}$ as $(\Delta_T, \Delta_Y, M_Y, \Delta_F, M_F) \to (0, 0, \infty, 0, \infty)$, $\forall s \in \mathcal{S}$.*

Proof. Given the properties of monotonicity, stability and consistency of the numerical scheme, this is a direct application of the result of Barles and Souganidis (1991).

References

Aït-Sahalia, Yacine, 1999, Transition densities for interest rate and other nonlinear diffusions, *Journal of Finance 54, 1361–1395.*

Aït-Sahalia, Yacine, and Per Mykland, 2003, The effects of random and discrete sampling when estimating continuous-time diffusions, *Econometrica 71, 483–549.*

Avellaneda, Marco, and Sasha Stoikov, 2008, High-frequency trading in a limit order book, *Quantitative Finance 8, 217–224.*

Barles, G., and P. Souganidis, 1991, Convergence of approximation schemes for fully nonlinear second order equations, *Asymptotic Analysis 4, 271–283.*

Baron, Matthew, Jonathan Brogaard, and Andrei Kirilenko, 2012, The trading profits of high frequency traders, *Working paper, University of Washington Foster School of Business.*

Brogaard, Jonathan, Björn Hagströmer, Lars L. Nordén, and Ryan Riodan, 2013a, Trading fast and slow: colocation and market quality, *Working paper, available at SSRN*

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riodan, 2013b, High frequency trading and price discovery, *Working paper, available at SSRN*

Buti, Sabrina, and Barbara Rindi, 2013, Undisclosed orders and optimal submission strategies in a limit order market, *Journal of Financial Economics 109, 797–812.*

Chen, Zhuliang, and Peter A. Forsyth, 2008, A numerical scheme for the impulse control formulation for pricing variable annuities with a guaranteed minimum withdrawal benefit (GMWB), *Nmerische Mathematik 109, 535–569.*

Cont, Rama, and Ekaterina Voltchkova, 2005, A finite difference scheme for option pricing in jump diffusion and exponential Lévy models, *Siam Journal of Numerical Analysis 43, 1596–1626.*

Crandall, Michael G., and Pierre-Louis Lions, 1983, Viscosity solutions of Hamilton-Jacobi equations, *Transactions of the American Mathematical Society 277, 1–42.*

Crandall, Michael G., Hitoshi Ishii, and Pierre-Louis Lions, 1992, User's guide to viscosity solutions of second order partial differential equations, *Bulletin of the American Mathematical Society 27, 1–67.*

Easley, David, Marcos M. López de Prado, and Maureen O'Hara, 2011, The Microstructure of the "Flash Crash": flow toxicity, liquidity crashes, and the probability of informed trading, *Journal of Portfolio Management 37, 118–128.*

Fleming, Wendell H., and Halil M. Soner, 2005, Controlled markov processes and viscosity solutions, second edition.

Gould, Martin D., Mason A. Porter, Stacy Williams, Mark McDonald, Daniel J. Fenn, and Sam D. Howison, 2013, Limit order books, *Working paper, available at Arxiv.*

Guilbaud, Fabien, and Huyên Pham, 2013, Optimal high-frequency trading with limit and market orders, *Quantitative Finance 13, 79–94.*

Hagströmer, Björn, and Lars L. Nordén, 2013, The diversity of high-frequency traders, *Journal of Financial Markets 16, 741–770.*

Hautsch, Nikolaus, and Ruihong Huang, 2011, Limit order flow, market impact and optimal order sizes: evidence from NASDAQ TotalView-ITCH data, *Working paper, available at SSRN*

Hautsch, Nikolaus, and Ruihong Huang, 2012, On the dark side of the market: Identifying and analyzing hidden order placements, *Working paper, available at SSRN*

Hasbrouck, Joel, and Gideon Saar, 2009, Technology and liquidity provision: The blurring of traditional definitions, *Journal of Financial Markets 12, 143–172.*

Hasbrouck, Joel, and Gideon Saar, 2013, Low-latency trading, *Journal of Financial Markets 16, 646–679.*

Hendershott, Terrence, and Ryan Riodan, 2013, Algorithmic trading and the market for liquidity, *Journal of Financial and Quantitative Analysis, forthcoming.*

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld, 2011, Does algorithmic trading improve liquidity? *Journal of Finance 66, 1–33.*

Hirschey, Nicholas, 2013, Do high-frequency traders anticipate buying and selling pressure?, *Working paper, available at SSRN*

Ho, Thomas, and Hans R. Stoll, 1981, Optimal dealer pricing under transactions and return uncertainty, *Journal of Financial Economics 9, 47–73.*

Karr, Alan, 1991, Point processes and their statistical inference, second edition.

Kirilenko, Andrei, Albert P. Kyle, Mehrdad Samadi, and Tugkan Tuzun, 2011, The flash crash: The impact of high frequency trading on an electronic market, *Technical report, available at SSRN*

Menkveld, Albert J., 2013, High-frequency trading and the *new market* makers, *Journal of Financial Markets 16, 712–740.*

Øksendal, Bernt, and Agnès Sulem, 2007, Applied stochastic control of jump diffusions, second edition.

Seydel, Roland C., 2009a, Impulse control for jump-diffusions: Viscosity solutions of quasi-variational inequalities and applications in bank risk management, *PhD thesis, University of Leipzig.*

Seydel, Roland C., 2009b, Existence and uniqueness of viscosity solutions for QVI associated with impulse control of jump-diffusions, *Stochastic Processes and their Applications 119, 3719–3748.*