High Frequency Market Making: Implications for Liquidity*

Yacine Aït-Sahalia†
Department of Economics
Bendheim Center for Finance
Princeton University
and NBER

Mehmet Sağlam‡
Department of Finance
Carl H. Lindner College of Business
University of Cincinnati

This Version: January 30, 2017

Abstract

We analyze the consequences for liquidity provision of competing market makers operating at high frequency. Competition increases overall liquidity and deters the fast market maker's use of order flow signals. Using various liquidity metrics, we find that the market maker provides more liquidity as he gets faster but shies away from it as volatility increases. We then provide a model-based analysis of the impact of four widely discussed policies designed to regulate high frequency trading: imposing a transactions tax, setting minimum-time limits before quotes can be cancelled, taxing the cancellations of limit orders, and replacing time priority with a pro rata or random allocation. We find that these policies are largely unable to induce high frequency market makers to provide liquidity that is robust across volatility events.

Keywords: Poisson Processes, Stochastic Optimal Control, High Frequency Trading, Market Making, Duopoly, Liquidity, Order Cancellations, Competition for Order Flow, Financial Market Regulation, Tobin Tax, Order Resting Time, Order Cancellation Tax, Pro Rata Allocation.

MSC 2000 Subject Classifications: Primary: 93E20; 91G80. Secondary: 60G55, 60J75.

*We are grateful to Jonathan Brogaard, Arseniy Kukanov, Costis Maglaras, Albert Menkveld, Ciamac Moallemi, Fabrice Rousseau and Duane Seppi for comments and suggestions, and to seminar and conference participants at UC San Diego, UC Riverside, University of Luxembourg, Harvard University, HEC Paris, MIT, University of Venice, University of Washington in Seattle, the Florence-Ritsumeikan Workshop on Stochastic Processes, the Humboldt-Copenhagen Conference, the AFFI Conference, the Banff High-Frequency Conference, the Okinawa Stochastic Processes Conference, MFA 2014, the NYU Five Star Conference and the AEA 2017 Annual Meetings. An earlier version of the paper circulated under the title "High Frequency Traders: Taking Advantage of Speed."

†Email: yacine@princeton.edu

‡Email: mehmet.saglam@uc.edu

1. Introduction

High-frequency traders (HFTs) have become a potent force in many markets, representing between 40 and 70% of the trading volume in US futures and equity markets, and slightly less in European, Canadian and Australian markets (see Biais and Woolley (2011)). HFTs develop and invest in a trading infrastructure designed to analyze a variety of trading signals and send orders to the marketplace in a fraction of a second. The potential profit from any single transaction resulting from an execution may be very tiny, and be achieved ex ante with a probability only slightly above 50%, but HFTs rely on this process being repeated thousands, if not more, times a day. As the law of large numbers and the central limit theorem relentlessly take their hold, profits ensue and presumably justify the HFTs' large investment in trading technology.

This significant amount of market activity has been accompanied by theoretical research addressing some of issues raised by the rapid development of high frequency trading. In particular, Foucault et al. (2016) extend Kyle's model by incorporating heterogeneity in the speed of information processing. Foucault et al. (2013) develop a model in which HFTs choose the speed at which to react to news, based on a trade-off between the advantages of trading first compared to the attention costs of following the news. Biais et al. (2015) analyze the arms race and equilibria arising in a model where traders choose whether to invest in fast trading technologies. Jovanovic and Menkveld (2010) study the effect of high frequency trading activity on welfare and adverse selection costs. Cvitanić and Kirilenko (2010) study the distribution of prices in a market before and after the introduction of HFTs. Other relevant papers include Pagnotta and Philippon (2011), Àlvaro Cartea and Penalva (2012), Jarrow and Protter (2012) and Moallemi and Sağlam (2013).

As the market structure and technology evolved, liquidity to financial markets is now primarily provided by high frequency trading firms as opposed to traditional market makers. We analyze in this paper the implications of this change for liquidity provision. For this purpose, we employ the model of a fully dynamically optimizing high frequency market maker developed in the companion paper Aït-Sahalia and Sağlam (2016), which we extend to allow for competition in the provision of liquidity. The model is in the tradition of inventory control problems, which originated in Amihud and Mendelson (1980) and Ho and Stoll (1981) for "traditional" market makers (see also Avellaneda and Stoikov (2008), Guilbaud and Pham (2013), Guéant et al. (2013), Àlvaro Cartea et al. (2014) and Hendershott and Menkveld (2014)). However, we depart from the classical models by endowing the high frequency market makers with effectively all the advantages: instead of market makers trying to protect themselves from a potentially better informed and faster order flow emanating from the exchange's customers, the market makers in our setup are both faster and better informed than the

customers, whom we call low frequency traders (LFTs).

The present paper provides several new results. First, using various liquidity metrics, we find that a faster market maker provides higher liquidity provision. This is largely consistent with the view that has emerged out of both the academic literature on HFTs and many public policy and industry analyses, namely that HFTs improve market quality by providing liquidity, contributing to price discovery, improving market efficiency and easing market fragmentation. This prediction of the model is compatible with the empirical findings in Hendershott et al. (2011), Hasbrouck and Saar (2013), Chaboud et al. (2010) and Menkveld (2013).

Second, we analyze the HFT's optimal provision of liquidity and equilibrium bid-ask spread as a function of fundamental price volatility, over which the HFT holds no informational advantage. One of the main theoretical predictions we make is that the HFT's liquidity provision is U-shaped as a function of volatility, first increasing as volatility attracts more LFTs, but then decreasing when price volatility increases beyond a certain level. Since this is precisely when large unexpected orders are likely to hit, markets can become fragile in volatile times, with imbalances arising because of inventories that intermediaries used to, but are no longer willing to temporarily hold. This U-shaped prediction of the model is consistent with the available empirical evidence. Brogaard et al. (2014) and Anand and Venkataraman (2016) find that HFT participation is higher on more volatile days. However, in extreme volatility events, such as flash crashes, Easley et al. (2011) and Kirilenko et al. (2010) show that some HFTs withdrew from their market-making roles.

Third, we show that the ability of HFTs to (imperfectly) predict the types of LFTs they face leads to a strategic widening of his quotes when an impatient trader is signaled to arrive, not unlike the ability of airlines to price-discriminate against business travelers. This results in a time-varying equilibrium bid-ask spread and is consistent with the common "unfairness" complaints of LFTs, whereby some orders are executed a penny away from the best prices despite what the trading screen suggested at the time the LFT submitted his order.

Fourth, we show that competition for order flow among HFTs results in splitting the rent extracted from LFTs, but that the overall liquidity provision increases and equilibrium bid-ask spreads decrease in the presence of competition, and LFTs tend to be better off. However, the HFT reduces his own liquidity provision to the market when compared to a monopoly situation. In the presence of competition, he may reduce his use of order flow signals to price discriminate against LFTs differing in their urgency to trade.

Finally, we provide the first formal, model-based, analysis of the impact of four widely discussed, and in some cases already implemented, HFT policies or regulations: imposing a transaction or Tobin tax on each trade; setting minimum rest times before limit orders can be cancelled; taxing the can-

cellations of limit orders; replacing price and time priority with a pro rata or random allocation. The built-in advantages of the HFTs in terms of speed and information prove hard to undo using these policies, and providing proper incentives to HFTs in terms of liquidity provision is also difficult. We find that both a transactions tax and a pro rata allocation scheme reduce the provision of liquidity. Imposing minimum rest times or cancellation taxes induces the HFT to quote more in low volatility environments, but then reduce his provision of liquidity when volatility is high. These policies lead to more liquidity when it is least needed, and less liquidity when it would be most needed. To summarize, the model predicts that these four policies, despite the varying degree of support they enjoy from regulators around the globe, are likely to be ineffective as far as making the provision of liquidity by HFTs more robust, or countercyclical, across the volatility cycle.

The paper is organized as follows. Section 2 develops the model of market making with queueing priority in the limit order book where the HFT is competing with a traditional market maker. Section 3 characterizes the optimal market making policy of the HFT. Section 4 develops the implications of the model for market structure, while Section 5 analyzes in the context of the model the impact of possible HFT policies. Section 6 concludes. Proofs and technical results are in the Appendix.

2. A Model of Market Making with Queueing Priority

We start by extending the model of high frequency market making with queueing priority based on Aït-Sahalia and Sağlam (2016) to include competition for order flow.

2.1. The Trading Environment

The market structure consists of a limit order book in which a single asset is traded in fixed lots of size 1. The asset transacts at discrete price levels around its fundamental value, X_t, with a tick size in the market $2C$ for some constant $C > 0$ (e.g., $C = \$0.005$). We assume that the asset's fundamental price, starting from a value X_0 which is a multiple of C, is subject over time to exogenous variability in the form of pure compound Poisson jumps with arrival rate σ and a Binomial distribution of jump sizes taking with equal probability values J and $-J$, where J is a multiple of C. Formally, we can write X_t as

$$X_t = X_0 + \sum_{i=1}^{N_t^\sigma} Z_i, \tag{2.1}$$

where Z_i are the independent jump sizes with zero expectation. Since the variance of X_t is proportional to σ, we will refer σ as the price volatility for brevity. The dynamics of X_t are illustrated in Figure 1.

There are two competing market makers: A traditional medium-frequency trader (MFT) and the

HFT. We assume that the MFT is nonstrategic and places limit orders at the best bid ($X_t - C$) and ask ($X_t + C$) at a Poisson rate of β.[1] When she places an order at either side, her order stays in the queue until it is met by an incoming market order. That is, the MFT never cancels an existing limit order. On the other side, the HFT is strategic and posts limit orders at the best bid ($X_t - C$) and ask ($X_t + C$) or one level removed from the best prices ($X_t - 3C$ and $X_t + 3C$). The HFT's trading technology determine his timing ability to alter his quotes (post new ones or cancel existing ones) upon the arrival of a Poisson process with arrival rate μ. Thus, the parameter μ is a measure of the HFT's speed.

We assume that both the MFT and the HFT offers at most one share to buy or sell. We use the following notation to denote the quoting decisions of the HFT and the MFT: $\ell_t^a = 1$ (resp. $\ell_t^b = 1$) will imply that the HFT has an active quote at the best ask (resp. bid) at time t, and similarly, $\ell_t^a = 2$ (resp. $\ell_t^b = 2$) will imply that the HFT has an active quote at the second-best ask (resp. bid) at time t. Finally, $\omega_t^a = 1$ (resp. $\omega_t^b = 1$) will imply that the MFT has an active quote at the best ask (resp. bid) at time t.

When the HFT decides to quote at the second-best prices, he obviously runs the risk that he will lose a transaction to the MFT even if the HFT had time priority, should the MFT places a limit order at the best price. Finally, $\ell_t^a = 0$ (resp. $\ell_t^b = 0$) means that the HFT either chooses not to quote to sell (resp. buy) or his most recent active order has been filled by a market order but the HFT has not yet submitted a new order due to the technological constraints. Figure 2 illustrates the possible quoting decisions by the HFT and the state of the order book at the bid side. The last two cases highlight the significance of the priority. If the MFT has a limit order in the best bid queue ($\omega^b = 1$) on top of the HFT (the sixth case), the HFT must wait for the MFT's order to be executed by a market order. Similarly, if the HFT has a limit order in the best bid queue ahead of MFT (the seventh case), the MFT will wait for the HFT's order to be either executed or cancelled by the HFT.

We assume that the limit order book contains exogenous depth supplied by competing liquidity providers at tick levels $X_t + 5C$ and/or $X_t - 5C$. This assumption is needed only to define a valid spread measure should the HFT completely withdraw from providing liquidity ($\ell_t^b = 0$ or $\ell_t^a = 0$). With this assumption, the bid-offer spread in the economy will be endogenously determined by the HFT's optimal quoting policy and can take values between $2C$, $4C$, $6C$, $8C$ and $10C$ for all 49 states of the order book.

The HFT's and the MFT's quotes interact with the incoming orders of LFTs. LFTs are of three types, characterized by a different willingness to wait to get a more desirable price: patient LFTs are only willing to buy (resp. sell) at price $X_t + C$ (resp. $X_t - C$). Impatient LFTs, on the other

[1] For increased complexity, the MFT can be also strategic but we expect the liquidity implications to be very similar.

hand, demand immediacy and are also willing to buy (resp. sell) at price $X_t + 3C$ (resp. $X_t - 3C$). Arbitrageurs are only willing to trade if they spot a particularly good deal, buying the asset below X_t or selling it above X_t. Patient and impatient traders and arbitrageurs arrive at random times according to Poisson processes with respective arrival rates λ_P, λ_I and λ_A; the sum of their arrivals produces the aggregate demand and supply functions λ^B and λ^S faced by the HFT. Figure 3 provides an illustration of these functions. In practice, if λ_A is sufficiently high, then arbitrageurs, although formally classified as LFTs here, act more like high frequency traders themselves who are usually referred as high frequency bandits (Menkveld (2016)): with high probability, someone will quickly take liquidity from the HFT by buying at a price Y below X and selling at a price Y above X.

The HFT has an informational advantage in the sense that he receives, upon the arrival of a Poisson process with rate θ, a pair of signals that are indicative of the direction of the next incoming market order (buy or sell), and about the type of the trader submitting the market order (patient or impatient). The two signals are i.i.d. Bernoulli random variable, $S^{\text{dir}} \in \{B, S\}$, imperfectly predicting the direction of the next incoming market order: B corresponds to a LFT order to buy and S refers to a LFT order to sell, and $S^{\text{type}} \in \{P, I\}$, with P and I refer to patient and impatient LFTs respectively. The accuracy level of the two signals are p and q respectively, both in $[1/2, 1)$. Each previous signal is cancelled by the arrival of either a new signal or a LFT order. The HFT receives no signals regarding the exogenous jumps in the asset's fundamental value and consequently cannot predict the arrival of the arbitrageurs either.

2.2. Objective Function of the HFT

The strategic HFT is risk-neutral, but controls his risk by penalizing inventory at a rate of $\Gamma|x_t|$ where Γ is a constant parameter of inventory aversion and x_t denotes his inventory. In practice, limiting or penalizing inventory is one of the primary sources of risk mitigation by HFTs.

The HFT's objective is to maximize his expected discounted rewards earned from transacting against the incoming order flow from LFTs, which earns him the bid-ask spread, and the potential penalty costs from holding an inventory. The HFT is also exposed to adverse selection costs when his stale quote gets picked primarily by arbitrageurs. Figure 1 illustrates how the HFT's quotes can become stale after a price jump. When the fundamental price of the security jumps up by $J = 4C$, the HFT's earlier quote at the pre-jump $X_t + C$ becomes stale and leads to an arbitrage opportunity for LFTs. Patient, impatient and as well as arbitrageurs would certainly like to buy the asset at the post-jump $X_t - 3C$. If they submit an order during this stale-quote period, a trade occurs, and the HFT would lose $3C$ to the LFT that submitted the order.

Let π denote any feasible policy that chooses ℓ^b and ℓ^a at decision times, T_k^q, and T_i^a be the ith sell

order submitted by LFT type $y_i^a \in \{P, I\}$ and T_j^b is the jth market buy order submitted by the LFT type $y_j^b \in \{P, I\}$ where i,j and k are positive integers. To track the most recent decision time by the HFT before the arrival of market orders, we define $\tau_i \equiv \max\{k : T_k^q < T_i^a\}$ and $\tau_j \equiv \max\{k : T_k^q < T_j^b\}$. Finally, let $f_i^b \in \{0, 1\}$ and $f_j^a \in \{0, 1\}$ be binary variables that show whether the HFT's order has execution priority at the arrival of market orders.

There are three potential outcomes in terms of the HFT's reward function when an LFT submits a buy or sell order. By symmetry, we focus on the bid side of the HFT's quotes, i.e., the ith sell order. The first case refers to a zero payoff due to no trade. If the HFT is not quoting at the bid side ($\ell_{\tau_i}^b = 0$), or the HFT's limit order does not have execution priority ($f_i^b = 0$) or the fundamental price has a positive jump after the most recent decision time ($X_{T_i^a} - X_{\tau_i} > 3C$), there will be no trade. If a patient LFT sends the ith sell order ($y_i^a = P$) and the HFT's quote is at the second-best bid ($\ell_{\tau_i}^b = 2$), then there is again no trade.

In the second case, the HFT gains C by trading with either a patient or an impatient LFT ($y_i^a \in \{P, I\}$) when he is quoting at the best bid ($\ell_{\tau_i}^b = 1$), his limit order has execution priority ($f_i^b = 1$) and the fundamental price has not changed since the HFT's most recent decision time ($X_{T_i^a} = X_{\tau_i}$). If there has been a negative jump during this period, the HFT would lose the jump amount mJ despite his spread gain where $X_{T_i^a} - X_{\tau_i} = -mJ$ with $m = 0, 1, 2 \ldots$.

In the third case, the HFT gains $3C$ if the order is submitted by an impatient LFT ($y_i^a = I$) when he is quoting at the second-best bid ($\ell_{\tau_i}^b = 2$), his limit order has execution priority ($f_i^b = 1$) and the fundamental price has not changed since the HFT's most recent decision time ($X_{T_i^a} = X_{\tau_i}$). If there has been a negative jump during this period, the HFT would again lose the jump amount mJ ($X_{T_i^a} - X_{\tau_i} = -mJ$).

We can summarize these cases with the following HFT gains function from LFTs' sell orders:

$$G^-(\ell, y, f, X_T, X) = \begin{cases} C - mJ & \text{if } \ell = 1, X_T - X = -mJ, f = 1 \\ 3C - mJ & \text{if } \ell = 2, y = I, X_T - X = -mJ, f = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

Similarly, we can obtain the symmetric $G^+(\ell, y, f, X_T, X)$ function to accommodate the gains from LFTs' buy orders. Putting it all together, the HFT has the following formal objective function:

$$\max_{\pi} \mathbb{E}^{\pi} \left[\sum_{i=1}^{\infty} e^{-DT_i^a} G^-(\ell_{\tau_i}^b, y_i^a, f_i^b, X_{T_i^a}, X_{\tau_i}) + \sum_{j=1}^{\infty} e^{-DT_j^b} G^+(\ell_{\tau_j}^a, y_j^b, f_j^a, X_{T_j^b}, X_{\tau_j}) - \Gamma \int_0^{\infty} e^{-Dt} |x_t| dt \right] \tag{2.3}$$

The first (resp. second) term is the HFT's gain from an incoming sell (resp. buy) order crossed against his existing limit order net of adverse selection costs due to stale quotes. The last term captures the HFT's inventory penalty costs.

3. Solution of the Model

The key advantage of the model's Poisson-based setup is that we can convert the existing continuous-time model to a discrete-time Markov Decision Process (MDP) by considering the arrival times of the LFTs' orders, signals, MFT's quotes and the HFT's decision time as the starting time of the new period. This effectively merges the different Poisson clocks into a single chronologically-ordered one (with arrival rate equal to the sum of the individual ones), and then performing a time change from the (random) Poisson clock to the corresponding discrete-time event clock.

3.1. Action and State Space

The state space after the discrete-time transformation is (x, s, b, a, e, j) where $x \in \{\ldots, -1, 0, 1, \ldots\}$ denotes the current holdings or inventory of the HFT. The second state variable consists of the order direction and LFT type signals, $s = \left(S^{\text{dir}}, S^{\text{type}}\right) \in \{BP, SP, BI, SI\}$. The next two states show the available active quotes at the bid side and the ask side. Each side of the limit order book, bid or ask queues, can be in one of seven states: $b = 00$, denoting that the bid queue is empty, with neither HFT nor MFT currently quoting; $b = 10$, denoting that the HFT is quoting at the best bid and MFT is not quoting; $b = 20$, denoting that the HFT is quoting at the second-best bid and MFT is not quoting; $b = 21$, denoting that the MFT is quoting at the best bid and the HFT is quoting at the second-best bid, but the MFT's order has price priority; $b = 11m$, denoting that both the MFT and the HFT are quoting at the best bid, but the MFT's order has time priority; $b = 11h$, denoting that both the MFT and the HFT are quoting at the best bid, but the HFT's order has time priority; $b = 01$, denoting that there is only an order by the MFT at the best bid, which has priority. The set of values in a is also similarly defined with another 7 potential values. Thus, the order book can be in one of 49 states.

Recall that the HFT can only make quoting decisions after μ-events in the merged Poisson clock. In the remaining Poisson arrivals, the HFT does not have the ability to change his existing quotes but can only maintain his quotes initiated at the most recent decision time. The fifth state variable, $e \in \{0, 1\}$, is therefore a binary state variable denoting when $e = 1$ that the discrete date in the merged clock corresponds to a $\mu-$event , in which case the HFT can revise his quotes; whereas $e = 0$ refers to the arrival of either a LFT order (λ), a signal event (θ) or a jump event (σ), in which cases the HFT cannot revise his quotes in this state. Finally, The sixth and last state variable, j, keeps

track of the jumps realized in the asset fundamental value since the HFT's last quoting action, with $j \in \{\ldots, -1, 0, 1, \ldots\}$.

The action taken by the HFT at each state is whether to quote at the second-best or best available prices, i.e., $d^b(x, s, b, a, e, j) \in \{0, 1, 2\}$ and $d^a(x, s, b, a, e, j) \in \{0, 1, 2\}$. When $e = 0$, the corresponding action taken by the HFT is determined by l so we can, as in the simplified model, consider these states as fake decision epochs that force the HFT to continue with existing active quotes. For example, if $b = 11m$, and $a = 20$, then $d^{b*}(x, s, 11m, 20, 0, j) = 1$ and $d^{a*}(x, s, 11m, 20, 0, j) = 2$. Thus, the relevant state for the HFT is when $e = 1$ and in this case the existing active quotes given by b and a are no longer binding. At this state, j also reverts to 0 as the HFT can now peg his quotes around the new fundamental price. For this purpose, we can suppress the last two states when we refer to the optimal market-making policy using $\ell^{b*}(x, s, b, a) = d^{b*}(x, s, b, a, 1, 0)$ and $\ell^{a*}(x, s, b, a) = d^{a*}(x, s, b, a, 1, 0)$.

We can reduce the state space for decision epochs further considering the fact that the existing quotes in the order book submitted by the HFT are irrelevant if they have no execution priority. So out of 7 possible states in each side of the book, the HFT needs to know whether the MFT is quoting or not and whether her orders have execution priority. These three cases are identified in the following subsets of the order book at each side: We let $l_b = 0$ ($l_a = 0$) to denote the case in which the MFT does not quote, i.e., $b = 00$, $b = 10$ and $b = 20$ ($a = 00$, $a = 10$ and $a = 20$). Similarly, we let $l_b = 1$ ($l_a = 1$ if the MFT quotes but the HFT has the execution priority, i.e., $b = 11h$ ($a = 11h$). Finally, we let $l_b = 2$ ($l_a = 2$) if $b = 01$, $b = 11m$ and $b = 21$ ($a = 01$, $a = 11m$ and $a = 21$) i.e., MFT quotes with execution priority. Finally, we let $l_b = 2$ ($l_a = 2$ if the MFT quotes but the HFT has the execution priority, i.e., $b = 11h$ ($a = 11h$). Thus, only 9 different cases are relevant for the HFT's quoting decisions from the complete 49 states of the order book. These states can be identified by $l \equiv l_b \times l_a$. Using this reduction, we will use $\ell^{b*}(x, s, l)$ and $\ell^{a*}(x, s, l)$ to characterize the optimal market making policy.

3.2. Optimal Market Making

We use two value functions to differentiate between quoting epochs from the rest. Let $V(x, s, b, a, e, j)$ denote the value function of the HFT. We can suppress some of the states in the case of $e = 1$ with $h(x, s, b, a) = V(x, s, b, a, 1, 0)$, as the HFT can revise his quotes at this time. Using the reduction in states in decision epochs, we can further concisely express this value function by differentiating between the state of the best bid or the ask. Using our earlier definition of $l \equiv l_b \times l_a$, the value function at decision epochs can be written as $h(x, s, l)$ where l is in $\{00, 01, 02, \ldots, 20, 21, 22\}$. Finally, we let $v(x, s, b, a, j)$ be the value function in the no-decision event types, i.e., a LFT order, signal, volatility or the MFT quoting events (i.e., when $e = 0$). In the Appendix, we derive the HJB equations that

the optimal value functions need to satisfy for optimality.

We can now characterize the optimal quoting policy of the HFT, showing that, it is based on asymmetric thresholds as a function of inventory, signal and the MFT's relevant quoting states. We provide the optimal quoting policy for the bid side of the market as the other case is symmetric.

Theorem 1. *The optimal quoting policy π^* of the HFT consists in quoting at the available bid and the ask prices according to a threshold policy, i.e., there exists state-dependent limits for both the bid and the ask side such that the optimal policy is monotonic with respect to inventory:*

$$\ell^{b*}(x, SP, l) = \begin{cases} 1 & x \leq U^1_{P,l} \\ 2 & U^1_{P,l} < x < U^0_{P,l} \\ 0 & x \geq U^0_{P,l} \end{cases}, \quad \ell^{b*}(x, BP, l) = \begin{cases} 1 & x \leq L^1_{P,l} \\ 2 & L^1_{P,l} < x < L^0_{P,l} \\ 0 & x \geq L^0_{P,l} \end{cases},$$

$$\ell^{b*}(x, SI, l) = \begin{cases} 1 & x \leq U^1_{I,l} \\ 2 & U^1_{I,l} < x < U^0_{I,l} \\ 0 & x \geq U^0_{I,l} \end{cases}, \quad \ell^{b*}(x, BI, l) = \begin{cases} 1 & x \leq L^1_{I,l} \\ 2 & L^1_{I,l} < x < L^0_{I,l} \\ 0 & x \geq L^0_{I,l} \end{cases}.$$

where l is in $\{00, 01, 02, 10, 11, 12, 20, 21, 22\}$.

The optimal quoting policy in the complete model can be interpreted as follows. The limits $U^0_{P,l}$ and $U^0_{I,l}$ are the relevant limits for the HFT to stop quoting for inventory considerations when the direction signal is aligned with the direction of quoting. The limits $L^0_{P,l}$ and $L^0_{I,l}$ are the relevant limits for the HFT to stop quoting for inventory considerations when the direction signal is opposite to the direction of quoting. These limits depend on the anticipated type of the LFT (patient or impatient) and the MFT's quotes and priority in the order book.

The additional set of limits, $U^1_{P,l}$, $U^1_{I,l}$, $L^1_{P,l}$ and $U^1_{I,l}$, that lets the HFT to decide between quoting at the best bid or at the second best bid. These limits are driven by the economic consideration based on the likelihood of trade given the accuracy of the signals and the relative ratio of patient LFTs to impatient LFTs, and the additional gain from transacting away from the best prices with an impatient LFT and the MFT's existing quotes in the order book.

Note that the presence of a monotonic structure in the optimal policy can greatly reduce the computational effort by using a structured algorithm to solve the MDP (see e.g., monotone policy iteration algorithm in Puterman (1994)).

3.3. *Example*

For illustrative purposes, we use the following realistic parameter values: the HFT will be able to make decisions in every 100 milliseconds which implies that $\mu = 600$ per minute; $C = \$0.005$ which makes the tick-size and the minimum spread to be \$0.01; the arrival rate of impatient LFTs on each side of the market is set to be $\lambda_I = 7.5$ per minute while the arrival rate of patient LFTs is given $\lambda_P = 22.5$ per minute which then implies that total arrival rate of LFTs on both sides of the market is 1 order per second; the arrival rate of arbitrageurs is given by, $\lambda_A = 300$ per minute on each side, which implies the same rate of arrivals as the HFT's decision events in aggregate. The discount rate is 10% per year which corresponds to roughly 10^{-6} per minute. For the accuracy of the signals, we set $p = 0.7$ and $q = 0.6$, i.e., the signal will predict the correct sign of the next market order with 70% chance and the type of the LFT submitting the next order with 60% chance. The signal will be subject to change $\theta = 30$ times per minute on average. The fundamental price will be subject to a jump occurring 10 times per minute, i.e., $\sigma = 10$. Each jump will be in the amount \$0.04 in either positive or negative direction with equal probability. Given $\mu \gg \sigma$, we use a truncated state space for the number of jumps since the HFT's last quoting action with letting $j \in \{-1, 0, 1\}$. We choose $\beta = 60$ for the speed of the MFT. Finally, $\Gamma = 0.05$ so that HFT is paying \$0.05 per minute for each non-zero inventory he is holding. This is of course a high value, necessary only to obtain tight inventory limits in order to illustrate the role these limits play.

With these parameters, the quoting limits are computed as:

	MFT States (l)								
Limits	00	01	02	10	11	12	20	21	22
$U^1_{P,l}$	0	0	0	0	0	0	0	0	-1
$U^0_{P,l}$	3	3	3	1	1	1	1	1	0
$U^1_{I,l}$	0	0	-1	0	0	0	0	0	-1
$U^0_{I,l}$	4	4	4	1	1	1	1	1	0
$L^1_{P,l}$	0	0	0	1	0	0	-1	-1	-1
$L^0_{P,l}$	2	2	2	2	1	1	0	0	0
$L^1_{I,l}$	0	0	0	1	0	0	-1	-1	-1
$L^0_{I,l}$	3	3	3	2	1	1	0	0	0

(3.1)

This example cleanly shows the impact of competition and execution priority on the HFT's order aggressiveness. Recall that U^0 and L^0 are the limits for the HFT to stop quoting to buy the asset for

inventory considerations. We observe that these limits decrease as soon as the MFT has a quote at the bid side (see e.g., $U^0_{P,l}$). The limits may drop further if the MFT's order has execution priority (see e.g., $L^0_{P,l}$). Interestingly, if the MFT has an order with execution priority at the ask side, this may also affect the HFT's quoting decision at the bid side. This is intuitive as the HFT may not offload a potential inventory quickly in the presence of existing MFT quotes at the ask side.

We can also compare the limits to those of the case without competition, i.e., $\beta = 0$. In this case, the quoting limits are computed as $U^1_P = 0$, $U^0_P = 4$, $U^1_I = -1$, $U^0_I = 5$, and $L^1_P = 0$, $L^0_P = 2$, $L^1_I = -1$, $L^0_I = 4$. Note that since there is no MFT quote, l drops from being a state variable. Comparing the limits for U^0 and L^0, we observe that in the presence of competition, the HFT's inventory limits tighten. The difference between U^1 and U^0 or L^1 and L^0 also imply that the HFT's ability to price discriminate via quoting at second-best prices is lower in the presence of competition. We examine these implications further in Section 4.4 by comparing the long-run expected bid-offer spread in both regimes.

4. Implications of the Model for Market Structure

In this section, we consider the implications of the model for the HFT's liquidity provision, cancellation or modification of orders, and how liquidity provision is affected by fundamental price volatility. We use the same parameter values given in Section 3.3. We use a monopoly model by setting $\beta = 0$ in Sections 4.1, 4.2 and 4.3, and a duopoly model by setting $\beta = 60$ in Section 4.4.

We first compute the optimal inventory limits and characterize the optimal policy of the trader according to Theorem 1. Under this optimal policy, the inventory of the trader will be in a finite set, $[-N^*, N^*]$. Therefore, under the optimal policy, the model is governed by a finite-state Markov Chain. Let P_{opt} be the probability transition matrix defined on this finite state space under the optimal policy. Since the Markov Chain is aperiodic and irreducible, a stationary distribution π exists for this Markov Chain, which solves $\pi P_{\mathsf{opt}} = \pi$.

We then compute the long-run liquidity metrics resulting from the equilibrium between the HFT's (or HFT's plus MFT's) quotes and the LFTs' orders under the Markov stationary distribution, π. We focus on the following three liquidity measures: the expected bid-offer spread, the fill rate of LFT orders and the probability that the HFT will quote at the best bid and ask prices, i.e., the probability of a one-tick market, which given the fixed quantity of one in the model captures the market depth provided by the HFT at the best prices.

4.1. HFT's Liquidity Provision and HFT's Speed

Does the HFT's speed have a positive impact on the market liquidity? Figure 5 illustrates that as the HFT gets faster (μ), market liquidity as measured by the three liquidity metrics is higher. This result is consistent with the empirical literature on HFT, which tends to suggest that HFTs have had a generally positive effect on spreads and depths (see, e.g., Hendershott et al. (2011), Hasbrouck and Saar (2013), Chaboud et al. (2010) and Menkveld (2013).)

In the model, as the HFT becomes faster, the various risks he is facing (getting caught with an out-of-date signal, getting caught with stale quotes, etc.) are reduced. Furthermore, getting faster also improves the likelihood that the HFT will be able to exploit signals that allow him to price-discriminate. Figure 6 illustrates that if the HFT's second signal accuracy (q) improves, then the average bid-offer spread in long-run equilibrium increases. This reflects the fact that the HFT is better able to price discriminate among types of LFTs and exploits this ability to engage in predatory trading against impatient LFTs. This result is consistent with the often-expressed complaints from LFTs regarding "phantom or fleeting liquidity", i.e., liquidity provided at the best bid or ask prices that suddenly disappears, leading to an execution one tick away from the best prices.

4.2. Quote Cancellations by the HFT

Limit order cancellations are widely observed empirically in high frequency data. Hasbrouck and Saar (2009) note that over one third of limit orders are cancelled within two seconds and term those "fleeting orders". Baruch and Glosten (2013) show that quote cancellations can emerge as an equilibrium strategy in a trading game. Angel et al. (2015) document that the ratio of quotes to trades, which was relatively stable at about 2 between 1993 and 2001, started increasing and is over 25 by 2013. Our model makes predictions that are consistent with this empirical fact: a high number of cancellations will occur as the HFT changes his quotes in responses to changes in the fundamental value, his inventory and the signals he receives.

Figure 7 and Figure 8 illustrate the HFT's quote revisions and cancellations while following his optimal policy (recall (3.1)). Since the optimal limits may be different for each possible signal, endogenous quote cancellations may occur for instance when the HFT receives a different signal than the one that was previously in effect. For example, when his inventory is zero or close to it, the HFT makes his quoting decision depending upon the signal predicting the type of incoming LFT. Figure 7 shows that if the HFT is expecting a patient LFT, he quotes at the best bid and ask but if the signal suggests that an impatient LFT is next to arrive, he quotes at the wider price levels. From the perspective of the LFTs, these lead to cancellations and a widening of the spread. Similarly, the signal

providing information about the likely direction of the next LFT order may induce the HFT to cancel his quotes to minimize the probability of a stale quote. The left panel in Figure 8 illustrates that if the signal changes to BP when the HFT is expecting to buy from a patient LFT, the HFT decides to cancel his quote at the second-best bid in order to lower the risk of leaving a stale quote. Similar withdrawal of an ask quote happens in the right panel of Figure 8.

4.3. Implications of Asset Price Volatility for the Provision of Liquidity

One fundamental issue regarding the HFT's provision of liquidity concerns not its quantity but its "quality". Possible definitions of that quality vary, but most include the notion that this liquidity is to be provided in a stable manner over time and over different market environments, consistent with the requirement to provide a "fair and orderly market" that is imposed on regulated specialists and market makers. Are unregulated HFTs fair weather liquidity providers ready to provide plenty of liquidity when the market is calm and doesn't really need it, only to remove it whenever the market becomes turbulent (and it would be needed)? This question is of central importance for market stability, and to understand the potential for systemic risk should HFTs suddenly suspend their provision of liquidity in response to a market shock, contributing to an amplification of that shock.

We quantify the change in the equilibrium bid-offer spread as a function of asset price volatility, measured by σ: see Figure 9. We find that the rate of liquidity provision by the HFT decreases as the fundamental price becomes more volatile, resulting in a higher spread in equilibrium. The model predicts that the HFT protects himself against unanticipated price jumps (over which he has no informational advantage) by quoting less frequently. The conclusion that it is optimal for the HFT to decrease his provision of liquidity when the price volatility risk increases has obvious consequences for the way markets can be expected to operate in times of stress, and for the potential need to regulate the provision of liquidity by market makers.

On the other hand, since λ^B and λ^S described in Section 2 aggregate the demand for and supply of the asset emanating from many different traders, who are likely to have heterogenous beliefs, trading motives and propensities, it is possible and even plausible for λ^B and λ^S to depend on the fundamental price volatility σ; if volatility leads to a higher intensity of trading, then $\partial\lambda^B/\partial\sigma \geq 0$ and $\partial\lambda^S/\partial\sigma \geq 0$. The resulting effect on the HFT's optimal quoting, and hence the equilibrium spread, is illustrated in Figure 9 in the case where $\lambda^B(Y-X,\sigma)$ and $\lambda^S(Y-X,\sigma)$ are both linearly increasing in σ, corresponding to a linear positive correlation between volatility and LFTs' supply and demand. Other cases are possible depending upon the assumptions made outside the model regarding the LFTs' motives for trading (portfolio rebalancing, limits reached, directional bets, etc.) that ultimately determine the dependence of λ^B and λ^S on σ.

Any positive dependence of λ^B and λ^S on σ makes it more attractive for the HFT to continue quoting, and in some cases increase his quoting in times of market stress, although as we see in Figure 9 the increase in the provision of liquidity is short-lived: as σ begins to increase, the HFT starts by providing more liquidity, resulting in a lower spread, but as σ keeps increasing the HFT starts to withdraw liquidity. The HFT's objective value is increasing, due both to the increase in trading opportunities with LFTs owing to the increased volatility, and to the higher profitability of each trade even when the HFT quotes less.

This U-shaped dependence of the equilibrium spread (or inverse U-shaped dependence of liquidity) on volatility is consistent with recent empirical evidence. Brogaard et al. (2014) and Anand and Venkataraman (2016) find that HFT participation is higher on more volatile days. However, in the presence of extreme volatility events, such as flash crashes, Easley et al. (2011) and Kirilenko et al. (2010) show that some HFTs may withdraw from their market-making roles.

4.4. Competition for Order Flow and Liquidity Provision

We next investigate the impact on liquidity of competition between the HFT and MFT, as they queue in the limit order book. We quantify this impact by comparing the equilibrium spread in the presence ($\beta = 60$) and absence ($\beta = 0$) of competition.

Figure 10 illustrates the impact of queueing on the HFT's value due to the presence of the MFT, as a function of the HFT speed in the monopoly and duopoly models. In the presence of the MFT, the HFT splits the rent by losing some of his profits to the MFT, as some of the LFTs' orders are now executed by the MFT. Comparing the slopes in the value function in both regimes, we see that lower speed is more detrimental when the HFT is extracting monopolistic rents in the absence of the MFT: with the MFT present, not being as fast is less costly for the HFT. This suggests that with competition, the HFT's incentive to increase his speed is lower.

We next study how the presence of the MFT affects market liquidity. Figure 11 shows that LFTs are better off when market makers compete, compared to the monopolistic HFT situation: the equilibrium bid-offer spread decreases. In the duopoly case, the HFT quotes less: after making a trade, either a sale or a buy, the HFT is now less likely to trade in the opposite side of the market immediately due to the presence of the MFT. Although the HFT provides less liquidity than he did as a monopolist, the total provision of liquidity facing the LFTs is higher in the duopoly situation. Also, competition with the MFT decreases the HFT's ability to price-discriminate, since removing his quotes at the best bid and ask prices to quote one tick away from them causes the HFT's to potentially lose his time priority in the best bid and ask queues to the MFT. This induces the HFT to quote more often at the best bid and ask, reducing the equilibrium spread and increasing the probability of a one-tick market.

5. Implications of The Model for HFT Regulations

Regulators worldwide are paying increasing attention to the impact of HFTs on the markets they supervise, and have started debating, developing and in some cases implementing policies designed primarily to limit any negative consequences stemming from the rise of HFTs. The basic premise is that markets should remain platforms to trade risk among investors with different needs and beliefs, and not become primarily race tracks with the potential to ultimately discourage slower users from participating, due to the perception of a lack of fairness.

In this section, we use the model to make predictions about the impact of some frequently proposed, and in some cases already enacted, regulatory policies. We analyze their impact on the HFT's objective value and provision of liquidity to the market. Although welcome, additional liquidity in good times when markets are already very liquid may not be particularly useful, and may even be counterproductive if it lulls agents into complacency only to dry up when markets experience volatility. So we view as a desirable outcome of a policy a reduced provision of liquidity by HFTs in good (low volatility) times, in exchange for an increased provision of liquidity in bad (high volatility) times. Such a countercyclical impact would be desirable for some of the same reasons that advocate for banks' capital requirements to increase in booms, in order to limit the provision of speculative credit, and decrease in recessions, in order to speed up the economic recovery.

We specifically examine the impact of four widely discussed HFT policies through the prism of our model: imposing a transaction tax on each trade, setting minimum-time limits before orders can be cancelled, taxing the cancellations of limit orders, and replacing time priority with a pro rata allocation of shares to all quotes in the book at that price level.

These policies, or combinations thereof, capture the main elements that have been proposed in various countries, and in some cases already implemented. In 2012, France introduced a 0.2% tax on transactions in large stocks, and a 0.01% tax on HFTs penalizing them for a high rate of order cancellations within a half-second[2],[3]. Similarly, in Italy, a tax of 0.02% on orders issued and then cancelled within half a second, once above a threshold, has been introduced[4]. The Deutsche Börse introduced a tax in 2012 that charges HFTs for high "order-to-trade" ratios[5] as does the London Stock Exchange[6]. Norwegian regulators too consider taxing traders who submit a large number of orders relative to their actual executions[7]. The CME Group, the world's largest futures exchange, has had for

[2] "Paris traders brace for financial transactions tax", Reuters, July 31, 2012.

[3] Somewhat predictably, many investors in France have avoided the tax by trading "contracts for difference" which allow them to profit from an asset's gain or loss without actuallys owning the shares.

[4] "All eyes on Italy's high-frequency rules", The Financial Times, February 19, 2013.

[5] "D Börse to charge for 'stupid algos'", The Financial Times, February 28, 2012.

[6] "Bourses play nice cop to head off speed-trade rules", Reuters, April 10, 2012.

[7] "Oslo Bors to charge for excessive orders", The Financial Times, May 24, 2012.

a number of years message volume caps, designed to prevent excessive numbers of orders from being placed[8], while Nasdaq and DirectEdge, two of the largest US stock exchanges have introduced fines to discourage excessive order placement[9]. Canadian regulators too began increasing the fees charged to HFTs that flood the market with orders[10], while Indian regulators are studying ways to curb HFTs[11]. On the other hand, Brazil appears more open to the influx of HFTs[12]. Australian regulators want HFTs to implement a "kill switch" to prevent future flash crashes, and are considering a tax charge, although they appear to take a more benevolent view of HFTs than some of their counterparts in Europe[13].

In January 2013, European Union finance ministers approved a transaction tax in Germany, France, Italy, Spain and seven other Eurozone countries[14]; the UK, concerned about the impact on the City, is opposed[15]. It seems unlikely at present that the initially far-reaching package will get implemented as proposed, if ever[16]. The German government has advanced legislation that would, among other things, force HFTs to register as such with the government[17] and limit their ability to rapidly place and cancel orders[18]. The European Parliament has voted to require HFTs to honor the quotes they submit for at least half a second; imposes a minimum half-second delay on executing orders in a bid; possible use of circuit breakers to interrupt a sudden market plunge; and fee structures that would discourage excessive algorithmic trades[19]. These rules could potentially apply to all 27 member states of the European Union if governments were to give their approval. In the US, the SEC and CFTC are discussing similar kinds of regulatory actions[20], while transaction tax legislation has been introduced in the Senate, although with little prospects of passage. Not surprisingly, many trade associations representing trading firms are opposing these proposals[21]. Other alternative policies involve bunching together incoming orders every few milliseconds, or randomizing their allocation ("scrambling"), so a HFT would face queuing risk, as well as a minimum rest time before a cancellation. EBS, one of the major trading platforms in the foreign exchange market, has discussed such a proposal with its users.

[8] CME Messaging Efficiency Program: http://www.cmegroup.com/globex/resources/cme-globex-messaging-efficiency-program.html

[9] "US bourses to fine HFT data-cloggers", The Financial Times, March 7, 2012.

[10] "Canada's 'hot' traders attract regulatory heat", The Financial Times, October 16, 2012.

[11] "India takes steps to rein in algos", The Financial Times, May 22, 2013.

[12] "Despite Risks, Brazil Courts the Millisecond Investor", The New York Times, May 22, 2013.

[13] "Australia finds HFT fears 'overstated'", The Financial Times, March 18, 2013.

[14] "Brussels proposes 30bn 'Tobin tax'", The Financial Times, February 14, 2013.

[15] "Britain challenges EU over 'Tobin tax'", The Financial Times, April 19, 2013.

[16] "Europe plans major scaling back of financial trading tax", Reuters, May 30, 2013.

[17] "Berlin forges ahead with trading controls", The Financial Times, September 25, 2012.

[18] "German Bundestag Passes Bill to Regulate High-frequency Trading", The Wall Street Journal, February 28, 2013.

[19] "EU Lawmakers Call for Enforced Delay on High-Frequency Trades", The Wall Street Journal, September 26, 2012.

[20] "High speed trading a stiff challenge for U.S. regulators", Reuters, May 19, 2013.

[21] "Traders see Europe's Tobin tax hurting savers", Reuters, October 11, 2012 and "Trading Clamps Spur Lobby Effort", The Wall Street Journal, March 24, 2013.

Pro rata matching is employed for instance at the Chicago Mercantile Exchange for Eurodollar futures contracts as an alternative to time priority. The objective is to reduce the value of being at the top of the queue for execution, an objective shared with the proposal for batch auctions (see, e.g., Budish et al. (2015)).

In all cases, we assume that the HFT continues to optimize taking into account the full regulatory environment, and we extend the model accordingly. We use the same parameter values as in Section 3.3 to facilitate the comparison with earlier applications. We use the monopoly model by setting $\beta = 0$ in Sections 5.1, 5.2 and 5.3, and the duopoly model by setting $\beta = 60$ in Section 5.4. This analysis remains dependent on the assumptions of the model, and excludes alternative responses by HFTs, such as simply moving their trading to an alternative non-regulated venue, instead of optimizing under the new constraints imposed by the regulation.

5.1. Tobin Tax: Taxing High Frequency Trades

The first policy we consider consists in taxing each trade that a HFT executes. Leaving aside the question of identifying HFT trades (perhaps by requiring HFT firms to register with the regulators, as has been proposed in Germany), a financial transactions tax is nothing new. Originally known as a "stamp duty," it was first implemented at the London Stock Exchange in the 17th century, was later advocated by Keynes on the grounds that speculation by uninformed traders increased volatility, and then by Tobin as a means of reducing currency fluctuations.

The argument in favor of a transactions tax is that financial trading is under-taxed relative to the rest of the economy; this encourages excessive trading, by HFTs in particular, which in turn undermines financial stability as the ability of HFTs to get out of the market quickly undermines the market's liquidity when it is most needed.

Of course, sophisticated traders may simply move their trading to financial instruments or jurisdictions not subject to the tax. Sweden for instance introduced a tax on the purchase or sale of stocks in 1984; the tax was repealed in 1990 after the country experienced a large displacement of trades.[22] A second argument often made against the tax is that it will depress economic activity by imposing a large burden on the financial sector. These two arguments are somewhat self-contradictory: either the tax is easily avoided so as to be inconsequential, or it imposes a large economic penalty, but not both together[23].

In the framework of our model, a transactions tax is straightforward to analyze. Suppose that the

[22] "Financial Transactions Taxes: The International Experience and the Lessons for Canada", by Marion G. Wrobel, Parliament of Canada Report, June 1996.

[23] "Europe should embrace a financial transaction tax", The Financial Times, May 28, 2013.

HFT pays κ dollars each time an LFT order crosses one of his limit orders. From the perspective of the HFT, the transaction tax, κ, merely reduces the gain that the HFT earns from each trade. We can analyze the impact of the tax policy on the objective value of the HFT and liquidity provision using our complete model.

Figure 12 displays three graphs illustrating objective value, equilibrium bid-offer spread with respect to the Tobin tax rate and equilibrium bid-offer spread as a function of volatility in the presence and absence of tax regulation. Figure 12 illustrates that the HFT's objective value is decreasing with higher taxes, i.e., with a maximum κ considered of 25 bps in line with the proposals being considered or already implemented in Europe. We observe that in the long-run transaction taxes do not incentivize the HFT to quote more on both sides of the market. If the taxes are high enough, in fact, the HFT's liquidity provision actually decreases, as seen in the top panel of Figure 12.

These predictions of the model are consistent with what has been observed in Italy following the introduction of the Tobin tax: the average daily trading volume for Italian-domiciled stocks has fallen by nearly 40% in March compared to January and February 2013.[24]. Lastly, we investigate how market liquidity, measured by the equilibrium bid-offer spread, changes with respect to volatility when the Tobin tax is implemented. The HFT's quoting has higher sensitivity to volatility compared to the absence of a Tobin tax, so the tax produces no improvement on that front either.

5.2. Speed Bumps for HFTs: Imposing a Minimum Rest Time Before a Quote Can Be Cancelled

Another possible policy consists in imposing a minimum time before a quote can be cancelled by the HFT. This minimum "rest time" is a widely discussed policy among regulators and exchanges, including European and Australian ones. The objective is to improve the provision of liquidity by HFTs by effectively forcing them to stand behind their quotes for at least a brief period. One concern about the reported higher liquidity due to HFT activity is that the provided liquidity is very short-lived, or "phantom", i.e., HFTs cancel many of their quotes before LFTs get a chance to trade with it.

We analyze the effect of minimum rest limits in our model as follows. Although the policy proposals all suggest a fixed waiting time, typically 500 milliseconds, it will come as no surprise in the context of the model that it is actually more convenient to analyze a version of the policy where the waiting time before a cancellation is random, itself derived from a Poisson process, with an expected value of 500 milliseconds. The arrival rate of that Poisson process controls the expected amount of waiting

[24] "Tax blow to Italian stock trading", The Financial Times, March 13, 2013.

time before a quote can be cancelled. That is, we suppose that each active quotes cannot be cancelled before a random time amount, τ^{cancel}, which is exponentially distributed with mean duration $1/\theta$.

Figure 13 illustrates the objective value of the HFT as a function of the average rest time, $1/\theta$, expressed in milliseconds. The limiting case $\theta \to \infty$, meaning that the minimum rest time is zero for active quotes, reverts to the base model. As the rest time increases, we observe in the upper left panel that the objective value of the HFT decreases: like any other constraint, a rest time does reduce the maximum achievable value.

Importantly, we find that this form of regulation is good for the overall liquidity as measured by equilibrium bid-offer spread. When the HFT has active quotes on both sides of the market, the minimum rest time comes into effect and forces the HFT not to cancel as frequently as before due to signal changes, including signals about the next incoming LFT's type. As he internalizes this constraint when optimizing his quoting policy, the HFT ends up initiating fewer quotes but they stay active sufficiently longer to result in a higher probability that he will have a quote active at any given point in time, and in a higher fill rate for the incoming LFTs' orders.

This policy fails however to result in a countercyclical improvement to liquidity: as shown in the bottom-panel graph in Figure 13, with rest times in effect, the HFT becomes more sensitive to market volatility, i.e., when a rest time is imposed, equilibrium bid-offer spread is lower when volatility is low, but higher when volatility is high, compared to the situation without one. The reason is that as σ increases, the potential risk from being caught with stale quotes increases, and is compounded by the imposition of a rest time which forces the HFT to honor a quote that he would rather not. The HFT then optimally reduces his quoting. In other words, rest times lead HFTs to provide more liquidity when the market does not really need it (more rain in a monsoon) but less when the market would really benefit from it (less rain in a drought).

5.3. Taxing Limit Order Cancellations

The third policy we consider consists in taxing the HFT whenever he cancels an existing quote. Unlike rest times, which make cancellations impossible within a certain time interval, this policy simply makes them costlier. The two are not directly comparable in that rest times are effectively an infinite tax on cancellations but only for a brief period, whereas a cancellation tax is a small tax that is in effect permanently. We assume that the HFT must pay ε percent as a penalty for each cancelled limit order. For example, if the HFT chooses action $\ell^b = 0$ or $\ell^a = 1$ when his existing quotes are $l = 11$, he will pay a cancellation tax of ε as he cancelled his existing quote on the bid side.

Figure 14 considers the impact of such cancellation taxes. We consider cases where ε is as high as 10 bps for each quote cancellation. We observe that the objective value of the HFT decreases in

the presence of cancellation taxes. The HFT tries to quote more at the best bid and the ask to avoid cancellation fees, which lowers the equilibrium bid-offer spread. A cancellation tax shares the same drawback as a minimum rest time: it improves market liquidity when volatility is low but when there is high volatility, the cancellation tax causes the HFT to quote less compared to the benchmark case.

5.4. Pro Rata or Random Allocation

The last policy we consider is pro rata allocation: as an alternative to price and time priority, or first in first out, pro rata allocation matches incoming orders against quotes by allocating shares proportionately to all quotes at the best price according to their size. This form of matching, employed for instance at the Chicago Mercantile Exchange for Eurodollar futures contracts, enables all liquidity providers to join the queue at a particular price level and have an opportunity to compete for the next fill at that price level, independently of their order's time priority. A closely-related alternative matching algorithm is one where, among the quotes in the book at the prevailing best price, the one to be executed is selected at random.

In our duopoly model with time priority, if both the HFT and the MFT have limit orders at the best price, the execution priority belongs to the liquidity provider who first submitted the order. Since quantities in the model are set at one share per transaction, we model pro rata/random allocation as follows. Instead of the first quote to enter the book being automatically selected for execution, an independent Bernoulli random variable is drawn and the order with the highest time-priority is selected with probability $\eta \in [0.5, 1]$. When $\eta = 1$, this reverts to our original model, i.e., pure time priority, whereas $\eta = 0.5$ completely eradicates time priority and corresponds to a fully random allocation. We also consider intermediary values for completeness.

Figure 15 considers the impact this policy on liquidity metrics. We find that the objective value of the HFT decreases in the presence of pro rata allocation. In the absence of this policy, HFT's orders often enjoy higher priority over the MFT's orders due to the HFT's speed advantage, which results in more trades for the HFT. However, once the pro rata policy in place, the HFT's speed advantage loses (part of) its significance and his objective value drops. Although this is indeed the objective of the policy, it results in less liquidity provision. Figure 15 illustrates that the HFT lowers his total liquidity provision under pro rata allocation, as it is now less likely for the HFT to trade with an LFT even if his order had time-priority. Since the HFT's objective is to make round-trip trades in minimum time, keeping the same quoting policy under pro rata allocation is now costly due to inventory risk. Consequently, the HFT is less willing to quote under a policy that diminishes his speed advantage. Specifically, the HFT will now find quoting at the second-best prices more attractive relative to the time-priority case as the HFT's probability of trading with a LFT at the best price is lower. A higher

incentive to quote at the second-best prices leads to higher equilibrium bid-offer spreads as shown in Figure 15. Finally, the bottom panel in Figure 15 illustrates that the HFT's quoting, although lower across the volatility range, has similar sensitivity to volatility compared to time priority, so pro rata matching produces no improvement in terms of making the provision of liquidity countercyclical with respect to volatility.

Aside from the model, note that if quantities were not fixed, we would expect the HFT to quote larger quantities under pro rata matching; this incentive creates a separate risk, that of liquidity providers quoting larger quantities than what they can safely absorb since pro rata allocation makes it less likely that they would be forced to do so. In the case of extreme market moves, such book padding strategies, although harmless in normal times, could result in large losses (see McPartland (2015)).

6. Conclusions

We propose a theoretical model of high frequency market making. We superpose different Poisson processes running on different time clocks to represent the arrival of different elements of market information and orders, resulting in a tractable and flexible framework where the optimal market making strategy of the HFT and the equilibrium between the HFT quotes and incoming LFTs' orders is fully characterized. The model reproduces many important stylized facts about HFTs. We find that the HFT's liquidity provision increases when he gets faster. We find that the optimal quoting policy of the HFT also leads to cancellation rates of his orders in the presence of informative signals about the order flow. The model also quantifies the impact of higher volatility on liquidity provision. We find that the HFT will provide less liquidity when volatility increases if the arrival rate of LFTs is independent of price volatility. In the presence of positive correlation between the arrival rates of the LFTs and fundamental price volatility, the model predicts an inverse U-shaped pattern of liquidity provision: a small increase in volatility leads to more liquidity, followed by a decrease as volatility increases further. When analyzing competition for order flow with a second market maker, we find that competing market makers split the rent extracted from LFTs, liquidity provision increases and LFTs tend to be better off.

Finally, we provide the first model-based analysis of the impact of four widely discussed HFT policies: imposing a transactions tax, setting minimum-time limits before orders can be cancelled, taxing the cancellations of limit orders, and replacing time priority with pro rata or random allocation. We assess these regulatory policies on the basis of their potential to induce the HFT to provide liquidity that is more resilient to increases in volatility, i.e., countercyclical with respect to volatility. We find

that none of the four policies result in an improvement. A transactions tax and pro rata matching result in less liquidity altogether. Minimum rest times and a cancellation tax result in more liquidity in good (low volatility) environments but less in bad (high volatility) environments, the opposite of the desired effect. Ultimately, our conclusion is that the microstructure-based speed and informational advantages of the HFT are difficult to even out.

References

Aït-Sahalia, Y., Sağlam, M., 2016. High frequency market making: Optimal quoting. Tech. rep., Princeton University.

Àlvaro Cartea, Jaimungal, S., Ricci, J., 2014. Buy low, sell high: A high frequency trading perspective. SIAM Journal on Financial Mathematics 5, 415–444.

Àlvaro Cartea, Penalva, J., 2012. Where is the value in high frequency trading? Quarterly Journal of Finance 2, 1–46.

Amihud, Y., Mendelson, H., 1980. Dealership market: Market-making and inventory. Journal of Financial Economics 8, 31–53.

Anand, A., Venkataraman, K., 2016. Market conditions, fragility and the economics of market making. Journal of Financial Economics 121, 327–349.

Angel, J., Harris, L., Spatt, C. S., 2015. Equity trading in the 21st century: An update. Quarterly Journal of Finance 5, 1–39.

Avellaneda, M., Stoikov, S., 2008. High-frequency trading in a limit order book. Quantitative Finance 8, 217–224.

Baruch, S., Glosten, L. R., 2013. Flickering quotes. Tech. rep., Working paper, Columbia University Graduate School of Business.

Biais, B., Foucault, T., Moinas, S., 2015. Equilibrium fast trading. Journal of Financial Economics 116, 292–313.

Biais, B., Woolley, P., 2011. High frequency trading. Tech. rep., Toulouse School of Economics and London School of Economics.

Brogaard, J. A., Hendershott, T., Riordan, R., 2014. High frequency trading and price discovery. Review of Financial Studies 27, 2267–2306.

Budish, E., Cramton, P., Shim, J., 2015. The high-frequency trading arms race: Frequent batch auctions as a market design response. Quarterly Journal of Economics 130, 1547–1621.

Chaboud, A., Chiquoine, B., Hjalmarsson, E., Vega, C., 2010. Rise of the machines: Algorithmic trading in the foreign exchange market. Tech. rep., IMF.

Cvitanić, J., Kirilenko, A., 2010. High frequency traders and asset prices. Tech. rep., Caltech.

Easley, D., de Prado, M. M. L., O'Hara, M., 2011. The microstructure of the "flash crash": Flow toxicity, liquidity crashes and the probability of informed trading. Journal of Portfolio Management 37, 118–128.

Foucault, T., Hombert, J., Rosu, I., 2016. News trading and speed. The Journal of Finance 71, 335–382.

Foucault, T., Kadan, O., Kandel, E., 2013. Liquidity cycles and make/take fees in electronic markets. The Journal of Finance 68, 299–341.

Guéant, O., Lehalle, C.-A., Fernandez-Tapia, J., 2013. Dealing with the inventory risk: A solution to the market making problem. Mathematics and Financial Economics 7, 477–507.

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

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

Hasbrouck, J., Saar, G., 2013. Low-latency trading. Journal of Financial Markets 16, 646–679.

Hendershott, T., Jones, C. M., Menkveld, A. J., 2011. Does algorithmic trading improve liquidity? The Journal of Finance 66, 1–33.

Hendershott, T., Menkveld, A. J., 2014. Price pressures. Journal of Financial Economics 114, 405–423.

Ho, T. S. Y., Stoll, H. R., 1981. Optimal dealer pricing under transactions and return uncertainty. Journal of Financial Economics 9, 47–73.

Jarrow, R. A., Protter, P., 2012. A dysfunctional role of high frequency trading in electronic markets. International Journal of Theoretical and Applied Finance 15 (3), 1–15.

Jovanovic, B., Menkveld, A. J., 2010. Middlemen in limit-order markets. Tech. rep., New York University and VU University Amsterdam.

Kirilenko, A., Kyle, A. P., Samadi, M., Tuzun, T., 2010. The flash crash: The impact of high frequency trading on an electronic market. Tech. rep., University of Maryland.

McPartland, J. W., 2015. Recommendations for equitable allocation of trades in high frequency trading environments. Journal of Trading 10, 81–100.

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

Menkveld, A. J., 2016. The economics of high-frequency trading: Taking stock. Annual Review of Financial Economics, forthcoming .

Moallemi, C., Sağlam, M., 2013. The cost of latency in high-frequency trading. Operations Research 61, 1070–1086.

Pagnotta, E., Philippon, T., 2011. Competing on speed. Tech. rep., NBER.

Puterman, M. L., 1994. Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons.

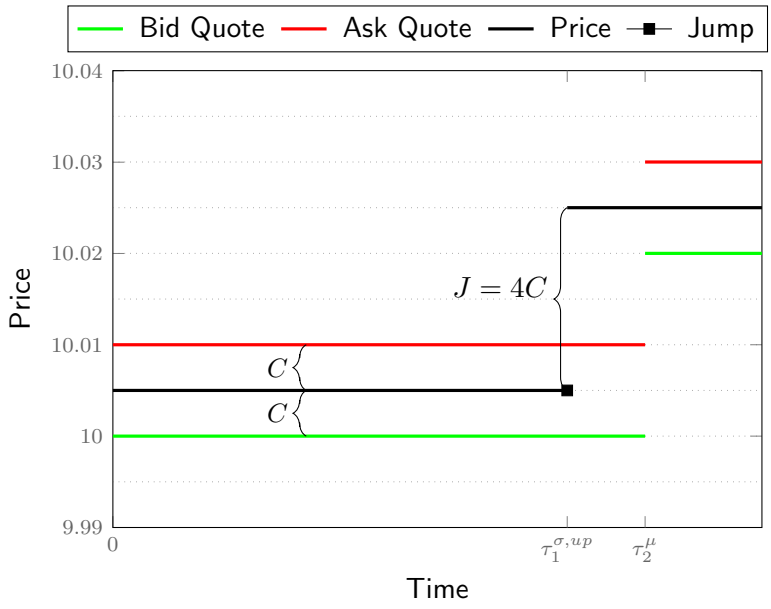

Fig. 1. Dynamics of the fundamental value: Example of a price jump.

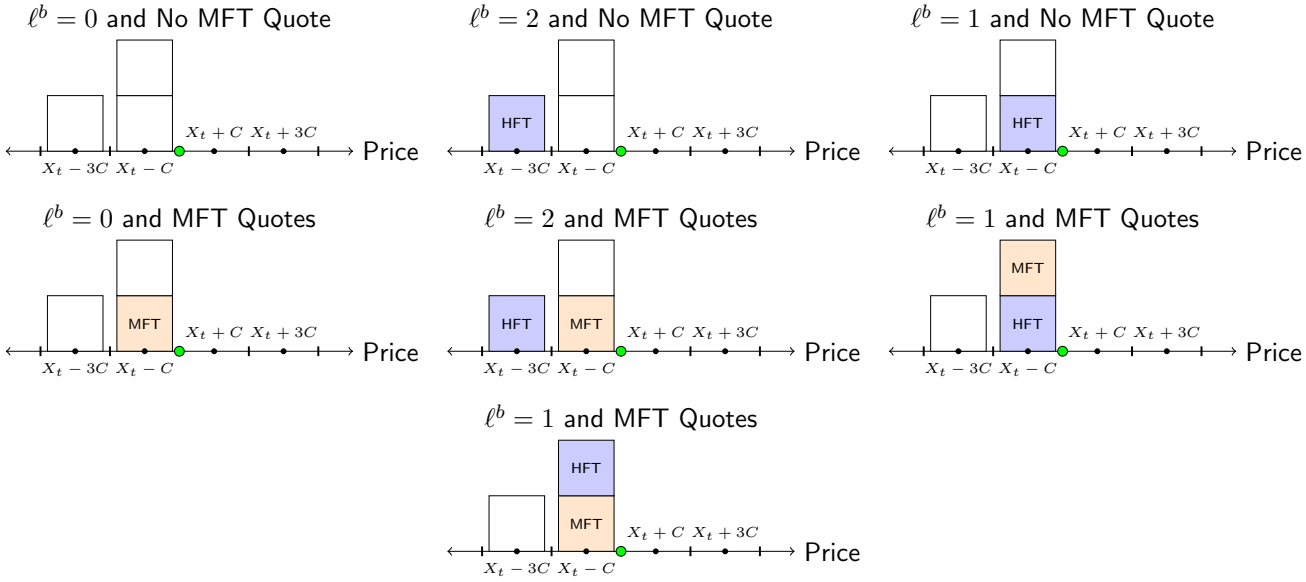

Fig. 2. The 7 possible states of the bid side of the book.

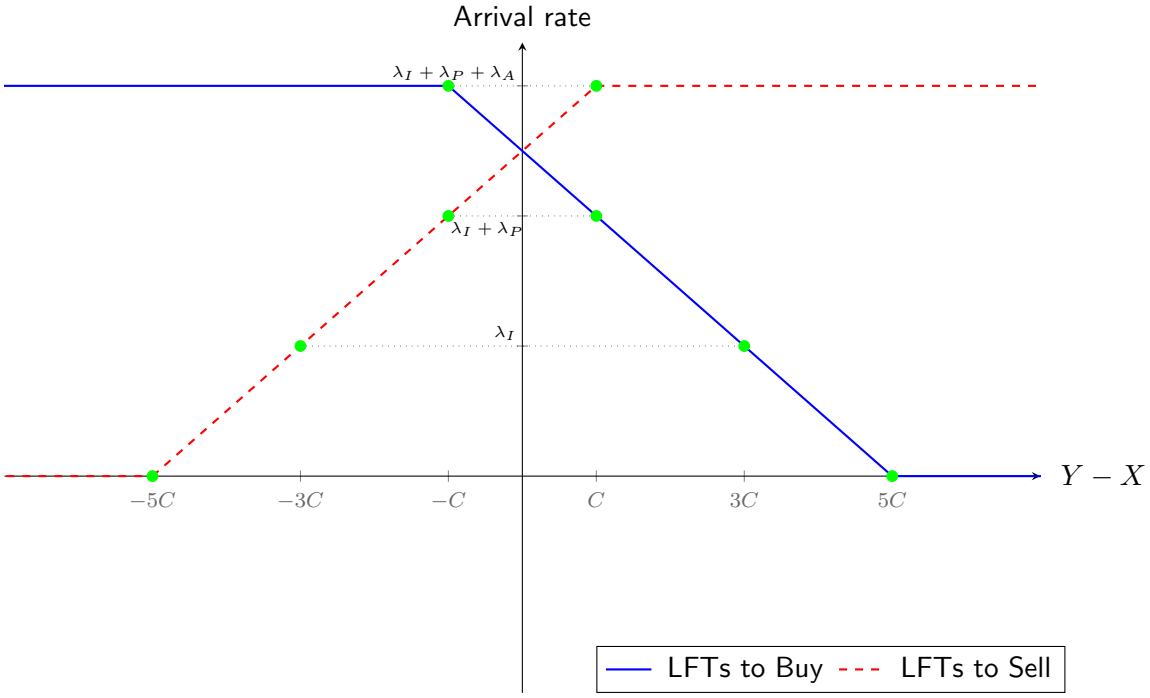

Fig. 3. Demand and supply curves from LFTs.

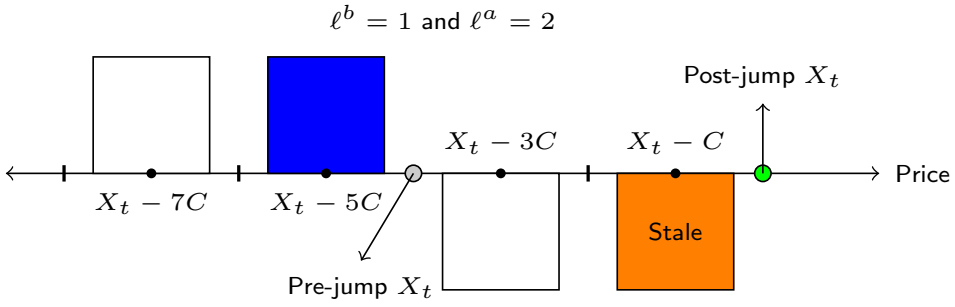

Fig. 4. Potential stale quotes of the HFT in the presence of price jumps.

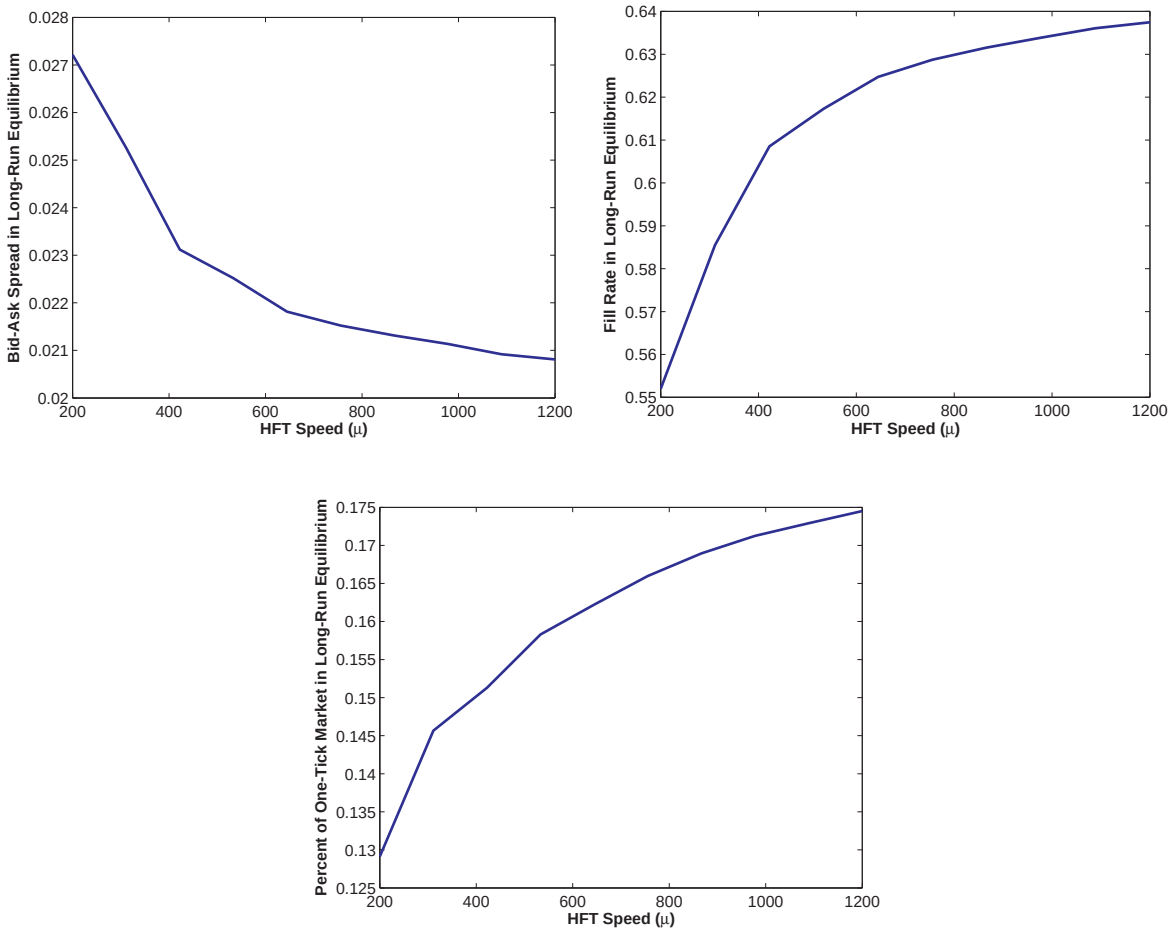

Fig. 5. Liquidity measures in long-run equilibrium as a function of the HFT's speed.

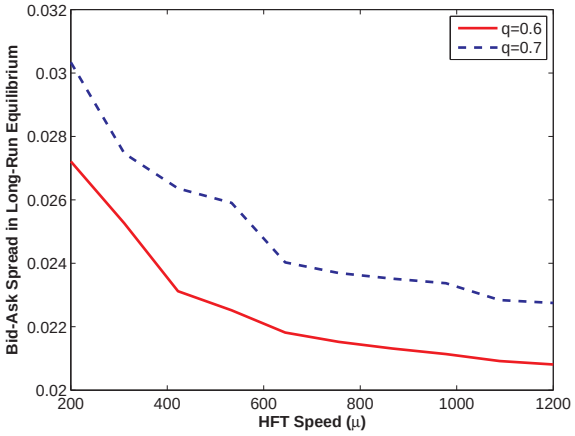

Fig. 6. Bid-offer spread in long-run equilibrium as a function of HFT's speed for two regimes of HFT's predictive ability regarding the LFT types.

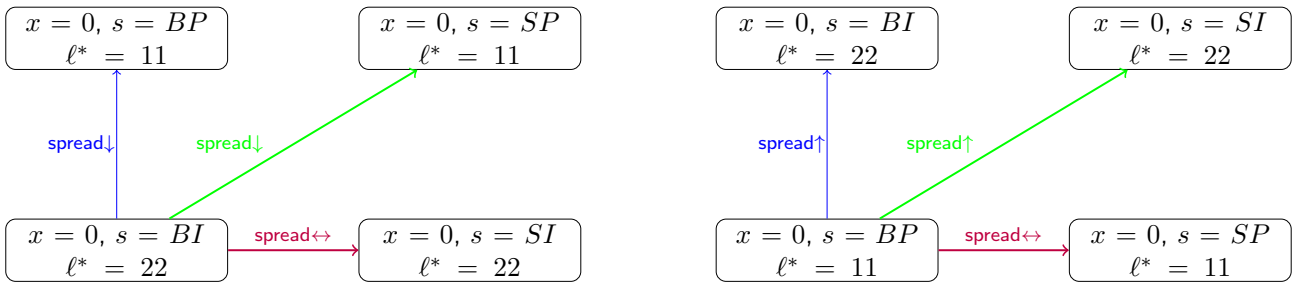

Fig. 7. HFT's quote changes under different inventory and signal states and their impact on the spread.

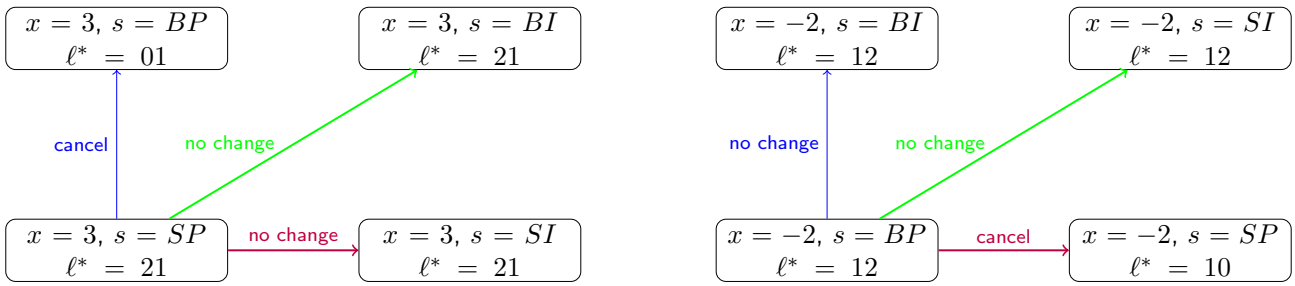

Fig. 8. HFT's order revisions and cancellations under different inventory and signal states.

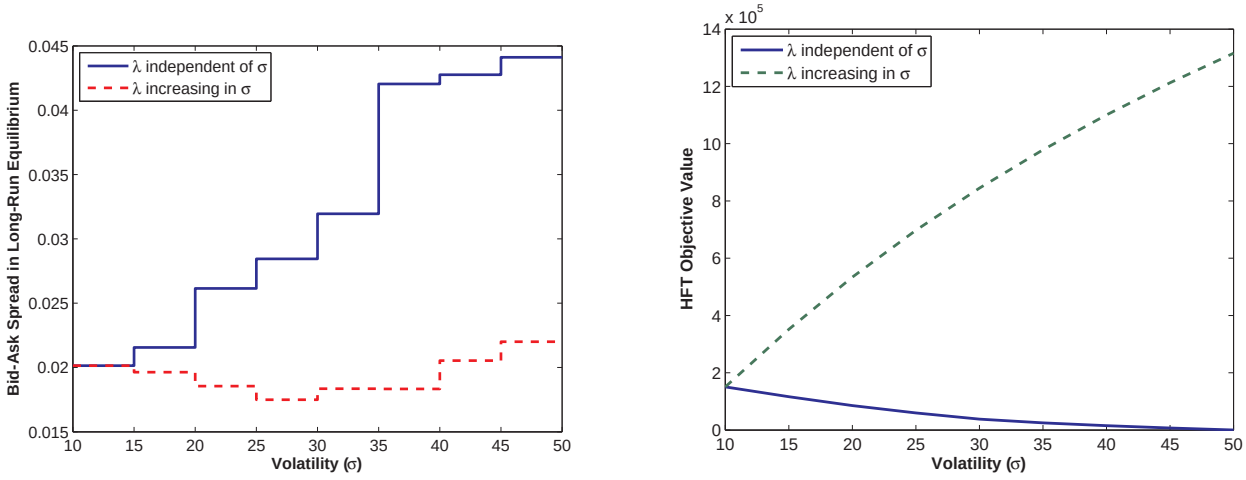

Fig. 9. Long-run equilibrium bid-offer spread as a function of price volatility.

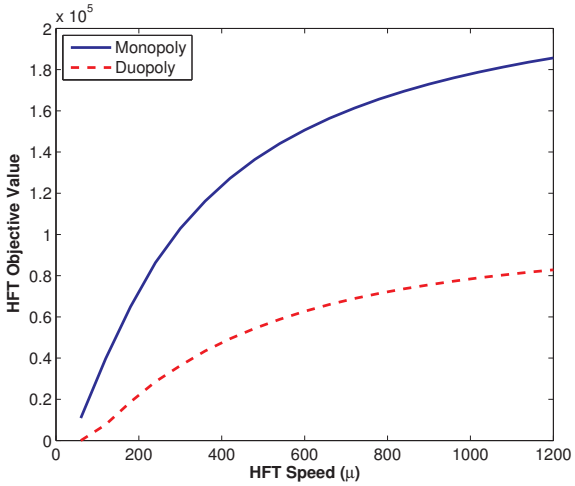

Fig. 10. HFT Optimal Values in the Monopoly (No Queuing) and Duopoly (Queuing) Situations.

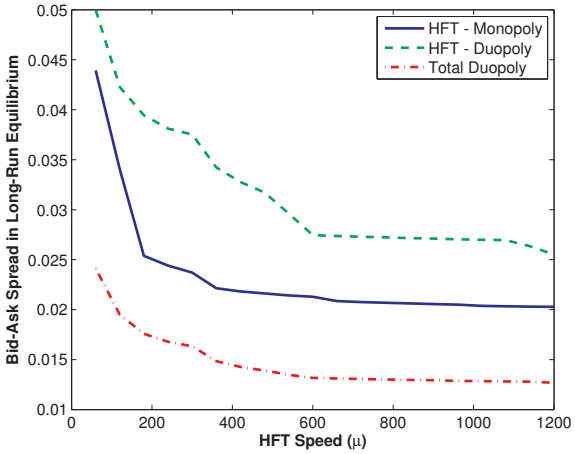

Fig. 11. Long-run equilibrium spread in the presence of the MFT as a function of the HFT's speed.

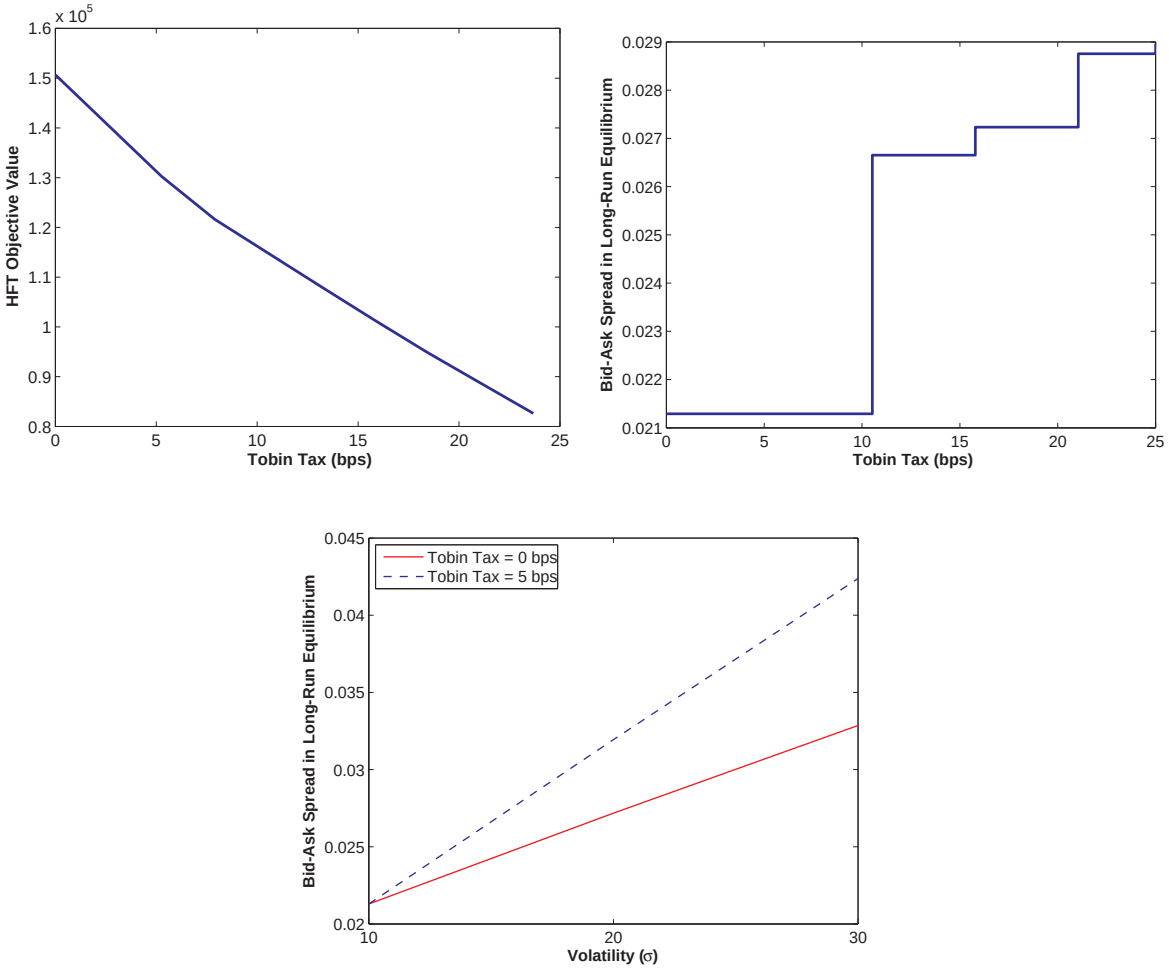

Fig. 12. Impact of a Tobin Tax

Notes: The top panels plot the effect of taxing transactions on the HFTs' value and his provision of liquidity. The bottom panel displays the sensitivity of the equilibrium bid-offer spread to volatility (in the form of price jumps) before and after the Tobin tax.

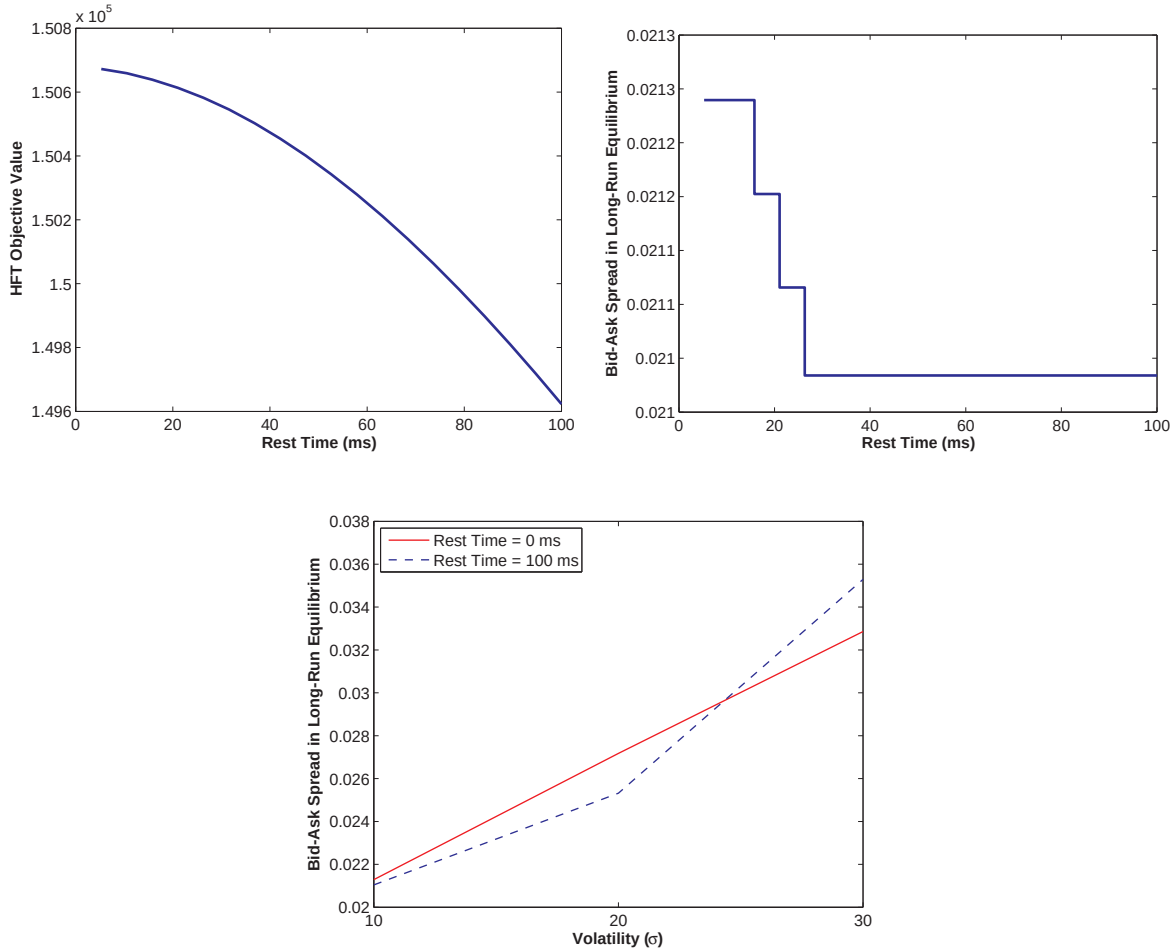

Fig. 13. Impact of a Mandatory Minimum Rest Time on HFT Quotes

Notes: The top panels plot the effect of minimum rest times on the HFTs' value and his provision of liquidity. The bottom panel displays the sensitivity of the equilibrium bid-offer spread to volatility (in the form of price jumps) before and after the minimum rest time.

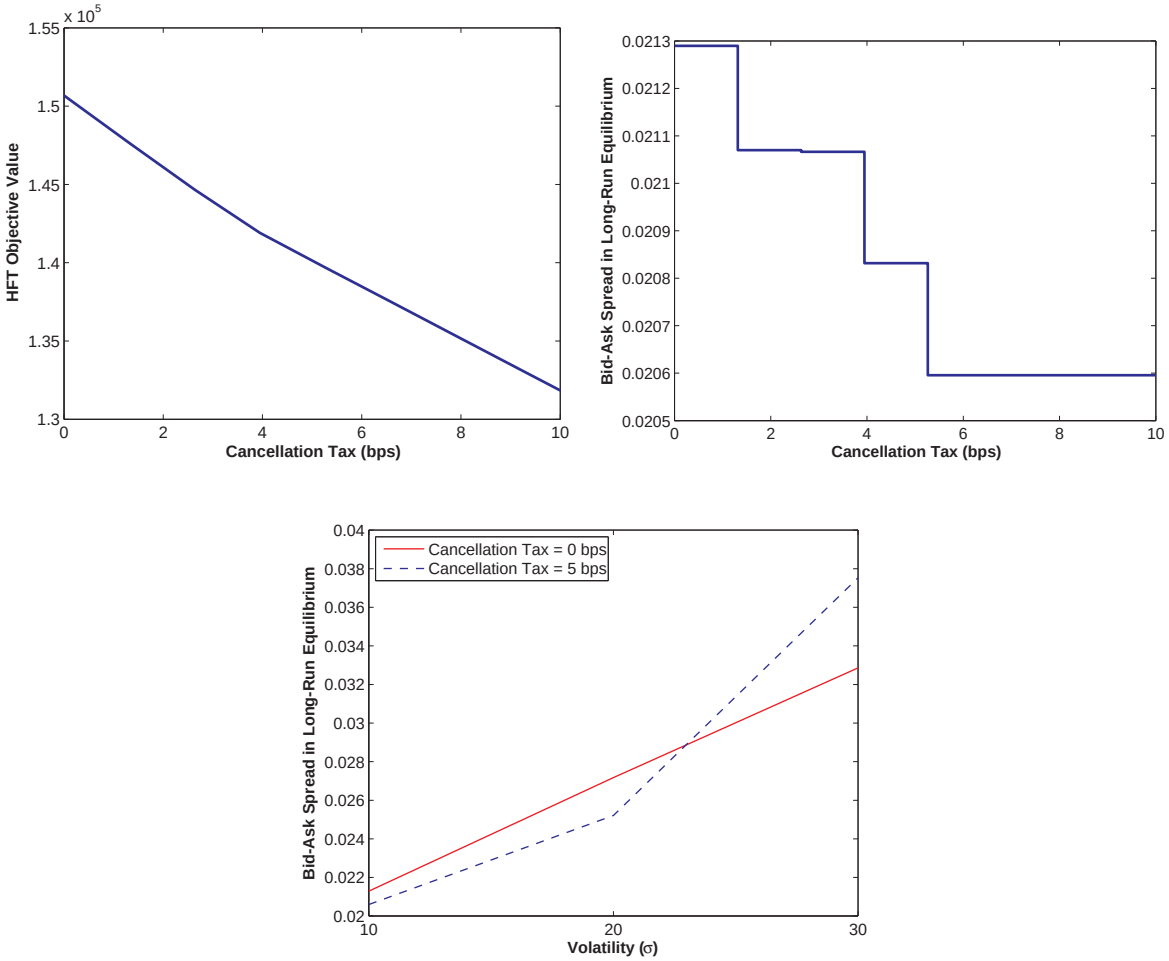

Fig. 14. Impact of a Cancellation Tax

Notes: The top panels plot the effect of taxing cancellations on the HFTs' value and his provision of liquidity. The bottom panel displays the sensitivity of the equilibrium bid-offer spread to volatility (in the form of price jumps) before and after the cancellation tax.

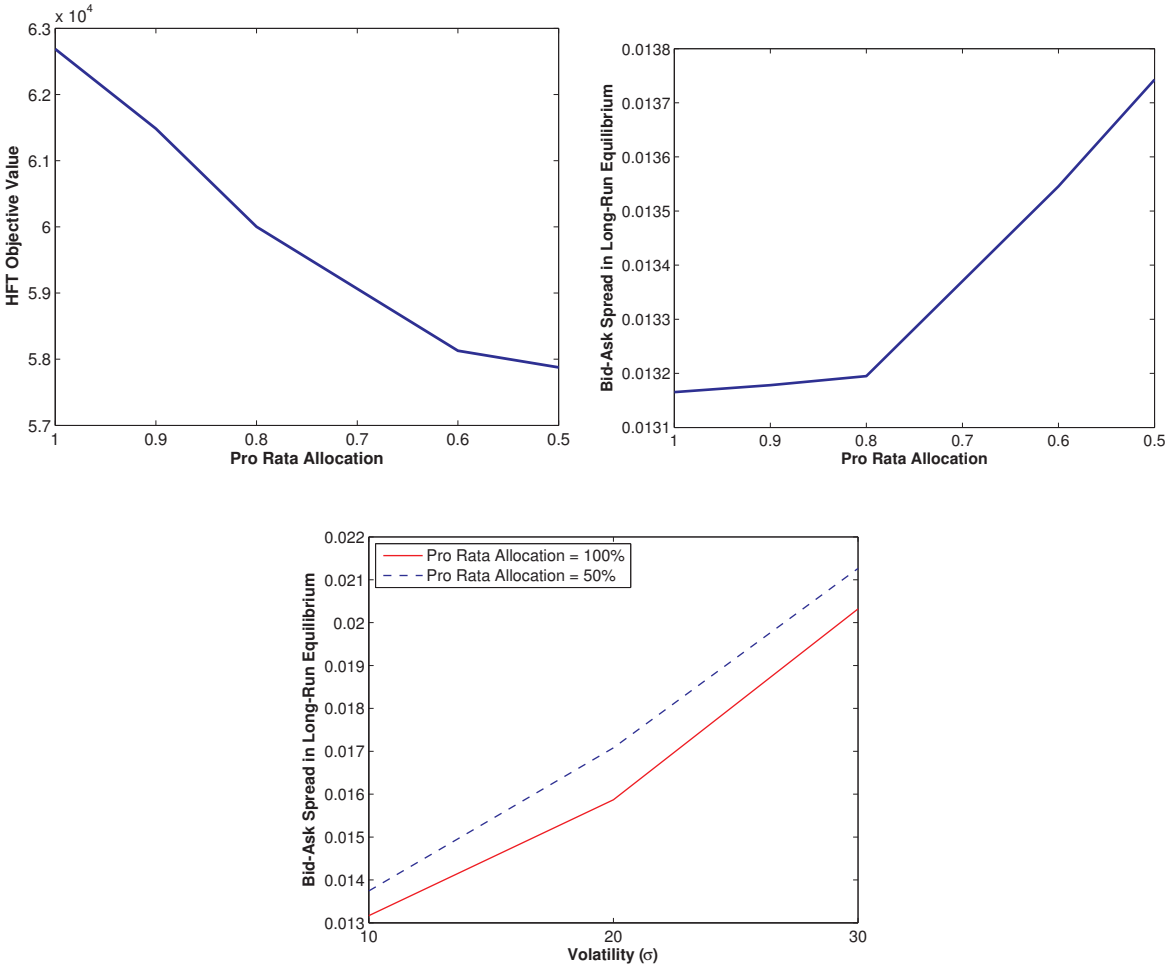

Fig. 15. Impact of Pro Rata Allocation

Notes: The top panels plot the effect of various pro rata levels on the HFTs' value and his provision of liquidity. The bottom panel displays the sensitivity of the equilibrium bid-offer spread to volatility (in the form of price jumps) before and after the pro rata policy with 50%.

Appendix for "High Frequency Market Making: Implications for Liquidity" Technical Results and Proofs

A. Discrete-time Transformation of the Model

We start by recalling the definition of a discounted infinite horizon Markov Decision Process (MDP), before showing that our continuous-time HFT optimization problem can be represented as such. A MDP is defined by a 4-tuple, $(I, A_i, \mathbb{P}(.|i,a), \mathbb{R}(.|i,a))$, in which I is the state space, A_i is the action space, i.e., the set of possible actions that a decision maker can take when the state is $i \in I$, $\mathbb{P}(.|i,a)$ is the probability transition matrix determining the state of the system in the next decision time, and finally $\mathbb{R}(.|i,a)$ is the reward matrix, specifying the reward obtained using action a when the state is in i. The HFT seeks a quoting policy that maximizes the expected discounted reward

$$v(i) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t \mathbb{R}(i_{t+1}|i_t, \pi(i_t))|i_0 = i\right], \tag{A.1}$$

where α is the discount rate. An admissible stationary policy π maps each state $i \in I$ to an action in A_i. Under mild technical conditions, we can guarantee the existence of optimal stationary policies (see Puterman (1994)). Conditioning on the first transition from i to i', we obtain the Hamilton-Jacobi-Bellman optimality equation

$$\begin{aligned} v(i) &= \max_{\pi}\left\{\sum_{i'} \mathbb{P}(i'|i,\pi(i))\left(\mathbb{R}(i'|i,\pi(i)) + \alpha\mathbb{E}\left[\sum_{t=1}^{\infty}\alpha^{t-1}\mathbb{R}(i_{t+1}|i_t,\pi(i_t))|i_1 = i'\right]\right)\right\} \\ &= \max_{\pi}\left\{\sum_{i'} \mathbb{P}(i'|i,\pi(i))\left(\mathbb{R}(i'|i,\pi(i)) + \alpha\mathbb{E}\left[\sum_{k=0}^{\infty}\alpha^{k}\mathbb{R}(i_{k+1}|i_k,\pi(i_k))|i_k = i'\right]\right)\right\} \\ &= \max_{a\in A_i}\left\{\sum_{i'} \mathbb{P}(i'|i,a)\left(\mathbb{R}(i'|i,a) + \alpha v(i')\right)\right\}. \end{aligned} \tag{A.2}$$

B. Transition Probabilities

We will calculate the transition probabilities at each state of the HFT in the presence of a competing MFT. First, note that the state transitions occur at a rate of $\mu + \lambda + \theta + \sigma + 2\beta$ where $\lambda \equiv 2(\lambda_P + \lambda_I + \lambda_A)$ and using uniformization we can have the same transition rate for all states and actions. Let $\mathbb{P}((x', s', b', a', e', j')|(x, s, b, a, e, j), (\ell^b, \ell^a))$ be the probability of reaching state (x', s', b', a', e', j') when the system is in state (x, s, b, a, e, j) and the trader takes the actions of ℓ^b and ℓ^a. Here b (a) is the

current state of the bid (ask) market.

First, we define our auxiliary variables. Let $\mathsf{pr}(s)$ denote the unconditional probability of receiving signal $s = \left(S^{\text{dir}}, S^{\text{type}}\right)$ right after the arrival of a market order or a signal.

$$\begin{aligned}
\mathsf{pr}(s) &= \sum_{i\in\{P,I\}} \sum_{j\in\{B,S\}} \mathbb{P}(M^{\text{type}} = i, M^{\text{dir}} = j)\mathbb{P}(s|M^{\text{type}} = i, M^{\text{dir}} = j) \\
&= \sum_{i\in\{P,I\}} \sum_{j\in\{B,S\}} (0.5/\lambda)\left(p\mathbb{1}\{s_1 = i\} + (1-p)\mathbb{1}\{s_1 \neq i\}\right)\left(q\mathbb{1}\{s_2 = j\} + (1-q)\mathbb{1}\{s_2 \neq j\}\right) \\
&\qquad\qquad \left(\lambda^I \mathbb{1}\{i = I\} + \lambda^P \mathbb{1}\{i = P\}\right) \\
&= \begin{cases} 0.5p\lambda^I/\lambda + 0.5(1-p)\lambda^P/\lambda & \text{if } s_1 = I, \\ 0.5(1-p)\lambda^I/\lambda + 0.5p\lambda^P/\lambda & \text{if } s_1 = P. \end{cases}
\end{aligned}$$

Let $\mathsf{m}_s^{s'}$ denote the conditional probability of receiving a market order with type s' (e.g., buy order submitted by an impatient LFT will be denoted by $s' = IB$) when the last signal appeared is s.

$$\begin{aligned}
\mathsf{m}_s^{s'} &= \frac{\mathbb{P}(M^{\text{type}} = s_1', M^{\text{dir}} = s_2')\mathbb{P}(S^{\text{type}} = s_1, S^{\text{dir}} = s_2|M^{\text{type}} = s_1', M^{\text{dir}} = s_2')}{\sum_{i\in\{P,I\}}\sum_{j\in\{B,S\}} \mathbb{P}(M^{\text{type}} = i, M^{\text{dir}} = j)\mathbb{P}(S^{\text{type}} = s_1, S^{\text{dir}} = s_2|M^{\text{type}} = i, M^{\text{dir}} = j)} \\
&= \left(p\mathbb{1}\{s_1 = s_1'\} + (1-p)\mathbb{1}\{s_1 \neq s_1'\}\right)\left(q\mathbb{1}\{s_2 = s_2'\} + (1-q)\mathbb{1}\{s_2 \neq s_2'\}\right) \\
&\left(\lambda^I \mathbb{1}\{s_1' = I\} + \lambda^P \mathbb{1}\{s_1' = P\}\right)/(\lambda \mathsf{pr}(s))
\end{aligned}$$

Suppose that the current state of the HFT is (x, s, b, a, e, j). Let $r = (\lambda + \mu + \theta + \sigma + 2\beta)$. In order to illustrate our methodology, we will provide a few examples of the transition probabilities.

If the HFT does not quote in either side of the market, the inventory level cannot change. Suppose that the MFT does not have any active quotes either so let the state be $(x, s, 00, 00, 1, j)$. If a new decision event arrives before the arrival of an LFT order, a signal or jump, the state for tracking jumps, j, reverts to zero. The MFT may send a limit order at the best bid or the ask during this time with probability β/r. Since the arrival of arbitrageurs will not change the state of the system, we also have the additional self-transition at the rate of $2\lambda_A/r$ to uniformize the model. Formally, we have

the following transition probabilities for $\mathbb{P}\left((x,s',b',a',e',j')|(x,s,00,00,1,j),(0,0)\right)$:

$$\begin{cases} \frac{(\lambda^{PI}+\theta)\mathsf{pr}(s')}{r} & \text{if } x=x', e'=0, b'=00, a'=00 \\ \frac{\sigma}{2r} & \text{if } x=x', e'=0, b'=00, a'=00, j'=j+1 \\ \frac{\sigma}{2r} & \text{if } x=x', e'=0, b'=00, a'=00, j'=j-1 \\ \frac{\mu}{r} & \text{if } x=x', e'=1, s=s', b'=00, a'=00, j'=0 \\ \frac{\beta}{r} & \text{if } x=x', e'=0, s=s', b'=00, a'=01, j'=j \\ \frac{\beta}{r} & \text{if } x=x', e'=0, s=s', b'=01, a'=00, j'=j \\ \frac{2\lambda_A}{r} & \text{if } x=x', e'=e, s=s', b'=b, a'=a, j'=j \\ 0 & \text{otherwise,} \end{cases}$$

where $\lambda^{PI} \equiv 2(\lambda_P + \lambda_I)$.

If the only quote at the bid side is from the MFT, the MFT will have priority for execution at the bid side. For example, suppose that the state is $(x,s,01,00,1,j)$ and the HFT decides to take action $(1,0)$. Since the MFT order has a priority, in the next epoch, this order can be executed and the HFT's order will now have the highest priority. Thus, there will be no possibility of an inventory increase compared to the monopoly model. Since the ask side is empty, the MFT can submit a new order at the ask side as well. In this case the stale quote risk still applies and will be applicable if $j<0$. Formally, we have the following transition probabilities for $\mathbb{P}\left((x,s',b',a',e',j')|(x,s,01,00,1,j),(1,0)\right)$:

$$\begin{cases} \frac{(\lambda^{\mathsf{sell}}+\theta)\mathsf{pr}(s')}{r} & \text{if } x=x', e'=0, b'=11m, a'=00, j\leq 0 \\ \frac{(\lambda^{\mathsf{sell}}+\lambda^{\mathsf{buy}}+\theta)\mathsf{pr}(s')}{r} & \text{if } x=x', e'=0, b'=11m, a'=00, j>0 \\ \frac{\lambda^{\mathsf{buy}}\mathsf{pr}(s')}{r} & \text{if } x=x', e'=0, b'=10, a'=00, j=0 \\ \frac{\left(\frac{\lambda_A}{2}+\lambda^{\mathsf{buy}}\right)\mathsf{pr}(s')}{r} & \text{if } x=x', e'=0, b'=10, a'=00, j<0 \\ \frac{\sigma}{2r} & \text{if } x=x', e'=0, s=s', b'=11m, a'=00, j'=j+1 \\ \frac{\sigma}{2r} & \text{if } x=x', e'=0, s=s', b'=11m, a'=00, j'=j-1 \\ \frac{\mu}{r} & \text{if } x=x', e'=1, s=s', b'=11m, a'=00, j'=0 \\ \frac{\beta}{r} & \text{if } x=x', e'=0, s=s', b'=11m, a'=01, j'=j \\ \frac{2\lambda_A+\beta}{r} & \text{if } x=x', e'=e, s=s', b'=b, a'=a, j'=j\geq 0 \\ \frac{\lambda_A+\beta}{r} & \text{if } x=x', e'=e, s=s', b'=b, a'=a, j'=j<0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\lambda^{\mathsf{buy}} = \lambda^{PI}(m_s^{IS}+m_s^{PS})$ and $\lambda^{\mathsf{sell}} = \lambda^{PI}(m_s^{IB}+m_s^{PB})$ denotes the corresponding intensity for HFT's buy and sell trade in absence of any jumps, respectively.

We now analyze the trade scenarios for one-sided quoting when HFT is quoting at the bid side with execution priority. For example, suppose that the state is $(x, s, 00, 00, 1, j)$ and the HFT decides to take action $(1, 0)$. The remaining one-sided HFT actions are also very similar. When the HFT's action is $(1, 0)$ he may increase his inventory by trading with the incoming market-sell order submitted by a patient or an impatient LFT, which occurs with probability $m_s^{IS} + m_s^{PS}$. We also need to account for the existence of stale quotes. Since the HFT is quoting at the bid side in this case, the stale quotes can only appear if $j < 0$. The MFT may submit a new order at the bid and the ask side as well. Formally, we have the following transition probabilities for $\mathbb{P}\left((x, s', b', a', e', j')|(x, s, 01, 00, 1, j), (1, 0)\right)$:

$$\begin{cases} \frac{(\lambda^{\mathsf{sell}}+\theta)\mathsf{pr}(s')}{r} & \text{if } x = x', e' = 0, b' = 10, a' = 00, j \leq 0 \\ \frac{(\lambda_P+\lambda_I+\theta)\mathsf{pr}(s')}{r} & \text{if } x = x', e' = 0, b' = 10, a' = 00, j > 0 \\ \frac{\lambda^{\mathsf{buy}}\mathsf{pr}(s')}{r} & \text{if } x + 1 = x', e' = 0, b' = 00, a' = 00, j = 0 \\ \frac{\left(\lambda_A+\lambda^{\mathsf{buy}}\right)\mathsf{pr}(s')}{r} & \text{if } x + 1 = x', e' = 0, b' = 00, a' = 00, j < 0 \\ \frac{\sigma}{2r} & \text{if } x = x', e' = 0, b' = 10, a' = 00, j' = j + 1 \\ \frac{\sigma}{2r} & \text{if } x = x', e' = 0, b' = 10, a' = 00, j' = j - 1 \\ \frac{\mu}{r} & \text{if } x = x', e' = 1, s = s', b' = 10, a' = 00, j' = 0 \\ \frac{\beta}{r} & \text{if } x = x', e' = 0, s = s', b' = 10, a' = 01, j' = j \\ \frac{\beta}{r} & \text{if } x = x', e' = 0, s = s', b' = 11h, a' = 00, j' = j \\ \frac{2\lambda_A}{r} & \text{if } x = x', e' = e, s = s', b' = b, a' = a, j' = j \geq 0 \\ \frac{\lambda_A}{r} & \text{if } x = x', e' = e, s = s', b' = b, a' = a, j' = j < 0 \\ 0 & \text{otherwise,} \end{cases}$$

If the system is observed at the arrival time of a market order or signal event, that is $e = 0$, the HFT cannot revise his quotes. We can accommodate these states in our model using fake decisions that merely sets the action to the existing quotes. In this case,

$$\begin{aligned} &\mathbb{P}\left((x', s', b', a', e', j')|(x, s, b, a, 0, j), (\ell^b, \ell^a)\right) \\ &= \begin{cases} \mathbb{P}\left((x', s', b', a', e', j')|(x, s, b, a, 1, j), (\ell^b, \ell^a)\right) & \text{if } \ell^b = \mathsf{hft}(b), \ell^a = \mathsf{hft}(a) \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

where hft is a mapping that extracts the HFT's action from 7 possible states with the following

definition:

$$\mathsf{hft}(l) = \begin{cases} 1 & \text{if } l = 11h, 11m, 10, \\ 2 & \text{if } l = 21, 20, \\ 0 & \text{if } l = 01, 00. \end{cases}$$

C. HFT's Reward Function

Let $\mathbb{R}\left((x', s', b', a', e', j')|(x, s, b, a, e, j), (\ell^b, \ell^a)\right)$ be the probability of reaching state (x', s', b', a', e', j') when the system is in state (x, s, b, a, e, j) and the trader takes the actions of ℓ^b and ℓ^a. We decompose this reward matrix into two parts: $\mathbb{R} = \mathbb{R}^{\text{trade}} + \mathbb{R}^{\text{inv}}$. The first part includes the rewards due to transactions which also include the adverse selection costs and the latter part denotes the inventory costs.

We would like to write the HFT's objective in (2.3) in the form of an MDP objective function as in (A.1). We first introduce the following notation. Let t_k be the time of the kth state transition due to a decision, signal or market order arrival (by convention $t_0 = 0$) and let τ_k be the length of this cycle, i.e., $\tau_k = t_k - t_{k-1}$. The sum of discounted rewards due to transactions can be tracked by

$$\sum_{k=1}^{\infty} e^{-Dt_k} \mathbb{R}^{\text{trade}}\left((x'_{t_k}, s'_{t_k}, b'_{t_k}, a'_{t_k}, e'_{t_k}, j'_{t_k})|(x_{t_{k-1}}, s_{t_{k-1}}, b_{t_{k-1}}, a_{t_{k-1}}, e_{t_{k-1}}, j_{t_{k-1}}), (\ell^b_{t_{k-1}}, \ell^a_{t_{k-1}})\right).$$

We define $\mathbb{R}^{\text{trade}}$ as follows. Let the execution priority states of the HFT be given by $\mathcal{P} = \{10, 20, 11h\}$. Then, let

$$\begin{aligned} &\mathbb{R}^{\text{trade}}\left((x', s', b', a', e', j')|(x, s, b, a, e, j), (\ell^b, \ell^a)\right) \\ &= \mathbb{1}\left(x + 1 = x'\right)\left((C - Jj^-)\mathbb{1}\left(\ell^b = 1, b \in \mathcal{P}\right) + (3C - Jj^-)\mathbb{1}\left(\ell^b = 2, b \in \mathcal{P}\right)\right) \\ &\quad + \mathbb{1}\left(x - 1 = x'\right)\left((C - Jj^+)\mathbb{1}\left(\ell^a = 1, b \in \mathcal{P}\right) + (3C - Jj^+)\mathbb{1}\left(\ell^a = 2, b \in \mathcal{P}\right)\right) \end{aligned}$$

where $j^+ \equiv \max(j, 0)$, $j^- \equiv \max(-j, 0)$. For simplicity in notation, let

$$\mathbb{R}^{\text{trade}}_{t_k} \equiv \mathbb{R}^{\text{trade}}\left((x'_{t_k}, s'_{t_k}, b'_{t_k}, a'_{t_k}, e'_{t_k}, j'_{t_k})|(x_{t_{k-1}}, s_{t_{k-1}}, b_{t_{k-1}}, a_{t_{k-1}}, e_{t_{k-1}}, j_{t_{k-1}}), (\ell^b_{t_{k-1}}, \ell^a_{t_{k-1}})\right).$$

We can take the expectation of the HFT's discounted earnings using the independence of each

cycle length, τ_i, which is an exponentially distributed random variable with mean $1/r$:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{\infty} e^{-Dt_k}\mathbb{R}^{\text{trade}}_{t_k}\right] = \sum_{k=1}^{\infty}\mathbb{E}\left[e^{-D\sum_{i=1}^{k}\tau_i}\right]\mathbb{E}\left[\mathbb{R}^{\text{trade}}_{t_k}\right] \\
&= \sum_{k=1}^{\infty}\mathbb{E}\left[e^{-D\tau_1}\right]^k\mathbb{E}\left[\mathbb{R}^{\text{trade}}_{t_k}\right] \\
&= \sum_{k=1}^{\infty}\left(\int_0^{\infty} re^{-(r+D)t}dt\right)^k\mathbb{E}\left[\mathbb{R}^{\text{trade}}_{t_k}\right] \\
&= \sum_{k=1}^{\infty}\left(\tfrac{r}{r+D}\right)^k\mathbb{E}\left[\mathbb{R}^{\text{trade}}_{t_k}\right] \\
&= \delta\sum_{k=0}^{\infty}\delta^k\mathbb{E}\left[\mathbb{R}^{\text{trade}}_{t_{k+1}}\right].
\end{aligned}$$

where δ is the "adjusted discount"factor given by $\frac{r}{r+D}$.

Discounted inventory costs in $\mathbb{R}^{\text{inv}}$ can be also computed by

$$\begin{aligned}
\mathbb{E}\left[\Gamma\int_0^{\infty} e^{-Dt}|x_t|dt\right] &= \Gamma\sum_{k=0}^{\infty}\mathbb{E}\left[\int_{t_k}^{t_{k+1}} e^{-Dt}|x_t|dt\right] \\
&= \Gamma\sum_{k=0}^{\infty}\mathbb{E}\left[\int_{t_k}^{t_{k+1}} e^{-Dt}dt\right]\mathbb{E}\left[|x_{t_k}|\right] \\
&= \tfrac{\Gamma}{D}\sum_{k=0}^{\infty}\mathbb{E}\left[e^{-Dt_k}\right]\left(1-\mathbb{E}\left[e^{-D\tau_{k+1}}\right]\right)\mathbb{E}\left[|x_{t_k}|\right] \\
&= \tfrac{\Gamma}{D}\sum_{k=0}^{\infty}\delta^k\left(\tfrac{D}{\lambda+\mu+D}\right)\mathbb{E}\left[|x_{t_k}|\right] \\
&= \tfrac{\Gamma}{r+D}\sum_{k=0}^{\infty}\left(\tfrac{r}{r+D}\right)^k\mathbb{E}\left[|x_{t_k}|\right].
\end{aligned}$$

We are now ready to define the total reward matrix. Let

$$\begin{aligned}
&\mathbb{R}\left((x',s',b',a',e',j')|(x,s,b,a,e,j),(\ell^b,\ell^a)\right) \\
&= \mathbb{1}\left(x+1=x'\right)\left((c-\phi j^-)\mathbb{1}\left(\ell^b=1,b\in\mathcal{P}\right)+(3c-\phi j^-)\mathbb{1}\left(\ell^b=2,b\in\mathcal{P}\right)\right) \\
&\quad+\mathbb{1}\left(x-1=x'\right)\left((c-\phi j^+)\mathbb{1}\left(\ell^a=1,b\in\mathcal{P}\right)+(3c-\phi j^+)\mathbb{1}\left(\ell^a=2,b\in\mathcal{P}\right)\right) \\
&\quad-\gamma|x|
\end{aligned}$$

where c, ϕ, and γ are defined as the "adjusted spread" "adjusted jump sizes " and "adjusted inventory aversion" parameters for the resulting discrete-time formulation of the model and given by

$$\delta \equiv \frac{r}{r+D}, \quad c \equiv \delta C, \quad \phi \equiv \delta J \ \text{ and } \ \gamma \equiv \frac{\Gamma}{r+D}. \tag{C.1}$$

Then, the HFT choose π to maximize the value function $V(x,s,b,a,e,j)$ which equals to

$$\mathbb{E}^{\pi}\left[\sum_{k=0}^{\infty}\delta^k\mathbb{R}\left((x'_{t_k},s'_{t_k},b'_{t_k},a'_{t_k},e'_{t_k},j'_{t_k})|(x_{t_{k-1}},s_{t_{k-1}},b_{t_{k-1}},a_{t_{k-1}},e_{t_{k-1}},j_{t_{k-1}}),(\ell^b_{t_{k-1}},\ell^a_{t_{k-1}})\right)\right], \tag{C.2}$$

starting from his initial state, (x,s,b,a,e,j). This derivation illustrates that our model can be cast in

the desired MDP form.

D. HFT's Value Function

We have now transformed our continuous-time problem into an equivalent discrete-time MDP. Using the Hamilton-Jacobi-Bellman optimality equations, $V(x,s,b,a,e,j)$ in (C.2) can be computed by solving the following set of equations:

$$V(x,s,b,a,e,j) = \max_{\ell^b,\ \ell^a} \Bigg\{ \sum_{(x',s',b',a',e',j')} \mathbb{P}\left((x',s',b',a',e',j')|(x,s,b,a,e,j),(\ell^b,\ell^a)\right) \times \\ \left[\mathbb{R}\left((x',s',b',a',e',j')|(x,s,b,a,e,j),(\ell^b,\ell^a)\right) + \delta V(x',s',b',a',e',j')\right] \Bigg\}. \tag{D.1}$$

By substituting the expressions for $\mathbb{P}$ and $\mathbb{R}$, we can obtain the implicit equations for the value functions of each state.

E. Optimal Market Making Solution

Proof Sketch of Theorem 1. Since the model is symmetric around the bid and ask side of the market, we can first eliminate the order direction signal from our state space. We have the following reduction for each $s_2 \in \{P, I\}$.

$$V(-x, Ss_2, b, a, e, j) = V(x, Bs_2, a, b, e, -j)$$

Using this result, we let $v(x,P,b,a,j) \equiv V(x,PS,b,a,0,j)$, $v(x,I,b,a,j) \equiv V(x,IS,b,a,0,j)$ and $h(x,P,\mathbf{rel}(b,a)) \equiv V(x,PS,b,a,1,0)$, $h(x,I,\mathbf{rel}(b,a)) \equiv V(x,IS,b,a,1,0)$ where $\mathbf{rel}(b,a)$ is a mapping that extracts the relevant information regarding the quoting of the MFT at both sides of the market. Recall that at both sides of the market, the HFT needs to know whether the MFT is quoting or not and with or without priority. So out of 7 possible states in each side of the book, the HFT needs to know whether the MFT is quoting or not and whether her orders have execution priority. These three cases are identified in the following subsets of the order book at each side: We let $l_b = 0$ ($l_a = 0$) to denote the case in which the MFT does not quote, i.e., $b = 00$, $b = 10$ and $b = 20$ ($a = 00$, $a = 10$ and $a = 20$). Similarly, we let $l_b = 1$ ($l_a = 1$ if the MFT quotes but the HFT has the execution priority, i.e., $b = 11h$ ($a = 11h$). Finally, we let $l_b = 2$ ($l_a = 2$) if $b = 01$, $b = 11m$ and $b = 21$ ($a = 01$, $a = 11m$ and $a = 21$) i.e., MFT quotes with execution priority. Finally, we let $l_b = 2$ ($l_a = 2$ if the MFT quotes but the HFT has the execution priority, i.e., $b = 11h$ ($a = 11h$). Thus, only 9 different

cases are relevant for the HFT's quoting decisions from the complete 49 states of the order book and $\mathbf{rel}(b,a)$ can be given by

	Ask (a)						
Bid (b)	00	10	20	21	01	$11m$	$11h$
00	00	00	00	02	02	02	01
10	00	00	00	02	02	02	01
20	00	00	00	02	02	02	01
21	20	20	20	22	22	22	21
21	20	20	20	22	22	22	21
21	20	20	20	22	22	22	21
21	10	10	10	12	12	12	11

(E.1)

Using the value iteration algorithm, we first establish by induction that value functions are concave in x. We also need to show that as inventory gets larger (smaller), less quoting at the bid (ask) side will be more and more attractive. Formally, we need to show that $v(x,s,00,a,j) - v(x,s,20,a,j)$, $v(x,s,01,a,j) - v(x,s,21,a,j)$, $v(x,s,21,a,j) - v(x,s,11m,a,j)$ and $v(x,s,21,a,j) - v(x,s,11h,a,j)$ will be nondecreasing in x for any fixed state a and the symmetric conditions for the ask side. We will refer to this condition as the "nondecreasing property." The threshold policy in Theorem 1 follows using this non-decreasing property by defining the appropriate limits.

Let $v^{(0)}(x,s,b,a,j) = 0$ for all (x,s,b,a,j). Then, in the base case, all value functions will include $\gamma|x|$ and an appropriate constant term that equals to the expectation of earning the spread plus the potential adverse selection cost as implied by the $\mathbb{R}^{\text{trade}}$. Therefore, the base case will satisfy the concavity and non-decreasing property. Assume that $v^{(n)}(x,s,b,a,j)$ satisfies the induction hypothesis. We will show the inductive step of the nondecreasing property using an illustrative example. The remaining cases are very similar.

First, we show that $v^{(n+1)}(x,P,20,00,j) - v^{(n+1)}(x,P,10,00,j)$ is nondecreasing in x. If $j \neq 0$,

$$
\begin{aligned}
v^{(n+1)}(x,P,20,00,j) - v^{(n+1)}(x,P,10,00,j) = \delta\Big\{ & \frac{\lambda^{\mathsf{sell}} + \theta}{r}\Big(\mathsf{pr}(PS)\left(v^{(n)}(x,P,20,00,j) - v^{(n)}(x,P,10,00,j)\right) \\
& + \mathsf{pr}(IS)\left(v^{(n)}(x,I,20,00,j) - v^{(n)}(x,I,10,00,j)\right) + \mathsf{pr}(PB)\left(z^{(n)}(x,I,20,00,j) - z^{(n)}(x,I,10,00,j)\right) \\
& + \mathsf{pr}(IB)\left(z^{(n)}(x,I,20,00,j) - z^{(n)}(x,I,10,00,j)\right)\Big) + \frac{\sigma}{2r}\Big(v^{(n)}(x,P,20,00,j-1) - v^{(n)}(x,P,10,00,j-1)\Big) \\
& + \frac{\beta}{r}\Big(v^{(n)}(x,P,21,00,j) - v^{(n)}(x,P,11h,00,j+1)\Big) + \frac{\beta}{r}\Big(v^{(n)}(x,P,20,01,j) + v^{(n)}(x,P,10,01,j)\Big) \\
& + \frac{\sigma}{2r}\Big(v^{(n)}(x,P,20,00,j+1) - v^{(n)}(x,P,10,00,j+1)\Big) + \frac{\lambda_A}{r}\Big(v^{(n)}(x,P,20,00,j) + v^{(n)}(x,P,10,00,j)\Big)\Big\}
\end{aligned}
$$

is also nondecreasing in x as each term satisfies the non-decreasing property via the induction hypothesis. If $j = 0$, we have the following additional terms:

$$\frac{\lambda^{PI} m_s^{PS}}{r}\Big(\mathsf{pr}(PS)\left(v^{(n)}(x,P,20,00,j) - v^{(n)}(x+1,P,00,00,j)\right) + \mathsf{pr}(IS)\left(v^{(n)}(x,I,20,00,j) - v^{(n)}(x+1,I,00,00,j)\right) \\ + \mathsf{pr}(PB)\left(z^{(n)}(x,I,20,00,j) - z^{(n)}(x+1,I,00,00,j)\right) + \mathsf{pr}(IB)\left(z^{(n)}(x,I,20,00,j) - z^{(n)}(x+1,I,00,00,j)\right)\Big)$$

which equals to

$$\frac{\lambda^{PI} m_s^{PS}}{r}\Big(\mathsf{pr}(PS)\left(v^{(n)}(x,P,20,00,j) - v^{(n)}(x,P,00,00,j)\right) + \mathsf{pr}(IS)\left(v^{(n)}(x,I,20,00,j) - v^{(n)}(x,I,00,00,j)\right) \\ + \mathsf{pr}(PB)\left(z^{(n)}(x,I,20,00,j) - z^{(n)}(x,I,00,00,j)\right) + \mathsf{pr}(IB)\left(z^{(n)}(x,I,20,00,j) - z^{(n)}(x,I,00,00,j)\right) \\ + \mathsf{pr}(PS)\left(v^{(n)}(x,P,00,00,j) - v^{(n)}(x+1,P,00,00,j)\right) + \mathsf{pr}(IS)\left(v^{(n)}(x,I,00,00,j) - v^{(n)}(x+1,I,00,00,j)\right) \\ + \mathsf{pr}(PB)\left(z^{(n)}(x,I,00,00,j) - z^{(n)}(x+1,I,00,00,j)\right) + \mathsf{pr}(IB)\left(z^{(n)}(x,I,00,00,j) - z^{(n)}(x+1,I,00,00,j)\right)\Big)$$

which is also nondecreasing in x as the first four terms satisfies the non-decreasing property via the induction hypothesis, and the last four terms satisfy the concavity in x via the induction hypothesis. Finally, we can conclude that each $v^{(n+1)}$ is concave in x by showing that $h^{(n)}$ is concave. This holds as $h^{(n)}$ can be expressed as the maximum of 9 value function of $v^{(n)}$ corresponding to the HFT's quoting actions. In each inventory region $h^{(n)}$ stays concave in x as we can write $h^{(n)}(x,s,l) - h^{(n)}(x+1,s,l)$ as positive sum of nondecreasing functions. □