High Frequency Traders: Taking Advantage of Speed

Yacine Aït-Sahalia[*]
Department of Economics
Bendheim Center for Finance
Princeton University
and NBER

Mehmet Saglam[†]
Department of Economics
Bendheim Center for Finance
Princeton University

September 26, 2013

Abstract

We propose a model of dynamic trading where a strategic high frequency trader receives an imperfect signal about future order flows, and exploits his speed advantage to optimize his quoting policy. We determine the provision of liquidity, order cancellations, and impact on low frequency traders as a function of both the high frequency trader's latency, and the market volatility. The model predicts that volatility leads high frequency traders to reduce their provision of liquidity. Finally, we analyze the impact of various policies designed to potentially regulate high frequency trading.

Keywords: High Frequency Trading, Liquidity, Order Cancellations, Competition for Order Flow, Tobin Tax, Order Resting Time, Order Cancellation Tax.
JEL Classification: G10.

[*] Email: yacine@princeton.edu
[†] Email: msaglam@princeton.edu

1. Introduction

High-frequency traders (HFTs) have become a potent force in many equity and futures markets. HFTs invest in and develop 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 their large investment in trading technology.

Speed has always been of the essence in financial markets. Traders were among the first to adopt the telegraph and then the telephone. Closer to us, "high frequency" meant anything intraday; then minute-by-minute transactions became the norm, quickly to be replaced by second-by-second time stamps. The time it now takes for an order to be sent and displayed on an electronic exchange, that is, the latency associated with the implementation of an order, is currently measured in milliseconds. The latest software and infrastructure developments are making it possible for the most cutting-edge trading firms to implement microsecond-based algorithms.

Measuring the extent of HFT activity is inherently a complex task. Although estimates vary across markets, trading venues and time period, it is generally thought that HFTs represent anywhere 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)). The trend points towards high frequency trading having perhaps plateaued in some of the markets where it was first introduced, but still expanding globally in new markets.

This significant amount of market activity has been accompanied by theoretical research addressing some of issues arising from the rapid rise of HFTs[1]. In particular, Foucault et al. (2012) consider a setting in which a HFT enjoys a speed advantage when gathering information. They compare two models of trading under asymmetric information which differ in the presence of an informed trader precessing information faster than the other market participants, and 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. They focus in

[1] See for example Pagnotta and Philippon (2011), Jarrow and Protter (2012), Moallemi and Saglam (2012), Pagnotta (2010), Cespa and Foucault (2008).

particular on liquidity cycles, where traders compete alternatively to make (by positing limit orders) and take (by sending market orders) liquidity from the market. Biais et al. (2011) 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. They find that the introduction of HFTs results in more mass around the center of the distribution and thinner tails.

The innovation in our paper is the modeling of a fully dynamically optimizing HFT. The dynamic optimization nature of the quoting process by HFTs is a departure from the existing literature, allowing us to address a new set of inherently dynamic questions: we use the model to study the quoting optimization by a HFT, initially in a monopolistic position, who enjoys a latency advantage and trades against many uninformed LFTs; quantify the HFT's latency advantage over LFTs; determine why HFTs cancel their orders at a very rapid rate; how their (endogenous) inventory constraints help shape their order placement and cancellation strategies; how the HFT's provision of liquidity can be expected to change in different market environments, such as high volatility ones; and how competition among HFTs can be expected to affect the provision of liquidity and the welfare of LFTs.

Additionally, a number of regulatory proposals have recently been discussed, and in some countries implemented, generally aiming at curbing the growth of trading by HFTs. Some of the proposed provisions include a transaction tax, limiting the ability of HFTs to cancel orders before some amount of time has elapsed, or taxing such cancellations. Using the model of the paper, we analyze the potential impact of these proposed regulations.

Technically, the innovation is the use of multiple, staggered, Poisson processes to represent the arrival of the various elements of the model: market orders by LFTs, signals to the HFT, changes in the asset's fundamental price, arrival of a competing trader who also provides liquidity, etc. This modeling device keeps the analysis of the dynamic optimization problem facing the HFT tractable and flexible, and makes it possible to analyze different market environments by varying the parameters of the model.

In our model, a HFT is able to (imperfectly) predict the upcoming order flow and exploit his speed advantage by placing and possibly canceling limit orders ahead of incoming market orders by LFTs. The HFT receives a private signal about the likely type of the next upcoming market order, buy or sell, and decides to quote or not to quote a limit order at the best available prices. These signals can be interpreted as the observation of microstructure-level market events that are informative about the direction of the order flow, such as limit order book imbalances, or new trades happening at different ticks in a correlated asset. HFT technology makes it possible not only to observe and process these signals at high frequency, but also to send in orders or cancellations, and trade, in response to them.

Our modeling choices are informed by some of the main empirical stylized facts that are known about HFTs (see e.g., Brogaard (2011) and Brunetti et al. (2011)). They include the fact that HFTs have a speed advantage in terms of order placement/cancellation; that there is a small number of HFTs relative to the mass of LFTs; that HFTs are recognizable by their high frequency of execution, a high number of trades, a high total volume traded, a small volume on each trade and the fact that they carry a low inventory as their primary risk control strategy. HFTs act primarily as market makers, unwilling to take directional bets. HFTs also tend to place many orders, with only some actually leading to execution, and many cancelled; they appear to systematically beat the odds when trading against LFTs; and they seem to exploit order flow information and generate trading signals on a very short time scale rather than longer-run information about the fundamental value of the asset.

This paper yields several new results. We show that, for fully optimizing HFTs, lower latency translates into higher profits, higher liquidity provision and higher cancellation rates in normal times. This is 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 have the potential to improve market quality by providing liquidity, contributing to price discovery, improving market efficiency and easing market fragmentation (see e.g., Hasbrouck and Saar (2010), Hendershott et al. (2011), Chaboud et al. (2010), Brogaard et al. (2012) and Menkveld (2013)).

However, what also emerges out of the model is also the fact that HFTs' provision of liquidity can be expected to decrease when price volatility picks up. 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 prediction of the model is particularly salient in light of the evidence that has emerged regarding flash crashes (see Easley et al. (2011) and Kirilenko et al. (2010)).

When we introduce competition for order flow with another trader, and derive the change in the HFT's optimal quoting policy, we find that the two market makers split the rent extracted from LFTs, liquidity provision increases and LFTs tend to be better off.

Finally, we analyze the possible impact of three widely discussed HFT policies: imposing a transaction tax on each trade, setting minimum-time limits before orders can be cancelled, and taxing the cancellations of limit orders. We find that, in the context of our model, imposing minimum time-limits and cancellation taxes induces the HFT to quote more on both sides of the market, whereas transaction taxes do not improve this measure of liquidity. One important finding is that when minimum time-limits are in effect, the fill rate of LFTs' market orders by the HFT does not decrease substantially in the presence of higher volatility, unlike the situation without minimum resting times.

The paper is organized as follows. Section 2 sets up the base model in continuous time. Section 3 derives the optimal quoting policy of the HFT. Section 4 provides a comparative statics analysis of the model, while Section 5 develops the implications of the model for market structure, including the HFT's provision of liquidity, order cancellations, and the fill rate of market orders. In Sections 6 and 7, we introduce two extensions of the base model: first, to price volatility, and second to duopolistic competition with another market maker. Section 8 analyzes in the context of the model the impact of possible HFT regulations. Section 9 concludes.

2. The Base Model

We consider a model where a HFT and a large number of LFTs are trading a single asset in an electronic limit order book. The LFTs are uninformed and submit market orders which arrive at random times according to a Poisson process with parameter λ.

By contrast, the HFT is market-making, employing only limit orders. At each instant $t = 0, 1, \ldots$, the fundamental price of the asset is denoted by S_t. The ask price is $S_t + C/2$ and the bid price is given by $S_t - C/2$ for a constant bid-ask spread $C > 0$. Initially, we assume that S_t is constant, so the only price moves are bid/ask bounces in response to transactions. We will relax this assumption in Section 6.

The HFT participates in the market by posting active limit orders to buy and sell at the best price in the book. We assume initially that the HFT enjoys a monopoly in terms of providing limit orders at the best price, an assumption we will relax in Section 7. Market orders are automatically and instantaneously executed against the limit orders. The objective of the HFT is to capture the spread as often as possible.

The quantity of each order, market or limit, is fixed at 1 lot, so the HFT is not optimizing over the quantity in each trade. Small volume on each trade matches what is observed empirically in markets that are popular with HFTs, such as the S&P500 eMini futures. Generally speaking, the quantity exchanged in each trade has been going down over time (see, e.g., Angel et al. (2010)).

The HFT can post limit orders at the best bid ($\ell^b = 1$) or the best ask price ($\ell^a = 1$). He may quote on either, both, or neither side of the market. He can also cancel his previously posted limit orders, thereby withdrawing from providing liquidity.

Besides speed, we endow the HFT with another advantage over LFTs. The HFT observes a signal s that is informative about the likely sign of the next incoming market order. The signal arrives as a Poisson process with rate μ, which is independent of the arrival process of the market orders. The arrival rate of signals received by the HFT is much larger than the arrival rate of the market orders by LFTs, i.e., $\mu \gg \lambda$. The arrival rate of the signal is larger when the HFT has better trading technology, i.e., he has the ability to process information at a higher rate.

The signal is an i.i.d. Bernoulli random variable, $s \in \{\text{“sell”}, \text{“buy”}\}$ with each being equally likely[2]. Each signal supersedes all previous signals. Conditional on a “sell” signal, the next market order by a LFT to buy will arrive with a rate of $p\lambda$ and the next market order to sell will arrive with a rate $(1-p)\lambda$ where $p \geq 1/2$, and vice versa if the signal is “buy.” Therefore, conditionally on the signal telling the HFT to “sell”, the next market order by a LFT will be a buy order with probability p, so if he followed his signal to sell the HFT can expect his quote to sell to be crossed with the incoming market order to buy with probability p, and a sell order with probability $1-p$. The parameter p measures the quality of the signal, starting from $p = 1/2$, where the signal is uninformative, and increasing all the way to $p = 1$, where the signal is perfectly predictive of the sign of the next incoming market order. With the signal equally likely to be “buy” or “sell”, p does not change the unconditional arrival rate of the market orders, which is $\lambda/2$ for both buy and sell orders sent by the LFTs.

This mechanism is meant to model the realistic situation where the HFT is able to extract information in real time from perhaps the current state of the limit order book, including any imbalances, from the recent trading patterns that may be predictive of future orders, or from data acquired ahead of other traders about trades about to be executed or confirmed, all obtained and processed on a very short (typically, millisecond) time scale. The idea is that the HFT is exploiting fleeting trading opportunities arising from the trading process itself.

The HFT makes quoting decisions immediately after observing either a signal or market order. Hence, on average, he is able to make new quoting decisions every $1/(\lambda + \mu)$ time units. When the arrival rate of signals increases, the HFT has the ability to make faster quoting revisions.

The HFT's position in the asset is denoted by x_t. This position can be positive or negative; in

[2] This can be generalized to an arbitrary Markov process, including one with serial correlation in the signal.

the case of a stock this means that we impose no restrictions on short selling, while in the case of a futures contract a positive (resp. negative) value of x_t denotes a long (resp. short) position in the contract. We assume that the HFT is risk-neutral, but is penalized for holding excess inventory at a rate of $\Gamma|x_t|$ where Γ is a constant parameter of inventory aversion. In practice, limiting or penalizing inventory is indeed one of the primary sources of risk mitigation by HFTs. It is also the reason that, despite the lack of competition and the constancy of the price in the basic model, the HFT will not systematically quote on both sides of the market and attempt to systematically capture every spread.

The HFT's objective is therefore to maximize his expected discounted rewards earned from the bid-ask spread minus the penalty costs from holding an inventory. The discount rate $D > 0$ is assumed to be constant. Let π be any feasible policy that chooses ℓ_t^b and ℓ_t^a at trading decision times. Formally, the HFT maximizes

$$\max_{\pi} \mathbb{E}^{\pi} \left[\frac{C}{2} \sum_{i=1}^{\infty} e^{-DT_i^{\mathsf{sell}}} \mathbb{1}\left(\ell^b_{T_i^{\mathsf{sell}}} = 1\right) + \frac{C}{2} \sum_{j=1}^{\infty} e^{-DT_j^{\mathsf{buy}}} \mathbb{1}\left(\ell^a_{T_j^{\mathsf{buy}}} = 1\right) - \Gamma \int_0^{\infty} e^{-Dt} |x_t| dt \right] \tag{2.1}$$

where T_i^{sell} is the ith market sell order and T_j^{buy} is the jth market buy order. The first term represents the HFT's spread gain from a market order to sell crossed against his bid limit order, the second the HFT's spread gain from a market order to buy crossed against his ask limit order, and the third his inventory penalty. We do not need to model explicitly any rebate provided by the exchange to market makers. From the perspective of the HFT in the model, it can be thought of as an increase in C.

To keep the model tractable and focus only on the essential, a number of elements are necessarily left out of the model; first, HFTs in our model are market makers and as such do not make a strategic choice between limit and market orders, but employ limit orders only; second, HFTs do not place orders larger than for one contract; third, HFTs' limit orders are always placed at the best bid and ask prices. The latter means that we exclude order placement strategies known as "quote stuffing" that place large numbers of quotes away from the best prices to falsely give the impression to other traders of an incoming imbalance, presumably without the intent of ever executing these orders. It also means that we preclude the use by HFTs of the now-banned "stub quotes", which are place-holding quotes far from the current market price, employed by market makers to post quotes without any desire to trade, but which may become relevant in a flash crash.

3. Dynamic Optimization by the HFT

We now turn to solving for the HFT's optimal strategy. The key advantage of our model's Poisson-based setup is that we can merge the two Poisson time clocks λ and μ into a single one, with signal or market orders arriving at a combined Poisson rate of $\lambda + \mu$, and determine the HFT's decisions at the resulting discrete random times. That is, the HFT's maximization problem in continuous-time can be equivalently converted into a tractable discrete-time problem using the *uniformization* methodology (see e.g., Puterman (1994)). We consider the continuous-time HFT problem at signal (arrival μ) or market order (arrival λ) arrival times. The HFT's trading decisions at each of these arrival times do not affect the timing of the next trading decision. Therefore, we can actually set-up our problem in discrete-time where HFT decisions are undertaken at fixed time intervals. We describe this methodology in the following section.

3.1. Discrete-time Equivalent Formulation

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 a MDP. 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 epoch, and finally $\mathbb{R}(.|i,a)$ is the reward matrix, specifying the reward obtained using action a when the state is in i. The decision maker seeks a policy that maximizes the expected discounted reward

$$v(i) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t \mathbb{R}(i_{t+1}|i_t, \pi(i_t))|i_0 = i\right], \tag{3.1}$$

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

$$\begin{aligned} v(i) &= \max_{\pi} \left\{ \sum_{i'} \mathbb{P}(i'|i,\pi(i)) \left(\mathbb{R}(i'|i,\pi(i)) + \alpha \mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^{t-1} \mathbb{R}(i_{t+1}|i_t,\pi(i_t))|i_1 = i' \right] \right) \right\} \\ &= \max_{\pi} \left\{ \sum_{i'} \mathbb{P}(i'|i,\pi(i)) \left(\mathbb{R}(i'|i,\pi(i)) + \alpha \mathbb{E}\left[\sum_{k=0}^{\infty} \alpha^{k} \mathbb{R}(i_{k+1}|i_k,\pi(i_k))|i_k = i' \right] \right) \right\} \\ &= \max_{a \in A_i} \left\{ \sum_{i'} \mathbb{P}(i'|i,a) \left(\mathbb{R}(i'|i,a) + \alpha v(i') \right) \right\}. \end{aligned} \tag{3.2}$$

3.1.1 State and Action Space

Specializing the general framework of MDPs to our problem, the state space can be represented by the pair of (x, s) where x denotes the current holdings of the HFT with $x \in \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ and s is the most recent signal received by the HFT, with $s \in \{1, -1\}$. Here "1" denotes the "buy" signal and "-1" denotes the "sell" signal. The corresponding action taken by the HFT at each state is whether to quote a limit order or not at the best bid and/or best ask, i.e., $\ell_t^b(x, s) \in \{0, 1\}$ and $\ell_t^a(x, s) \in \{0, 1\}$. Going from 0 to 1 means putting a new limit order in place; going from 1 to 0 means canceling an existing limit order; keeping the action at 1 means keeping an existing order alive longer; and keeping it at 0 means continuing not to quote on that side of the market.

3.1.2 Transition Probabilities

We now calculate the transition probabilities at each state of the HFT. First, note that the state transitions occur at a rate of $\lambda+\mu$ and the rate is the same for all states and actions. Let $\mathbb{P}((x', s')|(x, s), (\ell^b, \ell^a))$ be the probability of reaching state (x', s') when the system is in state (x, s) and the trader takes the actions of ℓ^b and ℓ^a. First, we define

$$\mathsf{pr}(s) = \begin{cases} p & \text{if } s = 1, \\ 1-p & \text{if } s = -1. \end{cases}$$

Suppose that the current state of the HFT is (x, s). Let τ^{sell} and τ^{buy} be the random arrival times of the market-sell and market-buy order, respectively. Note that τ^{sell} and τ^{buy} have exponential distributions with means $1/\left(\mathsf{pr}(s)\lambda\right)$ and $1/\left((1-\mathsf{pr}(s))\lambda\right)$, respectively. Similarly, let τ^{same} and τ^{opp} be the random arrival time of the same signal and opposite signal. Note that both these times have exponential distributions with mean $2/\mu$. Finally, let $\tau \equiv \min\left(\tau^{\mathsf{sell}}, \tau^{\mathsf{buy}}, \tau^{\mathsf{same}}, \tau^{\mathsf{opp}}\right)$ which has exponential distribution with mean $1/(\lambda + \mu)$. Using this notation, we provide the transition probabilities with respect to each action taken.

First, if the HFT does not quote, we obtain

$$\mathbb{P}\left((x', s')|(x, s), (0,0)\right) = \begin{cases} \frac{\mu/2}{\mu+\lambda} \equiv \mathbb{P}\left(\tau^{\mathsf{opp}} = \tau\right) & \text{if } x = x', s \neq s', \\ 1 - \frac{\mu/2}{\mu+\lambda} \equiv \mathbb{P}\left(\tau^{\mathsf{opp}} \neq \tau\right) & \text{if } x = x', s = s', \\ 0 & \text{otherwise.} \end{cases}$$

When the HFT does not quote, the only transition to a different state is due to the arrival of a signal of the opposite sign. The corresponding probability is given by

$$\begin{aligned} \mathbb{P}\left(\tau^{\mathsf{opp}} = \tau\right) &= \mathbb{P}\left(\tau^{\mathsf{opp}} = \min\left\{\tau^{\mathsf{opp}}, \min\left(\tau^{\mathsf{sell}}, \tau^{\mathsf{buy}}, \tau^{\mathsf{same}}\right)\right\}\right), \\ &= \int_0^\infty (\mu/2) e^{-\mu x/2} dx \int_x^\infty \left(\mu/2 + \lambda\right) e^{-(\mu/2+\lambda) y} dy, \\ &= \int_0^\infty (\mu/2) e^{-(\mu+\lambda)x} dx, \\ &= \frac{\mu/2}{\mu+\lambda}, \end{aligned}$$

where $\min\left(\tau^{\mathsf{sell}}, \tau^{\mathsf{buy}}, \tau^{\mathsf{same}}\right)$ is exponentially distributed with mean $1/(\mu/2 + \lambda)$. This assumption can be easily generalized. If $Z_1, \ldots, Z_n$ are each exponentially distributed with mean $1/z_i$, then $\mathbb{P}\left(Z_k = \min\left\{Z_1, \ldots, Z_n\right\}\right) = z_k/(z_1 + \ldots + z_n)$. We will use this fact in the remaining cases.

When the HFT takes action $(1,0)$, we have the following transition probabilities:

$$\mathbb{P}\left((x', s')|(x, s), (1,0)\right) = \begin{cases} \frac{\mu/2}{\mu+\lambda} \equiv \mathbb{P}\left(\tau^{\mathsf{opp}} = \tau\right) & \text{if } x = x', s \neq s', \\ \frac{\mathsf{pr}(s)\lambda}{\mu+\lambda} \equiv \mathbb{P}\left(\tau^{\mathsf{sell}} = \tau\right) & \text{if } x + 1 = x', s = s', \\ \frac{(1-\mathsf{pr}(s))\lambda + \mu/2}{\mu+\lambda} \equiv \mathbb{P}\left(\min\left(\tau^{\mathsf{buy}}, \tau^{\mathsf{same}}\right) = \tau\right) & \text{if } x = x', s = s', \\ 0 & \text{otherwise.} \end{cases}$$

When the HFT takes the action $(1,0)$, he may increase his inventory by trading with the incoming market-sell order, which occurs with probability $\mathsf{pr}(s)\lambda/(\mu+\lambda)$. Similarly, for HFT action $(0,1)$, we have

$$\mathbb{P}\left((x',s')|(x,s),(0,1)\right)=\begin{cases}\frac{\mu/2}{\mu+\lambda}\equiv\mathbb{P}\left(\tau^{\mathsf{opp}}=\tau\right) & \text{if } x=x', s\neq s',\\ \frac{(1-\mathsf{pr}(s))\lambda}{\mu+\lambda}\equiv\mathbb{P}\left(\tau^{\mathsf{buy}}=\tau\right) & \text{if } x-1=x', s=s',\\ \frac{\mathsf{pr}(s)\lambda+\mu/2}{\mu+\lambda}\equiv\mathbb{P}\left(\min\left(\tau^{\mathsf{sell}},\tau^{\mathsf{same}}\right)=\tau\right) & \text{if } x=x', s=s',\\ 0 & \text{otherwise.}\end{cases}$$

Finally, when the HFT quotes on both sides of the market, his inventory may increase with probability $\mathsf{pr}(s)\lambda/(\mu+\lambda)$, and decrease with probability $(1-\mathsf{pr}(s))\lambda/(\mu+\lambda)$:

$$\mathbb{P}\left((x',s')|(x,s),(1,1)\right)=\begin{cases}\frac{\mu/2}{\mu+\lambda}\equiv\mathbb{P}\left(\tau^{\mathsf{opp}}=\tau\right) & \text{if } x=x', s\neq s',\\ \frac{\mathsf{pr}(s)\lambda}{\mu+\lambda}\equiv\mathbb{P}\left(\tau^{\mathsf{sell}}=\tau\right) & \text{if } x+1=x', s=s',\\ \frac{(1-\mathsf{pr}(s))\lambda}{\mu+\lambda}\equiv\mathbb{P}\left(\tau^{\mathsf{buy}}=\tau\right) & \text{if } x-1=x', s=s',\\ \frac{\mu/2}{\mu+\lambda}\equiv\mathbb{P}\left(\tau^{\mathsf{same}}=\tau\right) & \text{if } x=x', s=s',\\ 0 & \text{otherwise.}\end{cases}$$

3.1.3 HFT's Reward Function

Let $R((x',s')|(x,s),(\ell^b,\ell^a))$ be the total reward achieved by the HFT when the system is in state (x,s), the HFT chooses quoting actions ℓ^b and ℓ^a and the system reaches the state (x',s'). We would like to write the HFT's objective in (2.1) in the form of an MDP objective function as in (3.1). We first introduce the following notation. Let t_k be the time of the kth state transition due to a signal or market order arrival (by convention $t_0=0$) and let τ_k be the length of this cycle, i.e., $\tau_k=t_k-t_{k-1}$.

We start with the first two terms in (2.1) that measure the spreads earned by the HFT when there is a trade. First observe that

$$\frac{C}{2}\sum_{i=1}^{\infty}e^{-DT_i^{\mathsf{sell}}}\mathbb{1}\left(\ell^b_{T_i^{\mathsf{buy}}}=1\right)+\frac{C}{2}\sum_{j=1}^{\infty}e^{-DT_j^b}\mathbb{1}\left(\ell^a_{T_j^b}=1\right)=\frac{C}{2}\sum_{k=1}^{\infty}e^{-Dt_k}R^+(x_{t_k}|x_{t_{k-1}},\ell^b_{t_{k-1}},\ell^a_{t_{k-1}}),$$

where we define $R^+(x'|x,\ell^b,\ell^a)=\mathbb{1}\left(x\neq x'\right)\mathbb{1}\left(\ell^a=1 \text{ or } \ell^b=1\right)$. 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 $1/(\lambda+\mu)$:

$$\begin{aligned}
\mathbb{E}\left[\tfrac{C}{2}\textstyle\sum_{k=1}^{\infty} e^{-Dt_k} R^{+}(x_{t_k}|x_{t_{k-1}}, \ell^b_{t_{k-1}}, \ell^a_{t_{k-1}})\right] &= \tfrac{C}{2}\textstyle\sum_{k=1}^{\infty} \mathbb{E}\left[e^{-D\sum_{i=1}^{k}\tau_i}\right] \mathbb{E}\left[R^{+}(x_{t_k}|x_{t_{k-1}}, \ell^b_{t_{k-1}}, \ell^a_{t_{k-1}})\right] \\
&= \tfrac{C}{2}\textstyle\sum_{k=1}^{\infty} \mathbb{E}\left[e^{-D\tau_1}\right]^k \mathbb{E}\left[R^{+}(x_{t_k}|x_{t_{k-1}}, \ell^b_{t_{k-1}}, \ell^a_{t_{k-1}})\right] \\
&= \tfrac{C}{2}\textstyle\sum_{k=1}^{\infty} \left(\int_0^{\infty}(\lambda+\mu)e^{-(\lambda+\mu+D)t}dt\right)^k \mathbb{E}\left[R^{+}(x_{t_k}|x_{t_{k-1}}, \ell^b_{t_{k-1}}, \ell^a_{t_{k-1}})\right] \\
&= \tfrac{C}{2}\textstyle\sum_{k=1}^{\infty} \left(\frac{\lambda+\mu}{\lambda+\mu+D}\right)^k \mathbb{E}\left[R^{+}(x_{t_k}|x_{t_{k-1}}, \ell^b_{t_{k-1}}, \ell^a_{t_{k-1}})\right] \\
&= \tfrac{\delta C}{2}\textstyle\sum_{k=0}^{\infty} \delta^k \mathbb{E}\left[R^{+}(x_{t_{k+1}}|x_{t_k}, \ell^b_{t_k}, \ell^a_{t_k})\right].
\end{aligned}$$

where we define the "adjusted discount factor," δ, by

$$\delta \equiv \frac{\lambda+\mu}{\lambda+\mu+D}. \tag{3.3}$$

Inventory costs in the third term of (2.1) can be simplified as

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

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

$$\mathbb{R}\left((x', s')|(x, s), (\ell^b, \ell^a)\right) = \frac{c}{2}\mathbb{1}\left(x \neq x'\right)\mathbb{1}\left(\ell^a = 1 \text{ or } \ell^b = 1\right) - \gamma|x| \qquad \forall (s, s'), \tag{3.4}$$

where

$$c \equiv \delta C \ \text{ and } \ \gamma \equiv \frac{\Gamma}{\lambda+\mu+D}, \tag{3.5}$$

become the "adjusted spread" and "adjusted inventory aversion" parameters for the discrete-time formulation. Then, the HFT maximizes

$$v(x, s) = \max_{\pi}\mathbb{E}^{\pi}\left[\sum_{k=0}^{\infty}\delta^k\mathbb{R}\left((x_{t_{k+1}}, s_{t_{k+1}})|(x_{t_k}, s_{t_k}), (\ell^b_{t_k}, \ell^a_{t_k})\right)\Big|(x_0, s_0) = (x, s)\right], \tag{3.6}$$

starting from his initial state, (x, s), which is in the requisite MDP form.

3.1.4 HFT's Value Function

We have now transformed our continuous-time problem into an equivalent discrete-time MDP. Using the Hamilton-Jacobi-Bellman optimality equation given in (3.2), $v(x,s)$ in (3.6) can be computed by solving the following set of equations:

$$v(x,s) = \max_{\ell^b,\ \ell^a} \left\{ \sum_{(x',s')} \mathbb{P}\left((x',s')|(x,s),(\ell^b,\ell^a)\right) \left\{ \mathbb{R}\left((x',s')|(x,s),(\ell^b,\ell^a)\right) + \delta v(x',s') \right\} \right\}. \tag{3.7}$$

By substituting the corresponding expressions for $\mathbb{P}\left(.|(x,s),(\ell^b,\ell^a)\right)$ and $\mathbb{R}\left(.|(x,s),(\ell^b,\ell^a)\right)$, the following lemma simplifies the implicit equations for the value functions in (3.7) and shows that the optimal decisions on ℓ^b and ℓ^a are separable:

Lemma 1.

$$\begin{aligned}
v(x,1) = -\gamma|x| + \delta\Bigg(& \frac{\mu/2}{\lambda+\mu}\left(v(x,1)+v(x,-1)\right) + \frac{p\lambda}{\lambda+\mu}\max\left\{\frac{c}{2\delta} + v(x+1,1), v(x,1)\right\} \\
& + \frac{(1-p)\lambda}{\lambda+\mu}\max\left\{\frac{c}{2\delta} + v(x-1,1), v(x,1)\right\}\Bigg) \\
v(x,-1) = -\gamma|x| + \delta\Bigg(& \frac{\mu/2}{\lambda+\mu}\left(v(x,1)+v(x,-1)\right) + \frac{p\lambda}{\lambda+\mu}\max\left\{\frac{c}{2\delta} + v(x-1,-1), v(x,-1)\right\} \\
& + \frac{(1-p)\lambda}{\lambda+\mu}\max\left\{\frac{c}{2\delta} + v(x+1,-1), v(x,-1)\right\}\Bigg)
\end{aligned}$$

Proof. We prove the result only for $v(x,1)$, since the second equation is derived analogously. Using (3.7), we substitute for the transition probabilities $\mathbb{P}$ and the reward function $\mathbb{R}$ derived in Section 3.1.2 and Section 3.1.3. We obtain the following expression involving a maximum where each term corresponds to the expected value taking actions $(0,0)$, $(1,0)$, $(0,1)$ and $(1,1)$ respectively.

$$\begin{aligned}
v(x,1) = \max\Bigg\{ & \tfrac{\mu/2}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,-1)\right) + \left(1 - \tfrac{\mu/2}{\lambda+\mu}\right)\left(-\gamma|x| + \delta v(x,1)\right), \\
& \tfrac{\mu/2}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,-1)\right) + \tfrac{p\lambda}{\lambda+\mu}\left(\tfrac{c}{2} - \gamma|x| + \delta v(x+1,1)\right) + \tfrac{\mu/2+(1-p)\lambda}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,1)\right), \\
& \tfrac{\mu/2}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,-1)\right) + \tfrac{(1-p)\lambda}{\lambda+\mu}\left(\tfrac{c}{2} - \gamma|x| + \delta v(x-1,1)\right) + \tfrac{\mu/2+p\lambda}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,1)\right), \\
& \tfrac{\mu/2}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,-1)\right) + \tfrac{\mu/2}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,1)\right) + \tfrac{p\lambda}{\lambda+\mu}\left(\tfrac{c}{2} - \gamma|x| + \delta v(x+1,1)\right) \\
& + \tfrac{(1-p)\lambda}{\lambda+\mu}\left(\tfrac{c}{2} - \gamma|x| + \delta v(x-1,1)\right) \Bigg\}.
\end{aligned}$$

We can rearrange the first three terms in the maximum and write them in the same form as the last term. This shows that we can actually separate the maximum into two maxima corresponding to the decisions to quote at the best bid or the best ask:

$$
\begin{aligned}
v(x,1) = \max \Bigg\{ & \tfrac{\mu/2}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,1)\right) + \tfrac{\mu/2}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,-1)\right) + \tfrac{p\lambda}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,1)\right) \\
& \qquad + \tfrac{(1-p)\lambda}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,1)\right), \\
& \tfrac{\mu/2}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,1)\right) + \tfrac{\mu/2}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,-1)\right) + \tfrac{p\lambda}{\lambda+\mu}\left(\tfrac{c}{2} - \gamma|x| + \delta v(x+1,1)\right) \\
& \qquad + \tfrac{(1-p)\lambda}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,1)\right), \\
& \tfrac{\mu/2}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,1)\right) + \tfrac{\mu/2}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,-1)\right) + \tfrac{p\lambda}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,1)\right) \\
& \qquad + \tfrac{(1-p)\lambda}{\lambda+\mu}\left(\tfrac{c}{2} - \gamma|x| + \delta v(x-1,1)\right), \\
& \tfrac{\mu/2}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,1)\right) + \tfrac{\mu/2}{\lambda+\mu}\left(-\gamma|x| + \delta v(x,-1)\right) + \tfrac{p\lambda}{\lambda+\mu}\left(\tfrac{c}{2} - \gamma|x| + \delta v(x+1,1)\right) \\
& \qquad + \tfrac{(1-p)\lambda}{\lambda+\mu}\left(\tfrac{c}{2} - \gamma|x| + \delta v(x-1,1)\right) \Bigg\} \\
= -\gamma|x| + \delta \Bigg(& \tfrac{\mu/2}{\lambda+\mu}\left(v(x,1) + v(x,-1)\right) + \tfrac{p\lambda}{\lambda+\mu} \max\left\{\tfrac{c}{2\delta} + v(x+1,1), v(x,1)\right\} \\
& \qquad + \tfrac{(1-p)\lambda}{\lambda+\mu} \max\left\{\tfrac{c}{2\delta} + v(x-1,1), v(x,1)\right\} \Bigg)
\end{aligned}
$$

□

Lemma 1 shows that the maximum taken over all possible actions in (3.7) can actually be separated into two maxima in each of which the HFT decides to quote or not to quote at the best bid or at the best ask. He quotes at the best bid to buy another share if $c/(2\delta) \geq v(x,s) - v(x+1,s)$ and similarly he quotes at the best ask to sell another share if $c/(2\delta) \geq v(x,s) - v(x-1,s)$. Note that when $s = 1$ (resp. $s = -1$), the probability of getting crossed by a market-sell (resp. market-buy) order is higher than getting crossed by a market-buy (resp. market-sell) order.

3.2. Optimal Market Making Policy by the HFT

In this section, we describe the optimal quoting policy of the HFT and show that it is based on thresholds. In the following theorem, we characterize the explicit structure of the quoting policy:

Theorem 1. *The optimal quoting policy π^* of the HFT consists in quoting at the best bid and the* *best ask according to a threshold policy, i.e., there exist L^* and U^* such that*

$$\ell^{b*}(x,1) = \begin{cases} 1 & \textit{when } x < U^* \\ 0 & \textit{when } x \geq U^* \end{cases} \qquad \ell^{a*}(x,1) = \begin{cases} 1 & \textit{when } x > L^* \\ 0 & \textit{when } x \leq L^* \end{cases}$$

$$\ell^{b*}(x,-1) = \begin{cases} 1 & \textit{when } x < -L^* \\ 0 & \textit{when } x \geq -L^* \end{cases} \qquad \ell^{a*}(x,-1) = \begin{cases} 1 & \textit{when } x > -U^* \\ 0 & \textit{when } x \leq -U^* \end{cases}$$

The limits L^* and U^* are functions of the model parameters, but not of the state.

We can interpret the result of Theorem 1 as follows. If the HFT receives a "buy" signal $(s = 1)$, he is going to act upon it by placing or keeping a limit order on the bid side of the book ($\ell^b = 1$) if his current long inventory is not already too high ($x < U^*$). But even if he receives a buy signal, he may place a limit order of the ask side of the book instead ($\ell^a = 1$), that is, offer to sell from his inventory, if his current long inventory is already too high ($x > L^*$).

In other words, as long as his current inventory is between the action thresholds L^* and U^*, the HFT quotes in order to capture the spread in the direction suggested by his signal: he quotes on the ask side if his signal says that the next market order will be a buy order, and on the bid side if his signal says that the next market order will be a sell order. But outside these thresholds, his inventory concerns take precedence and lead the HFT to potentially ignore his signal by systematically quoting on the ask side when his positive inventory is too high and on the bid side when his inventory is too negative, whatever the signal says.

Proof. Using the value iteration algorithm, we first establish by induction that $v(x, 1)$ and $v(x, -1)$ are concave functions. Let $v^{(0)}(x, 1) = 0$ and $v^{(0)}(x, -1) = 0$ for all x. Then, the base case states that

$$v^{(1)}(x,1) = -\gamma|x| + \frac{\lambda c}{2(\lambda+\mu)}, \qquad v^{(1)}(x,-1) = -\gamma|x| + \frac{\lambda c}{2(\lambda+\mu)}$$

which are both concave functions of x.

Assume that $v^{(n)}(x, 1)$ and $v^{(n)}(x, -1)$ are concave. Then,

$$v^{(n+1)}(x,1) = -\gamma|x| + \delta\Bigg(\tfrac{\mu/2}{\lambda+\mu}\left(v^{(n)}(x,1) + v^{(n)}(x,-1)\right) + \tfrac{p\lambda}{\lambda+\mu}\max\left\{\tfrac{c}{2\delta} + v^{(n)}(x+1,1), v^{(n)}(x,1)\right\} \\ + \tfrac{(1-p)\lambda}{\lambda+\mu}\max\left\{\tfrac{c}{2\delta} + v^{(n)}(x-1,1), v^{(n)}(x,1)\right\}\Bigg)$$

In order to establish the concavity of $v^{(n+1)}(x,1)$, we need the following lemma.

Lemma 2. *Fix $z \in \mathbb{R}$ and define $g(i) \triangleq \max\{z + f(i+1), f(i)\}$. $g(i)$ is concave if $f(i)$ is concave.*

Proof. Define $i^* \triangleq \min\{i : f(i) > z + f(i+1)\}$. For $i^* \leq i-2$, $g(i)-g(i+1) = f(i+1)-f(i+2) \leq z$ and is nondecreasing in i as $f(i)$ is assumed to be concave. If $i \geq i^*$, $g(i)-g(i+1) = f(i)-f(i+1) > z$ and is also nondecreasing in i due to concavity of $f(i)$. If $i = i^*-1$, $g(i)-g(i+1) = r+f(i+1)-f(i+1) = z$. Thus, for all i, $g(i) - g(i+1)$ is nondecreasing which makes $g(i)$ concave. □

A direct application of Lemma 2 also asserts that $g(i) \triangleq \max\{z + f(i-1), f(i)\}$ is also concave as it can be rewritten as $g(i) = z + \max\{f(i-1), f(i) - z\}$.

Using Lemma 2, $v^{(n+1)}(x,1)$ becomes concave as both $\max\left\{\frac{c}{2\delta} + v^{(n)}(x+1,1), v^{(n)}(x,1)\right\}$ and $\max\left\{\frac{c}{2\delta} + v^{(n)}(x-1,1), v^{(n)}(x,1)\right\}$ are concave. Since $v^{(n)}(x,1)$ converges to $v(x,1)$, $v(x,1)$ is concave in x. Due to the same structure in $v^{(n+1)}(x,-1)$, $v(x,-1)$ is also concave.

Using Lemma 1, $\ell_b^*(x,1) = 1$ if and only if $v(x,1) - v(x+1,1) \leq \frac{c}{2\delta}$. Since $v(x,1) - v(x+1,1)$ is nondecreasing in x, there exists U^* such that

$$\ell_b^*(x,1) = \begin{cases} 1 & x < U^*, \\ 0 & x \geq U^*. \end{cases}$$

Similarly, $\ell_a^*(x,1) = 1$ if and only if $v(x-1,1) - v(x,1) \geq -\frac{c}{2\delta}$. Since $v(x-1,1) - v(x,1)$ is nondecreasing in x, there exists L^* such that

$$\ell_s^*(x,1) = \begin{cases} 1 & x > L^*, \\ 0 & x \leq L^*. \end{cases}$$

In order to prove the last part of the theorem, we will use the following intuitive lemma which relates the value functions of different signals.

Lemma 3. *$v(x,1) = v(-x,-1)$.*

Proof. We use induction on the value iteration algorithm. Let $v^{(0)}(x,1) = 0$ and $v^{(0)}(x,-1) = 0$ for all x. Then, the base case states that

$$v^{(1)}(x,1) = -\gamma|x| + \frac{\lambda c}{2(\lambda+\mu)} \qquad v^{(1)}(x,-1) = -\gamma|x| + \frac{\lambda c}{2(\lambda+\mu)}$$

Thus, $v^{(1)}(x,1) = v^{(1)}(-x,1)$. Assume that $v^{(n)}(x,1)$ and $v^{(n)}(-x,-1)$ are equal. Then,

$$\begin{aligned}
v^{(n+1)}(x,1) &= -\gamma|x| + \delta\Bigg(\tfrac{\mu/2}{\lambda+\mu}\left(v^{(n)}(x,1)+v^{(n)}(x,-1)\right) + \tfrac{p\lambda}{\lambda+\mu}\max\left\{\tfrac{c}{2\delta}+v^{(n)}(x+1,1), v^{(n)}(x,1)\right\} \\
&\qquad\qquad + \tfrac{(1-p)\lambda}{\lambda+\mu}\max\left\{\tfrac{c}{2\delta}+v^{(n)}(x-1,1), v^{(n)}(x,1)\right\}\Bigg) \\
&= -\gamma|-x| + \delta\Bigg(\tfrac{\mu/2}{\lambda+\mu}\left(v^{(n)}(-x,-1)+v^{(n)}(-x,1)\right) + \tfrac{p\lambda}{\lambda+\mu}\max\Big\{\tfrac{c}{2\delta}+v^{(n)}(-x-1,-1), \\
&\qquad\qquad v^{(n)}(-x,-1)\Big\} + \tfrac{(1-p)\lambda}{\lambda+\mu}\max\left\{\tfrac{c}{2\delta}+v^{(n)}(-x+1,-1), v^{(n)}(-x,-1)\right\}\Bigg) \\
&= v^{(n+1)}(-x,-1)
\end{aligned}$$

where we use the induction hypothesis in the second equality. □

First, observe that L^* satisfies $v(L^*-1,1)-v(L^*,1) < -\frac{c}{2\delta}$ and $v(L^*,1)-v(L^*+1,1) \geq -\frac{c}{2\delta}$. Thus, using Lemma 3, we obtain $v(-L^*,-1) - v(-L^*+1,-1) > \frac{c}{2\delta}$ and $v(-L^*-1,1) - v(-L^*,-1) \leq \frac{c}{2\delta}$. Therefore, $-L^*$ is the threshold limit for our optimal bid quote, i.e.,

$$\ell_b^*(x,-1) = \begin{cases} 1 & x < -L^* \\ 0 & x \geq -L^* \end{cases}$$

Similarly, U^* satisfies $v(U^*-1,1) - v(U^*,1) \leq \frac{c}{2\delta}$ and $v(U^*,1) - v(U^*+1,1) > \frac{c}{2\delta}$. Thus, using Lemma 3, we obtain $v(-U^*,-1) - v(-U^*+1,-1) \geq -\frac{c}{2\delta}$ and $v(-U^*-1,1) - v(-U^*,-1) < \frac{c}{2\delta}$. Therefore, $-U^*$ is the threshold limit for our optimal ask quote, i.e.,

$$\ell_s^*(x,1) = \begin{cases} 1 & x > -U^* \\ 0 & x \leq -U^* \end{cases}$$

□

3.3. Computation of the HFT's Threshold Quoting Policy

In the previous section, we proved that the optimal market making policy involves thresholds. In this section, we exploit this solution structure and provide an efficient algorithm to solve for the threshold limits L^* and U^*, and the value functions v. We first prove that L^* and U^* are finite:

Lemma 4. *$L^* \in [-\frac{c/2}{\gamma(1-\delta)}, 0]$ and $U^* \in [0, \frac{c/2}{\gamma(1-\delta)}]$.*

Proof. L^* cannot be positive as it is strictly better for the trader to sell the unit and both earn $c/2$ and decrease the penalty cost. Similarly, U^* cannot be less than 0. We can obtain a lower bound for L^* quite easily. We know that the discounted expected cost between decision epochs is $\gamma|x|$. We know that the maximum discounted revenue from earning spreads is less than $\frac{c/2}{1-\delta}$. Thus, $L^* \geq -\frac{c/2}{\gamma(1-\delta)}$. Using the same reasoning, $U^* \leq \frac{c/2}{\gamma(1-\delta)}$. □

The following proposition provides a sufficient condition for threshold limits L and U to be optimal:

Proposition 1. *Let $N = \max(|L-1|, |U+1|)$. The following $2N+1$ equations uniquely determine the values of $v(-N,1), v(-N+1,1), \ldots, v(N-1,1), v(N,1)$. The function $v(x,1)$ is given for each value of the HFT's inventory x by:*

$$\begin{cases} -\gamma|x| + \delta\left(\frac{\mu/2}{\lambda+\mu}\left(v(x,1)+v(-x,1)\right) + \frac{p\lambda}{\lambda+\mu}\left(\frac{c}{2\delta} + v(x+1,1)\right) + \frac{(1-p)\lambda}{\lambda+\mu}v(x,1)\right), & x \in [-N, L], \\ -\gamma|x| + \frac{c\lambda/2}{\lambda+\mu} + \delta\left(\frac{\mu/2}{\lambda+\mu}\left(v(x,1)+v(-x,1)\right) + \frac{p\lambda}{\lambda+\mu}v(x+1,1) + \frac{(1-p)\lambda}{\lambda+\mu}v(x-1,1)\right), & x \in (L, U), \\ -\gamma|x| + \delta\left(\frac{\mu/2}{\lambda+\mu}\left(v(x,1)+v(-x,1)\right) + \frac{(1-p)\lambda}{\lambda+\mu}\left(\frac{c}{2\delta} + v(x-1,1)\right) + \frac{p\lambda}{\lambda+\mu}v(x,1)\right), & x \in [U, N], \end{cases}$$

If $v(L,1) - v(L-1,1) > \frac{c}{2\delta}$ and $v(U,1) - v(U+1,1) > \frac{c}{2\delta}$, then the threshold limits (L,U) are optimal.

Proof. We use the fact that under any threshold policy, (L,U), the inventory of the trader cannot go beyond the band of $[-N,N]$. When the inventory leaves the range of (L,U), the trader will only quote in one side of the market in order to pull the inventory back to this range. Since we know that threshold policy is optimal, we compute $v(x,1)$ for the specified regions and replace $v(x,-1)$ with $v(-x,1)$. At the end, we obtain the corresponding set of equations. Since the number of equations is equal to the number of unknowns, we can solve for all value functions. If the value functions satisfy the optimality equations of the quoting policy, (L,U) must be optimal. □

Using this sufficiency result, we can easily solve for the optimal threshold policies. We start by initializing both L and U to 0. We then decrement L and increment U until they satisfy the optimality equations given above.

We now have all the elements in place to compute the HFT's optimal strategy, and start by providing a simple illustration.

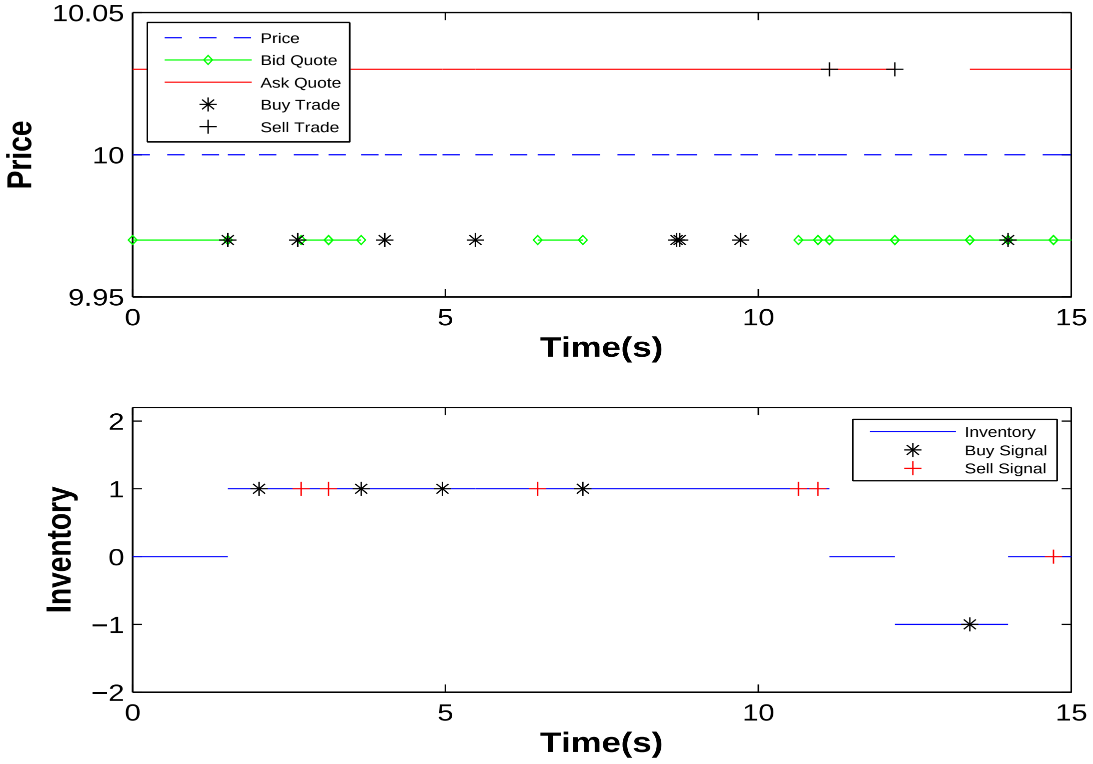

Fig. 1. Sample Price Path, HFT Optimal Quoting Policy and Transactions in the Base Model.

3.4. Illustration: A Simulated Path

Figure 1 shows a simulated set of arrival times of signals and market orders by LFTs, the HFT's optimal quoting strategy in response, and the resulting transactions and inventory of the HFT. The figure assumes the following parameter values: $C = \$0.06$, $\lambda = 24$ per minute, $p = 0.80$, $D = 0.05$, $\gamma = 0.20$, and $\mu = 50$ per minute. The optimal inventory limits that the HFT sets are $L^* = -2$ and $U^* = 1$.

We observe that when the HFT buys one share, buy signals force the HFT not to quote at the bid side. However, when the signal is sell, the HFT is in the quoting regime on both bid and ask sides. This pattern leads to a cancellation of the bid quotes whenever the signal changes from sell to buy. Similarly, when his inventory hits negative, the HFT starts not to quote at the ask side.

4. Optimal Policy Comparative Statics

In this section, we will provide a comparative statics analysis of the model. We are specifically interested in the dynamics of the HFT's objective value and his optimal trading/inventory critical limits, L^* ("Sell Limit") and U^* ("Buy Limit"), as a function of the model's parameters: the arrival rate of the LFTs, λ; the arrival rate of the HFT's signal, μ; the accuracy of the signal, p; the bid-ask spread, C; and the coefficient of HFT inventory aversion, Γ. These results serve primarily to explain intuitively how the model works, and the dependence of the HFT's optimal quoting policy on the different parameters.

In each of the following subsections, we vary one of the parameters of the model at a time, leaving the others fixed at the following values: the bid-offer spread, C, is \$0.10, the arrival rate of market orders, λ, is 100 (per minute), the arrival rate of signal, μ, is 100 (per minute), the signal accuracy, p is 0.80, the discount rate D is 0.5 and the coefficient of inventory aversion Γ is 0.2.

4.1. LFTs' Market Orders Arrival Rate

We start by reporting in Figure 2 the dependence of the HFT's optimal value and trading limits as a function of the arrival rate of the market orders submitted by the LFTs, λ. Having more LFTs in the market is unambiguously and monotonically good for the HFT, as their higher arrival rate increases his trading opportunities. On the right panel, recall that the optimal policy for the HFT consists in quoting according to the signal while inside the limits; quoting on the ask side of the book independently of the signal when his inventory is above the buy limit; and quoting on the bid side of the book independently of the signal when his inventory is below the sell limit. We see on the right panel of Figure 2 that the HFT is willing to widen the trading bands where he systematically follows his signal, since a higher value of λ implies that the bid-ask spread will get captured by the HFT at a higher rate, which in turn compensates the HFT for the additional potential inventory risk.

Note that the optimal policy limits shown on the right panel of Figure 2 change in discrete increments. This is a consequence of the fact that the HFT is holding and quoting an integer number of contracts, so the policy will only change if the underlying variable changes sufficiently to justify an integer change in the threshold limits.

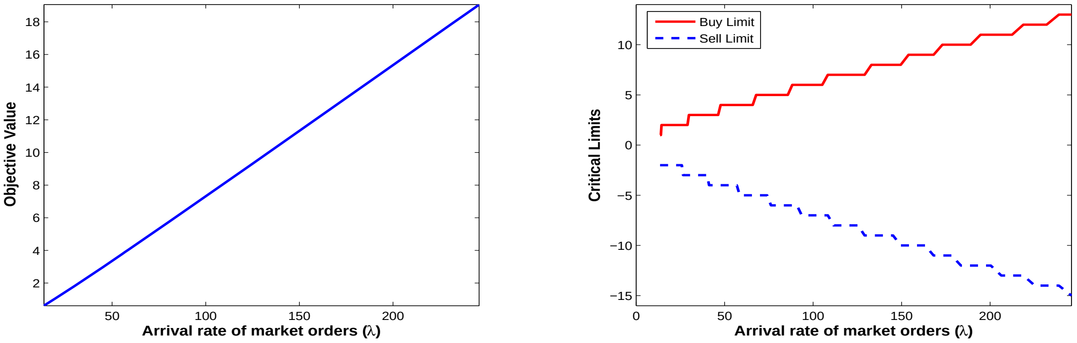

Fig. 2. Optimal Value and Trading Limits of the HFT as a Function of the Arrival Rate of LFTs.

4.2. HFT's Signal Arrival Rate

Figure 3 shows the dependence of the HFT's optimal value and his trading limits as a function of the arrival rate of the signal, μ. As the arrival rate of the signal increases, the value to the HFT increases, rapidly at first, and then reaches a plateau. This is because after some point the arrival of the signal is already much faster than the arrival rate of the market orders, λ, which is held fixed for this purpose. So the potential for achieving further profits is limited by the finite availability of LFTs: this is matching the empirically observed limits to HFT profitability due to the absence of sufficient trading opportunities, and is a limits-to-arbitrage type of situation. Beyond a certain point, better and better trading technology (higher μ) cannot deliver more trading opportunities when the pool of LFTs (λ) is bounded.

The HFT's optimal inventory limits gradually decrease in width with this increase in the signal arrival rate, as the higher rate of orders' updating align with the predictions for the next market order. The HFT adjusts his quotes according to his inventory at a higher frequency, which enables him to better manage his inventory risk.

4.3. Signal Accuracy

Next, we turn to the dependence of the HFT's optimal policy on the accuracy of the signal he receives, p. When $p = 1/2$, the signal is uninformative, as buy and sell orders are conditionally and unconditionally arriving at the same rate $\lambda/2$. When p increases towards 1, the signal becomes progressively more and more predictive of the side of the next incoming market order by a LFT. As the accuracy

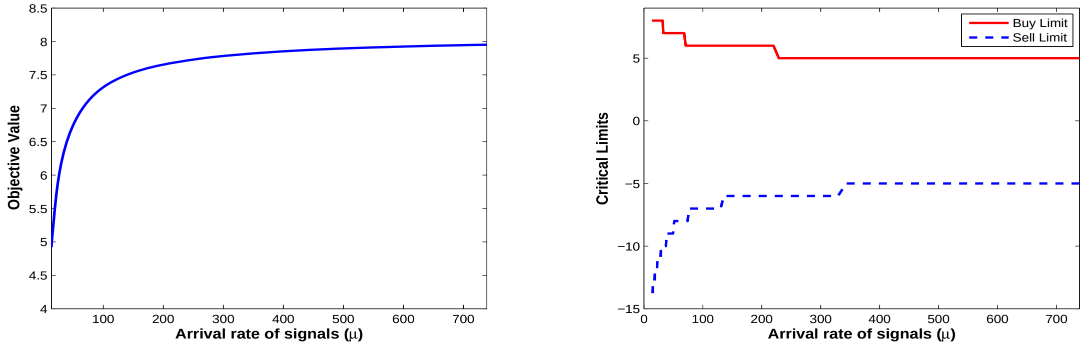

Fig. 3. Optimal Value and Trading Limits of the HFT as a Function of the Arrival Rate of the Signal.

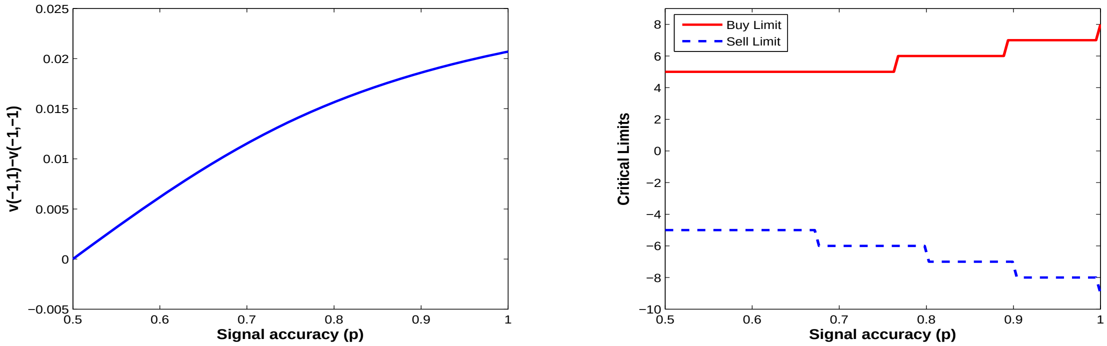

Fig. 4. Optimal Value and Trading Limits of the HFT as a Function of the Signal Accuracy.

of the signal increases, the HFT's value increases (left panel); the HFT is also more willing to trade in the direction of the signal and overrule his inventory concerns, resulting in a wider trading band (right panel).

4.4. Bid-Ask Spread

In Figure 5, we plot the optimal value and the inventory limits as a function of the bid-ask spread, C. When the bid-ask spread increases, it is intuitive that the HFT optimal value increases, as the positive part of his objective function in (2.1) is simply increased without any adverse consequences. We also observe that the HFT enlarges the width of his inventory limits: the increased reward in the form of a higher value of C makes the HFT more willing to tolerate additional inventory risk.

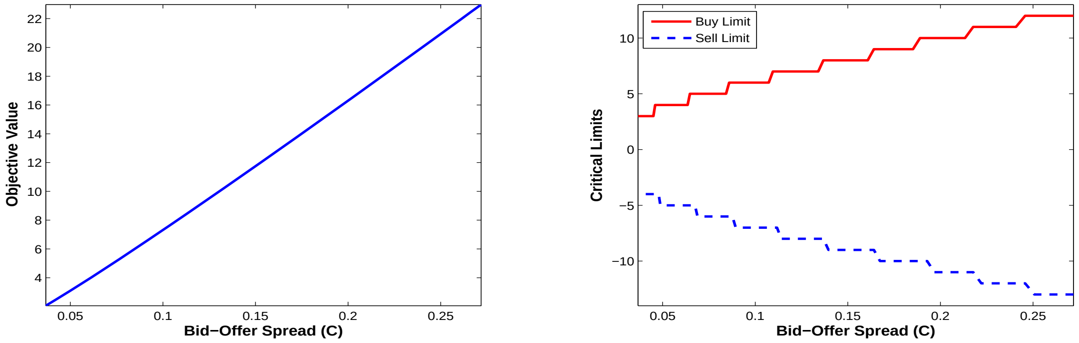

Fig. 5. Optimal Value and Trading Limits of the HFT as a Function of the Bid-Ask Spread.

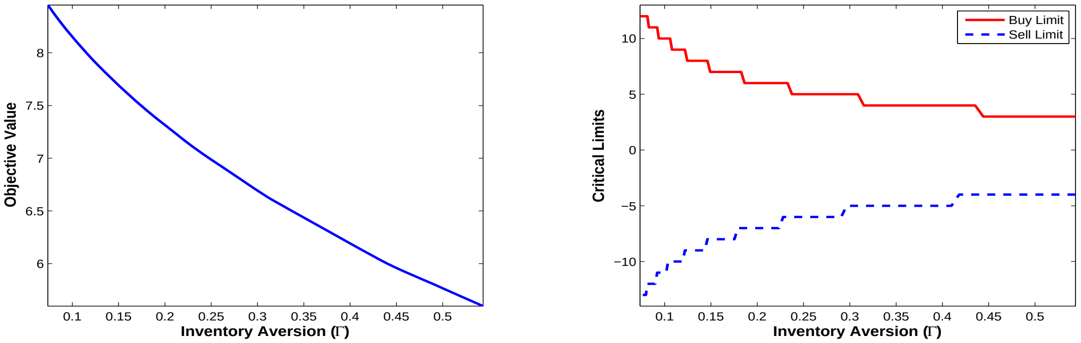

Fig. 6. Optimal Value and Trading Limits of the HFT as a Function of the Coefficient of Inventory Aversion.

4.5. HFT's Inventory Aversion

Figure 6 illustrates the HFT's optimal value and the inventory limits as a function of his coefficient of inventory aversion, Γ. Here, the comparative statics are the flip side of what we observed in Figure 5: as Γ increases, the HFT is penalized more heavily for holding inventory and the negative part of his objective function in (2.1) increases without any compensation. The HFT's optimal value decreases and the inventory limits become tighter.

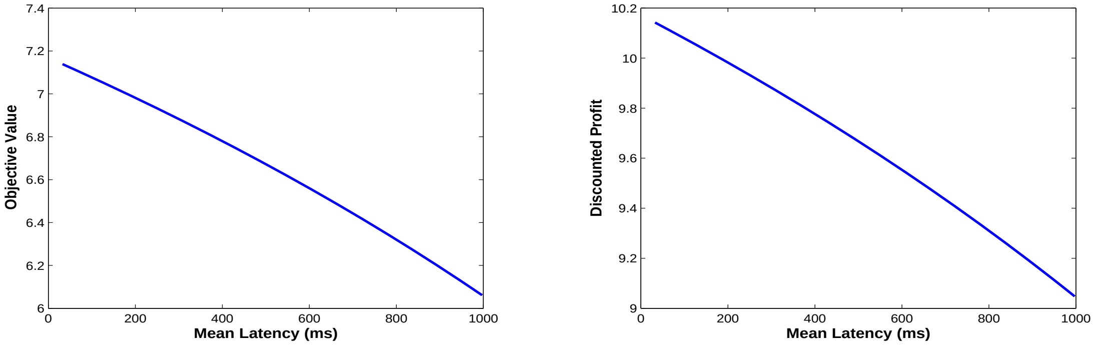

Fig. 7. Objective Value and Profit of the HFT as a Function of Mean Latency.

5. Implications of the Model for Market Structure

In this section, we consider empirical applications of our model. First, we will calibrate the model parameters using real data. Using this calibration, we will study the implications of the model for latency advantage, liquidity provision, welfare of LFTs and cancellation rates.

5.1. Main Calibration

In the upcoming empirical applications, we use the following values for our model parameters. We set the bid-offer spread, C, to be \$0.06, and arrival rate of market orders, λ to be 24 (per minute). We consider a range of μ values such that the mean latency of the HFT (i.e., $1/(\lambda + \mu)$) ranges from a few milliseconds to a second. Finally, we use $p = 0.80$, $D = 0.05$ and $\Gamma = 0.18$ so that the optimal inventory limits are empirically realistic. This calibration leads to $L^* = -2$ and $U^* = 1$, consistent with low HFT inventories.

5.2. HFT's Latency Advantage

Is having a more frequent signal coupled with better trading technology, i.e., lower latency, beneficial for the HFT? We answer this question by examining the impact of μ on the HFT's objective value and discounted profit.

Figure 7 illustrates the HFT's net discounted profit after inventory costs (objective value) and the corresponding discounted wealth generated by earning bid-offer spreads as a function of mean latency. We observe that decrease in mean latency provides higher trading profits to the HFT. Of course, these numbers are gross of the technological investment by the HFT that would most certainly be necessary to decrease his latency. The gross gain in HFT objective value when going from 1 second to 1 millisecond latency is roughly 18%, corresponding to a gain in net wealth of around 12%. This gain is due in the model to a wealth transfer from liquidity demanders, that is the LFTs, to the HFT.

5.3. Provision of Liquidity by the HFT

So, as he gets faster, the HFT gets richer. Yet, is he providing more liquidity to the LFTs? The embedded Markov Chain in the HFT's optimal market making strategy makes it possible to define and compute a long-run rate of quoting on either or both sides of the market by the HFT at the optimal policy, as follows. We first compute the optimal inventory limits, L^* and U^*, and characterize the optimal policy of the trader according to Theorem 1. Under this optimal policy, the inventory of the trader will be in the set, $[-N, N]$. At each inventory state, there are two possible signals, so under the optimal policy the process is governed by a finite-state Markov Chain with the following states. $(-N, 1), (-N, -1), \ldots, (0, 1), (0, -1), \ldots, (N, 1), (N, -1)$. Let P_{opt} be the probability transition matrix defined on this 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 $\pi P_{\mathsf{opt}} = \pi$. Let $\pi(x, s)$ be the stationary probability corresponding to the state (x, s).

According to the optimal policy, if $s = 1$, the HFT provides liquidity on both sides of the market when $x \in (L, U)$ and similarly if $s = -1$, the HFT provides liquidity on both sides of the market when $x \in (-U, -L)$. Let q_{quote} be the long-run probability of quoting to buy and sell at the same time. Then,

$$q_{\mathsf{quote}} = \sum_{x \in (L,U)} \pi(x, 1) + \sum_{x \in (-U,-L)} \pi(x, -1). \tag{5.1}$$

As the mean latency decreases, the HFT improves his ability to keep his inventory under control by predicting the order flow at a higher frequency. Consequently, the HFT has a higher chance of trading with a LFT, who is a liquidity taker, and thereby earning the spread. In order to do that, the HFT provides more liquidity to the market, on both sides, as his latency decreases, as shown in Figure 8. This result is consistent with the empirical literature on HFT, which tends to suggest that HFTs have a positive effect on the liquidity and depth in the market (see, e.g., Hendershott et al. (2011) and Menkveld (2013)).

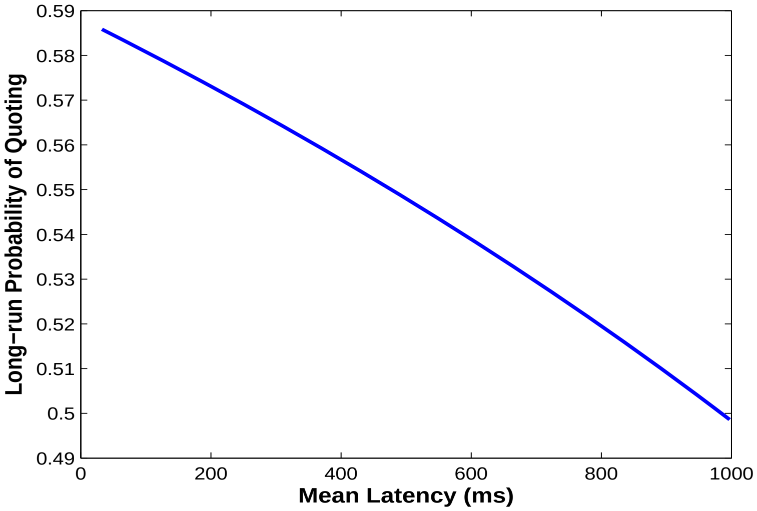

Fig. 8. Long-Run Probability of Quoting by the HFT as a Function of his Latency.

5.4. Are LFTs Hurt by the HFT's Speed Advantage?

So, more speed is good for the HFT's (gross) profits, and he provides more liquidity by increase the probability that he would quote on both sides of the market. How about the LFTs? In the model, LFTs employ market orders: they demand liquidity and immediacy. In light of this, our proxy for the welfare of LFTs is a function of the average fill rate that they are able to achieve for their market orders, thanks to the quotes provided by the HFT. We therefore now turn to the impact of the HFT's latency on the long-run market order fill rate.

Let q_{fill} be the unconditional long-run fill rate of a market order. q_{fill} is expected to be higher than q_{quote} as the market order may get filled when the HFT is quoting on a particular side of the market. Using the long-run stationary probabilities,

$$\begin{aligned} q_{\mathsf{fill}} = q_{\mathsf{quote}} &+ p\left(\sum_{x\in[-N,L]} \pi(x,1) + \sum_{x\in[U,N]} \pi(x,-1)\right) \\ &+ (1-p)\left(\sum_{x\in[-N,-U]} \pi(x,-1) + \sum_{x\in[-L,N]} \pi(x,1)\right). \end{aligned} \tag{5.2}$$

Figure 9 illustrates that as μ increases, more of the LFTs' orders get filled: the additional liquidity provided by a faster HFT is actually beneficial to the LFTs.

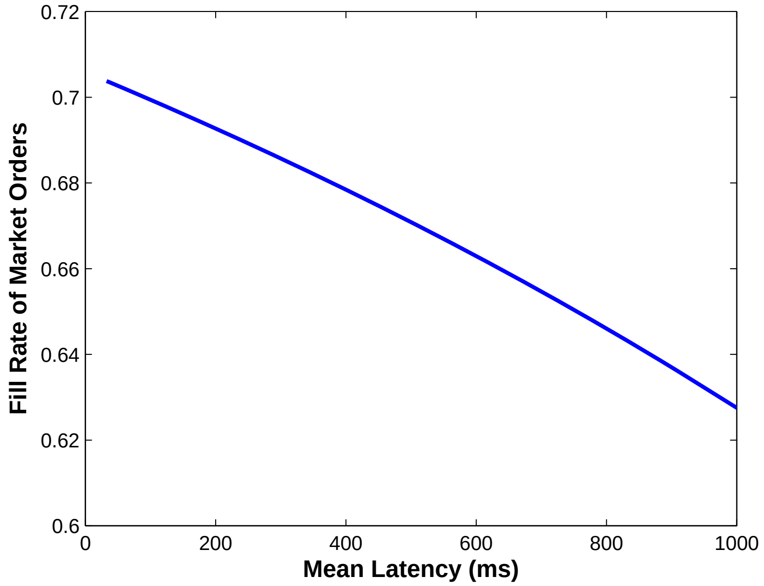

Fig. 9. Long-run Probability of LFTs' Orders Being Filled by the HFT.

5.5. *Order Cancellations by the HFT*

Order cancellations are widely observed in empirical 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". Our model is consistent with this empirical fact and can generate a similar behavior when optimal limits are not symmetric. In our model, an order cancellation consists in changing one of the elements of the pair (ℓ^b, ℓ^a), from 1 to 0 after the arrival of a signal event. Using the embedded Markov Chain under the optimal policy, we can find the long-run probability that an existing limit order will get cancelled by the HFT.

When $-L$ is not equal to U, an existing order may be cancelled when the HFT's signal fluctuates. Cancellations will occur on the specific states of the HFT. We define the cancellation region as a function of the signal as follows:

$$\mathcal{C}(s) \triangleq \begin{cases} (L,U)\backslash\left((L,U)\cap(-U,-L)\right) & \text{if } s=1, \\ (-U,-L)\backslash\left((L,U)\cap(-U,-L)\right) & \text{if } s=-1. \end{cases}$$

For example, when $L=-2$ and $U=1$, $\mathcal{C}(1)=\{1\}$ and $\mathcal{C}(-1)=\{-1\}$. Note that in these states, the HFT is quoting on both sides of the market but need to cancel one of the existing quotes when the signal changes as in the new state the HFT is no longer quoting at both sides. Consequently, a cancellation of the order will happen if the signal received by the HFT changes before a trade occurs. Let q_{cancel} be the unconditional long-run probability of an existing quote to be cancelled. We have:

$$q_{\mathsf{cancel}} = \frac{\mu/2}{\mu/2+\lambda}\left(\sum_{x\in\mathcal{C}(1)}\pi(x,1)+\sum_{x\in\mathcal{C}(-1)}\pi(x,-1)\right). \tag{5.3}$$

Figure 10 illustrates the increasing long-run cancellation rate as the mean latency of the trader decreases. Specifically, the long-run probability of order cancellations more than doubles to approximately 30% when going from 1 second to approximately 1 millisecond latency, as signal revisions become more frequent.

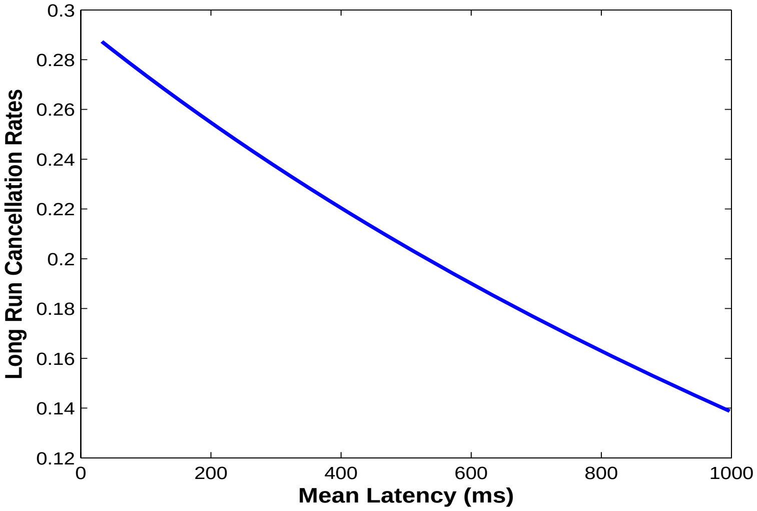

Fig. 10. Long-run Probability of Order Cancellation by the HFT as a Function of his Latency.

6. Extension 1: Price Volatility

So far, the HFT in the model provides liquidity to the market, and the faster he is, the better. However, one fundamental issue concerning the provision of liquidity by the HFT concerns the “quality” of that liquidity that is being provided. Possible definitions of that quality vary, but most include the notion that this liquidity is provided in a stable manner over time and over different market environments. Are HFTs fair weather liquidity providers ready to provide additional liquidity when the market 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, resulting in an amplification of that shock.

In our base model, the asset price was constant, which facilitated the analysis by reducing the number of possible state transitions. It is possible nevertheless to extend the model by introducing price volatility. Not surprisingly, assuming that price changes take the form of pure jumps, and that these jumps are driven by a Poisson process is the most convenient way to proceed, as it adds one more Poisson time clock to the two already considered (λ and μ). That is, let the fundamental price of the security, S_t, be given by a compound Poisson process

$$S_t = S_0 + \sum_{i=1}^{N_t} Y_i, \tag{6.1}$$

where N_t is a Poisson process with rate ζ counting the number of price jumps up to time t and Y_i is the stochastic magnitude of the ith jump with $\mathbb{E}[Y_i] = 0$, and a distribution subject to the requirement that S remains positive. The variance of the price process is proportional to ζ, and so the jump rate represents the level of price volatility.

Consistent with the notion that the HFT trades on the basis of short-lived market microstructure information, we assume that the HFT receives no information regarding these fundamental price movements; as before, his only signal is about the likely direction of the order flow. The optimal quoting behavior of the HFT retains the same structure with a modified equation for the value function, with revised transition probabilities due to price jumps. We can therefore derive the recursion for the value functions using a similar approach to that of Section 3. The following lemma provides the resulting dynamic programming equation:

Lemma 5.

$$\begin{aligned}
v(x,1) = -\gamma|x| + \delta\Bigg(& \frac{\mu/2}{r} v(x,-1) + \left(1 - \frac{\mu/2}{r} - \frac{\lambda}{r+\zeta}\right) v(x,1) \\
& + \frac{p\lambda}{r+\zeta} \max\left\{\frac{c}{2\delta} + v(x+1,1), v(x,1)\right\} \\
& + \frac{(1-p)\lambda}{r+\zeta} \max\left\{\frac{c}{2\delta} + v(x-1,1), v(x,1)\right\} \Bigg) \\
v(x,-1) = -\gamma|x| + \delta\Bigg(& \frac{\mu/2}{r} v(x,-1) + \left(1 - \frac{\mu/2}{r} - \frac{\lambda}{r+\zeta}\right) v(x,1) \\
& + \frac{p\lambda}{r+\zeta} \max\left\{\frac{c}{2\delta} + v(x-1,-1), v(x,-1)\right\} \\
& + \frac{(1-p)\lambda}{r+\zeta} \max\left\{\frac{c}{2\delta} + v(x+1,-1), v(x,-1)\right\} \Bigg)
\end{aligned}$$

where $r \equiv \lambda + \mu$, $\delta \equiv \frac{r}{r+D}$, $c \triangleq \delta C$ *and* $\gamma \equiv \frac{\Gamma}{r+D}$.

One advantage of our Poisson-based modeling approach is that the value function expressions in Lemma 1 and Lemma 5 retain the same structure. The main difference is that with the introduction of volatility, the probability of a trade in a single period, which is the only source of revenue for the HFT, decreases. When the price jump occurs, the HFT can no longer trade with the LFT (and earn the spread) with his existing limit orders.

The optimal quoting behavior of the HFT retains the same structure:

Theorem 2. *The optimal quoting policy* π^* *of the HFT consists in quoting at the best bid and the best ask according to a threshold policy, i.e., there exist* L^* *and* U^* *such that*

$$\ell^{b*}(x,1) = \begin{cases} 1 & \textit{when } x < U^* \\ 0 & \textit{when } x \geq U^* \end{cases} \qquad \ell^{a*}(x,1) = \begin{cases} 1 & \textit{when } x > L^* \\ 0 & \textit{when } x \leq L^* \end{cases}$$

$$\ell^{b*}(x,-1) = \begin{cases} 1 & \textit{when } x < -L^* \\ 0 & \textit{when } x \geq -L^* \end{cases} \qquad \ell^{a*}(x,-1) = \begin{cases} 1 & \textit{when } x > -U^* \\ 0 & \textit{when } x \leq -U^* \end{cases}$$

The limits L^* *and* U^* *are functions of the model parameters, but not of the state.*

Note that the limits L^* and U^*, now depend on the variability of the price, i.e., on the parameter ζ: for simplicity we use the same notation as before, but the limits are not the same as those of Theorem

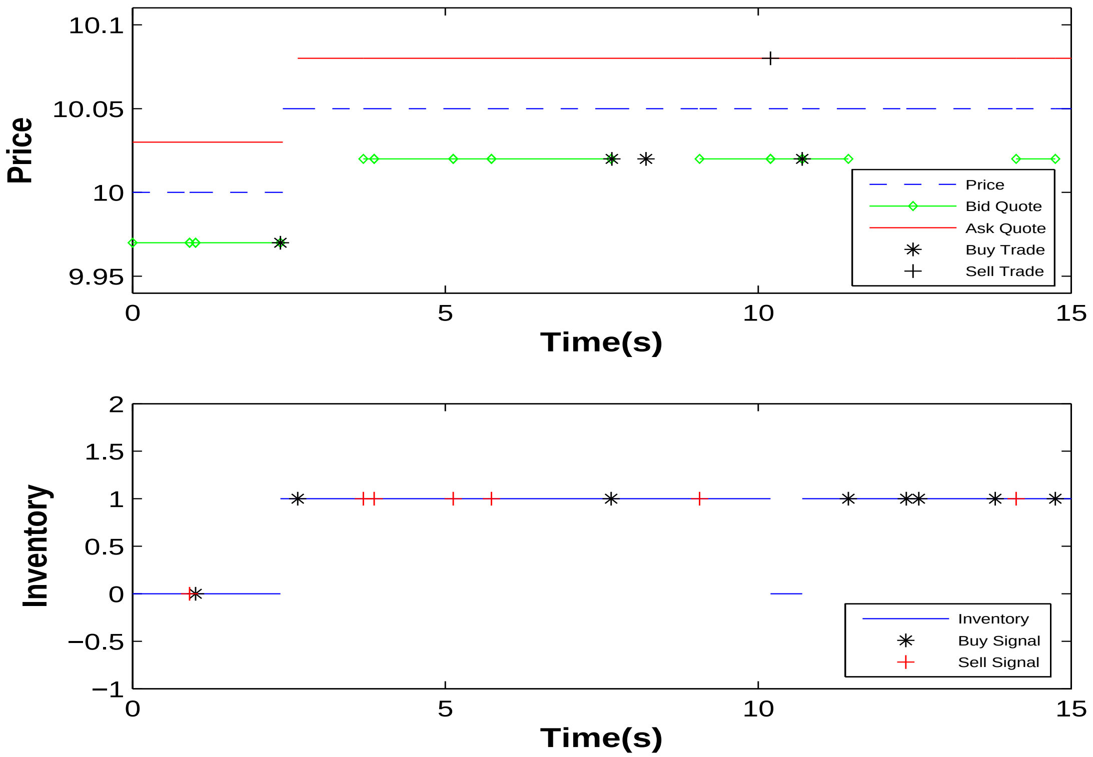

Fig. 11. Sample Price Path with Volsatility, HFT Optimal Quoting Policy and Transactions.

1 unless $\zeta = 0$. The proof of this theorem follows the same steps as the proof of Theorem 1, and is omitted to save space.

Figure 11 shows a simulated set of arrival times of signals and market orders by LFTs, the HFT's optimal quoting strategy and the resulting transactions and inventory of the HFT. The figure assumes the following parameter values: $C = \$0.06$, $\lambda = 24$ per minute, $p = 0.80$, $D = 0.05$, $\gamma = 0.35$, $\mu = 50$ per minute, and $\zeta = 5$ per minute. Optimal limits are $L^* = -2$ and $U^* = 1$ as in the base model. We observe that before the occurrence of the jump, HFT is quoting on the ask side as his inventory is at the the buy limit. When the price jump occurs, the HFT's existing ask quote becomes inactive and he waits till the next signal event to peg his limit order to the new mid-price.

Figure 12 shows how the HFT's optimal value and threshold limits change when there are jumps in prices, i.e., as a function of the arrival rate of price jumps, starting with 0 arrival, which matches the base model.

The impact of volatility on the quoting behavior of the HFT is another open question of interest.

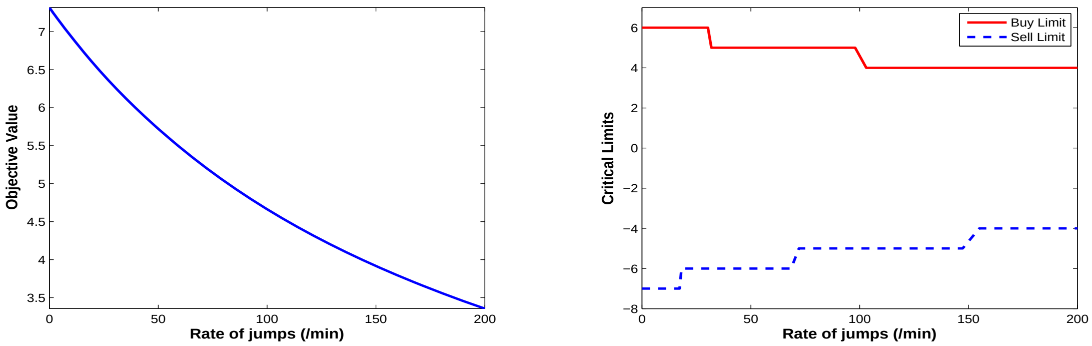

Fig. 12. Optimal Value and Trading Limits of the HFT as a Function of the Arrival Rate of Price Jumps.

This is an important concern for regulators. Specifically, during spikes in market volatility, we would like to examine whether HFTs continue providing liquidity at the same rate or whether they tend to withdraw their quotes from the marketplace. The model is capable of quantifying the change in overall quoting activity by the HFT as a function of volatility, as a comparative statics analysis with respect to the parameter ζ. A higher arrival rate ζ of price jumps is synonymous in the model with more volatility.

We set the mean latency of the trader to be 60 milliseconds. The rate of jumps ranges from 0 to 50 per minute where the latter limit corresponds to the high volatility regime in which the probability of a tick change is roughly two times the probability of a trade. Using this calibration, we can compute the long-run rate of quoting at both sides of the market as a function of jump rates.

Figure 13 illustrates the impact of volatility in the long-run probability of quoting. We observe that in the high volatility regime, the rate of liquidity provision by the HFT on both sides of the market decreases substantially. As the rate of jumps increases from 0 to 50, the long-run probability of quoting decreases from roughly 58% to 30%, which corresponds to a drop of nearly 50% with respect to the initial rate. The quoting rate by the HFT drops in stepwise increments as the rate of jumps crosses certain thresholds. This is due to the integer inventory limits. For instance, in Figure 13, when the rate of jumps is below 10 per minute approximately, the HFT sets the optimal limits as $L = -2$ and $U = 1$. When the rate crosses the 10 per minute level, the optimal policy requires quoting with inventory limits $L = -1$ and $U = 1$.

In other words, the model does predict that the HFT protects himself against unanticipated price

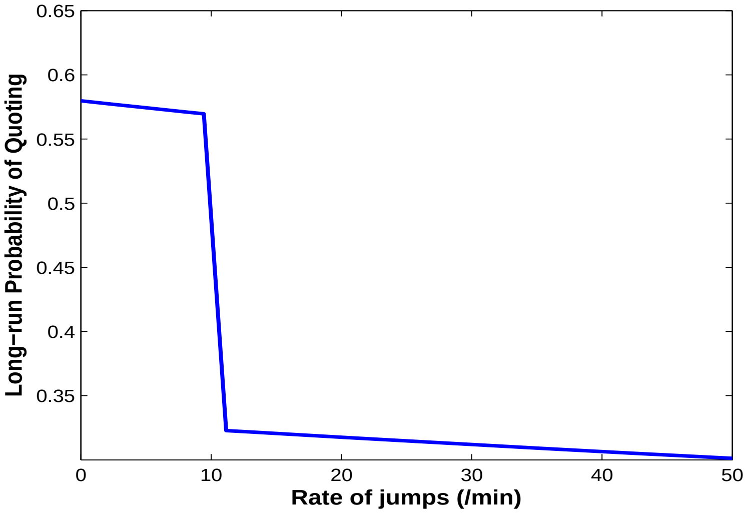

Fig. 13. Long-Run Probability of Quoting by the HFT as a Function of the Arrival Rate of Price Jumps.

jumps (over which he has no informational advantage) by quoting less frequently. We saw that in quiet markets, HFTs do contribute liquidity resulting in a gradual improvement in standard market quality metrics over time, which coincides with the growing prevalence of HFTs as market makers. But these statistics are measured over long horizons, and may fail to account for the temporary dislocations that appear in markets in the form of transitory minicrashes, since these are smoothed out over time.

7. Extension 2: Duopolistic Market Making with Queuing Priority

The baseline model includes only a single HFT, who is enjoying monopolistic rents as the sole provider of liquidity in the market. We now study a more realistic situation where the HFT is competing with a second fast liquidity provider in order to capture the bid-ask spread earned by executing LFTs' market orders. For simplicity, this second liquidity provider is a nonstrategic trader. This means that we are not solving a game between two strategic liquidity providers. Instead, we model the limit orders by the second trader as arriving on (perhaps not too surprisingly by this point in the paper) yet another Poisson clock, with rate β. The second trader's quoting actions are faster than the arrival rate of market orders by LFTs (λ), but for the sake of concreteness slower than the HFT's (μ), so we call this second trader a medium-frequency trader (MFT). The HFT remains fully strategic, taking now into account not only the arrival of the LFTs but now also the presence of the MFT he is competing with.

Both the MFT and the HFT trade with the LFTs' incoming market orders.

Analyzing this situation, even with the MFT behaving non-strategically, necessitates that we model an active limit order book, with price and time priority. There is now competition for priority and the fully strategic HFT needs to take into account the state of the limit order book and the priority or lack thereof that he would achieve when placing his own quotes.

We assume that the MFT, like the HFT, has at most one share to buy or sell at the best bid and ask queues. Suppose that order book events attributed to the MFT, which occur with a rate β, have equal likelihood to affect either the best bid or ask queue. There are two possible event types that the MFT can initiate in each queue. Consider the best bid queue. The MFT sends a limit order to the best bid queue with rate $\beta/2$ if he does not have an active order in this queue. On the other hand, if he does have an active limit order, he will cancel this order with rate $\beta/2$. The assumption of equal submission and cancellation rates by the MFT is done for simplicity only and can easily be relaxed. Fundamentally, this MFT is non-strategic; but the HFT is fully strategic. The analysis of the HFT's optimal strategy in this context proceeds by recognizing the presence of another Poisson time clock $(\lambda + \mu + \beta)$ in the model, with the HFT taking decisions at the arrival times of each event.

Incorporating the MFT into the model, the state of the best bid or ask queue can be in one of five states that are relevant for the HFT's optimization. We will use the following notation to describe each state:

- "ee" denotes that the queue is empty, with neither HFT nor MFT currently quoting;
- "hm" denotes that both the HFT and the MFT have active orders in the limit order book but the HFT has priority;
- "he" denotes that only HFT has an active order, which then has priority;
- "mh" denotes that both the MFT and the HFT have orders in the limit order book, but the MFT's order has higher priority;
- "me" denotes that there is only an order by the MFT in the book, which has priority.

Consequently, in this framework the best bid and ask queue can be in one of 25 states by interacting the 5 states for the bid queue with the 5 states for the ask queue. We can write the value function corresponding to each state. For example, for the case of empty bid and ask queues, the corresponding recursion for the value function is given by

$$
\begin{aligned}
v(x,1,\mathrm{ee},\mathrm{ee}) = -\gamma|x| + \delta \max \Bigg\{ & \tfrac{\mu/2}{r}\left(v(x,1,\mathrm{ee},\mathrm{ee}) + v(x,-1,\mathrm{ee},\mathrm{ee})\right) + \tfrac{\lambda}{r} v(x,1,\mathrm{ee},\mathrm{ee}) \\
& + \tfrac{\beta/2}{r}\left(v(x,1,\mathrm{me},\mathrm{ee}) + v(x,1,\mathrm{me},\mathrm{ee})\right), \\
& \tfrac{\mu/2}{r}\left(v(x,1,\mathrm{he},\mathrm{ee}) + v(x,-1,\mathrm{he},\mathrm{ee})\right) + \tfrac{p\lambda}{r}\left(\tfrac{c}{2\delta} + v(x+1,1,\mathrm{ee},\mathrm{ee})\right) \\
& + \tfrac{(1-p)\lambda}{r}\left(v(x,1,\mathrm{he},\mathrm{ee})\right) + \tfrac{\beta/2}{r}\left(v(x,1,\mathrm{hm},\mathrm{ee}) + v(x,1,\mathrm{he},\mathrm{me})\right), \\
& \tfrac{\mu/2}{r}\left(v(x,1,\mathrm{ee},\mathrm{he}) + v(x,-1,\mathrm{ee},\mathrm{he})\right) + \tfrac{p\lambda}{r}\left(v(x,1,\mathrm{ee},\mathrm{he})\right) \\
& + \tfrac{(1-p)\lambda}{r}\left(\tfrac{c}{2\delta} + v(x-1,1,\mathrm{ee},\mathrm{ee})\right) + \tfrac{\beta/2}{r}\left(v(x,1,\mathrm{ee},\mathrm{hm}) + v(x,1,\mathrm{me},\mathrm{hm})\right), \\
& \tfrac{\mu/2}{r}\left(v(x,1,\mathrm{he},\mathrm{he}) + v(x,-1,\mathrm{he},\mathrm{he})\right) + \tfrac{p\lambda}{r}\left(\tfrac{c}{2\delta} + v(x+1,1,\mathrm{ee},\mathrm{he})\right) \\
& + \tfrac{(1-p)\lambda}{r}\left(v(x-1,1,\mathrm{he},\mathrm{ee})\right) + \tfrac{\beta/2}{r}\left(v(x,1,\mathrm{he},\mathrm{hm}) + v(x,1,\mathrm{hm},\mathrm{he})\right) \Bigg\}
\end{aligned}
\tag{7.1}
$$

where $r \equiv \lambda + \mu + \beta$, $\delta \equiv \frac{r}{r+D}$, $c \triangleq \delta C$ and $\gamma \equiv \frac{\Gamma}{r+D}$.

The value function illustrates that the HFT has four actions to choose from. Each term in the maximum appearing in (7.1) is the HFT's payoff when he chooses (ℓ^b, ℓ^a) to be $(0,0)$, $(1,0)$, $(0,1)$ and $(1,1)$ respectively. In each case, we now consider the possibility that the MFT may submit a limit order at the best bid or best ask. If the HFT chooses to quote an order on either side of the market, he gains priority for execution as the book is empty. However, if he does not quote, then in the next period he may find the MFT's order in the book and his order will have lower time priority. We can solve for the optimal quoting policy of HFT in this framework. We employ the policy iteration algorithm (see, e.g., Puterman (1994)) by truncating the HFT's inventory after a certain limit. Since the interesting inventory regime of the HFT remains in a very narrow band around zero, the truncation has minimal impact.

7.1. *Duopoly: Splitting the Rent*

We first investigate the impact of priority due to queueing in the limit order book. We solve for the HFT's optimal policy and compute the optimal value he is able to achieve when competing for priority with the MFT, and compare to our baseline model where the HFT is a monopolistic provider of liquidity. We use the same calibration of the model parameters as in the illustrations contained in Section 2. In the absence of the second trader, $\beta = 0$. As for the MFT, we assume that $\beta = 50$. This calibration reflects the modeling choice that the MFT (β) is slower than the high frequency trader (μ), but the MFT's quoting actions are faster than the arrival rate of market orders by the LFTs (λ).

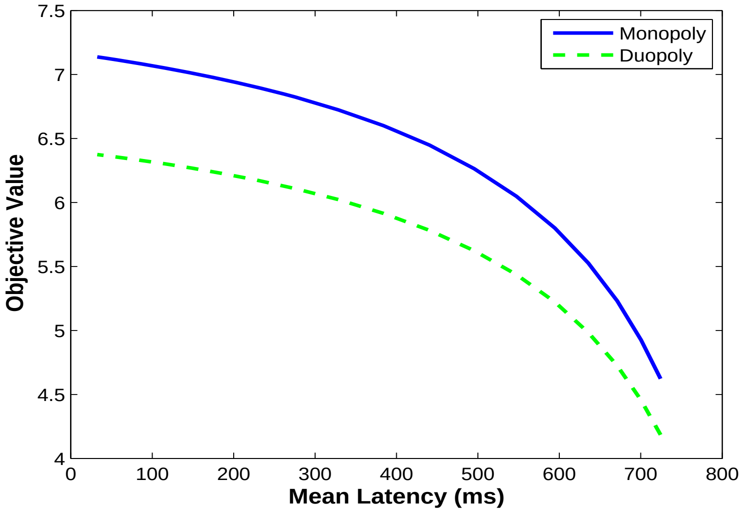

Fig. 14. HFT Optimal Values in the Monopoly (No Queuing) and Duopoly (Queuing) Situations.

Figure 14 illustrates the impact of queueing on the HFT's value due to the presence of the MFT, as a function of mean latency in the baseline (monopoly) and extended (duopoly) models. First, we observe that in the presence of the MFT, the HFT loses some of his profits to the MFT. Comparing the slopes in the value function in both regimes, we also observe that higher latency 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.

Looking now at the effect on the HFT and MFT, we see that they are effectively splitting the rent extracted from LFTs: as shown in Figure 15 the rent gets split between the two market makers. Not surprisingly, the faster the HFT (holding the speed of the MFT fixed), the more of the rent he is able to capture.

7.2. Duopoly vs. Monopoly: Implications for the LFTs

Figure 16 shows that LFTs are better off when market makers compete, compared to the monopolistic HFT situation: the fill rate for their orders increases, providing a key improvement to the liquidity of the market. The mean latency on the x-axis is the combined latency of the market makers, $1/(\mu+\lambda)$ in the monopoly situation and $1/\left(\mu+\lambda+\beta\right)$ in the duopoly situation.

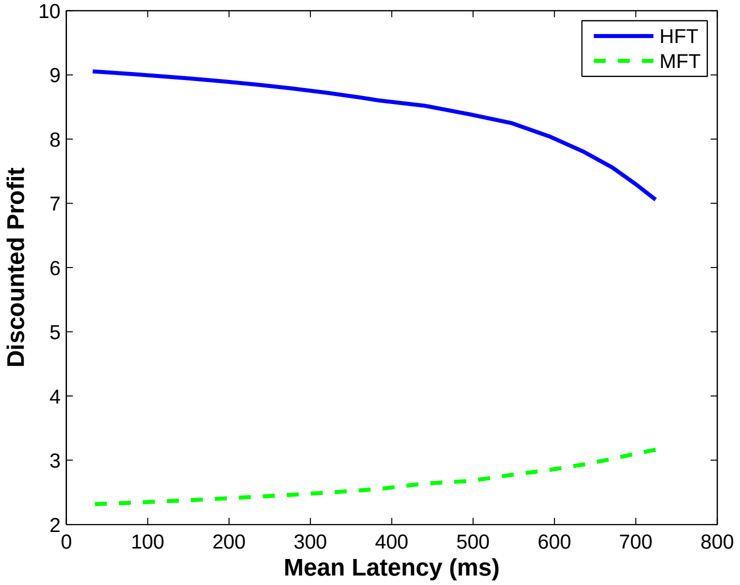

Fig. 15. HFT and MFT Duopolistically Splitting the Rent Extracted from LFTs

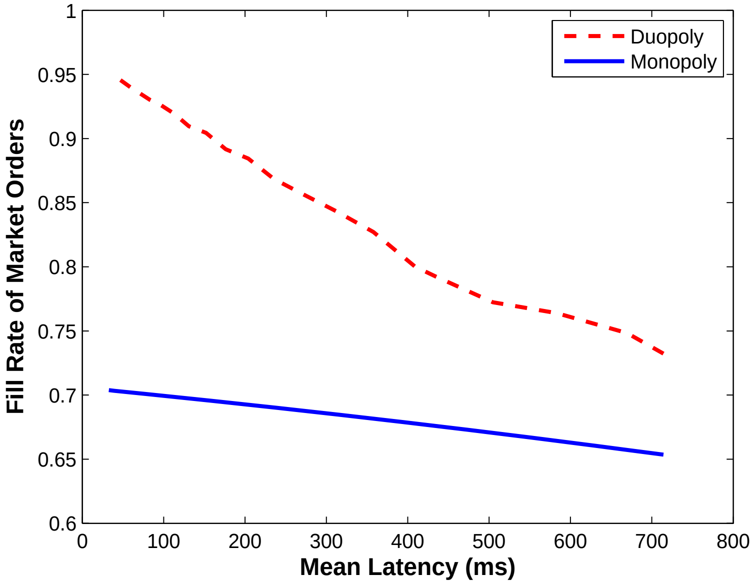

Fig. 16. Long-run Probability of LFTs' Orders Being Filled by Market Makers in the Monopoly (HFT) and Duopoly Cases (HFT and MFT).

8. Comparing Different HFT Regulations

Regulators around the globe are paying increasing attention to the impact of HFTs on the markets they supervise, and have started debating, developing and implementing policies designed to stem the rise of HFTs. The basic premise is that markets should remain primarily platforms to exchange risk, and not become race tracks with the potential to ultimately discourage slower users from participating.

In this section, we study the impact of some frequently proposed, and in some cases already enacted, regulatory policies on the HFT's objective value and provision of liquidity to the market. We view additional liquidity in good times as being largely useless, and even counterproductive especially if it dries 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 three 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, and taxing the cancellations of limit orders.

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[3],[4]. 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[5]. The Deutsche Börse introduced a tax in 2012 that charges HFTs for high "order-to-trade" ratios[6] as does the London Stock Exchange[7]. Norwegian regulators too consider taxing traders who submit a large number of orders relative to their actual executions[8]. The CME Group, the world's largest futures exchange, has had for a number of years message volume caps, designed to prevent excessive numbers of orders from being

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

[4] 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.

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

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

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

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

placed[9], while Nasdaq and DirectEdge, two of the largest US stock exchanges have introduced fines to discourage excessive order placement[10]. Canadian regulators too began increasing the fees charged to HFTs that flood the market with orders[11], while Indian regulators are studying ways to curb HFTs[12]. On the other hand, Brazil appears to welcome the influx of HFTs[13]. 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[14].

In January 2013, European Union finance ministers approved a transaction tax in Germany, France, Italy, Spain and seven other Eurozone countries[15]; the UK, concerned about the impact on the City, is opposed[16]. It seems unlikely at present that the initially far-reaching package will get implemented as proposed, if ever[17]. The German government has advanced legislation that would, among other things, force HFTs to register as such with the government[18] and limit their ability to rapidly place and cancel orders[19]. 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[20]. 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[21], 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[22].

We find that imposing minimum time-limits and cancellation taxes induce the HFT to quote more on both sides of the market, whereas transaction taxes do not improve this measure of liquidity. One important finding is that when minimum time-limits are in effect, the fill rate of market orders does not

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

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

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

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

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

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

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

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

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

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

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

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

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

[22] "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.

decrease substantially in the presence of higher volatility. Of course, this analysis remains dependent on the assumptions of the model, and excludes possible alternative responses by HFTs, such as trading on an alternative, non-regulated, venue.

8.1. 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, the law of unintended consequences may apply, as 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.[23] 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[24].

In the framework of our model, a transactions tax is straightforward to analyze. Suppose that the HFT pays $\kappa/2$ dollars each time a market order crosses one of his limit orders. From the perspective of the HFT, the transaction tax, κ, merely reduces the half bid-offer spread that the HFT earns from each trade. Defining the tax-adjusted bid-offer spread as $\tilde{C} \equiv C - \kappa$, we can analyze the impact of the tax policy on the objective value of the HFT and his long-run rate of quoting using the exact same equations developed so far. Figure 17 displays four graphs illustrating objective value, long-run fill rates, long-run quoting rates with respect to the Tobin tax rate and finally the long-run fill rates as a

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

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

function of volatility in the presence and absence of tax regulation. We use the same parameter values as in Section 4 to facilitate the comparison with earlier applications.

Figure 17 illustrates that the HFT's objective value is decreasing with higher taxes that are up to 5% of the bid-offer spread, C, i.e., with a maximum κ considered of 30 bps in line with the proposals being considered or already implemented in Europe. This is consistent with our earlier comparative statics for C, since this policy is equivalent from the point of view of the HFT to reducing the bid-offer spread earned by the HFT. 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 long-run rate of quoting actually decreases, as seen in Figure 17. 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.[25]. Lastly, we investigate how market liquidity, measured by long-run rate of fill-rates, changes with respect to volatility when the Tobin tax is implemented. Setting $\kappa = 15$ bps, we observe that the HFT's quoting has the same sensitivity to volatility when compared to the scenario in the absence of a Tobin tax.

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

Another possible policy consists in imposing a minimum time-limit on each limit order submitted by the HFT. This is a widely discussed policy among regulators and exchanges. European and Australian regulators have also recommended "resting-periods" for orders in order to slow-down HFTs' activity. One alternative policy might involve bunching together incoming orders every few milliseconds, so a HFT would face queuing risk, as well as a minimum waiting time before a cancellation. EBS, one of the major trading platforms in the foreign exchange market, has discussed such a proposal with its users.

We can incorporate random minimum time-limits in our model as follows. Although the policy proposals all suggest a fixed waiting time, typically of half a second, it will come as no surprise to the reader by this point of the paper that it is actually more convenient to analyze a version of the policy where the waiting time before a cancellation is allowed is random, and is itself derived from a Poisson process. The arrival rate of that Poisson process controls the expected amount of waiting time before a quote can be cancelled. That is, we suppose that each active limit order cannot be cancelled before

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

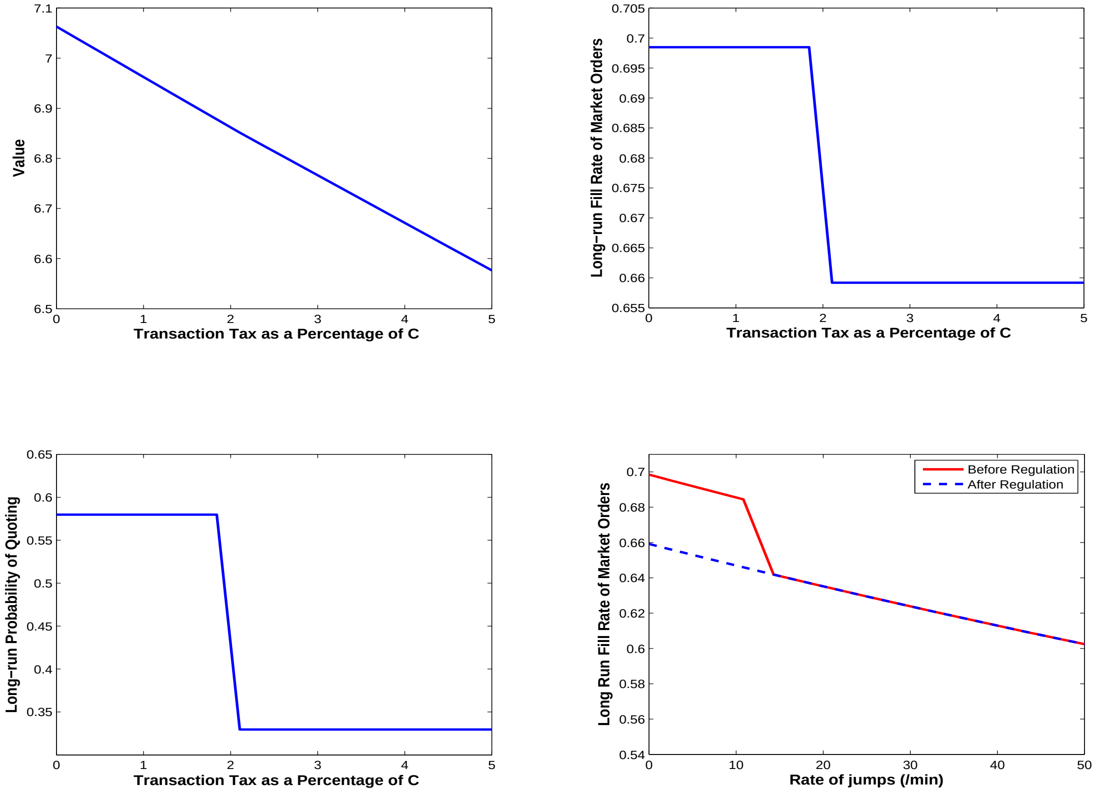

Fig. 17. Taxing High Frequency Trades.

Note: The left panels plot the effect of taxing transactions on the HFTs' value and his provision of liquidity. The right panels display the effect of a transaction tax on the fill-rate of LFTs' orders (upper panel) and the sensitivity of the fill rate to volatility (in the form of price jumps) before and after the minimum time regulation (lower panel).

a random time amount, τ^{cancel}, which is exponentially distributed with mean duration $2/\theta$. When the HFT is quoting on both sides of the market, the average time limit till a possible cancellation on either side of the market is $1/\theta$. In this case, state transitions will occur at rate $(\lambda + \mu + \theta)$ with respect to the base model.

The relevance of this cancellation constraint depends upon whether the HFT has in place a limit order or not. So we need to add one more state variable, $b_t \in \{00, 10, 01, 11\}$, which tracks the HFT's active limit orders that cannot yet be cancelled: $b = 00$ means that the HFT has no active limit orders that cannot be cancelled, 10 (resp. 01) means an active limit buy (resp. sell) order, and 11 active limit orders on both sides. When $b \neq 00$, this state variable will limit the available actions to the HFT. For example, when $b = 01$, the HFT must always choose $\ell^a = 1$.

Using the same discrete-time transformation principle for Poisson processes as in Section 3.1, we can obtain the optimality equations. The following lemma provides the optimality equations for $s = 1$:

Lemma 6.

$$
\begin{aligned}
v(x,1,00) = -\gamma|x| + \delta \max \Bigg\{ & \frac{\mu/2}{r} v(x,-1,00) + \left(1 - \frac{\mu/2}{r}\right) v(x,1,00), \\
& \frac{\mu/2}{r} v(x,-1,10) + \frac{p\lambda}{r}\left(\frac{c}{2\delta} + v(x+1,s,00)\right) \\
& + \frac{\mu/2 + (1-p)\lambda}{r} v(x,1,10) + \frac{\theta}{r} v(x,1,00), \\
& \frac{\mu/2}{r} v(x,-1,01) + \frac{(1-p)\lambda}{\lambda+\mu}\left(\frac{c}{2\delta} + v(x-1,1,00)\right) \\
& + \frac{\mu/2 + p\lambda}{r} v(x,1,01) + \frac{\theta}{r} v(x,1,00), \\
& \frac{\mu/2}{r} v(x,-1,11) + \frac{\mu/2}{r} v(x,1,11) \\
& + \frac{p\lambda}{r}\left(\frac{c}{2\delta} + v(x+1,1,01)\right) + \frac{\theta/2}{r} v(x,1,01) \\
& \quad + \frac{(1-p)\lambda}{r}\left(\frac{c}{2\delta} + v(x-1,1,10)\right) + \frac{\theta/2}{r} v(x,1,10) \Bigg\}.
\end{aligned}
$$

$$\begin{aligned}
v(x,1,10) = -\gamma|x| + \delta \max \Bigg\{ & \frac{\mu/2}{r} v(x,-1,10) + \frac{p\lambda}{r}\Big(\frac{c}{2\delta} + v(x+1,s,00)\Big) \\
& + \frac{\mu/2+(1-p)\lambda}{r} v(x,1,10) + \frac{\theta}{r} v(x,1,00), \\
& \frac{\mu/2}{r} v(x,-1,11) + \frac{\mu/2}{r} v(x,1,11) + \frac{p\lambda}{r}\Big(\frac{c}{2\delta} + v(x+1,1,01)\Big) + \frac{\theta/2}{r} v(x,1,01) \\
& + \frac{(1-p)\lambda}{r}\Big(\frac{c}{2\delta} + v(x-1,1,10)\Big) + \frac{\theta/2}{r} v(x,1,10) \Bigg\}.
\end{aligned}$$

$$\begin{aligned}
v(x,1,01) = -\gamma|x| + \delta \max \Bigg\{ & \frac{\mu/2}{r} v(x,-1,10) + \frac{p\lambda}{r}\Big(\frac{c}{2\delta} + v(x+1,s,00)\Big) \\
& + \frac{\mu/2+(1-p)\lambda}{r} v(x,1,10) + \frac{\theta}{r} v(x,1,00), \\
& \frac{\mu/2}{r} v(x,-1,11) + \frac{\mu/2}{r} v(x,1,11) + \frac{p\lambda}{r}\Big(\frac{c}{2\delta} + v(x+1,1,01)\Big) + \frac{\theta/2}{r} v(x,1,01) \\
& + \frac{(1-p)\lambda}{r}\Big(\frac{c}{2\delta} + v(x-1,1,10)\Big) + \frac{\theta/2}{r} v(x,1,10) \Bigg\}.
\end{aligned}$$

$$\begin{aligned}
v(x,1,11) = -\gamma|x| + \delta \Bigg\{ & \frac{\mu/2}{r} v(x,-1,11) + \frac{\mu/2}{r} v(x,1,11) + \frac{p\lambda}{r}\Big(\frac{c}{2\delta} + v(x+1,1,01)\Big) \\
& + \frac{\theta/2}{r} v(x,1,01) + \frac{(1-p)\lambda}{r}\Big(\frac{c}{2\delta} + v(x-1,1,10)\Big) + \frac{\theta/2}{r} v(x,1,10) \Bigg\}.
\end{aligned}$$

where $r \equiv \lambda + \mu + \theta$*,* $\delta \equiv \frac{r}{r+D}$*,* $c \triangleq \delta C$ *and* $\gamma \equiv \frac{\Gamma}{r+D}$*.*

Lemma 6 illustrates that the HFT has 4 actions to choose from only when $b = 00$. He only chooses the value of ℓ^a when $b = 10$ and only chooses the value of ℓ^b when $b = 01$. When $b = 11$, he has no choice other than to continue quoting at both sides of the market, as neither order can be cancelled yet. We can again solve for the optimal quoting policy of the HFT in this framework. We employ the policy iteration algorithm by truncating the HFT's minimum and maximum inventory.

Figure 18 illustrates the objective value of the HFT as a function of the average time-limit, $1/\theta$, expressed in milliseconds. Note that the limiting case $\theta \to \infty$, meaning that the minimum time-limit is zero for active quotes, reverts to our base model. As the average time-limit increases, we observe that the objective value of the HFT decreases. This tax policy is also good for overall liquidity as measured by the long-run fill rate of market orders and long-run probability of HFT quoting.

This tax policy has another desirable property: when the time limits are in effect, the HFT becomes less sensitive to the market volatility, i.e., his liquidity provision is more stable. The bottom-right graph in Figure 18 highlights this stability. Setting the average minimum time limit to 20 ms

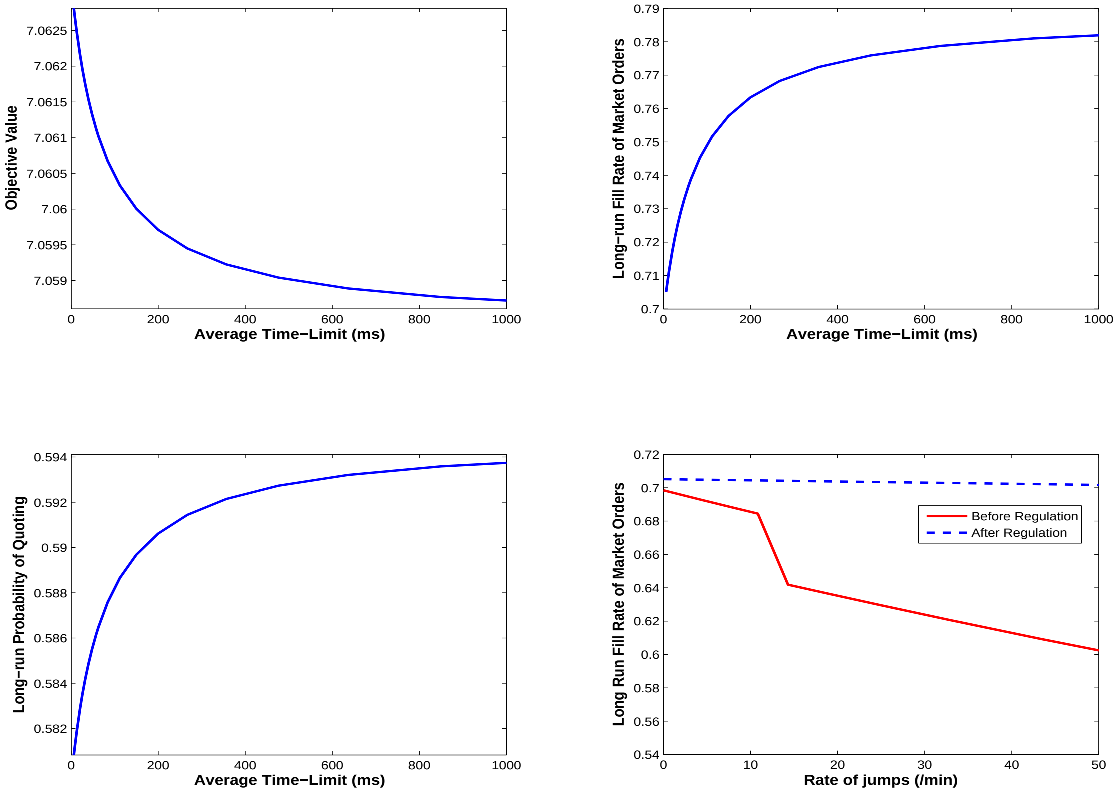

Fig. 18. Imposing a Minimum Time Before a Quote Can Be Cancelled.

Note: The left panels plot the effect of imposing minimum time limits before orders can be cancelled on the HFTs' value and his provision of liquidity. The right panels display the effect of the minimum time on the fill-rate of LFTs' orders (upper panel) and the sensitivity of the fill rate to volatility (in the form of price jumps) before and after the minimum time regulation (lower panel).

for each active quote, we observe that long-run fill rates decrease only slightly whereas in the absence of the regulation, the decrease is substantial.

8.3. Taxing Limit Order Cancellations

The last policy we consider consists in taxing the HFT whenever he cancels an existing quote. One concern about the reported higher liquidity due to HFT activity is that the provided liquidity is very short-lived, or "phantom", i.e., HFT cancels his quote before LFTs get a chance to trade with it. This tax proposal aims to make this type of HFT operation costlier.

Similar to our procedure in Section 8.2, we can incorporate this tax policy in our framework by including one more state variable, $b_t \in \{00, 10, 01, 11\}$, which tracks the HFT's most recent action.

We will assume that the HFT pays ε dollars as a penalty for each cancelled order. For example, if the HFT chooses action $(0,1)$ when $b = 10$, he will pay a cancellation tax of ε. Using the same discrete-time transformation as in Section 3.1, we can obtain the optimality equations. The following lemma provides the optimality equations for $s = 1$:

Lemma 7.

$$
\begin{aligned}
v(x,1,00) = -\gamma|x| + \delta \max \Bigg\{ & \frac{\mu/2}{r} v(x,-1,00) + \left(1 - \frac{\mu/2}{r}\right) v(x,1,00), \\
& \frac{\mu/2}{r} v(x,-1,10) + \frac{p\lambda}{r}\left(\frac{c}{2\delta} + v(x+1,s,00)\right) + \frac{\mu/2 + (1-p)\lambda}{r} v(x,1,10), \\
& \frac{\mu/2}{r} v(x,-1,01) + \frac{(1-p)\lambda}{\lambda+\mu}\left(\frac{c}{2\delta} + v(x-1,1,00)\right) + \frac{\mu/2 + p\lambda}{r} v(x,1,01), \\
& \frac{\mu/2}{r} v(x,-1,11) + \frac{\mu/2}{r} v(x,1,11) + \frac{p\lambda}{r}\left(\frac{c}{2\delta} + v(x+1,1,01)\right) \\
& + \frac{(1-p)\lambda}{r}\left(\frac{c}{2\delta} + v(x-1,1,10)\right) \Bigg\}.
\end{aligned}
$$

$$
\begin{aligned}
v(x,1,10) = -\gamma|x| + \delta \max \Bigg\{ & -\varepsilon + \frac{\mu/2}{r} v(x,-1,00) + \left(1 - \frac{\mu/2}{r}\right) v(x,1,00), \\
& \frac{\mu/2}{r} v(x,-1,10) + \frac{p\lambda}{r}\left(\frac{c}{2\delta} + v(x+1,s,00)\right) + \frac{\mu/2 + (1-p)\lambda}{r} v(x,1,10), \\
& -\varepsilon + \frac{\mu/2}{r} v(x,-1,01) + \frac{(1-p)\lambda}{\lambda+\mu}\left(\frac{c}{2\delta} + v(x-1,1,00)\right) + \frac{\mu/2 + p\lambda}{r} v(x,1,01), \\
& \frac{\mu/2}{r} v(x,-1,11) + \frac{\mu/2}{r} v(x,1,11) + \frac{p\lambda}{r}\left(\frac{c}{2\delta} + v(x+1,1,01)\right) \\
& + \frac{(1-p)\lambda}{r}\left(\frac{c}{2\delta} + v(x-1,1,10)\right) \Bigg\}.
\end{aligned}
$$

$$
\begin{aligned}
v(x,1,01) = -\gamma|x| + \delta \max \Bigg\{ & -\varepsilon + \frac{\mu/2}{r} v(x,-1,00) + \left(1 - \frac{\mu/2}{r}\right) v(x,1,00), \\
& -\varepsilon + \frac{\mu/2}{r} v(x,-1,10) + \frac{p\lambda}{r}\left(\frac{c}{2\delta} + v(x+1,s,00)\right) + \frac{\mu/2 + (1-p)\lambda}{r} v(x,1,10), \\
& \frac{\mu/2}{r} v(x,-1,01) + \frac{(1-p)\lambda}{\lambda+\mu}\left(\frac{c}{2\delta} + v(x-1,1,00)\right) + \frac{\mu/2 + p\lambda}{r} v(x,1,01), \\
& \frac{\mu/2}{r} v(x,-1,11) + \frac{\mu/2}{r} v(x,1,11) + \frac{p\lambda}{r}\left(\frac{c}{2\delta} + v(x+1,1,01)\right) \\
& + \frac{(1-p)\lambda}{r}\left(\frac{c}{2\delta} + v(x-1,1,10)\right) \Bigg\}.
\end{aligned}
$$

$$
\begin{aligned}
v(x,1,11) = -\gamma|x| + \delta \max\Bigg\{ & -2\varepsilon + \frac{\mu/2}{r} v(x,-1,00) + \Big(1 - \frac{\mu/2}{r}\Big) v(x,1,00), \\
& -\varepsilon + \frac{\mu/2}{r} v(x,-1,10) + \frac{p\lambda}{r}\Big(\frac{c}{2\delta} + v(x+1,s,00)\Big) + \frac{\mu/2 + (1-p)\lambda}{r} v(x,1,10), \\
& -\varepsilon + \frac{\mu/2}{r} v(x,-1,01) + \frac{(1-p)\lambda}{\lambda+\mu}\Big(\frac{c}{2\delta} + v(x-1,1,00)\Big) + \frac{\mu/2 + p\lambda}{r} v(x,1,01), \\
& \frac{\mu/2}{r} v(x,-1,11) + \frac{\mu/2}{r} v(x,1,11) + \frac{p\lambda}{r}\Big(\frac{c}{2\delta} + v(x+1,1,01)\Big) \\
& + \frac{(1-p)\lambda}{r}\Big(\frac{c}{2\delta} + v(x-1,1,10)\Big) \Bigg\}.
\end{aligned}
$$

where $r \equiv \lambda + \mu$, $\delta \equiv \frac{r}{r+D}$, $c \triangleq \delta C$ *and* $\gamma \equiv \frac{\Gamma}{r+D}$.

The optimality equations are similar to those of the base model except that the HFT is penalized when he cancels an active limit order. We numerically solve for the optimal quoting policy of the HFT in the presence of cancellation taxes. We again employ the policy iteration algorithm by truncating the HFT's minimum and maximum inventory.

Figure 19 considers the impact of such cancellation taxes. We consider cases where ε is as high as 5% of the bid-offer spread, C, i.e., 30 bps, as in Section 8.1. We observe that the objective value of the HFT decreases in the presence of cancellation taxes. When there are cancellation taxes, the HFT tries to set symmetric optimal inventory limits, if possible, to avoid any cancellation. That is, as soon as the tax is introduced the HFT changes the optimal limits from $L = -2$, $U = 1$ to $L = -2$, $U = 2$. With higher inventory limits, the HFT begins to quote more on both sides of the market, which increases the long run fill rates and probability of quoting.

One drawback of this tax policy is that the HFTs' liquidity provision is still sensitive to market volatility. Setting $\varepsilon = 15$ bps, we observe that the HFT's quoting has the same sensitivity to volatility when compared to the base scenario without any regulation. When there are price jumps, the HFT instantly lowers the inventory limits which adversely affect the long-run fill rate of market orders and hurts the LFTs.

9. Conclusions

We proposed a theoretical model of dynamic trading with a strategic HFT. The model is based on infinite horizon dynamic optimization. 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. We characterize the optimal market making strategy of the HFT.

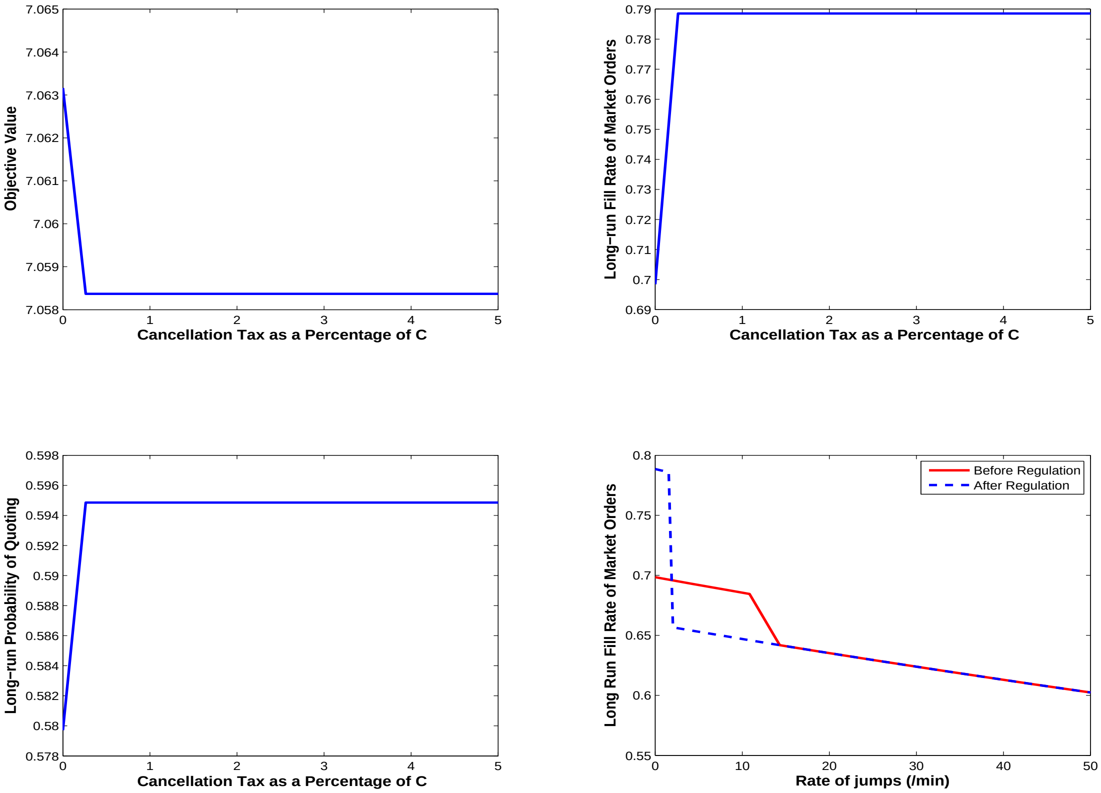

Fig. 19. Taxing Limit Order Cancellations.

Note: The left panels plot the effect of cancellation taxes on the HFTs' value and his provision of liquidity. The right panels display the effect of a cancellation tax on the fill-rate of LFTs' orders (upper panel) and the sensitivity of the fill rate to volatility (in the form of price jumps) before and after a cancellation tax regulation (lower panel).

The model reproduces many important stylized facts about HFTs that have been documented, including their propensity to place and cancel large numbers of orders and to provide plenty of liquidity when the market is quiet but suddenly decrease their provision of liquidity when volatility picks up. Under this policy, we find that the HFT's long-run rate of quoting on both sides of the market increases when he gets faster. We observe that lower latency also drives higher cancellation rates which could be as high as 30% in a realistically calibrated model. Our model also quantifies the impact of higher volatility on liquidity provision, with the HFT providing less liquidity when volatility increases.

Next, we introduced competition for order flow with a nonstrategic medium frequency trader, and computed the change in the HFT's optimal quoting policy. We find that when market makers compete, they split the rent extracted from LFTs, liquidity provision increases and LFTs tend to be better off.

Finally, we analyzed in the context of the model the impact of three widely discussed HFT policies: imposing a transaction tax on each trade, setting minimum-time limits before orders can be cancelled, and taxing the cancellations of limit orders. We find that imposing minimum time-limits and cancellation taxes induce the HFT to quote more on both sides of the market, whereas transaction taxes do not improve this measure of liquidity. One important finding is that when minimum time-limits are in effect, the fill rate of LFTs' market orders by the HFT does not decrease substantially in the presence of higher price volatility, unlike the situation without minimum resting times.

References

Angel, J., Harris, L., Spatt, C. S., 2010. Equity trading in the 21st century. Tech. rep., USC Marshall School of Business.

Biais, B., Foucault, T., Moinas, S., 2011. Equilibrium high-frequency trading. Tech. rep., Toulouse School of Economics.

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

Brogaard, J. A., 2011. High frequency trading and volatility. Tech. rep., University of Washington.

Brogaard, J. A., Hendershott, T., Riordan, R., 2012. High frequency trading and price discovery. Tech. rep., University of California - Berkeley.

Brunetti, C., Kirilenko, A., Mankad, S., 2011. Identifing high-frequency traders in electronic markets: Properties and forecasting. Tech. rep., Johns Hopkins University.

Cespa, G., Foucault, T., 2008. Insiders-outsiders, transparency, and the value of the ticker. Tech. rep., Department of Economics, Queen Mary, University of London.

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., 2012. News trading and speed. Tech. rep., HEC Paris.

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

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., 2010. Low-latency trading. Tech. rep., NYU Stern School of Business.

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

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.

Menkveld, A. J., 2013. High frequency trading and the new-market makers. Journal of Financial Markets, forthcoming .

Moallemi, C., Saglam, M., 2012. The cost of latency. Operations Research, forthcoming .

Pagnotta, E., 2010. Information and liquidity trading at optimal frequencies. Tech. rep., NYU Stern School of Business.

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, Inc.