Adrian, Tobias; Capponi, Agostino; Vogt, Erik; Zhang, Hongzhong

Working Paper

Intraday market making with overnight inventory costs

Staff Report, No. 799

Provided in Cooperation with:
Federal Reserve Bank of New York

Suggested Citation: Adrian, Tobias; Capponi, Agostino; Vogt, Erik; Zhang, Hongzhong (2016) : Intraday market making with overnight inventory costs, Staff Report, No. 799, Federal Reserve Bank of New York, New York, NY

This Version is available at:
https://hdl.handle.net/10419/175207

Federal Reserve Bank of New York
Staff Reports

Intraday Market Making with Overnight Inventory Costs

Tobias Adrian
Agostino Capponi
Erik Vogt
Hongzhong Zhang

Staff Report No. 799
October 2016

Intraday Market Making with Overnight Inventory Costs

Tobias Adrian, Agostino Capponi, Erik Vogt, and Hongzhong Zhang

Federal Reserve Bank of New York Staff Reports, no. 799
October 2016
JEL classification: G01, G12, G17

Abstract

The share of market making conducted by high-frequency trading (HFT) firms has been rising steadily. A distinguishing feature of HFTs is that they trade intraday, ending the day flat. To shed light on the economics of HFTs, and in a departure from existing market-making theories, we model an HFT that has access to unlimited leverage intraday but must fund any end-of-day inventory at an exogenously determined cost. Even though the inventory costs occur only at the end of the day, they impact intraday price and liquidity dynamics. This gives rise to an intraday endogenous price impact mechanism. As the end of the trading day approaches, the sensitivity of prices to inventory levels intensifies, making price impact stronger and widening bid-ask spreads. Moreover, imbalances of buy and sell orders may catalyze hikes and drops in prices, even under fixed supply and demand functions. Empirically, we show that these predictions are borne out in the U.S. Treasury market, where bid-ask spreads and price impact tend to rise toward the end of the day. Furthermore, price movements are negatively correlated with changes in inventory levels as measured by the cumulative net trading volume.

Key words: market microstructure, market liquidity, high-frequency trading, financial intermediation

---

Adrian,Vogt: Federal Reserve Bank of New York (e-mails: tobias.adrian@ny.frb.org, erik.vogt@ny.frb.org). Capponi, Zhang: Columbia University (e-mails: ac3827@columbia.edu, hz2244@columbia.edu).

1 Introduction

Over the last two decades, high-frequency trading firms (HFTs) have emerged as a new and significant class of financial intermediaries. While representing only a small fraction of trading fifteen years ago, HFTs now account for around half of the volume in major equity, Treasury, foreign exchange, and associated futures markets (see Securities and Exchange Commission (2010), Joint Staff Report (2015), and BIS (2011)). Furthermore, HFT-driven information flows, as measured by the messages sent to and from exchanges, has in some venues reached 80 percent of total message traffic in normal times (Joint Staff Report (2015)). This increased presence of algorithmic and high frequency trading is profoundly impacting various facets of market quality, which has been the subject of intense research (see Herndeshott et al. (2011), Menkveld (2013), Brogaard et al. (2014), and Herndeshott and Menkveld (2014)).

One key area in which HFTs affect market quality is through liquidity provision, or market making. Like traditional dealers, HFTs provide liquidity to the market, in the sense of temporarily warehousing securities to intermediate buyers and sellers through time and across markets. However, in contrast to traditional dealers, HFTs perform intermediation services by trading on their own account and hence differ materially in terms of their funding structures. Thus, without the use of external funding through money markets, debt markets, and other liabilities, HFT balance sheets tend to be substantially smaller than those of traditional sell-side market makers. Smaller balance sheets enforce the need to keep positions small and short-lived in order to limit the amount of capital held in margin accounts (Menkveld (2016)). As a consequence, most of the trading that HFTs undertake occurs intraday, with positions largely unwound at the end of each trading day. The practice of ending the day flat is in fact often used as a defining characteristic of market making HFTs (see Securities and Exchange Commission (2010), Joint Staff Report (2015), Menkveld (2016), SEC letters (2010)).

This paper studies intraday market making with the added objective of ending the trading day flat. Specifically, we present a model of a market making HFT who dynamically places bid and ask prices in order to maximize end-of-day profits, but with the additional goal of unwinding its positions before markets close. The HFT in this context can be interpreted as having access to unlimited leverage intraday while facing an exogenously determined cost that is proportional to its end-of-day inventory. We show that even though the inventory cost is assessed only at the end of the day, the HFT's intertemporal hedging demand due to the inventory cost impacts liquidity and trade price dynamics throughout the day. Intuitively, the end of day constraint induces a nontrivial feedback mechanism between inventory levels and prices. The adjustment of inventory by the market maker and price changes reinforce each other, putting high pressure on prices when fast and extreme imbalances in inventory occur. At the same time, the zero intraday trading costs allow the

HFT to trade aggressively, thus leading to a compression of bid ask spreads on average.

Our model produces a number of insights that extend the intuition from existing inventory control problems. First, the degree to which ask and bid prices are impacted by the HFT's inventory in our setup is primarily determined by the shadow value of the overnight inventory constraint. The shadow value of the overnight inventory constraint therefore represents a new time-varying component of the bid-ask spread beyond those previously identified in the literature (see Stoll (1989), Glosten and Harris (1988)). Second, because the HFT intermediates between randomly arriving investors, the bid-ask spreads in our model also turn out to be functions of the arrival intensities of buyers and sellers. Narrow bid-ask spreads correspond to high investor arrival intensities. A liquid market with narrow bid ask spreads is therefore defined as one where buyers and sellers arrive with high intensity, which allows the HFT to have greater control over its inventory. Third, and most importantly, a more severe end-of-day inventory constraint will cause prices to be more sensitive to inventory, which gives rise to an intraday *endogenous price impact*. As time approaches the end of the trading day, prices become more sensitive to inventory levels, leading to stronger price impact and larger bid-ask spreads. When compared to a benchmark model without end-of-day trading costs where bid-ask spreads are evenly distributed throughout the day, the model with end-of-day trading costs generates bid-ask spreads and trade prices that are more sensitive to order flow imbalances.

We verify the testable implications of our model using high frequency U.S. Treasury data. The intraday pattern of bid-ask spreads is strongly supported by the data. We also find a highly significant negative relationship between inventory and prices, a feature that was especially pronounced on October 15, 2014, when the so called *Flash Rally* occurred in the U.S. Treasury market. During such flash events, bid-ask spreads remain tight despite sharp declines in depth, as trading intensities increase. Moreover, the sensitivity of prices to the inventory level tends to increase as time approaches the day's end. Furthermore, the bid-ask spread trajectory also tends to increase as time approaches the day's end, reflecting the HFT's effort to control the end of day inventory.

To quantify the economic impact of the HFT's price determination, we perform comparative statics analyses on two measures of price stability: the largest bid-ask spread throughout the trading day and the worst case deviation of traded prices from a benchmark equilibrium price. These measures reflect the disadvantage that end investors accrue through the HFT's inefficient intermediation activities. We also consider the maximum drawdown of the midquote price, which captures the stability of the financial market intermediated by the HFT. Higher arrival rates of buy and sell orders reduce bid-ask spreads and price deviation from the equilibrium at the expense of increasing price variations. If the end-of-day inventory cost is very high, the maximum drawdown of the midquote price decreases as the intensity of order arrivals increases. This is because the HFT puts more effort into controlling its inventory and treats higher arrival

rates as opportunities for managing inventory. On the other hand, if the end-of-day inventory cost is low, the maximum drawdown increases with the arrival rates. Under these circumstances, the HFT mainly focuses on exploiting trading profits when the intensity of order arrivals increase, and is not much concerned with inventory control.

We also conduct a welfare analysis based on the surplus earned by buy and sell investors, and the maximized objective function of the HFT. Our analysis suggests that a socially desirable system is obtained when buyers and sellers arrive with high intensities (i.e. the market is highly liquid) and intermediated by an HFT with low end of day inventory costs. Higher arrival rates of buy and sell orders result in more trading opportunities for buyer and seller investors. Moreover, they help the HFT reduce the price impact coming from overnight inventory costs. On the other hand, a higher penalty for inventory holdings intensifies the impact of tail risk on welfare, especially if the arrival rates are high.

The rest of the paper is organized as follows. We explain our contribution to the existing literature on high-frequency trading in Section 2. We discuss the relevant institutional details of HFTs in Section 3. We introduce the theory of HFT market making in Section 4. Section 5 formulates the decision making problem of the HFT and the solution for the optimal bid and ask price trajectories. We provide an empirical analysis testing the implications of our model in Section 6. Section 7 presents an analysis of price stability, and Section 8 discusses welfare. Section 9 concludes. Proofs are in the appendix.

2 Literature Review

There are several microstructure models that analyze the price impact of trades and the determinants of bid-ask spreads. The revenues of market makers, who provide liquidity, must offset their incurred costs. These costs can be inventory costs (e.g. Stoll (1980) and Amihud and Mendelson (1980)) or adverse selection costs (e.g. Kyle (1985) and Glosten and Milgrom (1985)).

Models with inventory costs predict that dealers set quotes in order to maintain their inventory level around a target level. The early works by Stoll (1980) assume that the market makers are risk-averse. Stoll considers a mean-variance market maker wishing to optimize his or her expected profit from bid-ask spreads, and to quickly find offsetting transactions in order to minimize inventory risk. Thus, large positive inventories can be reduced by lowering ask prices, and negative inventories can be unwound by setting higher bids. Stoll's model predicts a linear relation between prices and inventories. Amihud and Mendelson (1980) consider a dynamic model in which dealers are risk neutral and have hard constraints on their inventory. In their model, buy and sell orders arrive according to a Poisson process with price-dependent rates. They find that inventories have an impact on equilibrium prices, and that trading volume decreases as the inventory

approaches the long or short limit. Furthermore, they find that the bid-ask spread widens. Recently, Aït-Sahalia and Saglam (2016) develop a model in which a risk-neutral high frequency trader maximizes its expected reward minus a penalty cost for holding inventory throughout the trading period. The HFT decides whether to quote on one side (buy or sell), both sides, or not quote in each point in time, and it is allowed to cancel orders. Trades are executed at the best bid and ask quotes when market orders arrive, earning the HFT a fixed spread in each transaction.

While Stoll (1980) presents a static model, we are studying the dynamic implications of end of day inventory costs. Amihud and Mendelson (1980) also consider a dynamic model, but assume an infinite time horizon and restrict inventory levels to be inside a prespecified interval throughout the trading day. The model of Aït-Sahalia and Saglam (2016) contrasts with our approach in that we allow bid and ask quotes to be endogenously chosen, taking into account demand functions of buy and sell investors, as well as inventory costs incurred at the end of the day.

The seminal contributions by Kyle (1985) and Glosten and Milgrom (1985) derive the equilibrium prices in a model with information asymmetries and a monopolistic market maker. The latter must cover losses from transactions with traders who have access to superior information by charging a spread. In Kyle (1985)'s model, order flow is driven by uninformed traders, who only trade for liquidity purposes and hence prices do not reflect full information. His model predicts a linear market depth, i.e. prices vary linearly with the aggregate order flow. The setting of our model is closely related to another canonical model of market microstructure proposed by Glosten and Milgrom (1985). Differently from Kyle (1985), in which the monopolistic market maker fills aggregate order imbalances, in Glosten and Milgrom (1985) the market maker observes the orders submitted by informed and uninformed traders, who arrive in the market according to a Poisson process with exogenously specified intensities. Bid and ask quotes are optimally chosen by the market maker based on the probability that an arrival order is informed.

Admati and Pfleiderer (1988) allow liquidity traders to be strategic on the time of their trades. Their model includes intraday effects, while earlier literature focused primarily on day to day liquidity dynamics. The model of Admati and Pfleiderer (1988) reproduces certain empirical facts, such as the U-shaped pattern of trading volume throughout the day. Danilova and Julliard (2015) develop a rational expectation equilibrium model that explains volatility, liquidity, and trading activity by the degree of asymmetric information and trading frictions. Volatility information is released to the market at trading times that, due to traders' strategic choices, differ from calendar times. The model makes predictions about volatility, price quotes, tightness, depth, resilience, and trading activity which are borne out in high frequency trading data. Foucault et al. (2016) present a model of high frequency trading where dealers receive public high frequency news about fundamentals, while speculators have private signals about long term fundamental values. They

show that high frequency trading on the public information only arises when the speculator is fast relative to the dealer, meaning that he or she can trade on forecasted price movements before the dealer receives the news. The price process features a volatility component that is driven by the speculator's instantaneous forecast of news.

In contrast to Kyle (1985), Glosten and Milgrom (1985), Admati and Pfleiderer (1988), Danilova and Julliard (2015), and Foucault et al. (2016), we abstract from asymmetric information and strategic behavior in order to focus on the role of the end of day inventory constraint. In particular, our theory explicitly aims at capturing the intraday price and liquidity impact of the end of day inventory cost.

Empirical studies have analyzed the relationship between trades, prices and bid-ask spreads using transaction data. Glosten and Harris (1988) and Hasbrouck (1988) decompose bid-ask spreads into two components, reflecting compensation for inventory costs and adverse selection costs, which arise from the presence of informed traders. They find that, in contrast to the transitory spread component explained by inventory considerations, the permanent component explained by information asymmetries is significant for large trades but not for small ones. Hasbrouck and Saar (2013) show that low latency improves market quality by reducing bid-ask spreads, the total price impact of trades, and short-term volatility. Herndeshott et al. (2011) also come to similar conclusions and find that for large stocks, algorithmic trading reduces bid-ask spreads and adverse selection and also improves price discovery. Brogaard et al. (2014) find that algorithmic trading facilitates price discovery as high frequency traders trade in the direction of permanent price changes and in the opposite direction of transitory pricing errors. Herndeshott and Menkveld (2014) analyze the transitory component of price changes, defined as pressures temporarily moving prices away from fundamentals, and relate them to the HFT's inventory using data from the New York Stock exchange. In their model, the HFT trades off revenue loss coming from price pressures with price risk coming from a state of nonzero inventory. Menkveld (2013) studies the trading strategy of a large high-frequency trader accounting for utility and cost of holding inventory during periods of pressure. Chaboud et al. (2014) analyze the effects that algorithmic trading has on the informational efficiency of foreign exchange prices, showing that it speeds up price discovery but at the same time imposes higher adverse selection costs on slower trades. These findings are in line with Biais et al. (2015) who show that high speed technology enable fast traders to retrieve information before slow traders, generating adverse selection, and thus negative externalities. We also refer to Jones (2013) and Menkveld (2016) for reviews of theoretical and empirical research in high-frequency trading.

3 HFT Inventory Costs

To the best of our knowledge, our study is the first to explicitly consider the impact of an end of day inventory cost on intraday pricing and liquidity dynamics. The HFT's desire to end the day with little to no inventory is the key distinguishing feature of our model relative to the existing literature on market making. This desire to end the day "flat" appears to be a universally agreed upon characteristic of HFTs. For example, in their concept release on equity market structure, the Securities and Exchange Commission (2010), p. 45, defines HFTs as professional traders that engage in a large number of transactions intraday and that possess common characteristics, which include: "(1) the use of extraordinarily high-speed and sophisticated computer programs for generating, routing, and executing orders; (2) use of co-location services and individual data feeds offered by exchanges and others to minimize network and other types of latencies; (3) very short time-frames for establishing and liquidating positions; (4) the submission of numerous orders that are cancelled shortly after submission; and (5) ending the trading day in as close to a flat position as possible (that is, not carrying significant, unhedged positions over-night)." Moreover, the Joint Staff Report (2015) on the U.S. Treasury market flash event on October 15, 2014, goes further by saying that the desire to end the day flat differentiates HFTs from traditional bank dealers as market makers, who in contrast "routinely end trading sessions with sizable long or short positions both in the cash and futures markets." In his recent survey, Menkveld (2016) corroborates these statements, highlighting that HFTs are best thought of as a new type of financial intermediary, who trade a lot intraday but avoid carrying a position overnight.

Several empirical studies are strongly supportive of this characterization. Using data on a Dutch equity trading venue, Jovanovic and Menkveld (2011) are able to identify a participant with a unique broker ID that trades very frequently, representing roughly every third trade on the venue. They suggest that "What makes this broker an HFT, though, is that his net position over the trading day is zero almost half of the sample days." A figure of this broker's net inventory was reproduced in the survey by Biais and Woolley (2011), which shows periods of autocorrelated positive and negative inventory eventually ending at exactly zero at close. Biais and Woolley (2011) suggest that this behavior is emblematic of HFTs, who are differentiated from other market participants by their short investment horizons: "The key difference is the holding period or investing horizon. That of HFT ranges between milliseconds and hours. Their entire positions are closed at the end of each trading day." This behavior is supported in a separate study conducted by Benos and Sagade (2016), who analyze proprietary participant-level data from U.K. equity markets over a four-month period. Their findings suggest that "HFTs generally end the day with a relatively flat position," with the mean HFT having a volume-weighted end-of-day position corresponding to 5% of their total intraday volume.

In the U.S. Treasury market, the Joint Staff Report (2015) found that "...a significant share of PTF

[HFT] activity focuses on the provision of short-term liquidity on both sides of the market, and as such their high observed trading volume in the Treasury market does not translate into net changes in their positions across a trading session. An analysis of account-level data in the Treasury futures market over a number of days that include October 15 shows that more than 80 percent of trading in the 10- and 30-year contracts represented short-term intraday turnover." The Joint Staff Report (2015) furthermore shows that the median HFT ended the trading day with an absolute position of less than 5% of their total intraday volume. In the Treasury futures market, this figure shrinks to 1%.

The December 2015 Senior Credit Officer Survey on Dealer Financing Terms by the Board of Governors of the Federal Reserve System (2016) summarizes answers to special questions on intraday and overnight credit extended to HFTs. Overnight positions of HFTs are reported to be de minimis when compared to intraday positions. Importantly, intraday exposure management is primarily done via exposure limits, not margining. In addition to this survey, evidence from the margining documentation by central clearing platforms and exchanges paints a complementary picture on the limited usage of intraday margin. Central clearing counterparties tend to compute variation margins at discrete times during the day, or at the end of the day. This evidence suggests that intraday inventory costs might be close to, or exactly, zero, depending on where the HFT is trading.

This evidence on the intraday credit risk management of exchanges, central clearing counterparties, and dealers clearly suggests incentives for HFTs to carry little inventory overnight. The academic literature discusses a few additional underlying reasons for closing out positions at the end of the trading day. Brogaard and Garriott (2015) suggest a risk management motive, as HFTs wish to avoid exposure to the risk that asset values might change overnight. While such a motive may purely be driven by risk aversion, it may also be driven by the desire to avoid overnight margin requirements or other funding costs. For example, overnight positions might have to be funded in the repo or securities lending markets, requiring haircuts. Furthermore, a reduction in inventory results in a reduction of the HFT's value-at-risk, which in turn reduces any overnight margining costs. Indeed, brokers typically require additional initial and maintenance margins for positions held overnight.[1] The effect of increased overnight margining costs is a need for the HFT to deleverage before the end of the trading day.

The main contribution of this paper is to be the first to study the implications of HFTs' zero-overnight-inventory motive. We show that an HFT that manages its inventory with an eye towards ending the day flat behaves differently *at all hours of the day* compared to a benchmark market maker without such an end-of-day objective. The intuition follows from a backward induction argument, which implies that a one-time, end-of-day inventory cost will be factored into trading decisions that recursively trace back to the start of

[1] See, for example, https://gdcdyn.interactivebrokers.com/en/index.php?f=marginnew&p=overview1

the trading day. These trading decisions, in turn, have implications for both price and liquidity dynamics throughout the course of the trading day. In particular, we show that because price impact endogenously steepens with the strength of the zero-overnight-inventory motive, sudden intraday price moves or flash events can be amplified by the end-of-day inventory constraint. In markets that are increasingly intermediated by HFTs, this type of dynamic raises potential financial stability considerations.

4 The Model

We consider a model in which the trading day runs from time zero to T and is divided into T steps. There is a single asset traded in an electronic limit order book: an HFT sells for buy orders and buys for sell orders. The HFT is the only counterparty available for trade when an arrival event occurs. Buy and sell orders arrive in the market according to a Bernoulli process. The staggered arrival of buy and sell orders creates a supply and demand for immediacy, a concept first introduced in the finite-period model of Grossman and Miller (1988). Buy orders create selling pressure that pulls prices down and away from the equilibrium. Sell orders instead create buying pressure to bring prices back up to the equilibrium. Throughout the paper, we use $\tilde{b}$ to denote the bid price, and $\tilde{a}$ to denote the ask price.

4.1 Buy and Sell Orders

The arrival of orders is modeled as a Bernoulli process. In each time step $s \in \{0, 1, \dots T\}$, a trading order arrives with probability π. We split the arrivals of the Bernoulli process into two new arrival processes. Each arrival of the original Bernoulli process goes into the first of the two new processes, referred to as the buy order arrival process and denoted by N^{BO}, with probability $\frac{\pi^{BO}}{\pi}$. It goes into the second of the two new processes, referred to as the sell order arrival process and denoted by N^{SO}, with probability $\frac{\pi^{SO}}{\pi}$. Therefore, we have $\pi = \pi^{BO} + \pi^{SO}$.

We use $\tilde{q}$ to denote the minimum price at which a sell order is placed, and by $\tilde{p}$ the maximum price at which a buy order is placed. $\tilde{q}$ is the reservation price for sell orders, and can be interpreted as a stop loss. $\tilde{p}$ can be viewed symmetrically for buy orders.

For a given ask price x, the number of shares demanded by buy investors is given by

$$Q^{BO}(x) = c\,(\tilde{p} - x)^{+}\,. \tag{1}$$

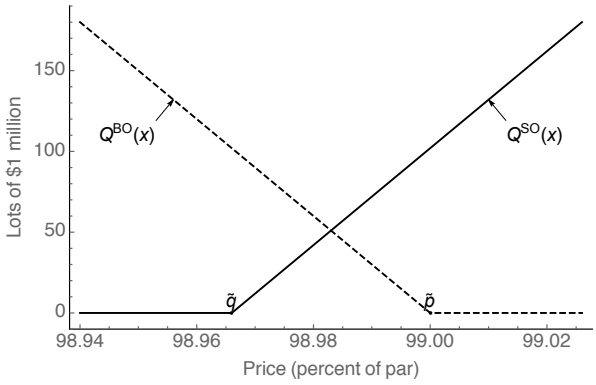


Figure 1: **Example Demand and Supply Functions of Buy and Sell Investors.** This figure illustrates the supply and demand functions of non-HFT investors. The price (x-axis) is in terms of percentage of par, and the quantity (y-axis) is measured in lots of \$1 million. We set $\tilde{p} = 99$, $\tilde{q} = 98.966$ and the slope $c = 30$ (lots of \$1 million per basepoint of par). The quantity supplied by a sell investor (solid) is an increasing function of price within the stop loss thresholds. Similarly, the quantity demanded by buy investors (dashed) declines in price within the same thresholds.

For a given bid price x, the number of shares supplied by sell investors is given by

$$Q^{SO}(x) = c\,(x - \tilde{q})^{+}. \tag{2}$$

Above, we have assumed that both the demand and supply curve have the same slope c. Such an assumption is driven by an empirical analysis of the limit order book data in the U.S. Treasury market. Note that the demand and supply functions Q^{BO} and Q^{SO} are reduced forms for preferences, beliefs, investment objectives, and hedging motives of buyers and sellers.

For future reference, we introduce the *equilibrium price* $\bar{p}$ defined as the price at which demand and supply intersect in a frictionless market. A direct calculation shows that this is given by

$$\bar{p} = \frac{\pi^{BO}\tilde{p} + \pi^{SO}\tilde{q}}{\pi}. \tag{3}$$

This may also be interpreted as the price in a hypothetical market where the market maker minimizes expected order imbalances. More precisely, it may be easily verified that such a price corresponds to the solution of the following minimization problem:

$$\min_{x} \mathbb{E}\left[\left(Q^{BO}(x)\,N_t^{BO} - Q^{SO}(x)\,N_t^{SO}\right)^2\right], \tag{4}$$

where N_t^{BO} and N_t^{SO} denote, respectively, the number of buy and sell orders arrived by time t.

HFT

In our model of market making, the HFT optimally chooses bid and ask prices through time. The wealth of the HFT at time t, W_t, is given by the initial cash holdings of the HFT, W_0, plus the cumulative gains from trades with buyers, minus the cumulative expenses from trades with sellers. Specifically, a trade with a buy investor at time t results in $Q^{BO}(\tilde{a}_t)$ shares of the asset sold at price $\tilde{a}_t$; likewise, a trade with a sell investor at time t results in $Q^{SO}(\tilde{b}_t)$ shares of the asset bought at price $\tilde{b}_t$. The wealth at time t is given by

$$W_t = W_0 + \sum_{s=1}^{t} \tilde{a}_s \, Q^{BO}(\tilde{a}_s) \, \Delta N_s^{BO} - \sum_{s=1}^{t} \tilde{b}_s \, Q^{SO}(\tilde{b}_s) \, \Delta N_s^{SO}, \tag{5}$$

where we set $\Delta N_s^{SO} := N_s^{SO} - N_{s-1}^{SO}$ and $\Delta N_s^{BO} := N_s^{BO} - N_{s-1}^{BO}$. The HFT's inventory accumulated in the interval $[0, t]$ is given by the sum of shares bought from the seller, minus the sum of shares sold to the buyer until time t. That is,

$$I_t = \underbrace{\sum_{s=1}^{t} Q^{SO}(b_s) \, \Delta N_s^{SO}}_{\substack{\text{Shares bought} \\ \text{from sell investors}}} - \underbrace{\sum_{s=1}^{t} Q^{BO}(a_s) \, \Delta N_s^{BO}}_{\substack{\text{Shares sold} \\ \text{to buy investors}}}. \tag{6}$$

We model the objective of the HFT in reduced form. The HFT is risk neutral and maximizes its end of day wealth W_T, but is subject to an end of day inventory cost of size λI_T^2. Such a cost is the novel aspect in our study and will play a major role in the forthcoming analysis. As discussed earlier in Section 3, HFTs tend to have de minimus balance sheets, thus making any overnight inventory costly to carry. We additionally assume that end of day inventory is valued at the equilibrium price $\bar{p}$. This amounts to assuming that inventory is marked at $\bar{p} I_T$.[2]

Altogether, this leads to the following maximization problem for the HFT:

$$\max_{(\tilde{a}_\cdot, \tilde{b}_\cdot) \in (\mathbb{R}_+^2)^T} \mathbb{E}\left[W_T - \lambda I_T^2 + \bar{p} I_T \right], \tag{7}$$

subject to the budget constraint (5).

The HFT's problem amounts to optimally choosing the ask and bid paths $(\tilde{a}_\cdot, \tilde{b}_\cdot) = (\tilde{a}_t, \tilde{b}_t)_{t=1}^T$ which maximize the expected utility from terminal wealth net of overnight inventory costs. The ask $\tilde{a}_t$ and bid $\tilde{b}_t$ are decided based on the information available by time $t-1$.

The above described model is related to previously proposed market making models of inventory management. These include the monopolistic market making model proposed by Amihud and Mendelson (1980), where a specialist has balance sheet costs throughout the trading horizon which is assumed to be infinite. In

[2] This could also reflect expected prices in an overnight market in which the HFT does not participate.

their model, the dealer is constrained to hold the inventory within a pre-specified interval at all times, and optimally chooses the bid-ask prices to maximize the long term growth rate of his or her wealth process. As in our model, buyer or sellers arrive randomly; however they can only trade one unit of the asset in each trade. Aït-Sahalia and Saglam (2016) also consider a perpetual decision making problem for an HFT who incurs intraday costs for holding inventory. In their model, buyers and sellers arrive at random times and the HFT decides whether to transact with one of them so as to maximize its expected discounted payoff. In each trade, the HFT earns a fixed bid-ask spread. The strategy of the HFT is time homogeneous and of threshold type. By contrast, as we demonstrate in the forthcoming sections, the forward looking nature of the end-of-day constraint has a strong impact on the intraday price and spread dynamics in our model.

5 The Control Problem

This section studies the dynamic optimization problem of the HFT. The primary state variable in our decision making problem is the inventory level of the HFT. As inventory increases, so does the shadow cost of the end-of-day inventory constraint. Furthermore, that shadow cost rises throughout the day, and reaches its highest value just before the day's end. As a result, the HFT will try to maintain a smaller inventory position as time progresses to avoid bearing such an increasing cost.

We will first formulate the dynamic programming problem. We then characterize the optimal price setting behavior. We also present comparative statics in a sequence of propositions, which serve as a basis for the empirical analysis conducted in the following section.

5.1 Dynamic Programming Formulation

The value function of the control problem, defined as the HFT's continuation value at time t given its current level of wealth w and of inventory i, is given by

$$V(t,w,i) := \sup_{(\tilde{a}_\cdot,\tilde{b}_\cdot)\in(\mathbb{R}_+^2)^T} \mathbb{E}\left[U\big(W_T^{(\tilde{a},\tilde{b})}, I_T^{(\tilde{a},\tilde{b})}\big)|W_t^{(\tilde{a},\tilde{b})} = w, I_t^{(\tilde{a},\tilde{b})} = i\right], \tag{8}$$

where the end-of-day utility of the HFT is

$$U(w,i) = w - \lambda i^2 + \bar{p}i.$$

The state variables $(W_t^{(\tilde{a},\tilde{b})}, I_t^{(\tilde{a},\tilde{b})})_{t=0}^{T-1}$ are given by the wealth and inventory processes of equations (5) and (6). By virtue of the dynamic programming principle (see Puterman (1994)), for $0 \le t \le u \le T$, we have

that

$$V(t,w,i) = \sup_{(\tilde{a},\tilde{b})\in\mathbb{R}^2_+} \mathbb{E}\left[V\big(u, W_u^{(\tilde{a},\tilde{b})}, I_u^{(\tilde{a},\tilde{b})}\big)|W_t^{(\tilde{a},\tilde{b})} = w, I_t^{(\tilde{a},\tilde{b})} = i\right]. \tag{9}$$

Intuitively, the value $V(t,w,i)$ gives the optimal expected utility at a future time instant.

From our pre-specified supply and demand curves, we know that an incoming buy order at time t will reduce the HFT's inventory by $Q^{BO}(\tilde{a}_t)$, while increasing the wealth of the HFT by $\tilde{a}_t \times Q^{BO}(\tilde{a}_t)$. Likewise, an incoming sell order at time t will increase the HFT's inventory by $Q^{SO}(\tilde{b}_t)$, while reducing the wealth of the HFT by $\tilde{b}_t \times Q^{SO}(\tilde{b}_t)$. Therefore, for any (Markov) control on the ask and bid prices $(\tilde{a}_t, \tilde{b}_t)_{t=1}^T$, the controlled state process $(W_t^{(\tilde{a},\tilde{b})}, I_t^{(\tilde{a},\tilde{b})})_{t=0}^T$ constitutes a controlled Markov process. Specifically, given the state $\{W_{t-1}^{(\tilde{a},\tilde{b})} = w, I_{t-1}^{(\tilde{a},\tilde{b})} = i\}$ and the control pair $(\tilde{a}_t, \tilde{b}_t)$, we have the time t transition probability of the wealth and inventory given by

$$(W_t^{(\tilde{a},\tilde{b})}, I_t^{(\tilde{a},\tilde{b})}) = \begin{cases} (w + f^{BO}(\tilde{a}_t), i - Q^{BO}(\tilde{a}_t)), & \text{with probability } \pi^{BO}, \\ (w - f^{SO}(\tilde{b}_t), i + Q^{SO}(\tilde{b}_t)), & \text{with probability } \pi^{SO}, \\ (w, i), & \text{with probability } 1 - \pi^{BO} - \pi^{SO}, \end{cases} \tag{10}$$

where we have introduced the following notation:

$$f^{BO}(x) = x\, Q^{BO}(x), \qquad f^{SO}(x) = x\, Q^{SO}(x). \tag{11}$$

Clearly, the time when the Markov process transits is completely determined by the exogenously given arrival sequences of buy orders and sell orders. Yet, as seen from (10), the control on ask and bid prices influences the possible states reached after a trade, and hence serves as an effective means of controlling inventory to the HFT.

From equations (9) and (10), we obtain the following Bellman equation:

$$V(t-1,w,i) = V(t,w,i) + \sup_{(\tilde{a}_\cdot,\tilde{b}_\cdot)\in(\mathbb{R}^2_+)^T} H(t,\tilde{a},\tilde{b}), \tag{12}$$

with terminal condition $V(T,w,i) = U(w,i)$, where H denotes the Hamiltonian given by

$$H(t,\tilde{a},\tilde{b}) := \pi^{BO}[V(t,w+f^{BO}(\tilde{a}), i - Q^{BO}(\tilde{a})) - V(t,w,i)] + \pi^{SO}[V(t,w-f^{SO}(\tilde{b}), i + Q^{SO}(\tilde{b})) - V(t,w,i)].$$

The linearity of the value function V in the wealth variable w suggests that we can rewrite

$$V(t,w,i) = w + F(t,i) \tag{13}$$

where $F(t,i) \equiv V(t,0,i)$, is the optimal expected utility of an HFT who possesses zero wealth and an inventory level i at time t (more precisely, after the trade at time t, if it occurs). At the end of the trading day, i.e. at $t = T$, the remaining inventory I_T is valued at the price $\bar{p}$, hence we have that $F(T,i) = -\lambda i^2 + \bar{p}i$. From (12), we deduce that for $1 \leq t \leq T$, the function F solves the following nonlinear equation:

$$F(t-1,i) = F(t,i) + \sup_{(\tilde{a},\tilde{b})\in\mathbb{R}_+^2} \tilde{H}(t,\tilde{a},\tilde{b}), \tag{14}$$

where the new Hamiltonian $\tilde{H}$ is defined as

$$\tilde{H}(t,\tilde{a},\tilde{b}) := \pi^{BO}[f^{BO}(\tilde{a}) + F(t,i-Q^{BO}(\tilde{a})) - F(t,i)] + \pi^{SO}[-f^{SO}(\tilde{b}) + F(t,i+Q^{SO}(\tilde{b})) - F(t,i)]. \tag{15}$$

The Hamiltonian $\tilde{H}(t,\tilde{a},\tilde{b})$ measures the utility of the HFT, as seen from time $t-1$, of choosing bid and ask prices $(\tilde{b},\tilde{a})$ at time t and then trading optimally for the remainder of the day. In particular, suppose the HFT has inventory level i at time $t-1$, then an incoming buy order at time t will increase the net wealth of the HFT by $f^{BO}(\tilde{a}_t)$, and leave the inventory of the HFT at $i - Q^{BO}(\tilde{a}_t)$. This is worth $F(t, i-Q^{BO}(\tilde{a}_t))$ in utility terms. Symmetrically, under the same circumstances, an incoming sell order will reduce the wealth of the HFT by $f^{SO}(\tilde{b}_t)$, and leave its inventory at $i + Q^{SO}(\tilde{b}_t)$. This is worth (in utility) $F(t, i+Q^{SO}(\tilde{b}_t))$. Consequently,the main objective is to choose the optimal ask $\tilde{a}$ and bid $\tilde{b}$ so as to best control the inventory level while at the same time maximizing proceeds from buy and sell trades.

We now make a change of variables

$$a = c\tilde{a}, b = c\tilde{b}, p = c\tilde{p}, q = c\tilde{q},$$

which will be useful for subsequent analyses. Going forward, we will refer to a and b as the scaled ask and bid prices. Note that they share the same unit as the HFT's inventory level. With this notation, the optimization problem in (14) may be written as

$$F(t-1,i) = F(t,i) + H_t(i), \tag{16}$$

where $H_t(i)$ is the optimized Hamiltonian

$$H_t(i) := \sup_{(a,b)\in\mathbb{R}_+^2} \left\{ \pi^{BO} \left[\frac{1}{c} a(p-a)^+ + F(t, i-(p-a)^+) - F(t,i) \right] + \pi^{SO} \left[-\frac{1}{c} b(b-q)^+ + F(t, i+(b-q)^+) - F(t,i) \right] \right\}. \tag{17}$$

5.2 Optimal Price Policies and their Dependence on Inventory

This section studies the dependence of bid and ask prices on the inventory levels at a specific time. We determine the bid and ask prices which maximize the expected utility of the HFT by solving the Bellman equation (16). In terms of our scaled bid and ask variables, note that $(b-q)^+$ is the quantity that the HFT purchases from sell investors, while $(p-a)^+$ is the quantity that the HFT sells to buy investors. Our methodology is based on a backward induction algorithm that involves a certain invariant convex property of the function $F(t,i)$, which we establish. To that end, we assume that for $t = 1, 2, \ldots, T$ the function $F(t,i)$ is strictly concave and continuously differentiable in i with a derivative mapped onto $\mathbb{R}$. Notice that these properties mean that the function $F(t,i)$ behaves like a quadratic function with a negative leading coefficient. In particular, when $|i|$ is very large, the optimal expected utility $F(t,i) \ll 0$ and the marginal optimal exepcted utility $\partial_i F(t,i) > 0$ if $i < 0$ and $\partial_i F(t,i) < 0$ if $i > 0$, which will translate into reducing inventory of size $|i|$ through trading.

Using this function $F(t,i)$, we will derive the optimal ask $\tilde{a}_t^*$ and the optimal bid price $\tilde{b}_t^*$, as well as their monotonicity with respect to the inventory level. We then prove that the differential and convex properties that $F(t,i)$ possesses carry over to $F(t-1,i)$. Hence, an induction argument will establish the results for all t (see Proposition 5.3 for the details).

In the remainder of the section, we determine the optimal ask and the optimal bid prices. First, it can be clearly seen that even though $F(t,i)$ is assumed to be continuously differentiable in i, the objective function in (16) is not differentiable in a, b. Thus, to apply the first order condition, we will need to consider a simplified, smoothed version of (16), and then relate the optimum of this simplified optimization problem to the original problem. Specifically, consider candidate optimal scaled ask and bid prices as follows

$$a_t^*(i) = \max\{a_t(i), 0\}, \tag{18}$$

$$b_t^*(i) = \max\{b_t(i), 0\}. \tag{19}$$

Above, the functions $a_t(i), b_t(i)$ are the solutions to the unconstrained version of the dynamic optimization

problem in equation (16), i.e. without the constraint $a, b \geq 0$ and without the plus sign in the demand and supply functions:

$$\sup_{(a,b)\in\mathbb{R}^2} \left\{ \pi^{BO} \left[\frac{1}{c} a(p-a) + F(t, i-(p-a)) - F(t,i) \right] + \pi^{SO} \left[-\frac{1}{c} b(b-q) + F(t, i+(b-q)) - F(t,i) \right] \right\}. \tag{20}$$

The relaxation in (20) makes the optimization problem analytically tractable: Because the function $F(t,i)$ is strictly concave in i, we notice that for each fixed i, the mapping $a \mapsto \frac{1}{c}a(p-a) + F(t, i-(p-a)) - F(t,i)$ is strictly concave in a. Likewise, the mapping $b \mapsto -\frac{1}{c}b(b-q) + F(t, i+b-q) - F(t,i)$ is strictly concave in b. Hence, for each fixed i, there is a unique optimal pair $(a_t(i), b_t(i))$ that maximizes the unconstrained Hamiltonian in (20) (notice that it may still occur that $a_t(i) < b_t(i)$). In addition, $a_t(i), b_t(i)$ are the solutions of the decoupled system of first order conditions given by

$$\partial_i F(t, i-p+a_t(i)) + \frac{1}{c}(p - 2a_t(i)) = 0, \tag{21}$$

$$\partial_i F(t, i+b_t(i)-q) + \frac{1}{c}(q - 2b_t(i)) = 0. \tag{22}$$

Equations (21) and (22) capture the key aspects of the HFT's dynamic optimization problem. If there were no end of day inventory motives, i.e. no overnight funding costs λi^2 and no end-of-day inventory value $\bar{p}i$, the HFT would simply solve a myopic optimization problem. To do so, the HFT would set the (scaled) ask price a so as to equate the marginal benefit of raising a, given by $\frac{1}{c}(p - 2a_t(i))$, to zero. The benefit arises from higher profits earned by the HFT when it sells. The solution to this static problem would be to simply set $a = \frac{1}{2}c\tilde{p}$, where we have used that $p = c\tilde{p}$. Hence, the ask price would be proportional to the upper reservation price $\tilde{p}$ with proportionality constant equal to the demand slope c. Similarly, the HFT would set the (scaled) bid price b so as to equate the marginal benefit of raising b, $-\frac{1}{c}(q - 2b_t(i))$, to 0. This yields $b = \frac{1}{2}c\tilde{q}$.

The distinguishing feature of our model is the end-of-day inventory motive, which has implications on the intraday pricing behavior. This is reflected in the terms $\partial_i F(t, i-p+a_t(i))$ and $\partial_i F(t, i+b_t(i)-q)$ appearing in equations (21) and (22). These terms represent the instantaneous cost of holding inventory in the dynamic programming problem. These derivative terms therefore drive a wedge between the myopic and the forward looking dynamic optimization problem. The wedge is graphically illustrated in Figure 2. The figure compares the ask price in the static, myopic, problem with the corresponding price in dynamic problem. Setting a higher ask a, relative to the myopic case,[3] impacts the shadow value of the end-of-day

[3] In fact, if setting ask $\tilde{a} = \frac{\tilde{p}}{2}$ and bid $\tilde{b} = \frac{\tilde{q}}{2}$, the HFT will mostly likely end up with a large negative inventory, because sellers will not supply any inventory to it at a price lower than $\tilde{q}$.

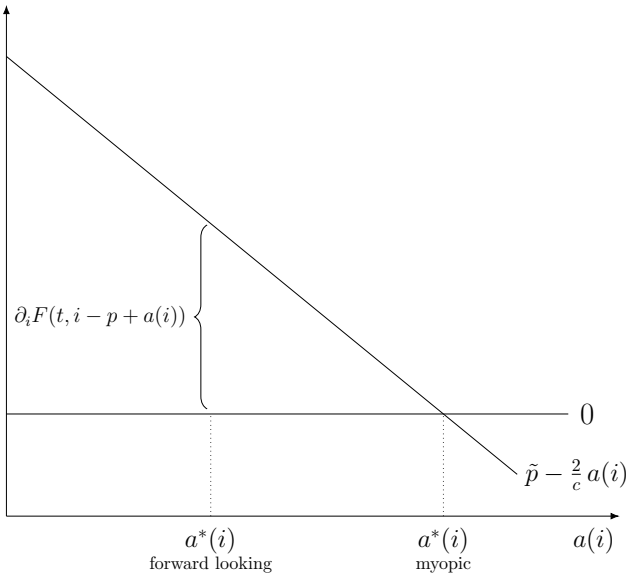


Figure 2: **Impact of End-of Day Inventory Motives on HFT Intraday First-Order Conditions.** End-of-day considerations drive a wedge between the marginal benefit of a trade occurring at time t versus the marginal utility at the post-trade inventory level for the remaining period of the day. In the absence of overnight inventory costs and end-of-day mark-to-market gains, the HFT solves a myopic decision problem and chooses the ask price $a^*(i)$ so that the marginal cost of increasing the ask, $\tilde{p} - \frac{2}{c}\, a(i)$, equals zero. However, because ask prices also affect the present value of overnight inventory costs and end-of-day mark-to-market gains through $\partial_i F(t, i-p+a(i))$, the choice of $a^*(i)$ in the forward looking dynamic optimization problem differs from the corresponding choice in a myopic decision problem.

inventory motives. An analogous wedge arises for the bid. The $\partial_i F$ terms play therefore a crucial role in determining how end-of-day inventory motives impact intraday bid and ask quoting decisions of the HFT.

We can quantify the sensitivity of $F(t,i)$ with respect to inventory levels. When i is negative and very small, $\partial_i F(t,i) > 0$; when i is positive and very large, $\partial_i F(t,i) < 0$; and for intermediate levels of inventory i (either negative or positive), $\partial_i F(t,i)$ is strictly decreasing. Hence, the wedge between the myopic and the dynamic optimization problem will be such that both the ask price a and the bid price b are decreasing in i. Even though bid and ask prices cannot be explicitly written as a function of the inventory level, we can characterize the properties of the solution to the first-order conditions (21) and (22) using the above discussed properties of $F(t,i)$. We will also present a recursive algorithm that allows us to solve for a and b numerically.

Lemma 5.1. *Fix any $t = 1, 2, \ldots, T$, we have*

$$a_t(i) = G_t^{-1}\left(\frac{p-2i}{c}\right) - i + p, \tag{23}$$

$$b_t(i) = G_t^{-1}\left(\frac{q-2i}{c}\right) - i + q, \tag{24}$$

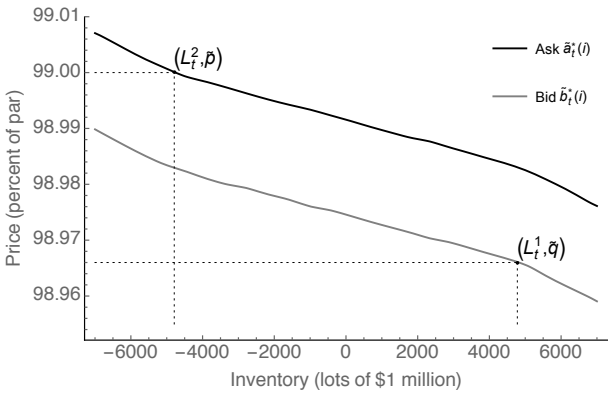


Figure 3: **The Optimal Price Policy Functions.** The optimal policy functions (of the current inventory level i) for bid and ask prices, $\tilde{b}_t^*(i)$ and $\tilde{a}_t^*(i)$, at a fixed time t. We take $T-t$ to be equal to one-thousandth of a time step. When the inventory is low (i.e. $i \leq L_t^2$), the ask price is higher than $\tilde{p}$, so that buyers do not trade with the HFT, but sellers trade with it and sell $Q^{SO}(\tilde{a}_t^*(i))$ shares in each trade (see Figure 1). When the inventory is high (i.e. $i \geq L_t^1$), the bid price is higher than $\tilde{q}$, so that sellers do not trade with the HFT, but buyers trade with it and purchase $Q^{BO}(\tilde{b}_t^*(i))$ shares in each trade. When the inventory is in the active trading region (i.e. $L_t^2 < i < L_t^1$), both ask and bid prices are between $\tilde{q}$ and $\tilde{p}$, and the HFT can trade with both counter-parties and earn a positive bid-ask spread. Moreover, for these moderate inventory levels, both the ask and bid price functions are roughly linear in the inventory level, hence their slope can be measured by the reciprocal of the width of the active trading region, $L_t^1 - L_t^2$. A detailed analysis of the inventory boundaries is given in the remainder of the section.

where G_t^{-1} is the i-inverse function of a strictly decreasing function defined as

$$G_t(i) := \partial_i F(t,i) - \frac{2}{c} i.$$

The mappings $i :\mapsto a_t(i)$ and $i :\mapsto b_t(i)$ are all strictly decreasing, continuous, and mapping onto $\mathbb{R}$. Moreover, for all $\lambda > 0$ we have

$$\frac{1}{2}(p-q) < a_t(i) - b_t(i) < p - q. \tag{25}$$

Lemma 5.1 implies a number of features for the candidate ask and bid quotes given in (18) and (19). First and the foremost, it shows that both $a_t^*(i)$ and $b_t^*(i)$ are *continuous, non-increasing functions* of the inventory level at time $t-1$. Intuitively, as the inventory gets larger, the HFT would like to offload inventory so as to reduce the penalty incurred for a large inventory position. To that end, the HFT wants to sell a larger number of shares to the buyer, which can be facilitated by setting a low ask $a_t^*(i)$. At the same time, it wants to reduce the bid so that the sell investor is only willing to supply it a small number of shares (or none) and its inventory thus does not increase much. Secondly, when the reservation prices of the buy and sell investor, $\tilde{p}$ and $\tilde{q}$, become closer, the bid-ask spread will shrink accordingly.

The candidates $a_t^*(i)$ and $b_t^*(i)$ are indeed the optimal solution for the problem (17), as formalized in the next lemma.

Lemma 5.2. *The optimal ask and bid prices are given by* $\tilde{a}_t^* := a_t^*(i)/c$ *and* $\tilde{b}_t^*(i) := b_t^*(i)/c$*, where* $a_t^*(i)$ *and* $b_t^*(i)$ *are given in* (18) *and* (19) *(see Figure 3).*

5.3 Intertemporal Analysis of Optimal Price Policies

This section investigates the dynamic behavior of the optimal ask and bid price as time moves towards the end of the day. We know from Lemma 5.1 that both the optimal ask price and the optimal bid price depend on the HFT's inventory level. Next, we want to identify the critical inventory thresholds, i.e. the levels at which the HFT decides to post ask and bid prices equal to $\tilde{p}$ and $\tilde{q}$ respectively, so as to shut down trades with buy and sell investors. Specifically, we define the critical inventory boundaries L_t^1 and L_t^2 as the unique solutions to the following equations:

$$b_t^*(L_t^1) = q, \quad a_t^*(L_t^2) = p. \tag{26}$$

Equivalently, using the system of first-order conditions given by equations (21) and (22), we obtain (recall that $p = c\,\tilde{p}$ and $q = c\,\tilde{q}$)

$$\begin{cases} \partial_i F(t, L_t^1) - \tilde{q} = 0, \\ \partial_i F(t, L_t^2) - \tilde{p} = 0. \end{cases} \tag{27}$$

We then have that the optimal bid price $\tilde{b}_t^*(i)$ is always lower than or equal to $\tilde{q}$ when the inventory level $i \geq L_t^1$. Because $Q^{SO}(x) = 0$ for all $x \leq \tilde{q}$, we deduce that the HFT only trades with buyers, i.e. only sells, when its inventory is higher than the critical level L_t^1. We henceforth refer to all inventory levels i larger than L_t^1 as the *sell only region*. When the HFT's inventory is in the sell only region, the HFT's main objective is to unload its inventory as quickly as possible. To that end, the HFT sets a low ask price to encourage trading with the buyers, while essentially shutting down bidding by setting the bid lower than or equal to $\tilde{q}$, the reservation price for sellers. This behavior is consistent with actual inventory risk management strategies of HFTs in practice, who have been known to suspend trading if an undesirable threshold inventory level is reached.[4]

Likewise, the optimal ask price $\tilde{a}_t^*(i)$ is always higher than or equal to $\tilde{p}$ when the inventory level $i \leq L_t^2$. Because $Q^{BO}(x) = 0$ for all $x \geq \tilde{p}$, we deduce that the HFT only trades with sellers, i.e. only buys, when its inventory is lower than the critical level L_t^2. We henceforth refer to all i smaller than L_t^2 as the *buy only region*. In the buy only region, the HFT only trades with sell investors to build up inventory by setting the ask price higher than $\tilde{p}$ and the bid price to a high level.[5] In both the sell only region and the buy

[4] This behavior is described, for example, in Aït-Sahalia and Saglam (2016), who claim that in the attempt of limiting the size of the inventory for risk mitigation purposes, the HFT does not necessary quote on both sides of the market.

[5] One-sided trades also arise in Aït-Sahalia and Saglam (2016), in which a monopolistic HFT with a positive inventory may

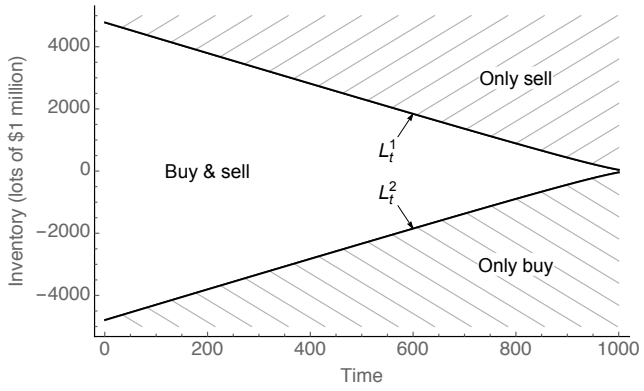


Figure 4: **The Critical Inventory Thresholds.** When the HFT's accumulated inventory crosses the thresholds L_t^i, $i = 1, 2$, the HFT's trading activity will change qualitatively. Inventories are measured in lots of \$1 million. When the HFT's inventory level is between L_t^1 and L_t^2, it actively trades with both counter-parties. If the HFT's inventory level is higher than L_t^1, then it only trades with buyers to unload its inventory. On the other hand, if the HFT's inventory level is lower than L_t^2, then it only trades with sellers to build up its inventory. As time passes, the active trading region, defined as the inventory levels at which the HFT both buys and sells, shrinks as the impact of the end-of-day inventory constraint materializes. This means that the HFT will make more efforts to keep its inventory inside the active trading region, in order to avoid any one-side trade near the day's end.

only region, trading activity degenerates to one-sided trades, thus the bid-ask spread no longer acts as an appropriate measurement of liquidity under these conditions.

The one-sided trading activity is also present in the model of Amihud and Mendelson (1980), in which the dealer buys only when the inventory reaches the lower extreme of the admissible inventory interval and sells only when the inventory reaches the upper extreme of such an interval. This differs from our setup because in which the active trade region shrinks endogenously with time (see Proposition 5.4 below) and is sensitive to the severity of the end-of-day constraint. In Amihud and Mendelson (1980), by contrast, the admissible inventory interval is exogenously fixed beforehand.

Recall that Lemma 5.1 asserts that the scaled bid-ask spread $a_t^*(i) - b_t^*(i)$ stays inside the open interval $(\frac{p-q}{2}, p-q)$ for all i, and thus $a_t(L_t^1) < b_t(L_t^1) + p - q = p = a_t(L_t^2)$, so that the boundaries L_t^1, L_t^2 satisfy

$$L_t^1 > L_t^2.$$

When the HFT's inventory level i is between L_t^2 and L_t^1, it actively trades with both buyers and sellers. We refer to this range of inventories i as the *active trading region* (see Figure 4).

The next proposition, proven in the appendix, shows that the value function $F(t, i)$ has certain time-invariant properties.

stop quoting on one side of the market because of inventory aversion. This may occur if their equilibrium condition, requiring the arrival probability of buyers and sellers to be the same, is violated so so that the HFT may accumulate inventory over time.

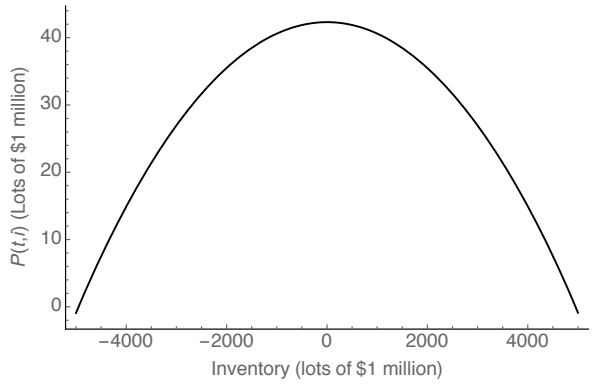


Figure 5: **The Optimal Present Values of Inventory Constraint.** The value function $F(t,i)$ measures the optimal expected utility of the HFT at T, as seen from time t, and given that the level of inventory at t is i. Suppose the HFT can cash out its inventory immediately in a secondary market at a price $\bar{p}$ per share, then the residual $P(t,i) = F(t,i) - \bar{p}i$ gives the optimal present value of the inventory constraint. In particular, at time T, $P(T,i) = -\lambda i^2$ gives the end-of-day inventory cost. This figure plots $P(t,i)$ for $t = 1$, as a function of i. The concavity of $P(1,i)$ in i is smaller than that of $P(T,i)$, -2λ, (see Proposition 5.5 for more details), because the impact of the end-of-day inventory constraint fades away as the time remaining till the day's end increases.

Proposition 5.3. *For any $t = 1, 2, \ldots, T$, $F(t-1,i)$ is strictly i-concave, continuously differentiable, and with a i-derivative mapping onto $\mathbb{R}$ which admits the following recursive representation*

$$\begin{aligned}
\partial_i F(t-1,i) &= \mathbb{E}\left[\partial_i F(t, I_t^{(\tilde{a}^*,\tilde{b}^*)}) | I_{t-1}^{(\tilde{a}^*,\tilde{b}^*)} = i\right] \\
&= \begin{cases}
(1-\pi^{BO})\partial_i F(t,i) + \pi^{BO}\partial_i F(t, i-p+a_t^*(i)), & i \geq L_t^1 \\
(1-\pi^{BO}-\pi^{SO})\partial_i F(t,i) + \pi^{BO}\partial_i F(t, i-p+a_t^*(i)) + \pi^{SO}\partial_i F(t, i+b_t^*(i)-q), & L_t^1 > i > L_t^2 \\
(1-\pi^{SO})\partial_i F(t,i) + \pi^{SO}\partial_i F(t, i+b_t^*(i)-q), & L_t^2 \geq i
\end{cases}
\end{aligned} \tag{28}$$

where $a_t^*(i)$ and $b_t^*(i)$ be defined as in equation (18) and (19), respectively.

Equation (28) is intuitive. First, the marginal utility at $t-1$, $\partial_i F(t-1,i)$, is the conditional expectation of the time t marginal utility $\partial_i F(t, I_t^{(\tilde{a}^*,\tilde{b}^*)})$, assuming that trading occurs at the optimal ask price $\tilde{a}_t^*(i)$ and at the optimal bid price $\tilde{b}_t^*(i)$. Recall that $a_t^*(i)$ and $b_t^*(i)$ are stationary points of (16), so that we can evaluate the derivative of $F(t-1,i)$ with respect to i at that stationary point. Depending on the level of inventory, the expected marginal utility at t is determined by one-sided trades (only with sellers if the inventory is in the buy-only region and only with buyers if the inventory is in the sell-only region) or by two-sided trades if the inventory is in the active trading region. Moreover, consider the time t optimal expected utility of the

HFT in (16), net of inventory holdings and valued at the equilibrium price, i.e.

$$P(t,i) := F(t,i) - \bar{p}i. \tag{29}$$

We can interpret $P(t,i)$ as the *optimal present value of the inventory constraint* (see Figure 5). Then $\partial_i P(t,i)$ gives the marginal gain for the HFT stemming from an infinitesimal change in inventory. By the i-concavity of $F(t,i)$, we know that $\partial_i P(t,i)$ is continuous, and strictly decreasing. Because $\bar{p} \in (\tilde{q}, \tilde{p})$, we know from (27) that $\partial_i P(t,i) > 0$ if $i < L_t^2$, and $\partial_i P(t,i) < 0$ if $i > L_t^1$. Hence, as argued above, if the inventory is very large ($i > L_t^1$), it will be beneficial to reduce the inventory level, while if the inventory is very low ($i < L_t^2$), it will be beneficial to increase it. Furthermore, because

$$\partial_i P(t-1,i) - \partial_i P(t,i) = \partial_i F(t-1,i) - \partial_i F(t,i),$$

we know from (28) that if the current level of inventory is large, say, $i > \max\{L_t^1, L_{t-1}^1\}$, we have $\partial_i P(t,i) < \partial_i P(t-1,i) < 0$. That is, the sensitivity of the present value of the inventory constraint with respect to the inventory level becomes smaller as the time remaining to the day's end increases. In other words, the marginal benefit of reducing the inventory materializes as time progresses. Likewise, if the current inventory level is low, say, $i < \min\{L_t^2, L_{t-1}^2\}$, we have $\partial_i P(t,i) > \partial_i P(t-1,i) > 0$. Hence, this sensitivity also gets lower as the time remaining until the day's end increases. Consequently, the HFT exhibits different trading behaviors as time progresses. To see this, let us look again at Figure 4 and observe a very interesting feature of optimal market making: the size of the active trading region increases as the time remaining until the day's end increases. Intuitively, this is because the shadow cost of the end of day inventory constraint is lower given that the HFT has more time to build or offload the inventory, i.e. to execute multiple round-trip trades, before the day closes. As time approaches the end of the trading day, the HFT may need to shut down trades either with sell investors if it has an excessively large positive inventory or with buy investors if it faces a short inventory position of large size. For this reason, we observe that both the sell only region (L_t^1, ∞) and the buy only region $(-\infty, L_t^2)$ "grow" as time approaches the day's end. We formalize these statements in the following result.

Proposition 5.4. *The sequence $(L_t^1)_{t=1}^T$ is positive, and strictly decreasing, while the sequences $(L_t^2)_{t=1}^T$ is negative, and strictly increasing. In particular, $L_T^1 = \frac{\bar{p}-\tilde{q}}{2\lambda}$ and $L_T^2 = -\frac{\tilde{p}-\bar{p}}{2\lambda}$ (see Figure 4).*

The larger the active trading region, the more aggressively the HFT can trade as it is less concerned about the inventory constraint. By (21) and (22), if the inventory level $i \in [L_T^2, L_T^1]$, we have the end of day

or time-T optimal price policy functions

$$\tilde{a}_T^*(i) = \frac{1}{1+\lambda c}\left(\tilde{p}(\frac{1}{2}+\lambda c) + \frac{\bar{p}}{2} - \lambda i\right)^+, \tag{30}$$

$$\tilde{b}_T^*(i) = \frac{1}{1+\lambda c}\left(\tilde{q}(\frac{1}{2}+\lambda c) + \frac{\bar{p}}{2} - \lambda i\right)^+. \tag{31}$$

Hence, both the ask and the bid prices are linear in the inventory level with slopes $-\frac{\lambda}{1+\lambda c}$ at day's end. When we are sufficiently far from the day's end, the dependence of the optimal bid and ask profiles on the inventory level is roughly linear, with the "slope" being inversely proportional to the size of the active trading region, $L_t^1 - L_t^2$ (see Figure 3 for the illustration). By Proposition 5.4, this region gets larger as more time is left until the day's end (see also Figure 4). This implies that the sensitivity of ask and bid prices on the inventory level, measured by the "slope" of the corresponding price policy functions, becomes weaker as the time remaining until the day's end increases. Moreover, as long as both the ask and bid prices are positive, the bid-ask spread at time T is given by

$$\tilde{a}_T^*(i) - \tilde{b}_T^*(i) = \frac{\frac{1}{2}+\lambda c}{1+\lambda c}(\tilde{p}-\tilde{q}) =: B(\lambda), \tag{32}$$

which is independent of the inventory i. For convenience, we have defined it as a function of λ, and denoted it by $B(\lambda)$. When t is far away from T, the "slope" of the price policy functions is smaller in absolute value, and hence we expect that the bid-ask spread is also smaller earlier in the day. We will explore these phenomena in depth in the next subsection.

To understand the impact of a severe inventory constraint, we use the explicit formulas in (30), (31) and (32) to study the ask and bid price policy functions, as well as the bid-ask spread at time T. As the inventory constraint tightens, i.e. λ gets larger, both $\tilde{a}_T^*(i)$ and $\tilde{b}_T^*(i)$ become more sensitive to the inventory level i. In the limiting case $\lambda \to \infty$, $\tilde{a}_T^*(i) \to (\tilde{p} - \frac{1}{c}i)^+$ and $\tilde{b}_T^*(i) \to (\tilde{q} - \frac{1}{c}i)^+$, and the bid-ask spread tends to its maximum value, $\tilde{p}-\tilde{q}$, if both bid and ask prices are strictly positive. Hence, when the inventory constraint is very severe, the HFT will only perform one-side trades, either with buy or sell investors. Specifically, if the inventory i at time $T-1$ is positive, then $\tilde{a}_T^*(i) < \tilde{p}$ and $\tilde{b}_T^*(i) < \tilde{q}$. Given the demand function (1) and the supply function (2), the HFT only trades with buy investors to reduce inventory. If the inventory i at time $T-1$ is negative, then $\tilde{a}_T^*(i) > \tilde{p}$ and $\tilde{b}_T^*(i) > \tilde{q}$, and the HFT only trades with sell investors to increase inventory. If $i = 0$, then $\tilde{a}_T^*(i) = \tilde{p}$ and $\tilde{b}_T^*(i) = \tilde{q}$, i.e. the HFT shuts down trading with both parties at time $T-1$ so as to maintain its flat inventory level. This behavior is also consistent with the result in Proposition 5.4, because it can be directly seen that both L_T^1 and L_T^2 converge to 0, as $\lambda \to \infty$.

5.4 Endogenous Price Impact and Widening Bid-Ask Spread

This section highlights the mechanism through which price impact arises endogenously in our model. It also studies the dependence of the bid-ask spread on the severity of the overnight inventory constraint and the passage of time. We start with the following result which provides bounds for the bid-ask spread, and also for its sensitivity to changes in inventory levels.

Proposition 5.5. *Let $(\lambda_t^1)_{t=1}^T$ and $(\lambda_t^2)_{t=1}^T$ be sequences of positive numbers, defined recursively according to*

$$\begin{cases} \lambda_{t-1}^1 = \lambda_t^1\left(1 - \min\{\pi^{BO}, \pi^{SO}\}\dfrac{\lambda_t^1 c}{1+\lambda_t^1 c}\right), & t = 2,3,\dots,T, \\ \lambda_{t-1}^2 = \lambda_t^2\left(1 - (\pi^{BO}+\pi^{SO})\dfrac{\lambda_t^2 c}{1+\lambda_t^2 c}\right), & t = 2,3,\dots,T, \\ \lambda_T^1 = \lambda_T^2 = \lambda. \end{cases} \tag{33}$$

Suppose that $\tilde{a}_T(L_{t_0}^1) \geq 0$ for some $t_0 = 1,2,\dots,T$,[6] then for any $t = t_0, t_0+1,\dots,T$, the optimal price policy functions satisfy

$$-\frac{\lambda_t^1}{1+\lambda_t^1 c} \leq \frac{\tilde{a}_t^*(i_1) - \tilde{a}_t^*(i_2)}{i_1 - i_2}, \frac{\tilde{b}_t^*(i_1) - \tilde{b}_t^*(i_2)}{i_1 - i_2} \leq -\frac{\lambda_t^2}{1+\lambda_t^2 c}, \quad \text{for any } L_t^1 \geq i_1 > i_2 \geq L_t^2, \tag{34}$$

and the bid-ask spread satisfies

$$B(\lambda_t^2) \leq \tilde{a}_t^*(i) - \tilde{b}_t^*(i) \leq B(\lambda_t^1), \quad \text{for any } L_t^1 \geq i \geq L_t^2, \tag{35}$$

where the function $B(\cdot)$ is defined in (32). Moreover, for any $t = t_0, t_0+1,\dots,T$, we have

$$-2\lambda_t^1 \leq \frac{\partial_i F(t,i_1) - \partial_i F(t,i_2)}{i_1 - i_2} \leq -2\lambda_t^2, \quad \text{for all } L_{t_0^1} \geq i_1 > i_2. \tag{36}$$

Proposition 5.5 gives a range for the "slopes" of the optimal ask and bid price policy functions with respect to the inventory level, and for the bid-ask spread, in terms of two positive sequences $(\lambda_t^1)_{t=1}^T$ and $(\lambda_t^2)_{t=1}^T$. These two sequences also provide bounds for the i-concavity of the value function $F(t,i)$ (see Eq. (36)). Differently from the day's end, in which the i-concavity of $F(T,i)$ is measured by the severity of the overnight inventory cost λ, the concavity of $F(t,i)$ depends on the optimal trading behavior of the market maker, which is influenced by time and inventory level. More specifically, the trading activities of the HFT reduce the concavity of the value function, which mainly comes from the overnight inventory cost. When

[6] This means that we do not consider extreme inventory levels at which the HFT is even willing to pay a price to buyers to dump its inventory (i.e. the condition can be dropped if the HFT is allowed to set a negative ask price when necessary). Because $(L_t^1)_{t=1}^T$ is decreasing in t, this implies that the inequality holds for $t, t+1,\dots,T$. Under the set of parameters used in Figure 4, the condition is satisfied for all $t = 1,2,\dots T$.

the HFT can trade with both counter-parties, the probability of a trade is $\pi^{BO} + \pi^{SO}$, and the concavity is reduced the most (see (33), (28), and Figure 4).

Both sequences $(\lambda^1_{T-t+1})_{t=1}^T$ and $(\lambda^2_{T-t+1})_{t=1}^T$ are strictly decreasing, but at a pace much slower than an exponential rate,[7] since the percentage decay rate decreases as the sequence gets smaller. Because $\frac{\lambda}{1+\lambda c}$ and $B(\lambda)$ are increasing in λ for $\lambda > 0$, we deduce from Proposition 5.5 that both the sensitivity of the optimal ask and bid price policy functions to the inventory level, as well as the bid-ask spread, are lower when the time-to-close $T-t$ increases. This is consistent with the dynamic behavior of the critical inventory thresholds reported in Figure 4. Moreover, from Eq. (33), we deduce that the speed of the time decay of the sequences $(\lambda^1_{T-t+1})_{t=1}^T$ and $(\lambda^2_{T-t+1})_{t=1}^T$ depends on the arrival probabilities π^{BO} and π^{SO}. The larger π^{BO} and π^{SO} are, the faster these two sequences decrease.

The economic intuition behind the mechanism described above is as follows: as the market becomes more liquid, i.e. buy and sell orders arrive more frequently, the bid-ask spread at a fixed time before day's end declines. Recall that the overnight inventory cost λ only makes the bid-ask spread at T larger, thus a more liquid market makes the inventory constraint fade away faster as the time until the day's end increases. Moreover, as time approaches the end of the day, the growing concern about the inventory constraint discourages the HFT from trading actively. Hence, it sets a larger spread to reduce the quantity traded in each time step (see Figure 6), but at the same time, the per-unit trading profit increases from each buy-and-sell roundtrip. Such a trading behavior reflects the tradeoff faced by the HFT between making trading profits and holding a non-zero inventory at the end of day. This prediction of our model is well-reflected in the treasury trading data we study in the forthcoming section (see Figure 11).

As discussed above, the relationship between the optimal ask price/bid price and the inventory level is neither linear nor time-homogenous. Nonetheless, for a short period of time, we can gain insights into the short-term dynamics of bid and ask prices by simplifying the optimal price policy functions, making them depend linearly on inventory levels and ignoring the time-dependence of the function's coefficients. We know that a linear relation between prices and inventory levels only holds at the day's end. If we are sufficiently far from the day's end, then we can still use the lower and upper bounds for the slope of ask and bid price policy functions given in Eq. (34). These two bounds coincide at the day's end. As argued in the discussion after Proposition 5.5 (see in particular the corresponding footnote), the lower and upper bounds behave approximately as the reciprocal of time to maturity and thus are still of the same order. This implies an approximate linear relation between prices and inventories locally (in time and inventory levels).

[7] In fact, it can be proved using standard analysis that, both sequences are decreasing polynomially, or more precisely, the reciprocal sequences $(1/\lambda^1_{T-t+1})_{t=1}^\infty$ and $(1/\lambda^2_{T-t+1})_{t=1}^\infty$ increase linearly as $t \to \infty$. This is in line with Figure 4, where we have observed that both $(L^1_{T-t+1})_{t=1}^T$ and $(L^2_{T-t+1})_{t=1}^T$ are roughly linear in t.

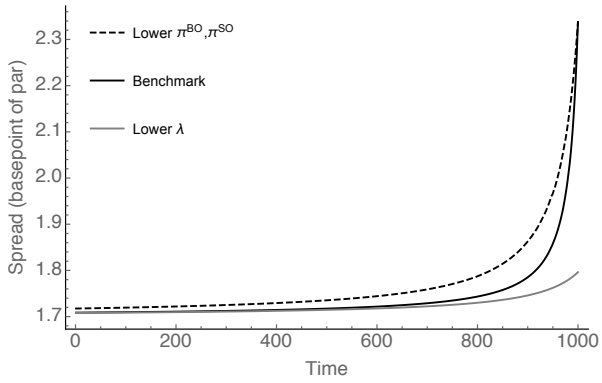


Figure 6: **Optimal Bid-Ask Spread at Zero Inventory Level.** This figure plots the optimal bid-ask spread when the inventory level is zero, $\tilde{a}_t^*(0) - \tilde{b}_t^*(0)$, using the demand and supply functions specified in Figure 1 and for three different triplets $(\lambda, \pi^{BO}, \pi^{SO})$. Define $\lambda_0 := 0.02$ per lot of \$100 million and $\pi_0 := 10\%$. (a) Benchmark case (solid black): $\lambda = \lambda_0$ and $\pi^{BO} = \pi^{SO} = \pi_0$; (b) Lower π^{BO}, π^{SO} case (dashed black): $\lambda = \lambda_0$ and $\pi^{BO} = \pi^{SO} = 0.5\pi_0$; (c) Lower λ case (solid gray): $\lambda = 0.1\lambda_0$ and $\pi^{BO} = \pi^{SO} = \pi_0$. The higher are the arrival probabilities (i.e. larger π^{BO} and π^{SO}), the smaller is the bid-ask spread. For fixed arrival probabilities of market orders, the more severe is the end-of-day inventory constraint (i.e. larger λ), the wider is the bid-ask spread. When the severity of the end-of-day inventory constraint is low (i.e. very small λ), the bid-ask spread is nearly flat throughout the whole trading period. As the time left till the end of the day increases, the bid-ask spread approaches $(\tilde{p} - \tilde{q})/2$ (the spread when there are no end-of-day inventory constraints) under all parameter settings.

In particular, for the optimal ask price, we have:

$$\tilde{a}_t^* = \beta_0 - \beta_1 I_{t-1},$$

where β_0 and β_1 are constants and $\beta_1 > 0$.[8] Taking the difference of the above equation at two consecutive time points, we have

$$\tilde{a}_{t+1}^* - \tilde{a}_t^* = -\beta_1 (I_t - I_{t-1}) = -\beta_1 \left(Q^{SO}(\tilde{b}_t^*) \Delta N_t^{SO} - Q^{BO}(\tilde{a}_t^*) \Delta N_t^{BO} \right).$$

Thus, if a buy order is executed at time t (i.e. a buyer arrives at t and $Q^{BO}(\tilde{a}_t^*) > 0$), then the ask price will increase by $\beta_1 Q^{BO}(\tilde{a}_t^*)$, which is proportional to the size of the trade taking place at time t, $Q^{BO}(\tilde{a}_t^*)$. Likewise, if there is a sell order coming at time t, then the ask price will decrease by $\beta_1 Q^{SO}(\tilde{b}_t^*)$, in an effort to invite a larger sized trade with a buyer in the next time step so as to balance the inventory. Hence, price impact arises endogenously in our model. Furthermore, in conjunction with Proposition 5.5, we also deduce that the price impact coefficient, β_1, tends to be larger as time moves towards the day's end, which is indeed well-documented in our empirical analysis (see Section 6).[9]

[8] A similar argument can be made for the bid price.

[9] Admati and Pfleiderer (1988) argue that traders prefer to trade when the market is "thick", or, when the price impact of their trades is small. In our model, buyers and sellers arrive according to Bernoulli processes with fixed intensities. Nonetheless,

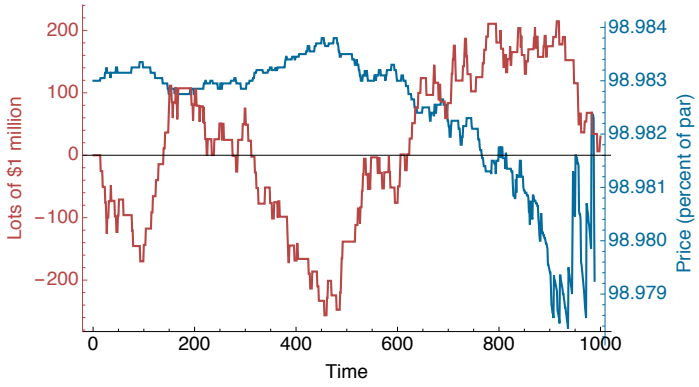


Figure 7: **Simulated Inventory and Midquote Paths.** A simulated inventory path and the corresponding midquote price process consisting of 1000 time steps. We have used the demand and supply functions given in Figure 1, arrival probabilities $\pi^{BO} = \pi^{SO} = 10\%$, the end-of-day inventory cost $\lambda = 0.02$ per lot of \$100 million, and a zero initial inventory. Based on a simulated sequence of arrivals of buy and sell orders, the optimally controlled inventory of the HFT (shown in red), starts at 0 and ends at around 0. The dark blue line illustrates the midquote price trajectory. As shown in earlier sections, the price trajectory is *negatively* correlated with that of the HFT's inventory level, and this dependence increases as time approaches the day's end. Moreover, during the course of the day, the HFT's inventory has experienced multiple round trips (excursions) over and below 0.

If a much larger number of buyers, relative to sellers, arrives during a short period of time, then the price impact generated from trades with buy investors will quickly push ask and bid prices upward, hence resulting in a *flash rally*. Likewise, if many more sellers arrive relative to buyers, the price impact resulting from the asset sales will quickly drive the price down, hence resulting in a *flash crash*.

The time varying and endogenous nature of the price-impact and bid-ask spreads are distinguishing features of our model and are generated by the forward looking nature of the inventory constraint. By contrast, in the model of Amihud and Mendelson (1980), the price policy functions are time homogeneous because both bid and ask are independent of time due to the perpetual nature of the dealer's decision making problem. In the model proposed by Aït-Sahalia and Saglam (2016), both prices and spreads are independent of inventory levels and time, again due to the perpetual setting of their problem. In both of these models, a unit trading size is considered, hence price impact is not endogenous.

5.5 The Zero Inventory Constraint

We consider the limiting case of our model in which the HFT does not have any end-of-day constraint. If $\lambda = 0$, then we have $F(T, i) = \bar{p}i$, and can recovery the optimal ask and bid price policy function at time T

our model predictions can be reconciled with those of Admati and Pfleiderer (1988). This is because both the price impact and the bid-ask spread increase as time approaches the day's end, and under these market conditions buyers and sellers would arrive less often according to Admati and Pfleiderer (1988).

by taking limits of (30) and (31) as $\lambda \to 0$:

$$\tilde{a}^*_T(i) = \frac{1}{2}(\tilde{p} + \bar{p}), \tag{37}$$

$$\tilde{b}^*_T(i) = \frac{1}{2}(\tilde{q} + \bar{p}). \tag{38}$$

Clearly, $\tilde{q} < \tilde{b}^*_T(i) < \tilde{a}^*_T(i) < \tilde{p}$ for all i, and the bid-ask spread at time T is given by $B(0) = \frac{1}{2}(\tilde{p} - \tilde{q})$. Furthermore, using mathematical induction, we can show that $\partial_i F(t, i) = \bar{p}$ for all $t = 1, \ldots, T$. Hence, (37) and (38) give the optimal ask and bid price policy functions for all t. Therefore, in the absence of the end-of-day constraint, prices are independent of inventory levels (unless the buyer/seller's preference, i.e. the supply/demand curve changes) throughout the whole trading period, and consequently there will be no price impact. These strikingly different qualitative features highlight the significant role played by the end-of-day inventory constraint in determining the intraday price dynamics, which will be tested via an empirical analysis in the next section.

6 Empirical Analysis and Testable Implications

The U.S. Treasury market is one of the most liquid and largest security market globally, and HFTs represent a significant proportion of trading activity in the Treasury market (see Joint Staff Report (2015)). This suggests that the Treasury market provides an appropriate testing ground for our model. We examine the following *Testable Implications* from our model:

(TI-1) There is a significant positive relationship between both bid and ask prices (and hence midquotes) and the negative of the HFT's inventory (see Lemma 5.1 and Lemma 5.2). Thus, bid and ask prices negatively co-move with the HFT's inventory.

(TI-2) The dependence of bid and ask prices on HFT's inventory becomes stronger as time approaches the day's end (see Proposition 5.5). Because of the HFT's inventory management motives, each trade of the HFT generates price impact. Moreover, the price impact is the largest at the day's end (see Proposition 5.5 and discussions following it).

(TI-3) Due to the endogenous price impact, flash events can occur if significant trading in one direction is followed by significant trading in the opposite direction during a short time interval (this follows from Lemma 5.1, and Lemma 5.2 applied to short time intervals).

(TI-4) The bid-ask spread tends to increase as time approaches the day's end (see Proposition 5.5).

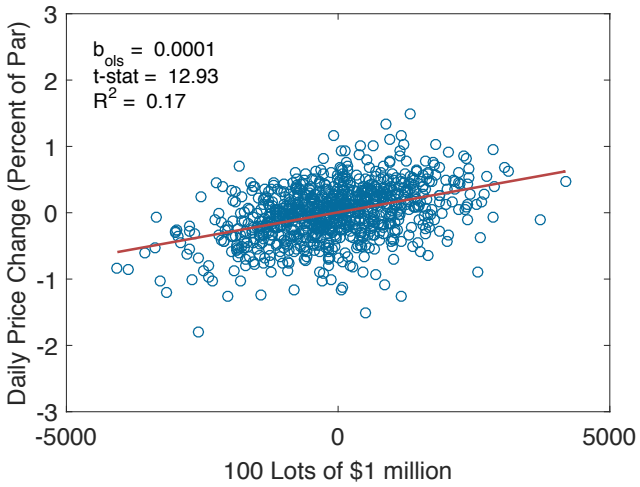


Figure 8: **10-Year Treasury Prices and Cumulative Net Volume.** This figure scatters daily cumulative net volume against daily 10-year Treasury midnight-to-close price changes. Cumulative net volume is measured in 100 lots of $1 million and represent the cumulative sum of intraday tick-by-tick net dollar volumes on BrokerTecfrom midnight until close at 17:30. The red line is the OLS projection of daily price changes on cumulative net volume, with slope coefficient, Newey-West t-statistic, and R-squared reported in the top left corner. The daily correlation between cumulative net volumes and returns is 41%. The sample consists of daily observations from 4/2/2012 to 10/30/2015.

We examine each of these Testable Implications (TI-1 through TI-4) in turn, using high-frequency intraday data from BrokerTec, a major U.S. Treasury ECN (electronic communications network) that accounts for about 60% of electronic trading activity in the cash market for U.S. Treasury securities. The availability of detailed trade and limit order book data time-stamped to the millisecond make this data ideally suited to examine intraday trading and liquidity patterns.[10]

To test (TI-1), we construct a proxy for HFT inventory as the negative cumulative net volume (buy minus sell). As Brogaard et al. (2014) point out, "if HFTs' inventory positions are close to zero overnight, then their inventories can be measured by accumulating their buying and selling activity in each stock from opening to each point in time." While our data does not provide participant-level information, in markets where a large share of transactions are through an intermediate HFT (as the Joint Staff Report (2015) suggests), the cumulative net volume can be considered as a proxy for the negative of the HFT's inventory, because each buy order reduces the inventory of the HFT by the same amount, and each sell order increases the inventory of the HFT by an equal amount. With this measure in hand, (TI-1) implies that bid and ask prices (and hence midquotes) should be negatively correlated with HFT inventory and hence positively correlated with cumulative net volume, which is indeed what we find. Figure 8 shows that a regression of price changes on cumulative net volume changes yields a statistically strong relationship between this inventory measure and

[10] A detailed overview of the microstructure of the BrokerTec ECN is provided by Fleming et al. (2014).

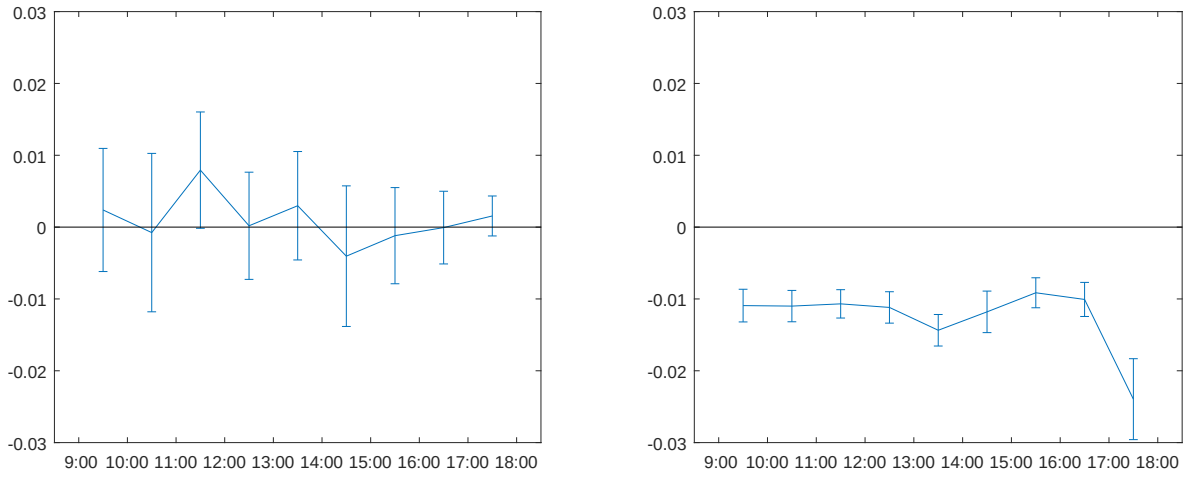


Figure 9: **Inventories and 10-Year Treasury Prices by Hour.** This figure plots intercepts (left panel) and slopes (right panel) of regressions of 10-year Treasury price changes on changes in market maker inventory, as proxied by cumulative net volume (sell minus buy). For a fixed hour of each day in the sample (say 9:00-10:00am), we obtain the midquote price change over that hour, as well as the inventory change over the same hour. This results in 858 price and inventory changes that are used in the 9:00-10:00am regression. The process is repeated for each hour of the day, resulting in price and inventory changes by hour. Since the market closes at 17:30, the last regression uses price and inventory changes from 17:00-17:30. Coefficient estimates and 99% Newey-West confidence intervals are plotted in the figure. Prices are measured in percent of par, and inventories are measured in 100 lots of $1 million. The underlying data use intraday orderbook observations from 4/2/2012 to 10/30/2015.

midquotes, with a Newey-West t-statistic of more than 12. In terms of magnitudes, the regression shows that an intraday cumulative net dollar volume change of about $1 billion in notional par value corresponds to an increase in 10-year Treasury prices of about 0.01 percent of par.

Table 1: **Price Changes and Inventories: Tests of Equal Slopes on Intraday vs Close**

This table tests the null hypothesis that the relationship between price changes and inventories is the same intraday and near the close of the trading day. That is, we run regressions of 10-year Treasury price changes on changes in market maker inventory, as proxied by cumulative net sell minus buy volume. For a fixed hour of each day in the sample (say 9:00-10:00am), we obtain the midquote price change over that hour, as well as the inventory change over the same hour. This results in 858 price and inventory changes that are used in the 9:00-10:00am regression. The process is repeated for each hour of the day, resulting in price and inventory changes by hour. Since the market closes at 17:30, the last regression uses price and inventory changes from 17:00-17:30. The difference between the slope coefficient from the 17:00 - 17:30 regression and the slope coefficients from each prior hour is reported, along with the associate t-statistic and p-value.

Time of Day i	$\beta_{close} - \beta_i$	t-stat	p-value
9:00 - 10:00	-0.01	[-6.43]	(0.000)
10:00 - 11:00	-0.01	[-6.43]	(0.000)
11:00 - 12:00	-0.01	[-6.67]	(0.000)
12:00 - 13:00	-0.01	[-6.34]	(0.000)
13:00 - 14:00	-0.01	[-4.76]	(0.000)
14:00 - 15:00	-0.01	[-5.76]	(0.000)
15:00 - 16:00	-0.01	[-7.40]	(0.000)
16:00 - 17:00	-0.01	[-6.81]	(0.000)

(TI-2) suggests intraday variation in the relationship between midquote prices and HFT inventory. In particular, the HFT's desire to end the day flat causes the relationship between quoted prices and inventory to steepen near the close of trading. To test this implication, we ran the same regression as in (TI-1), restricted to each hour of the active trading day.[11] Figure 9 confirms that the negative relationship between prices and HFT inventory holds intraday, and is statistically significant. Most importantly, the figure shows a strong steepening of the relationship between our measure of HFT inventory and prices near close. Table 1 confirms through a two-sample test that the steepening of the relationship near close is statistically significant relative to any prior hour in the trading day.

To examine (TI-3), we note that flash events are characterized by a rapid and large fall or rise in prices followed by a reversal in a short window of time. In our model, a large price increase can occur during a short time window if an intense buying, relative to selling, pressure is realized. The reversal in prices would then be accompanied by intense selling, relative to buying, pressure. The run-up in prices and the reversal can be explained by the same mechanism described in (TI-1), but applied to a short window of time so that it can be interpreted as a flash event. That is, during a period of intense buying pressure, cumulative net volumes rise, which means the HFT is shedding its inventory quickly. This increases price impact because the HFT revises its quotes upward in the attempt of discouraging further declines in inventory. A sudden reversal in order flow would then generate the opposite price impact. For prices to return to the same levels as those observed before the flash event, the volume of reversal trading would have to be of similar magnitude as the price run-up volume. When we plot the relationship between quoted prices and cumulative net volume during the October 15, 2014, flash event, we see that this is indeed the case. Figure 10 shows that the price run-up was accompanied by high buy volume relative to sell volume – a difference of about 800 lots of $1 million from 9:30 to 9:38am. Conversely, the subsequent reversal in prices from 9:38 to 9:44am was accompanied by about 900 lots sold relative to those bought. Importantly, we note that at least for this particular flash event, there were no discontinuities in the price path, in the sense of small volumes causing large price changes.

Finally, (TI-4) has implications for bid-ask spreads, which are expected to increase towards the close of trading. Indeed, the left panel of Figure 11 shows that, on average, across our full sample of trading days, bid-ask spreads are low from 9:00am onwards and rise sharply heading into close. Moreover, the right panel of Figure 11 shows that this is an empirical regularity that is maintained on almost all of the trading days in in our sample.

[11] As Fleming et al. (2014) show, trading activity rises sharply in this market at 8:00am ET. However, due to the prevalence of pre-announced, liquidity-distorting 8:30am news releases, we focus on the hours of 9:00 to closing at 17:30 ET.

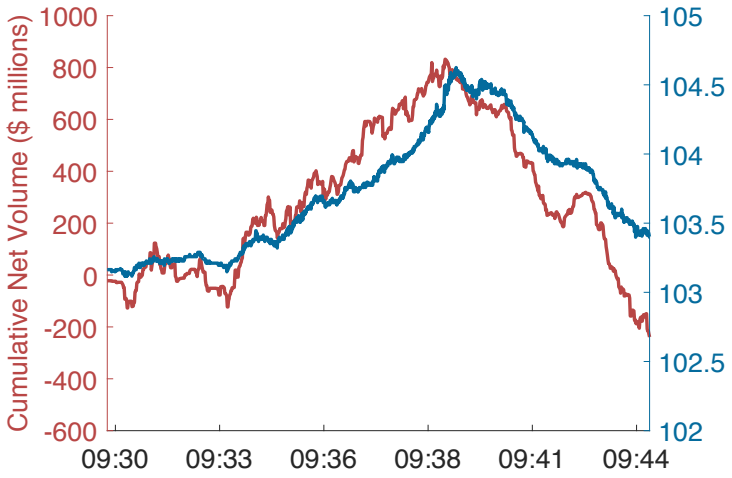


Figure 10: **Cumulative Net Volume and Prices during a Flash Event.** This figure plots cumulative net volume (buy minus sell) and mid-quote prices during the flash event on October 15, 2014.

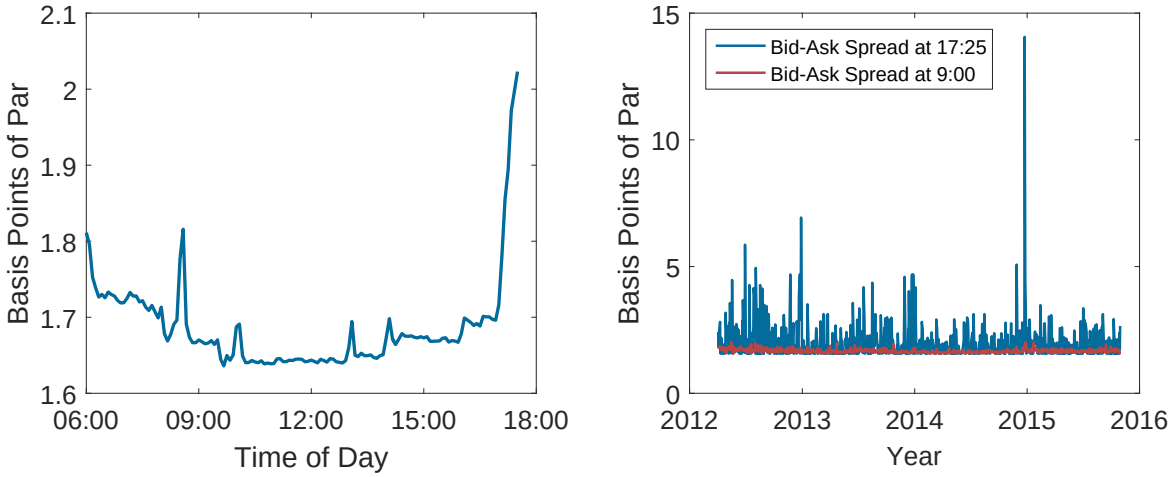


Figure 11: **Bid-Ask Spreads Intraday and over Time.** The left panel plots intraday bid-ask spreads for 10-year Treasuries. Using tick-by-tick order book data, bid-ask spreads are aggregated to 5-minute frequencies and then averaged across days. The spike at 8:30am coincides with major economic releases in the U.S. The right panel plots bid-ask spreads for 10-year Treasuries as recorded near the start of active trading 9:00 and near close at 17:25. The sample constists of high-frequency intraday observations from 4/2/2012 to 10/30/2015.

7 Price Stability

The overnight inventory constraint steepens the price impact function. We therefore expect more volatile price paths as a function of the tightness of the inventory constraint. To investigate the impact of the constraint on measures of price stability, we perform comparative statics with respect to the two key parameters: the overnight funding cost λ, and the arrival probabilities π^{BO}, π^{SO}. We compute and analyze three measures of price stability: the maximum deviation of a traded price from the long-run equilibrium price $\bar{p}$,

the maximum drawdown of the mid-price, and the maximum bid-ask spread during trading.

The maximum price deviation from the equilibrium price $\bar{p}$ is defined as

$$\max\{\bar{p} - \text{ minimum traded bid price, maximum traded ask price } - \bar{p}\}.$$

The maximum drawdown of the mid-price is defined by

$$\max_{1\le t\le T}\left(\max_{1\le s\le t} S_s - S_t\right),$$

where $S_t = \frac{1}{2}(\tilde{a}_t^* + \tilde{b}_t^*)$ is the mid-price. Moreover, we measure the maximum bid-ask spread at trading times, i.e., the spread whenever a trade occurs either with a buy or sell investor. Throughout the section, we use the demand and supply functions given as in Figure 1.

For each $(\lambda, \pi^{BO}, \pi^{SO})$ triplet, we run ten-thousand simulations starting with a zero initial inventory. On each path, we compute all three of these measures. The three panels of Table 2 present sample mean estimates for the expected maximum deviation of a traded price from the equilibrium price, the expected maximum bid-ask spread, and the maximum drawdown. The numbers in parenthesis are estimates for the standard deviation of sample means.

Table 2 shows that the maximum bid-ask spread is increasing in the overnight inventory cost λ (Panel A). This suggests that the higher overnight inventory cost is passed on to the end investors in the form of more extreme spikes in bid-ask spreads. Furthermore, Table 2 shows that prices are less stable when overnight inventory costs λ are high. This result holds for both measures of price stability (Panels B and C). This finding reflects the fact that price impact is an increasing function of the overnight inventory cost.

Panel A of Table 2 also shows that bid-ask spreads are declining in the arrival probabilities π^{BO}, π^{SO}, which would be expected from more and more liquid markets. High arrival probabilities induce more trading activities and opportunities, and hence the maximum deviation from the equilibrium value is also large when arrival probabilities are larger (Panel B). However, high arrival probabilities can induce larger drawdowns or reduce them, depending on the level of overnight inventory cost (see Panel C). Specifically, for smaller λ, the expected maximum drawdown is increasing with arrival probabilities; for larger λ, the expected maximum drawdown is decreasing with arrival probabilities. This means that, whether a highly liquid market will alleviate or aggravate maximum drawdowns, depends on whether the overnight inventory cost is high or low.

The plots in Figure 12 indicate that the distribution of the maximum bid-ask spread is heavily skewed to the left, with a mode at the largest value. As the arrival probabilities increase, the maximum bid-ask spread is more concentrated at this mode. This is expected because the maximum bid-ask spread along any

We use the demand and supply functions given as in Figure 1, and conduct comparative statics on the inventory cost coefficient λ (per lot of \$100 million) and the arrival probabilities π^{BO}, π^{SO}.

Table 2:
Panel A: Sample mean estimates of the expected maximum bid-ask spread at trading. Numbers in parenthesis are estimates for the standard deviation of the sample means. All quantities are expressed in basepoint of par.
Panel B: Expected maximum deviation from the equilibrium price, $\bar{p}$. Numbers in parenthesis are estimates for the standard deviation of the sample means. All quantities are expressed in basepoint of par.
Panel C: Expected maximum drawdown of the mid-price. Numbers in parenthesis are estimates for the standard deviation of the sample means. All quantities are expressed in basepoint of par.

Panel A: Max Spread	$\lambda = 0.002$	$\lambda = 0.01$	$\lambda = 0.02$	$\lambda = 0.04$	$\lambda = 0.08$
$\pi^{BO} = \pi^{SO} = 5\%$	1.7880 (2.3033e-05)	2.0064 (2.0108e-04)	2.1691 (3.7665e-04)	2.3700 (5.8740e-04)	2.5909 (7.6314e-04)
$\pi^{BO} = \pi^{SO} = 7\%$	1.7881 (2.2992e-05)	2.0076 (2.0184e-04)	2.1708 (3.8002e-04)	2.3707 (5.9166e-04)	2.5922 (7.7200e-04)
$\pi^{BO} = \pi^{SO} = 10\%$	1.7884 (2.2898e-05)	2.0093 (2.0300e-04)	2.1733 (3.8408e-04)	2.3748 (5.9812e-04)	2.5954 (7.8088e-04)
$\pi^{BO} = \pi^{SO} = 20\%$	1.7891 (2.2188e-05)	2.0149 (2.0571e-04)	2.1832 (3.9414e-04)	2.3872 (6.1458e-04)	2.6047 (7.9631e-04)
$\pi^{BO} = \pi^{SO} = 30\%$	1.7899 (2.1403e-05)	2.0217 (2.0702e-04)	2.1946 (4.0415e-04)	2.4036 (6.3242e-04)	2.6256 (8.1021e-04)
Theoretical spread at T	1.7962	2.0923	2.3375	2.6273	2.9000
Panel B: Max deviation	$\lambda = 0.002$	$\lambda = 0.01$	$\lambda = 0.02$	$\lambda = 0.04$	$\lambda = 0.08$
$\pi^{BO} = \pi^{SO} = 5\%$	0.3565 (9.7284e-07)	0.4350 (2.0269e-04)	0.4784 (2.7300e-04)	0.5267 (3.9184e-04)	0.5767 (6.1118e-04)
$\pi^{BO} = \pi^{SO} = 7\%$	0.3583 (9.9658e-07)	0.4361 (2.0533e-04)	0.4795 (2.7873e-04)	0.5279 (4.0312e-04)	0.5793 (6.3825e-04)
$\pi^{BO} = \pi^{SO} = 10\%$	0.3599 (1.0191e-04)	0.4376 (2.0893e-04)	0.4812 (2.8595e-04)	0.5300 (4.1296e-04)	0.5813 (6.4236e-04)
$\pi^{BO} = \pi^{SO} = 20\%$	0.3619 (1.0486e-04)	0.4398 (2.1502e-04)	0.4846(2.9775e-04)	0.5346 (4.3024e-04)	0.5869 (6.6160e-04)
$\pi^{BO} = \pi^{SO} = 30\%$	0.3626 (1.0532e-04)	0.4414 (2.1752e-04)	0.4876 (3.0546e-04)	0.5394 (4.4557e-04)	0.5934 (6.7936e-04)
Panel C: Max drawdown	$\lambda = 0.002$	$\lambda = 0.01$	$\lambda = 0.02$	$\lambda = 0.04$	$\lambda = 0.08$
$\pi^{BO} = \pi^{SO} = 5\%$	0.2142 (2.9570e-04)	0.4450 (5.9802e-04)	0.5704 (8.0682e-04)	0.7030 (1.2477e-03)	0.8408 (2.2192e-03)
$\pi^{BO} = \pi^{SO} = 7\%$	0.2202 (3.0202e-04)	0.4483 (6.0427e-04)	0.5729 (8.2453e-04)	0.7050 (1.2941e-03)	0.8454 (2.3461e-03)
$\pi^{BO} = \pi^{SO} = 10\%$	0.2246 (3.0812e-04)	0.4494 (6.0748e-04)	0.5727 (8.2378e-04)	0.7076 (1.2656e-03)	0.8402 (2.1197e-03)
$\pi^{BO} = \pi^{SO} = 20\%$	0.2288 (3.1391e-04)	0.4451 (6.1501e-04)	0.5655 (8.4144e-04)	0.6951 (1.2806e-03)	0.8423 (2.2391e-03)
$\pi^{BO} = \pi^{SO} = 30\%$	0.2293 (3.1616e-04)	0.4377 (6.1930e-04)	0.5527 (8.4952e-04)	0.6851 (1.3058e-03)	0.8289 (2.1909e-03)

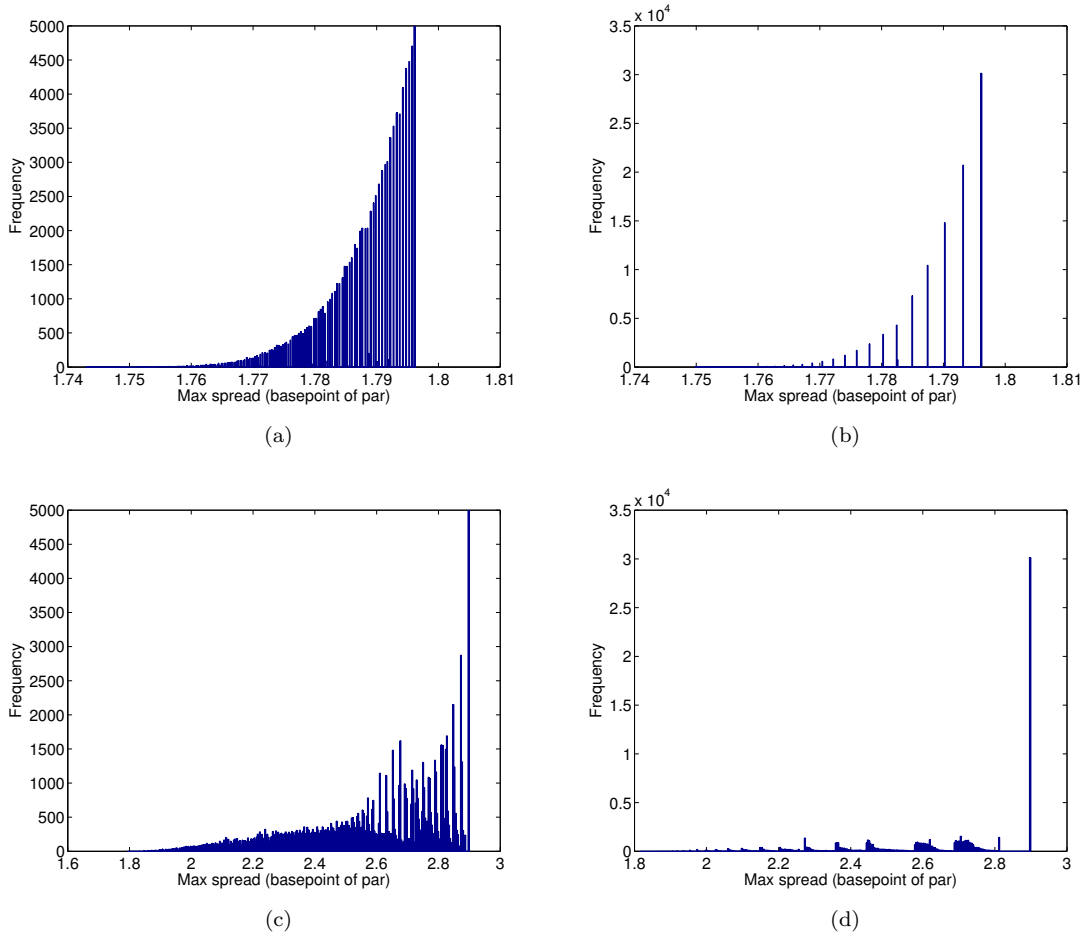


Figure 12: **Distribution of Maximum Bid-Ask Spread.** This figure plots the simulated distribution of the maximum bid-ask spread for four different triplets $(\lambda, \pi^{BO}, \pi^{SO})$. Recall that λ measures the severity of the overnight inventory cost (per lot of \$100 million), and π^{BO} and π^{SO} measure the arrival probabilities of buy and sell investors, respectively. Panel (a): $\lambda = 0.002$ and $\pi^{BO} = \pi^{SO} = 5\%$; Panel (b): $\lambda = 0.002$ and $\pi^{BO} = \pi^{SO} = 30\%$; Panel (c): $\lambda = 0.08$ and $\pi^{BO} = \pi^{SO} = 5\%$; Panel (d): $\lambda = 0.08$ and $\pi^{BO} = \pi^{SO} = 30\%$.

arrival sequence of buy and sell orders tends to be realized at the end of the trading day. At that time, it is explicitly given by the equation $B(\lambda)$ given in Lemma 5.5.

Figure 13 indicates that the distribution of the maximum deviation from the equilibrium price is skewed to the right. Higher arrival probabilities induces more trading activity and hence a larger maximum deviation. As the overnight inventory cost λ becomes larger, the maximum deviation is highly concentrated to the left, but has a very long tail to the right. This means that the HFT's motive to avoid a high overnight inventory cost can create larger price distortions in the market.

Figure 14 suggests that the distribution of the maximum drawdown is most skewed to the right. Higher λ means higher price impact and hence more price fluctuation leading to higher drawdown. Moreover, for a

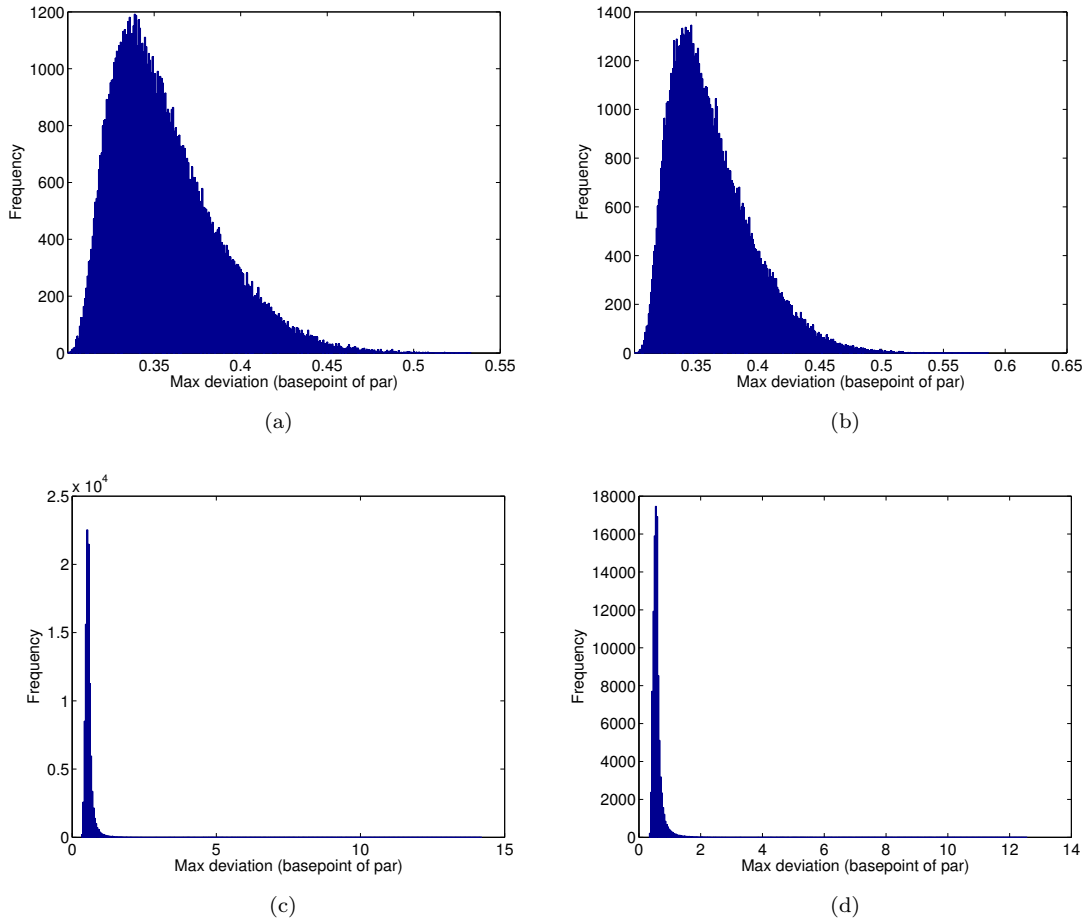


Figure 13: **Distribution of Maximum Deviation from Equilibrium.** This figure plots the simulated distribution of the maximum deviation of traded prices from the equilibrium value $\bar{p}$ for four triplets $(\lambda, \pi^{BO}, \pi^{SO})$. Recall that λ measures the severity of the overnight inventory cost (per lot of \$100 million), and π^{BO} and π^{SO} measure the arrival probabilities of buy and sell investors, respectively. Panel (a): $\lambda = 0.002$ and $\pi^{BO} = \pi^{SO} = 5\%$; Panel (b): $\lambda = 0.002$ and $\pi^{BO} = \pi^{SO} = 30\%$; Panel (c): $\lambda = 0.08$ and $\pi^{BO} = \pi^{SO} = 5\%$; Panel (d): $\lambda = 0.08$ and $\pi^{BO} = \pi^{SO} = 30\%$.

small λ, larger arrival probabilities result in more trading and profit opportunities for the HFT, and hence a larger maximum drawdown; for a large λ, larger arrival probabilities give more opportunities to manage inventory, and hence the HFT will exert higher effort to control its inventory, which results in a smaller maximum drawdown.

The above analysis was conducted by taking averages across sample paths. Next, we examine a specific sample path in which sellers arrive at higher intensity than buyers. As we have argued before, asymmetric arrival intensities are a typical feature of flash events. We simulate buy and sell order arrivals, and then analyze the impact of three different end of day inventory cost: $\lambda = 0.002$; $\lambda = 0.02$; and $\lambda = 0.08$. As it appears from the simulated trajectories of the mid-price reported in Figure 15, higher end of day inventory

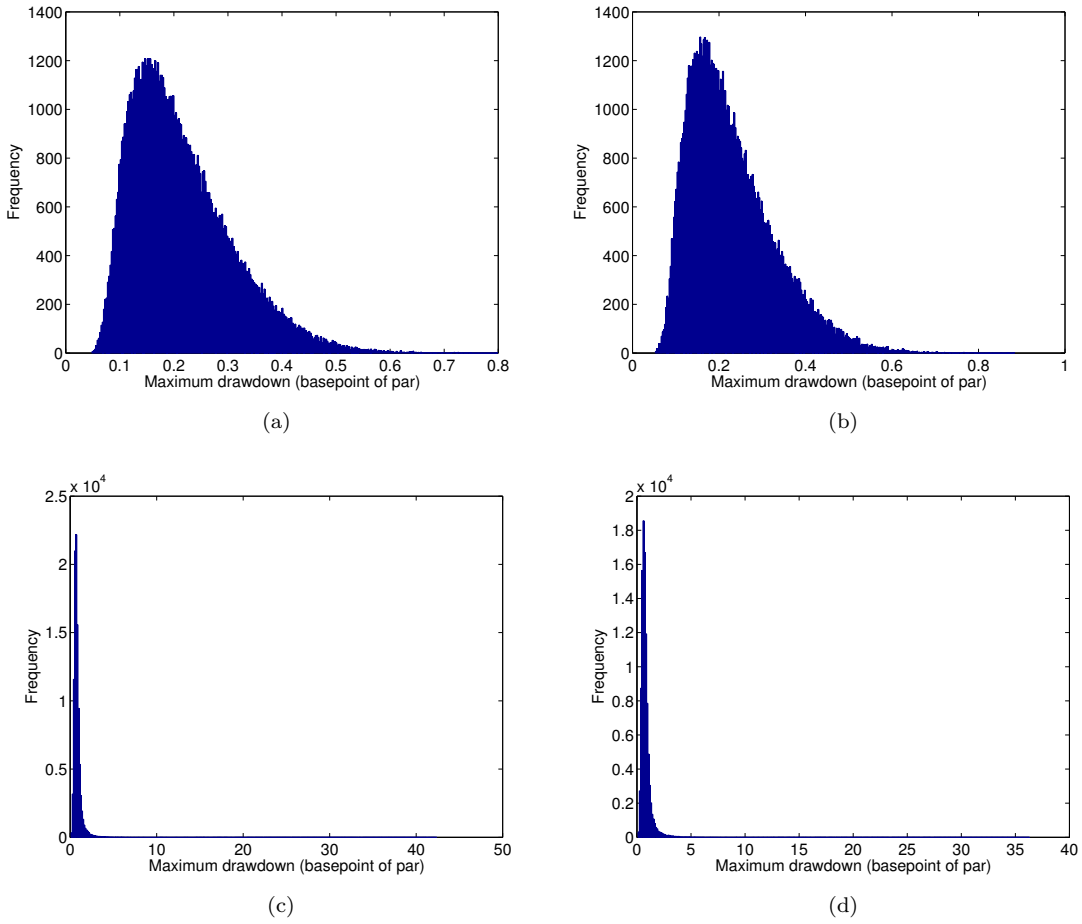


Figure 14: **Distribution of Maximum Price Drawdown.** This figure plots simulated distributions of the maximum drawdown of the midquote price for four different triplets $(\lambda, \pi^{BO}, \pi^{SO})$. Recall that λ measures the severity of the overnight inventory cost (per lot of \$1 million), and π^{BO} and π^{SO} measure the arrival probabilities of buy and sell investors, respectively. Panel (a): $\lambda = 0.002$ and $\pi^{BO} = \pi^{SO} = 5\%$; Panel (b): $\lambda = 0.002$ and $\pi^{BO} = \pi^{SO} = 30\%$; Panel (c): $\lambda = 0.08$ and $\pi^{BO} = \pi^{SO} = 5\%$; Panel (d): $\lambda = 0.08$ and $\pi^{BO} = \pi^{SO} = 30\%$.

costs amplify the downward pressure on prices caused by the arrival imbalance.

8 Welfare

This section conducts a welfare analysis. The demand and supply functions (1), (2) are reduced form specifications for the optimal trading strategy of end investors that we do not explicitly model. In order to compute a welfare measure for those end investors, we use the notion of consumer surplus. More specifically, if there is a buy order arriving at time s, then we measure the surplus for this order as the area between the

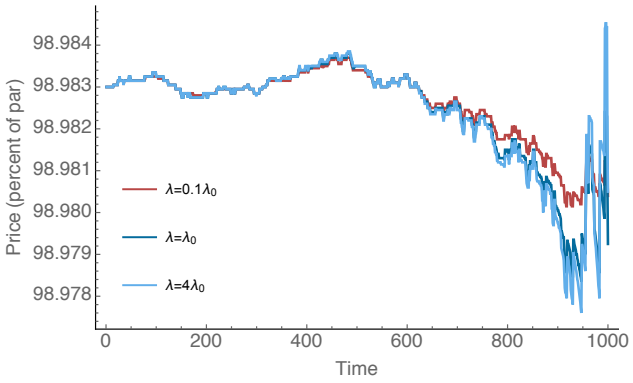


Figure 15: **Simulated Price Trajectories for Different Overnight Inventory Costs.** This figure plots simulated trajectories of the midquote price under the demand and supply functions specified as in Figure 1. We choose symmetric arrival probabilities $\pi^{BO} = \pi^{SO} = 10\%$. We consider three different values of λ: $\lambda = 0.002$ (per lot of \$100 million) for the red line; $\lambda = 0.02$ (per lot of \$100 million) for the dark blue line; and $\lambda = 0.08$ (per lot of \$100 million) for the light blue line. For each choice of λ, we consider the same arrival sequence of buy and sell orders. Evidently, the larger λ, and the more volatile the price, especially near the day's end.

traded quantity $Q^{BO}(\tilde{a}_s^*)$ and the demand curve $Q^{BO}(x)$:

$$\int_{\tilde{a}_s^*}^{\tilde{p}} [Q^{BO}(\tilde{a}_s^*) - Q^{BO}(x)]dx = \frac{c}{2}(\tilde{p} - \tilde{a}_s^*)^2.$$

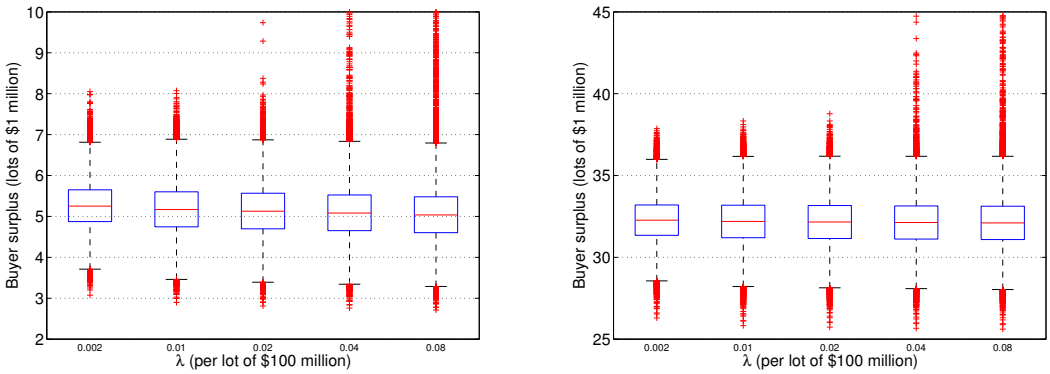
The cumulative surplus for all buy orders is then defined as the sum of all such areas when buyers arrive. The surplus for sellers is defined following the same logic. We can thus define the buy and the sell investor surplus respectively as

$$Surplus^{BO} = \frac{c}{2}\sum_{s=1}^{T}(\tilde{p} - \tilde{a}_s^*)^2 \Delta N_s^{BO}, \tag{39}$$

$$Surplus^{SO} = \frac{c}{2}\sum_{s=1}^{T}(\tilde{b}_s^* - \tilde{q})^2 \Delta N_s^{SO}, \tag{40}$$

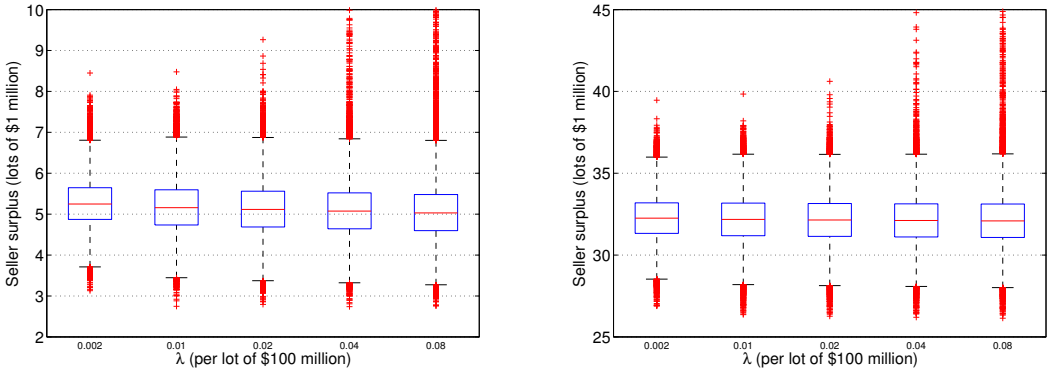
Figures 16 and 17 show that the surpluses for buyers and sellers are slightly higher when λ is small, but as π^{BO}, π^{SO} become larger, the surpluses are increased for all λ-values, and λ appears to play a smaller role for the surpluses' median, lower and upper 25% quantiles. However, when λ larger, there are more outliers, and the distribution of the surpluses becomes strongly positive skewed, especially when π^{BO}, π^{SO} are large. As λ measures the severity of the overnight funding costs, these results suggest that the HFT passes those costs onto the end investor, especially when the market is less liquid (lower π^{BO}, π^{SO}). Hence smaller overnight funding costs and a liquid market environment are both beneficial for end investors, as one might expect.

For the HFT, we use the objective function of which it is maximizing the expectation, (7), as a measure



(a) (b)

Figure 16: Box plots of the buyer's surplus for ten triplets $(\lambda, \pi^{BO}, \pi^{SO})$. (Certain outliers have been truncated to show at a higher resolution the distribution around the middle 50% quartile.) Panel (a): $\pi^{BO} = \pi^{SO} = 5\%$; Panel (b): $\pi^{BO} = \pi^{SO} = 30\%$.



(a) (b)

Figure 17: Box plots of the seller's surplus for ten different triplets $(\lambda, \pi^{BO}, \pi^{SO})$. (Certain outliers have been truncated to show details of the middle 50% quartile.) Panel (a): $\pi^{BO} = \pi^{SO} = 5\%$; Panel (b): $\pi^{BO} = \pi^{SO} = 30\%$.

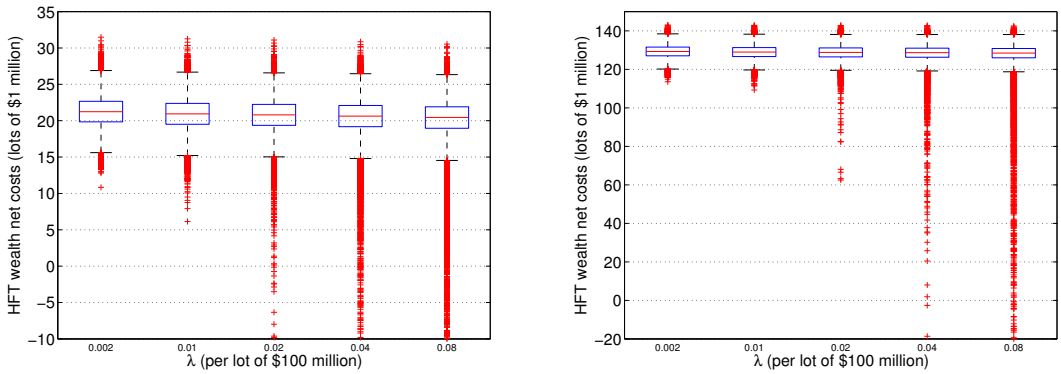
of welfare.

$$W_T - \lambda I_T^2 + \bar{p} I_T$$

This corresponds to the *optimal achievable wealth*, as it depends on the optimally chosen bid and ask prices, and is netted of the end-of-day inventory cost.

Figure 18 suggests that the median welfare of the HFT increases as the buy and sell order arrival probabilities (π^{BO}, π^{SO}) become larger, and decreases as the end-of-day costs λ increases. When π^{BO} and π^{SO} are at 30% (very liquid market), λ no longer plays a major role in determining the median, the lower and the upper 25% quantiles of the HFT's welfare. However, a larger λ always causes more outliers and

tail risks. It is also noteworthy that the HFT's welfare is strongly negatively skewed when λ is large. This reflects the increased challenge for the HFT to control its inventory in order to avoid the overnight funding cost.



(a) (b)

Figure 18: Box plots of the HFT's wealth minus overnight inventory cost for ten different triplets $(\lambda, \pi^{BO}, \pi^{SO})$. (Certain outliers have been truncated to show details of the middle 50% quartile.) Panel (a): $\pi^{BO} = \pi^{SO} = 5\%$; Panel (b): $\pi^{BO} = \pi^{SO} = 30\%$.

9 Conclusion

The HFT sector has grown rapidly in recent years. High frequency, automated, and algorithmic trading now account for the majority of volume in U.S. Treasury, global equity, and foreign exchange markets. While a number of researchers have investigated the HFT sector from an empirical angle, there is little theory that explores conceptual differences between dealer models of market making and HFT models of market making.

Our paper is an attempt to fill this gap along one dimension, namely, by studying the importance of the overnight inventory cost for the determination of price and liquidity dynamics. The distinguishing feature of our approach is to assume that the HFT does not face any constraints during the day, but faces an inventory cost at the end of the day. The HFT thus has a strong incentive to end the day with little to no inventory, which is a hallmark of their short investment horizons.

We characterize the optimal market making behavior of such an HFT and conduct comparative statics relative to the magnitude of end of day costs, as well as the degree of competition. The optimal price setting strategy of the HFT gives rise to bid-ask spreads and price impact metrics that tend to rise towards the end of the day, even though arrival rates of buyers and sellers are constant, a feature that is present in intraday U.S. Treasury data. Importantly, both bid-ask spread and price impact arise endogenously as functions of

inventory, time of day, and the magnitude of the overnight inventory cost. Even though trading is costless intraday, and the HFT only faces the inventory cost at the end of the day, equilibrium bid-ask spreads and price impact depend on the overnight inventory cost at all times during the day. The steepening of price impact due to the end of day constraint leads to more volatile price paths intraday.

Appendix

A Algorithm

In this section, we show that, if the i-derivative of value function $F(t,i)$ is piecewise linear, then so are $a_t^*(i)$ and $b_t^*(i)$. As a consequence, by Proposition 5.3 we deduce that $\partial_i F(t-1,i)$ is also piecewise linear. Since piecewise linear functions are computationally tractable (through their slopes and kinks), we obtain an efficient and semi-analytical backward algorithm that computes $\partial_i F(t,i)$, the optimal ask $a_t^*(i)$ and optimal bid $b_t^*(t)$ from the i-derivative of $F(T,i) = -\lambda i^2 + \bar{p}i$.

To begin, notice that a piece-wise linear function $f(x)$ can be uniquely identified by specifying the left slope (left derivative) at the first kink, the right slope (right derivative) at the last kink, and the value of the function at all intermediate kinks. In the sequel, we will use extensively the following representation

$$\{s_{-\infty}, (x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_k, f(y_k)), s_\infty\},$$

where $s_{-\infty}$ denotes the left derivative at the first kink, s_∞ denotes the right derivative at the last kink, and $(x_1, f(x_1)), \ldots (x_k, f(x_k))$ are the x-y pairs of the function at the intermediate kinks such that $x_1 < x_2 < \ldots < x_k$. The above representation is equivalent to

$$f(x) = [s_{-\infty}(x - x_1) + f(x_1)]1_{\{x \le x_1\}} + [s_\infty(x - x_k) + f(x_k)]1_{\{x > x_k\}} + \sum_{m=1}^{k-1} \left(\frac{x - x_m}{x_{m+1} - x_m} f(x_{m+1}) + \frac{x_{m+1} - x}{x_{m+1} - x_m} f(x_m) \right) 1_{\{x \in (x_m, x_{m+1}]\}}.$$

Let us assume that $\partial_i F(t,i)$ is piece-wise linear with k kinks. Then, it can be fully determined by the following:

$$\{s_{-\infty}, (i_1, \partial_i F(t,i_1)), \ldots, (i_k, \partial_i F(t,i_k)), s_\infty\}$$

where $s_{\pm\infty} < 0$ stand for the slope of $\partial_i F(t,i)$ for sufficiently large or small x. Recall the function $G_t(i)$ in Eq. (5.1) Then $G_t(i)$ admits a similar representation

$$\left\{ s_{-\infty} - \frac{2}{c}, (i_1, \partial_i F(t,i_1) - \frac{2}{c} i_1), \ldots, (i_k, \partial_i F(t,i_k) - \frac{2}{c} i_k), s_\infty - \frac{2}{c} \right\}$$

Using the above representation, we can determine the inverse G_t^{-1} using its representation (the x-coordinates are reversed because G_t is decreasing):

$$\left\{ \frac{1}{s_\infty - \frac{2}{c}}, (\partial_i F(t,i_k) - \frac{2}{c} i_k, i_k), \ldots, (\partial_i F(t,i_1) - \frac{2}{c} i_1, i_1), \frac{1}{s_{-\infty} - \frac{2}{c}} \right\}$$

As a result, $G_t^{-1}\left(\frac{p-2i}{c}\right)$ has representation

$$\left\{ \frac{1}{1 - \frac{s_{-\infty} c}{2}}, \left(\frac{p}{2} - \frac{c}{2} \partial_i F(t,i_1) + i_1, i_1 \right), \ldots, \left(\frac{p}{2} - \frac{c}{2} \partial_i F(t,i_k) + i_k, i_k \right), \frac{1}{1 - \frac{s_\infty c}{2}} \right\}$$

Using the linear relation between G_t^{-1} and $a_t(i)$ given in equation (23), we obtain that $a_t(i)$ is also piecewise linear and admits the representation

$$\left\{ \frac{1}{\frac{2}{s_{-\infty}c} - 1}, \left(\frac{p}{2} - \frac{c}{2}\partial_i F(t, i_1) + i_1, \frac{p}{2} + \frac{c}{2}\partial_i F(t, i_1)\right), \right.$$
$$\left. \ldots, \left(\frac{p}{2} - \frac{c}{2}\partial_i F(t, i_k) + i_k, \frac{p}{2} + \frac{c}{2}\partial_i F(t, i_k)\right), \frac{1}{\frac{2}{s_{\infty}c} - 1} \right\}$$

For the bid price, we use (23) and (24) to obtain that

$$b_t(i) = a_t(i + \frac{p-q}{2}) - \frac{p-q}{2}. \tag{41}$$

Hence, the optimal scaled bid function $b_t(i)$ is piecewise linear with representation

$$\left\{ \frac{1}{\frac{2}{s_{-\infty}c} - 1}, \left(\frac{q}{2} - \frac{c}{2}\partial_i F(t, i_1) + i_1, \frac{q}{2} + \frac{c}{2}\partial_i F(t, i_1)\right), \right.$$
$$\left. \ldots, \left(\frac{q}{2} - \frac{c}{2}\partial_i F(t, i_k) + i_k, \frac{c}{2}\partial_i F(t, i_k)\right), \frac{1}{\frac{2}{s_{\infty}c} - 1} \right\}.$$

We then use the expression for the derivative of the value function with respect to inventory given in equation (28). In fact, from the proof of Proposition 5.3, we know that (28) can be equivalently expressed as

$$\partial_i F(t-1, i) = \begin{cases} (1-\pi^{BO})\partial_i F(t,i) + \pi^{BO}\partial_i F(t, i-p), & i \geq L_t^0 \\ (1-\pi^{BO})\partial_i F(t,i) + \dfrac{\pi^{BO}}{c}(2a_t(i) - p), & L_t^0 > i \geq L_t^1 \\ (1-\pi^{BO}-\pi^{SO})\partial_i F(t,i) + \dfrac{\pi^{BO}}{c}(2a_t(i)-p) + \dfrac{\pi^{SO}}{c}(2b_t(i)-q), & L_t^1 > i > L_t^2 \\ (1-\pi^{SO})\partial_i F(t,i) + \dfrac{\pi^{SO}}{c}(2b_t(i)-q), & L_t^2 > i, \end{cases} \tag{42}$$

where L_t^0 is the inventory level such that $a_t(L_t^0) = 0$, i.e. it solves

$$\partial_i F(t, L_t^0 - p) + \tilde{p} = 0. \tag{43}$$

Given piecewise linear functions $\partial_i F(t,i)$, $a_t^*(i)$ and $b_t^*(t)$, we can conveniently use the above expression to obtain a representation for function $\partial_i F(t-1, i)$, which is necessarily also piecewise linear. Iterating this step, we can obtain the optimal ask $a_t^*(i)$, the optimal bid b_t^* and $\partial_i F(t,i)$ for all $i = 1, 2, \ldots, T$.

It is worth remarking that, as more recursive steps are taken, the structure of $\partial_i F(t,i)$ can become quite complex. This is because more and more kinks will arise because of horizontal translations, or additions between piecewise linear functions. One way to improve the efficiency of the algorithm and memory management is to work with a finite domain for the inventory, by only storing kinks within a pre-specified inventory interval, if we are only interested to know the optimal price policy functions when the inventory level is within this interval. We refer to the proof of Proposition 5.5 below for why this is legitimate.

B Proofs

Proof that (3) *solves the minimization problem* (4). We only consider $x \in [\tilde{q}, \tilde{p}]$, because the price outside of this interval will result in a one-sided market. Conditioning on $N_t^{BO} + N_t^{SO} = n$, we know that N_t^{BO} follows

a Binomial distribution with parameter $(n, \frac{\pi^{BO}}{\pi})$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\left(Q^{BO}(x)\,N_t^{BO} - Q^{SO}(x)\,N_t^{SO}\right)^2 | N_t^{BO} + N_t^{SO} = n\right] \\
=&\mathbb{E}\left[\left(Q^{BO}(x)\,N_t^{BO} + Q^{SO}(x)\,N_t^{BO} - Q^{SO}(x)n\right)^2 | N_t^{BO} + N_t^{SO} = n\right] \\
=&c^2\mathbb{E}\left[\left((\tilde{p}-\tilde{q})N_t^{BO} - (x-\tilde{q})n\right)^2 | N_t^{BO} + N_t^{SO} = n\right].
\end{aligned}$$

It follows that the optimal x that minimize the above quantity must solve the first-order condition:

$$(\tilde{p}-\tilde{q})\mathbb{E}[N_t^{BO}|N_t^{BO} + N_t^{SO} = n] - (x-\tilde{q})n = 0,$$

which is equivalent to

$$(\tilde{p}-\tilde{q})n\frac{\pi^{BO}}{\pi} - (x-\tilde{q})n = 0,$$

leading to (3). □

Proof of Lemma 5.1. For fixed t and all $i \in \mathbb{R}$, because $F(t,i)$ is assumed to be strictly concave in i (its derivative is decreasing), we know that $G_t(i)$ is strictly decreasing in i. Furthermore, we have assumed that $\partial_i F(t,i)$ is increasing in i and maps onto $\mathbb{R}$, thus we know that $G_t(-\infty) = \infty = -G_t(\infty)$.

After algebraic manipulations, we deduce that the following relations hold:

$$0 = \partial_i F(t, i-p+a_t(i)) + \frac{1}{c}(p - 2a_t(i)) = G_t(i-p+a_t(i)) - \frac{1}{c}(p-2i)$$

Using the definition of inverse functions, we obtain

$$i - p + a_t(i) = G_t^{-1}\left(\frac{p-2i}{c}\right)$$

leading to (23). Similarly, from

$$0 = \partial_i F(t, i+b_t(i)-q) + \frac{1}{c}(q - 2b_t(i)) = G_t(i+b_t(i)-q) - \frac{q-2i}{c},$$

we obtain (24).

To show that the mapping $i :\mapsto a_t(i)$ is strictly decreasing, we consider $i_1 < i_2$, then we have

$$\partial_i F(t, i_2 - p + a_t(i_1)) + \frac{1}{c}(p - 2a_t(i_1)) < \partial_i F(t, i_1 - p + a_t(i_1)) + \frac{1}{c}(p - 2a_t(i_1)) = 0,$$

where we used the fact that $\partial_i F(t,i)$ is strictly decreasing in order to get the first inequality. On the other hand, the mapping $a :\mapsto \partial_i F(t, i_2 - p + a) + \frac{1}{c}(p-2a)$ is clearly strictly decreasing, and $a_t(i_2)$ is the zero of this mapping. So we must have

$$a_t(i_2) < a_t(i_1).$$

Applying the same argument to equation $\partial_i F(t, i+b-q) - \frac{2b}{c} = 0$, we obtain the same result for $b_t(i)$.

Finally, notice that

$$\frac{p-2i}{c} = \frac{q-2i}{c} + \frac{p-q}{c} > \frac{q-2i}{c}.$$

By (23), (24) and monotonicity of G_t^{-1}, we have

$$a_t(i) - b_t(i) = G_t^{-1}\left(\frac{q-2i}{c} + \frac{p-q}{c}\right) - G_t^{-1}\left(\frac{q-2i}{c}\right) + p - q < p - q. \tag{44}$$

On the other hand, using (41) we have

$$a_t(i) - b_t(i) = \frac{p-q}{2} + a_t(i) - a_t(i + \frac{p-q}{2}) > \frac{p-q}{2},$$

where the inequality follows because $a_t(i)$ is strictly decreasing in i. This completes the proof. □

Proof of Lemma 5.2. The functions $a_t(i)$ and $b_t(i)$ in Lemma 5.1 give the optimal scaled "ask" and "bid" which solve the optimization problem

$$\sup_{a,b}\{\pi^{BO}[\frac{1}{c}a(p-a) + F(t, i-p+a) - F(t,i)] + \pi^{SO}[-\frac{1}{c}b(b-q) + F(t, i+b-q) - F(t,i)]\}.$$

From the fact that the objective function in this unconstrained problem is concave in a and b, we deduce that, for each fixed $i \in \mathbb{R}$, the optimal scaled ask and bid price should be given by $a_t(i)$ and $b_t(i)$ as long as $0 \le a_t(i) < p$ and $b_t(i) > q$. Using again the strict i-concavity of $F(t,i)$, we know that, for an i such that $a_t(i) \ge p$, there cannot be any a such that $0 < a < p$ that maximizes the Hamiltonian in (16), instead, any $a \ge p$, such as $a_t(i)$ itself, is a maximizer of this Hamiltonian. Therefore, the optimal scaled ask and bid prices are given by the expressions in equations (18) and (19). □

Proof of Proposition 5.3. As seen in Lemma 5.1, we always have $a_t(i) > b_t(i)$ for $a_t(i)$ and $b_t(i)$ in (23) and (24), respectively. Let L_t^0 be the inventory level where $a_t(i) = 0$, i.e. (43) holds. Then for all $i \ge L_t^0$, we have $a_t(i) \le 0$, and hence $b_t^*(i) = a_t^*(i) = 0$.

We use the optimal bid and ask prices determined above to show that the optimized Hamiltonian $H_t(i)$ in (17) is smooth in i (this requires verifying the smooth fit at the boundary points L_t^0, L_t^1, L_t^2). In the sequel we discuss the form of the optimized Hamiltonian, $H_t(i)$, in four different regions of the inventory.

1. For $i \ge L_t^0$, we have $a_t^*(i) = b_t^*(i) = 0$. This means that the inventory of the HFT is so large that it sells to buyers for free in the attempt of reducing inventory, so

$$\begin{aligned} H_t(i) = &\sup_{(a,b)\in(\mathbb{R}^+)^2}\left\{\pi^{BO}[\frac{a(p-a)^+}{c} + F(t, i-(p-a)^+)] + \pi^{SO}[-\frac{b(b-q)^+}{c} + F(t, i+(b-q)^+)]\right\} \\ &- (\pi^{BO}+\pi^{SO})F(t,i) \\ =&\pi^{BO}[F(t,i-p) - F(t,i)]. \end{aligned}$$

   It follows that

$$H_t'(i) = \pi^{BO}[\partial_i F(t, i-p) - \partial_i F(t,i)], \tag{45}$$

   which is positive because $\partial_i F(t,i)$ is decreasing in i.

2. For $L_t^0 > i \ge L_t^1$, we have $b_t^*(i) < q$ and $p > a_t^*(i) = a_t(i) > 0$, thus the HFT only trades with buy investors and

$$H_t(i) = \pi^{BO}\left[\frac{1}{c}a_t(i)(p - a_t(i)) + F(t, i-p+a_t(i)) - F(t,i)\right].$$

   Recall that $a_t(i)$ is strictly decreasing in i, so that we can take the differential of H_t with respect to i:

$$\begin{aligned} dH_t(i) =&\pi^{BO}\left[\frac{1}{c}(p-2a_t(i))da_t(i) + \partial_i F(t, i-p+a_t(i))(di + da_t(i)) - \partial_t F(t,i)di\right] \\ =&\pi^{BO}[\frac{1}{c}(p-2a_t(i)) + \partial_i F(t, i-p+a_t(i))da_t(i) + \pi^{BO}[\partial_i F(t, i-p+a_t(i)) - \partial_i F(t,i)]di \\ =&\pi^{BO}[\partial_i F(t, i-p+a_t(i)) - \partial_i F(t,i)]di \end{aligned}$$

where the last step follows from the first-order condition that $a_t(i)$ satisfies (see Eq. (21)). Hence, $H_t(i)$ is also differentiable, with

$$H_t'(i) = \pi^{BO}[\frac{1}{c}(2a_t(i) - p) - \partial_i F(t,i)] = \pi^{BO}[\partial_i F(t, i - p + a_t(i)) - \partial_i F(t,i)], \tag{46}$$

which is positive because $\partial_i F(t,i)$ is decreasing in i. By evaluating (45) and (46) at $i = L_t^0$, we notice that $H_t'(i)$ is continuous at $i = L_t^0$.

3. For $L_t^1 > i > L_t^2$, we have $q < b_t^*(i) = b_t(i) < a_t^*(i) = a_t(i) < p$, and the HFT actively trade with both buyers and sellers, so

$$\begin{aligned} H_t(i) =& \pi^{BO}[\frac{1}{c}a_t(i)(p - a_t(i)) + F(t, i - p + a_t(i))] + \pi^{SO}[-\frac{1}{c}b_t(i)(b_t(i) - q) + F(t, i + b_t(i) - q)] \\ & - (\pi^{BO} + \pi^{SO})F(t,i). \end{aligned}$$

It follows that (by the definitions of $a_t(i)$ and $b_t(i)$)

$$\begin{aligned} H_t'(i) =& \pi^{BO}[\partial_i F(t, i - p + a_t^*(i)) - \partial_i F(t,i)] + \pi^{SO}[\partial_i F(t, i + b_t^*(i) - q) - \partial_i F(t,i)] \\ =& \pi^{BO}[\frac{1}{c}(2a_t^*(i) - p) - \partial_i F(t,i)] + \pi^{SO}[\frac{1}{c}(2b_t^*(i) - q) - \partial_i F(t,i)]. \end{aligned} \tag{47}$$

By evaluating (46) and (47) at $i = L_t^1$, we notice that $H_t'(i)$ is continuous at $i = L_t^1$.

4. For $i \le L_t^2$, we have $a_t^*(i) = a_t(i) \ge p$ and $b_t^*(i) = b_t(i) > q$. So the HFT only trades with sell investors to increase its inventory, and

$$H_t(i) = \pi^{SO}[-\frac{1}{c}b_t(i)(b_t(i) - q) + F(t, i + b_t(i) - q) - F(t,i)]. \tag{48}$$

Recall that $b_t(i)$ is strictly decreasing in i, we can take the differential of H_t with respect to i and obtain:

$$\begin{aligned} dH_t(i) =& \pi^{SO}\left[\frac{1}{c}(q - 2b_t(i))db_t(i) + \partial_i F(t, i + b_t(i) - q)(di + db_t(i)) - \partial_t F(t,i)di\right] \\ =& \pi^{SO}[\frac{1}{c}(q - 2b_t(i)) + \partial_i F(t, i + b_t(i) - q)db_t(i) + \pi^{SO}[\partial_i F(t, i + b_t(i) - q) - \partial_i F(t,i)]di \\ =& \pi^{SO}[\partial_i F(t, i + b_t(i) - q) - \partial_i F(t,i)]di, \end{aligned} \tag{49}$$

where in the last equality we have used the first order condition satisfied by $b_t(i)$ (see Eq. (22)). Thus

$$H_t'(i) = \pi^{SO}[\frac{1}{c}(2b_t(i) - q) - \partial_i F(t,i)] = \pi^{SO}[\partial_i F(t, i + b_t(i) - q) - \partial_i F(t,i)], \tag{50}$$

which is negative because $\partial_i F(t,i)$ is decreasing in i. By evaluating (47) and (50) at $i = L_t^2$, we notice that $H_t'(i)$ is also continuous at $i = L_t^2$.

From (16) we know that $F(t-1,i)$ is also continuously differentiable in i, and

$$\partial_i F(t-1,i) = \partial_i F(t,i) + H_t'(i).$$

From the above analysis we deduce that $\partial_i F(t-1,i)$ is C^1, and that its derivative with respect to the inventory level is given by equation (28). The expression in (28) can be equivalently written more explicitly as (42). As both $a_t(i)$ and $b_t(i)$ are strictly decreasing in i, we deduce immediately from (42) that $\partial_i F(t-1,i)$ is also strictly decreasing, hence $F(t-1,i)$ is strictly concave. Moreover, it is clear that $\lim_{i\to\infty} \partial_i F(t,i) = -\infty$. As $b_t(i) > q$ for all $i < L_t^2$ and using the expression in the last line of (42) we have

$$\partial_i F(t-1,i) = (1 - \pi^{SO})\partial_i F(t,i) + \pi^{SO}\frac{1}{c}(2b_t(i) - q),$$

we know that

$$\lim_{i\to-\infty}\partial_i F(t-1,i)=\infty.$$

This completes the proof. □

Proof of Proposition 5.4. To prove the monotonicity of the sequences at hand, we notice that for any $t=2,\dots,T$

$$\begin{cases} H_t'(i)>0, & \forall i\geq L_t^1 \text{ (see, in the proof of Proposition 5.3)}\\ H_t'(i)<0, & \forall i\leq L_t^2 \text{ (see, in the proof of Proposition 5.3)}\\ \partial_i F(t-1,i)=\partial_i F(t,i)+H_t'(i). \end{cases}$$

It follows that (using (27) at time t and $t-1$),

$$\begin{aligned}\partial_i F(t-1,L_t^1)&=\partial_i F(t,L_t^1)+H_t'(L_t^1)>\tilde{q}=\partial_i F(t-1,L_{t-1}^1),\\ \partial_i F(t-1,L_t^2)&=\partial_i F(t,L_t^2)+H_t'(L_t^2)<\tilde{p}=\partial_i F(t-1,L_{t-1}^2).\end{aligned}$$

Because $\partial_i F(t-1,i)$ is strictly decreasing in i, we know that $L_{t-1}^1>L_t^1$ and $L_{t-1}^2<L_t^2$.[12] Finally, a straightforward calculation using (30) and (31) yields

$$L_T^1=\frac{\bar{p}-\tilde{q}}{2\lambda}>0,\quad L_T^2=-\frac{\tilde{p}-\bar{p}}{2\lambda}<0.$$

Hence, the sequence $(L_t^1)_{t=1}^T$ is positive and $(L_t^2)_{t=1}^T$ is negative. □

Lemma B.1. *For any t, the mappings $i:\mapsto i+a_t(i)$ and $i:\mapsto i+b_t(i)$ are strictly increasing.*

Proof. Recall that $a_t(i)$ solves

$$\partial_i F(t,i-p+a_t(i))+\frac{p-2a_t(i)}{c}=0. \tag{51}$$

For $i_1>i_2$, $a_t(i_1)<a_t(i_2)$, so by the monotonicity of $\partial_i F(t,i)$ in i, we have

$$i_1-p+a_t(i_1)>i_2-p+a_t(i_2), \tag{52}$$

thus we have $i_1+a_t(i_1)>i_2+a_t(i_2)$. The claim about $i+b_t(i)$ can be obtained directly using (41). □

Proof of Proposition 5.5. Let us first allow for a negative ask price. Then the recursive equation that $\partial_i F(t,i)$ satisfies, (42), needs a slight modification. Specifically,

$$\partial_i F(t-1,i)=\begin{cases}(1-\pi^{BO})\partial_i F(t,i)+\dfrac{\pi^{BO}}{c}(2a_t(i)-p), & i\geq L_t^1\\ (1-\pi^{BO}-\pi^{SO})\partial_i F(t,i)+\dfrac{\pi^{BO}}{c}(2a_t(i)-p)+\dfrac{\pi^{SO}}{c}(2b_t(i)-q), & L_t^1>i>L_t^2\\ (1-\pi^{SO})\partial_i F(t,i)+\dfrac{\pi^{SO}}{c}(2b_t(i)-q), & L_t^2>i,\end{cases} \tag{53}$$

Essentially, this allows the HFT to unload its inventory by even paying a price to the buyer, if the inventory level is extremely high. While this change affects the HFT's trading behavior in this extremely adverse scenario, it dose not alter the optimal strategy when the inventory level is moderate, as discussed later in the proof.

For any $i_1>i_2$, using the first-order condition (21), we have

$$\partial_i F(t,i_1-p+a_t(i_1))-\partial_i F(t,i_2-p+a_t(i_2))=\frac{2}{c}[a_t(i_1)-a_t(i_2)]. \tag{54}$$

[12] It is worth mentioning that the same argument can be applied to establish that the sequence $(l_t^x)_{t=1}^T$ is strictly decreasing, where l_t^x is the unique root to $\partial_i F(t,l_t^x)=x$ for a fixed $x<\tilde{q}$. On the other hand, if $x>\tilde{p}$, then the sequence $(l_t^x)_{t=1}^T$ is strictly increasing.

Dividing both sides of (54) by $(i_1 - p + a_t(i_1)) - (i_2 - p + a_t(i_2)) = i_1 - i_2 + a_t(i_1) - a_t(i_2)$, which is positive by Lemma B.1 above, we obtain

$$\frac{\partial_i F(t, i_1 - p + a_t(i_1)) - \partial_i F(t, i_2 - p + a_t(i_2))}{(i_1 - p + a_t(i_1)) - (i_2 - p + a_t(i_2))} = \frac{2}{c} \frac{\frac{a_t(i_1) - a_t(i_2)}{i_1 - i_2}}{\frac{a_t(i_1) - a_t(i_2)}{i_1 - i_2} + 1}. \tag{55}$$

We suppose that, for fixed $t = 2, \ldots, T$, there are positive constants $\lambda_t^1 > \lambda_t^2 > 0$ such that, for all $j_1 > j_2$

$$-2\lambda_t^1 \leq \frac{\partial_i F(t, j_1) - \partial_i F(t, j_2)}{j_1 - j_2} \leq -2\lambda_t^2. \tag{56}$$

Using (55) and (56) we obtain that

$$-\frac{\lambda_t^1 c}{1 + \lambda_t^1 c} \leq \frac{a_t(i_1) - a_t(i_2)}{i_1 - i_2} \leq -\frac{\lambda_t^2 c}{1 + \lambda_t^2 c}, \text{ so } -\frac{2\lambda_t^1}{1 + \lambda_t^1 c} \leq \frac{2}{c} \frac{a_t(i_1) - a_t(i_2)}{i_1 - i_2} \leq -\frac{2\lambda_t^2}{1 + \lambda_t^2 c}. \tag{57}$$

Likewise, for any $i_1 > i_2$, we have

$$-\frac{\lambda_t^1 c}{1 + \lambda_t^1 c} \leq \frac{b_t(i_1) - b_t(i_2)}{i_1 - i_2} \leq -\frac{\lambda_t^2 c}{1 + \lambda_t^2 c}, \text{ so } -\frac{2\lambda_t^1}{1 + \lambda_t^1 c} \leq \frac{2}{c} \frac{b_t(i_1) - b_t(i_2)}{i_1 - i_2} \leq -\frac{2\lambda_t^2}{1 + \lambda_t^2 c}. \tag{58}$$

In other words, the graphs of $\frac{2}{c} a_t(i)$ and $\frac{2}{c} b_t(i)$ are less steep than that of $\partial_i F(t, i)$ in absolute value. Using (53), (57) and (58), we deduce that

$$-2\lambda_t^1 \left(1 - \min\{\pi^{BO}, \pi^{SO}\} \frac{\lambda_t^1 c}{1 + \lambda_t^1 c}\right) \leq \frac{\partial_i F(t, i_1) - \partial_i F(t, i_2)}{i_1 - i_2} \leq -2\lambda_t^2 \left(1 - \frac{\pi \lambda_t^2 c}{1 + \lambda_t^2 c}\right), \tag{59}$$

where $\pi = \pi^{BO} + \pi^{SO}$. Because $\lambda_T^1 = \lambda_T^2 = \lambda$, from (56) and (59) we deduce that we can choose

$$\begin{cases} \lambda_{t-1}^1 = \lambda_t^1 \left(1 - \min\{\pi^{BO}, \pi^{SO}\} \frac{\lambda_t^1 c}{1 + \lambda_t^1 c}\right), & t = 2, 3, \ldots, T, \\ \lambda_{t-1}^2 = \lambda_t^2 \left(1 - \pi \frac{\lambda_t^2 c}{1 + \lambda_t^2 c}\right), & t = 2, 3, \ldots, T, \\ \lambda_T^1 = \lambda_T^2 = \lambda, \end{cases} \tag{60}$$

such that (56), (57) and (58) hold for all $t = 1, 2, \ldots, T$. Furthermore, from (21) and (22) we have

$$\partial_i F(t, i - p + a_t(i)) - \partial_i F(t, i + b_t(i) - q) = \frac{2}{c}[a_t(i) - b_t(i)] + \tilde{q} - \tilde{p}.$$

Using the same analysis as above, we deduce that

$$B(\lambda_t^1) > (\tilde{p} - \tilde{q}) \frac{\frac{1}{2} + \lambda_t^2 c}{1 + \lambda_t^2 c} > \frac{a_t(i) - b_t(i)}{c} > (\tilde{p} - \tilde{q}) \frac{\frac{1}{2} + \lambda_t^2 c}{1 + \lambda_t^2 c} = B(\lambda_t^2). \tag{61}$$

We now prove that the simplification of allowing for a negative ask price does not impact the optimal price policy functions in the active trading region. Indeed, the optimal bid and ask price functions are computed using data on the i-derivative of the value function, $\partial_i F(t, i)$, which is a local property of $F(t, i)$. Therefore, as long as the information needed to get the correct price policy functions over the active trading region is not altered when we allow for a negative ask price, we can still use the conclusions drawn from the above analysis. We exploit this intuition formally below.

First, to get the ask price at time t_0 in the active trading region $[L_{t_0}^2, L_{t_0}^1]$, we will need data on $\partial_i F(t_0, i - p + a_{t_0}(i))$ for all i in this region. Because $a_{t_0}(i) \leq p$ in this domain, it suffices to know data on $\partial_i F(t_0, i)$ for all $i \leq L_{t_0}^1$. On the other hand, for the scaled bid price, only values higher than q are relevant (to determine trading quotes and updating $\partial_i F(t_0 - 1, i)$), i.e. we only need to know data on $\partial_i F(t_0, i - q + b_{t_0}(i))$ for all

$i \leq L^1_{t_0}$. But by Lemma B.1, we know that $i :\mapsto i + b_{t_0}(i)$ is strictly increasing, so again, it suffices to know data on $\partial_i F(t_0, i)$ for all $i \leq L^1_{t_0}$.

Second, to get $\partial_i F(t_0, i)$ for all $i \leq L^1_{t_0}$, we use (42) to reduce it to data at time $t_0 + 1$. We now need to guarantee that we are indifferent to the decision of allowing or not for a negative ask price when the inventory $i \geq L^0_{t_0+1}$, where $L^0_{t_0+1}$ is defined in (43) as the critical inventory level at which the optimal ask price equals 0 at time $t_0 + 1$. To this end, we follow the proof of Proposition 5.4 and the footnote therein to prove that the sequence $(L^0_t)_{t=1}^T$ is strictly decreasing. On the other hand, a straightforward calculation yields that $L^0_T = \frac{\tilde{p}+\bar{p}}{2\lambda} + p$, so we deduce that $L^1_{t_0} \leq L^0_t$ for all $t = t_0, t_0 + 1, \ldots, T$. Hence the first line in (42) is not invoked for the recursion from time t_0 to $t_0 + 1$. Thus, we will need data on $\partial_i F(t_0 + 1, i)$ for all $i \leq L^1_{t_0}$, data on $a_{t_0+1}(i)$ for all $L^2_{t_0+1} \leq i \leq L^1_{t_0}$, and those on $b_{t_0+1}(i)$ for all $i \leq L^1_{t_0+1} < L^1_{t_0}$.

Similar as above, information on $a_{t_0+1}(i)$ for all $L^2_{t_0+1} \leq i \leq L^1_{t_0}$ can be inferred from that of $\partial_i F(t_0 + 1, i - p + a_{t_0+1}(i))$ for all $i \leq L^1_{t_0}$, or that about $\partial_i F(t_0 + 1, i)$ for all $i \leq L^1_{t_0}$, thanks to Lemma B.1. Likewise, information on $b_{t_0+1}(i)$ for all $i \leq L^1_{t_0+1}$ can be inferred from data on $\partial_i F(t_0 + 1, i)$ for $i \leq L^1_{t_0}$.

Thirdly, to get $\partial_i F(t_0 + 1, i)$ for $i \leq L^1_{t_0}$, we use (42) again, and the first line in (42) is not invoked because we know $L^1_{t_0} < L^0_{t_0+2}$, etc. Iterating this argument until we hit time T, we conclude that to make these recursive arguments work, the initial data needed is information on $\partial_i F(T, i)$ for all $i \leq L^1_{t_0}$. At T, we know that

$$\partial_i F(T, i) = -2\lambda i + \bar{p}.$$

Regressing backward, we notice that in each step needed to retrieve information about $\partial_i F(t, i)$, $t = t_0, t_0 + 1, \ldots, T$, we never used the first line in Eq. (42). In other words, for our objective at hand, we can legitimately draw the same conclusions by just considering the simplified version (53).

The above argument can be developed further to improve the efficiency of the algorithm introduced in Appendix A. In fact, in the analysis above we only need to know the information about $\partial_i F(t, i)$ for all $L^2_{t_0} \leq i \leq L^1_{t_0}$ and $t = t_0, t_0 + 1, \ldots, T$, if our objective is to know the optimal price policy function at time t_0 for the active trading region. In the actual implementation of the algorithm, we have exploited this observation and only stored the kinks of $\partial_i F(t, i)$ whose x-coordinates are in a pre-specified, sufficiently large interval $[A, B]$ (i.e. $A \ll L^1_2$ and $\frac{\tilde{p}+\bar{p}}{2\lambda} + p \gg B \gg L^1_1$). □

References

Admati, A. and Pfleiderer, P. (1988). A theory of intraday patterns: Volume and price variability. *Review of Financial Studies*, 1(1):3–40.

Aït-Sahalia, Y. and Saglam, M. (2016). High frequency traders: Taking advantage of speed. *NBER Working Paper*, 19531.

Amihud, Y. and Mendelson, H. (1980). Dealership market: Market-making with inventory. *Journal of Financial Economics*, 8(1):31–53.

Benos, E. and Sagade, S. (2016). Price discovery and the cross-section of high-frequency trading. *Journal of Financial Markets, forthcoming.*

Biais, B., Foucault, T., and Moinas, S. (2015). Equilibrium fast trading. *Journal of Financial Economics*, 116(2):292–313.

Biais, B. and Woolley, P. (2011). High frequency trading. *Working Paper, London School of Economics and Toulouse School of Economics.*

BIS (2011). High-frequency trading in the foreign exchange market. *Bank for International Settlements*, pages 1–37.

Board of Governors of the Federal Reserve System (2016). Senior credit officer opinion survey on dealer financing terms. December 2015.

Brogaard, J. and Garriott, C. (2015). High-frequency trading competition. *Working Paper, Foster School of Business, University of Washington.*

Brogaard, J., Hendershott, T., and Riordan, R. (2014). High-frequency trading and price discovery. *Review of Financial Studies*, 27(8):2267–2306.

Chaboud, A., Chiquoine, B., Hjalmarsson, E., and Vega, C. (2014). Rise of the machines: Algorithmic trading in the foreign exchange market. *Journal of Finance*, 69(5):2045–2084.

Danilova, A. and Julliard, C. (2015). Information asymmetries, volatility, liquidity, and the Tobin tax. *Working Paper, London School of Economics*.

Fleming, M., Mizrach, B., and Nguyen, G. (2014). The microstructure of a u.s. treasury ecn: The brokertec platform. *Working Paper, Federal Reserve Bank of New York Staff Reports*, 381.

Foucault, T., Hombert, J., and Roşu, I. (2016). News trading and speed. *Journal of Finance*, 71(1):335–382.

Glosten, L. and Milgrom, P. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14:71–100.

Glosten, L. R. and Harris, L. E. (1988). Estimating the components of the bid/ask spread. *Journal of financial Economics*, 21(1):123–142.

Grossman, S. and Miller, M. (1988). Liquidity and market structure. *Journal of Finance*, 43(3):617–633.

Hasbrouck, J. (1988). Trades, quotes, inventories, and information. *Journal of Financial Economics*, 22:229–252.

Hasbrouck, J. and Saar, G. (2013). Low latency trading. *Journal of Financial Markets*, 16:646–679.

Herndeshott, T., Jones, C., and Menkveld, A. (2011). Does algorithmic trading improve liquidity? *Journal of Finance*, 66:1–33.

Herndeshott, T. and Menkveld, A. (2014). Price pressures. *Journal of Financial Economics*, 114:405–423.

Joint Staff Report (2015). The U.S. Treasury Market on October 15, 2014. U.S. Department of the Treasury, Board of Governors of the Federal Reserve System, U.S. Securities and Exchange Commission, U.S. Commodity Futures Trading Commission.

Jones, C. (2013). What do we know about high-frequency trading? *Columbia Business School Research Paper*, 13-11.

Jovanovic, B. and Menkveld, A. J. (2011). Middlemen in limit-order markets. *Working paper, VU University of Amsterdam*.

Kyle, A. (1985). Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335.

Menkveld, A. (2013). High frequency trading and the new-market makers. *Journal of Financial Markets*, 16(4):712–740.

Menkveld, A. (2016). The economics of high-frequency trading: Taking stock. *Forthcoming in Annual Review of Financial Economics*, 8.

Puterman, M. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc.

SEC letters (2010). Tradeworx, inc. public commentary on sec market structure concept release. *Tradeworx, Inc.*

Securities and Exchange Commission (2010). Concept release on equity market structure. Release No. 34-61358.

Stoll, H. (1980). The supply of dealer services in securities markets. *Journal of Finance*, 33(4):1133–1151.

Stoll, H. R. (1989). Inferring the components of the bid-ask spread: Theory and empirical tests. *The Journal of Finance*, 44(1):115–134.