Federal Reserve Bank of New York
Staff Reports

Intraday Market Making with Overnight Inventory Costs

Tobias Adrian
Agostino Capponi
Michael Fleming
Erik Vogt
Hongzhong Zhang

Staff Report No. 799
October 2016
Revised March 2020

Intraday Market Making with Overnight Inventory Costs

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

Federal Reserve Bank of New York Staff Reports, no. 799
October 2016, revised March 2020
JEL classification: G12, G17, G23

Abstract

The U.S. Treasury market is highly intermediated by nonbank principal trading firms (PTFs). Limited capital forces PTFs to end the trading day roughly flat. We construct a continuous time market making model to analyze the trade-off faced by a profit-maximizing firm with overnight inventory costs, and develop closed-form representations of the optimal price policy functions. Our model reveals that bid-ask spreads widen as the end of the trading day approaches, and that increases in order arrival rates do not always lead to higher price volatility. Our empirical analysis shows that Treasury security trading costs increase as the close of trading approaches, consistent with model predictions.

Key words: market microstructure theory, market liquidity, market making, financial intermediation

---

Fleming: Federal Reserve Bank of New York (email: michael.fleming@ny.frb.org). Adrian: International Monetary Fund (email: tadrian@imf.org). Capponi: Columbia University (email: ac3827@columbia.edu). Vogt: Citadel LLC (email: erik.vogt@citadel.com). Zhang: Columbia University (email: hz2244@columbia.edu). This work was completed when Adrian and Vogt were at the Federal Reserve Bank of New York. The authors thank Yacine Aït-Sahalia, Yakov Amihud, Jean-Edouard Colliard, Thierry Foucault, Terry Hendershott, Stefano Lovo, Albert Menkveld, Ioanid Rosu, Gideon Saar, Pete Kyle, and seminar participants at VU Amsterdam, the University of Michigan, the Fields Institute, the IAQF/Thalesian Society, the University of Chicago Market Microstructure Conference, and the 2nd Eastern Conference in Mathematical Finance for valuable comments and suggestions. They also thank Francisco Ruela for excellent research assistance. The research of Agostino Capponi and Hongzhong Zhang has been supported by a NSF-DMS:1716145 grant. The views expressed in this paper are those of the authors and do not necessarily represent the position of the Federal Reserve Bank of New York, the Federal Reserve System, or the International Monetary Fund, including its management and executive directors.

To view the authors' disclosure statements, visit
https://www.newyorkfed.org/research/staff_reports/sr799.html.

Introduction

Since the turn of the century, new market makers have emerged across financial markets.[1] For U.S. Treasury securities, in particular, the Joint Staff Report of the U.S. Department of the Treasury, the Board of Governors of the Federal Reserve System, the Federal Reserve Bank of New York, the U.S. Securities and Exchange Commission, and the U.S. Commodity Futures Trading Commission (Joint Staff Report (2015)) provides an unprecedented glimpse into the market's evolving structure. Through the use of non-public data including participant names, the report identifies a new class of highly active non-bank intermediaries in the secondary market, collectively termed principal trading firms (PTFs).[2] Absent from this market before the mid-2000s, PTFs now account for over half of the volume on electronic trading platforms and close to 80 percent of message traffic.

Many PTFs are active liquidity providers, submitting passive orders to central limit order books and trading on their own accounts. PTFs tend to employ limited capital, relying on prime brokers for their funding and clearing needs. To limit the amount of capital held in margin accounts, PTFs keep their positions small and short-lived (Menkveld (2016)). It follows that PTFs largely unwind positions by the end of the trading day, even though they trade actively intraday. The practice of ending the day flat is in stark contrast to bank dealers, which tend to carry significant positions overnight.[3]

Our paper focuses on the overnight inventory management motive as a distinguishing characteristic of market making PTFs. To shed light on this characteristic, we consider a model consisting of a representative PTF – intermediating between randomly arriving buyers and sellers – which dynamically places bid and ask prices in order to maximize end-of-day profits, but with the additional objective of unwinding its positions before the market closes.

[1] See, for example, Securities and Exchange Commission (2010) and Menkveld (2013) for equity markets and Bank for International Settlements (2011) and Chaboud et al. (2014) for foreign exchange markets.

[2] The report's characterization of PTFs ("principal investor, deploys proprietary automated trading strategies, low latency typically key element of trading strategies", p. 50), is similar to that commonly used to describe high-frequency trading firms (HFTs). For consistency, and because low latency is not a feature of our model, we use the PTF and not the HFT terminology throughout the paper.

[3] Aggregated dealer positions in U.S. Treasuries are reported in the FR 2004 Weekly Report of Dealer Positions, Transactions, and Financing available at the Federal Reserve Bank of New York's website: https://www.newyorkfed.org/markets/primarydealers

In this context, the PTF can be interpreted as having access to unlimited financing intraday while facing an exogenously specified cost for remaining inventory at the end of the day. Our main contribution is to rigorously analyze the arising trade-off: the PTF balances profits from crossing the bid-ask spread against the present value, or shadow cost, of incurring inventory costs at the end of the day.

We derive closed-form representations for the PTF's optimal value function and price policy functions, and show that the PTF's intertemporal hedging demand feeds into liquidity and trade price dynamics throughout the day. Our model predicts that price impact and bid-ask spreads will rise toward the end of the trading day as the need of reaching a zero inventory target becomes stronger. Thus, in markets in which a significant proportion of liquidity providers operate with end-of-day inventory constraints, one would expect to see a similar widening of bid-ask spreads and rise in price impact. To verify this model implication, we analyze an anonymized data set of the same interdealer electronic communications network (ECN) reviewed in the Joint Staff Report (2015). We find statistically robust evidence of widening bid-ask spreads and heightened price impact as the end of the trading day approaches, consistent with our model predictions.

As a gauge of welfare, we study the extent to which the end-of-day inventory motive affects the value of the PTF, and surplus of buyers and sellers. Our analysis reveals that, while the PTF's overnight inventory costs always negatively affect the PTF's value as well as buyer/seller's surplus, the magnitude of the loss depends nontrivially on the arrival rate of market orders. Specifically, in markets where orders arrive more frequently, the buyer/seller's surplus per trade and the PTF's value have very little sensitivity to changes in the overnight inventory cost. By contrast, in markets with low order arrival rates, these quantities are highly sensitive to changes in the overnight inventory cost. These findings suggest that market makers that seek to limit their overnight inventories do not necessarily decrease buyer/seller's surplus, as long as the market is intrinsically active.

The endogenous price impact generated by the end-of-day inventory cost adds novel economic insights on market quality relative to the existing literature. We find that higher overnight inventory costs unambiguously lead to realizations of wider spreads. However, the

effects on price volatility are more subtle. The magnitude of the overnight inventory costs and the time of day can significantly tilt the relationship between the arrival rates of orders and measures of price volatility. On the one hand, higher arrival rates mitigate the impact of the overnight inventory cost on prices, and hence lower the resulting volatility, if trades occur earlier in the day. On the other hand, higher arrival rates also increase the trading frequency, which raises price volatility if the price impact is large, especially near the end of the trading day. Moreover, if the overnight inventory cost is very high, the instantaneous volatility exhibits a small, sudden drop before the day's close. This is because the aversion towards holding residual inventory near the day's close is so strong that the PTF trades very low volumes to avoid fluctuations in its inventory levels.

Our paper contributes to the broad literature on market microstructure theory and market making. In contrast to existing models of inventory-based market making (e.g. Hendershott and Menkveld (2014)), time-to-close is a key state variable in the market maker's value function in our setup. We show that the proposed finite horizon approach is crucial to explaining intraday price and liquidity dynamics observed in high-frequency U.S. Treasury data. Furthermore, we show that the intraday market maker's aversion to overnight inventory generates price impact and bid-ask spread dynamics that are absent in other finite-time horizon approaches or models of market making under asymmetric information (e.g. Kyle (1985), Glosten and Milgrom (1985) and Admati and Pfleiderer (1988)). Our model is well-suited to capture end-of-day dynamics for assets with little asymmetric information like Treasuries (there is arguably less asymmetric information about the value of public debt than, for instance, about the equity value of individual firms).

Bradfield (1979) studies a discrete-time dynamic model for market making of a profit maximizing specialist who targets an end-of-day inventory level. Similar to our paper, he finds that inventory hedging motives of the specialist increase price variability as the trading day unfolds. In contrast to our findings, he shows that the specialist maintains his inventory at the targeting level if his limit order book is in its average position, which results in a flat price trajectory. Furthermore, he does not analyze the impact of inventory control on bid-ask spreads, price impact, and welfare of market participants.

The structure of our market making setup is quite unique, and it requires solving an inventory control problem driven by jump processes. In contrast to existing literature on this topic (e.g. Bayaraktar and Ludkovski (2012) and Cartea and Jaimungal (2015)), the control state variable is the size of the jumps and not the jump intensity. Stochastic control problems of this kind are associated with non-dominated probability measure changes, and are described by a second order backward stochastic differential equation (see Soner et al. (2012)). By deriving analytical representations for the optimal strategies and associated quantities, we are able to obtain economic insights on the endogenous price impact after constructing an explicit mapping from inventories/volumes to prices.

A methodological contribution of our paper is the solution of the inventory control problem. Our solution concept requires an extension of techniques used for standard market making problems, because the end-of-day inventory constraint generates dynamic hedging motives. The proposed methodology opens the door to solving nonstandard control problems, i.e., those driven by pure-jump systems in which the control is applied at random times on the size of jumps.

The rest of the paper is organized as follows. We discuss institutional details of PTFs in Section 1. We introduce the market making model with overnight costs in Section 2. Section 3 formulates the PTF's decision making problem and performs an intertemporal analysis of the optimal bid and ask price policies. Section 4 presents a comparative statics analysis for price stability measures and welfare of market participants. Section 5 provides empirical evidence of our model predictions against U.S. Treasury data. Section 6 concludes. Technical proofs are delegated to the Appendix.

1 The Importance of Overnight Inventory Costs

We believe our paper is the first to explicitly consider the impact of an end-of-day inventory cost on intraday pricing, liquidity dynamics, and welfare of market participants. As shown in the next section, the overnight inventory cost represents the disutility of the market maker from carrying inventory overnight, and therefore captures a preference for ending the day

"flat". The desire to end the day flat is an agreed upon characteristic of PTFs. For example, in its concept release on equity market structure, the Securities and Exchange Commission (2010), p. 45, describes "professional traders acting in a proprietary capacity that engage in strategies which generate a large number of trades on a daily basis." The concept release describes the common characteristics of these firms, including "ending the trading day in as close to a flat position as possible (that is, not carrying significant, unhedged positions overnight)." Menkveld (2016) corroborates these statements, highlighting that PTFs (HFTs) are best thought of as a new type of financial intermediary that trades in large volumes intraday but avoids carrying positions overnight. Cvitanic and Kirilenko (2010) argue that high-frequency traders manage inventories to ensure that no positions are carried overnight after markets close.

Duffie and Ashcraft (2007)'s empirical study is supportive of this characterization and presents findings that can be viewed as testable implications of our model.[4] While not having an explicit model, they predicate their empirical analysis on the idea that, towards the end of the day, traders in the federal funds market are more desperate to run their inventory levels towards the target values, and therefore adjust the prices they quote, and are willing to accept, accordingly.[5] Quoting their paper:

> "Banks do not have much incentive to hold reserve balances in large amounts at the close of the business day because these balances do not earn interest from the Fed. Unnecessary end-of-day balances could have been exchanged for interest-bearing overnight assets such as federal funds loans or reverse purchase agreements. During the business day, financial institutions are permitted to have negative balances at a below-market interest rate in their accounts ... Motivated in part by discussions with federal funds traders, we find that federal funds trading is significantly more sensitive to balances in the last hour of the day. For

[4] A separate empirical study by Benos and Sagade (2016) analyzes proprietary data from U.K. equity markets over a four-month period and finds that "HFTs generally end the day with a relatively flat position", with a volume-weighted end-of-day position corresponding to 5% of their total intraday volume on average.

[5] Duffie and Ashcraft (2007) invoke models from the over-the-counter search literature, where trading opportunities are random and valuations of different agents load more significantly on individual properties when there are less trading opportunities, and conversely are more similar when the asset can still be traded.

example, at some large banks, federal funds traders responsible for targeting a small, nonnegative, end-of-day balance ask other profit centers of their banks to avoid large unscheduled transactions (for example currency trades) near the end of the day .... Once a federal funds trader has a reasonable estimate of the day's yet-to-be-executed send and receive transactions, he or she can adjust pricing and trading negotiations with other banks to push the bank's balances in the desired direction. We show evidence of this behavior and, further, find that lending is more active when federal funds rate volatility in the trailing half hour is high."

We believe our paper is the first to provide theoretical support for the temporal pattern highlighted in the above quote. Existing market making models of inventory management are not able to generate this pattern. For instance, in Amihud and Mendelson (1980)'s monopolistic market making model, a specialist incurs costs of inventory replenishment throughout an infinite trading horizon. In their model, the dealer is constrained to hold the inventory within a pre-specified interval at all times, and optimally chooses bid/ask prices to maximize the long-term/stationary growth rate of its wealth process. Hendershott and Menkveld (2014) use a classical stochastic optimal linear regulator framework, and effectively solve a discrete-time perpetual optimization problem with a quadratic intraday inventory cost, coming from the aversion of the market maker to the risk of distributed dividends. As in our model, the demand and supply functions of buyers and sellers are linear and exogenously specified. Because of the perpetual nature of the problem, both in Amihud and Mendelson (1980) and Hendershott and Menkveld (2014), the optimal bid-ask spread and price pressures are time-homogenous. In contrast, the time-to-close plays a crucial role in our model because an unbalanced inventory position held by the PTF near the day's close may generate high price instability: the management of the inventory executed by the PTF to restore a position close to flat may lead to large purchases and sales, which in turn cause high fluctuations in ask and bid prices.

The December 2015 Senior Credit Officer Survey on Dealer Financing Terms of the Board of Governors of the Federal Reserve System (2016) summarizes answers to special

questions on intraday and overnight credit extended to PTFs. Overnight positions of PTFs 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 of 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 on the intraday credit risk management of exchanges, central clearing counterparties, and dealers suggests that intraday inventory costs might be close to, or exactly, zero, depending on when the PTF is trading, and that PTFs face incentives to carry little inventory overnight.

Our model predicts that PTFs become more and more hesitant to post competitive quotes and transact toward the end of the day, reflecting their aversion to holding positions overnight. We would therefore expect activity to shift toward standard bank dealers at this time, and we present empirical evidence consistent with this prediction later in the paper. Moreover, as shown in the Joint Staff Report (2015), the median absolute end-of-day position for PTFs is 4.4% of its daily volume, but the comparable figure for bank-dealers is 19.0%, consistent with the idea that PTFs are especially averse to holding overnight positions. The median maximum absolute intraday position for PTFs is 15.3% versus 28.3% for bank-dealers, suggesting that PTFs are willing to hold substantial positions intraday.

The academic literature discusses a few additional underlying reasons for closing out positions at the end of the trading day. Brogaard and Garriott (2019) suggest a risk management motive, as PTFs 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 PTF'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.[6]

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

2 The Model

This section introduces our mathematical model of market making. We consider a trading day which runs continuously from time 0 through T. We assume a single dealer model, i.e., the PTF market maker is the sole price maker. We do not consider the interactions between PTFs resulting from their competition over bid/ask prices or traded quantities.

There is a single asset traded in a dealer market. A PTF market maker always takes the other side of arriving market orders. It sells for (market) buy orders and buys for (market) sell orders. Buy and sell orders arrive in the market according to two independent Poisson processes. The staggered arrival of buy and sell orders implies that the PTF is the only counterparty available for trade when an order arrives.

In the next two subsections, we introduce the model of buy/sell order arrival and the PTF objectives. Section 2.1 introduces the demand function and arrival times of buy and sell orders. Section 2.2 introduces the PTF's objective function, which is used to determine the bid price b, and the ask price a. Throughout the paper, we fix a complete filtered probability space $(\Omega, \mathbb{P}, \mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]})$ capturing all randomness.

2.1 Buy and Sell Orders

Buy orders arrive with a deterministic time-varying intensity $\pi^B(t) > 0$, and we use N^B to denote the $\mathbb{F}$-adapted non-homogeneous Poisson process which counts the number of such arrivals. Similarly, the arrival of sell orders is described by an independent $\mathbb{F}$-adapted non-homogeneous Poisson process N^S with a deterministic time-varying intensity $\pi^S(t) > 0$. We assume that the fundamental price of the asset follows a Brownian motion with volatility $\sigma > 0$:

$$dS_t = \sigma dB_t, \quad t \geq 0, \tag{1}$$

where $(B_t)_{t\geq 0}$ is a standard $\mathbb{F}$-Brownian motion, independent of the Poisson processes N^B and N^S, and $S_0 > 0$ is a positive constant. We use $-p < 0$ to denote the minimum price (relative to the fundamental) at which a sell order is placed, and $p > 0$ the maximum price

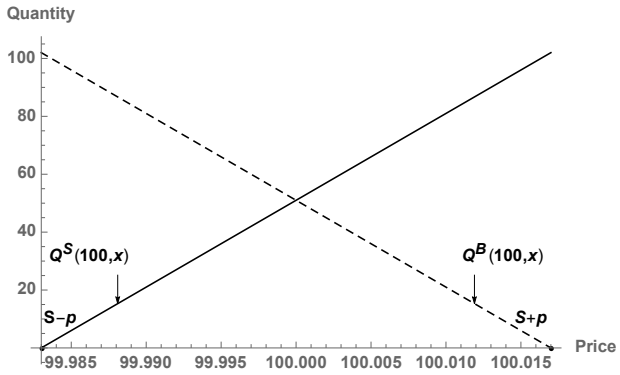


Figure 1: **Example of Demand and Supply Functions.** This figure illustrates the supply and demand functions of buyers and sellers. The price (x-axis) is in terms of percentage of par, and the quantity (y-axis) is measured in lots of \$1 million. We set the current fundamental price $S = 100$, $p = 0.017$ and slope $c = 30$ (lots of \$1 million per basis point of par). The quantity supplied by a seller (solid) is an increasing function of price. Similarly, the quantity demanded by a buyer (dashed) is a decreasing function of price.

(relative to the fundamental) at which a buy order is placed. That is, at time t, $S_t - p$ is the reservation price for sell orders and can be interpreted as a stop loss, while $S_t + p$ can be viewed symmetrically for buy orders. For a given ask price x at time $t \in [0, T]$, the number of shares demanded by buyers is given by

$$Q^B(S_t, x) = c\,(S_t + p - x)\,, \tag{2}$$

where $c > 0$ is a constant. For a given bid price x at time $t \in [0, T]$, the number of shares supplied by sellers is given by

$$Q^S(S_t, x) = c\,(x + p - S_t). \tag{3}$$

Above, we have assumed that both the demand and supply curves have the same slope c.[7] Note that the demand and supply functions Q^B and Q^S are reduced form models for (stochastic) preferences, beliefs, investment objectives, and hedging motives of buyers and sellers. A similar form of linear demand and supply function for buy and sell orders is considered by Hendershott and Menkveld (2014), where the traded quantity depends on

[7]This assumption can be relaxed at the expenses of sacrificing analytical tractability, but without qualitatively changing the conclusions of our analysis.

the deviation of the quoted price from the unobserved efficient price. In contrast to our setup, buyers and sellers arrive synchronously at all times in their model, and they do not focus on intraday trading patterns. Our model assumes deterministic continuous arrival rates for buyers and sellers, but makes the demand and supply function price-dependent. In the context of optimal liquidation, Bayaraktar and Ludkovski (2012) and Cartea and Jaimungal (2015) model order arrivals through a point process whose intensity depends on the liquidation price. Hence, while in our model a lower bid price reduces liquidity supply, in their model a similar effect may be obtained by reducing the arrival rate of sell orders.

2.2 PTF

In our model of market making, the PTF optimally chooses bid and ask prices through time. The PTF's cash holdings at time t, W_t, are given by its initial cash holdings, W_0, plus its trading revenue, i.e., cumulative proceeds from trades with buyers less cumulative outlays from trades with sellers. Specifically, a trade with a buyer at time t results in $Q^B(S_t, a_t)$ shares of the asset sold at price a_t; likewise, a trade with a seller at time t results in $Q^S(S_t, b_t)$ shares of the asset purchased at price b_t. This leads to the following expression for the PTF's cash holdings at time t:

$$W_t^{(a,b)} = W_0 + \int_0^t a_u \, Q^B(S_u, a_u) \, dN_u^B - \int_0^t b_u \, Q^S(S_u, b_u) \, dN_u^S, \tag{4}$$

where we set $dN_u^S := N_u^S - N_{u-}^S$ and $dN_u^B := N_u^B - N_{u-}^B$, and W_0 is a constant. The PTF's inventory accumulated in the interval $[0, t]$ is given by the number of shares purchased from sellers minus the number of shares sold to buyers until time t. That is,

$$I_t^{(a,b)} = I_0 + \underbrace{\int_0^t Q^S(S_u, b_u) \, dN_u^S}_{\substack{\text{Shares purchased} \\ \text{from sellers}}} - \underbrace{\int_0^t Q^B(S_u, a_u) \, dN_u^B}_{\substack{\text{Shares sold} \\ \text{to buyers}}}, \tag{5}$$

where I_0 is a constant.

The PTF maximizes the expected value of its end-of-day wealth, which is given by the sum

of its cash holdings W_T and the value of its end-of-day inventory marked at the fundamental price S_T, minus the cost for holding end-of-day inventory. Altogether, this leads to the following maximization problem for the PTF:

$$\max_{(a.,b.)\in\mathcal{A}_{0,T}} \mathbb{E}\left[W_T^{(a,b)} + S_T I_T^{(a,b)} - \lambda\left(I_T^{(a,b)}\right)^2\right], \tag{6}$$

subject to the budget constraint (4) and inventory dynamics (5). Above, $\mathbb{E}[\cdot]$ is the expectation operator under the probability measure $\mathbb{P}$, $W_T^{(a,b)} + S_T I_T^{(a,b)}$ is the PTF's total end-of-day wealth, and $\lambda > 0$ is a constant quantifying the severity of the overnight inventory cost, which is assumed to be quadratic in the size of inventory held at T. For on-the-run securities, in particular, prices remain near par, so a \$3 million par value position has a market value of roughly \$3 million, and variation in the market value of a position is closely approximated by variation in the par value of the position.[8] In other markets such as equities, where shares can trade for arbitrary prices, one would need to penalize the dollar amount of the inventory.

Because $S_T I_T^{(a,b)} - \lambda\left(I_T^{(a,b)}\right)^2 = I_T^{(a,b)}\left(S_T - \lambda I_T^{(a,b)}\right)$, one can interpret the inventory penalty as driven by a linear instantaneous price impact generated when selling the whole terminal inventory $I_T^{(a,b)}$ via a market order at time T. We use $\mathcal{A}_{0,T}$ to denote the collection of all admissible controls (which will be specified below) over the time period $[0,T]$.

The PTF's problem amounts to optimally choosing the ask and bid trajectories $(a.,b.) = (a_t,b_t)_{t\in[0,T]}$ which maximize the expected end-of-day wealth net of overnight inventory costs. The ask a_t and bid b_t are decided based on the information available before any trading at t occurs. Formally, we have:

Definition 2.1. *Let $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]}$ be the filtration generated by $(S_t, N_t^B, N_t^S)_{t\in[0,T]}$, then the admissible control set $\mathcal{A}_{0,T}$ includes all real-valued, $\mathbb{F}$-predictable, left continuous ask and bid strategies over the period $[0,T]$, such that*

$$\int_0^T \left((a_u)^2 + (b_u)^2\right) du < \infty. \tag{7}$$

[8] Treasury securities are issued at a price close to par. The securities are then only on the run for a short time (e.g., one month in the case of the 5-year note), until another security of the same original maturity is issued. As a result, there is not much time for their prices to move away from their initial values.

From the linear demand and supply functions (2)-(3), dynamics of cash holdings and inventory (4)-(5), we immediately realize that the condition in (7) ensures that the first order variation of $W_t^{(a,b)}$ has bounded expectation, and $\mathbb{E}[(I_t^{(a,b)})^2] < \infty$ and $\int_0^T \mathbb{E}[(I_t^{(a,b)})^2]dt < \infty$. In particular, the condition requires that the expected overnight cost of any admissible strategy must be finite.

Note that our PTF only provides liquidity, and is the only counterparty available for trade when an order arrives. It employs limit orders only, and always crosses against market orders. In a model with informational asymmetries, Rosu (2019) studies the optimal order choice of an informed trader, who dynamically decides between a market order or a limit order, based on the magnitude of privately observed mispricing. In his model, the informed trader submits a market order if there is extreme mispricing, and a limit order otherwise. Moreover, our model considers a single monopolistic dealer. In a possible future extension to a competitive environment, PTFs could have different beliefs and each PTF might submit limit orders that could be crossed by other PTFs.

3 The Control Problem

In this section, we analyze the control problem solved by the PTF. We formulate the PTF's dynamic optimization problem in terms of the Hamilton-Jacobi Bellman (HJB) equation in Section 3.1. We characterize its optimal solution in Section 3.2. We present an intertemporal analysis of the price policies and formulate the theoretical predictions of the model through formal statements in Section 3.3. We discuss the endogenous nature of the price impact in our model in Section 3.4.

3.1 Dynamic Programming Formulation

The value function of the control problem, defined as the PTF's optimal expected utility at time t given its current cash holdings w and inventory level i, is given by

$$V_t := \operatorname*{ess\,sup}_{(a.,b.)\in\mathcal{A}_{t,T}} \mathbb{E}\left[U\big(S_T, W_T^{(a,b)}, I_T^{(a,b)}\big)\big|\mathcal{F}_t\right], \quad \mathbb{P}\text{-a.s.}$$

where $\mathcal{A}_{t,T}$ is the collection of admissible strategies in Definition 2.1 restricted in the time interval $[t,T]$, and the PTF's end-of-day utility is

$$U(S,w,i) = w + Si - \lambda i^2.$$

By virtue of the dynamic programming principle, for $0 \le t \le u \le T$, we have that

$$V_t = \operatorname*{ess\,sup}_{(a,b)\in\mathcal{A}_{t,T}} \mathbb{E}\left[V_u\big|\mathcal{F}_t\right], \quad \mathbb{P}\text{-a.s.} \tag{8}$$

To determine the optimal value V_0 and the optimal price policies, we first restrict our attention to admissible strategies that are Markovian in the fundamental price S, the cash holdings W, and the inventory level I (we will later show in Theorem 3.2 that the optimal price policies are indeed Markovian). Then there exists a deterministic function v such that

$$V_t = v(t, S_t, W_t^{(a,b)}, I_t^{(a,b)}), \quad \mathbb{P}\text{-a.s..} \tag{9}$$

From our pre-specified supply and demand curves, we know that an incoming buy order at time t will reduce the PTF's inventory by $Q^B(S_t, a_t)$, and increase its cash holdings by $a_t\, Q^B(S_t, a_t)$. Likewise, an incoming sell order at time t will increase the PTF's inventory by $Q^S(S_t, b_t)$, but reduce its cash holdings by $b_t\, Q^S(S_t, b_t)$. Therefore, for any Markovian admissible control on the ask and bid prices $(a_t, b_t)_{t\in[0,T]}$, the controlled state process $(W_t^{(a,b)}, I_t^{(a,b)})_{t\in[0,T]}$ constitutes a pure jump process. Specifically, given the state $\{S_{t-} = S, W_{t-}^{(a,b)} = w, I_{t-}^{(a,b)} = i\}$ and the control pair (a_t, b_t), we have the time t transition

of the cash holdings and inventory processes given by

$$(W_t^{(a,b)}, I_t^{(a,b)}) = \begin{cases} (w + a_t\, Q^B(S, a_t), i - Q^B(S, a_t)), & \text{with probability } \pi^B(t)dt, \\ (w - b_t\, Q^S(S, b_t), i + Q^S(S, b_t)), & \text{with probability } \pi^S(t)dt, \\ (w, i), & \text{with probability } 1 - (\pi^B(t) + \pi^S(t))dt. \end{cases} \tag{10}$$

The time when a transition occurs is thus completely determined by the arrival sequences of buy 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 for the PTF to control inventory. From equations (1), (8), (9) and (10), we obtain the HJB equation satisfied by the function v:

$$\partial_t v + \frac{1}{2}\sigma^2 \partial_S^2 v + \sup_{(a.,b.)\in\mathbb{R}^2} H(t, S, w, i, a, b) = 0, \tag{11}$$

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

$$\begin{aligned} H(t, S, w, i, a, b) \;:=\; & \pi^B(t)[v(t, S, w + a\, Q^B(S, a), i - Q^B(S, a)) - v(t, S, w, i)] \\ & + \pi^S(t)[v(t, S, w - b\, Q^S(S, b), i + Q^S(S, b)) - v(t, S, w, i)]. \end{aligned}$$

3.2 Optimal Price Policies

We determine the Markovian bid and ask price policies which maximize the PTF's expected utility by solving the HJB equation (11). The linearity of the value function U in the cash holdings variable w suggests that we can rewrite

$$v(t, S, w, i) = w + F(t, S, i),$$

where $F(t, S, i) = v(t, S, 0, i)$ is the optimal expected utility of a PTF which starts with zero cash holdings and an inventory level i at time t (more precisely, after the trade at time t, if it

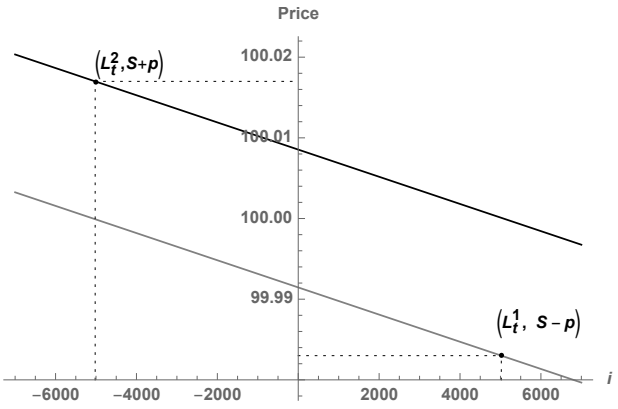


Figure 2: **The Optimal Price Policy Functions.** The optimal policy functions (of the current inventory level i) for bid (gray) and ask (black) prices, $b_t^*(S,i)$ and $a_t^*(S,i)$, at a fixed time $t=0$, with $S=100$. We set $T=10$. When the inventory is low (i.e. $i \leq L_0^2$), the ask price is higher than $S+p$, so that buyers do not buy from the PTF, but sellers sell $Q^S(S, b_0^*(S,i))$ shares to the PTF in each trade (see Figure 1). When the inventory is high (i.e. $i \geq L_0^1$), the bid price is lower than $S-p$, so that sellers do not sell to the PTF, but buyers purchase $Q^B(S, a_0^*(S,i))$ shares from the PTF in each trade. When the inventory is in the active trading region (i.e. $L_0^2 < i < L_0^1$), both ask and bid prices are between $S-p$ and $S+p$, and the PTF can trade with both buyers and sellers and earn a positive bid-ask spread. Moreover, for these moderate inventory levels, it is shown in Theorem 3.2 that both the ask and bid price functions are linear in the inventory level, hence their slope can be measured by the reciprocal of the width of the active trading region, $L_0^1 - L_0^2$. A detailed analysis of the inventory boundaries is given in the next subsection.

occurs). Our methodology exploits the concave property of the function $F(t,S,i)$, which we will establish in Theorem 3.2. For the moment, let us assume that for $t \in [0,T]$, the function $F(t,S,i)$ is strictly concave and continuously differentiable in i with a derivative mapped onto $\mathbb{R}$. This means, broadly speaking, that the function $F(t,S,i)$ behaves like a quadratic function with a negative leading coefficient. In particular, when $|i|$ is very large, the optimal expected utility $F(t,S,i) \ll 0$ and the marginal optimal expected utility $\partial_i F(t,S,i) > 0$ if $i < 0$ and $\partial_i F(t,S,i) < 0$ if $i > 0$. Using this function $F(t,S,i)$, we will derive the optimal ask price a_t^* and the optimal bid price b_t^*, as well as their monotonicity properties with respect to the inventory level.

Lemma 3.1. *For any $t \in [0, T)$, we have*

$$a_t^*(S,i) = S + \frac{1}{c} G_{t,S}^{-1}\left(S + p - \frac{2i}{c}\right) - \frac{i}{c} + p, \tag{12}$$

$$b_t^*(S,i) = S + \frac{1}{c} G_{t,S}^{-1}\left(S - p - \frac{2i}{c}\right) - \frac{i}{c} - p, \tag{13}$$

where $G_{t,S}^{-1}$ is the i-inverse function of a strictly decreasing function $G_{t,S}$:

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

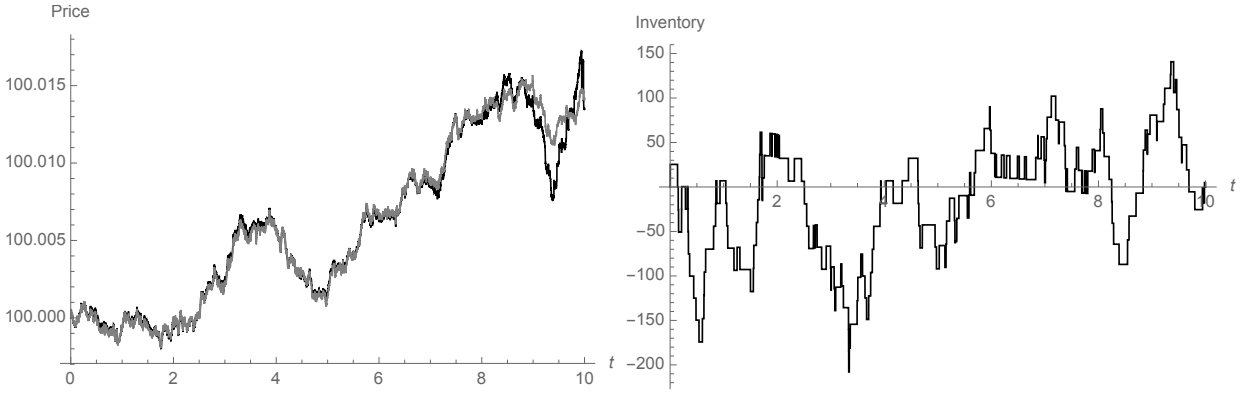
The mappings $i \mapsto a_t(S,i)$ and $i \mapsto b_t(S,i)$ are all strictly decreasing, continuous, and mapping onto $\mathbb{R}$. Moreover, for all $\lambda > 0$ we have

$$p < a_t^*(S,i) - b_t^*(S,i) < 2p. \tag{14}$$

Lemma 3.1 implies a number of properties for the optimal ask and bid prices. First and foremost, it shows that both $a_t^*(S,i)$ and $b_t^*(S,i)$ are *continuous, non-increasing functions* of the inventory level at time $t-$ (see Figure 2). This can be intuitively understood as follows. As its inventory gets larger, the PTF would like to offload inventory to reduce the penalty for holding a large inventory position at the close. To that end, the PTF wants to sell a larger number of shares to the buyer, and consequently sets a low ask price $a_t^*(S,i)$. At the same time, it wants to reduce the bid so that the seller is only willing to supply a small number of shares (or none) and its inventory thus does not increase much. Second, when the reservation prices of the buyers and sellers, $S + p$ and $S - p$, are close, the bid-ask spread is narrow (see (14)).

3.3 Intertemporal Analysis of Optimal Price Policies

We know from Lemma 3.1 that the optimal ask and bid prices both depend on the PTF's inventory level (see Figure 3 for a simulation of the price and inventory trajectories). Next, we want to identify the critical inventory thresholds, i.e., the levels at which the PTF decides



(a) Mid-Quote and the Fundamental Price

(b) Inventory

Figure 3: **Simulated Trajectories of the Mid-Quote Price, Fundamental Price, and PTF's Inventory Level.** Simulated trajectories of the fundamental price process (gray in Panel (a)), the mid-quote price process (black in Panel (a)), and the corresponding inventory path (Panel (b)). The demand and supply functions are those specified in Figure 1. We choose the arrival rates $\pi^B(t) = \pi^S(t) \equiv 10$, the end-of-day inventory cost $\lambda = 0.02$ per \$100 million par, and assume a zero initial inventory. We set the annualized volatility of the fundamental price process $\sigma = 3.75\%$. The trajectory of the price pressure process (the difference between the mid-quote and the fundamental price process) is *negatively* correlated with that of the PTF's inventory process, consistent with what is theoretically shown in earlier sections. The strength of this dependence increases as the day's close approaches. Noticeably, over the course of the day, the PTF's inventory crosses above and below 0 multiple times.

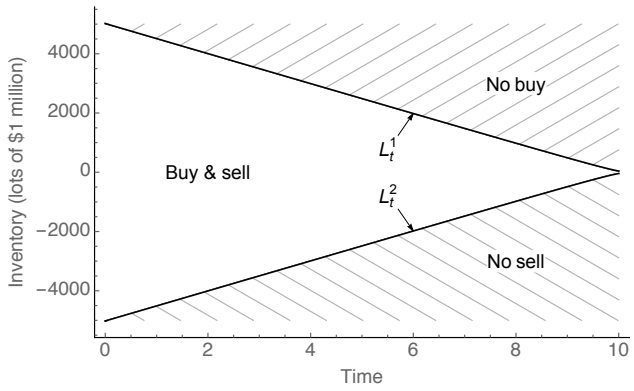


Figure 4: **The Critical Inventory Thresholds.** When the PTF's accumulated inventory crosses the thresholds L_t^i, $i = 1, 2$, the PTF's trading activity will qualitatively change. When the PTF's inventory level (in lots of \$1 million) is between L_t^1 and L_t^2, the PTF actively trades with both buyers and sellers. As time passes, the active trading region specified by the inventory levels at which the PTF trades with both buyers and sellers shrinks, because the impact of the end-of-day inventory cost becomes stronger. This means that the PTF will manage its inventory so as to keep it inside the active trading region, in order to avoid any one-sided trading near the day's close. Notice that, when the arriving intensity are constants, i.e. $\pi^S(t) \equiv \pi^S > 0, \pi^B(t) \equiv \pi^B > 0$ (shown in the figure), the boundaries L_t^1 and L_t^2 are essentially linear. The underlying reason is that the function $\alpha(t) \sim \frac{1}{c(\pi^B+\pi^B)(T-t)}$ when $t \ll T$. See (16) and Proposition 3.5 below.

to post ask and bid prices equal to $S_t + p$ and $S_t - p$ so as to prevent trading with buyers and sellers, respectively. Specifically, we define the critical inventory boundaries L_t^1 and L_t^2 as the unique solutions to the following equations:

$$b_t^*(S_t, L_t^1) = S_t - p, \quad a_t^*(S_t, L_t^2) = S_t + p. \tag{15}$$

Recall that Lemma 3.1 asserts that the bid-ask spread $a_t^*(S_t, i) - b_t^*(S_t, i)$ stays inside the open interval $(p, 2p)$ for all i, and thus $a_t(S_t, L_t^1) < b_t(S_t, L_t^1) + 2p = p = a_t(S_t, L_t^2)$, so that the boundaries L_t^1, L_t^2 satisfy

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

When the inventory level i is between L_t^2 and L_t^1, the PTF actively trades both with buyers and sellers. We refer to this range of inventories as the *Buy & Sell region* (see Figure 4). If the PTF's inventory level i is higher than L_t^1, then it only trades with buyers to unload its inventory. We refer to this range of inventories as the *No Buy region*. Conversely, if

the PTF's inventory level i is lower than L_t^2, then it only trades with sellers to build up its inventory. We refer to this range of inventories as the *No Sell region*. This result shows that, because of aversion to inventory, a monopolistic PTF may not always quote on both sides of the market and capture every spread, even in the absence of competition from other PTFs.

The main theorem (Theorem 3.2 below), proven in the Appendix, derives a closed-form expression for the value function $v(t, S, w, i)$, and shows that it possesses certain time-invariant properties with direct economic interpretations. We stress that the value function also determines the optimal value when the price policies are not restricted to the Markovian class.

Theorem 3.2. *Let $\alpha(t)$ be the unique negative root to the following equation* [9]

$$\alpha'(t) = -c(\pi^B(t) + \pi^S(t))\frac{\alpha^2(t)}{1 - c\alpha(t)}, \quad t \in [0, T), \tag{16}$$

with terminal condition $\alpha(T) = -\lambda$. Define $v(t, S, w, i) = w + F(t, S, i)$, with

$$F(t, S, i) = \alpha(t)i^2 + Si + cp^2 \int_t^T \frac{(\pi^B(u) + \pi^S(u))}{4(1 - c\alpha(u))} du. \tag{17}$$

Then the PTF's optimal value at time t is given by $V_t = v(t, S_t, W_t, I_t)$, $\mathbb{P}$-a.s. where W_t and I_t are the PTF's cash holdings and inventory level, respectively, at time $t \in [0, T]$. Moreover, the optimal price policy functions are given by

$$a_t(S, i) = S + \frac{(1 - 2c\alpha(t))}{2(1 - c\alpha(t))}p + \frac{\alpha(t)}{1 - c\alpha(t)}i, \tag{18}$$

$$b_t(S, i) = S - \frac{(1 - 2c\alpha(t))}{2(1 - c\alpha(t))}p + \frac{\alpha(t)}{1 - c\alpha(t)}i. \tag{19}$$

That is, this strategy beats all Markovian/non-Markovian admissible strategies in solving the optimization (6).

[9] The ordinary differential equation for $\alpha(t)$ is obtained by imposing that $v(t, S, w, i) = w + F(t, S, i)$ satisfies the HJB equation (11), where the function $F(t, S, i)$ given in (17) is quadratic in i (see the proof in the Appendix). It follows from Eq. (16) that if the terminal condition $\alpha(T) = -\lambda < 0$, then $\alpha'(t) < 0$ for all $t \in [0, T]$. Hence, $\alpha(t)$ is a decreasing function of time throughout the whole interval.

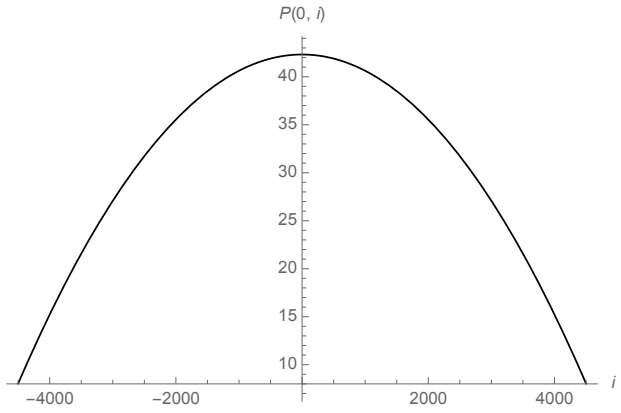


Figure 5: **The Optimal Present Values of the Inventory Constraint.** The value function $F(t, S, i)$ measures the PTF's optimal expected utility at T, as seen from time t, and given that the level of inventory at t is i. Suppose that the PTF can cash out its inventory immediately in the secondary market at the fundamental price $S_t = S$ per share. Then the residual value $P(t, i) = F(t, S, i) - Si$, which is independent of S as seen from (17) or (20), gives the optimal present value of the end-of-day inventory cost. This figure plots $P(t, i)$ (in lots of \$1 million) for $t = 0$, as a function of i (in lots of \$1 million), given constant arriving intensities $\pi^B(t) \equiv \pi^B > 0, \pi^S(t) \equiv \pi^S > 0$.

At any time $t \in [0, T]$, the value function $F(t, S, i)$ is strictly concave in i. This means that the PTF is always averse to holding inventory, because it accounts at any point in time for the cost incurred at T if it holds residual inventory. The intercept of $F(t, S, i)$, given by $F(t, S, 0)$, quantifies the optimal expected trading revenue less the overnight cost if the PTF begins with a zero inventory at time t. The positivity of $F(t, S, 0)$ for all times $t \in [0, T)$ indicates that the PTF will trade throughout the entire day, even if its inventory returns to zero and the time is close to the end of the day. Moreover, consider the time t optimal expected utility of the PTF given in (17), net of inventory holdings valued at the current fundamental price, i.e.,

$$P(t, i) := F(t, S, i) - Si = \alpha(t)i^2 + cp^2 \int_t^T \frac{(\pi^B(u) + \pi^S(u))}{4(1 - c\alpha(u))} du. \tag{20}$$

We can interpret $P(t, i)$ as the *optimal present value of the inventory cost* (see Figure 5). Then $\partial_i P(t, i)$ gives the PTF's marginal gain stemming from an infinitesimal change in its inventory. In particular, from (20) we notice that $\partial_i P(t, i) > 0$ if $i < 0$ and $\partial_i P(t, i) < 0$ if $i > 0$, hence it is always beneficial for the PTF to trade in a way that targets a zero inventory

position. Hence, the PTF's inventory will exhibit mean-reversion around the zero inventory level during the trading period $[0, T]$, see Figure 3.

In what follows, we focus on the optimal policy determined in Theorem 3.2, and drop the subscript (a^*, b^*) in the notation for the inventory level. The following corollary formally characterizes the dynamics of the PTF's inventory process.

Corollary 3.3. *The inventory process under the optimal market making, $(I_t)_{t\in[0,T]}$, follows the linear stochastic differential equation*

$$\begin{aligned} dI_t &= \left(\frac{p(\pi^S(t)-\pi^B(t))}{2(1-c\alpha(t))} + \frac{2(\pi^B(t)+\pi^S(t))\alpha(t)}{1-c\alpha(t)} I_{t-}\right) c\, dt \\ &\quad - \frac{p-2\alpha(t)I_{t-}}{2(1-c\alpha(t))} c\,(dN_t^B - \pi^B(t)dt) + \frac{p+2\alpha(t)I_{t-}}{2(1-c\alpha(t))} c\,(dN_t^S - \pi^S(t)dt), \end{aligned} \tag{21}$$

for all $t \in [0, T]$. The expected inventory level is given by

$$\begin{aligned} \mathbb{E}[I_t] &= I_0 \exp\left(\int_0^t \frac{2c(\pi^B(u)+\pi^S(u))\alpha(u)}{1-c\alpha(u)} du\right) \\ &\quad + \int_0^t \frac{cp(\pi^S(u)-\pi^B(u))}{2(1-c\alpha(u))} \exp\left(\int_u^t \frac{2c(\pi^B(v)+\pi^S(v))\alpha(v)}{1-c\alpha(v)} dv\right) du. \end{aligned} \tag{22}$$

Corollary 3.3 shows that when the buy and sell orders arrive at the same intensity, i.e., $\pi^B(t) \equiv \pi^S(t)$ for all $t \in [0, T]$, the PTF's expected inventory converges to 0 as time moves forward. Mathematically, this is because the exponent of the first term in (22) is the integral of a strictly negative function. It reflects the effectiveness of the PTF's inventory control strategy to avoid paying high overnight costs. Furthermore, if the initial inventory $I_0 = 0$, then the expected inventory stays at 0 at all times. This means that the PTF has, on average, a neutral inventory position at any point in time. We know from (16) that $\alpha(t)$ only depends on the sum of arrival rates $\pi^B(t) + \pi^S(t)$. Hence, in the case of an asymmetric market, i.e., $\pi^B(t) = \pi_0(t) + \epsilon(t), \pi^S(t) = \pi_0(t) - \epsilon(t)$ for some $\epsilon(t) \in (-\pi_0(t), \pi_0(t))$, the expected inventory of the market maker is given by

$$\mathbb{E}[I_t] = I_0 \exp\left(\int_0^t \frac{4c\pi_0(u)\alpha(u)}{1-c\alpha(u)} du\right) - \int_0^t \frac{cp\,\epsilon(u)}{1-c\alpha(u)} \exp\left(\int_u^t \frac{4c\pi_0(v)\alpha(v)}{1-c\alpha(v)} dv\right) du.$$

Recall that $\alpha(t) < 0$ for $t \in [0, T]$, and its value only depends on $\pi_0(t)$ and not on $\epsilon(t)$. Compared with the case of a symmetric market, i.e., $\epsilon(t) \equiv 0$, the expected inventory of the market maker at any point in time is always higher if sell orders arrive more frequently, i.e., $\epsilon(t) < 0$, and always lower if buy orders arrive more frequently, i.e., $\epsilon(t) > 0$. The above considerations highlight the importance of inventory control through price determination in markets with asymmetric arrivals. If buy and sell orders arrive at the same rate, then the optimal price policy needs to guarantee that the PTF's inventory mean reverts to a flat position. This is no longer the case in an asymmetric market. Even if the PTF starts with zero inventory, the optimal price policy implies that the inventory builds up if sell orders arrive more frequently than than buy orders.

Even though $\alpha(t)$ does not admit a closed-form representation, we can still analyze its properties implicitly via the nonlinear equation (16) which defines it.

Lemma 3.4. *The function $\alpha(t)$ defined in* (16) *is strictly decreasing over the interval $[0, T]$. Suppose $\pi^B(t) \equiv \pi^B$ and $\pi^S(t) \equiv \pi^S$ are positive constants. For fixed $t \in [0, T)$ and $\lambda > 0$, $|\alpha(t)|$ is strictly decreasing with $\pi^B + \pi^S$; while for fixed $t \in [0, T]$ and π^B, π^S, $|\alpha(t)|$ is strictly increasing with λ.*

We know from Theorem 3.2 that the function $\alpha(t)$ plays a pivotal role in the PTF's price policy and value functions. Lemma 3.4 states that the PTF exhibits different trading behavior as time progresses. Such a behavior is reflected, for example, in the evolution of the Buy & Sell region of the PTF's inventory (Figure 4), which highlights a key feature of optimal market making: the size of the Buy & Sell region increases as the time remaining until the day's end $(T - t)$ increases. Intuitively, this can be understood from the fact that the shadow cost of the end-of-day inventory will be lower if the PTF has more time to build or offload its inventory, i.e., to execute multiple round-trip trades, before the day's close. As the end of the trading day approaches, the PTF may need to stop trading with sellers if it has an excessive long position or with buyers if it has an excessive short position. For this reason, we observe that both the No Buy region (L_t^1, ∞) and the No Sell region $(-\infty, L_t^2)$ "grow" as the end of the day approaches . We formalize these statements in the following

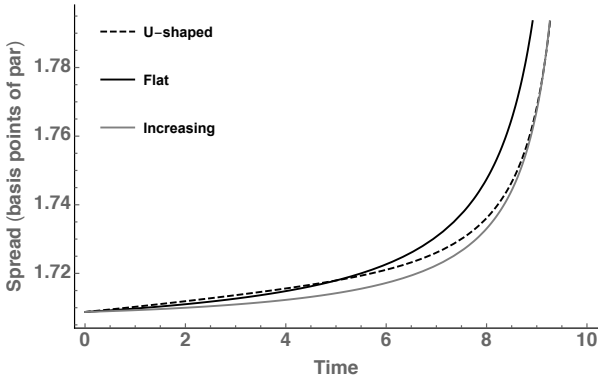


Figure 6: **Optimal Bid-Ask Spread.** This figure plots the optimal bid-ask spread, $a_t^* - b_t^*$, using the demand and supply functions specified in Figure 1. Define $\lambda_0 := 0.02$ per \$100 million par. (a) Flat benchmark (solid black): $\pi^B(t) = \pi^S(t) \equiv 10$; (b) U-shaped (dashed black): $\pi^B(t) = \pi^S(t) = 7 + 0.36(t-5)^2$; (c) Increasing (solid gray): $\pi^B(t) = \pi^S(t) = 5 + t$. The spreads at time T are all equal to $\frac{1+2c\lambda}{1+c\lambda}p$. In all cases, as the time until the end of the day increases, the bid-ask spread approaches p (the spread when there are no end-of-day inventory costs) under all parameter settings.

corollary.

Corollary 3.5. *The upper inventory boundary $(L_t^1)_{t\in[0,T]}$ is positive and strictly decreasing, while the lower inventory boundary $(L_t^2)_{t\in[0,T]}$ is negative and strictly increasing. In particular, $L_t^1 = -\frac{p}{2\alpha(t)}$ and $L_t^2 = \frac{p}{2\alpha(t)}$ for all $t \in [0,T]$ (see Figure 4).*

If the PTF starts off within the Buy & Sell region, then it will try to stay within this region to avoid one-sided trading. It does so by adjusting its prices to encourage larger sized trading orders in its favorable direction. Since the arrival rate of orders is constant, such a market making behavior moves the PTF's inventory towards zero. Notice that the resulting effect is essentially the same as trading more frequently when the PTF's inventory is unbalanced - a phenomenon often observed in practice.

The larger the active trading region, the more aggressively the PTF can trade because it is less concerned about the end-of-day inventory cost. This is also reflected in the price policy functions. Indeed, we see from Theorem 3.2 that the optimal bid and ask price policy functions are linear in the inventory level, with "slope" given by $\alpha(t)/(1 - c\alpha(t))$. Thus, Lemma 3.4 asserts that the sensitivity of bid and ask prices to the inventory level becomes weaker as the time remaining until the day's end increases.

Moreover, the bid-ask spread features a nontrivial temporal pattern, as a result of the PTF's market making activities. In particular, it follows from (18) and (19) that the bid-ask spread is given by

$$a_t(S,i) - b_t(S,i) = \frac{1 - 2c\alpha(t)}{1 - c\alpha(t)} p, \tag{23}$$

which is independent of the inventory level and the fundamental price. This result contrasts with O'Hara and Oldfield (1986), who consider an optimal market making problem in which each day consists of n trading periods and the market maker maximizes utility over an infinite number of trading days. In their model, the market maker profits from the bid-ask spread but faces end-of-day inventory effects. If the market maker is short, it pays a broker loan rate on borrowing overnight, but if it is long, it receives an interest rate from lending to brokers overnight.[10] The equilibrium bid-ask spread in O'Hara and Oldfield (1986) is thus influenced by overnight inventory, whereas it is not in our model. The following result follows from Lemma 3.4.

Corollary 3.6. *The bid-ask spread is strictly increasing with time. Suppose $\pi^B(t) \equiv \pi^B$ and $\pi^S(t) \equiv \pi^S$ are constants. For a fixed time $t \in [0, T)$, the bid-ask spread is strictly increasing with λ, and strictly decreasing with $\pi^B + \pi^S$.*

Observe that the optimal bid-ask spread depends on order arrivals only through $\alpha(t)$, which in turn depends on the total arrival rate $\pi^B(t) + \pi^S(t)$, but not on $\pi^B(t)$ and $\pi^S(t)$ individually. Hence, whether or not arrival rates are symmetric does not affect the optimal bid-ask spread. However, a higher arrival rate towards the end of the trading day implies narrower bid-ask spreads, because it provides more opportunities for the PTF to manage its inventory (see Figure 6).

The economic intuition for the results in Corollary 3.6 is as follows: as the market becomes more active, i.e., buy and sell orders arrive more frequently, the bid-ask spread at a fixed time before day's close narrows. Recall that the overnight inventory cost λ only contributes to widen the bid-ask spread at T, thus a more active market makes the end-of-day inventory

[10] O'Hara and Oldfield (1986) model the overnight market as a competitive repurchase market in which short parties borrow securities and lend cash, while long parties lend securities and borrow cash. In contrast, we do not model the overnight market.

cost fade away faster as the time to the day's close increases. Moreover, as the end of the day approaches, the growing concern about the inventory cost discourages the PTF from trading actively. Hence, it sets a wider 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. This trading behavior reflects the tradeoff faced by the PTF between making trading profits and holding a non-zero inventory at the end of the day. It follows directly from the ordinary differential equation (16) that the higher the order arrival rates π^B and π^S, the faster $\alpha(t)$ will decrease as $T-t$ increases. As a result (see also Corollary 3.6), the bid-ask spread will also narrow faster. Compared with the flat and increasing pattern of arrival rates, the U-shaped pattern tends to have the widest bid-ask spreads and lower arrival rates at intermediate times. As the end of the day approaches, arrival rates are the lowest under the flat arrival pattern. This also explains why the spread curves associated with the U-shaped and flat arrival patterns cross, resulting in the widest bid-ask spread under a flat arrival rate.

Our prediction on the temporal pattern of bid-ask spreads contrasts with that of Ho and Stoll (1981). In their model, the monopolistic market maker is averse to both fluctuations in the market price of its inventory and to uncertainty in the execution time of transactions. By contrast, our PTF is risk neutral with respect to the cash proceeds of transactions, and only pays a cost on the size of its overnight inventory. At any point in time, it does not face any aversion with respect to the value of inventory at the terminal time.[11]

To understand the impact of a high overnight inventory cost, we use the explicit formulas in (18), (19) and (23) to study the ask and bid price policy functions, as well as the bid-ask spread. As the end-of-day inventory cost increases, i.e. λ gets larger, we obtain that $|\alpha(t)|$ also becomes larger via a standard comparison argument in (16). This in turn implies that both $a_t^*(S,i)$ and $b_t^*(S,i)$ become more sensitive to the inventory level i, and the bid-ask spread widens. In the limiting case $\lambda \to \infty$, $a_t^*(S,i) \to S + p - \frac{1}{c}i$ and $b_t^*(S,i) \to S - p - \frac{1}{c}i$, and the bid-ask spread tends to its maximum value, $2p$. Given that the reservation prices

[11] We also note that Ho and Stoll (1981)'s prediction holds only for the case of a sufficiently short time horizon, whereas our prediction holds true for any time horizon.

for the buyers and sellers also differ by $2p$, we know that the Buy & Sell region for the PTF is then empty at all times. The same conclusion can be drawn from Corollary 3.5, which implies that $L_t^1 = L_t^2 = 0$, i.e., the PTF will not be willing to trade at all if its initial inventory is already 0.

3.4 Endogenous Temporary Price Impact

Price impact arises endogenously in our model, and can be analytically quantified and explained. As can be directly seen from Theorem 3.2, the relationship between the optimal ask/bid price and the inventory level is linear, and the slope coefficient of this linear relationship is time-dependent. In particular, from the optimal ask and bid price policy functions, we deduce that the mid-quote price, $P_t := \frac{1}{2}(a_t^* + b_t^*)$, is given by

$$P_t = S_t + \frac{\alpha(t)}{1 - c\alpha(t)} I_t. \tag{24}$$

Taking the differential of the above expression with respect to t, we obtain

$$dP_t = dS_t - \frac{c(\pi^B + \pi^S)\alpha^2(t)}{(1 - c\alpha(t))^3} I_t dt + \frac{\alpha(t)}{1 - c\alpha(t)} dI_t. \tag{25}$$

Equation (25) describes the dynamics of the mid-quote price. It indicates that the instantaneous change of the mid-quote price is negatively related to the current level of the PTF's inventory. Moreover, the price jumps every time a trade is executed. Specifically, if a buy order is fulfilled at time t i.e. $dI_t = -Q^B(S_t, a_t^*) < 0$, then the mid-quote price will increase by $\frac{\alpha(t)}{1-c\alpha(t)} dI_t > 0$, which is proportional to the size of the trade taking place at time t, $Q^B(S_t, a_t^*)$. Likewise, if there is a sell order coming at time t, then the mid-quote price will decrease by $-\frac{\alpha(t)}{1-c\alpha(t)} dI_t$, in an effort to invite a larger sized trade with a buyer to balance inventory. Because the mid-quote price moves by $-\frac{\alpha(t)}{1-c\alpha(t)}$ multiples of the trading size whenever a trade is executed, we refer to this ratio as the *price impact coefficient*. In conjunction with Lemma 3.4, we deduce that the price impact coefficient tends to be larger as time moves towards the day's end, a phenomenon we document in our empirical analysis

(see Section 5).[12]

The time varying and endogenous nature of the price impact coefficient and bid-ask spreads is a distinguishing feature of our model, and is driven by the forward looking nature of the end-of-day inventory cost. By contrast, in 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. Moreover, a unit trading size is considered in their model, hence an identical price impact is observed in every trade. Models with information asymmetries, such as Glosten and Milgrom (1985), can also generate an endogenous price impact from trades (of a unit size) when the traded asset has a binary valued fundamental price (being either high or low). In contrast to our current setup, their price impact mechanism is generated from a Bayesian learning mechanism: namely, an incoming buy (sell) order signifies a stronger belief that the underlying security has a high (low) fundamental price, and leads the market maker to adjust prices for the next step according to this updated belief. Moreover, Glosten and Milgrom (1985) also predict that, as more trades take place, both ask and bid price will converge to the true fundamental price, so bid-ask spreads become narrower, not wider, as time progresses.

4 Comparative Statics

We study the sensitivity of price stability and welfare measures to the severity of the overnight inventory cost λ, as well as to the arrival rates $\pi^B(t)$ and $\pi^S(t)$ of buy and sell orders. Section 4.1 provides comparative statics for the price volatility process. Section 4.2 studies the dependence of the buyer/seller's surplus, and the PTF's value, on the inventory cost and order arrival frequency. Throughout the section, we assume that the PFT starts with zero inventory, i.e., $I_0 = 0$.

[12] 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, 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 the day's end approaches, which, according to Admati and Pfleiderer (1988), should cause buyers and sellers to arrive less frequently.

4.1 Price Stability

The overnight inventory cost steepens the market maker's price impact function. We therefore expect price trajectories to be more volatile when market makers face higher inventory costs. We measure price stability through the quadratic variation of the mid-quote price, given by $P_u = \frac{1}{2}(a_u^* + b_u^*)$ at time u, during the period $[0, t]$, and denote it by $QV_t \equiv \langle P \rangle_t$. Formally, for an equidistant partition of the time interval $[0, t]$ into n subintervals, we have[13]

$$QV_t = \lim_{n\to\infty} \sum_{k=1}^{n} \left(P_{\frac{k}{n}t} - P_{\frac{(k-1)}{n}t} \right)^2.$$

Equations (1) and (24) imply that the PTF's market making activities subject to the overnight costs will increase the quadratic variation of the mid-quote price. In particular, the quadratic variation of the mid-quote price stays the same as that of the fundamental price if overnight costs are absent. More specifically, we have

Proposition 4.1. *Set $\pi^B(t) = \pi_0(t) + \epsilon(t), \pi^S(t) = \pi_0(t) - \epsilon(t)$ for some $\epsilon(t) \in (-\pi_0(t), \pi_0(t))$. Then the expected quadratic variation during the period $[0, t]$, for any $t \in [0, T]$, is given by*

$$\mathbb{E}[QV_t] = \sigma^2 t + \int_0^t \left(\frac{\pi_0(u) c^2 \alpha^2(u)}{2(1 - c\alpha(u))^4} \left(p^2 + 4\alpha^2(u) \mathbb{E}[I_u^2] \right) - \frac{2\epsilon(t)\, c^2 p \alpha(u)}{(1 - c\alpha(u))^2} \mathbb{E}[I_u] \right) du,$$

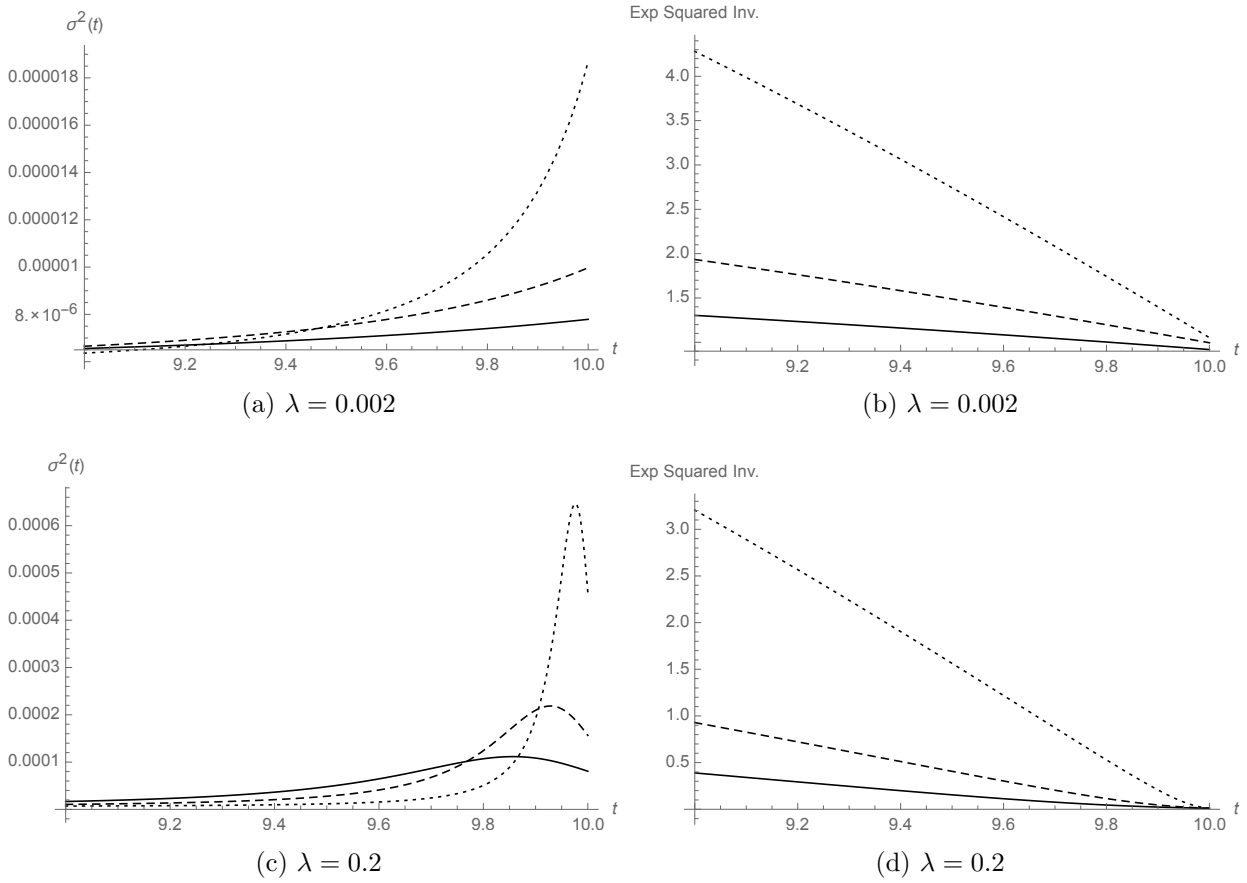
where

$$\mathbb{E}[I_t^2] = \int_0^t \left(\frac{\pi_0(u)(cp)^2}{2(1 - c\alpha(u))^2} + 2 \left(\pi_0(u) \frac{c^2 p \alpha(u)}{(1 - c\alpha(u))^2} - \epsilon(u) \frac{cp}{1 - c\alpha(u)} \right) \mathbb{E}[I_u] \right) \times \exp\left(\int_u^t \frac{2\pi_0(v) c\alpha(v)(2 - c\alpha(v))}{(1 - c\alpha(v))^2} dv \right) du. \quad (26)$$

Quadratic variation is a natural measure for price volatility. Specifically, it allows us to introduce an instantaneous squared volatility measure

$$\sigma^2(t) := \frac{d}{dt} \mathbb{E}[QV_t] = \sigma^2 + \frac{\pi_0(t) c^2}{2} \frac{\alpha^2(t)}{(1 - c\alpha(t))^4} \left(p^2 + 4\alpha^2(t) \mathbb{E}[I_t^2] \right) - \frac{2\epsilon(t)\, c^2 p \alpha(t)}{(1 - c\alpha(t))^2} \mathbb{E}[I_t]. \quad (27)$$

[13] This limit can be relaxed to any partition as long as the length of the longest subinterval eventually converges to 0.



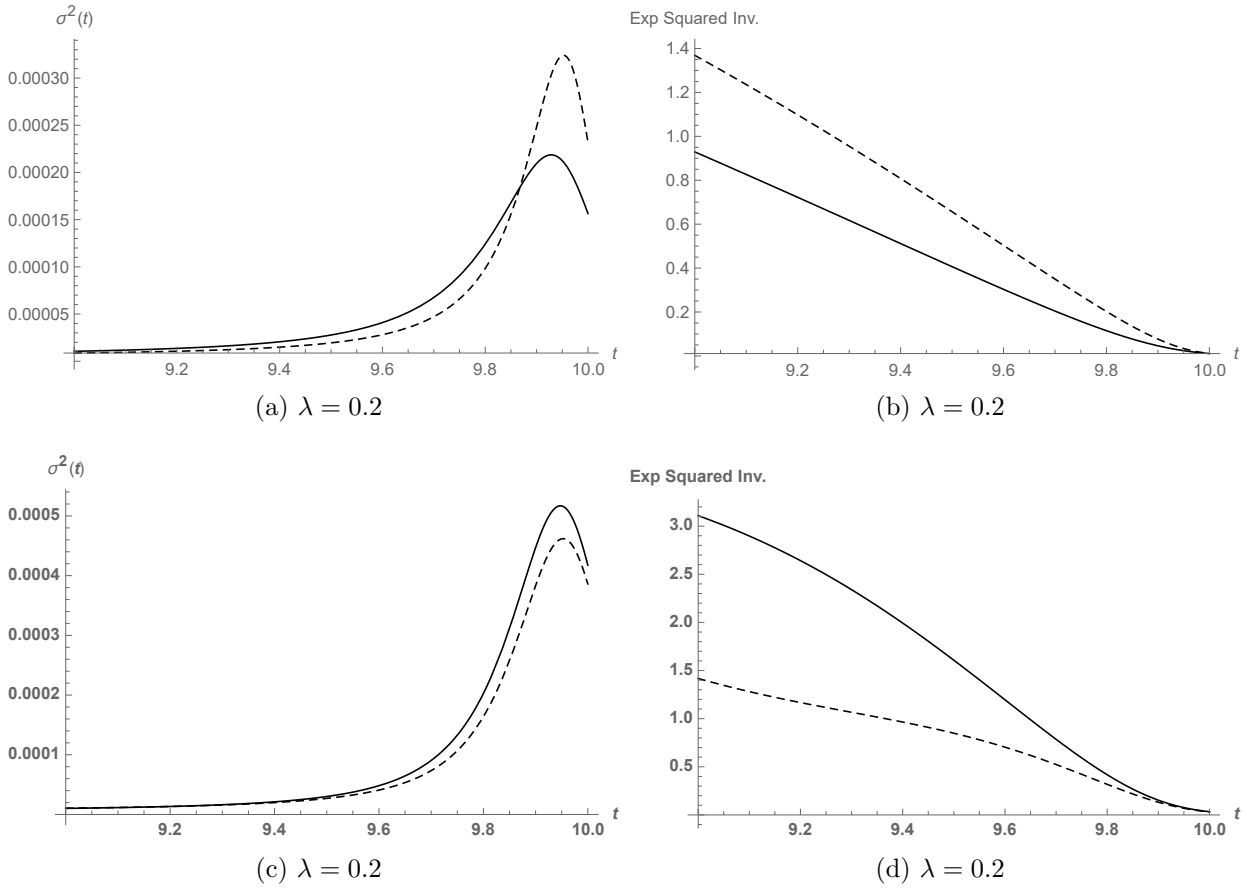
(a) $\lambda = 0.002$ (b) $\lambda = 0.002$

(c) $\lambda = 0.2$ (d) $\lambda = 0.2$

Figure 7: **Squared Volatilities and Squared Inventories for Symmetric Order Arrival Rates.** In all panels, dotted lines are for order arrival rates $\pi^B(t) = \pi^S(t) = 30$, dashed lines are for $\pi^B(t) = \pi^S(t) = 10$, and solid lines are for $\pi^B = \pi^S = 5$. Panels (a) and (c) plot, respectively, the squared volatility $\sigma^2(t)$ for the low overnight inventory cost $\lambda = 0.002$, and the high overnight cost $\lambda = 0.2$, during the last 10% of the trading day. Panels (b) and (d) plot the corresponding expected squared inventory of the PTF under the low, respectively high, inventory cost. We set the annualized volatility of the fundamental price process $\sigma = 3.75\%$.

We conduct a comparative statics analysis of the squared volatility measure $\sigma^2(t)$ with respect to the severity of the overnight inventory cost λ (per \$100 million par) and the arrival rate π. We use the demand and supply functions given in Figure 1.

Our analysis shows that both for the case of symmetric and time varying asymmetric arrival rates (see Figures 7 and 8), the overnight inventory cost plays two distinct roles: it generates an intensifying price impact intraday and an aversion towards holding excessive inventory near the day's close. For a low level of overnight costs, e.g. $\lambda = 0.002$, the aversion towards inventory holdings is dominated by the phenomenon of intensifying price impact,



(a) $\lambda = 0.2$ (b) $\lambda = 0.2$

(c) $\lambda = 0.2$ (d) $\lambda = 0.2$

Figure 8: **Squared Volatilities and Squared Inventories for Time Varying Arrival Rates.** In all panels, the overnight inventory cost λ is fixed at 0.2 (per \$100 million par). Panels (a) and (c) plot, respectively, the squared volatility $\sigma^2(t)$ during the last 10% of the trading day. Panels (b) and (d) plot the corresponding expected squared inventory of the PTF. In panels (a) and (b), the solid lines are for the constant arrival rate $\pi^B(t) = \pi^S(t) \equiv 10$, the dashed lines are for arrival rates $\pi^B(t) = \pi^S(t) = t + 5$. In panels (c) and (d), the solid lines are for arrival rates $\pi^B(t) = t + 5$, $\pi^S(t) = 15 - t$, and the dashed lines are for arrival rates $\pi^B(t) = 7 + 0.36(t-5)^2$, $\pi^S(t) = 13 - 0.36(t-5)^2$. We set the annualized volatility of the fundamental price process $\sigma = 3.75\%$.

and thus the instantaneous squared volatility process increases over time for all three levels of arrival rates. In contrast, if overnight inventory costs are high, e.g. $\lambda = 0.2$, the aversion towards holding inventory becomes the prevailing force and leads the PTF to drastically reduce its trading volume near the day's close. This brings down the instantaneous squared volatility consistently across the levels of arrival rates.

Figure 7 shows the dynamics of the instantaneous squared volatility process defined in (27), and of the second moment of the inventory process for the case of symmetric time homogeneous arrival rates. The plot in Panel (a) of Figure 7, shows that the squared

volatility increases over time, consistent with the bid-ask spread pattern highlighted in Figure 6. However, different from the bid-ask spread, the instantaneous squared volatility is not monotone in the order arrival rates π^B, π^S. In particular, as π^B, π^S increase, the squared volatility may decrease if the trading time is far from the day's close, and increase if the trading time is near the day's close. This phenomenon can be attributed to the complex relation between order arrival rates and the dynamics of price trajectories. On the one hand, higher order arrival rates lead to higher trading volumes, and hence generate more volatility. On the other hand, as the intensities of order arrivals get higher, the impact of the overnight costs (see Theorem 3.2, Lemma 3.4, and Section 3.4) is attenuated faster as the trading time gets further away from the day's close, reducing both price impact and price volatility. The overall trend depends on which of these two counteracting forces prevail: our analysis shows that as the market becomes more active, the squared volatility decreases in earlier times of the day but increases at later times of the day. Panel (b) of Figure 7 reflects the PTF's aversion towards holding excessive inventory. Both in the case of symmetric and asymmetric arrival rates, as the end of the trading day approaches the size of the PTF's expected squared inventory decreases.

Panels (c) and (d) of Figure 7 show the dynamics of volatility and expected squared inventory for higher overnight costs. Interestingly, apart from the increase in price volatility of Panel (c) compared with Panel (a), we observe that the squared volatility always drops near the day's close. Comparing Panel (b) with Panel (d), we deduce that this difference can be explained by the PTF's aversion towards holding excessive inventory near the day's close. For large values of λ, the PTF chooses to trade very little towards the close, preferring to maintain a near zero inventory rather than profiting from trade executions. As a consequence of the very low trading activity, the price volatility declines.

Figure 8 shows that the temporal pattern of order arrivals strongly affects inventory management and price volatility. The top panels of the figure indicate that, as order arrival rates increase towards the end of the trading day, the PTF's expected squared inventory and price volatility also increase. Despite the symmetry in the arrival of orders, a higher intensity increases the variance of inventory positions and thus results in higher instantaneous

squared volatility. The bottom panels of the figure deal with the case of time varying, but asymmetric, arrival rates. We consider two temporal patterns. In the first case, buy orders increase linearly over time while sell orders decrease linearly over time. In the second case, buy orders arrive at a high intensity at the beginning and end of the trading day but with a moderate intensity during the trading day, while sell orders follow a reverse U-shaped pattern. In both cases, the PTF has a higher imbalance in its inventory relative to the case of symmetric arrivals, which results in increased volatility and inventory compared to the case of symmetric arrivals. Noticeable in the case of U-shaped arrival rates, the expected inventory and volatility are lower. This can be understood from the fact that the total level of imbalance is lower as the expected inventory mean-reverts twice towards zero before the end of the day, hence making it easier for the PTF to control inventory.

The above analysis is based on analytical results derived in Proposition 4.1. Next, we use stochastic simulation to examine a specific sample path in which a larger number of sell orders, relative to buy orders, is realized. As discussed earlier in the paper, asymmetric arrival of buy/sell orders has been identified in the empirical literature as a typical characteristic of flash events. We simulate buy and sell order arrivals, and then analyze the effects of two different end-of-day inventory costs: $\lambda = 0.002$; and $\lambda = 0.2$.

As shown from the simulated trajectories of the mid-quote price reported in Figure 9, higher end-of-day inventory costs amplify the downward pressure on prices caused by the order imbalance. The figure suggests a strong link between the magnitude of the overnight inventory cost and the volatility of price trajectories. As evident from the figure, managing inventory is more challenging when arrival rates are asymmetric, and the failure to do so effectively may result in large overnight positions (compare to the case when $\lambda = 0$ in the top panels). This explains why the price exhibits abrupt downward jumps when the overnight cost is high. By contrast, the PTF accumulates, on average, zero inventory if order arrivals are symmetric. As a result, prices can be used to manage inventory without inducing big volatility spikes towards the end of the trading day (compare price trajectories in the left panels of Figure 9).

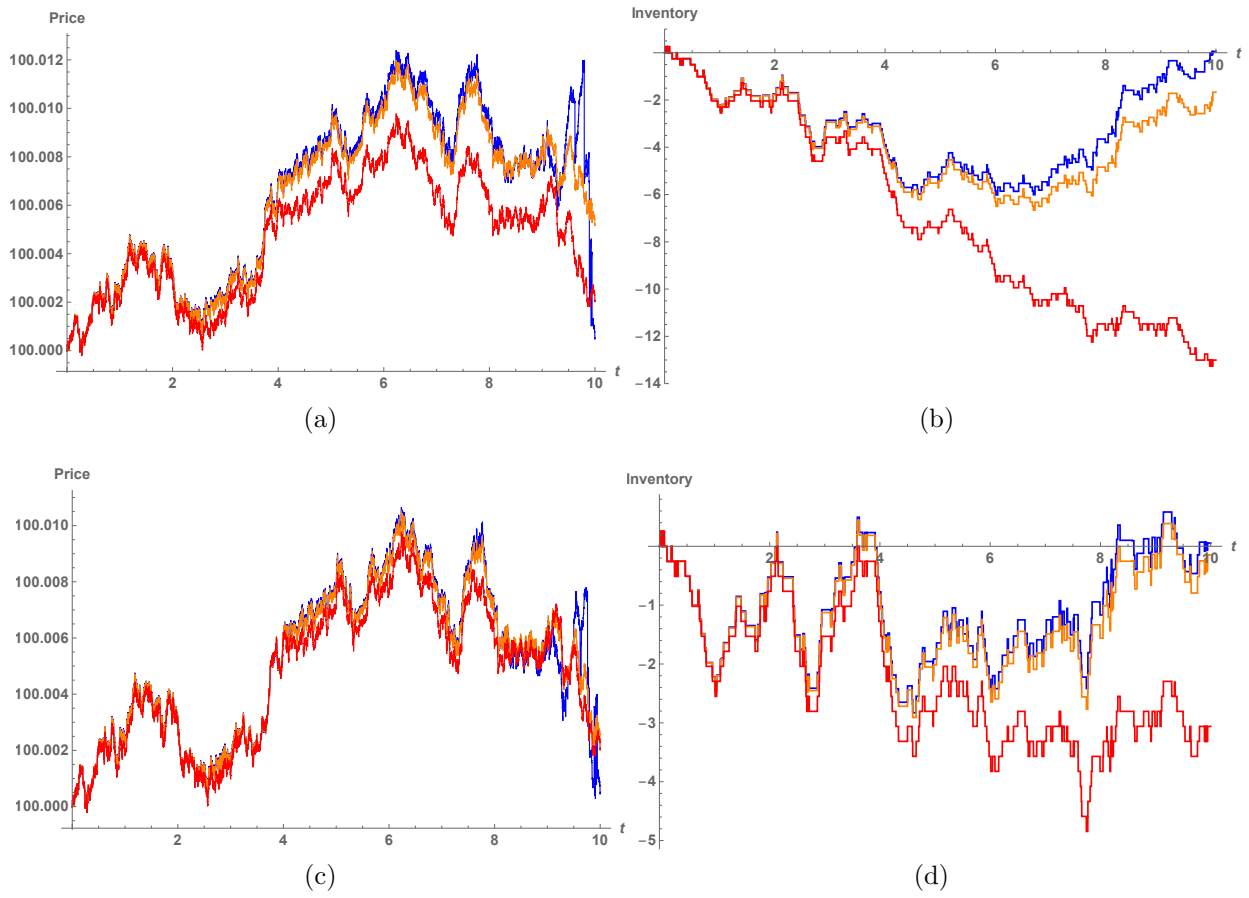


Figure 9: **Simulated Price and Inventory Trajectories for Different Overnight Inventory Costs.** This figure plots the trajectories of the mid-quote price and inventory under the demand and supply functions specified as in Figure 1 for two different values of λ: $\lambda = 0.002$ (per \$100 million par) in the orange line; $\lambda = 0.2$ (per \$100 million par) in the blue line. The fundamental price, which can be viewed as the mid-quote price when $\lambda = 0$, is shown in red. Top panels: asymmetric arrival rates $\pi^B(t) = 12, \pi^S(t) = 8$. Bottom panels: symmetric arrival rates $\pi^B(t) = 10, \pi^S(t) = 10$. We set the annualized volatility of the fundamental price process $\sigma = 3.75\%$. For each choice of λ, we consider the same arrival sequence of buy and sell orders. Noticeably, the larger the λ, the more volatile the price trajectory, especially near the day's close.

4.2 Welfare of Market Participants

We now conduct a comparative statics analysis on the surplus of the buyers and sellers, and the value of the PTF. Throughout this section, we set $\pi^B(t) = \pi^S(t) = \pi_0$, and $I_0 = 0$. The surplus for a buyer or a seller is defined as follows. If there is a buy order arriving at time t, then we measure the surplus for this order as the area of the region bounded by the traded

quantity $Q^B(S_t, a_t^*)$, the demand curve $Q^B(S_t, x)$, and the reservation price $S_t + p$:

$$\int_{a_t^*}^{S_t+p} [Q^B(S_t, a_t^*) - Q^B(S_t, x)]dx = \frac{c}{2}(S_t + p - a_t^*)^2.$$

Similarly, the surplus for a sell order arriving at time t is measured by $\frac{c}{2}(b_s^* - S_t + p)^2$. If $\lambda = 0$, the PTF acts as a period-by-period monopolist or monopsonist and implements the constant ask and bid price policy functions $S \pm \frac{1}{2}p$ (see footnote 9). This yields a constant surplus per trade for the buyers/sellers equal to $\frac{c}{8}p^2$.

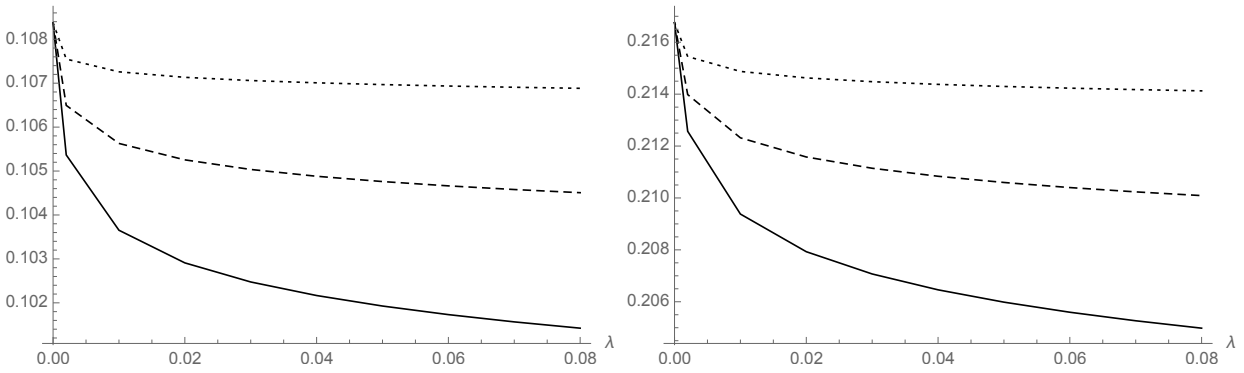
To assess the effect of the end-of-day inventory cost λ on surplus, we compare the surplus per trade for buy and sell orders with those in the benchmark case $\lambda = 0$. Recall that in contrast to the benchmark case, a positive inventory cost λ will generate price impact from trades. This in turn leads to a stochastic price process (which may be higher or lower than the corresponding price when $\lambda = 0$) and random surplus in every trade. We thus use the ratio between the expected total surplus for buy or sell orders and the expected number of buy or sell trades during the whole period to measure the surplus per trade for each type of order. These are respectively defined as

$$\begin{aligned} ASurplus^B :=& \frac{c}{2\pi_0 T}\mathbb{E}\left[\int_0^T (S_t + p - a_t^*)^2 dN_t^B\right], \\ ASurplus^S :=& \frac{c}{2\pi_0 T}\mathbb{E}\left[\int_0^T (b_t^* - S_t + p)^2 dN_t^S\right]. \end{aligned}$$

We also compute the ratio between the PTF market maker's value function at the end of the day and the expected number of trades, i.e.,

$$AValue^M := \frac{v(0, 0, S, 0)}{2\pi_0 T}.$$

This quantity measures the PTF's realized utility per trade under the price trajectory determined by the optimal trading actions. A direct application of Theorem 3.2 yields the following



(a) Buyer/Seller's Surplus per Trade (b) PTF's Value per Trade

Figure 10: **Plots of Average Buyer or Seller, and PTF's Value per Trade.** In all panels, dotted lines refer to $\pi_0 = 30$, dashed lines refer to $\pi_0 = 10$, and solid lines refer to $\pi_0 = 5$. Panel (a): buyer or seller surplus per trade; Panel (b): PTF's value per trade.

Proposition 4.2. *The surplus per trade for buyers and sellers is given by*

$$ASurplus^B = ASurplus^S = \frac{c}{2T}\int_0^T \left(\frac{p^2}{4(1-c\alpha(t))^2} + \frac{\alpha^2(t)}{(1-c\alpha(t))^2}\mathbb{E}[I_t^2]\right)dt,$$

where $\mathbb{E}[I_t^2]$ *is given in* (26)*. The value per trade of the PTF is given by*

$$AValue^M = \frac{c}{4T}\int_0^T \frac{p^2}{1-c\alpha(t)}dt.$$

We analyze the changes in realized surplus for buyers/sellers, and the value of the PTF, relative to the benchmark case of $\lambda = 0$, as we vary the overnight cost parameter λ and the order arrival rate parameter π.

Figure 10 reports the average surplus per trade for buyers or sellers, and the value of the PTF, for different choices of λ and π. First, the two panels suggest that, for any $\lambda > 0$, higher arrival rates improve the surplus per trade of buyers/sellers and the value of the PTF: the dotted curves ($\pi_0 = 30$) lie strictly above the dashed ($\pi_0 = 10$) and solid ($\pi_0 = 5$) ones. Second, because all confidence bands are downward sloping, larger overnight inventory costs λ have a negative effect on the buyer/seller's surplus and PTF's value. Third, increases in the overnight inventory cost relative to the benchmark case $\lambda = 0$ have the greatest negative impact on buyer/seller's surplus and PTF value if the arrival rates are low ($\pi_0 = 5$).

Altogether, these results suggest that while the PTF's overnight inventory cost has a negative effect on the aggregate welfare, its magnitude depends crucially on market activity. When buyers and sellers arrive intraday at a high intensity, the buyer/seller's surplus and PTF's value per trade for small $\lambda > 0$ is not too different from the surplus per trade in the benchmark case $\lambda = 0$. In contrast, the buyer/seller's surplus and PTF's value per trade in markets with low trading activity is notably more sensitive to changes in the overnight inventory cost. The loss in surplus per trade when λ goes from zero to a slightly positive value is especially striking for buyers and sellers. This suggests that in markets with low trading activity, the PTF is posting higher asks and lower bids per trade relative to the static monopoly and monopsony setting. Panel (b) in the figure shows, however, that these widened bid-ask spreads are not sufficient to compensate the PTF for the incurred inventory costs, as the PTF is unable to achieve the utility realized in the benchmark case. The PTF thus appears unable to fully transfer the overnight inventory cost to buyers and sellers.

5 Empirical Evidence from Treasury Data

Our theory has implications for markets that are intermediated by liquidity providers with overnight inventory constraints. Based on the findings of the Joint Staff Report (2015), we expect the BrokerTec interdealer ECN for U.S. Treasury securities to represent such a market. Using non-public data from BrokerTec, the Joint Staff Report (2015) found that PTFs have become leading liquidity providers in the interdealer Treasury market.[14] Moreover, the report shows that PTFs carry significantly less inventory overnight than bank-dealers, implying a strong end-of-day inventory aversion. We thus use BrokerTec data to analyze whether price and liquidity dynamics are consistent with the activities of liquidity providers facing overnight inventory constraints. We describe the data set in Section 5.1. We empirically test the model prediction on the relation between price changes and inventory changes in Section 5.2. We analyze the model predictions on intraday patterns of bid-ask

[14] Activity in the Treasury market is roughly split between the interdealer segment and the dealer-to-customer segment (Brain et al. (2018)). Participation in the interdealer segment was historically limited to dealers (hence the name), but expanded to include PTFs and hedge funds in the mid 2000s.

spread in Section 5.3. Section 5.4 provides further empirical analysis on day-of-week effects and Section 5.5 assesses possible alternative explanations for our empirical findings.

5.1 Description of Data Set

Our data is from the BrokerTec interdealer ECN. Nearly all interdealer trading of on-the-run U.S. coupon securities occurs via electronic platforms among which BrokerTec accounts for about 80% of trading.[15] In contrast to the Joint Staff Report (2015), our BrokerTec data set is anonymized to protect the identities and strategies of participating firms. Nonetheless, the data set is very rich, containing every electronic message submitted to the trading platform for the 1,376 trading days between January 2, 2013 and June 30, 2018, time-stamped to the millisecond (through Tuesday, March, 31, 2015) or microsecond (from Wednesday, April, 1, 2015), and representing the same information available to computer-based traders in real-time. The on-the-run 10-year Treasury note is the focus of our empirical analysis, following the Joint Staff Report (2015). Over our sample period, daily trading volume in the note averages $40.7 billion and the daily number of trades averages just over 16,000.

The BrokerTec platform operates as an electronic limit order market, in which buyers are matched to sellers without human intervention (Fleming et al. (2018)). Traders send in orders (minimum size $1 million par value) that can be aggressive (liquidity taking or market orders) or passive (liquidity providing or limit orders). Limit orders remain in the book until canceled or lifted. In addition to shown limit orders, BrokerTec allows traders to submit iceberg orders, which conceal part of the total quantity the trader is willing to transact at the posted price. BrokerTec also features a workup protocol by which each market order triggers a temporary phase during which market participants can transact additional volume at the same price as the original order (Fleming and Nguyen (2019)).

We use the BrokerTec message data to reconstruct the limit order book. The reconstructed limit order book data allows us to calculate bid-ask spreads at the inside tier of the

[15] Electronic brokers account for 87% of trading in on-the-run coupon securities that occurs through interdealer brokers (Brain et al. (2018)). According to Greenwich Associates (https://www.greenwich.com/blog/does-cme-own-us-treasury-market), based on 2017 Q4 data, BrokerTec's market share in the electronic interdealer market is 80%.

book as well as changes in the bid-ask midpoint (or mid-quote price). To analyze bid-ask spreads, we calculate the average spread for each five-minute interval across all order book snapshots within each interval. To analyze price impact, we relate changes in the bid-ask midpoint for each 30-minute interval to our measure of market maker inventory changes over the same interval. The BrokerTec platform operates 22-23 hours per day during the week, but our focus is on New York trading hours, 7:30 a.m. - 5:30 p.m., when the market is most liquid (Fleming (1997)).

Because our data set does not identify individual market participants, we proxy for market makers' aggregate inventory as the negative of the observed cumulative net volume, a measure that is often also referred to as order imbalance. Intuitively, this proxy captures the difference between the number of aggressive buy and sell orders submitted to BrokerTec. If this number is zero, then the number of aggressive buy orders (that is, buyer-initiated trades) equals the number of aggressive sell orders (seller-initiated trades), which translates into no inventory change for intermediaries as a group. By contrast, a large positive cumulative net volume or order imbalance means that more aggressive buy orders were placed than aggressive sell orders. To satisfy this demand, a representative intermediary that started the day with zero inventory is therefore interpreted as having sold short the difference, and thus carries a negative inventory position.

We expect this proxy for market maker inventory to perform well. Boehmer et al. (2018) present evidence that PTF intraday trading activity has a strong factor structure, suggesting that PTFs pursue a handful of similar strategies. This finding suggests that it is reasonable to treat many firms with highly correlated market making strategies as a single, representative firm. Furthermore, treating order imbalances as a proxy for market maker inventory has found support in the literature (see, for example, Chordia et al. (2002)). Lastly, our analysis explicitly excludes the period around the most important scheduled macroeconomic announcements, when the asymmetric information component of order flow is likely to be the highest, allowing us to focus on inventory-based interpretations of order flow.

To produce our cumulative net volume measure, we must classify trades as buys or sells. Fortunately, the side that initiated a trade is a field in the BrokerTec data set, allowing us

to classify trades directly. This is an advantage over indirect classification algorithms used in the literature.

5.2 The Relation between Prices and Inventory

Our theory predicts that inventory changes and price changes are negatively related. To see this explicitly, we derive a discrete-time version of (25). Fixing a given time increment $\Delta t > 0$, we consider the Euler scheme

$$P_{t+\Delta t} - P_t \approx S_{t+\Delta t} - S_t - \frac{c(\pi^B + \pi^S)\alpha^2(t)}{(1 - c\alpha(t))^3} I_t \, \Delta t + \frac{\alpha(t)}{1 - c\alpha(t)}(I_{t+\Delta t} - I_t). \tag{28}$$

We simplify (28) to facilitate the empirical analysis. In particular, notice that

$$\begin{cases} (S_{t+\Delta t} - S_t)/\sqrt{\Delta t} \sim N(0, \sigma^2), \\ -\dfrac{c(\pi^B + \pi^S)\alpha^2(t)}{(1 - c\alpha(t))^3} I_t \Delta t = O(\Delta t), \\ \dfrac{\alpha(t)}{1 - c\alpha(t)}(I_{t+\Delta t} - I_t) = O(1) \text{ if there are trades in period } (t, t + \Delta t]. \end{cases}$$

This gives rise to a regression equation of the form

$$P_{n\Delta t} - P_{(n-1)\Delta t} = \beta_n (I_{n\Delta t} - I_{(n-1)\Delta t}) + \gamma_n + \epsilon_n,$$

for $n = 1, 2, \ldots, \lfloor T/\Delta t \rfloor$, where ϵ_n's are i.i.d. $N(0, \sigma^2 \Delta t)$, γ_n's are constant coefficients, and β_n's are negative coefficients that are decreasing with n.

The steepening of the slope coefficients β_n intraday, i.e., as n increases and time approaches the day's close, is the distinguishing feature of our theory relative to existing market-making models with inventory constraints. To test this prediction empirically, we regress half hour price changes on our proxy for inventory changes for each half hour interval of the day, exploiting variation across days to identify the slope and intercept parameters. Specifically, for each half hour window, starting with 9 a.m. - 9:30 a.m., we obtain the mid-quote price change over that half hour for each day, as well as the inventory change over

the same half hour.[16] This results in 1,376 price and inventory changes for the 9 a.m. - 9:30 a.m. regression, one for each day in our sample.[17] The process is repeated for subsequent half hour intervals, allowing us to estimate price-inventory sensitivities over the course of the trading day. Since the BrokerTec ECN closes at 5:30 p.m., the last regression uses data over the 5 p.m. - 5:30 p.m. window. While these regressions can in principle be performed on shorter intervals, we note that very high frequency estimates would result in intervals containing zero buy and sell orders. Thus, the half hour frequency strikes a middle ground between allowing a sufficient number of buy and sell orders to arrive, and at the same time examining intraday variation in price-inventory sensitivities.

The results in Figure 11 show a statistically significant negative relationship between price changes and inventory changes for each half hour between 9 a.m. and 5:30 p.m. Most importantly, however, the figure shows a strong steepening of price-inventory sensitivities near the market close, consistent with the theoretical predictions in Section 3.4. To verify whether this steepening is statistically significant, Table 1 reports the results of t-tests on the differences between near-close (5 p.m. - 5:30 p.m.) and earlier intraday price impact coefficients. The table shows that the slope of the price-inventory relationship is not time-homogenous as predicted by existing theories, but statistically steeper toward the end of the day. Note that the left panel in Figure 11 shows that the intercepts from the regressions of price changes on the market maker's inventory changes are close to zero, suggesting that price and inventory changes are linked by an approximate linear relationship over short time windows, as described by equation (24).

5.3 The Intraday Pattern of Bid-Ask Spreads

We proceed to test the empirical implications of our model for bid-ask spreads. Corollary 3.6 asserts that bid-ask spreads are narrower intraday than near the close. This relation

[16] We deliberately start this analysis after the period surrounding the most important scheduled macroeconomic announcements, such as the employment report. These announcements, which are released at 8:30 a.m., are associated with worse liquidity and high information asymmetry (see, for example, Fleming and Remolona (1999) and Green (2004)).

[17] There are as few as 1,344 observations available for some of the half hour intervals because of early market closes, which occur around holidays.

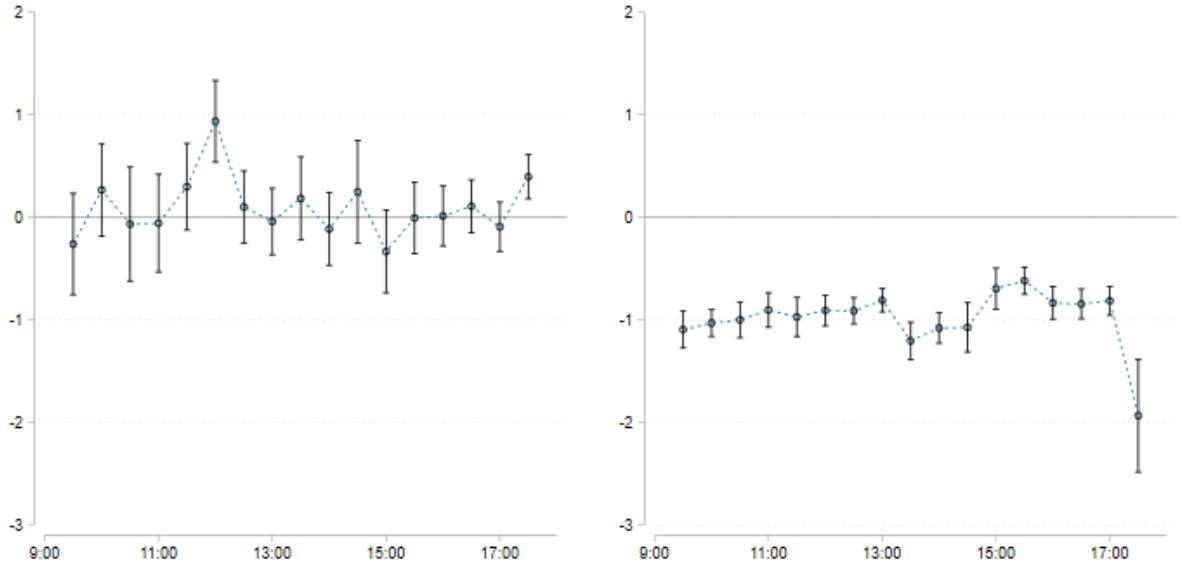


Figure 11: **The Relation between 10-Year Treasury Price and Inventory Changes.** This figure plots intercepts γ_n (left panel) and slopes β_n (right panel) from regressions of 10-year Treasury price changes on changes in market maker inventory, as proxied by the negative of cumulative net volume. For each half hour interval, starting with 9 a.m. - 9:30 a.m., we calculate the mid-quote price change over that half hour for each day, as well as the inventory change over the same half hour. We then regress the price changes on the inventory changes. Analogous regressions are run for each subsequent half hour interval of the trading day. Since the market closes at 5:30 p.m., the last regression uses price and inventory changes from 5 p.m. - 5:30 p.m. The figure also plots 99% Newey-West confidence intervals for the coefficient estimates. Prices are measured in basis points of par and inventories are measured in hundreds of millions of dollars par. The sample period is January 2, 2013 - June 29, 2018.

is also graphically illustrated in Figure 6, which shows a pattern of model-implied bid-ask spreads that are tight for most of the trading day and then widen rapidly near the close. Our analysis of BrokerTec bid-ask spreads reveals a similar pattern: The left panel of Figure 12 shows that, on average, bid-ask spreads are relatively narrow from 7:30 a.m. - 5 p.m., hovering around 1.7 basis points of par.[18] As the end of the day approaches, however, bid-ask spreads widen sharply, mimicking the rapid widening in bid-ask spreads predicted by our theoretical model and visualized in Figure 6. Moreover, the right panel of Figure 12 shows that this is an empirical regularity that is maintained on almost all of the trading days in our sample. That is, spreads near the market close (5:25 p.m. - 5:30 p.m.) are nearly always

[18]The spikes at 8:30 a.m. and 10 a.m. are associated with scheduled macroeconomic announcements at those times, the 1 p.m. spike with Treasury auction closes, and the 2 p.m. and 2:15 p.m. spikes with Federal Open Market Committee announcements.

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

This table reports the results of tests of the null hypothesis that the relationship between price changes and inventory changes is the same intraday as near the close of the trading day. Slope coefficients β_n are from regressions of 10-year Treasury price changes on changes in market maker inventory, as proxied by the negative of the cumulative net volume. For each half hour interval, starting with 9 a.m. - 9:30 a.m., we obtain the mid-quote price change over that half hour for each day, as well as the inventory change over the same half hour. We then regress the price changes on the inventory changes. Analogous regressions are run for each subsequent half hour interval of the trading day. Since the market closes at 5:30 p.m., the last regression uses price and inventory changes from 5 p.m. - 5:30 p.m. The difference between the slope coefficient for the 5 p.m. - 5:30 p.m. interval and the slope coefficient from each earlier half hour interval is reported, along with the associated t-statistic and p-value. Prices are measured in basis points of par and inventories are measured in hundreds of millions of dollars par. The sample period is January 2, 2013 - June 29, 2018.

Time of Day i	$\beta_{close} - \beta_i$	t-stat	p-value
9:00 - 9:30	-0.86	[-3.76]	(0.00)
9:30 - 10:00	-0.89	[-3.95]	(0.00)
10:00 - 10:30	-0.90	[-3.94]	(0.00)
10:30 - 11:00	-1.00	[-4.43]	(0.00)
11:00 - 11:30	-1.01	[-4.37]	(0.00)
11:30 - 12:00	-1.08	[-4.80]	(0.00)
12:00 - 12:30	-1.02	[-4.55]	(0.00)
12:30 - 13:00	-1.18	[-5.27]	(0.00)
13:00 - 13:30	-0.76	[-3.31]	(0.00)
13:30 - 14:00	-0.89	[-3.94]	(0.00)
14:00 - 14:30	-0.85	[-3.56]	(0.00)
14:30 - 15:00	-1.29	[-5.57]	(0.00)
15:00 - 15:30	-1.37	[-6.09]	(0.00)
15:30 - 16:00	-1.12	[-4.91]	(0.00)
16:00 - 16:30	-1.13	[-5.00]	(0.00)
16:30 - 17:00	-1.10	[-4.83]	(0.00)

wider than spreads earlier in the day (e.g., 9 a.m. - 9:05 a.m.). Together with steepening price impacts, the widening of bid-ask spreads toward the end of the day is consistent with the hypothesis that liquidity providing PTFs on the BrokerTec ECN are averse to carrying inventory overnight.

5.4 Price Impact by Day of Week

A natural extension of the previous exercises is to ask whether the end-of-day inventory effects vary by day of the week. In particular, Fridays are unique from the perspective of an intraday trader in that the next trading session for managing risk and inventory is 49-50 hours away. By contrast, from Monday to Thursday, the overnight trading session in

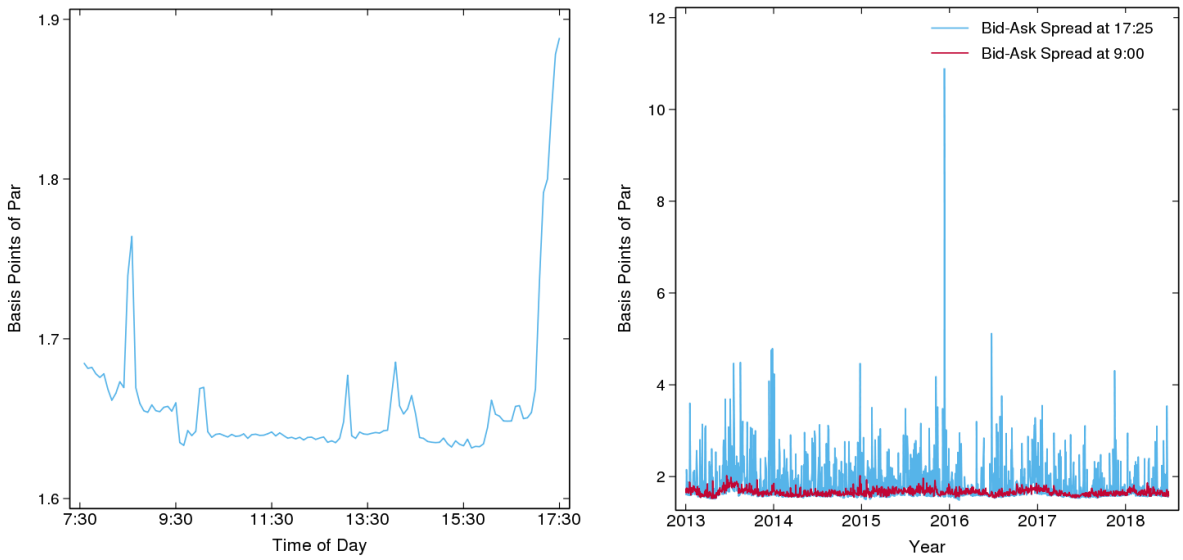


Figure 12: **Bid-Ask Spreads: Intraday and in the Time Series.** The left panel plots intraday bid-ask spreads for the 10-year Treasury note, for the January 2, 2013 - June 29, 2018 sample period. Using tick-by-tick order book data, bid-ask spreads are averaged for each 5 minute interval of each day, and then averaged across days for each 5 minute interval. The right panel plots the time series of bid-ask spreads for the 10-year Treasury note early in the New York trading day (9 a.m. - 9:05 a.m.) and near the market close (5:25 p.m. - 5:30 p.m.)

Treasuries opens 1-2 hours after the close, allowing some opportunity to trade out of residual inventories (albeit with higher transaction costs).[19] To the extent that news is revealed over the weekend, one might expect PTFs to show an aversion to potentially large Friday-to-Monday price swings, and therefore display a stronger desire to shed inventory before the close on Friday.[20]

To investigate this hypothesis, we partition our data set into a Friday-only subsample and a pooled Monday through Thursday subsample. For each subsample, we isolate the same intraday time intervals starting with 9 a.m. to 9:30 a.m., and regress price changes on changes in our inventory proxy, including a constant. We repeat this process for each subsequent half hour interval until 5:30 p.m.

[19]The reason the gap to the next trading session is not constant is because the market close is fixed relative to New York trading hours (5:30 p.m.) and the market open is fixed relative to Tokyo's market open (8:30 a.m. local time), and Japan does not have daylight saving time.

[20]We thank the referee for making this suggestion.

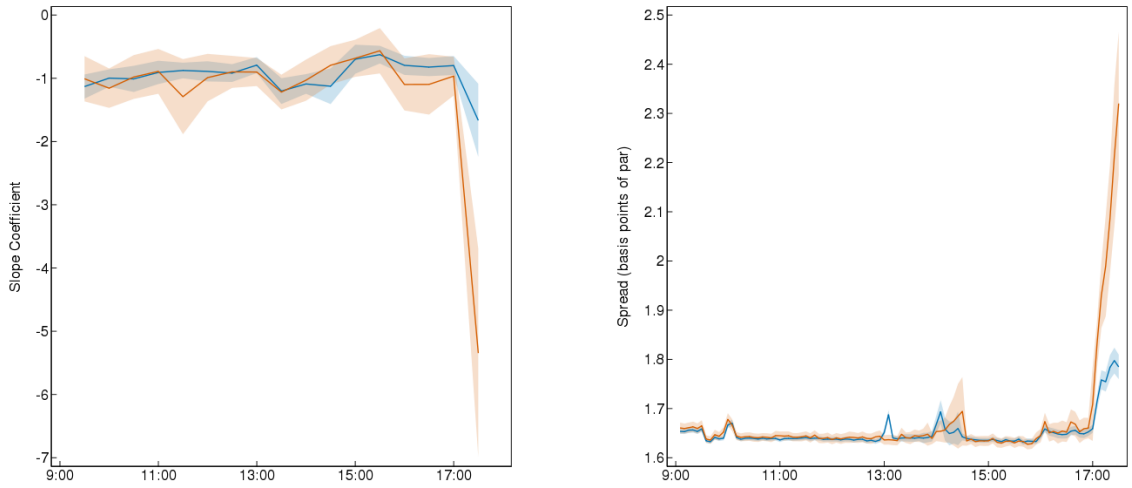


Figure 13: **Price Impact and Bid-Ask Spreads by Day of Week.** The left panel plots slopes β_n from regressions of 10-year Treasury price changes on changes in market maker inventory, as proxied by the negative of the cumulative net volume, for subsamples of Friday-only observations (orange) and pooled Monday through Thursday observations (blue). For each half hour interval, starting with 9 a.m. - 9:30 a.m., we obtain the mid-quote price change over that half hour for each day, as well as the inventory change over the same half hour. We then regress the price changes on the inventory changes for our Friday-only observations and for our Monday through Thursday observations. Analogous regressions are run for each subsequent half hour interval of the trading day. Since the market closes at 5:30 p.m., the last regressions use price and inventory changes from 5 p.m. - 5:30 p.m. The right panel plots intraday bid-ask spreads for the 10-year Treasury note on Fridays (orange) and on other days of the week (blue). Using tick-by-tick order book data, bid-ask spreads are averaged for each 5 minute interval of each day, and then averaged across days for each 5 minute interval for the subsamples of Friday-only observations and pooled Monday through Thursday observations. In both panels, shaded areas represent 99% confidence intervals based on standard errors (robust standard errors in left panel), and the sample period is January 2, 2013 - June 29, 2018.

The left panel of Figure 13 plots the resulting slope estimates by time of day, including their 99% confidence intervals. The slope estimates show a clear negative price-inventory relationship that persists throughout the day and strengthens toward the close, consistent with Figure 11. Moreover, the Friday closing (5 p.m. - 5:30 p.m.) price impact coefficient is approximately three times larger in magnitude than the closing price impact coefficient for the Monday through Thursday subsample.

The right panel of Figure 13 plots representative bid-ask spreads, obtained by averaging bid and ask quotes within each five-minute interval on each day and then further averaging

them across days for each five-minute interval within each subsample. Here, too, the end-of-day widening of bid-ask spreads is significantly greater on Fridays than it is for the other days of the week. Through the lens of the model, these price impact and bid-ask spread patterns provide support for the hypothesis that PTFs have an especially strong aversion to holding inventory over the weekend.

5.5 Alternative Explanations

Might there be alternative explanations for our findings? Fleming (1997) shows that Treasury market trading volume trails off as the end of the trading day approaches. Market participants exiting the market at the end of the day, reflecting reduced demand for liquidity provision, might naturally cause spreads to widen and price impact to increase. Indeed, we find average trading volume in the 5 p.m. - 5:30 p.m. interval to be lower than that in any other half hour interval during New York trading hours (albeit higher than in any half hour interval before 2 a.m.).

One way to address this question is to consider how PTFs behave relative to other market participants near the end of the trading day. Unfortunately, our dataset is anonymized, as mentioned earlier, precluding us from identifying specific firms or even firm types. That said, we can proxy for firm type by observed behavior in the market. In particular, we can proxy for the percent of trading volume accounted for by PTFs by the percent of trading volume that appears to be low latency. We define a trade as low latency if it occurs with 0.01 seconds of the preceding trade, too short of an interval to reflect human reaction.[21] Our measure is motivated by Hasbrouck and Saar (2013)'s definition of low latency activity as strategies that respond to market events in the millisecond environment and is similar to the approach used by Salem et al. (2018), who also examine data from BrokerTec.

As shown in Figure 14, our proxy for the percent of trading volume accounted for by PTFs is fairly steady across most of the trading day, and then declines sharply near the end of the trading day. That is, not only is volume declining at the end of the trading day, but

[21] The measure is necessarily imprecise because low latency trades are an imperfect proxy for PTF trades and because a trade that appears low latency may occur shortly after another trade by chance, or because both trades are reacting to an earlier market event and arriving at the platform at about the same time.

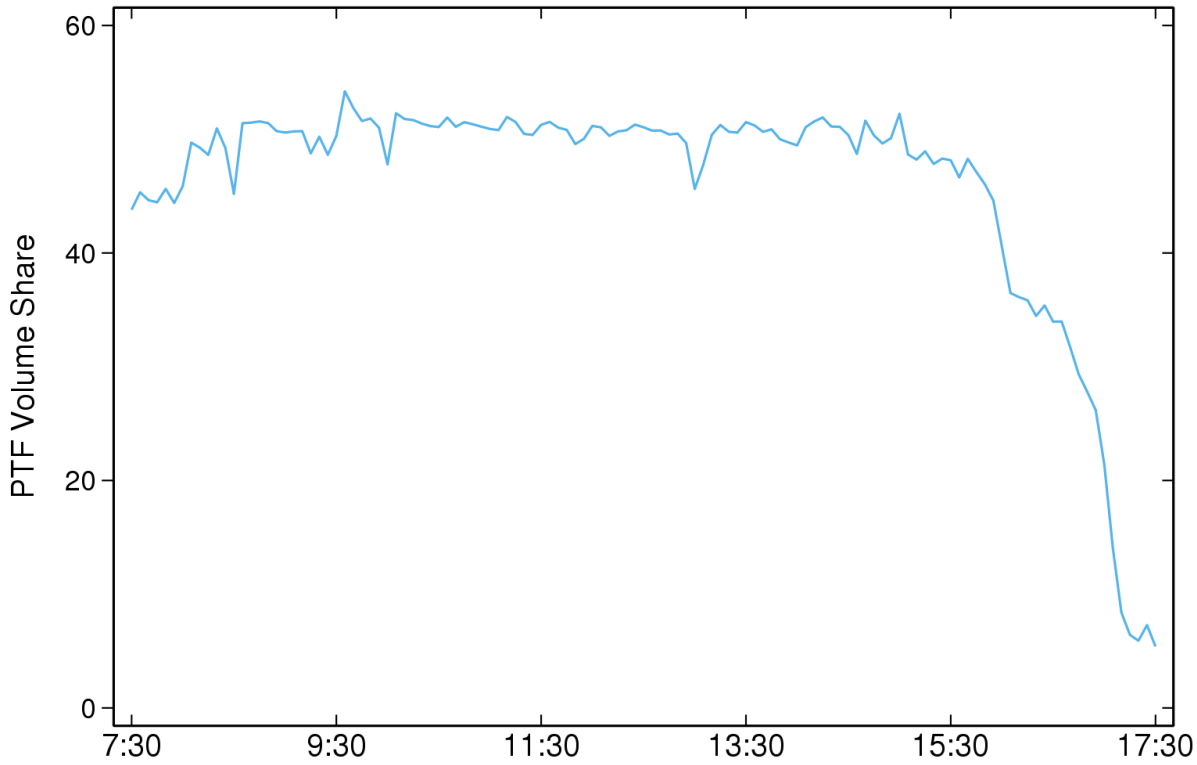


Figure 14: **PTF Trading Volume Share over the Trading Day.** This figure plots the share of trading volume over the trading day attributable to PTFs, as proxied by the share of trading volume that is estimated to be low latency. For each 5 minute interval of each day, we calculate the share of trading volume attributable to trades that execute within 0.01 seconds of the preceding trade. We then average these shares across days for each 5 minute interval. The sample period is April 1, 2015 - June 29, 2018.

PTFs' share of volume is seemingly also declining. Hesitance to take on new positions as the market close approaches may be causing PTFs to widen their spread and/or exit the market, leaving a greater share of market making to firms with larger balance sheets and lower inventory costs, such as broker-dealers.

Another possible explanation for the observed end-of-day empirical patterns is that they reflect the earlier (5 p.m.) close of futures and options trading. Cross-market activity accounts for a large share of trading in both the Treasury cash and futures markets (Dobrev and Schaumburg (2015)). The cessation of futures and options trading on the Chicago Board of Trade at 5 p.m. precludes cash market traders from immediately hedging positions in these other markets and might cause market makers to widen their spreads regardless of

the pending cash market close. We do in fact observe a discrete worsening of market liquidity right at 5 p.m., as measured by reduced order book depth, for example.

That said, we also observe a widening of bid-ask spreads in the minutes before 5 p.m. when both markets are open, and also in the minutes before 5:30 p.m., when only the cash market is open. These patterns are especially evident when we calculate bid-ask spreads for larger sized trades.[22] The general pattern of liquidity worsening as the end of the trading day approaches is thus replicated to a lesser degree in the period before the futures market close at 5 p.m. and again before the cash market close at 5:30 p.m.

6 Conclusion

Principal trading firms now account for a large share of liquidity provision in the secondary U.S. Treasury market (Joint Staff Report (2015)). The literature suggests that PTFs unwind their trading books before the end of the day, because they face overnight funding costs that are substantial relative to their limited capital bases. To study the effect of overnight inventory aversion, we propose a continuous time model of intraday market making in which a PTF balances trading profits from crossing the bid-ask spread against the costs of holding residual inventory at the end of the trading day.

We characterize the optimal market making behavior of a PTF and provide closed-form expressions for the PTF's value function as well as for its optimal price policy functions. The PTF's optimal price setting strategy gives rise to endogenous price impact and bid-ask spreads that vary as a function of the PTF's shadow value of the overnight inventory cost. Inventory aversion leads to bid-ask spreads and price impact that rise as the end of the trading day approaches, even though buy and sell orders arrive at a constant rate. Higher overnight inventory costs steepen the price impact function of the market maker, and lead to a higher volatility of the price trajectories. A higher intensity of order arrivals attenuates

[22]Fleming et al. (2018) find that 97% of inside spreads for the on-the-run 2-year note equal the minimum tick size (using BrokerTec tick data for 2010-2011). To detect changes in liquidity that are otherwise masked by the minimum spread for a $1 million trade, we also calculate the spread for a "large" trade, defined as a trade size at the 95th percentile (based on all trades over the sample period), allowing for greater variation in the spread.

the shadow cost of overnight inventory and reduces price volatility if the trading time is sufficiently far from the day's close. On the other hand, a more active market intensifies price impact and increases price volatility if trading occurs towards the end of the day.

We also study price stability and provide comparative statics measuring the surplus per trade for buyers, sellers, and the PTF's value. Our results suggest that, while overnight inventory costs increase price instability and reduce buyer/seller's surplus, these effects can be either offset or reversed depending on market activity. In markets with high order arrival rates, increases in the overnight inventory costs have a negligible impact on the buyer/seller's surplus per trade and the PTF's value. In contrast, overnight inventory costs have the most detrimental effect in markets with low order arrival rates.

The theoretical predictions of our model are supported by empirical evidence based on U.S. Treasury data. Our analysis confirms a statistically significant negative relationship between intraday changes in prices and changes in inventory holdings, as well as the intensification of price impact and the widening of bid-ask spreads toward the day's close. The finding that PTFs' share of trading activity seems to decline as the end of the trading day approaches provides further support for the hypothesis that these temporal patterns are driven by PTFs' aversion to holding excessive overnight inventory.

A Proofs

Proof of Lemma 3.1. For fixed t and S, because $F(t,S,i)$ is assumed to be strictly concave in i (its derivative is decreasing), we know that $G_{t,S}(i)$ is strictly decreasing in i. Furthermore, we have assumed that $\partial_i F(t,S,i)$ is increasing in i and maps onto $\mathbb{R}$, thus we know that $G_{t,S}(-\infty) = \infty = -G_{t,S}(\infty)$.

To prove the results, we determine the optimal ask and bid prices by maximizing the Hamiltonian H. Because the Hamiltonian H can be written as the sum of a function of the ask price a and another function of the bid price b, we can separately determine the optimal ask and bid prices.

For the optimal ask price, since the function $F(t,S,i)$ is strictly concave in i, we know that for each fixed i, the mapping

$$\begin{aligned} a \mapsto\ & v(t,S,w + a\,Q^B(S,a), i - Q^B(S,a)) - v(t,S,w,i) \\ &= ca(S+p-a) + F(t,S,i - c(S+p-a)) - F(t,S,i) \end{aligned}$$

is strictly concave in a. Likewise, the mapping

$$\begin{aligned} b \mapsto\ & v(t,S,w - b\,Q^S(S,b), i + Q^S(S,b)) - v(t,S,w,i) \\ &= -cb(b - S + p) + F(t,S,i + c(b - S + p)) - F(t,S,i) \end{aligned}$$

is strictly concave in b. Hence, for each fixed i, there is a unique optimal pair $(a_t(S,i), b_t(S,i))$ that maximizes the Hamiltonian in (11) (notice that it may still occur that $a_t(S,i) < b_t(S,i)$). In addition, $a_t(S,i), b_t(S,i)$ are the solutions of the decoupled system of first order conditions given by

$$\begin{aligned} \partial_i F(t,S,i - c(S + p - a_t(S,i))) + S + p - 2a_t(S,i) = 0, \\ \partial_i F(t,S,i + c(b_t(S,i) - S + p)) + S - p - 2b_t(S,i) = 0. \end{aligned}$$

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

$$\begin{aligned} 0 &= \partial_i F(t,S,i - c(S + p - a_t(S,i))) + S + p - 2a_t(S,i) \\ &= G_{t,S}(i - c(S + p - a_t(S,i))) - (S + p - \frac{2i}{c}). \end{aligned}$$

Using the definition of inverse functions, we obtain

$$i - c(S + p - a_t(S,i)) = G_{t,S}^{-1}\left(S + p - \frac{2i}{c}\right),$$

leading to (12). Similarly, from

$$0 = \partial_i F(t,S,i + c(b_t(S,i) - S + p)) + S - p - 2b_t(S,i) = G_{t,S}(i + c(b_t(i) - S + p)) - (S - p - \frac{2i}{c}),$$

we obtain (13).

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

$$\begin{aligned} \partial_i F(t,S,i_2 - c(S + p - a_t(S,i_1))) + S + p - 2a_t(S,i_1) \\ < \partial_i F(t,S,i_1 - c(S + p - a_t(S,i_1))) + S + p - 2a_t(S,i_1)) = 0, \end{aligned}$$

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

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

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

Finally, notice that

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

By (12), (13) and the monotonicity of $G^{-1}_{t,S}$, we have

$$a_t(S,i) - b_t(S,i) = \frac{1}{c}G^{-1}_{t,S}\bigg(-p - \frac{2i}{c} + 2p\bigg) - \frac{1}{c}G^{-1}_{t,S}\bigg(-p - \frac{2i}{c}\bigg) + 2p < 2p.$$

On the other hand, we notice that

$$b_t(S,i) = a_t(S, i+cp) - p,$$

which immediately implies

$$a_t(S,i) - b_t(S,i) = p + a_t(S,i) - a_t(S,i+cp) > p,$$

where the inequality follows from the fact that $a_t(S,i)$ is strictly decreasing in i. This completes the proof. □

Proof of Theorem 3.2. We begin by considering a quadratic ansatz for the value function. We conjecture that the value function is of the form $\tilde{v}(t,S,w,i) = w + A(t)i^2 + B(t)i + C(t)$, where $A(t), B(t), C(t)$ are differentiable functions of time t. Because the terminal value is known to be $U(S,w,i) = w + Si - \lambda i^2$, it must hold that $A(T) = -\lambda, B(T) = S, C(T) = 0$. Plugging this conjectured value function into the HJB equation, we obtain

$$\begin{aligned}\partial_t\widetilde{v} + \frac{1}{2}\sigma^2\partial_S^2\widetilde{v} = -\sup_{a,b}\big\{&\pi^B(t)[\widetilde{v}(t,S,w + a\,Q^B(S,a), i - Q^B(S,a)) - \widetilde{v}(t,S,w,i)] \\ &+\pi^S(t)[\widetilde{v}(t,S,w - b\,Q^S(S,b), i + Q^S(S,b)) - \widetilde{v}(t,S,w,i)]\big\}, \quad (29)\end{aligned}$$

where the supremum is attained if and only if $a = a^*_t(S,i)$ and $b = b^*_t(S,i)$ as in (18) and (19) Assuming $A(t) < 0$ holds for all $t \le T$, and matching the coefficients of $i^2, i, 1$ on both sides of (29), we obtain that

$$A(t) = \alpha(t), B(t) = S, C(t) = cp^2\int_t^T \frac{(\pi^B(u) + \pi^S(u))du}{2(1 - c\alpha(u))},$$

where $\alpha(t)$ satisfies $\alpha(T) = -\lambda$ and solves an ODE $\alpha'(t) + c(\pi^B(t) + \pi^S(t))\alpha^2(t)/(1 - c\alpha(t)) = 0$. That is, $\alpha'(t) = f(t, \alpha(t))$ where the slope field

$$f(t,\alpha) = -(\pi^B(t) + \pi^S(t))\frac{c\alpha^2}{1 - c\alpha}.$$

Using standard argument we can show that the ODE has a unique solution for any nonzero terminal condition $\alpha(T)$ and the solution is always negative if $\alpha(T) < 0$.

Conversely, because α solves the above ODE, and it is negative, the conjectured $\widetilde{v}$ solves the HJB equation given by $\widetilde{v}(T,S,w,i) = U(S,w,i) = w + Si - \lambda i^2$, and (29). Moreover, the integrability condition (7) ensures that the stochastic integrals with respect to the Brownian motion and the compensated Poisson processes are true martingales. Hence, for any (Markovian or non-Markovian) admissible control $(a.,b.) \in \mathcal{A}_{0,T}$, the process $(\widetilde{v}(t, S_t, W_t^{(a,b)}, I_t^{(a,b)}))_{t\in[0,T]}$

is a supermartingale. Therefore, for any $t \in [0,T)$, we have $\mathbb{P}$-a.s. that

$$\begin{aligned}\widetilde{v}(t,S_t,W_t,I_t) \geq & \sup_{(a.,b.)\in\mathcal{A}_{t,T}} \mathbb{E}\left[\tilde{v}(T,S_T,W_T^{(a,b)},I_T^{(a,b)})|\mathcal{F}_t\right] \\ = & \sup_{(a.,b.)\in\mathcal{A}_{t,T}} \mathbb{E}\left[U(S_T,W_T^{(a,b)},I_T^{(a,b)})|\mathcal{F}_t\right] \\ = & V_t.\end{aligned}$$

On the other hand, from the proofs of Corollary 3.3 and Proposition 4.1 we know that $\mathbb{E}\left[|I_t^{(a^*,b^*)}|\right] < \infty$ and $\mathbb{E}\left[\left(I_t^{(a^*,b^*)}\right)^2\right]$ is a continuous bounded function over $[0,T]$. Since the conjectured price policies (a^*,b^*) are linear in the inventory, it is straightforward to verify that they are admissible strategies. By construction, we know that when $a_u = a_u^*(S,I_u)$ and $b_u = b_u^*(S,I_u)$ for all $u \in [0,T]$, $(\widetilde{v}(t,S_t,W_t^{(a^*,b^*)},I_t^{(a^*,b^*)}))_{t\in[0,T]}$ is a true martingale, so we must have

$$\widetilde{v}(t,S_t,W_t,I_t) = V_t, \quad \mathbb{P}\text{-a.s.}.$$

The optimal price policy functions follow from Lemma 3.1. This completes the proof. □

Proof of Corollary 3.3. The stochastic differential equation (SDE) for the inventory dynamics given by

$$\begin{aligned}dI_t \;=\; & \left(\frac{p(\pi^S(t)-\pi^B(t))}{2(1-c\alpha(t))} + \frac{2(\pi^B(t)+\pi^S(t))\alpha(t)}{1-c\alpha(t)} I_{t-}\right) c\,dt \\ & -\frac{p-2\alpha(t)I_{t-}}{2(1-c\alpha(t))} c\left(dN_t^B - \pi^B(t)dt\right) + \frac{p+2\alpha(t)I_{t-}}{2(1-c\alpha(t))} c\left(dN_t^S - \pi^S(t)dt\right)\end{aligned}$$

follows from (5), (18) and (19). Let us now show that $\mathbb{E}\left[|I_t|\right] < \infty$ for all $t \in [0,t]$, so $\mathbb{E}[I_t]$ is well-defined. To that end, we notice that I_t satisfies a SDE of the form

$$I_t - I_0 = \int_0^t (c_1(u) + c_2(u)I_u)dN_u^B + \int_0^t (c_3(u) + c_4(u)I_u)dN_u^S,$$

where $c_k(u)$, $k = 1,2,3,4$ are four continuous functions of time u. Using standard arguments based on the triangular inequality, localization and the monotone convergence theorem, we obtain that

$$\mathbb{E}|I_t - I_0| \leq \int_0^t (d_1(t) + d_2(t)\mathbb{E}|I_t - I_0|)dt,$$

where $d_1(t)$ and $d_2(t)$ are two nonnegative continuous functions. By Grönwall's inequality, we know that $\mathbb{E}|I_t - I_0| < \infty$ and so $\mathbb{E}|I_t| < \infty$ for all $t \in [0,T]$.

As a consequence, we can denote the expected inventory by $g(t) = \mathbb{E}[I_t]$, and treat it as a solution to an ODE:

$$g'(t) = \frac{cp(\pi^S(t)-\pi^B(t))}{2(1-c\alpha(t))} + \frac{2c(\pi^B(t)+\pi^S(t))\alpha(t)}{1-c\alpha(t)} g(t).$$

Solving the above ODE we obtain that

$$g(t) = g(0)\exp\left(\int_0^t \frac{2c(\pi^B(t)+\pi^S(t))\alpha(u)}{1-c\alpha(u)}du\right) + \int_0^t \frac{cp(\pi^S(t)-\pi^B(t))}{2(1-c\alpha(u))}\exp\left(\int_u^t \frac{2c(\pi^B(t)+\pi^S(t))\alpha(v)}{1-c\alpha(v)}dv\right)du.$$

This completes the proof. □

Proof of Lemma 3.4. Let $h(x) = -\frac{1}{x} + c\log x$ for all $x > 0$, which is a strictly increasing function. Then $\alpha(t)$ satisfies

$$h(\lambda) - h(|\alpha(t)|) = c\int_t^T (\pi^B(u)+\pi^S(u))du, \quad t \in [0,T]. \tag{30}$$

Because the right hand side of (30) is strictly decreasing in t, we know that $h(|\alpha(t)|)$ must be strictly increasing, hence $|\alpha(t)| = -\alpha(t)$ is strictly increasing. Suppose $\pi^B(t) \equiv \pi^B$ and $\pi^S(t) \equiv \pi^S$ are constants. For fixed $t \in [0,T)$ and $\lambda > 0$, as we increase $\pi^B + \pi^S$, we see that $h(|\alpha(t)|)$ decreases strictly, so does $|\alpha(t)|$. For fixed $t \in [0,T]$ and π^B, π^S, we see from (30) that $h(|\alpha(t)|)$ increases strictly, so does $|\alpha(t)|$. □

Proof of Corollary 3.5. The formulas for L_t^1 and L_t^2 follow directly from (15), (18) and (19). Then the monotonicity of $(L_t^1)_{t\in[0,T]}$ and $(L_t^2)_{t\in[0,T]}$ follows form Lemma 3.4. □

Proof of Proposition 4.1. We begin by deriving the second moment of the inventory $\mathbb{E}[I_t^2]$. To that end, we use (5) and Itô-Lévy lemma to obtain that

$$\begin{aligned} dI_t^2 =& [(I_t - Q^B(S_t, a_t^*(S_t,I_t)))^2 - (I_t)^2]dN_t^B + [(I_t + Q^S(S_t, b_t^*(S_t,I_t)))^2 - (I_t)^2]dN_t^S \\ =& \left(\frac{(cp)^2}{4(1-c\alpha(u))^2} + \frac{c\alpha(t)(2-c\alpha(t))}{(1-c\alpha(t))^2}I_t^2 - [\frac{cp}{(1-c\alpha(t))} - \frac{c^2p\alpha(t)}{(1-c\alpha(t))^2}]I_t\right)dN_t^B \\ &+ \left(\frac{(cp)^2}{4(1-c\alpha(u))^2} + \frac{c\alpha(t)(2-c\alpha(t))}{(1-c\alpha(t))^2}I_t^2 + [\frac{cp}{(1-c\alpha(t))} + \frac{c^2p\alpha(t)}{(1-c\alpha(t))^2}]I_t\right)dN_t^S. \end{aligned}$$

Let $g_2(t) = \mathbb{E}[I_t^2]$, then we have $g_2(0) = 0$, and let $g_1(t) = \mathbb{E}[I_t]$. Using the same argument as in the proof of Corollary 3.3, we can first prove that $g_2(t) < \infty$ for all $t \in [0,T]$. Then we know that $g_2(\cdot)$ solves ODE

$$g_2'(t) = \frac{\pi_0(t)}{2}\frac{(cp)^2}{(1-c\alpha(t))^2} + 2\pi_0(t)\frac{c\alpha(t)(2-c\alpha(t))}{(1-c\alpha(t))^2}g_2(t) + 2\left(\frac{\pi_0(t)c^2p\alpha(t)}{(1-c\alpha(t))^2} - \frac{\epsilon(t)cp}{1-c\alpha(t)}\right)g_1(t).$$

The solution to the differential equation above is given by

$$g_2(t) = \int_0^t \left(\frac{\pi_0(u)(cp)^2}{2(1-c\alpha(u))^2} + 2\left(\frac{\pi_0(u)c^2p\alpha(t)}{(1-c\alpha(t))^2} - \frac{\epsilon(u)cp}{1-c\alpha(t)}\right)g_1(u)\right) \times \exp\left(\int_u^t \frac{2\pi_0(v)c\alpha(v)(2-c\alpha(v))}{(1-c\alpha(v))^2}dv\right)du.$$

Let $\langle I\rangle_t$ be the quadratic variation of the inventory process during the interval $[0,t]$. Then

we have

$$\begin{aligned}
d\langle I\rangle_t =&[Q^B(S_t, a_t(S_t, I_t))]^2 dN_t^B + [Q^S(S_t, b_t(S_t, I_t))]^2 dN_t^S \\
=&\left(\frac{(cp)^2}{4(1-c\alpha(t))^2} + \frac{c^2\alpha^2(t)}{(1-c\alpha(t))^2} I_t^2 - \frac{c^2 p\alpha(t)}{(1-c\alpha(t))^2} I_t\right) dN_t^B \\
&+ \left(\frac{(cp)^2}{4(1-c\alpha(t))^2} + \frac{c^2\alpha^2(t)}{(1-c\alpha(t))^2} I_t^2 + \frac{c^2 p\alpha(t)}{(1-c\alpha(t))^2} I_t\right) dN_t^S.
\end{aligned}$$

It follows that

$$\mathbb{E}[\langle I\rangle_t] = \int_0^t \left(\frac{\pi_0(u)(cp)^2}{2(1-c\alpha(u))^2} + \frac{2\pi_0(u)c^2\alpha^2(u)}{(1-c\alpha(u))^2} g_2(u) - \frac{2\epsilon(u)\, c^2 p\alpha(t)}{(1-c\alpha(t))^2} g_1(t)\right) du.$$

On the other hand, from (1) and (24), we know that

$$\begin{aligned}
\mathbb{E}[QV_t] =&\sigma^2 t + \mathbb{E}[\int_0^t \frac{\alpha^2(u)}{(1-c\alpha(u))^2} d\langle I\rangle_u] \\
=&\sigma^2 t + \int_0^t \frac{\alpha^2(u)}{(1-c\alpha(u))^2} d\mathbb{E}[\langle I\rangle_u] \\
=&\sigma^2 t + \int_0^t \frac{\alpha^2(u)}{(1-c\alpha(u))^2} \left(\frac{\pi_0(u)(cp)^2}{2(1-c\alpha(u))^2} + \frac{2\pi_0(u)c^2\alpha^2(u)}{(1-c\alpha(u))^2} g_2(u) - \frac{2\epsilon(u)\, c^2 p\alpha(u)}{(1-c\alpha(u))^2} g_1(u)\right) du.
\end{aligned}$$

This completes the proof. □

Proof of Proposition 4.2. From (18) we deduce that

$$(S_t + p - a_t^*)^2 = \frac{p^2}{4(1-c\alpha(t))^2} + \frac{\alpha^2(t)}{(1-c\alpha(t))^2} I_t^2 - \frac{p\alpha(t)}{(1-c\alpha(t))^2} I_t.$$

Hence, we obtain

$$\mathbb{E}[(S_t + p - a_t^*)^2] = \frac{p^2}{4(1-c\alpha(t))^2} + \frac{\alpha^2(t)}{(1-c\alpha(t))^2} \mathbb{E}[I_t^2].$$

On the other hand, an application of Fubini's theorem yields

$$\begin{aligned}
ASurplus^B =&\frac{c}{2\pi_0 T} \mathbb{E}\left[\int_0^T (S_t + p - a_t^*)^2 dN_t^B\right] \\
=&\frac{c}{2T} \int_0^T \mathbb{E}[(S_t + p - a_t^*)^2] dt \\
=&\frac{c}{2T} \int_0^T \left(\frac{p^2}{4(1-c\alpha(t))^2} + \frac{\alpha^2(t)}{(1-c\alpha(t))^2} \mathbb{E}[I_t^2]\right) dt.
\end{aligned}$$

Repeating the same analysis above, one can derive the corresponding expression for $ASurplus^S$. The expression for $AValue^M$ follows immediately from Theorem 3.2. □

References

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

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

Bank for International Settlements (2011). High-frequency trading in the foreign exchange market. *Markets Committee Papers*, (5).

Bayaraktar, E. and Ludkovski, M. (2012). Liquidation in limit order books with controlled intensity. *Mathematical Finance*, 24(4):627–650.

Benos, E. and Sagade, S. (2016). Price discovery and the cross-section of high-frequency trading. *Journal of Financial Markets*, 30:54–77.

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

Boehmer, E., Li, D., and Saar, G. (2018). The competitive landscape of high-frequency trading firms. *Review of Financial Studies*, 31(6):2227–2276.

Bradfield, J. (1979). A formal dynamic model of market making. *Journal of Financial and Quantitative Analysis*, 14(2):275–291.

Brain, D., De Pooter, M., Dobrev, D., Fleming, M., Johansson, P., Keane, F., Puglia, M., Rodrigues, A., and Shachar, O. (2018). Breaking down TRACE volumes further. *Liberty Street Economics, Federal Reserve Bank of New York*, November 29, 2018.

Brogaard, J. and Garriott, C. (2019). High-frequency trading competition. *Journal of Financial and Quantitative Analysis*, 54(4):1469–1497.

Cartea, A. and Jaimungal, S. (2015). Risk metrics and fine tuning of high frequency trading strategies. *Mathematical Finance*, 25(3):576–611.

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.

Chordia, T., Roll, R., and Subrahmanyam, A. (2002). Order imbalance, liquidity, and market returns. *Journal of Financial Economics*, 65(1):111–130.

Cvitanic, J. and Kirilenko, A. (2010). High frequency traders and asset prices. Working paper, California Institute of Technology and Imperial College London.

Dobrev, D. and Schaumburg, E. (2015). High-frequency cross-market trading in U.S. Treasury markets. *Liberty Street Economics, Federal Reserve Bank of New York*, August 19, 2015.

Duffie, D. and Ashcraft, A. (2007). Systemic illiquidity in the federal funds market. *American Economic Review*, 97(2):221–225.

Fleming, M. (1997). The round-the-clock market for U.S. Treasury securities. *Federal Reserve Bank of New York Economic Policy Review*, 3:9–32.

Fleming, M., Mizrach, B., and Nguyen, G. (2018). The microstructure of a U.S. Treasury ECN: The BrokerTec platform. *Journal of Financial Markets*, 40:2–22.

Fleming, M. and Nguyen, G. (2019). Price and size discovery in financial markets: Evidence from the U.S. Treasury securities market. *Review of Asset Pricing Studies*, 9(2):256–295.

Fleming, M. and Remolona, E. (1999). Price formation and liquidity in the U.S. Treasury market: The response to public information. *Journal of Finance*, 54(5):1901–1915.

Glosten, L. and Milgrom, P. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1):71–100.

Green, T. C. (2004). Economic news and the impact of trading on bond prices. *Journal of Finance*, 59(3):1201–1233.

Hasbrouck, J. and Saar, G. (2013). Low latency trading. *Journal of Financial Markets*, 16(4):646–679.

Hendershott, T. and Menkveld, A. (2014). Price pressures. *Journal of Financial Economics*, 114(3):405–423.

Ho, T. and Stoll, H. (1981). Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics*, 9(1):47–73.

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, Federal Reserve Bank of New York, U.S. Securities and Exchange Commission, U.S. Commodity Futures Trading Commission.

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. *Annual Review of Financial Economics*, 8:1–24.

O'Hara, M. and Oldfield, G. (1986). The microeconomics of market making. *Journal of Financial and Quantitative Analysis*, 21(4):2603–2619.

Rosu, I. (2019). Liquidity and information in limit order markets. *Journal of Financial and Quantitative Analysis, Forthcoming*.

Salem, M., Younger, J., and St John, H. (2018). Fast and furious: The link between rapid trading and volatility in U.S. rates markets. J.P. Morgan.

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

Soner, H., Touzi, N., and Zhang, J. (2012). Wellposedness of second order backward SDEs. *Probability Theory and Related Fields*, 153(1-2):149–190.