Journal of Financial Markets 43 (2019) 1–30

Contents lists available at ScienceDirect

Journal of Financial Markets

journal homepage: www.elsevier.com/locate/finmar

Fast and slow informed trading☆

Ioanid Roşu

HEC Paris, 1 Rue de la Libération, 78351, Jouy-en-Josas, France

ARTICLE INFO

Article history:
Received 10 May 2018
Revised 13 February 2019
Accepted 15 February 2019
Available online 23 February 2019

Classification codes:
G14
D82

Keywords:
Trading volume
Inventory
Volatility
High-frequency trading
Price impact
Mean reversion

ABSTRACT

I develop a model in which traders receive a stream of private signals, and differ in their information processing speed. In equilibrium, the fast traders (FTs) quickly reveal a large fraction of their information. If a FT is averse to holding inventory, his optimal strategy changes considerably as his aversion crosses a threshold. He no longer takes long-term bets on the asset value, gets most of his profits in cash, and generates a "hot potato" effect: after trading on information, the FT quickly unloads part of his inventory to slower traders. The results match evidence about high-frequency traders.

1. Introduction

Today's markets are increasingly characterized by the continuous arrival of vast amounts of information. A media article about high-frequency trading reports on the hedge fund firm Citadel: "Its market data system, for example, contains roughly 100 times the amount of information in the Library of Congress. [...] The signals, or alphas, that prove to have predictive power are then translated into computer algorithms, which are integrated into Citadel's master source code and electronic trading program." ("Man vs. Machine," CNBC.com, Sept. 13, 2010). The sources of information from which traders obtain these signals usually include company-specific news and reports, economic indicators, stock indexes, prices of other securities, prices on various other trading platforms, limit order book changes, as well as various "machine readable news" and even "sentiment"

☆ I thank two anonymous referees, Gideon Saar (the editor), Kerry Back, Laurent Calvet, Thierry Foucault, Johan Hombert, Pete Kyle, Stefano Lovo, Victor Martinez, Daniel Schmidt, Dimitri Vayanos, and Jiang Wang for their suggestions. I am also grateful to finance seminar participants at Copenhagen Business School, HEC Paris, Univ. of Durham, Univ. of Leicester, Univ. Paris Dauphine, Univ. Madrid Carlos III, ESSEC, KU Leuven, Aalto; and to conference participants at the AFFI Eurofidai 2015 meetings, CEPR Gerzensee 2015 meetings, European Finance Association 2014 meetings, American Finance Association 2013 meetings, Society for Advancement of Economic Theory in Portugal, Central Bank Microstructure Conference in Norway, "Trading in Electronic Markets" Conference in Toulouse, Market Microstructure Many Viewpoints Conference in Paris, and the Labex ECODEC Workshop in Finance, for valuable comments. I acknowledge financial support from the Investissements d'Avenir Labex (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047).

E-mail address: rosu@hec.fr.

https://doi.org/10.1016/j.finmar.2019.02.003
1386-4181/

indicators.[1]

At the same time, financial markets have seen in recent years the spectacular rise of algorithmic trading, and in particular of high-frequency trading. Hendershott et al. (2011) report that from a starting point near zero in the mid-1990s, high-frequency trading rose to as much as 73% of trading volume in the United States in 2009. Chaboud et al. (2014) consider various foreign exchange markets and find that starting from essentially zero in 2003, algorithmic trading rose by the end of 2007 to approximately 60% of the trading volume for the euro-dollar and dollar-yen markets, and 80% for the euro-yen market. This coincidental arrival raises the question whether or not at least some of the high-frequency traders (HFTs) process information and trade very quickly in order to take advantage of their speed and superior computing power. Recent empirical evidence suggests that this is indeed the case (e.g., Brogaard et al., 2014; Brogaard et al., 2015; Benos and Sagade, 2016; Kirilenko et al., 2017; Boehmer et al., 2018b; Hirschey, 2018; Baron et al., 2019). Nevertheless, despite the large role played by HFTs in the current financial landscape, there has been relatively little progress in explaining their strategies in connection with information processing.

I consider the following questions regarding HFTs: What are the optimal trading strategies of HFTs who process information? Why do HFTs account for such a large share of the trading volume? What explains the race for speed among HFTs? What are the effects of HFTs on measures of market quality, such as liquidity and price volatility? How can HFTs' order flow anticipate future order flow and returns? What explains the "intermediation chains" or "hot potato" effects found among HFTs? (e.g., Weller, 2012; Kirilenko et al., 2017.) Why do some HFTs have low inventories? Regarding the last question, some identify HFTs as traders with both high trading volume *and* low inventories (e.g., SEC, 2010; Kirilenko et al., 2017). Then, a natural question arises: Why would having low inventories be part of the definition of HFTs?

In this paper, I provide a theoretical model of informed trading with speed differences that parsimoniously addresses these questions. The word "speed" in my context refers not to the speed of trading, which is arguably less important in modern trading platforms, but rather to the speed of receiving and processing information. To analyze informed trading at different speeds, I start with Kyle's (1985) model and modify it along several dimensions.[2] First, the asset's fundamental value is not constant but follows a random walk process, and each risk-neutral informed trader, or speculator, gradually receives signals about the asset value increments. Second, there are multiple speculators who differ in their speed, by receiving their signal with a lag. Third, each speculator can trade only on lagged signals with a lag of at most m, where m is an exogenously given number.

It is the last assumption that sets my model apart from previous models of informed trading. A key effect of this assumption is to prevent the "rat race" phenomenon identified by Holden and Subrahmanyam (1992), by which traders with identical information reveal their information so quickly that the equilibrium breaks down at the "high frequency" limit, when the number of trading rounds approaches infinity. In my model, the speculators reveal only a fraction of their total private information, and this has a stabilizing effect on the equilibrium. Economically, one can think of this assumption as equivalent to having a positive information processing cost per signal (and per trading round).[3] Indeed, since one of my results is that the value of information decays fast, even a tiny information processing cost would make speculators optimally ignore their signals after a sufficiently large number of lags m.

To simplify the analysis, I focus on the particular case when $m = 1$ when speculators can trade using only their current signal and its lagged value. Thus, there are two types of speculators: "fast traders" (FTs), who observe the signal instantly; and "slow traders" (STs), who observe the signal after one lag. In this case, the equilibrium can be described in closed form.[4]

I find that the fast traders generate most of the trading volume, volatility, and profits. To understand why, suppose that nine FTs decide what weight to use on the last signal they receive. Because the dealer sets a price function that is linear in the aggregate order size, each FT faces a Cournot-type problem and trades such that his price impact is on average 10% of his signal. That brings the expected aggregate price impact to 90% of the signal, and leaves on average only 10% of the signal unknown to the dealer. Thus, once the STs observe the lagged signal, they now have much less private information to exploit. Moreover, the ST profits are further diminished by competition with FTs, who also trade on the lagged signal. Empirically, Baron et al. (2019) find that the profits of HFTs are concentrated among a small number of incumbents, and their profits are correlated with speed. An additional consequence of this result is anticipatory trading: the order flow of fast traders predicts the order flow of slow traders in the next period. Thus, the speculator order flow autocorrelation is positive, although it is small if the number of fast traders is large. Empirically, Brogaard (2011) finds that the autocorrelation of aggregate HFT order flow is indeed small and positive. Also, using NASDAQ data on HFTs, Hirschey (2018) finds that HFT order flow anticipates future order flow.

A related result is that volume, volatility, and liquidity increase with the number of FTs. First, more competition from FTs makes the prices more informative overall, and thus increases liquidity (measured by the inverse price impact coefficient, as in Kyle, 1985). As the market is more liquid, FTs face a lower price impact, and therefore trade even more aggressively. This creates an amplification mechanism that allows the aggregate FT trading volume to be increasing roughly linearly with the number

[1] "Math-loving traders are using powerful computers to speed-read news reports, editorials, company Web sites, blog posts and even Twitter messages—and then letting the machines decide what it all means for the markets." ("Computers That Trade on the News," The New York Times, Dec. 22, 2010).

[2] As Kyle (1985), I assume that informed traders are market takers and thus submit only market orders; this is a plausible assumption for informed HFTs (e.g., Brogaard et al., 2014). In addition, I argue that the model is also able to describe HFTs who behave like market makers, as I later show that fast traders (with sufficiently large inventory costs) partially trade in the opposite direction to the slower traders, and thus in effect provide liquidity to them.

[3] Intuitively, information processing is costly because speculators need to avoid trading on stale information, and this involves (i) constantly monitoring public information to verify that their signal has not been incorporated into the price, and (ii) extracting the predictable part of their signal from past order flow, so that speculators trade only on the unpredictable (non-stale) part.

[4] In the Internet Appendix, I verify numerically that the main results of the particular case, $m = 1$, carry through to the general case ($m \geq 1$).

of FTs. The effect of FTs on volatility is more muted but still positive; this is because in my model price volatility is bounded above by the fundamental volatility of the asset. Empirically, in line with my theoretical results, Zhang (2010), Hendershott et al. (2011), and Boehmer et al. (2018a) document that HFTs exert a positive effect on liquidity. Moreover, Zhang (2010), and Boehmer et al. (2018a) find a positive effect of HFTs on volatility. Note, however, that my model is more likely to apply only to the subcategory of informed, market taking HFTs, and not to all HFTs. The results should therefore be interpreted with caution.

Despite being able to match several stylized facts about HFTs in the model, a few questions remain. Why do many HFTs have low inventories, both intraday and at the day close?[5] Why do HFTs engage in "hot potato" trading (or "intermediation chains"), in which HFTs pass their inventories to other traders?[6] What is the role of speed in explaining these phenomena?

To provide some theoretical guidance on these issues, I extend the benchmark model described above to include one trader with inventory costs. These costs can arise from risk aversion or from capital constraints, but I take a reduced form approach and assume the costs are quadratic in inventory, with a coefficient called "inventory aversion" (see Madhavan and Smidt, 1993). I call this additional trader the Inventory-averse Fast Trader, or IFT.[7] I call this extension the "model with inventory management." In addition to choosing the weight on his current signal, the IFT also chooses the rate at which he mean reverts his inventory to zero each period. Suppose the IFT does inventory management (i.e., chooses a positive rate of inventory mean reversion), but not necessarily optimally.

The first effect of inventory management is that the IFT keeps all his profits in cash. To see this, suppose the IFT chooses a coefficient of mean reversion of 1%. This translates into the inventory being reduced by a fraction of 1% in each trading round. Therefore, the IFT's inventory tends to become small over many rounds, and because the model is set in the high-frequency limit (in continuous time), the inventory becomes in fact negligible.[8] I call this result the "low inventory effect."

The second effect is that the IFT no longer makes profits by betting on the fundamental value of the asset. This stands in sharp contrast to the behavior of a risk-neutral speculator, such as the fast trader in the benchmark model (with no IFT). Indeed, the FT accumulates inventory in the direction of his information, since he knows his signals are correlated with the asset's liquidation value. By contrast, although the IFT initially trades on his current signal, he subsequently fully reverses the bet on that signal by removing a fraction of his inventory in each trading round. Thus, the IFT's direct profit from each signal eventually decays to zero. I call this result the "information decay effect."

The third effect of inventory management is that, in order to make a profit, the IFT must (i) anticipate the slow trading, and (ii) trade in the opposite direction to slow trading. By "slow trading" here I simply mean the part of order flow that involves the speculators' lagged signals.[9] To understand this effect, consider how the IFT uses a given signal. The information decay effect means that the IFT's eventual profit from betting on his signal are zero. Therefore, the IFT must benefit from inventory reversal. Since any trade has a price impact, inventory reversal generates a profit only if it gets pooled with order flow in the opposite direction, so that the IFT's price impact is negative. But in order for this profit to exist on average, the opposite order flow must come from speculators who use lagged signals, that is, from slow trading. I call this result the "hot potato effect," or the "intermediation chain effect."[10]

The reason behind this terminology is that the IFT's current signal (the "potato") produces undesirable inventory (is "hot") and must be passed on to slower traders in order to produce a profit. Thus, speed is important to the IFT. Without slower trading, there is no hot potato effect, and the IFT makes a negative expected profit from any trading strategy that mean reverts his inventory to zero. Note also that the hot potato generates a complementarity between the IFT and slow traders: stronger inventory mean reversion by the IFT reduces the price impact of the STs, who can trade more aggressively. However, more aggressive trading by the STs allows stronger mean reversion from the IFT.

Fig. 1 illustrates the optimal behavior of the IFT as a function of his inventory aversion coefficient.[11] There are two contrasting types of behavior, depending on how his inventory aversion compares to a threshold. Below the threshold, the IFT behaves like a risk-neutral speculator, and makes money from taking fundamental bets on his signals. The only difference is that with increasing inventory aversion, he optimally reduces the weight on his signal, to reduce his inventory costs. He does not mean revert his inventory at all, because of the information decay effect: indeed, even a very small inventory mean reversion would eventually destroy all profits from the fundamental bets. Above the threshold, the IFT's optimal behavior changes dramatically:

[5] SEC (2010) characterizes HFTs by their "very short time-frames for establishing and liquidating positions" and argues that HFTs end "the trading day in as close to a flat position as possible (that is, not carrying significant, unhedged positions over-night)." See also Menkveld (2013), Brogaard et al. (2015), and Kirilenko et al. (2017).

[6] Lyons (1997) studies a model of "hot potato" trading among dealers in FX markets. Weller (2012) analyzes both theoretically and empirically "intermediation chains" in which uninformed HFTs unwind inventories to slower, fundamental traders. Glode and Opp (2016) study intermediation chains theoretically in OTC markets with asymmetric information. Kirilenko et al. (2017) mention a hot potato effect during the Flash Crash episode of May 6, 2010, when some HFTs churned out their inventories very quickly to trade with other HFTs.

[7] The IFT is assumed fast because without slower traders managing inventory is not profitable. The case of several IFTs is discussed in the Subsection 5.5 in the Internet Appendix, but the results are qualitatively similar.

[8] Formally, the inventory follows an autoregressive process, hence its variance has the same order as the variance of the signal, which at high frequencies is negligible.

[9] A subtle point is that slow trading does not need to come from actual slow traders. Slow trading can also arise from fast traders who use their lagged signals as part of their optimal trading strategy.

[10] In my simplified framework, the intermediation chain only has one link, between the IFT and the slow traders. In Subsection 5.6 in the Internet Appendix, I provide an extension in which speculators use more than one lag for their signals, and I obtain an intermediation chain with two links.

[11] Inventory aversion is similar to risk aversion, but solving the model with a risk-averse fast trader would be considerably more difficult.

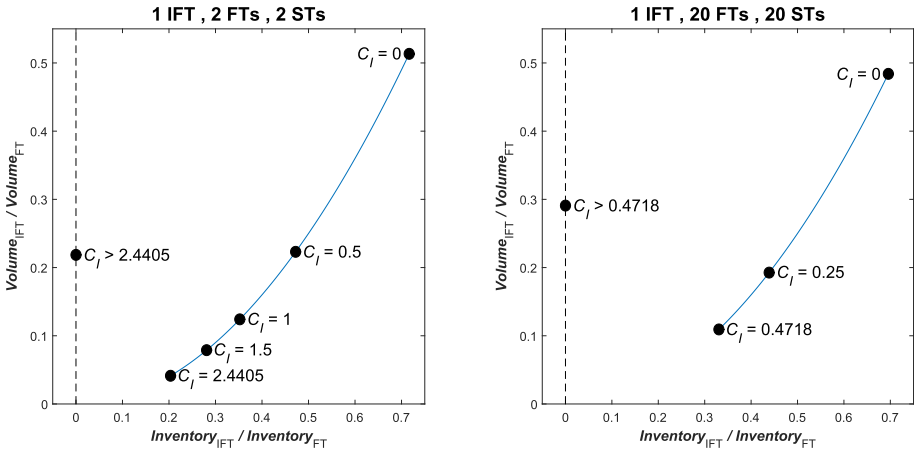

Fig. 1. Optimal inventory mean reversion. This figure shows the inventory and trading volume of an inventory-averse fast trader (IFT) for different values of his inventory aversion coefficient C_I, when the IFT competes with N_F fast traders (FTs) and N_S slow traders (STs). On the horizontal axis is the IFT's inventory, measured by the square root of his average expected squared position in the stock, relative to a FT's inventory. On the vertical axis is the IFT's trading volume, measured by the instantaneous variance of his trading strategy, relative to a FT's trading volume. The IFT's trading strategy is his best response, taking as fixed the equilibrium behavior of the FTs and STs as described in Theorem 3 below, with parameters $\sigma_w = 1, \sigma_u = 1$, and with $N_F = N_S$ equal to either 2 or 20.

he trades more aggressively on his signal, and at the same time engages in quick inventory mean reversion. As a result, compared to below the threshold, his trading volume spikes up yet his inventory remains essentially zero at all times. Note that the threshold at which the behavior discontinuity occurs is decreasing in the number of fast traders or slow traders, as both provide more of the slow trading necessary for the IFT to manage his inventory. Thus, even with small values of the inventory aversion coefficient, the IFT can find it optimal to engage in inventory management and keep all his profits in cash.

My results speak to the literature on high-frequency trading. One may think that in practice HFTs have very low inventories because either (i) HFTs have very high risk aversion, or (ii) HFTs do not have superior information and wish to maintain zero inventory to avoid averse selection on their positions in the risky asset. My results suggest that this is not necessarily the case. Indeed, Fig. 1 suggests (and I rigorously prove in Proposition 7) that in the limit when the number of speculators is large, the threshold inventory aversion converges to zero, and the optimal mean reversion rate is close to one. In other words, even with low inventory aversion, the IFT chooses very large mean reversion. Yet, even at these high rates of mean reversion, the IFT does not lose more than about 50% of his average profits from inventory management (the advantage being that he has all his profits in cash).

I predict that in practice the fast speculators are sharply divided into two categories. In both categories, speculators trade with a large volume. However, in one category speculators accumulate inventory by taking fundamental bets. In the other category, speculators have very low inventories; they initially trade on their signals but then quickly pass on part of their inventory to slower traders. These covariance patterns produce testable implications of the model.

The division of fast speculators into two categories appears consistent with the empirical findings of Kirilenko et al. (2017), who study trading activity in E-mini S&P 500 futures during several days around the Flash Crash of May 6, 2010. The "opportunistic traders" described in their paper resembles the risk-neutral fast traders in my model: opportunistic traders have large volume, appear to be fast, and accumulate relatively large inventories. By contrast, the HFTs in their paper, while they are also fast and trade in large volume, keep very low inventories. Indeed, the HFTs in their sample liquidate 0.5% of their aggregate inventories on average each second.

1.1. Related literature

This paper contributes to the literature on trading with asymmetric information. I show that competition among informed traders, combined with noisy trading strategies, produces a large informed trading volume and a quick information decay.[12] The market is very efficient because competition among informed traders makes them trade aggressively on their common information. This intuition is present in Holden and Subrahmanyam (1992), and Foster and Viswanathan (1996). The former finds that the competition among informed traders is so strong, that in the continuous-time limit there is no equilibrium in smooth strategies. My contribution to this literature is to show that there exists an equilibrium in noisy strategies. This rests on two key assumptions: (i) noisy information, i.e., speculators learn over time by observing a stream of signals, and (ii) finite lags, i.e., speculators only use a signal for a fixed number of lags, which is plausible if there is a positive information processing cost per signal.

[12] A speculator's strategy is "smooth" if the volatility generated by that speculator's trades is of a lower magnitude compared to the volatility from noise trading; and "noisy" if the magnitudes are the same.

Without the finite lags assumption, noisy information by itself does not generate noisy strategies, as Back and Pedersen (1998) show. Chau and Vayanos (2008) and Caldentey and Stacchetti (2010) find that noisy information coupled with either model stationarity or a random liquidation deadline produces strategies that are still smooth as in Kyle (1985), but towards the high-frequency limit they have almost infinite weight. Thus, the market in these papers is nearly strong-form efficient, which makes speculators' strategies appear noisy (there is no actual equilibrium in the limit). By contrast, in my model, the market is not strong-form efficient even in the limit, yet strategies are noisy. Foucault et al. (2016) propose a model in which a single speculator receives a signal one instant before public news. The speculator's strategy is noisy, but for a different reason than in my model: the speculator optimally trades with a large weight on his forecast of the news.[13]

My paper also contributes to the rapidly growing literature on high-frequency trading, e.g., Cartea and Penalva (2012), Weller (2012), Hoffmann (2014), Biais et al. (2015), Budish et al. (2015), Foucault et al. (2016), Aït-Sahalia and Sağlam (2017), Du and Zhu (2017), Li (2017), Pagnotta and Philippon (2018); see also the survey by Menkveld (2016). In much of this literature, it is the speed difference that has a large effect in equilibrium. The usual model setup has certain traders who are faster in taking advantage of an opportunity that disappears quickly. As a result, traders enter into a winner-takes-all contest, in which even the smallest difference in speed has a large effect on profits.[14] By contrast, my results regarding volume and volatility remain true even if all informed traders have the same speed. This is because in my model the need for speed arises endogenously, from competition among informed traders. In my model, being "slow" simply means trading on lagged signals. Since in equilibrium speculators also use lagged signals (the unanticipated part, to be precise), in some sense all traders are slow as well. Yet, it is true in my model that a genuinely slower trader makes less money, since he can only trade on older information that has already lost much of its value.

The results in this paper regarding the optimal inventory of informed traders are, to my knowledge, new. Theoretical models of inventory usually attribute inventory mean reversion to passive market makers, who do not possess superior information, but are concerned with absorbing order flow (e.g., Ho and Stoll, 1981; Madhavan and Smidt, 1993; and Hendershott and Menkveld, 2014). This paper shows that an informed investor with inventory costs (the "IFT") can display behavior that makes him appear like a market maker, even though he only submits market orders, as in Kyle (1985). Indeed, in my model the IFT does not take fundamental bets, passes his risky inventory to slower traders (the hot potato effect), and keeps all his money in cash. To obtain these results, even a small inventory aversion of the IFT suffices, but only if enough slow trading exists.

A related paper is Hirshleifer et al. (1994). In their 2-period model, risk-averse speculators with a speed advantage first trade to exploit their information, after which they revert their position because of risk aversion, while the slower speculators trade in the same direction as the initial trade of the faster speculators. The focus of Hirshleifer et al. (1994) is different, as they are interested in information acquisition and explaining behavior such as "herding" and "profit taking." My goal is to analyze the inventory problem of fast informed traders in a fully dynamic context, and to study the properties of the resulting optimal strategies.

The paper is organized as follows. In Section 2, I describe the model setup. In Section 3, I solve for the equilibrium in the particular case with two categories of traders: fast and slow, and discuss the effect of fast and slow traders on various measures of market quality. In Section 4, I introduce an extension of the benchmark model in which a new trader (the IFT) has inventory costs. Then, I analyze the IFT's optimal strategy and its effect on equilibrium. In Section 5, I discuss the robustness of the main results to various extensions. Section 6 concludes. All proofs are in the Appendix or the Internet Appendix. In the Internet Appendix, I provide solutions for the equilibrium in the general case, and analyze several modifications and extensions of the benchmark model.

2. Benchmark model

I set the trading model in discrete time in order to describe the equilibrium as the number of trading periods approaches infinity, and the setup approaches a continuous-time model on $[0, 1]$.[15] I thus consider a discrete model with T periods, where the time interval $\Delta t = \frac{1}{T}$ is the discrete analog of the infinitesimal interval dt of continuous time. Trading takes place at times th, where $t = 1, 2, \ldots, T$ and $h = \Delta t > 0$ (e.g, Chau and Vayanos, 2008). The level of a variable v at the time th is denoted by v_t, and its change is denoted by $\Delta v_t = v_t - v_{t-1}$.

The liquidation value of the asset is v_T, where:

$$v_T = \sigma_v B_T^v = \sum_{t=1}^{T} \sigma_v \, \Delta B_t^v, \tag{1}$$

[13] In Cao et al. (2015) traders' strategies are also noisy: informed traders must disclose their orders immediately after trading, and therefore optimally obfuscate their signal by adding a large noise component to their trades.

[14] See, e.g., the model with speed differences of Biais et al. (2015), or the model of news anticipation of Foucault et al. (2016). Other models feature differential access to fundamental information, e.g., Bernhardt and Miao (2004) and Albuquerque and Miao (2014), or differential access to price information, e.g., Cespa and Foucault (2014) and Easley et al. (2016). These other papers, however, do not address directly the effect of speed on traders' strategies and their profits.

[15] Alternatively, one can consider a continuous-time model over $[0, 1]$ where the trading intervals are of infinitesimal length dt (e.g., Foucault et al., 2016). In that case, however, the trading strategies are not usual Itô processes (since some traders use lagged signals), and thus traders' profits cannot be computed with the Itô integral. My solution is to consider the discretized model, and define traders' profits as the limit when the number of periods approaches infinity.

where B^v is a (continuous) Brownian motion over $[0, 1]$, and $\sigma_v > 0$ is a constant called the "fundamental volatility." I interpret v_T as the "long-run" value of the asset; in the high-frequency world, this can be taken to be the asset value at the end of the trading day. The increments Δv_t are then the short-term changes in value due to the arrival of new information. The risk-free rate is assumed zero.

There are three types of market participants: (a) $N \geq 1$ risk-neutral speculators, who observe the flow of information at different speeds, as described below; (b) noise traders; and (c) one competitive risk-neutral dealer, who sets the price at which trading takes place.

2.1. Information and speed

Speculators have the same trading speed, but differ in the speed of processing information. To abstract away from the issue of forecasting the forecasts of others, as described by Foster and Viswanathan (1996), I assume that speculators receive the same signal each period, but differ in the number of lags at which they receive the signal. At $t = 0$, there is no information asymmetry between the speculators and the dealer, as $v_0 = 0$. Subsequently, each speculator receives the following flow of signals:

$$\Delta s_t = \Delta v_t + \Delta \eta_t, \text{ with } \Delta \eta_t = \sigma_\eta \, \Delta B^\eta_t, \tag{2}$$

where $t = 1, 2, \ldots, T$ and B^η is a Brownian motion over $[0, 1]$ independent from all other variables. Denote:

$$w_t = \mathsf{E}(v_T \mid \{s_\tau\}_{\tau \leq t}) \tag{3}$$

the expected value conditional on the information flow until t. I call w_t the "value forecast," or simply "forecast." Because there is no initial information asymmetry, $w_0 = 0$. Denote by σ_w the instantaneous volatility of w_t, or the "forecast volatility." The increment of the forecast w_t, and the forecast variance are, respectively,

$$\Delta w_t = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2} \, \Delta s_t, \quad \sigma_w^2 = \frac{\mathsf{Var}(\Delta w_t)}{\Delta t} = \frac{\sigma_v^4}{\sigma_v^2 + \sigma_\eta^2}. \tag{4}$$

When deriving empirical implications, I call σ_w the "signal precision," as a precise signal (small σ_η) corresponds to a large σ_w.

Speculators obtain their signal with a lag $\ell \in \{0, 1, 2, \ldots, T - 1\}$. A "$\ell$-speculator" is a trader who at $t = 1, 2, \ldots, T$ observes the signal from ℓ periods before, $\Delta s_{t-\ell}$.

2.2. Trading and prices

At each $t = 1, 2, \ldots, T$, denote by Δx^i_t the market order submitted by speculator $i = 1, \ldots, N$ at t, and by Δu_t the market order submitted by the noise traders, which is of the form $\Delta u_t = \sigma_u \Delta B^u_t$, where B^u is a Brownian motion independent from all other variables. Then, the aggregate order flow executed by the dealer at t is:

$$\Delta y_t = \sum_{i=1}^{N} \Delta x^i_t + \Delta u_t. \tag{5}$$

The dealer is risk-neutral and competitive, hence she executes the order flow at a price equal to her expectation of the liquidation value conditional on her information. Let $\mathcal{I}_t = \{y_\tau\}_{\tau < t}$ be the dealer's information set just before trading at t. The order flow at date t, Δy_t, executes at:

$$p_t = \mathsf{E}\left(v_T \middle| \, \mathcal{I}_t \cup \Delta y_t\right). \tag{6}$$

Together with the price, another important quantity is the dealer's expectation at t of the k-lagged signal Δw_{t-k}:

$$z_{t-k,t} = \mathsf{E}\left(\Delta w_{t-k} \mid \mathcal{I}_t\right). \tag{7}$$

2.3. Equilibrium definition

In general, a trading strategy for an ℓ-speculator is a process followed by his risky asset position, x_t, which is measurable with respect to his information set $\mathcal{J}^{(\ell)}_t = \{y_\tau\}_{\tau < t} \cup \{s_\tau\}_{\tau \leq t-\ell}$. For a given trading strategy, the speculator's expected profit π_τ, from date τ onwards, is:

$$\pi_\tau = \mathsf{E}\left(\sum_{t=\tau}^{T} (v_T - p_t)\Delta x_t \mid \mathcal{J}^{(\ell)}_\tau\right). \tag{8}$$

As in Back et al. (2000), I focus on linear equilibria in which the trading strategy has a particular dependence on the traders' forecasts. Specifically, I consider strategies that are linear in the unpredictable part of their signals[16]:

$$\widetilde{\Delta w}_{t-k,t} = \Delta w_{t-k} - z_{t-k,t}, \quad k = \ell,\ \ell+1,\ \ldots \tag{9}$$

I restrict strategies to exclude signals older than a fixed number of lags m (which is allowed to depend on the speculator's speed parameter ℓ). This assumption can be justified by costly information processing, as explained at the end of this section. Formally, the ℓ-speculator's strategy is of the form:

$$\Delta x_t = \gamma_{\ell,t}\widetilde{\Delta w}_{t-\ell,t} + \gamma_{\ell+1,t}\widetilde{\Delta w}_{t-\ell-1,t} + \cdots + \gamma_{m,t}\widetilde{\Delta w}_{t-m,t}. \tag{10}$$

To focus on the equilibrium behavior when $\Delta t = \frac{1}{T}$ is small, I require that the ℓ-speculator's strategy is the discretization of a continuous-time strategy on $[0,1]$. Recall that the subscript t refers to the actual time $\frac{t}{T} \in [0,1]$. I thus require that the coefficients $\gamma_{k,t}$ of the strategy in (10) are continuous functions of time.[17] To indicate that this is a continuous-time strategy, I use differential notation:

$$\mathrm{d}x_t = \gamma_{\ell,t}\widetilde{\mathrm{d}w}_{t-\ell,t} + \gamma_{\ell+1,t}\widetilde{\mathrm{d}w}_{t-\ell-1,t} + \cdots + \gamma_{m,t}\widetilde{\mathrm{d}w}_{t-m,t}, \tag{11}$$

where t is still regarded as an element of $\{1,2,\ldots,T\}$. If instead one regards $t \in (0,1]$, then the subscript $t-k$ should be replaced by $t-k\mathrm{d}t$.[18] In the rest of the paper, I preserve the ambiguity of the notation in (11), but to avoid confusion I often write integrals over $t \in (0,T]$, and set $T = 1$.

For the strategies in (11), I define the expected profit as the (possibly infinite) limit of the discrete sums in (8) when T approaches infinity. With a slight abuse of notation, I use the integral sign to denote this limit[19]:

$$\pi_\tau = \mathsf{E}_\tau\left(\int_\tau^1 (v_T - p_t)\mathrm{d}x_t\right) = \lim_{T\to\infty} \mathsf{E}\left(\sum_{t\geq \tau T}^{T} (v_T - p_t)\Delta x_t \mid \mathcal{J}_\tau^{(\ell)}\right). \tag{12}$$

A linear equilibrium is such that: (i) each speculator chooses the coefficients $\gamma_{k,t}$ in the trading strategy (11) to maximize his expected trading profit (12) given the dealer's pricing policy, and (ii) the dealer's pricing policy given by (6) and (7) is consistent with the equilibrium speculator trading strategies.

Finally, the speculators take the covariance structure of $z_{t-k,t}$ to be independent of their strategy. More precisely, for all $j,k \geq 0$, the speculators consider the number:

$$Z_{j,k,t} = \mathsf{Cov}\left(\Delta w_{t-j}, z_{t-k,t}\right) \tag{13}$$

to depend only on j, k, and t. Thus, the covariance terms $Z_{j,k,t}$ are computed by the dealer, as part of her (publicly known) pricing rules.[20]

2.4. Model notation

If all speculators in the model have a strategy of the form (11) with the same $m \geq 0$, I call it the "benchmark model" with m lags, and write $\mathcal{M}_m$. I focus on the particular case with $m = 1$ lags. In this setup, the 0-speculators are called the "fast traders," and the 1-speculators are called the "slow traders." Thus, I also call $\mathcal{M}_1$ the "model with fast and slow traders."

If some ℓ-speculators have strategies of the form (11) with different m_ℓ, I call this the general model with m lags, where m is the maximum of all m_ℓ. An important case is the general model with $m = 1$ lags in which 0-speculators (fast traders) only trade on their current signal ($m_0 = 0$) and the 1-speculators (slow traders) only use their lagged signal ($m_1 = 1$). I call this the "general benchmark model," and denote it by $\mathcal{M}_{0,1}$. In Section 3, I solve for the equilibrium in both $\mathcal{M}_1$ and $\mathcal{M}_{0,1}$, and show that $\mathcal{M}_1$ can be regarded as a particular case of $\mathcal{M}_{0,1}$.

[16] Intuitively, if the strategy had a predictable component, the dealer's price would adjust and reduce the speculators' profit. The unpredictability of the speculators' strategies can be proved quite generally, following Kyle (1985), as long as the speculators and the dealer are risk-neutral.

[17] This requirement implies that the coefficients $\gamma_{k,t}$ are deterministic, and hence known at $t = 0$. Similar assumptions are made in other continuous-time models, e.g., Back et al. (2000). More generally, one can choose $\gamma_{k,t}$ to be integrable (but deterministic) functions of t.

[18] Indeed, as t corresponds to the actual time $t' = \frac{t}{T} \in (0,1]$, and 1 corresponds to $\frac{1}{T} = \Delta t$ with its infinitesimal version dt, if follows that $t-k$ corresponds to the actual time $t' - kdt$.

[19] One may be tempted to define the integral inside the expectation as an Itô integral, but this does not work, as x_t and p_t are not Itô processes. I thank the referees for pointing this out.

[20] For instance, the price impact coefficient λ_t in the dealer's pricing rule $\Delta p_t = \lambda_t \Delta y_t$ is computed using the covariance term $\mathsf{Cov}(w_t, \Delta y_t)$ (see equation (54)). Hence, even though a speculator affects Δy_t by his strategy, he can consider the covariance term $\mathsf{Cov}(w_t, \Delta y_t)$ to be independent of his strategy. I further discuss this assumption in Section 5.

2.5. Information processing

The assumption that speculators cannot use lagged signals beyond a given bound can be justified by introducing an information processing cost $\Delta > 0$ per individual signal and per unit of time. More precisely, I consider an alternative model in which an ℓ-speculator can use all past signals, but must pay a fixed cost $\Delta_\ell dt$ each time he trades with a nonzero weight ($\gamma_{k,t}$) on his k-lagged signal (see equation (11)). Then, intuitively, because the value of information decays with the lag, and the speculator does not want to accumulate too large a cost, he must stop using lagged signals beyond an upper bound. In Result 1, I show that for a particular value of Δ, the alternative model is equivalent to $\mathcal{M}_1$.

In choosing speculator strategies as in (11), I make two implicit assumptions: that speculators (i) must process each signals individually, and (ii) cannot use their signals to learn about other speculators' forecasts. These assumptions can be justified by introducing specific information processing costs, but it is important for the intuition of the model to provide separate justification. Assumption (i) essentially prevents speculators from simply relying on free public aggregate signals, such as the price, to shortcut the learning process. This is because in reality prices may contain other relevant information about the fundamental value, along which the speculators are adversely selected.[21] Assumption (ii) is made for convenience, to avoid the problem of forecasting the forecasts of others described by Foster and Viswanathan (1996). This is not an issue in the benchmark model $\mathcal{M}_1$, but does become a problem when speculators use signals of lag at least two. Even then, I show in an extension of the model (Subsection 2.2 in the Internet Appendix) that the main predictions of the benchmark model remain qualitatively the same. In Section 5, I discuss assumptions (i) and (ii) in more detail.

3. Fast and slow traders

In this section, I analyze the important case in which speculators use signals with a maximum lag of one. There are two types of speculators: (i) the Fast Traders (FTs), who observe the signal with no delay (called 0-speculators in Section 2); and (ii) the Slow Traders (STs), who observe the signal with a delay of one lag (called 1-speculators). As in (11), the trading strategy of FTs and STs is of the form:

$$dx_t = \gamma_t(dw_t - z_{t,t}) + \mu_t(dw_{t-1} - z_{t-1,t}), \quad t \in (0, T], \tag{14}$$

where $T = 1$. Note that the weight γ_t must be zero for a ST. There are two possibilities: either the FT can trade on both the current and the lagged signals, or the FT can trade only on the current signal (i.e., the FT's weight γ_t must be zero).[22] The former case is the benchmark model $\mathcal{M}_1$. The latter case is the general benchmark model $\mathcal{M}_{0,1}$.

Note that the FT's current signal (dw_t) is orthogonal on the past order flow, hence the dealer sets $z_{t,t} = 0$. To simplify notation, let $\widetilde{dw}_{t-1} = \widetilde{dw}_{t-1,t}$ be the unanticipated part at t of the lagged signal. Then, the trading strategy in (14) can be written as:

$$dx_t = \gamma_t dw_t + \mu_t \widetilde{dw}_{t-1}, \quad \text{with} \quad \widetilde{dw}_{t-1} = dw_{t-1} - z_{t-1,t}. \tag{15}$$

3.1. Equilibrium

I solve for the equilibrium of the model $\mathcal{M}_1$ in closed form. One important implication is that the FTs and STs trade identically on their lagged signal (μ_t is the same for all). Therefore, if one requires the FTs to use only their current signal (as in $\mathcal{M}_{0,1}$) and introduce an equal number of additional STs, then the aggregate behavior remains essentially the same. Hence, the model $\mathcal{M}_1$ can be regarded as a particular case of $\mathcal{M}_{0,1}$, which justifies calling $\mathcal{M}_{0,1}$ the "general benchmark model." In fact, the latter model can also be solved in closed form, by using essentially the same formulas.

Theorem 1 shows that a closed-form linear equilibrium of the model exists. The equilibrium is symmetric, in the sense that the FTs have identical trading strategies, and so do the STs. I also provide asymptotic results when the number of FTs is large.

Theorem 1. *Let $N_F > 0$ be the number of FTs and $N_S \geq 0$ the number of STs, and define $N_L = N_F + N_S$ (the number of lag traders). Then, there exists a symmetric linear equilibrium with constant coefficients, such that for all $t \in (0, T]$:*

$$dx_t^F = \gamma dw_t + \mu \widetilde{dw}_{t-1}, \quad dx_t^S = \mu \widetilde{dw}_{t-1},$$

$$\widetilde{dw}_{t-1} = dw_{t-1} - \rho dy_{t-1}, \quad dp_t = \lambda dy_t,$$

where the coefficients γ, μ, ρ, λ are:

[21] I formalize this intuition in Section 4 in the Internet Appendix, where I introduce an orthogonal dimension of the fundamental value, and show that trading strategies that rely on prices make an average loss.

[22] Intuitively, this can occur if the FT must pay a higher processing cost per signal than the ST; see Footnote 26.

$$\gamma = \frac{1}{\lambda}\,\frac{1}{N_F+1}, \quad \mu = \frac{1}{\lambda}\,\frac{1}{N_L+1}\,\frac{1}{1+b},$$
$$\rho = \frac{\sigma_w}{\sigma_u}\,\sqrt{(1-a)(a-b^2)}, \quad \lambda = \rho\,\frac{N_F}{N_F-b},$$

with $\omega = 1 + \frac{1}{N_F}\,\frac{N_L}{N_L+1} \in [1,2)$, $b = \frac{1}{2}\left((\omega^2 + 4\frac{N_L}{N_L+1})^{1/2} - \omega\right) \in [0, b_\infty)$, $a = \frac{N_F-b}{N_F+1} \in (0,1)$, *with the following asymptotic limits when* N_F *is large:* $\omega_\infty = a_\infty = 1, b_\infty = \frac{1}{2}(\sqrt{5}-1), \lambda_\infty = \rho_\infty = \frac{\sigma_w}{\sigma_u}\,\frac{1}{\sqrt{N_F}}$.[23] *The number* b *is increasing in both* N_F *and* N_S.

Theorem 1 implies that FTs and STs trade with the same intensity (μ) on their lagged signals. This is true because the current signal dw_t is uncorrelated with the lagged signal $\widetilde{dw}_{t-1}$, which implies that the FTs and the STs get the same expression for the expected profit that comes from the lagged signal.[24]

I now discuss some comparative statics regarding the optimal weights γ and μ (for brevity, I omit the proofs). The FTs' optimal weight γ is decreasing in the number of FTs, yet it is increasing in the number of STs. The first statement simply reflects that, when the number of FTs is larger, these traders must divide the pie into smaller slices. The same logic applies to the coefficient on the lagged signal: μ is decreasing in both N_F and N_S, as the FTs and STs compete in trading on their common lagged signal. This last intuition also shows that the FTs' weight γ is increasing in the number of STs. Indeed, when there is more competition from STs, the FTs have an incentive to trade more aggressively on their current signal, as the STs have not yet observed this signal.

In Corollary 1, I state a few implications of Theorem 1 that help provide more intuition for the equilibrium.

Corollary 1. *In equilibrium, the following formulas hold:*

$$\lambda\,\overline{\gamma} = \frac{N_F}{N_F+1}, \quad \lambda\,\overline{\mu} = \frac{1}{1+b}\,\frac{N_L}{N_L+1},$$
$$\frac{\mathrm{Var}\left(\widetilde{dw}_t\right)}{dt} = (1-a)\,\sigma_w^2 = \frac{1+b}{N_F+1}\,\sigma_w^2, \quad \frac{\mathrm{Cov}\left(\widetilde{dw}_t, w_t\right)}{dt} = \frac{1-a}{1+b}\,\sigma_w^2 = \frac{\sigma_w^2}{N_F+1}. \tag{18}$$

The first equation in (18) implies that $\lambda\overline{\gamma}dw_t = \frac{N_F}{N_F+1}dw_t$, which shows that most of the current signal (dw_t) is incorporated into the price by the FTs. The intuition comes from the Cournot nature of the equilibrium. Indeed, when trading on the current signal, the benefit of each FT increases linearly with the intensity of trading γ on his signal, while the price at which he eventually trades increases linearly with the aggregate quantity demanded. Given that the price impact of the other $N_F - 1$ FTs aggregates to $\frac{N_F-1}{N_F+1}dw_t$, the FT is a monopsonist against the residual supply curve, and trades such that his price impact is half of $\frac{2}{N_F+1}dw_t$, that is, his price impact equals $\frac{1}{N_F+1}dw_t$.

After incorporating $\frac{N_F}{N_F+1}dw_t$ in trading round t, the FTs must compete with the STs for the remaining $\frac{1}{N_F+1}dw_t$ in the next trading round. As explained before, the speculators must trade a multiple of the unanticipated part of the lagged signal, $\widetilde{dw}_t = dw_t - \rho dy_t$. Thus, when trading on the lagged signal, the benefit of each speculator—fast or slow—increases linearly with the intensity of trading μ, and is proportional to the covariance $\mathrm{Cov}\left(\widetilde{dw}_t, w_t\right)$. At the same time, each speculator faces a price that increases linearly with the aggregate quantity demanded, and which is proportional to the lagged signal variance $\mathrm{Var}\left(\widetilde{dw}_t\right)$. The argument is now similar to the Cournot one above, except that everything gets multiplied by the ratio $\mathrm{Cov}(\widetilde{dw}_t, w_t)/\mathrm{Var}(\widetilde{dw}_t)$, which according to (18) is equal to $1/(1+b)$. This justifies the second equation in (18). It also implies that in the case of the lagged signal, only a fraction $1/(1+b)$ of it is incorporated the price by the speculators.

In Proposition 1, I compute the expected profits of the FTs and the STs.

Proposition 1. *In equilibrium, the expected profit at* $t = 0$ *of a FT and a ST satisfies, respectively,*

$$\frac{\pi^F}{\sigma_w^2} = \frac{\gamma}{N_F+1} + \frac{1}{N_F+1}\,\frac{\mu}{N_L+1}, \quad \frac{\pi^S}{\sigma_w^2} = \frac{1}{N_F+1}\,\frac{\mu}{N_L+1}. \tag{19}$$

Thus, the FT-to-ST expected profit ratio is $\frac{\pi^F}{\pi^S} = 1 + \frac{(N_L+1)^2(1+b)}{N_F+1}$, *which implies that* $\frac{\pi^S}{\pi^F} \approx \frac{N_F}{(N_F+N_S)^2}\,\frac{1}{1+b_\infty}$ *when* N_F *is large.*

[23] If X is a variable that depends on N_F, I say that X_∞ is the asymptotic value of a number X, and write $X \approx X_\infty$, whenever the ratio X/X_∞ converges to 1 as N_F approaches infinity.

[24] This result does not generalize to the case when there are more lags ($m > 1$). In Section 1 in the Internet Appendix, there is a positive autocorrelation between the signals of higher lags, which reflects a more complicated covariance structure. Mathematically, this translates into the covariance matrix A having nonzero entries $A_{i,j}$ when $i > j \geq 1$.

Thus, even if there is only one ST (i.e., $N_S = 1$), the ST's profit is small compared to a FT's profit. The reason is that FTs also trade on their lagged signals, and thus compete with the ST.[25] Indeed, FTs compete for trading on dw_t only among themselves, while they also compete with the STs for trading on the lagged signal $\widetilde{dw}_{t-1}$.

Finally, Proposition 1 gives an estimate for the information processing cost Δ that would be sufficient to discourage speculators from trading on lagged signals beyond one, if that were not imposed by the model. I state the following numerical result.

Result 1. *Consider the alternative setup with N_F FTs and N_S STs, in which each speculator can use past signals at any lag, but must pay for each signal (used with nonzero weight) an information processing cost $\delta = \frac{1}{N_F+1}\frac{\mu}{N_L+1}\sigma_w^2$. Then, the alternative model is equivalent to the benchmark model $\mathcal{M}_1$.*

I now consider the general benchmark model $\mathcal{M}_{0,1}$, in which the FTs use only the current signal, while the STs use only the lagged signal.[26] The strategies of the FTs and STs are, respectively, of the form $dx_t^F = \gamma_t dw_t$ and $dx_t^S = \mu_t \widetilde{dw}_{t-1}$, where $\widetilde{dw}_{t-1} = dw_{t-1} - \rho_t dy_{t-1}$. The dealer sets the price using the rule $dp_t = \lambda_t dy_t$. Let $N_F \geq 1$ be the number of FTs and $N_L \geq 0$ the number of STs.

Corollary 2 shows that the model $\mathcal{M}_1$ with N_F FTs and N_S STs produces essentially the same outcome as the benchmark model $\mathcal{M}_{0,1}$ with N_F FTs and $N_L = N_F + N_S$ STs.

Corollary 2. *Consider (a) the model $\mathcal{M}_1$ with $N_F \geq 1$ FTs and $N_S \geq 0$ STs, and (b) the model $\mathcal{M}_{0,1}$ with N_F FTs and $N_L = N_F + N_S$ STs. Then, the equilibrium coefficients γ, μ, λ, and ρ in the two models are identical.*

Because of this equivalence, in the rest of the paper I also call the model $\mathcal{M}_{0,1}$ the benchmark model. There are two important cases:

- If $N_L \geq N_F$, the benchmark model is equivalent to the model $\mathcal{M}_1$ with N_F FTs and $N_S = N_L - N_F$ STs;
- If $N_L = 0$, the benchmark model is the model $\mathcal{M}_0$, with 0 lags.

3.2. Market quality

I next examine the effect of fast and slow trading on various measures of market quality. Following Corollary 2, I consider the benchmark model in which $N_F \geq 1$ FTs trade only on the current signal, and $N_L \geq 0$ STs trade on the lagged signal. I define "fast trading" as the speculators' aggregate trading on their current signal, and "slow trading" as the speculators' aggregate trading on their lagged signal.

To define measures of market quality, I first decompose the aggregate speculator order flow into fast trading and slow trading. Denote by $d\overline{x}_t$ the aggregate speculator order flow. Let $\overline{\gamma}$ be the aggregate weight on the current signal (dw_t), and $\overline{\mu}$ the aggregate weight on the lagged signal ($\widetilde{dw}_{t-1}$). I decompose the aggregate speculator order flow $d\overline{x}_t$ into two components:

$$d\overline{x}_t = \underbrace{\overline{\gamma}\, dw_t}_{\text{Fast Trading}} + \underbrace{\overline{\mu}\, \widetilde{dw}_{t-1}}_{\text{Slow Trading}}, \quad \text{with} \quad \overline{\gamma} = N_F\gamma, \quad \overline{\mu} = N_L\mu. \tag{20}$$

I call $b = \rho\overline{\mu}$ (defined in Theorem 1) the "slow trading coefficient." Then, slow trading exists (is nonzero) only if the number of traders who use their lagged signal is positive, or equivalently if $b > 0$. Note that the case when there is no slow trading coincides with the model $\mathcal{M}_0$ with 0 lags from Section 2. In that case, N_F FTs use only their current signal.

I now define the measures of market quality. Recall that the dealer sets a price that changes in proportion to the total order flow $dy = d\overline{x}_t + du_t$:

$$dp_t = \lambda\, dy_t = \lambda\left(\overline{\gamma}\, dw_t + \overline{\mu}\, \widetilde{dw}_{t-1} + du_t\right). \tag{21}$$

First, as it is standard in the literature, I measure illiquidity by the price impact coefficient λ. Thus, the market is considered illiquid if the price impact of a unit of trade is large, that is, if the coefficient λ is large.

Second, I define trading volume as the infinitesimal variance of the aggregate order flow dy_t, that is, $TV = \sigma_y^2 = \frac{\text{Var}(dy_t)}{dt}$. I argue that this is a measure of trading volume. Indeed, in each trading round the actual aggregate order flow is given by dy_t. Thus, one can interpret trading volume as the absolute value of the order flow: $|dy_t|$. From the theory of normal variables, the average trading volume is given by $\mathsf{E}\left(|dy_t|\right) = \sqrt{\frac{2}{\pi}}\,\sigma_y$. As $TV = \sigma_y^2$, it follows that TV is monotonic in $\mathsf{E}\left(|dy_t|\right)$, and thus TV can be used a measure of trading volume. Using (21), the trading volume satisfies:

[25] If instead the FTs traded only on their current signal, and only the ST used his lagged signal, then the formula (19) would still be correct if one set $N_L = 1$. In that case, the profit ratio $\pi^F/\pi^S = 1 + 4(1+b)/(N_F+1)$ would still be larger than one.

[26] As in Result 1, $\mathcal{M}_{0,1}$ is equivalent to an alternative setup with information processing costs, in which (i) the STs pay the cost Δ from Result 1, while (ii) the FTs pay a cost slightly higher than Δ. Indeed, if a FT paid Δ, he would be indifferent between using his lagged signal and not using it, while with a slightly higher cost, he would be strictly worse off and would ignore his lagged signal.

$$TV = \overline{\gamma}^2\,\sigma_w^2 + \overline{\mu}^2\,\sigma_{\widetilde{w}}^2 + \sigma_u^2, \quad \text{with} \quad \sigma_{\widetilde{w}}^2 = \frac{\mathrm{Var}\left(\widetilde{\mathrm{d}w}_t\right)}{\mathrm{d}t}. \tag{22}$$

The trading volume measure TV can be decomposed into the speculator trading volume and the noise trading volume: $TV = TV^s + TV^n$, with $TV^s = \overline{\gamma}^2\sigma_w^2 + \overline{\mu}^2\sigma_{\widetilde{w}}^2$ and $TV^n = \sigma_u^2$.

Third, I define price volatility as the square root of the instantaneous price variance, that is, $\sigma_p = \left(\frac{\mathrm{Var}(\mathrm{d}p_t)}{\mathrm{d}t}\right)^{1/2}$. From (21), it follows that the instantaneous price variance can be computed simply as the product of the illiquidity measure λ and the trading volume $TV = \sigma_y^2$. Thus,

$$\sigma_p^2 = \lambda^2\,TV = \lambda^2\,\left(\overline{\gamma}^2\,\sigma_w^2 + \overline{\mu}^2\,\sigma_{\widetilde{w}}^2 + \sigma_u^2\right). \tag{23}$$

Fourth, I define price informativeness as a measure inversely related to the forecast error variance $\Sigma_t = E((w_t - p_{t-1})^2)$. Thus, if prices are informative, they stay close to the forecast w_t (i.e., the variance Σ_t is small). In Section 1 in the Internet Appendix, in the general model with at most m lagged signals ($\mathcal{M}_m$), I show that Σ_t evolves according to $\Sigma_t' = \sigma_w^2 - \sigma_p^2$, where σ_p^2 is the price variance (Proposition IA.1). Therefore, since Σ_t' is inversely monotonic in the price variance, I do not use it as a separate measure of market quality.

Fifth, the speculator participation rate is defined as the ratio of speculator trading volume over total trading volume:

$$SPR = \frac{TV^s}{TV} = \frac{\overline{\gamma}^2\,\sigma_w^2 + \overline{\mu}^2\,\sigma_{\widetilde{w}}^2}{\overline{\gamma}^2\,\sigma_w^2 + \overline{\mu}^2\,\sigma_{\widetilde{w}}^2 + \sigma_u^2}. \tag{24}$$

SPR can also be interpreted as the fraction of price variance due to the speculators.

Proposition 2 provides explicit formulas for the measures of market quality. As before, I use asymptotic notation when N_F is large: $X \approx Y$ stands for $\lim_{N_F\to\infty} \frac{X}{Y} = 1$.

Proposition 2. *In the benchmark model with $N_F \geq 1$ FTs and $N_L \geq 0$ STs, the price impact coefficient, trading volume, price volatility, and speculator participation rate satisfy:*

$$\begin{aligned} &\lambda = \frac{\sigma_w}{\sigma_u}\,\frac{\sqrt{(1+b)(a-b^2)}}{\sqrt{N_F+1}}\,\frac{N_F}{N_F-b}, \quad TV = \sigma_u^2(N_F+1)\frac{a}{(1+b)(a-b^2)}, \\ &\sigma_p^2 = \sigma_w^2\,\frac{N_F^2}{(N_F+1)(N_F-b)}, \quad SPR = a + \frac{b^2(1+b)}{N_F-b}, \end{aligned} \tag{25}$$

where $b^2 + b\left(1 + \frac{1}{N_F}\,\frac{N_L}{N_L+1}\right) = \frac{N_L}{N_L+1}$, and $a = \frac{N_F-b}{N_F+1}$.

Panel A of Fig. 2 shows how the four measures of market quality vary with the number of FTs (N_F), while holding the number of STs (N_L) constant. Panel B of Fig. 2 shows how the four measures of market quality vary with N_L, while holding N_F constant. All four market quality measures vary in the same direction with respect to N_F and N_L. Nevertheless, the number of FTs has a much stronger effect on these measures than the number of STs.

To get more intuition about the effect of fast trading on market quality, I consider the simplest case, when $N_L = 0$. Since all speculators trade only on their current signal, this case coincides with the model $\mathcal{M}_0$ as defined in Section 2. In this model, there is no slow trading ($\overline{\mu} = 0$), hence the slow trading coefficient b is zero. Moreover, $a = \frac{N_F-b}{N_F+1} = \frac{N_F}{N_F+1}$. Thus, one can solve the model $\mathcal{M}_0$ by using Proposition 2. Nevertheless, in Proposition 3, I solve for the equilibrium of $\mathcal{M}_0$ independently.

Proposition 3. *Consider the model $\mathcal{M}_0$, with N_F FTs whose trading strategy is of the form $dx_t = \gamma_t dw_t$. Then, the optimal coefficient γ is constant and equal to $\gamma = \frac{1}{\lambda}\,\frac{1}{N_F+1} = \frac{\sigma_u}{\sigma_w}\,\frac{1}{\sqrt{N_F}}$. The price impact coefficient, trading volume, price volatility, and speculator participation rate satisfy, respectively,*

$$\lambda = \frac{\sigma_w}{\sigma_u}\,\frac{\sqrt{N_F}}{N_F+1}, \quad TV = \sigma_u^2(N_F+1), \quad \sigma_p^2 = \sigma_w^2\frac{N_F}{N_F+1}, \quad SPR = \frac{N_F}{N_F+1}. \tag{26}$$

Using Proposition 3, I discuss the effect of the number N_F of FTs on the measures of market quality. First, note that I obtain the same qualitative results for Proposition 3 as those displayed in Fig. 2. Namely, illiquidity is decreasing in N_F, while the other three measures are increasing in N_F.

An important consequence of Proposition 3 is that the speculator participation rate can be made arbitrarily close to 1 if the number of FTs is large. In that case, noise trading volatility is only a small part of the total volatility. This stands in sharp contrast for instance with the models of Kyle (1985) or Back et al. (2000), in which virtually all instantaneous price volatility is generated by the noise traders at the high-frequency limit (in continuous time).

The market is more efficient when the number of FTs is large. Indeed, in the Proof of Proposition 3, I show that the rate of change of the forecast error variance Σ' is constant and equal to $\frac{\sigma_w^2}{N_F+1}$. Since by assumption there is no initial informational

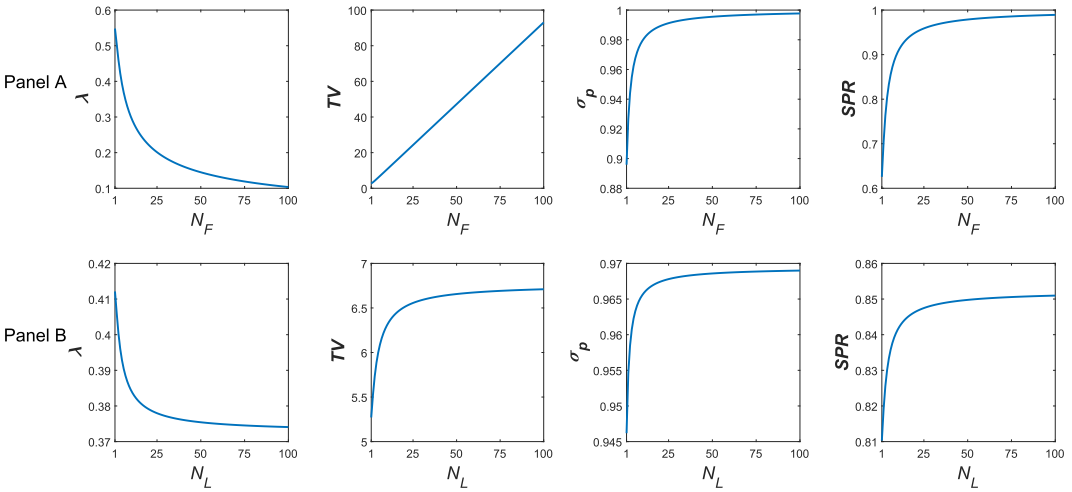

Fig. 2. Market quality with fast and slow traders. This figure shows the dependence of four market quality measures on the number of FTs (N_F) and the number of traders that use lagged signals (N_L). The four measures are: (i) the illiquidity λ, (ii) the trading volume TV, (iii) the price volatility σ_p, and (iv) the speculator participation rate SPR. In Panel A, I plot the four market quality measures against N_F, while N_L remains fixed at $N_L = 5$. In Panel B, I plot the four market quality measures against N_L, when N_F remains fixed at $N_L = 5$. The other parameters are $\sigma_w = 1$ and $\sigma_u = 1$.

asymmetry ($\Sigma_0 = 0$), it follows that $\Sigma_t \leq \frac{\sigma_w^2}{N_F+1}$ for all t. In other words, the price stays close to the fundamental value at all times. Thus, a larger number N_F of FTs, rather than destabilizing the market, makes the market more efficient.

The trading volume TV strongly increases with the number of FTs. This occurs because the competition among the FTs makes them trade more aggressively, and as a result they reveal more information. As explained below, this lowers the traders' price impact, which has an amplifying effect on trading. As a result, the trading volume grows essentially linearly in the number of FTs (see equation (26)). Moreover, the speculator participation rate SPR also increases in N_F, since SPR is the fraction of trading volume caused by the speculators.

Surprisingly, a larger number of FTs makes the market more liquid, as more information is revealed when there are more competing speculators. This appears to contradict the fact that more informed trading should increase adverse selection. To understand the source of this apparent contradiction, note that illiquidity is measured by the price impact λ of one unit of volume. But, while the trading volume TV increases in N_F in an unbounded way, its price impact is bounded by magnitude of the signal dw_t.[27] Thus, the price impact per unit of volume actually decreases, indicating that prices are more informative. This makes the market more liquid overall. This result is consistent with the empirical studies of Zhang (2010), Hendershott et al. (2011), and Boehmer et al. (2018a).

To understand the effect of FTs on the price volatility σ_p, consider the pricing formula $dy_t = \lambda dy_t$, which implies $\sigma_p^2 = \lambda^2 TV$. There are two effects of N_F on the price volatility σ_P. First, the trading volume TV increases in N_F, which has a positive effect on σ_P. Second, price impact λ decreases in N_F, which has a negative effect on σ_P. The first effect is slightly stronger than the second, hence the net effect is that price volatility σ_P increases in N_F. This result is consistent with the empirical studies of Zhang (2010) and Boehmer et al. (2018a).

A few caveats are in order. First, when discussing the effects of HFTs on market quality, the studies mentioned above do not proxy HFT activity by the number of HFTs present in the market, but by the HFTs' turnover or intensity of order-related message traffic. An answer to this concern is that, as already noted, trading volume does increase in the number of FTs. Second, in my paper I do not model "passive" HFTs, that is, HFTs that offer liquidity via limit orders. Therefore, it is possible that an increase in the number of passive HFTs decreases price volatility, which would cancel the opposite effect of the number of "active" HFTs. For instance, Hasbrouck and Saar (2013) document that HFTs exert a negative effect on volatility, possibly because they also consider passive HFTs, which by providing liquidity have a stabilizing effect on price volatility. Moreover, Chaboud et al. (2014) find essentially no relation. In my model, the dependence of price volatility on N_F is weak, which may explain the mixed results in the empirical literature.

3.3. Anticipatory trading

I start by analyzing the autocorrelation of the components of the order flow. Since the dealer is competitive and risk-neutral, the total order flow dy_t has zero autocorrelation. However, because the dealer cannot identify the part of the order flow that

[27] In Section 1 in the Internet Appendix, I make this intuition rigorous in the general case; see the discussion surrounding Proposition IA.4.

comes from speculators, the speculator order flow can in principle be autocorrelated.

As in Subsection 3.2, the aggregate speculator order flow decomposes into its fast trading and slow trading components:

$$\mathrm{d}\bar{x}_t = \underbrace{\mathrm{d}\bar{x}_t^F}_{\text{Fast Trading}} + \underbrace{\mathrm{d}\bar{x}_t^S}_{\text{Slow Trading}}, \quad \text{with} \quad \mathrm{d}\bar{x}_t^F = \bar{\gamma}\,\mathrm{d}w_t, \quad \mathrm{d}\bar{x}_t^S = \bar{\mu}\,\widetilde{\mathrm{d}w}_{t-1}, \tag{27}$$

where $\bar{\gamma} = N_F\gamma$ and $\bar{\mu} = N_L\mu$. As before, by definition, slow trading exists if $b = \rho\bar{\mu} > 0$, or equivalently if $N_L > 0$.

I define speculator order flow autocorrelation by $\mathrm{Corr}\left(\mathrm{d}\bar{x}_t, \mathrm{d}\bar{x}_{t+1}\right)$. Because $\mathrm{d}\bar{x}_{t+1}^F$ is orthogonal to both components of $\mathrm{d}\bar{x}_t^F$, I obtain the decomposition:

$$\rho_{\bar{x}} = \mathrm{Corr}\left(\mathrm{d}\bar{x}_t, \mathrm{d}\bar{x}_{t+1}\right) = \underbrace{\frac{\mathrm{Cov}\left(\mathrm{d}\bar{x}_t^F, \mathrm{d}\bar{x}_{t+1}^S\right)}{\mathrm{Var}(\mathrm{d}\bar{x}_t)}}_{\text{Anticipatory Trading}} + \underbrace{\frac{\mathrm{Cov}\left(\mathrm{d}\bar{x}_t^S, \mathrm{d}\bar{x}_{t+1}^S\right)}{\mathrm{Var}(\mathrm{d}\bar{x}_t)}}_{\text{Expectation Adjustment}}. \tag{28}$$

The "anticipatory trading" part is denoted by ρ_{AT}, and the "expectation adjustment" part by ρ_{EA}. The first component arises because fast trading at t anticipates slow trading at $t+1$. Indeed, there is a positive correlation between fast trading at t and slow trading at $t+1$ ($\bar{\mu}\widetilde{\mathrm{d}w}_t$). The second component arises because slow trading at $t+1$ is based on lagged signals, adjusted by subtracting the dealer's expectation which incorporates past lagged signals. Because of this expectation adjustment, the slow order flow is negatively autocorrelated. Formally, slow trading at $t+1$ ($\bar{\mu}\widetilde{\mathrm{d}w}_t$) is proportional to the lagged signal minus the dealer's expectation, $\widetilde{\mathrm{d}w}_t = \mathrm{d}w_t - \rho\mathrm{d}y_t$. But the dealer's expectation is proportional to the total order flow at t, which includes the previous slow trading ($\mathrm{d}y_t = \bar{\gamma}\,\mathrm{d}w_t + \bar{\mu}\,\widetilde{\mathrm{d}w}_{t-1} + \mathrm{d}u_t$). One obtains:

$$\rho_{\bar{x}} = \rho_{AT} + \rho_{EA}, \quad \text{with} \quad \rho_{AT} = \frac{}{\bar{\mu}\bar{\gamma}}\frac{\mathrm{Var}\left(\mathrm{d}w_t\right)}{\mathrm{Var}(\mathrm{d}\bar{x}_t)}, \quad \rho_{EA} = -\rho\bar{\mu}^3\frac{\mathrm{Var}\left(\widetilde{\mathrm{d}w}_{t-1}\right)}{\mathrm{Var}(\mathrm{d}\bar{x}_t)}. \tag{29}$$

Proposition 4 provides explicit formulas for the two components of the speculator order flow autocorrelation.

Proposition 4. *Consider the benchmark model with $N_F \geq 1$ FTs and $N_L \geq 0$ STs. Then, the speculator order flow autocorrelation and its components satisfy:*

$$\rho_{\bar{x}} = \frac{b(b+1)(a-b^2)}{a^2+b^2(1-a)}\,\frac{1}{N_F+1}, \quad \frac{\rho_{AT}}{\rho_{\bar{x}}} = \frac{a}{a-b^2}, \quad \frac{\rho_{EA}}{\rho_{\bar{x}}} = -\frac{b^2}{a-b^2}, \tag{30}$$

where a and b are as in Proposition 2. Moreover, $\rho_{\bar{x}}$ is strictly positive if and only if slow trading exists, that is, if and only if $N_L > 0$.

One implication of Proposition 4 is that, as long as there is slow trading, the speculator order flow autocorrelation $\rho_{\bar{x}}$ is nonzero. To understand why, note that both the anticipatory trading component and the expectation adjustment component depend on the presence of slow trading. Formally, if there is no slow trading, $\bar{\mu} = 0$ implies that both components of the speculator order flow autocorrelation are zero.

Fig. 3 shows how the speculator order flow autocorrelation ($\rho_{\bar{x}}$) and its anticipatory trading component (ρ_{AT}) depend on the number of FTs (N_F) for four different values of the number of STs ($N_L = 1, 3, 5, 20$). Both $\rho_{\bar{x}}$ and ρ_{AT} are decreasing in N_F. Indeed, when the number of FTs is large, there is only $\frac{1}{N_F+1}$ of the signal left in the next period for the STs. Hence, one should expect the autocorrelation to decrease by the order of $\frac{1}{N_F+1}$, which is indeed the case. For instance, when $N_L = 5$, the speculator order flow autocorrelation is 22.56% when there is one FT, but decreases to 2.84% when there are 20 FTs. My results are consistent with the empirical literature on HFTs. For instance, Brogaard (2011) finds that the autocorrelation of aggregate HFT order flow is small but positive.

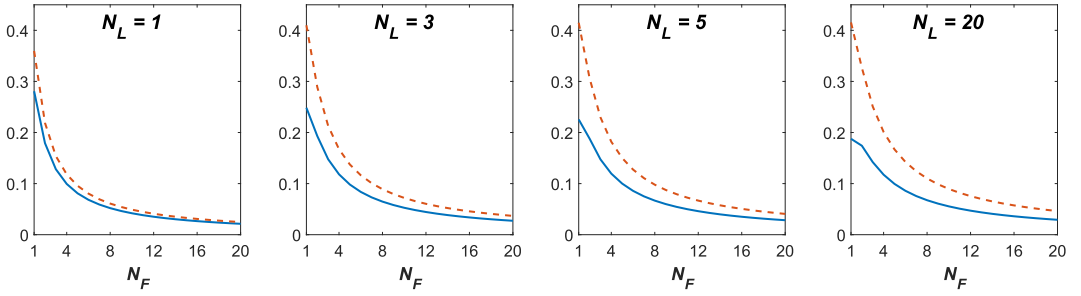

Fig. 3. Speculator order flow autocorrelation. This figure shows the speculator order flow autocorrelation $\rho_{\bar{x}}$ (solid line) and the anticipatory trading component ρ_{AT} (dashed line) as a function of the number of FTs (N_F). The four graphs correspond to four values of the number of speculators using their lagged signal: $N_L = 1, 3, 5, 20$.

The anticipatory trading component ρ_{AT} is increasing in the number N_L of STs (to see this, fix for instance $N_F = 10$ in each of the four graphs in Fig. 3). The intuition is simple: when the number of STs is larger, fast trading in each period can better predict the slow trading the next period, hence the correlation ρ_{AT} is larger. Using NASDAQ data on HFTs, Hirschey (2018) finds that HFT order flow anticipates non-HFT order flow. But the NASDAQ defines HFTs along several criteria including the use of large trading volume and low inventories. In my model, these characteristics may describe the FTs, but not the STs.[28] Thus, if I interpreted FTs in my model as HFTs and STs as non-HFTs, my previous results would imply that HFT order flow anticipates non-HFT order flow.

4. Inventory management

In this section, I analyze the inventory problem of fast traders. In the benchmark model, speculators are risk-neutral and therefore are not concerned about their inventories. I thus modify the model by introducing a type of trader called "Inventory-averse Fast Trader" (IFT). The expected utility of the IFT is defined as in Section 2 (see the discussion before equation (12)), but I introduce a penalty that depends on the IFT's inventory x_t in the risky asset:

$$U = \mathsf{E}\left(\int_0^T (v_T - p_t)\mathrm{d}x_t\right) - C_I \mathsf{E}\left(\int_0^T x_t^2 \mathrm{d}t\right), \tag{31}$$

where $T = 1$, and $C_I > 0$ is a constant called the trader's inventory aversion coefficient. I do not identify the exact source of inventory costs for this type of trader, but the costs can be thought to arise either from capital constraints or from risk aversion.[29]

I call the resulting setup the model with inventory management. To get some intuition for this model, I first solve for the optimal strategy of the IFT in a partial equilibrium framework, taking as fixed the behavior of the other speculators and the dealer. The solution is provided in closed form. I continue with a general equilibrium analysis, and show that the equilibrium remains qualitatively the same. I study the properties of the general equilibrium, as well as the effect of the inventory management on market quality.

4.1. Setup

I consider a model with $N_F + 1$ FTs (who trade only on their current signal) and N_L STs, but I replace one risk-neutral FT with an IFT with utility as in (31).[30] Thus, there are N_F FTs, N_L STs, and one IFT.

To simplify the presentation, I assume directly that the speculators' strategies have constant coefficients, and that the dealer has pricing rules as in the benchmark model. Thus, the FT $i = 1, \ldots, N_F$ has a trading strategy of the form $\mathrm{d}x_{i,t}^F = \gamma_i \mathrm{d}w_t$, while the ST $j = 1, \ldots, N_L$ has a trading strategy of the form $\mathrm{d}x_{j,t}^S = \mu_j \widetilde{\mathrm{d}w}_{t-1}$. The coefficient λ is chosen so that the dealer breaks even, meaning that her expected profit is zero.[31]

Since the IFT has quadratic inventory costs, it is plausible to expect that his optimal trading strategy is linear in the inventory.[32] Therefore, I assume that the IFT's strategy is of the following type:

$$\mathrm{d}x_t = -\Theta x_{t-1} + G\,\mathrm{d}w_t, \tag{32}$$

with constant coefficients $\Theta \in [0, 2)$ and $G \in \mathbb{R}$. Equivalently, the IFT's inventory x_t follows an $AR(1)$ process $x_t = \phi x_{t-1} + G\mathrm{d}w_t$, with an autoregressive coefficient $\phi = 1 - \Theta \in (-1, 1]$.[33]

If $\Theta > 0$, in each trading round the IFT removes a fraction Θ of his current inventory, with the goal of bringing his inventory eventually to zero. One measure of how quickly the inventory mean reverts to zero is the "inventory half-life." This is defined as the average number of periods (of length $\mathrm{d}t$) that the process needs to halve the distance from its mean, i.e.,

$$\text{Inventory Half} - \text{Life} = \frac{\ln(1/2)}{\ln(\phi)}\,\mathrm{d}t = \frac{\ln(1/2)}{\ln(1-\Theta)}\,\mathrm{d}t. \tag{33}$$

Hence, the IFT's inventory half-life is of the order of $\mathrm{d}t$. This in practice can be short (minutes, seconds, milliseconds), which means that when $\Theta > 0$ the IFT does very quick, "real-time" inventory management.

I next discuss the different types of inventory management. In Subsection 4.3, I find that there is a discontinuity between the cases $\Theta = 0$ and $\Theta > 0$. Thus, I introduce a new case in which Θ is infinitesimal and of the form $\Theta = \theta \mathrm{d}t$, with $\theta \in (0, \infty)$.

[28] In Section 4, a FT with sufficiently large inventory costs (called the IFT) has large trading volume and infinitesimal inventory, while STs have a smaller trading volume and relatively large inventories.

[29] Like inventory aversion, risk aversion generates a quadratic penalty on inventory, but it generates other terms as well (e.g., Hendershott and Menkveld, 2014). Therefore, solving the model with risk-averse traders would be considerably more difficult.

[30] In Subsection 5.6 in the Internet Appendix, I introduce more than one inventory-averse trader, which makes the problem more complicated, but does not change qualitatively the main results.

[31] Because of inventory management, the aggregate order flow is no longer unpredictable by the dealer. Nevertheless, the only source of predictability is the IFT's inventory, and as shown later, this inventory in equilibrium is very small due to mean reversion. In Subsection 5.4 in the Internet Appendix, I show that the equilibrium does not change qualitatively if one properly accounts for inventory predictability.

[32] This is standard in the literature, e.g., Ho and Stoll (1981), Madhavan and Smidt (1993), or Hendershott and Menkveld (2014).

[33] A standard result is that the $AR(1)$ process becomes explosive (with infinite mean and variance) if ϕ is outside $[-1, 1]$, or equivalently if Θ is outside $[0, 2]$.

This intermediate inventory management regime continuously connects the other two. Thus, there are three different cases (regimes):

- $\Theta = 0$, the "neutral regime:" the IFT's strategy is of the form $\mathrm{d}x_t = G\mathrm{d}w_t$, similar to the strategy of a (risk-neutral) FT.
- $\Theta > 0$, the "quick regime:" the IFT's strategy is of the form $\mathrm{d}x_t = -\Theta x_{t-1} + G\mathrm{d}w_t$. The inventory half-life is of the order of $\mathrm{d}t$.
- $\Theta = \theta \mathrm{d}t$, the "smooth regime:" the IFT's strategy is of the form $\mathrm{d}x_t = -\theta x_{t-1}\mathrm{d}t + G\mathrm{d}w_t$, with $\theta \in (0, \infty)$.[34] The inventory half-life $\frac{\ln(1/2)}{\ln(1-\theta \mathrm{d}t)}\,\mathrm{d}t = \frac{\ln(2)}{\theta}$, which is much larger than the inventory half-life in the quick regime.

In Subsection 4.3, I show that the smooth regime continuously connects the cases $\Theta = 0$ (neutral regime) with the case $\Theta > 0$ (quick regime). More precisely, $\theta = 0$ in the smooth regime coincides with $\Theta = 0$, while the limit when $\theta \nearrow \infty$ in the smooth regime coincides with the limit when $\Theta \searrow 0$ in the quick regime. In general, I show that the smooth regime is never optimal for the IFT, and therefore I can focus on the comparison between the neutral and the quick regimes.

4.2. Zero inventories

In the quick regime, the IFT's inventory follows an autoregressive process: $x_t = \phi x_{t-1} + G\mathrm{d}w_t$ with coefficient $\phi \in (-1, 1)$. Thus, the variance of the IFT's inventory is $\mathsf{Var}(x_t) = \mathsf{Var}(\mathrm{d}w_t)/(1 - \phi^2)$; however, the variance of the increment $\mathrm{d}w_t$ is equal to $\sigma_w^2 \mathrm{d}t$, hence it is infinitesimal, and therefore so is the inventory x_t.[35] Thus, in the continuous-time limit, the inventory is essentially zero at all times. This fact can also be seen from the formula (33), which shows that the inventory half-life in the quick regime is a multiple of the infinitesimal time increment $\mathrm{d}t$.

In general, the expected profit of any speculator satisfies:

$$\pi = \mathsf{E}\int_0^T (v_T - p_t)\mathrm{d}x_t = \underbrace{\mathsf{E}\left(v_T(x_T - x_0)\right)}_{\text{Inventory Component}} + \underbrace{\mathsf{E}\int_0^T (-p_t)\,\mathrm{d}x_t}_{\text{Cash Component}}. \tag{34}$$

The inventory component is the expected profit due to the accumulation of inventory in the risky asset. This does not translate into cash profits until the liquidation date T. The cash component is the expected profit that comes from changes in the cash account due to trading.

Proposition 5 provides a useful formula in the case of a speculator who has zero inventories, and who therefore gets all his profits from the cash component.

Proposition 5. *Consider a speculator with trading strategy $\mathrm{d}x_t$ for $t \in (0, T]$, such that the initial and final inventories are zero, i.e., $x_0 = 0$ and $x_T = 0$ almost surely. Then the speculator's expected profit is:*

$$\pi_c = \mathsf{E}\int_0^T x_{t-1}\,\mathrm{d}p_t. \tag{35}$$

Thus, whenever inventory management results in zero inventories for the speculator, the trading strategy is only profitable when the inventory level (x_{t-1}) forecasts the subsequent change in price ($\mathrm{d}p_t$). In linear equilibria, the price change must be proportional to the part of the aggregate order flow unanticipated by the dealer. Therefore, according to Proposition 5, the speculator must be able to forecast the unanticipated aggregate order flow. This can occur only if the subsequent order flow contains a component that is correlated with the speculator's past trading.

I now define "slower trading" as the part of the aggregate order flow that is positively correlated with the speculator's past inventory. Proposition 5 then shows that the speculator makes positive profits while keeping zero inventory only if there is slower trading.

In the case of the IFT, his inventory is zero at all times, so Proposition 5 can be applied. Note that there is indeed slower trading coming from the STs (as long as $N_L > 0$): the inventory of the IFT at $t - 1$ contains $G\mathrm{d}w_{t-1}$, which is positively correlated with the aggregate order flow at t via the orders of the ST, $\mu_j \widetilde{\mathrm{d}w}_{t-1}$. Hence, it is possible for the IFT to make positive profits while keeping zero inventory.

4.3. IFT and inventory management

Consider the inventory management model with N_F FTs, N_L STs, and one IFT. In this subsection, I solve for the optimal strategy of the IFT in a partial equilibrium analysis, keeping the behavior of the other players fixed. The behavior of the other players is analyzed in Subsection 4.4.

[34] This is called an Ornstein-Uhlenbeck process.

[35] See also equation (69) in the Appendix, where I show that $\mathsf{E}(x_t^2)$ is of the order of dt.

I thus fix the coefficients γ and μ that describe the strategies of the FTs and STs, and the coefficients λ and ρ that describe the dealer's pricing rules. Suppose the IFT has a trading strategy as in (32): $\mathrm{d}x_t = -\Theta x_{t-1} + G\mathrm{d}w_t$, which is not necessarily optimal. The expected profit of the IFT can then be written as $\pi = \mathsf{E}\int_0^T (w_t - p_t)\mathrm{d}x_t$. Since $w_t = w_{t-1} + \mathrm{d}w_t$ and $p_t = p_{t-1} + \lambda \mathrm{d}y_t$, one has the following decomposition:

$$\pi = \underbrace{G\mathsf{E}\int_0^T (\mathrm{d}w_t - \lambda \mathrm{d}y_t)\mathrm{d}w_t}_{\pi_0} - \underbrace{\Theta\mathsf{E}\int_0^T (w_{t-1} - p_{t-1})x_{t-1}}_{\ell_r} + \underbrace{\Theta\mathsf{E}\int_0^T x_{t-1}\mathrm{d}p_t}_{\pi_a}. \tag{36}$$

The first term, denoted by π_0, is the IFT's expected profit when $\Theta = 0$, which reflects the profits that result from exploiting his signals ($\mathrm{d}w_t$). The second term, denoted by ℓ_r, is the informational loss that comes from inventory mean reversion: indeed, by reducing inventory by Θx_{t-1} each period, there is an expected loss coming from the correlation of x_{t-1} with the remaining informational advantage $w_{t-1} - p_{t-1}$. Put differently, by managing inventory the IFT trades against his previous signals. The third term, denoted by π_a, is the profit that comes from anticipation of slow trading: at time t, the IFT reduces his inventory by Θx_{t-1}, exactly when the STs submit a market order in the opposition direction (which is part of the current aggregate order flow $\mathrm{d}y_t$). Note that the third term is equal to $\Theta\pi_c$, where π_c is the expected profit of a speculator who keeps all his profits in cash: see equation (35).

When the IFT mean reverts his inventory ($\Theta > 0$), his inventory is zero, and his profit is $\pi = \pi_c$. Equation (36) then implies that $\ell_r = \pi_0 - (1 - \Theta)\pi_c$. This implies that mean reversion fully erases all the profits obtained from the IFT's trading on his signals. To understand the intuition for this result, suppose the IFT observes a new signal $\mathrm{d}w_t$. Initially, the IFT trades on his signal ($G\mathrm{d}w_t$), but subsequently he fully reverses his trade by unloading a positive fraction of his inventory each period. Therefore, the only way for the IFT to make money is to ensure that the inventory reversal is done at a profit. This can occur for instance if the IFT expects that when he sells (Θx_{t-1}), other traders buy even more, and as a result his overall price impact is negative. (The profit from this activity is exactly the anticipation profit π_a.) But this is only possible if there are STs, as their lagged signals can be predicted by the IFT.

In order to formalize this last result, I define additional coefficients:

$$\gamma^- = N_F\gamma, \quad \overline{\mu} = N_L\mu, \quad a^- = \rho\gamma^-, \quad b = \rho\overline{\mu}, \quad R = \frac{\lambda}{\rho}. \tag{37}$$

The next result provides an explicit formula for the IFT's expected profit.

Proposition 6. *Let $\mathrm{d}x_t = -\Theta x_{t-1} + G\mathrm{d}w_t$ be the IFT's strategy (not necessarily optimal), with $\Theta > 0$, and hence $\phi = 1 - \Theta \in (-1, 1)$. Suppose $b \in (-1, 1)$. Then, the IFT has all his profits in cash. His expected profit π satisfies:*

$$\pi = \lambda\left(\overline{\mu}G\frac{1 - a^-}{1 + \phi b} - G^2\frac{b + \frac{1}{1+\phi}}{1 + \phi b}\right)\sigma_w^2. \tag{38}$$

Proposition 6 shows that, as a result of keeping all his profits in cash, the IFT behaves very differently compared to risk-neutral speculators such as the FTs: while the risk-neutral speculator trades directly on his private information, the IFT benefits only indirectly, from timing his trades and unloading his inventory to slower traders. Indeed, equation (38) shows that in the absence of slow trading ($\overline{\mu} = 0$), the IFT makes negative expected profits.

Equation (38) also explains how the IFT's profit depends on the coefficients $a^- = N_F\rho\gamma$ and $b = N_L\rho\mu$, which respectively measure the amount of fast trading and slow trading. When there is more fast trading (a^- is higher), the IFT's profit is smaller because of increased competition from FTs. When there is more slow trading (b is higher), there is a larger benefit ($\sigma_w^2 RG(1 - a^-)\frac{b}{1+\phi b}$) that comes from providing liquidity to STs, but also a larger cost ($\sigma_w^2\lambda G^2\frac{b+1/(1+\phi)}{1+\phi b}$). This cost arises from the fact that when the IFT at t provides liquidity to the STs, these do not trade on the lagged signal ($\mathrm{d}w_{t-1}$) but rather on its unanticipated part ($\widetilde{\mathrm{d}w}_{t-1} = \mathrm{d}w_{t-1} - \rho\mathrm{d}y_{t-1}$), which reduces the IFT's profits.[36]

I now describe the optimal strategy of the IFT. Recall that beside the expected profit, the IFT's utility also includes a penalty cost that is quadratic in the inventory: $C_I\mathsf{E}\left(\int_0^T x_t^2\mathrm{d}t\right)$. This penalty is not relevant when $\Theta > 0$, because in that case the IFT has zero inventory. When $\Theta = 0$, however, the penalty can be considerable, depending on the inventory aversion coefficient C_I. Theorem 2 describes the IFT's optimal strategy when the slow trading coefficient b is above a threshold: $b > \frac{\sqrt{17}-1}{8} \approx 0.3904$.

[36] It turns out that, compared to the benefit, the cost is more strongly increasing in b, hence the optimal G is actually decreasing in b (see equation (40)).

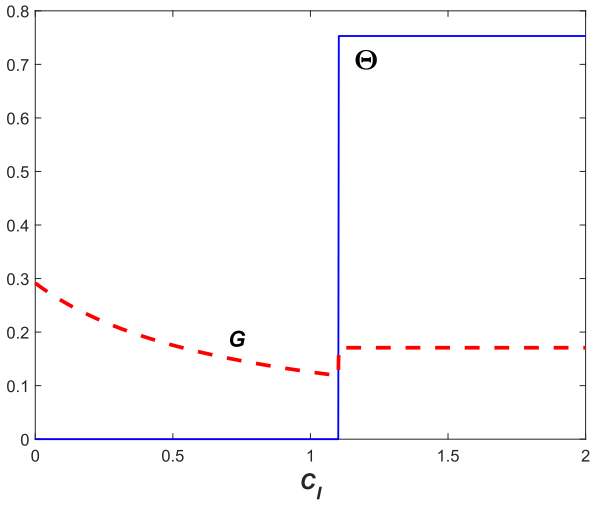

Fig. 4. Optimal IFT inventory management. This figure shows the coefficients of the IFT's optimal trading strategy ($dx_t = -\Theta x_{t-1} + G dw_t$) in the inventory management model with $N_F = 5$ FTs and $N_L = 5$ STs. On the horizontal axis is the IFT's inventory aversion, C_I. The parameter values are $\sigma_w = 1$ and $\sigma_u = 1$. For the model coefficients, I use the equilibrium values from Subsection 4.4: $a^- = 0.7088$, $b = 0.5424$, $\lambda = 0.3782$, and $\rho = 0.3439$. The formulas for G, Θ, and $\overline{C}_I$ are computed using Theorem 2.

This condition is true if for instance there are $N_F \geq 1$ FTs and $N_L \geq 2$ STs.[37]

Theorem 2. *In the inventory management model, suppose the coefficients satisfy the following inequalities:* $0 \leq a^-, b < 1$ *and* $\lambda, \rho > 0$. *In addition, suppose* $b > \frac{\sqrt{17}-1}{8} \approx 0.3904$.[38] *Let* $\overline{C}_I = 2\lambda\left(\frac{(1-Ra^-)^2(1+\sqrt{1-b})^2}{R^2 b(1-a^-)^2} - 1\right)$. *Then, if* $C_I < \overline{C}_I$, *the optimal strategy of the IFT is to set:*

$$\Theta = 0, \quad G = \frac{1 - Ra^-}{2\lambda + C_I}. \tag{39}$$

If $C_I > \overline{C}_I$, *the optimal strategy of the IFT is to set:*

$$\Theta = 2 - \frac{\sqrt{1-b}}{b} \in (0, 2), \quad G = \frac{1 - a^-}{2\rho\left(1 + \frac{1}{\sqrt{1-b}}\right)}. \tag{40}$$

Theorem 2 implies that there are two different types of optimal behavior for the IFT, depending on how his inventory aversion compares to a threshold value ($\overline{C}_I$).

1. (Neutral regime) If the inventory aversion coefficient is small (below $\overline{C}_I$), the IFT sets $\Theta = 0$ and controls his inventory by choosing his weight G. As his inventory aversion gets larger, the IFT reduces his inventory costs by decreasing G. The tradeoff is that a smaller G also reduces expected profits. The behavior of the IFT when $\Theta = 0$ is essentially the same as the behavior of a FT.
2. (Quick regime) If the inventory aversion is large (above $\overline{C}_I$), the IFT manages his inventory by choosing a positive mean reversion coefficient ($\Theta > 0$). There is no longer a tradeoff between expected profit and inventory costs, as the IFT has zero inventory costs. Hence, the IFT chooses the weight G and the mean reversion Θ to maximize expected profit (more details below).

Thus, a small change in the IFT's inventory aversion can have a large effect on the IFT's behavior. Fig. 4 shows the coefficients of the optimal strategy when there are $N_F = 5$ FTs and $N_S = 5$ STs. When the IFT's inventory aversion rises above the threshold $\overline{C}_I = 0.1021$, his optimal mean reversion coefficient jumps from $\Theta = 0$ to $\Theta = 0.7530$. Moreover, his optimal weight jumps from $G = 0.1186$ (the left limit of G at the threshold) to $G = 0.1708$ (the constant value of G above the threshold).

To get more intuition for the discontinuity in the IFT's optimal trading strategy, I examine the smooth regime in connection with the neutral and the quick regimes. Recall that the IFT's trading strategy is of the form $dx_t = -\Theta x_{t-1} + G dw_t$, where either (i) $\Theta = 0$ (neutral regime), (ii) $\Theta = \theta dt$ (smooth regime), or (iii) $\Theta > 0$ (quick regime). I then verify that the IFT's expected

[37] If instead $b < 0.3904$, one can show that a similar analysis holds (see Subsection 5.1 in the Internet Appendix). The IFT still manages inventory but the optimal Θ is at its lowest possible value, denoted by 0_+. This value is the same as $\theta = \infty$ in the smooth regime.

[38] In equilibrium (Subsection 4.4), I obtain the following numerical results: the condition $b < 1$ is always satisfied, and the condition $b > \frac{\sqrt{17}-1}{8}$ is equivalent to having (i) $N_L \geq 2$ and (ii) $N_L \geq 6$ if $N_F = 0$.

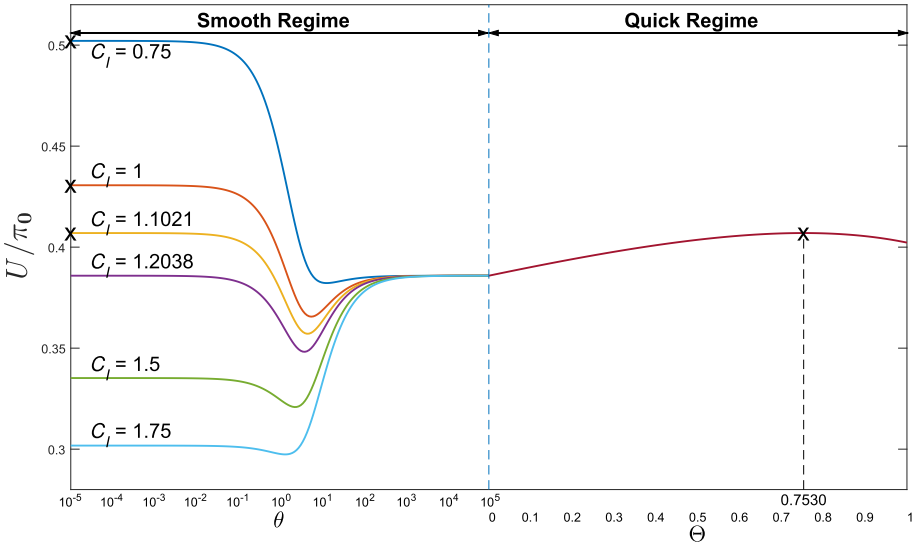

Fig. 5. IFT inventory management and utility. This figure shows the maximum normalized expected utility of the IFT for a fixed mean reversion rate, in the inventory management model with $N_F = 5$ FTs and $N_L = 5$ STs. On the horizontal axis is the IFT's mean reversion rate given by (i) θ from the IFT's trading strategy, $dx_t = -\theta x_{t-1}\,dt + G dw_t$ (the smooth regime), or (ii) Θ from the IFT's trading strategy, $dx_t = -\Theta x_{t-1} + G dw_t$ (the quick regime). On the vertical axis is the IFT's maximum expected utility U when G varies and Θ (or θ) is fixed, normalized by the maximum expected profit π_0 when G varies and $\Theta = 0$ (the neutral regime). The other parameter values are $\sigma_w = 1$ and $\sigma_u = 1$. For the model coefficients, I use the equilibrium values $a^- = 0.7088$, $b = 0.5424$, $\lambda = 0.3782$, $\rho = 0.3439$.

utility, along with its components described above, varies continuously across the three regimes. More formally, if I denote by $U(\Theta)$ the IFT's expected utility in either of the three regimes, $\lim_{\theta\to 0} U(\theta dt) = U(0)$ and $\lim_{\theta\to\infty} U(\theta dt) = \lim_{\Theta\to 0} U(\Theta)$ (see Subsection 6.1 in the Internet Appendix). Thus, the smooth regime indeed continuously connects the neutral regime with the quick regime.

Fig. 5 shows the expected utility U as a function of Θ across the smooth and quick regimes. To simplify the presentation, instead of considering $U = U(\Theta, G)$ as a function of both Θ and G, I only consider the value of G that maximizes U given Θ.[39] Formally, if $U(\Theta, G)$ indicates the dependence of U on both Θ and G, in Fig. 5 I plot U/π_0, where $U = \max_G U(\Theta, G)$ and $\pi_0 = \max_G \pi_0(G)$. Fig. 5 shows that the IFT's utility indeed changes continuously from the smooth regime to the quick regime. Moreover, when the inventory aversion coefficient C_I varies, there are two cases:

- If $C_I < 1.1021$, the maximum U is attained at $\Theta = 0$.
- If $C_I > 1.1021$, the maximum U is attained at $\Theta = 2 - \frac{\sqrt{1-b}}{b} \approx 0.7530$.

I thus confirm a result proved in Theorem 2: when the inventory aversion C_I crosses the threshold $\overline{C}_I = 1.1021$, the optimal Θ jumps discontinuously from 0 to 0.7530. As observed in Fig. 5, the reason for this discontinuity is that in the smooth regime the optimum θ is either zero or infinity, but never in between.

To understand the intuition behind this last fact, I describe in more detail how the IFT's utility changes with Θ. By definition, this utility is equal to the expected profit minus the quadratic penalty on inventory. From (36),

$$U = \pi_0 - \underbrace{\mathrm{E}\int_0^T (w_{t-1} - p_{t-1}) x_{t-1}\Theta}_{\ell_r} + \underbrace{\Theta \mathrm{E}\int_0^T x_{t-1}\mathrm{d}p_t}_{\pi_a} - \underbrace{C_I \mathrm{E}\int_0^T x_{t-1}^2 \mathrm{d}t}_{\ell_i}. \tag{41}$$

The term π_0 does not depend on Θ, and is the same for the smooth and quick regimes. The loss ℓ_r that comes from inventory mean reversion is positive in both regimes: indeed, in both cases the IFT trades against his own past signal (more precisely, he trades against the part of the signal that was not yet incorporated into the price: $w_t - p_t$). The loss ℓ_r is increasing in Θ (or θ): the more the IFT mean reverts his inventory, the larger the corresponding informational loss. The third term, π_a, is zero in the smooth regime, while it is positive only in the quick regime: this is because the IFT, who can anticipate slow trading, can benefit from providing liquidity to STs only when he reverts a large enough part of his inventory, that is, when Θ is not infinitesimal (the quick regime). The fourth term, the inventory penalty ℓ_i, is positive in the smooth regime, but starts decreasing fast in θ

[39] In all regimes, the expected utility is quadratic and concave in G. In the quick regime, U is given by equation (38). In the smooth regime, equation (IA.538) in Section 6 in the Internet Appendix implies that $\frac{U}{\sigma_w^2} = G\left((1 - Ra^-) - F_\theta\left(1 - R\frac{a^- + b}{1+b}\right)\right) - \frac{G^2}{2}\left(2\lambda\left(1 - \frac{F_\theta}{1+b}\right) + F_{2\theta}\left(\frac{\lambda}{1+b} + \frac{C_I}{\theta}\right)\right)$, where $F_\theta = 1 - \frac{1-e^{-\theta}}{\theta}$.

when this coefficient is sufficiently large, and it approaches zero in the limit. Thus, in the quick regime, ℓ_i is zero, as the IFT's inventory is zero at all times.

I next explain why the IFT's maximum utility U in the smooth regime only occurs either at $\theta = 0$ or at $\theta = \infty$ (see Fig. 5). Initially, when the mean reversion coefficient θ is small, an increase in θ raises the informational loss ℓ_r from trading against his own signals, while the associated reduction in inventory does not significantly diminish the penalty ℓ_i (which is quadratic in inventory). However, when θ is large, the inventory penalty is reduced more dramatically and contributes to a rise in utility as θ approaches infinity. Because of the drop in utility in the middle range of θ, the IFT's maximum expected utility in the smooth regime can only occur at either of the endpoints (0 or ∞).[40]

I also explain why the IFT's maximum utility U in the quick regime is realized at an interior Θ (equal to 0.7530 in Fig. 5). First, when Θ is in the quick regime, the inventory penalty ℓ_i is zero, hence there are only two nonzero terms that depend on Θ: the informational loss ℓ_r that comes from mean reversion, and the gain π_a that comes from anticipating STs. As Θ increases, the mean reversion loss ℓ_r increases (it was already positive in the smooth regime), but the anticipatory gain π_a increases as well (it was zero in the smooth regime). When Θ is small, the term ℓ_r dominates and U is increasing in Θ. When Θ is large, the term π_a dominates and U is decreasing in Θ. As a result, U has an interior optimum in the quick regime.

Thus, depending on their inventory aversion, the fast speculators fall into two sharply different categories. In both categories, speculators generate relatively large trading volume. But in one category (when the speculators' inventory aversion is low) the speculators make fundamental bets and accumulate inventories, while in the other category speculators mean revert their inventories very quickly, and keep their profits in cash. My results appear consistent with the "opportunistic traders" and the "high-frequency traders" described in Kirilenko et al. (2017). Both opportunistic traders and HFTs have large volume and appear to be fast. But while opportunistic traders have relatively large inventories, the HFTs in their sample (during several days around the Flash Crash of May 6, 2010) liquidate 0.5% of their aggregate inventories on average each second. This implies that HFT inventories have an $AR(1)$ half-life of a little over 2 min.

I next examine how the IFT's optimal strategy is correlated with slow trading. Proposition 6 shows that if there is no slow trading, the IFT cannot make positive profits. Theorem 2 shows that with enough slow trading, the IFT can manage inventory and make positive profits (see equation (83) in the Appendix). In the previous discussion, I argue that this is possible only if the IFT trades in the opposite direction to the slow trading. Corollary 3 shows that this is indeed the case.

Corollary 3. *Suppose the IFT is sufficiently inventory-averse ($C_I > \overline{C}_I$). Denote by $\mathrm{d}\overline{x}_t^S = \overline{\mu}\widetilde{\mathrm{d}w}_{t-1}$ the slow trading component of the speculator order flow. Then, the IFT's optimal strategy is negatively correlated with slow trading:*

$$\mathrm{Cov}\left(\mathrm{d}x_t, \mathrm{d}\overline{x}_t^S\right) = -\Theta\mathrm{Cov}\left(x_{t-1}, \mathrm{d}\overline{x}_t^S\right) < 0. \tag{42}$$

I call this phenomenon the "hot potato" effect, or the "intermediation chain" effect. The intuition is that the IFT's current signal generates undesirable inventory and must be passed on to slower traders in order to produce a profit. The passing of inventory can be thought as the beginning of an intermediation chain. Weller (2012) and Kirilenko et al. (2017) document such hot potato effects among high-frequency traders.

4.4. Equilibrium results

In this subsection, I solve for the full equilibrium of the inventory management model. For simplicity, I assume that the IFT is sufficiently inventory-averse, meaning that his inventory aversion is above a certain threshold (formally, above the threshold value $\overline{C}_I$ from Theorem 2). Then, Theorem 3 shows that the solution can be expressed almost in closed form, except for the slow trading coefficient b, which satisfies a non-linear equation in one variable.

Theorem 3. *Consider the inventory management model with one sufficiently averse IFT, N_F FTs, and N_L STs. Suppose there is an equilibrium in which the speculators' strategies are: $\mathrm{d}x_t = -\Theta x_{t-1} + G\mathrm{d}w_t$ (the IFT), $\mathrm{d}x_t^F = \gamma\mathrm{d}w_t$ (the FTs), $\mathrm{d}x_t^S = \mu\widetilde{\mathrm{d}w}_{t-1}$ (the STs); and the dealer's pricing rules are: $\mathrm{d}p_t = \lambda\mathrm{d}y_t$, $\widetilde{\mathrm{d}w}_t = \mathrm{d}w_t - \rho\mathrm{d}y_t$. Denote the coefficients R, a^-, and b as in (37). Suppose $\frac{\sqrt{17}-1}{8} < b < 1$. Then, the equilibrium coefficients satisfy* equations (84)–(86) *in the Appendix.*

Conversely, suppose that equations (84)–(86) *have a real solution such that $\frac{\sqrt{17}-1}{8} < b < 1$, $a < 1$, and $\lambda > 0$. Then, the speculators' strategies and the dealer's pricing rules with these coefficients provide an equilibrium of the model.*

Rather than relying on numerical results to study the equilibrium, I start by providing asymptotical results when the number of FTs and STs is large. The advantage to this approach is that the asymptotic results can be expressed in closed form, and thus help provide a clearer intuition for the equilibrium. Let $\overline{C}_I$ be the threshold aversion from Theorem 2. Let π be the expected profit of a sufficiently averse IFT ($C_I \geq \overline{C}_I$), and $\pi^{C_I=0}$ be the maximum expected profit of a risk-neutral IFT ($C_I = 0$), where the behavior of the other speculators and the dealer is taken to be the same. Let γ_0 be the benchmark FT weight, and $\pi_0^F = \frac{\gamma_0}{N_F+2}\sigma_w^2$ the benchmark profit of a FT, as in Proposition 1. I use the asymptotic notation: $X \approx X_\infty$ stands for $\lim_{N_F,N_L\to\infty} \frac{X}{X_\infty} = 1$. Then,

[40] This result is true in general: see the numerical Result IA.1 in Section 6 in the Internet Appendix.

Proposition 7 provides an asymptotic description of the equilibrium.

Proposition 7. *Consider (i) the inventory management model with one sufficiently averse IFT, N_F FTs, and N_L STs, and (ii) the benchmark model with $N_F + 1$ FTs and N_L STs. Then, the equilibrium coefficients γ, μ, λ, ρ are asymptotically equal across the two models when N_F and N_L are large. Also, $a \approx 1, b \approx b_\infty = 0.6180$, and the following asymptotic formulas hold:*

$$\Theta \approx 1, \quad \frac{G}{\gamma_0} \approx 1 - b_\infty = 0.3820, \quad \frac{\pi}{\pi_0^F} \approx 2b_\infty - 1 = 0.2361,$$
$$\frac{\pi}{\pi^{C_I=0}} \approx \frac{4}{5} b_\infty = 49.44\%, \quad \overline{C}_I \approx \frac{1+5b_\infty}{2} \lambda_\infty \approx 2.0451 \frac{\sigma_w}{\sigma_u} \frac{1}{\sqrt{N_F+1}}.$$

The first implication of Proposition 7 is that model with inventory management is asymptotically the same as the benchmark model when both N_F and N_L are large. This is not surprising, since when there are many other speculators, the IFT has a relatively smaller and smaller role in the limit.

The behavior of the IFT is more surprising. First, when there are many other speculators, the IFT's inventory mean reversion becomes extreme (Θ approaches 1). This means that the IFT's inventory half-life becomes essentially zero, as the IFT removes most of his inventory each period. This extreme mean reversion is possible because the existence of a sufficient amount of slow trading allows the hot potato effect to generate positive profits for the IFT. Furthermore, the equation $\pi \approx 49.44\% \times \pi^{C_I=0}$ implies that even under extreme inventory mean reversion ($\Theta \approx 1$), the IFT can trade so that he only loses on average about 50% of his maximum expected profits corresponding to an inventory aversion of zero (i.e., gets about half of the maximum profit of a FT).[41]

The equation $\overline{C}_I \approx 2.0451 \frac{\sigma_w}{\sigma_u} \frac{1}{\sqrt{N_F+1}}$ implies that the threshold inventory aversion above which the IFT chooses to mean revert his inventory becomes very small when the number of competing FTs is large. This is perhaps counterintuitive, since one may think that the IFT chooses fast inventory mean reversion because he has very high inventory aversion. This is not the case, however. Indeed, even when the IFT has small inventory aversion, a sufficient amount of slow trading is enough to convince the IFT to engage in very fast inventory mean reversion. This is because inventory management is a zero/one proposition. Once the IFT engages in inventory management ($\Theta > 0$), any profits from fundamental bets become zero, and the hot potato effect is the sole source of profits.

I now compare the IFT with the other speculators. For the IFT, I consider the following variables: $TV_x = \mathrm{Var}(\mathrm{d}x_t)/\mathrm{d}t$, the IFT's trading volume, measured by his order flow variance (as in Subsection 3.2); $\rho_x = \mathrm{Corr}(\mathrm{d}x_t, \mathrm{d}x_{t+1})$, the IFT's order flow autocorrelation; and $\beta_{x,\overline{x}^S} = \mathrm{Cov}(\mathrm{d}x_t, \mathrm{d}\overline{x}_t^S)/\mathrm{Var}(\mathrm{d}\overline{x}_t^S)$, the regression coefficient of the IFT's strategy ($\mathrm{d}x_t$) on the slow trading component ($\mathrm{d}\overline{x}_t^S$). I also consider: TV_{x^F}, the individual FT volume; $TV_{\overline{x}^F}$, the aggregate FT volume; $TV_{\overline{x}^S}$, the aggregate ST volume; $\rho_{\overline{x}^F}$, the aggregate FT order flow autocorrelation; and $\rho_{\overline{x}^S}$, the aggregate ST order flow autocorrelation.

Proposition 8 provides formulas for all these quantities, as well as asymptotic limits when both N_F and N_L are large. Note that some of these results provide new testable implications, regarding the relation between trading volume, order flow covariance, and inventory.

Proposition 8. *For a sufficiently averse IFT, the variables defined above satisfy the following formulas:*

$$\begin{aligned}
&\frac{TV_x}{TV_{x^F}} = \frac{2G^2}{(1+\phi)\gamma^2} \approx 4 - 6b_\infty = 0.2918, \quad \frac{TV_{\overline{x}^S}}{TV_{\overline{x}^F}} = \frac{b^2(1-a)}{(a^-)^2} \approx \frac{b_\infty}{N_F+1}, \\
&\rho_x = -\frac{\Theta}{2} \approx -\frac{1}{2}, \quad \rho_{\overline{x}^F} = 0, \quad \rho_{\overline{x}^S} \approx -b_\infty = -0.6180, \\
&\beta_{x,\overline{x}^S} = -\frac{\Theta(1-a^-)}{2b(1+2\sqrt{1-b})} \approx -\frac{3+b_\infty}{5(N_F+1)} = -\frac{0.7236}{N_F+1}.
\end{aligned} \tag{44}$$

The last formula illustrates the hot potato effect. The IFT's order flow has a negative beta on the STs' aggregate order flow, which means that the IFT and the STs trade in opposite directions. As the number of FTs becomes larger, there is more information released to the public by the trades of the FTs, hence there is less room for slow trading. As a result, the hot potato effect is less intense when there is a large number of FTs.

Proposition 8 implies that in the limit when N_F and N_L are large, the IFT's trading volume is about 30% of the individual FT trading volume. This implies that the IFT's trading volume is comparable to that of a regular FT. By contrast, just as in the benchmark model, the volume coming from STs is much smaller than the volume coming from FTs. This confirms the intuition that in an empirical analysis that selects traders based on volume, the IFT and the FTs are in the category with large trading volume, while the STs are in the category with small trading volume.

If one compares order flow autocorrelations, one sees that the IFT is similar to the STs, but not to the FTs. Indeed, the IFT and the STs have negative and large order flow autocorrelation. By contrast, the FTs have zero order flow autocorrelation.[42] Finally, if one compares inventories, the IFT has infinitesimal inventory, while the variance of the other speculators' inventory increases

[41] This recalls the saying attributed to Joseph Kennedy (the founder of the Kennedy dynasty) that "I would gladly give up half my fortune if I could be sure the other half would be safe."

[42] Even if the FTs were allowed to trade on lagged signals, the autocorrelation of their order flow would still be very small (of the order of $\frac{1}{N_F+1}$).

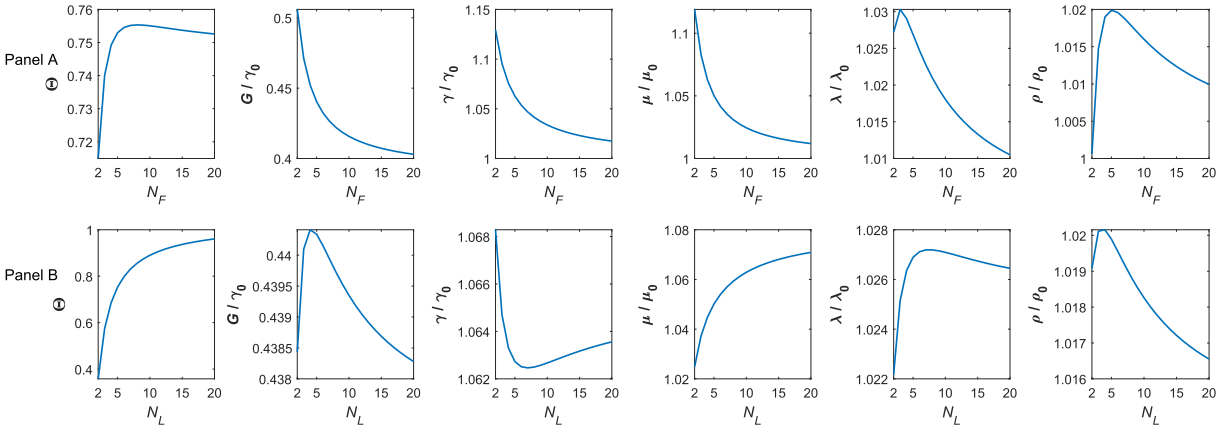

Fig. 6. Equilibrium coefficients with inventory management. This figure shows several equilibrium coefficients that arise in the inventory management model. If X is a variable in the inventory management model, I denote by X_0 the corresponding variable in the benchmark model. On the vertical axis I consider the following (normalized) coefficients: Θ, $\frac{G}{\gamma_0}$, $\frac{\gamma}{\gamma_0}$, $\frac{\mu}{\mu_0}$, $\frac{\lambda}{\lambda_0}$, and $\frac{\rho}{\rho_0}$. In Panel A, I plot each coefficient against the number of FTs (N_F), while fixing $N_L = 5$. In Panel B, I plot the six coefficients against the number of STs (N_L), while fixing $N_F = 5$. The other parameters are $\sigma_w = 1$ and $\sigma_u = 1$.

over time.[43] Nevertheless, the STs' inventories are smaller relative to FTs' inventories, since the STs have smaller volume.

I next present some numerical results for the equilibrium coefficients. Fig. 6 shows the equilibrium coefficients Θ, G, γ, μ, λ, and ρ, when N_F and N_L vary.[44] Some of these coefficients are normalized by the corresponding coefficient in the benchmark model.

As expected, the mean reversion coefficient Θ is increasing in the number of STs (N_L). This is because the IFT needs slow traders in order to make profits. The IFT's weight G is less than half the benchmark weight γ_0, indicating that the IFT shifts towards inventory management in order to make profits. This leaves more room for fundamental profits, which explains why both the FTs and the STs are better off with inventory management than in the benchmark model (γ/γ_0 and μ/μ_0 are both above one), despite the price impact λ being larger than in the benchmark ($\lambda/\lambda_0 > 1$). The reason why the market is more illiquid in the inventory management model is that the IFT trades much less intensely on his signal (G is less than half of γ_0), and therefore the informational efficiency is lower. To see directly that the market is less informationally efficient in the inventory management model, I use the fact that in my model price volatility is a proxy for informational efficiency (see the discussion in Subsection 3.2). Then, I verify numerically that indeed $\sigma_p/\sigma_{p,0} < 1$, which implies that with inventory management the market is less informationally efficient.

5. Robustness and extensions

In this section, I discuss several model assumptions, and verify that the results remain qualitatively the same when these assumptions are relaxed.

An important assumption of the benchmark model is that speculators use strategies of the type (11) that are linear in the unanticipated part of signals up to a certain lag. In Section 2, I justify this by an information processing cost per signal. But by choosing strategies as in (11), I also implicitly assume that speculators (i) must process each signals individually, and (ii) cannot use their signals to learn about the other speculators' forecasts.

Assumption (i) can be relaxed by allowing other combinations of past signals, such as the price to be part of the trading strategy.[45] Thus, one can add a Kyle term of the form $\beta_t(w_t - p_t)dt$ to the strategy in (11). In that case, the speculator might be adversely selected if the price p_t is also affected by speculators who learn about other components of the fundamental value. To formalize this intuition, in Section 4 in the Internet Appendix, I introduce an orthogonal dimension of the fundamental value, and show that trading strategies with a Kyle component add relatively little to a speculator's profit, and, depending on the parameter values, often lead to a loss.

Assumption (ii) can be relaxed by allowing slower speculators to learn about faster speculators' forecasts. This is not an issue in the benchmark model $\mathcal{M}_1$: indeed, STs do not need to learn at t about the FTs' previous signal dw_{t-1}, because at t they already learn dw_{t-1} perfectly. Nevertheless, in the model $\mathcal{M}_2$ (in which speculators use signals with a maximum lag of two), the slowest traders at t observe the double-lagged signal dw_{t-2} but could also get a noisy signal about dw_{t-1} by observing the aggregate order flow at $t-1$. Thus, the slowest traders at t observe a signal is of the form $\overline{\gamma}dw_{t-1} + du_t$, which is the aggregate

[43] For the IFT, $\text{Var}(x_t) = \frac{G^2}{1-\phi^2}\sigma_w^2\,dt$ (see equation (68) in the Appendix), while for the FT, $\text{Var}(x_t^F) = t\sigma_w^2$, as the FT's inventory follows a random walk.

[44] I consider $N_F, N_L \geq 2$. The reason is that in order to apply Theorem 2, one needs to have $b > \frac{\sqrt{17}-1}{8}$. This is true in equilibrium if $N_F, N_L \geq 2$.

[45] One need not worry about finite combinations of signals, since it is plausible in that case that one needs to pay attention to the individual signals that are part of the combination. But, as the model is set in continuous time, the price is an infinite combination of past signals.

order flow dy_{t-1} minus the part $\overline{\mu}\widetilde{dw}_{t-2} + \overline{\nu}\widetilde{dw}_{t-3}$ already observed by the slowest traders at t.[46] I argue that such learning from the order flow is difficult and risky, because then the slowest traders must know (and be confident about) the aggregate trading coefficients of everyone else. Nevertheless, in Subsection 2.2 in the Internet Appendix, I consider an extension of $\mathcal{M}_2$ in which the slowest traders can learn from the order flow at no cost. In this extension, I verify that the main results of the benchmark model are robust. In particular, even after learning from the order flow, the slowest traders' profits are an order of magnitude smaller than those of the other traders.

One related extension is to allow signals that are not perfectly correlated. In that case, one would encounter the phenomenon of Foster and Viswanathan (1996), that speculators need to forecast the forecasts of others. Even though I have not been able to solve such an extension, the intuition would likely remain very similar. Indeed, as observed by Back et al. (2000), when signals are not perfectly correlated, initially speculators trade very aggressively on the common part of their signals (the "rat race"). Because in my model speculators drop their signals after a few lags, it is plausible that the speculators would trade as if they had nearly identical signals.

Another way of justifying the trading strategy in (11) is to add (lagged) public news to the benchmark model. Suppose at every t a public signal, called "news", is revealed about the k-lagged signal dw_{t-k}. In that case, Foucault et al. (2016) show that the optimal strategy of a (fast) speculator must be of the form $dx_t = \beta_t(w_t - p_t)dt + \gamma_t^0 dw_t + \cdots + \gamma^k \widetilde{dw}_{t-k}$. Thus, the only difference between this strategy and the strategy in (11) is the Kyle term $\beta_t(w_t - p_t)dt$. But, as seen before, this term can be ignored if speculators want strategies that protect against the price containing information about other components of the fundamental value.

In Section 3 in the Internet Appendix, I consider an extension of the model $\mathcal{M}_1$ in which the speculators' signals are made public with a lag $k = 2$. For this extension, the precision of public news, which is measured by the ratio σ_w/σ_v (see equation (4)), becomes a parameter that continuously connects the benchmark model $\mathcal{M}_1$ with a strong-form efficient model in which the fundamental value is revealed with lag 2. It turns out that for most values of the news precision parameter (σ_w/σ_v less than 0.8), the speculators' equilibrium behavior in this extension is much closer to the benchmark model $\mathcal{M}_1$ than to the strong-form efficient model. Moreover, for the same parameter values, the contribution of the public news to price variance is usually less than 1% of the price variance due to the order flow. Thus, in the first approximation, public news can be ignored, and the results in the model $\mathcal{M}_1$ are robust to this extension.

Another issue in the benchmark model is the assumption (13) that the signal covariances do not depend on the speculators' strategies and are set by the dealers (e.g., the price impact coefficient λ depends on the covariance of the forecast w_t with the aggregate order flow dy_t, but is set by the dealer). To estimate the effect of this assumption, in Section 7 in the Internet Appendix, I analyze an extension $\mathcal{D}_1$, which is a discrete version of $\mathcal{M}_1$ in which in addition I allow these covariances to depend on speculators' strategies. Then, by taking the limit of $\mathcal{D}_1$ when the time interval approaches zero, one sees that the limit differs from $\mathcal{M}_1$ by a term of the order of $1/(N_F + N_S)^2$. Numerical results show that this difference is indeed very small.

One potential extension of the model is to consider traders who process information at different frequencies, that is, they receive signals every L periods, which would justify calling them high-frequency traders. Such a model appears too complicated to solve in closed form, and even numerically. Nevertheless, this alternative model appears to be a mixture of two types of models, one of which is essentially my benchmark model. To see this, suppose that $L = 2$ for low frequency traders (LFTs), and $L = 1$ for high-frequency traders (HFTs). Then, when t is even ($t = 2k$), both LFTs and HFTs receive updates, while when t is odd ($t = 2k + 1$), only the HFTs receive updates. So the proposed extension would be a mixture of the following two models: (i) one model (corresponding to t even) in which multiple informed traders use their current signals; (ii) one model (corresponding to t odd) in which HFTs use their current signal, while LFTs can only use their lagged signal. This is essentially my benchmark model with FTs and STs, if one assumes that larger lags are not used. Intuitively, my main results are likely to be true in the proposed setup. For example, HFTs make larger profits than LFTs because when t is odd, the HFTs have an informational advantage and use their current signals, while the LFTs use their lagged signals, which produces lower profits.[47]

I also examine several extensions of the model with inventory management, and show that the main results are robust. First, in Subsection 5.3 in the Internet Appendix, I consider a more general IFT strategy that has a component of trading on the lagged signal: $dx_t = -\Theta x_{t-1} + G dw_t + M\widetilde{dw}_{t-1}$.[48] Second, in Subsection 5.4 in the Internet Appendix, the dealer takes into account that the aggregate order flow has a predictable component, coming from the mean reversion term $-\Theta x_{t-1}$. In this extension, the dealer no longer sets (as in Section 4) the price change at $dp_t = \lambda_t dy_t$, where λ_t is determined by her expected profit being zero, but she correctly sets $dp_t = \lambda_t \widetilde{dy}_t$, where $\widetilde{dy}_t$ is the unanticipated part of the aggregate order flow at t.

The main result in Section 4 is that a sufficiently inventory-averse IFT optimally engages in quick inventory mean reversion, and in effect provides liquidity to the STs. In Subsection 5.5 in the Internet Appendix, I show in an extension to multiple IFTs that this result holds even if all FTs become IFTs, and the only speculators remaining are slow. In that case, the FTs (that are only IFTs) no longer speculate on the long-term value, but just pass their inventory (the "hot potato") to the slower traders.

[46] Here I denote by $\overline{\gamma}$, $\overline{\mu}$, and $\overline{\nu}$ the aggregate coefficients on the signals of lag 0, 1, and 2, respectively.

[47] Also, a HFT with inventory costs (call him the IFT) would also find it optimal to use quick mean reversion for his inventory, at least at times when t is odd, but most likely at all times (as long as the IFT's trading is correlated with aggregate trading next period, which can also come from the HFTs when they trade on their lagged signals).

[48] In Subsection 6.3 in the Internet Appendix, I also show that smooth strategies of the form $dx_t = -\theta x_{t-1} dt + G dw_t + M\widetilde{dw}_{t-1}$ are never optimal when $\theta \in (0, \infty)$.

Finally, in Subsection 5.6 of the benchmark model $\mathcal{M}_2$ that has both an inventory-averse fast trader (IFT) and an inventory-averse medium trader (IMT). Note that in $\mathcal{M}_2$ there are three types of speculators: fast traders (FTs), who at t observe $\mathrm{d}w_t$; medium traders (MTs), who at t observe $\mathrm{d}w_{t-1}$; and slow traders (STs), who at t observe $\mathrm{d}w_{t-2}$. In this extension, I consider an IFT with strategy $\mathrm{d}x_t = -\Theta x_{t-1} + G\mathrm{d}w_t$, and an IMT with strategy $\mathrm{d}z_t = -\Omega z_{t-1} + H\widetilde{\mathrm{d}w}_{t-1}$. This creates an intermediation chain with two links: (i) the IFT, who provides liquidity to the IMT and the MTs; and (ii) the IMT, who provides liquidity to the STs. Compared to the situation in which one link is missing, the chain with two links has the effect that G decreases for the IFT and H increases for the IMT. Intuitively, the IFT trades less aggressively on his signal (that is, G is lower) because the IFT now does not benefit as much from slower trading: the IMT (who is part of the slower trading) is not as aggressive as a regular MT in trading on his signal. By contrast, the IMT trades more aggressively on his signal (that is, H is higher) because for the IMT the liquidity provision by the IFT decreases the IMT's relative price impact from trading on his signal and thus makes him more aggressive.

6. Conclusion

I have presented a theoretical model in which traders continuously receive signals over time about the value of an asset, but only use each signal for a finite number of lags (which can be justified by an information processing cost per signal). I find that competition among speculators reveals much private information to the public, and the value of information decays fast. Therefore, a trader who is just one instant slower than the other traders loses the majority of the profits by being slow. Another consequence is that the market is very efficient and liquid. As a feedback effect, because of the small price impact (high market liquidity), the informed traders are capable of trading even more aggressively. In equilibrium, the fast speculator trading volume is very large and dominates the overall trading volume. I also considered an extension of the model in which a fast speculator, called the inventory-averse fast trader (IFT), has quadratic inventory costs. I find that a sufficiently averse IFT has a very different behavior compared to a risk-neutral fast trader. The IFT keeps his profits in cash, makes no fundamental bets on the value of the risky asset, and quickly passes his inventory to slow traders, who use their lagged signals. This hot potato effect is possible because the existence of slower traders more than reverses the price impact of the IFT.

Appendix A. Proofs

I start with some notation preliminaries. Recall that $t - k$ is notation for $t - k\mathrm{d}t$, and

$$T = 1. \tag{45}$$

For a process X_t, I denote by $\sigma_{X,t}$ the instantaneous volatility of X_t, which is the limit $\lim_{\Delta t\to 0} \frac{\mathrm{Var}(\Delta X_t)}{\Delta t}$, if this limit exists. In general, a tilde above a symbol denotes normalization by σ_w or σ_w^2. For instance, if σ_u is the instantaneous volatility of the noise trader order flow, and σ_y the instantaneous volatility of the total order flow, denote:

$$\widetilde{\sigma}_u = \frac{\sigma_u}{\sigma_w}, \quad \widetilde{\sigma}_y = \frac{\sigma_y}{\sigma_w}. \tag{46}$$

If $\mathrm{d}x_t$ is a trading strategy, $t \in (0, T]$, let $\widetilde{\pi}$ be the normalized expected profit at $t = 0$:

$$\widetilde{\pi} = \frac{1}{\sigma_w^2}\mathsf{E}\left(\int_0^T (w_t - p_t)\mathrm{d}x_t\right). \tag{47}$$

For covariances, a tilde above a symbol means normalization by both σ_w^2 and $\mathrm{d}t$:

$$\widetilde{\mathsf{Var}}\left(\widetilde{\mathrm{d}w}_t\right) = \frac{\mathsf{Var}\left(\widetilde{\mathrm{d}w}_t\right)}{\sigma_w^2\mathrm{d}t} = A_t, \quad \widetilde{\mathsf{Cov}}\left(w_t, \widetilde{\mathrm{d}w}_t\right) = \frac{\mathsf{Cov}\left(w_t, \widetilde{\mathrm{d}w}_t\right)}{\sigma_w^2\mathrm{d}t} = B_t. \tag{48}$$

Proof of Theorem 1. I look for an equilibrium with the following properties: (i) the equilibrium is symmetric, in the sense that the FTs have identical trading strategies, and the same for the STs; (ii) the equilibrium coefficients are constant with respect to time.

To solve for the equilibrium, in the first step the dealer's pricing functions are taken as given, and I solve for the optimal trading strategies for the FTs and STs. In the second step, the speculators' trading strategies are taken as given, and I compute the dealer's pricing functions. In Section 2, I assumed that the speculators take the signal covariance structure as given (see equation (13)). In the current context, this means that the speculators consider the following covariances A_t and B_t from (48) as fixed constants. Thus, in the rest of the Appendix, I assume that the dealer also sets A and B, in addition to setting λ and ρ.

Speculators' optimal strategy (γ, μ)

Since I search for an equilibrium with constant coefficients, I assume that the speculators take as given the dealer's pricing rules $\mathrm{d}p_t = \lambda\mathrm{d}y_t$ and $z_{t-1,t} = \rho\mathrm{d}y_{t-1}$, and also the covariances $A = \widetilde{\mathsf{Var}}(\widetilde{\mathrm{d}w}_t)$ and $B = \widetilde{\mathsf{Cov}}(w_t, \widetilde{\mathrm{d}w}_t)$.

Consider a FT, indexed by $i = 1, \ldots, N_F$. He chooses $\mathrm{d}x_t^i = \gamma_t^i\mathrm{d}w_t + \mu_t^i\widetilde{\mathrm{d}w}_{t-1}$, and assumes that at each $t \in (0, T]$, the price

satisfies[49]:

$$\mathrm{d}p_t = \lambda\,\mathrm{d}y_t, \quad \text{with} \quad \mathrm{d}y_t = \left(\gamma_t^i + \gamma_t^{-i}\right)\mathrm{d}w_t + \left(\mu_t^i + \mu_t^{-i}\right)\widetilde{\mathrm{d}w}_{t-1} + \mathrm{d}u_t, \tag{49}$$

where the superscript "$-i$" indicates the aggregate quantity from the other speculators. Since $\mathrm{d}w_t$ and $\widetilde{\mathrm{d}w}_{t-1}$ are both orthogonal on the public information set $\mathcal{I}_t$, and $p_{t-1}\in\mathcal{I}_t$, it follows that $\mathrm{d}x_t^i$ is orthogonal to p_{t-1} as well. The normalized expected profit of FT i at $\tau\in[0,T)$ satisfies:

$$\begin{aligned}\widetilde{\pi}_\tau^F &= \frac{1}{\sigma_w^2}\mathsf{E}\int_\tau^T \left(w_t - p_{t-1} - \lambda\left((\gamma_t^i+\gamma_t^{-i})\mathrm{d}w_t + (\mu_t^i+\mu_t^{-i})\widetilde{\mathrm{d}w}_{t-1} + \mathrm{d}u_t\right)\right)\mathrm{d}x_t^i \\ &= \int_\tau^T \left(\gamma_t^i - \lambda\gamma_t^i\left(\gamma_t^i+\gamma_t^{-i}\right) + \mu_t^i B - \lambda\mu_t^i\left(\mu_t^i+\mu_t^{-i}\right)A\right)\mathrm{d}t.\end{aligned} \tag{50}$$

This is a pointwise optimization problem, hence it is enough to consider the profit at $\tau = 0$, and maximize the expression over γ_t^i and μ_t^i. The solution of this problem is $\lambda\gamma_t^i = \frac{1-\lambda\gamma_t^{-i}}{2}$, and $\lambda\mu_t^i = \frac{B/A-\lambda\mu_t^{-i}}{2}$. The ST $j = 1,\dots,N_S$ solves the same problem, only the coefficient on $\mathrm{d}w_t$ is $\gamma_t^j = 0$. Thus, all $\gamma\prime$s are equal for the FTs, and all $\mu\prime$s are equal for the FTs and STs. I also find that they are constant, and since $N_L = N_F + N_S$, one obtains:

$$\gamma = \frac{1}{\lambda}\frac{1}{1+N_F}, \quad \mu = \frac{B/A}{\lambda}\frac{1}{1+N_L}. \tag{51}$$

Dealer's pricing rules (λ, ρ, A, B)

The dealer takes the speculators' strategies as given, and assumes that the aggregate order flow is of the form:

$$\mathrm{d}y_t = \mathrm{d}u_t + \overline{\gamma}\,\mathrm{d}w_t + \overline{\mu}\,\widetilde{\mathrm{d}w}_{t-1}, \quad \text{with} \quad \overline{\gamma} = N_F\,\gamma, \quad \overline{\mu} = N_L\,\mu. \tag{52}$$

Moreover, the dealer assumes that, in their trading strategy, the speculators set:

$$\widetilde{\mathrm{d}w}_{t-1} = \mathrm{d}w_{t-1} - \rho_*\,\mathrm{d}y_{t-1}. \tag{53}$$

Later I require that in equilibrium the dealer's pricing coefficient ρ coincides with the coefficient ρ_* used by the speculators.

Since the order flow $\mathrm{d}y_t$ is orthogonal to the dealer's information set $\mathcal{I}_t$, the dealer sets λ_t, ρ_t, A_t, and B_t, such that the following equations are satisfied:

$$\begin{aligned}\lambda_t &= \frac{\widetilde{\mathrm{Cov}}(w_t,\mathrm{d}y_t)}{\widetilde{\mathrm{Var}}(\mathrm{d}y_t)} = \frac{\overline{\gamma}+\overline{\mu}B_{t-1}}{\sigma_{y,t}^2}, \quad \mathrm{d}p_t = \lambda_t\mathrm{d}y_t, \\ \rho_t &= \frac{\widetilde{\mathrm{Cov}}(\mathrm{d}w_t,\mathrm{d}y_t)}{\widetilde{\mathrm{Var}}(\mathrm{d}y_t)} = \frac{\overline{\gamma}}{\sigma_{y,t}^2}, \quad \widetilde{\mathrm{d}w}_t = \mathrm{d}w_t - \rho_t\mathrm{d}y_t, \\ \sigma_{y,t}^2 &= \widetilde{\mathrm{Var}}(\mathrm{d}y_t^2) = \widetilde{\sigma}_u^2 + \overline{\gamma}^2 + \overline{\mu}^2 A_{t-1}, \\ B_t &= \widetilde{\mathrm{Cov}}\left(w_t, \mathrm{d}w_t - \rho_*\mathrm{d}y_t\right) = (1-\rho_*\overline{\gamma}) - \rho_*\overline{\mu}B_{t-1}, \\ A_t &= \widetilde{\mathrm{Var}}\left(\mathrm{d}w_t - \rho_*\mathrm{d}y_t\right) = 1 - 2\rho_*\overline{\gamma} + \rho_*^2\sigma_{y,t}^2 \\ &= 1 - 2\rho_*\overline{\gamma} + \rho_*^2(\widetilde{\sigma}_u^2+\overline{\gamma}^2) + \rho_*^2\overline{\mu}^2 A_{t-1}.\end{aligned} \tag{54}$$

Consider the last equation in (54), $A_t = 1 - 2\rho_*\overline{\gamma} + \rho_*^2(\widetilde{\sigma}_u^2+\overline{\gamma}^2) + \rho_*^2\overline{\mu}^2 A_{t-1}$, which is a recursive equation in A_t. Then, Lemma A.1 implies that A does not depend on t, as long as $|\rho_*\overline{\mu}| < 1$. But, since the dealer takes the speculators' strategies as given, I can use the equilibrium condition $\rho_*\overline{\mu} = b \in (0,1)$. The same method shows that B does not depend on t. Moreover, Lemma A.1 can be used to compute the constant values of A and B:

$$A = \frac{(1-\rho_*\overline{\gamma})^2 + \rho_*^2\widetilde{\sigma}_u^2}{1-(\rho_*\overline{\mu})^2}, \quad B = \frac{1-\rho_*\overline{\gamma}}{1+\rho_*\overline{\mu}}. \tag{55}$$

Then, equation (54) shows that λ, ρ, and $\widetilde{\sigma}_y$ are independent on t as well.

Equilibrium conditions

[49] By the assumption (13), the speculators take the covariance structure as computed by the dealer. By construction, the lagged signal $\widetilde{\mathrm{d}w}_{t-1}$ is orthogonal to the dealer's information set at time t, which includes the price p_{t-1}, hence the covariance $\mathrm{cov}(\widetilde{\mathrm{d}w}_{t-1}, p_{t-1})$ is set to zero.

I now use the equations derived above to solve for the equilibrium values of γ, μ, λ, $\rho = \rho_*$, A, B, and $\widetilde{\sigma}_y$. Denote:

$$a = \rho\overline{\gamma}, \quad b = \rho\overline{\mu}, \quad R = \frac{\lambda}{\rho}. \tag{56}$$

From (55), one gets $A = \frac{(1-a)^2+\rho^2\widetilde{\sigma}_u^2}{1-b^2}$. Then, substitute A in $\widetilde{\sigma}_y^2 = \widetilde{\sigma}_u^2 + \overline{\gamma}^2 + \overline{\mu}^2 A$ from (54) to obtain $\rho^2\widetilde{\sigma}_y^2 = \frac{\rho^2\widetilde{\sigma}_u^2+(a^2+b^2-2ab^2)}{1-b^2}$. To summarize,

$$B = \frac{1-a}{1+b}, \quad A = \frac{(1-a)^2+\rho^2\widetilde{\sigma}_u^2}{1-b^2}, \quad \rho^2\widetilde{\sigma}_y^2 = \frac{\rho^2\widetilde{\sigma}_u^2+(a^2+b^2-2ab^2)}{1-b^2}. \tag{57}$$

Using (54), one gets $R = \frac{\lambda}{\rho} = \frac{\overline{\gamma}+\overline{\mu}B}{\overline{\gamma}} = \frac{a+b\frac{1-a}{1+b}}{a} = \frac{a+b}{a(1+b)}$. Also, the equation for ρ implies $\rho = \frac{\overline{\gamma}}{\widetilde{\sigma}_y^2} = \frac{\rho a}{\rho^2\widetilde{\sigma}_y^2}$. Using the formula for $\rho^2\widetilde{\sigma}_y^2$ in (57), one computes $\rho^2\widetilde{\sigma}_u^2 = (1-a)(a-b^2)$. Using this formula, one obtains $\rho^2\widetilde{\sigma}_y^2 = a$ and $A = 1 - a$. To summarize,

$$R = \frac{\lambda}{\rho} = \frac{a+b}{a(1+b)}, \quad \rho^2\widetilde{\sigma}_u^2 = (1-a)(a-b^2), \quad \rho^2\widetilde{\sigma}_y^2 = a, \quad A = 1-a. \tag{58}$$

From (51), one has $\frac{N_F}{N_F+1} = \lambda\overline{\gamma} = \frac{\lambda}{\rho}a = \frac{a+b}{1+b}$. From this, $a = \frac{N_F-b}{N_F+1}$, and $B = \frac{1-a}{1+b} = \frac{\frac{1+b}{N_F+1}}{1+b} = \frac{1}{N_F+1}$. Also, $\frac{B}{A}\frac{N_L}{N_L+1} = \lambda\overline{\mu} = \frac{\lambda}{\rho}b = \frac{b(a+b)}{a(1+b)}$. Since $\frac{B}{A} = \frac{1}{1+b}$, one has $\frac{N_L}{N_L+1} = \frac{b(a+b)}{a}$, or $\frac{a}{b(1+b)}\frac{N_L}{N_L+1} = \frac{a+b}{1+b}$. The two formulas for $\frac{a+b}{1+b}$ imply $b(1+b)\frac{N_F}{N_F+1} = a\frac{N_L}{N_L+1}$. To summarize,

$$a = \frac{N_F-b}{N_F+1}, \quad B = \frac{1}{N_F+1}, \quad b(1+b)\frac{N_F}{N_F+1} = \frac{N_F-b}{N_F+1}\frac{N_L}{N_L+1}. \tag{59}$$

From $\frac{\lambda}{\rho}a = \frac{N_F}{N_F+1}$ and $a = \frac{N_F-b}{N_F+1}$, one gets $\frac{\lambda}{\rho} = \frac{N_F}{N_F-b}$, as stated.

From (59), one obtains the quadratic equation $b^2 + b\omega = \frac{N_L}{N_L+1}$, with $\omega = 1 + \frac{1}{N_F}\frac{N_L}{N_L+1}$. One solution of this quadratic equation is $b = \frac{\omega+\left(\omega+4\frac{N_L}{N_L+1}\right)^{1/2}}{2} \geq 1$, which leads to a negative $\widetilde{\sigma}_y^2$ (see (57)). Thus, I choose the other solution, $b = \frac{-\omega+\left(\omega+4\frac{N_L}{N_L+1}\right)^{1/2}}{2} \geq 0$. Let $b_\infty = \frac{\sqrt{5}-1}{2}$. Since $b_\infty^2 + b_\infty = 1$ and $\omega \geq 1$, one has $b_\infty^2 + b_\infty\omega \geq 1$. Moreover, since $b^2 + b\omega = \frac{N_L}{N_L+1} < 1$, one gets $b^2 + b\omega < b_\infty^2 + b_\infty\omega$. But the function $b^2 + b\omega$ is strictly increasing in b when $b \geq 0$, hence one obtains $b < b_\infty$. Thus, $b \in [0, b_\infty)$, as stated in Theorem 1. I also obtain $a = \frac{N_F-b}{N_F+1} \in (0,1)$. The proof of the exact formulas in (17) is complete.

I next derive the asymptotic formulas in (17). When N_F is large, note that $a = \frac{N_F}{N_F-b} \approx a_\infty = 1$, $\omega = 1 + \frac{1}{N_F}\frac{N_L}{N_L+1} \approx \omega_\infty = 1$. Therefore, one also gets $b \approx b_\infty = \frac{\sqrt{5}-1}{2}$. It is simple to verify that the formulas for γ_∞, μ_∞, λ_∞, and ρ_∞ are as stated in (17).

I next show how b depends on N_F and N_L (the dependence on N_S is the same as the dependence on $N_L = N_F + N_S$). Consider the function $F(\beta,\omega) = \sqrt{\omega^2+4\beta} - \omega$, and note that $b = F(\beta,\omega)/2$, with $\beta = \frac{N_L}{N_L+1}$ and $\omega = 1 + \frac{\beta}{N_F}$. One computes $\frac{\partial\beta}{\partial N_F} = \frac{\partial\beta}{\partial N_L} = \frac{1}{(N_L+1)^2}$, $\frac{\partial\omega}{\partial N_F} = -\frac{N_L(N_L+1)-N_F}{N_F^2(N_L+1)^2} < 0$, $\frac{\partial\omega}{\partial N_L} = \frac{1}{N_F(N_L+1)^2} > 0$. Also, $\frac{\partial F}{\partial\beta} = \frac{2}{\sqrt{\omega^2+4\beta}} > 0$, and $\frac{\partial F}{\partial\omega} = \frac{\beta}{\sqrt{\omega^2+4\beta}} - 1 = -\frac{b}{\sqrt{\omega^2+4\beta}} < 0$. Then, $\frac{\partial(2b)}{\partial N_F} = \frac{\partial F}{\partial\beta}\cdot\frac{\partial\beta}{\partial N_F} + \frac{\partial F}{\partial\omega}\cdot\frac{\partial\omega}{\partial N_F} > 0$, and $\frac{\partial(2b)}{\partial N_L} = \frac{\partial F}{\partial\beta}\cdot\frac{\partial\beta}{\partial N_L} + \frac{\partial F}{\partial\omega}\cdot\frac{\partial\omega}{\partial N_L} = \frac{1}{(N_L+1)^2\sqrt{\omega^2+4\beta}}\left(2 - \frac{b}{N_F}\right) > 0$, where the last inequality follows from $b \in (0,1)$.

I end the analysis of the equilibrium conditions, by proving several more useful inequalities for a and b. Let $\beta_F = \frac{N_F}{N_F+1}$, and recall that $\beta = \frac{N_L}{N_L+1}$. Then, b satisfies the quadratic equation $b^2 + b\omega = \beta$, with $\omega = 1 + \frac{\beta}{N_F}$. Start with the straightforward inequality $\beta < \beta_F + 1$, and multiply it by β_F. One gets $\beta\beta_F < \beta_F^2 + \beta_F$. Since $\beta_F = 1 - \frac{\beta_F}{N_F}$, one gets $\beta(1 - \frac{\beta_F}{N_F}) < \beta_F^2 + \beta_F$, or equivalently $\beta < \beta_F^2 + \beta_F(1 + \frac{\beta}{N_F})$. Since $b^2 + b\omega = \beta$ and $\omega = 1 + \frac{\beta}{N_F}$, one gets $b^2 + b\omega < \beta_F^2 + \beta_F\omega$. Because the function $f(x) = x^2 + x\omega$ is increasing in $x \in (0,1)$, one has $b < \beta_F = \frac{N_F}{N_F+1}$. This inequality is equivalent to $N_F - b > N_F b$. Dividing by $N_F + 1$, one gets $a = \frac{N_F-b}{N_F+1} > \frac{N_F b}{N_F+1} = b\beta_F$. But I have already shown that $\beta_F > b$, hence $a > b\beta_F > b^2$. To summarize,

$$b < \frac{N_F}{N_F+1}, \quad a > b^2. \tag{60}$$

Lemma A.1 can now be used to show that the coefficients A and B are constant. Indeed, in the proof of the Theorem, both A_t and B_t satisfy recursive equations of the form $X_t = \alpha + \beta X_{t-1}$, with $\beta \in (-1,1)$. Then, Lemma A.1 implies that X_t converges to a fixed number $\frac{\alpha}{1-\beta}$, regardless of the starting point. But, since the model is set in continuous time, and $t + 1$ actually stands

for $t + \mathrm{d}t$, the convergence occurs in an infinitesimal amount of time. Thus, X_t is constant for all t, and that constant is equal to $\frac{\alpha}{1-\beta}$. ■

I now state the Lemma that is used in the Proof of Theorem 1.

Lemma A.1. *Let $X_1 \in \mathbb{R}$, and consider a sequence $X_t \in \mathbb{R}$, which satisfies the following recursive equation:*

$$X_t - \beta X_{t-1} = \alpha, \quad t \geq 2. \tag{61}$$

Then the sequence X_t converges to $\overline{X} = \frac{\alpha}{1-\beta}$, regardless of the initial value of X_1, if and only if $\beta \in (-1, 1)$.

Proof. First, note that $\overline{X}$ is well defined as long as $\beta \neq 1$. Let $Y_t = X_t - \overline{X}$. Then, the new sequence Y_t satisfies the recursive equation $Y_t - \beta Y_{t-1} = 0$, which has the following general solution:

$$Y_t = C\beta^t, \quad t \geq 1, \quad \text{with} \quad C \in \mathbb{R}. \tag{62}$$

Thus, Y_t is convergent for any values of C if and only if all $\beta \in (-1, 1]$. But when $\beta = 1$, the value of $\overline{X}$ is not defined. This finishes the proof. ■

Proof of Corollary 1. In the Proof of Theorem 1, equation (51) implies $\lambda\overline{\gamma} = \frac{N_F}{N_F+1}$, $\lambda\overline{\mu} = \frac{B}{A}\frac{N_I}{N_L+1}$. But from (57) and (58), one has $\frac{B}{A} = \frac{1}{1+b}$, which proves the first row in (18). The second row in (18) just rewrites the formulas for A and B from equations (57) and (58). ■

Proof of Proposition 1. From Corollary 1, $\lambda\overline{\gamma} = \frac{N_F}{N_F+1}$ and $\lambda\overline{\mu} = \frac{B}{A}\frac{N_L}{N_L+1}$. From (50), the equilibrium normalized expected profit of the FT is:

$$\widetilde{\pi}^F = \gamma - \lambda\gamma\overline{\gamma} + \mu B - \lambda\mu\overline{\mu}A = \gamma\left(1 - \frac{N_F}{N_F+1}\right) + B\mu\left(1 - \frac{N_L}{N_L+1}\right) \tag{63}$$

From (59), $B = \frac{1}{N_F+1}$, which proves the desired formula for π^F. The profit of the ST is the same as for the FT, but with $\gamma = 0$. The last statement now follows from the asymptotic results in Theorem 1. ■

Justification of Result 1. According to Proposition 1, $\Delta\mathrm{d}t$ is the expected profit that speculators get per unit of time $\mathrm{d}t$ from trading on their lagged signal ($\widetilde{\mathrm{d}w}_{t-1}$). Given that all speculators break even on this lag, they would not trade on any signal with a larger lag, as this would cost them the same (Δ), but would bring a lower profit. For this last statement, I use Proposition IA.3 in Section 1 in the Internet Appendix, which shows numerically and asymptotically that the profit generated by lagged signals is decreasing in the number of lags. ■

Proof of Corollary 2. One simply follows the Proof of Theorem 1 to solve for the equilibrium in the $\mathcal{M}_{0,1}$ model. The key step is to observe that in Theorem 1 the FT's choice of μ is the same as the ST's choice of μ, and therefore it does not matter who does the optimization, as long as the total number of speculators using their lagged signal is the same. ■

Proof of Proposition 2. Since $1 - a = \frac{1+b}{N_F+1}$, equation (17) implies that $\lambda = \rho\frac{N_F}{N_F-b} = \frac{\sigma_w}{\sigma_u}\sqrt{(1-a)(a-b^2)}\frac{N_F}{N_F-b} = \frac{\sigma_w}{\sigma_u}\frac{\sqrt{(1+b)(a-b^2)}}{\sqrt{N_F+1}}\frac{N_F}{N_F-b}$, which proves the first equation in (25).

By definition, the trading volume is $TV = \sigma_y^2$. From (58), $TV = \sigma_y^2 = \widetilde{\sigma}_y^2\sigma_w^2 = \frac{a\sigma_w^2}{\rho^2}$. From (17), $\rho^2 = \frac{\sigma_w^2}{\sigma_u^2}(1-a)(a-b^2)$, hence $TV = \sigma_u^2\frac{a}{(1-a)(a-b^2)}$. Substituting $1 - a = \frac{1+b}{N_F+1}$, one gets $TV = \sigma_u^2(N_F+1)\frac{a}{(1+b)(a-b^2)}$, which proves the second equation in (25).

The price volatility is $\sigma_p^2 = \lambda^2 TV = \left(\frac{\lambda}{\rho}\right)^2\rho^2 TV = \left(\frac{\lambda}{\rho}\right)^2 a\sigma_w^2$. From (17), $\frac{\lambda}{\rho} = \frac{N_F}{N_F-b}$, hence $\sigma_p^2 = \left(\frac{N_F}{N_F-b}\right)^2\frac{N_F-b}{N_F+1}\sigma_w^2 = \frac{N_F^2}{(N_F+1)(N_F-b)}\sigma_w^2$, which proves the third equation in (25).

The speculator participation rate is $SPR = \frac{\overline{\gamma}^2\sigma_w^2 + \overline{\mu}^2\sigma_{\widetilde{w}}^2}{TV} = \frac{\rho^2(\overline{\gamma}^2\sigma_w^2 + \overline{\mu}^2\sigma_{\widetilde{w}}^2)}{a\sigma_w^2}$. Since $\rho\overline{\gamma} = a$, $\rho\overline{\mu} = b$, and $\sigma_{\widetilde{w}}^2 = (1-a)\sigma_w^2$, one gets $SPR = \frac{a^2+b^2(1-a)}{a}$. This proves the last equation in (25), since $\frac{1-a}{a} = \frac{1+b}{N_F-b}$. ■

Proof of Proposition 3. As in Theorem 1, I start with the FTs' choice of optimal trading strategy. Each FT $i = 1, \ldots, N_F$ observes $\mathrm{d}w_t$, and chooses $\mathrm{d}x_t^i = \gamma_t^i\mathrm{d}w_t$ to maximize the expected profit:

$$\pi_0 = \mathsf{E}\left(\int_0^T \left(w_t - p_{t-1} - \lambda_t(\mathrm{d}x_t^i + \mathrm{d}x_t^{-i} + \mathrm{d}u_t)\right)\mathrm{d}x_t^i\right) = \int_0^T \gamma_t^i\sigma_w^2\mathrm{d}t - \lambda_t\gamma_t^i\left(\gamma_t^i + \gamma_t^{-i}\right)\sigma_w^2\mathrm{d}t, \tag{64}$$

where the superscript "$-i$" indicates the aggregate quantity from the other FTs. This is a pointwise quadratic optimization problem, with solution $\lambda_t\gamma_t^i = \frac{1-\lambda_t\gamma_t^{-i}}{2}$. Since this is true for all $i = 1, \ldots, N_F$, the equilibrium is symmetric and one computes $\gamma_t = \frac{1}{\lambda_t}\frac{1}{1+N_F}$.

The dealer takes the FTs' strategies as given, thus assumes that the aggregate order flow is of the form $\mathrm{d}y_t = \mathrm{d}u_t + N_F\gamma_t\mathrm{d}w_t$.

To set λ_t, the dealer sets p_t such that $\mathrm{d}p_t = \lambda_t \mathrm{d}y_t$, with $\lambda_t = \frac{\mathrm{Cov}(w_t, \mathrm{d}y_t)}{\mathrm{Var}(\mathrm{d}y_t)} = \frac{N_F\gamma_t\sigma_w^2}{\sigma_u^2 + N_F^2\gamma_t^2\sigma_w^2}$. This implies $\lambda_t^2\sigma_u^2 + (N_F\gamma_t\lambda_t)^2\sigma_w^2 = N_F\gamma_t\lambda_t\sigma_w^2$. But $N_F\lambda_t\gamma_t = \frac{N_F}{N_F+1}$. Hence, $\lambda_t^2\sigma_u^2 + \left(\frac{N_F}{N_F+1}\right)^2\sigma_w^2 = \frac{N_F}{N_F+1}\sigma_w^2$, or $\lambda_t^2\sigma_u^2 = \frac{N_F}{(N_F+1)^2}\sigma_w^2$, which implies the formula $\lambda = \frac{\sigma_w}{\sigma_u}\frac{\sqrt{N_F}}{N_F+1}$. I then compute $\gamma_t = \frac{1}{\lambda_t}\frac{N_F}{1+N_F} = \frac{\sigma_u}{\sigma_w}\frac{1}{\sqrt{N_F}}$.

One has $TV = \sigma_y^2 = N_F^2\gamma^2\sigma_w^2 + \sigma_u^2$. But $N_F\gamma = \frac{\sigma_u}{\sigma_w}\sqrt{N_F}$, hence $TV = \sigma_u^2(1+N_F)$. Next, $\sigma_p^2 = \lambda^2 TV = \frac{\sigma_w^2}{\sigma_u^2}\frac{N_F}{(N_F+1)^2}\sigma_u^2(N_F+1) = \sigma_w^2\frac{N_F}{N_F+1}$. Also, $SPR = \frac{TV-\sigma_u^2}{TV} = \frac{\sigma_u^2(N_F+1)-\sigma_u^2}{\sigma_u^2(N_F+1)} = \frac{N_F}{N_F+1}$.

Finally, one computes Σ'. From the formula above for λ, one gets $\mathrm{Var}(\mathrm{d}p_t) = \lambda^2\mathrm{Var}(\mathrm{d}y_t) = \lambda\mathrm{Cov}(w_t, \mathrm{d}y_t) = \mathrm{Cov}(w_t, \mathrm{d}p_t)$. Since $\Sigma_t = \mathrm{Var}(w_t - p_{t-1}) = \mathsf{E}\left((w_t - p_{t-1})^2\right)$, one computes $\Sigma_t' = \frac{1}{\mathrm{d}t}\mathsf{E}\left(2(\mathrm{d}w_{t+1} - \mathrm{d}p_t)(w_t - p_{t-1}) + (\mathrm{d}w_{t+1} - \mathrm{d}p_t)^2\right) = -2\frac{\mathrm{Cov}(w_t, \mathrm{d}p_t)}{\mathrm{d}t} + \sigma_w^2 + \frac{\mathrm{Var}(\mathrm{d}p_t)}{\mathrm{d}t} = \sigma_w^2 - \sigma_p^2 = \frac{\sigma_w^2}{N_F+1}$. ■

Proof of Proposition 4. I use the formulas from the Proof of Theorem 1. Since $\widetilde{\mathrm{d}w}_t$ is orthogonal on $\mathrm{d}y_t$, one has $\widetilde{\mathrm{Cov}}\left(\widetilde{\mathrm{d}w}_t, \mathrm{d}w_t\right) = \widetilde{\mathrm{Cov}}\left(\widetilde{\mathrm{d}w}_t, \widetilde{\mathrm{d}w}_t\right) = A = 1 - a = \frac{1+b}{N_F+1}$. Then, $\widetilde{\mathrm{Cov}}\left(\widetilde{\mathrm{d}w}_t, \widetilde{\mathrm{d}w}_{t-1}\right) = \widetilde{\mathrm{Cov}}\left(\mathrm{d}w_t - \rho\overline{\gamma}\mathrm{d}w_t - \rho\overline{\mu}\widetilde{\mathrm{d}w}_{t-1}, \widetilde{\mathrm{d}w}_{t-1}\right) = -\rho\overline{\mu}A$. Therefore,

$$\begin{aligned}\widetilde{\mathrm{Cov}}\left(\mathrm{d}\overline{x}_{t+1}, \mathrm{d}\overline{x}_t\right) &= \widetilde{\mathrm{Cov}}\left(\overline{\gamma}\mathrm{d}w_{t+1} + \overline{\mu}\widetilde{\mathrm{d}w}_t, \overline{\gamma}\mathrm{d}w_t + \overline{\mu}\widetilde{\mathrm{d}w}_{t-1}\right) = \overline{\mu\gamma}A + \overline{\mu}^2(-bA)\\ \widetilde{\mathrm{Var}}\left(\mathrm{d}\overline{x}_t\right) &= \widetilde{\mathrm{Var}}\left(\overline{\gamma}\mathrm{d}w_t + \overline{\mu}\widetilde{\mathrm{d}w}_{t-1}\right) = \overline{\gamma}^2 + \overline{\mu}^2 A.\end{aligned} \tag{65}$$

By multiplying both the numerator and denominator by ρ^2, one computes:

$$\rho_{\overline{x}} = \frac{\overline{\mu\gamma}A}{\overline{\gamma}^2 + \overline{\mu}^2 A} - \frac{b\overline{\mu}^2 A}{\overline{\gamma}^2 + \overline{\mu}^2 A} = \frac{ab(1-a)}{a^2 + b^2(1-a)} - \frac{b^3(1-a)}{a^2 + b^2(1-a)} = \rho_{AT} + \rho_{EA}. \tag{66}$$

Then, $\rho_{\overline{x}} = \frac{ab-b^3}{a^2+b^2(1-a)}(1-a) = \frac{(a-b^2)b}{a^2+b^2(1-a)}\frac{1+b}{N_F+1}$, which implies the desired formulas.

I next prove that $\rho_{\overline{x}} > 0$ if and only if there is slow trading. When there is no slow trading, $b = \rho\overline{\mu} = 0$, hence $\rho_{\overline{x}} = 0$. When there is slow trading, I show that $\rho_{\overline{x}} = \frac{b(b+1)(a-b^2)}{a^2+b^2(1-a)}\frac{1}{N_F+1} > 0$. Indeed, one has $b > 0$, $a < 1$, and from equation (60), $a - b^2 > 0$. ■

Proof of Proposition 5. By definition, $\mathrm{d}(x_t p_t) = x_t p_t - x_{t-1}p_{t-1} = p_t\mathrm{d}x_t + x_{t-1}\mathrm{d}p_t$. Integrating this equality, one gets $x_T p_T - x_0 p_0 = \int_0^T p_t\mathrm{d}x_t + \int_0^T x_{t-1}\mathrm{d}p_t$. But $x_T = x_0 = 0$ (almost surely), hence $-\int_0^T p_t\mathrm{d}x_t = \int_0^T x_{t-1}\mathrm{d}p_t$. One also has $\int_0^T v_T\mathrm{d}x_t = v_T(x_T - x_0) = 0$. Thus, $\pi = \mathsf{E}\int_0^T (v_T - p_t)\mathrm{d}x_t = -\mathsf{E}\int_0^T p_t\mathrm{d}x_t = \mathsf{E}\int_0^T x_{t-1}\mathrm{d}p_t$. ■

Proof of Proposition 6. If x_t is the IFT's inventory in the risky asset, denote:

$$\Omega_t^{xx} = \frac{\mathsf{E}\left(x_t^2\right)}{\sigma_w^2\mathrm{d}t}, \quad X_t = \frac{\mathsf{E}\left(x_t\widetilde{\mathrm{d}w}_t\right)}{\sigma_w^2\mathrm{d}t}, \quad Z_t = \frac{\mathsf{E}\left(x_{t-1}\mathrm{d}y_t\right)}{\sigma_w^2\mathrm{d}t}. \tag{67}$$

Since $\Theta > 0$, one has $\Theta \in (0,2)$, or $\phi = 1 - \Theta \in (-1,1)$. From (32), x_t satisfies the recursive equation $x_t = \phi x_{t-1} + G\mathrm{d}w_t$. One computes $\Omega_t^{xx} = \frac{\mathsf{E}((x_t)^2)}{\sigma_w^2\mathrm{d}t} = \frac{\mathsf{E}((\phi x_{t-1} + G\mathrm{d}w_t)^2)}{\sigma_w^2\mathrm{d}t} = \phi^2\Omega_{t-1}^{xx} + G^2$. Since $\phi^2 \in (-1,1)$, I apply Lemma A.1 to the recursive formula $\Omega_t^{xx} = \phi^2\Omega_{t-1}^{xx} + G^2$. Then, Ω_t^{xx} is constant and equal to:

$$\Omega^{xx} = \frac{G^2}{1-\phi^2} = \frac{G^2}{\Theta(1+\phi)}, \tag{68}$$

which is the usual variance formula for the $AR(1)$ process. From (68) it follows that:

$$\mathsf{E}\left(x_t^2\right) = \Omega^{xx}\sigma_w^2\mathrm{d}t, \tag{69}$$

which implies that the inventory is infinitesimal. It follows that the inventory costs are zero, and all the profits are in cash. Also, the IFT's expected utility is the same as his expected profit. As the initial inventory is $x_0 = 0$, one has that $x_T = 0$, and Proposition 5 implies that the IFT's expected profit is:

$$\pi_{\Theta>0} = \lambda\mathsf{E}\int_0^T x_{t-1}\mathrm{d}y_t = \lambda\int_0^T Z_t\mathrm{d}t. \tag{70}$$

The order flow at t is $\mathrm{d}y_t = -\Theta x_{t-1} + \overline{\gamma}\mathrm{d}w_t + \overline{\mu}\widetilde{\mathrm{d}w}_{t-1} + \mathrm{d}u_t$, with $\overline{\gamma} = \gamma^- + G$. Then, Z_t is a function of X_{t-1}:

$$Z_t = \frac{\mathsf{E}\left(x_{t-1}\mathrm{d}y_t\right)}{\sigma_w^2\mathrm{d}t} = -\Theta\Omega_{t-1}^{xx} + \overline{\mu}X_{t-1} = -\frac{G^2}{1+\phi} + \overline{\mu}X_{t-1}. \tag{71}$$

The recursive formula for X_t is $X_t = \frac{\mathsf{E}(x_t \widetilde{\mathrm{d}w_t})}{\sigma_w^2 \mathrm{d}t} = \frac{\mathsf{E}((\phi x_{t-1} + G\mathrm{d}w_t)(\mathrm{d}w_t - \rho \mathrm{d}y_t))}{\sigma_w^2 \mathrm{d}t} = -\phi\rho Z_t + G - G\rho\overline{\gamma} = -\phi\rho\overline{\mu}X_{t-1} + \phi\frac{\rho G^2}{1+\phi} + G - G\rho\overline{\gamma} = -\phi b X_{t-1} + G(1-a^-) - \frac{\rho G^2}{1+\phi}$. By assumption, $0 \leq b < 1$, hence $\phi b \in (-1,1)$. Lemma A.1 implies that X_t is constant and equal to:

$$X = \frac{G(1-a^-) - \frac{\rho G^2}{1+\phi}}{1+\phi b}. \tag{72}$$

From (71), Z_t is also constant and satisfies:

$$Z = \overline{\mu}X - \frac{G^2}{1+\phi} = \overline{\mu}G\frac{1-a^-}{1+\phi b} - G^2\frac{b+\frac{1}{1+\phi}}{1+\phi b}. \tag{73}$$

From (70), the IFT's expected profit is:

$$\widetilde{\pi}_{\Theta>0} = \lambda Z = \lambda\left(\overline{\mu}G\frac{1-a^-}{1+\phi b} - G^2\frac{b+\frac{1}{1+\phi}}{1+\phi b}\right). \tag{74}$$

This finishes the proof. ■

Proof of Theorem 2. Let $\Theta = 0$. Then, the IFT's strategy is of the form $\mathrm{d}x_t = G\mathrm{d}w_t$. The IFT's expected profit is $\pi_{\Theta=0} = \mathsf{E}\int_0^T (w_t - p_t)\mathrm{d}x_t = \mathsf{E}\int_0^1 \left(w_{t-1} - p_{t-1} + \mathrm{d}w_t - \lambda \mathrm{d}y_t\right)\left(G\mathrm{d}w_t\right) = \mathsf{E}\int_0^1 \left(\mathrm{d}w_t - \lambda\mathrm{d}y_t\right)\left(G\mathrm{d}w_t\right) = \mathsf{E}\int_0^1 \left(\mathrm{d}w_t - \lambda\overline{\gamma}\mathrm{d}w_t\right)\left(G\mathrm{d}w_t\right) = G(1-\lambda\overline{\gamma})\sigma_w^2$. But $\lambda\overline{\gamma} = \lambda G + \lambda\gamma^- = \lambda G + Ra^-$. The IFT's normalized expected profit is:

$$\widetilde{\pi}_{\Theta=0} = G(1-\lambda\overline{\gamma}) = G(1-Ra^-) - \lambda G^2. \tag{75}$$

To compute the IFT's inventory costs, denote by $\Omega_t^{xx} = \frac{\mathsf{E}(x_t^2)}{\sigma_w^2}$. One computes $\frac{\mathrm{d}\Omega_t^{xx}}{\mathrm{d}t} = \frac{1}{\sigma_w^2 \mathrm{d}t}\mathsf{E}\left(2x_{t-1}\mathrm{d}x_t + (\mathrm{d}x_t)^2\right) = \frac{1}{\sigma_w^2 \mathrm{d}t}\mathsf{E}\left(2Gx_{t-1}\mathrm{d}w_t + G^2(\mathrm{d}w_t)^2\right) = G^2$. Since $\Omega_0^{xx} = 0$, the solution of this first order ODE is $\Omega_t^{xx} = tG^2$, for all $t \in [0,1]$. Hence, the inventory costs are equal to:

$$C_I\mathsf{E}\int_0^1 x_t^2\mathrm{d}t = C_I G^2\int_0^1 t\mathrm{d}t = \frac{C_I}{2}G^2. \tag{76}$$

From (75) and (76), the IFT's normalized expected utility when $\Theta = 0$ is:

$$\widetilde{U}_{\Theta=0} = G(1-Ra^-) - G^2\left(\lambda + \frac{C_I}{2}\right). \tag{77}$$

The function $\widetilde{U}_{\Theta=0}$ attains its maximum at $G = \frac{1-Ra^-}{2\lambda+C_I} = \frac{1-Ra^-}{2\lambda\left(1+\frac{C_I}{2\lambda}\right)}$, as stated in Theorem 2. The maximum value is:

$$\widetilde{U}^{\max}_{\Theta=0} = \frac{(1-Ra^-)^2}{2(2\lambda+C_I)}. \tag{78}$$

Let $\Theta > 0$, which is equivalent to $\phi = 1 - \Theta \in (-1,1)$. In the Proof of Proposition 6, I have computed the IFT's expected profit (see (38)), and shown that the IFT's inventory costs are zero. Hence, the IFT's expected utility is the same as his expected profit, and satisfies $\widetilde{U}_{\Theta>0} = \widetilde{\pi}_{\Theta>0} = \frac{\lambda}{\rho}\left(bG\frac{1-a^-}{1+\phi b} - \rho G^2\frac{b+\frac{1}{1+\phi}}{1+\phi b}\right)$. The first order condition with respect to G implies that at the optimum:

$$G = \frac{b(1-a^-)}{2\rho\left(b+\frac{1}{1+\phi}\right)}, \tag{79}$$

which expresses the optimum G as a function of ϕ. The second order condition for a maximum is $\lambda\frac{b+\frac{1}{1+\phi}}{1+\phi b} > 0$, which follows from $\lambda > 0, b \in [0,1)$, and $\phi \in (-1,1)$. For the optimum G, the normalized expected utility (profit) of the IFT is:

$$\widetilde{U}_{\Theta>0} = \frac{(Rb(1-a^-))^2}{4\lambda(1+\phi b)\left(b+\frac{1}{1+\phi}\right)}. \tag{80}$$

I now analyze the function:

$$f(\phi) = (1+\phi b)\left(b+\frac{1}{1+\phi}\right) \quad\Longrightarrow\quad f'(\phi) = \frac{b^2(1+\phi)^2 + b - 1}{(1+\phi)^2}. \tag{81}$$

The polynomial in the numerator has two roots:

$$\phi_1 = -1 + \frac{\sqrt{1-b}}{b} \quad \phi_2 = -1 - \frac{\sqrt{1-b}}{b}. \tag{82}$$

By assumption $b < 1$, hence both roots are real. Clearly, $\phi_2 < -1$. I show that $\phi_1 \in (-1, 1)$. First, note that ϕ_1 is decreasing in b. For $b = 1$, one has $\phi_1 = -1$, while for $b = \frac{\sqrt{17}-1}{8}$ (which satisfies $4b^2 + b = 1$) one has $\phi_1 = 1$. Since by assumption $\frac{\sqrt{17}-1}{8} < b < 1$, it follows that indeed $\phi_1 \in (-1, 1)$. Thus, $f'(\phi)$ is negative on $(-1, \phi_1)$ and positive on $(\phi_1, 1)$. Hence, $f(\phi)$ attains its minimum at $\phi = \phi_1$, which implies that the normalized expected utility $\widetilde{U}_{\Theta>0}$ from (80) attains its maximum at $\phi = \phi_1$, or $\Theta = 2 - \frac{\sqrt{1-b}}{b}$, as stated in Theorem 2. Also, if one substitutes $\phi = \phi_1$ in (79), one gets $G = \frac{1-a^-}{2\rho\left(1+\frac{1}{\sqrt{1-b}}\right)}$, as stated in Theorem 2. The maximum value (over both G and Θ) is:

$$\widetilde{U}^{\max}_{\Theta>0} = \frac{(Rb(1-a^-))^2}{4\lambda b(1+\sqrt{1-b})^2}. \tag{83}$$

To determine the cutoff value for the inventory aversion coefficient C_I, set $\widetilde{U}^{\max}_{\Theta=0} = \widetilde{U}^{\max}_{\Theta>0}$. From (78) and (83), algebraic manipulation shows that the cutoff value is $\overline{C}_I = 2\lambda\left(\frac{(1-Ra^-)^2(1+\sqrt{1-b})^2}{R^2 b(1-a^-)^2} - 1\right)$, as stated in Theorem 2. ■

Proof of Corollary 3. Let $\Theta > 0$. In the context of Theorem 2, $b > \frac{\sqrt{17}-1}{8} > 0$ and $\rho > 0$, hence $\overline{\mu} = \frac{b}{\rho} > 0$. The IFT's strategy is $\mathrm{d}x_t = -\Theta x_{t-1} + G\mathrm{d}w_t$, while the slow trading component is $\mathrm{d}\overline{x}^S_t = \overline{\mu}\widetilde{\mathrm{d}w}_{t-1}$. Since $\mathrm{d}w_t$ is orthogonal to $\widetilde{\mathrm{d}w}_{t-1}$, $\mathrm{Cov}(\mathrm{d}x_t, \mathrm{d}\overline{x}^S_t) = -\Theta\mathrm{Cov}(x_{t-1}, \mathrm{d}\overline{x}^S_t) = -\Theta\overline{\mu}\mathrm{Cov}(x_{t-1}, \widetilde{\mathrm{d}w}_{t-1})$. This proves the equality in (42). Since $\Theta > 0$ and $\overline{\mu} > 0$, one needs to prove the inequality $\mathrm{Cov}(x_{t-1}, \widetilde{\mathrm{d}w}_{t-1}) > 0$. But $\mathrm{Cov}(x_{t-1}, \widetilde{\mathrm{d}w}_{t-1}) = X\sigma_w^2\mathrm{d}t$ (see (67)). From (72), $X = \frac{G(1-a^-) - \frac{\rho G^2}{1+\phi}}{1+\phi b}$. Substituting the optimal G and $\phi = 1 - \Theta$ from Theorem 2, one obtains $X = \frac{(1-a^-)^2}{4(1+\sqrt{1-b})}$. As in Theorem 2, $a^-, b \in [0, 1)$, hence $X > 0$ and the proof is complete. ■

Proof of Theorem 3. Consider the following implicit equation in b:

$$\frac{2b(1+b)(2B+1)}{n_L} = \frac{Q}{B^2(a^- + b)} + \frac{3bB + 2b^2B - 1 - b}{b}(1 - a^-) - 2, \tag{84}$$

where the following substitutions are made[50]:

$$\begin{aligned}
&n_F = \frac{N_F}{N_F + 1}, \quad n_L = \frac{N_L}{N_L + 1}, \quad B = \frac{1}{\sqrt{1-b}},\\
&q = (B+1)(2(B^2-1) - n_F(3B^2-2)),\\
&a^- = \frac{-q \pm \sqrt{q^2 + n_F B^5((4-n_F)B + 2(2-n_F))}}{B^2((4-n_F)B + 2(2-n_F))},\\
&Q = B^3(a^-)^2 + 2(3B^3 + 3B^2 - 2B - 1)a^- + (B^3 + 2B^2 - 2).
\end{aligned} \tag{85}$$

I write the equations for the other coefficients:

$$\begin{aligned}
&R = \frac{4(B+1)B^2(a^- + b)}{Q}, \quad a = \frac{(2B+1)a^- + 1}{2(B+1)}\\
&\rho^2 = \left((a - b^2) + \frac{2bB - 1}{2B+1}(1-a)\right)(1-a)\frac{\sigma_w^2}{\sigma_u^2}, \quad \lambda = R\rho\\
&\Theta = 2 - \frac{\sqrt{1-b}}{b}, \quad G = \frac{1-a}{\rho(2B+1)}, \quad \gamma = \frac{a^-}{\rho N_F}, \quad \mu = \frac{b}{\rho N_L}.
\end{aligned} \tag{86}$$

The proof is now left to Subsections 5.1 and 5.2 in the Internet Appendix. ■

Proof of Proposition 7. See Subsection 5.2 in the Internet Appendix. ■

Proof of Proposition 8. See Subsection 5.2 in the Internet Appendix. ■

[50] To be rigorous, I have included the case when a^- is negative. However, numerically this case never occurs in equilibrium, because it leads to $\lambda < 0$, which contradicts the FT's second order condition (IA.315) in Section 5 in the Internet Appendix.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.finmar.2019.02.003.

References

Aït-Sahalia, Yacine, Sağlam, Mehment, 2017. High Frequency Market Making: Optimal Quoting. Working Paper, June 14.
Albuquerque, Rui, Miao, Jianjun, 2014. Advance information and asset prices. J. Econ. Theor. 149, 236–275.
Back, Kerry, Cao, Henry, Willard, Gregory, 2000. Imperfect competition among informed traders. J. Finance 55, 2117–2155.
Back, Kerry, Pedersen, Hal, 1998. Long-lived information and intraday patterns. J. Financ. Mark. 1, 385–402.
Baron, Matthew, Brogaard, Jonathan, Hagströmer, Björn, Kirilenko, Andrei, 2019. Risk and return in high-frequency trading. J. Financ. Quant. Anal. forthcoming.
Benos, Evangelos, Sagade, Satchit, 2016. Price discovery and the cross-section of high-frequency trading. J. Financ. Mark. 30, 54–77.
Bernhardt, Dan, Miao, Jianjun, 2004. Informed trading when information becomes stale. J. Finance 59, 339–390.
Biais, Bruno, Foucault, Thierry, Moinas, Sophie, 2015. Equilibrium fast trading. J. Financ. Econ. 116, 292–313.
Boehmer, Ekkehart, Fong, Kingsley, Wu, Julie, 2018a. Algorithmic Trading and Market Quality: International Evidence. Working Paper, July.
Boehmer, Ekkehart, Li, Dan, Saar, Gideon, 2018b. The competitive landscape of high-frequency trading firms. Rev. Financ. Stud. 31, 2227–2276.
Brogaard, Jonathan, 2011. High Frequency Trading and its Impact on Market Quality. Working Paper, December.
Brogaard, Jonathan, Hagströmer, Björn, Nordén, Lars, Ryan, Riordan, 2015. Trading fast and slow: colocation and market quality. Rev. Financ. Stud. 28 3367–3366.
Brogaard, Jonathan, Hendershott, Terrence, Riordan, Ryan, 2014. High-frequency trading and price discovery. Rev. Financ. Stud. 27, 2267–2306.
Budish, Eric, Cramton, Peter, Shim, John, 2015. The high-frequency trading arms race: frequent batch Auctions as a market design response. Q. J. Econ. 130, 1547–1622.
Cao, Huining Henry, Ma, Yuan, Ye, Dongyan, 2015. Disclosure, Learning, and Coordination. Working Paper, April 20.
Caldentey, René, Stacchetti, Ennio, 2010. Insider trading with a random deadline. Econometrica 78, 245–283.
Cartea, Álvaro, Penalva, José, 2012. Where is the value in high frequency trading? Q. J. Finance 2, 1–46.
Cespa, Giovanni, Foucault, Thierry, 2014. Sale of price information by exchanges: does it promote price discovery? Manag. Sci. 60, 148–165.
Chaboud, Alain, Chiquoine, Benjamin, Hjalmarsson, Erik, Vega, Clara, 2014. Rise of the machines: algorithmic trading in the foreign exchange market. J. Finance 69, 2045–2084.
Chau, Minh, Vayanos, Dimitri, 2008. Strong-form efficiency with monopolistic insiders. Rev. Financ. Stud. 18, 2275–2306.
Du, Songzi, Zhu, Haoxiang, 2017. What is the optimal trading frequency in financial markets? Rev. Econ. Stud. 84, 1606–1651.
Easley, David, O'Hara, Maureen, Yang, Liyan, 2016. Differential access to price information in financial markets. J. Financ. Quant. Anal. 51, 1071–1110.
Foster, Douglas, Viswanathan, S., 1996. Strategic trading when agents forecast the forecast of others. J. Finance 51, 1437–1478.
Foucault, Thierry, Hombert, Johan, Roşu, Ioanid, 2016. News trading and speed. J. Finance 71, 335–382.
Glode, Vincent, Opp, Christian, 2016. Asymmetric information and intermediation chains. Am. Econ. Rev. 106, 2699–2721.
Hasbrouck, Joel, Saar, Gideon, 2013. Low-latency trading. J. Financ. Mark. 16, 646–679.
Hendershott, Terrence, Jones, Charles, Menkveld, Albert, 2011. Does algorithmic trading improve liquidity? J. Finance 66, 1–33.
Hendershott, Terrence, Menkveld, Albert, 2014. Price pressures. J. Financ. Econ. 114, 405–423.
Hirschey, Nicholas, 2018. Do High-Frequency Traders Anticipate Buying and Selling Pressure? Working Paper, May.
Hirshleifer, David, Subrahmanyam, Avanidhar, Titman, Sheridan, 1994. Security analysis and trading patterns when some investors receive information before others. J. Finance 49, 1665–1698.
Ho, Thomas, Stoll, Hans, 1981. Optimal dealer pricing under transactions and return uncertainty. J. Financ. Econ. 9, 47–73.
Hoffmann, Peter, 2014. A dynamic limit order market with fast and slow traders. J. Financ. Econ. 113, 156–169.
Holden, Craig, Subrahmanyam, Avanidhar, 1992. Long-lived private information and imperfect competition. J. Finance 47, 247–270.
Kirilenko, Andrei, Kyle, Albert, Samadi, Mehrdad, Tuzun, Tugkan, 2017. The Flash Crash: the impact of high frequency trading on an electronic market. J. Finance 72, 967–998.
Kyle, Albert, 1985. Continuous auctions and insider trading. Econometrica 53, 1315–1335.
Li, Wei, 2017. High Frequency Trading with Speed Hierarchies. Working Paper, January.
Lyons, Richard, 1997. A simultaneous trade model of the foreign exchange hot potato. J. Int. Econ. 42, 275–298.
Madhavan, Ananth, Smidt, Seymour, 1993. An analysis of changes in specialist inventories and quotations. J. Finance 48, 1595–1628.
Menkveld, Albert, 2013. High frequency trading and the new-market makers. J. Financ. Mark. 16, 712–740.
Menkveld, Albert, 2016. The economics of high-frequency trading: taking stock. Ann. Rev. Financ. Econ. 8, 1–24.
Pagnotta, Emiliano, Philippon, Thomas, 2018. Competing on speed. Econometrica 86, 1067–1115.
SEC, 2010. Concept Release on Equity Market Structure. Release No. 34-61358; File No. S7-02-10.
Weller, Brian, October 7, 2012. High Frequency Intermediation. Working Paper.
Zhang, X. Frank, 2010. High Frequency Trading, Stock Volatility, and Price Discovery. Working Paper, November.