Fast and Slow Informed Trading*

Ioanid Roșu†

October 17, 2018

Abstract

This paper develops 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.

Keywords: Trading volume, inventory, volatility, high frequency trading, price impact, mean reversion.

---


†HEC Paris, Email: rosu@hec.fr.
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, September 13th 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” 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.² This coincidental arrival raises the question whether or not at least some of the HFTs do 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.³ But, despite the large role played by high frequency traders (HFTs) in the current financial landscape, there has been relatively little progress in explaining their strategies in connection with information processing.

We 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?

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,” New York Times, December 22nd 2010).
² Hendershott, Jones, and Menkveld (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, Chiquoine, Hjalmarsson, and Vega (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.
How can HFT order flow anticipate future order flow and returns? What explains the “intermediation chains” or “hot potato” effects found among HFTs (see Kirilenko et al. 2017, or Weller 2012)? Why do some HFTs have low inventories? Regarding the last question, some recent literature identifies HFTs as traders with both high trading volume and low inventories (see Kirilenko et al. 2017, SEC 2010). But then, a natural question arises: why would having low inventories be part of the definition of HFTs?

In this paper, we provide a theoretical model of informed trading with speed differences which parsimoniously addresses these questions. The word “speed” in our 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, we start with the Kyle (1985) model and modify it along several dimensions. 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, in the sense that some speculators receive 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 our model apart from previous models of informed trading. A key effect of this assumption is to prevent the “rat race” phenomenon discovered 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 our model, the speculators reveal only a fraction of their total private information, and this has a stabilizing effect on the equilibrium. Economically, we think of this assumption as equivalent to having a positive information processing cost per signal (and per trading round). Indeed, since one of our results is that the value of information decays fast,

---

4As in Kyle (1985), we assume that informed traders are market takers and thus submit only market orders; this is a plausible assumption for informed HFTs (see Brogaard, Hendershott, and Riordan 2014). We argue though that the model may also describe HFTs who behave like market makers, as we later show that fast traders (with sufficiently large inventory costs) partially trade in opposite direction to the slower traders, and thus in effect provide liquidity to these traders.

5Intuitively, 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
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, we restrict our attention to 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 (or FTs), who observe the signal instantly; and slow traders (or STs), who observe the signal after one lag. In this case, the equilibrium can be described in closed form.\(^6\)

Our first result is 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 have received. 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. (2018) find out 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 high-frequency traders, Hirschey (2017) finds that HFT order flow anticipates future order flow.

A related result is that volume, volatility and liquidity are increasing with the number of FTs. First, more competition from FTs makes the prices more informative overall, and thus increases liquidity (measured, as in Kyle 1985, by the inverse price impact coefficient). 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

\(6\)In the Internet Appendix we verify numerically that the main results of the particular case \( m = 1 \) carry through to the general case \((m > 1)\).
the aggregate FT trading volume to be increasing roughly linearly with the number of FTs. The effect of FTs on volatility is more muted but still positive; this is because in our model price volatility is bounded above by the fundamental volatility of the asset. Empirically, in line with our theoretical results, Hendershott, Jones, and Menkveld (2011), Boehmer, Fong, and Wu (2015), and Zhang (2010) document a positive effect of HFTs on liquidity. Moreover, the last two papers find a positive effect of HFTs on volatility. We should point out, however, that our model is more likely to apply only to the subcategory of informed, market taking HFTs, and not to all HFTs. Our results should therefore be interpreted with caution.

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

To provide some theoretical guidance on these issues, we extend our benchmark model to include one trader with inventory costs. These costs can arise from risk aversion or from capital constraints, but we take a reduced form approach and assume the costs are quadratic in inventory, with a coefficient called inventory aversion (see Madhavan and Smidt 1993). We call this additional trader the Inventory-averse Fast Trader, or IFT. We 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. Without discussing yet optimality, suppose the IFT does inventory management, i.e., chooses a positive rate of inventory mean reversion. What are the effects of this choice?

---

7SEC (2010) characterizes HFTs by their “very short time-frames for establishing and liquidating positions” and argue 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 Kirilenko et al. (2017), Brogaard et al. (2015), or Menkveld (2013).

8Weller (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 would churn out their inventories very quickly to trade with other HFTs.

9The IFT is assumed fast because without slower traders it is not profitable to manage inventory. The case of several IFTs is discussed in the Internet Appendix 5.5, but the results are qualitatively similar.
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, IFT’s inventory tends to become small over many rounds, and because our model is set in the high frequency limit (in continuous time), the inventory becomes in fact negligible.\(^{10}\) We 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 each trading round. Thus, IFT’s direct revenue from each signal eventually decays to zero. We 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 we simply mean the part of order flow that involves the speculators’ lagged signals.\(^{11}\) To understand this effect, consider how the IFT uses a given signal. The information decay effect means that IFT’s final revenues from betting on his signal are zero. Therefore, the IFT must benefit from inventory reversal. Since any trade has price impact, inventory reversal makes a profit only if gets pooled with order flow in the opposition direction, so that IFT’s price impact is negative. But in order to be expected profit, the opposite order flow must come from speculators who use lagged signals, i.e., from slow trading. We call this result the hot potato effect, or the intermediation chain effect.\(^{12}\)

\(^{10}\)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.

\(^{11}\)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.

\(^{12}\)In our simplified framework, the intermediation chain only has one link, between the IFT and the slow traders. But we conjecture that in a model where speculators use more than one lag for their signals, the intermediation chains become longer, depending on the number of lags.
Figure 1: Optimal Inventory Mean Reversion. This figure plots 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 IFT’s inventory, measured by the square root of his average expected squared position in the stock, relative to FT’s inventory. On the vertical axis is IFT’s trading volume, measured by the instantaneous variance of his trading strategy, relative to 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 by Theorem 3 below, with parameters $\sigma_w = 1$, $\sigma_u = 1$, and with $N_F = N_S$ equal to either 2 or 20.

The reason behind this terminology is that 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. But more aggressive trading by the STs allows stronger mean reversion from the IFT.

Figure 1 illustrates the optimal behavior of the IFT as a function of his inventory aversion coefficient.\textsuperscript{13} 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

\textsuperscript{13}Inventory aversion is similar to risk aversion, but solving the model with a risk averse fast trader would be considerably more difficult.
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 revenues from the fundamental bets. Above the threshold, IFT’s optimal behavior changes dramatically: 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.

Our 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. Our results suggest that this is not necessarily the case. Indeed, Figure 1 suggests (and we 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 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).

We predict that in practice the fast speculators are sharply divided into two categories. In both categories speculators trade with a large volume. But 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 our model.

The division of fast speculators in two categories appears consistent with the empirical findings of Kirilenko et al. (2017), who study trading activity in the 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 our risk neutral fast traders: opportunistic traders have large volume, appear to be fast, and accumulate relatively large inventories. By contrast the “high frequency traders” in their paper, while they are also fast and trade in large volume, keep very low inventories. Indeed, HFTs in their sample liquidate 0.5% of their aggregate inventories on average each second.

**Related Literature**

Our paper contributes to the literature on trading with asymmetric information. We show that competition among informed traders, combined with noisy trading strategies, produces a large informed trading volume and a quick information decay.\(^{14}\) 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 Subrahmanyan (1992) and Foster and Viswanathan (1996). The former paper finds that the competition among informed traders is so strong, that in the continuous time limit there is no equilibrium in smooth strategies. Our 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.

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 our model the market is not strong-form efficient even in the limit, yet strategies are noisy. Foucault, Hombert, and Roșu (2016) propose a model in which a single speculator receives a signal one instant before public news. The speculator’s strategy is noisy, but

\(^{14}\)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.
for a different reason than in our model: the speculator optimally trades with a large
weight on his forecast of the news.\textsuperscript{15}

Our paper also contributes to the rapidly growing literature on High Frequency
Trading.\textsuperscript{16} In much of this literature, it is the speed \textit{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.\textsuperscript{17} By contrast, our results regarding volume and volatility remain true even
if all informed traders have the same speed. This is because in our model the need for
speed arises endogenously, from competition among informed traders. In our 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 our 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.

Our results regarding the optimal inventory of informed traders are, to our knowl-
dge, 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.\textsuperscript{18} Our 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 our model the
IFT does not take fundamental bets, passes his risky inventory to slower traders (the

\textsuperscript{15}In Cao, Ma, and Ye (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.

\textsuperscript{16}See Biais, Foucault, and Moinas (2015), Aït-Sahalia and Sağlam (2017), Budish, Cramton, and
Pagnotta and Philippon (2018), Weller (2012), Cartea and Penalva (2012); see also the survey by
Menkveld (2016), and the references therein.

\textsuperscript{17}See, e.g., the model with speed differences of Biais, Foucault, and Moinas (2015), or the model
of news anticipation of Foucault, Hombert, and Roșu (2016). There are other models with differential
access to fundamental information, e.g., Albuquerque and Miao (2014), and Bernhardt and Miao (2004);
or with differential access to price information, e.g., Cespa and Foucault (2014), and Easley, O’Hara,
and Yang (2016). These other models, however, do not directly study the effect of speed on traders’
strategies and their profits.

\textsuperscript{18}See Ho and Stoll (1981), Madhavan and Smidt (1993), Hendershott and Menkveld (2014), as well
as many references therein.
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, Subrahmanyam, and Titman (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, Subrahmanyam, and Titman (1994) is different, as they are interested in information acquisition and explaining behavior such as “herding” and “profit taking.” Our 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. Section 2 describes the model setup. Section 3 solves for the equilibrium in the particular case with two categories of traders: fast and slow, and discusses the effect of fast and slow traders on various measures of market quality. Section 4 introduces and extension of the baseline model in which a new trader (the IFT) has inventory costs. Then, it analyzes IFT’s optimal strategy and its effect on equilibrium. Section 5 discusses the robustness of our main results to various extensions. Section 6 concludes. All proofs are in the Appendix or the Internet Appendix. The Internet Appendix solves for the equilibrium in the general case, and analyzes several modifications and extensions of our baseline model.

2 Benchmark Model

We set our trading model in discrete time, with the eventual goal of describing the equilibrium as the number of trading periods approaches infinity, and the setup approaches a continuous time model on $[0, 1]$. We 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 $h = \Delta t > 0$ (see, e.g., Chau and Vayanos 2008). The level of a variable $v$ at the time $th$ is denoted

\[ v(\theta(\cdot)) = v(th, \theta(\cdot)) \]

Alternatively, we can consider a continuous time model over $[0, 1]$ where the trading intervals are of infinitesimal length $dt$ (see, e.g., Foucault, Hombert, and Roşu 2016). But in that case, 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. Thus, our solution is to consider the discretized model, and define traders’ profits as the limit when the number of periods approaches infinity.
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

$$v_T = \sigma_v B^\nu_T = \sum_{t=1}^T \sigma_v \Delta B^\nu_t,$$

(1)

where $B^\nu$ is a (continuous) Brownian motion over $[0, 1]$, and $\sigma_v > 0$ is a constant called the fundamental volatility. We 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 $\Delta v_t$ are then the short term changes in value due to the arrival of new information. The risk-free rate is assumed to be 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.

**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), we 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:

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

(2)

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

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

(3)

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

\[
\Delta w_t = \frac{\sigma_w^2}{\sigma^2 + \sigma^2_\eta} \Delta s_t, \quad \sigma_w^2 = \frac{\text{Var}(\Delta w_t)}{\Delta t} = \frac{\sigma^4}{\sigma^2 + \sigma^2_\eta}.
\]

When deriving empirical implications, we call \( \sigma_w \) the signal precision, as a precise signal (small \( \sigma_\eta \)) corresponds to a large \( \sigma_w \).

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 \( \ell \) periods before, \( \Delta s_{t-\ell} \).

**Trading and Prices.** At each \( t = 1, 2, \ldots, T \), denote by \( \Delta x^i_t \) the market order submitted by speculator \( i = 1, \ldots, N \) at \( t \), and by \( \Delta u_t \) 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

\[
\Delta y_t = \sum_{i=1}^N \Delta x^i_t + \Delta u_t.
\]

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 \( t \), \( \Delta y_t \), executes at

\[
p_t = \mathbb{E}(v_T | \mathcal{I}_t \cup \Delta y_t).
\]

Together with the price, another important quantity is the dealer’s expectation at \( t \) of the \( k \)-lagged signal \( \Delta w_{t-k} \):

\[
z_{t-k,t} = \mathbb{E}(\Delta w_{t-k} | \mathcal{I}_t).
\]

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

\[
\pi_\tau = \mathbb{E} \left( \sum_{t=\tau}^{T} (v_T - p_t) \Delta x_t \bigg| \mathcal{J}_\tau^{(\ell)} \right).
\]
As in Back, Cao, and Willard (2000), we focus on linear equilibria in which the trading strategy has a particular dependence on the traders’ forecasts. Specifically, we consider strategies which are linear in the unpredictable part of their signals,\(^{20}\)

\[
\Delta w_{t-k,t} = \Delta w_{t-k} - z_{t-k,t}, \quad k = \ell, \ell+1, \ldots
\]  

(9)

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

\[
\Delta x_t = \gamma_{\ell,t} \Delta w_{t-\ell,t} + \gamma_{\ell+1,t} \Delta w_{t-\ell-1,t} + \cdots + \gamma_{m,t} \Delta w_{t-m,t}.
\]  

(10)

As we are interested in the equilibrium behavior when \(\Delta t = \frac{1}{T}\) is small, we require that the \(\ell\)-speculator’s strategy is the discretization of a continuous-time strategy on \([0, 1]\). Recall that the subscript \(t\) refers to the actual time \(\frac{t}{T} \in [0, 1]\). We thus require that the coefficients \(\gamma_{k,t}\) of the strategy in (10) are continuous functions of time.\(^{21}\) To indicate that this is a continuous-time strategy, we use differential notation:

\[
dx_t = \gamma_{\ell,t} \dw_{t-\ell,t} + \gamma_{\ell+1,t} \dw_{t-\ell-1,t} + \cdots + \gamma_{m,t} \dw_{t-m,t},
\]  

(11)

where \(t\) is still regarded as an element of \(\{1, 2, \ldots, T\}\). If instead we regard \(t \in (0, 1]\), then the subscript \(t-k\) should be replaced by \(t-k dt\).\(^{22}\) In the rest of the paper, we preserve the ambiguity of the notation in (11), but to avoid confusion we often write integrals over \(t \in (0, T]\), and set \(T = 1\).

For the strategies in (11), we 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,

\(^{20}\)Intuitively, if the strategy had a predictable component, the dealer’s price would adjust and reduce the speculator’s profit. The unpredictability of speculators’ strategies can be proved quite generally, following Kyle (1985), as long as the speculators and the dealer are risk neutral.

\(^{21}\)This requirement implies that the coefficients \(\gamma_{k,t}\) are deterministic, and hence known at \(t = 0\). Other continuous time models such as Back, Cao, and Willard (2000) make similar assumptions. More generally, we can choose \(\gamma_{k,t}\) to be integrable (but deterministic) functions of \(t\).

\(^{22}\)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 \(t' - k dt\).
we use the integral sign to denote this limit:23

\[ \pi_\tau = E_\tau \left( \int_\tau^1 (v_T - p_t) dx_t \right) = \lim_{T \to \infty} E \left( \sum_{t \geq \tau T} (v_T - p_t) \Delta x_t \bigg| \mathcal{F}_\tau^{(t)} \right). \quad (12) \]

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 speculators’ 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 numbers

\[ Z_{j,k,t} = \text{Cov}(\Delta w_{t-j}, z_{t-k,t}) \quad (13) \]

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.24

**Model Notation.** If all speculators in the model have a strategy of the form (11) with the same \( m \geq 0 \), we call it the *benchmark model* with \( m \) lags, and write \( \mathcal{M}_m \). In the paper, we 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, we also call \( \mathcal{M}_1 \) the *model with fast and slow traders*.

If some \( \ell \)-speculators have strategies of the form (11) with different \( m_\ell \), we call this the general model with \( m \) lags, where \( m \) is the maximum of all \( m_\ell \). We are particularly interested in 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 \( (m_1 = 1) \). We call this the *general benchmark model* and denote it by \( \mathcal{M}_{0,1} \). In Section 3, we 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 \( \mathcal{M}_{0,1} \).

---

23 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. We thank the referees for pointing out this fact to us.

24 For instance, the price impact coefficient \( \lambda_t \) in the dealer’s pricing rule \( \Delta p_t = \lambda_t \Delta y_t \) is computed using the covariance term \( \text{Cov}(w_t, \Delta y_t) \) (see equation (A10)). Hence, even though a speculator affects \( \Delta y_t \) by his strategy, he can consider the covariance term \( \text{Cov}(w_t, \Delta y_t) \) to be independent of his strategy. We further discuss this assumption in Section 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, we consider an alternative model in which an $\ell$-speculator can use all past signals, but must pay a fixed cost $\delta_{\ell} \cdot dt$ 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 we show that for a particular value of $\delta$ the alternative model is equivalent to $\mathcal{M}_1$.

In choosing speculator strategies as in (11), we 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 to simply rely 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. Assumption (ii) is made for convenience, to avoid the problem of forecasting the forecasts of others (see 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, we show in an extension of the model (Internet Appendix Section 2.2) that the main predictions of the benchmark model remain qualitatively the same. In Section 5 we discuss assumptions (i) and (ii) in more detail.

3 Fast and Slow Traders

In this section, we 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, or FTs, who observe the signal with no delay (called 0-speculators in Section 2); and (ii) the Slow

---

25 We formalize this intuition in Internet Appendix Section 4, where we introduce an orthogonal dimension of the fundamental value, and show that trading strategies that rely on prices make an average loss.
Traders, or 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

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

where \( T = 1 \). Note that the weight \( \gamma_t \) 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., FT’s weight \( \gamma_t \) must be zero.\(^{26}\) The former case is the benchmark model \( \mathcal{M}_1 \). The latter case is the general benchmark model \( \mathcal{M}_{0,1} \).

Note that 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 \( \tilde{dw}_t = \tilde{dw}_{t-1} \) be the unanticipated part at \( t \) of the lagged signal. Then, the trading strategy in (14) can be written as

\[ dx_t = \gamma_t dw_t + \mu_t \tilde{dw}_{t-1}, \quad \text{with} \quad \tilde{dw}_{t-1} = dw_{t-1} - z_{t-1,t}. \quad (15) \]

### 3.1 Equilibrium

In this section we 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 (\( \mu_t \) is the same for all). Therefore, if we require 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 \( \mathcal{M}_{0,1} \), and we are justified in 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 there exists a closed-form linear equilibrium of the model. The equilibrium is symmetric, in the sense that the FTs have identical trading strategies, and so do the STs. We also provide asymptotic results when the number \( N_F \) of fast traders is large. If \( X \) is a variable that depends on \( N_F \), we say that \( X_{\infty} \) is the asymptotic value of a number \( X \) and write \( X \approx X_{\infty} \) whenever the ratio \( X/X_{\infty} \) converges to 1 as

\[^{26}\text{Intuitively, this can occur if the FT must pay a higher processing cost per signal than the ST; see Footnote 29.}\]
$N_F$ approaches infinity.

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

$$
\begin{align*}
\text{dx}_t^F &= \gamma dw_t + \mu \tilde{dw}_{t-1}, \\
\text{dx}_t^S &= \mu \tilde{dw}_{t-1}, \\
\tilde{dw}_{t-1} &= dw_{t-1} - \rho dy_{t-1}, \\
\text{dp}_t &= \lambda dy_t,
\end{align*}
$$

(16)

where the coefficients $\gamma$, $\mu$, $\rho$, $\lambda$ are given by:

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

(17)

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

Theorem 1 implies that FTs and STs trade with the same intensity ($\mu$) on their lagged signals. This is true because the current signal $dw_t$ is uncorrelated with the lagged signal $\tilde{dw}_{t-1}$, which implies that the FTs and the STs get the same expression for the expected profit that comes from the lagged signal.\(^{27}\)

We now discuss some comparative statics regarding the optimal weights $\gamma$ and $\mu$ (for brevity, we omit the proofs). The fast traders’ optimal weight $\gamma$ is decreasing in the number of fast traders, yet it is increasing in the number of slow traders. The first statement simply reflects that, when the number of fast traders is larger, these traders must divide the pie into smaller slices. The same logic applies to the coefficient on the lagged signal: $\mu$ is decreasing in both $N_F$ and $N_S$, as the fast and slow traders compete in trading on their common lagged signal. This last intuition also shows that the fast traders’ weight $\gamma$ is increasing in the number of slow traders. Indeed, when there is

\(^{27}\)This result does not generalize to the case when there are more lags ($m > 1$). In Internet Appendix Section 1, we see that 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 non zero entries $A_{i,j}$ when $i > j \geq 1$. 

18
more competition from slow traders, the fast traders have an incentive to trade more aggressively on their current signal, as the slow traders have not yet observed this signal.

The next result helps to get more intuition for the equilibrium.

**Corollary 1.** In equilibrium we have the following formulas

\[
\begin{align*}
\lambda \gamma &= \frac{N_F}{N_F + 1}, \\
\lambda \mu &= \frac{1}{1 + b} \frac{N_L}{N_L + 1}, \\
\text{Var}(\tilde{d}w_t) &= (1 - a) \sigma_w^2 = \frac{1 + b}{N_F + 1} \sigma_w^2, \\
\text{Cov}(\tilde{d}w_t, w_t) &= \frac{1 - a}{1 + b} \sigma_w^2 = \sigma_w^2 \frac{N_F}{N_F + 1}.
\end{align*}
\]

(18)

The first equation in (18) implies that \(\lambda \gamma d_w = \frac{N_F}{N_F + 1} d_w_t\), which shows that most of the current signal \((d_w_t)\) is incorporated into the price by the fast traders. The intuition comes from the Cournot nature of the equilibrium. Indeed, when trading on the current signal, the benefit of each of each FT increases linearly with the intensity of trading \(\gamma\) 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 \(N_F - 1\) fast traders aggregates to \(\frac{N_F - 1}{N_F + 1} d_w_t\), the FT is a monopsonist against the residual supply curve, and trades such that his price impact is half of \(\frac{2}{N_F + 1} d_w_t\), i.e., his price impact equals \(\frac{1}{N_F + 1} d_w_t\).

After incorporating \(\frac{N_F}{N_F + 1} d_w_t\) in trading round \(t\), the fast traders must compete with the slow traders for the remaining \(\frac{1}{N_F + 1} d_w_t\) in the next trading round. As explained before, the speculators must trade a multiple of the unanticipated part of the lagged signal, \(\tilde{d}w_t = d_w_t - \rho d_y_t\). Thus, when trading on the lagged signal, the benefit of each speculator—fast or slow—increases linearly with the intensity of trading \(\mu\), and is proportional to the covariance \(\text{Cov}(\tilde{d}w_t, w_t)\). 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 \(\text{Var}(\tilde{d}w_t)\). The argument is now similar to the Cournot one above, except that everything gets multiplied by the ratio \(\text{Cov}(\tilde{d}w_t, w_t) / \text{Var}(\tilde{d}w_t)\), 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 by the speculators into the price.
We use the results in Theorem 1 to compute the expected profits of the fast traders and the slow traders.

**Proposition 1.** *The expected profit of the FTs and STs at* \( t = 0 \) *from their equilibrium strategies are given, respectively, by:*

\[
\pi_F^{\sigma^2_w} = \frac{\gamma}{N_F + 1} + \frac{1}{N_F + 1} \frac{\mu}{N_L + 1}, \quad \pi_S^{\sigma^2_w} = \frac{1}{N_F + 1} \frac{\mu}{N_L + 1}.
\]

(19)

*The ratio of fast to slow profits therefore satisfies* \( \frac{\pi_F}{\pi_S} = 1 + \frac{(N_L + 1)^2(1 + b)}{N_F + 1} \), *which implies that when* \( N_F \) *is large,* \( \frac{\pi_S}{\pi_F} \approx \frac{N_F}{(N_F + N_S)^2} \frac{1}{1 + b} \).

Thus, even if there is only one ST (i.e., \( N_S = 1 \)), the ST profits are small compared to the FT profits. The reason is that FTs trade also on their lagged signals, and thus compete with the STs.\(^{28}\) Indeed, FTs compete for trading on \( dw_t \) only among themselves, while they also compete with the STs for trading on the lagged signal \( \tilde{dw}_{t-1} \).

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

**Result 1.** *Consider the alternative setup with* \( N_F \) *fast speculators and* \( N_S \) *slow speculators, 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* \( \delta = \frac{1}{N_F + 1} \frac{\mu}{N_L + 1} \sigma^2_w \). *Then, the alternative model is equivalent to the benchmark model* \( M_1 \).

We now consider the *general benchmark model* \( M_{0,1} \), in which the fast traders use only the current signal, while the slow traders use only the lagged signal.\(^{29}\) The strategies of the FTs and STs are, respectively, of the form \( dx^F_t = \gamma_t dw_t \) and \( dx^S_t = \mu_t \tilde{dw}_{t-1} \), where \( \tilde{dw}_{t-1} = dw_{t-1} - \rho_t dy_{t-1} \). 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.

\(^{28}\)If instead we FTs traded only on their current signal, and only the slow trader used his lagged signal, then the formula (19) would still be correct if we set \( N_L = 1 \). In that case, the profit ratio \( \pi^F/\pi^S = 1 + 4(1 + b)/(N_F + 1) \) is still larger than one.

\(^{29}\)As in Result 1, \( M_{0,1} \) is equivalent to an alternative setup with information processing costs, in which (i) the STs pay the cost \( \delta \) from Result 1, while (ii) the FTs pay a cost slightly higher than \( \delta \). Indeed, if a FT paid \( \delta \), 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.
The next result shows that the model $M_1$ with $N_F$ fast traders and $N_S$ slow traders produces essentially the same outcome as the benchmark model $M_{0,1}$ with $N_F$ fast traders and $N_L = N_F + N_S$ slow traders.

**Corollary 2.** Consider (a) the model $M_1$ with $N_F \geq 1$ fast traders and $N_S \geq 0$ slow traders; and (b) the model $M_{0,1}$ with $N_F$ fast traders and $N_L = N_F + N_S$ slow traders. Then, the equilibrium coefficients $\gamma, \mu, \lambda, \rho$ in the two models are identical.

Because of this equivalence, in the rest of the paper we also call the model $M_{0,1}$ the benchmark model. There are two important particular cases:

- If $N_L \geq N_F$, the benchmark model is equivalent to the model $M_1$ with $N_F$ fast traders and $N_S = N_L - N_F$ slow traders;

- If $N_L = 0$, the benchmark model is the model $M_0$, with 0 lags.

### 3.2 Market Quality

We now study the effect of fast and slow trading on various measures of market quality. Following Corollary 2, we consider the benchmark model in which $N_F \geq 1$ fast traders trade only on the current signal, and $N_L \geq 0$ slow traders trade on the lagged signal. We 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, we first decompose the aggregate speculator order flow into fast trading and slow trading. Denote by $d\tilde{x}_t$ be the aggregate speculator order flow. Let $\tilde{\gamma}$ be the aggregate weight on the current signal $(dw_t)$, and $\tilde{\mu}$ the aggregate weight on the lagged signal $(\tilde{dw}_{t-1})$. We decompose the aggregate speculator order flow $d\tilde{x}_t$ into two components: *fast trading*, which represents the aggregate trading on the current signal; and *slow trading*, which represents the aggregate trading on the lagged signal:

$$
d\tilde{x}_t = \tilde{\gamma} dw_t + \tilde{\mu} \tilde{dw}_{t-1}, \text{ with } \tilde{\gamma} = N_F \gamma, \tilde{\mu} = N_L \mu. \quad (20)
$$

As in Theorem 1, we define $b = \rho \tilde{\mu}$.

We call $b$ 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 $M_0$ with 0 lags from Section 2. In that case $N_F$ fast traders use only their current signal.

We now define the measures of market quality. Recall that the dealer sets a price that changes in proportion to the total order flow $d y = d x_t + d u_t$:

$$d p_t = \lambda d y_t = \lambda \left( \tilde{\gamma} d w_t + \tilde{\mu} \tilde{d} w_{t-1} + d u_t \right), \quad (21)$$

First, as it is standard in the literature, we define *illiquidity* to be the price impact coefficient $\lambda$. Thus, the market is considered illiquid if the price impact of a unit of trade is large, i.e., if the coefficient $\lambda$ is large.

Second, we define *trading volume* as the infinitesimal variance of the aggregate order flow $d y_t$, that is, $TV = \sigma_y^2 = \frac{\text{Var}(d y_t)}{d t}$. We argue that this is a measure of trading volume. Indeed, in each trading round the actual aggregate order flow is given by $d y_t$. Thus, one can interpret trading volume as the absolute value of the order flow: $|d y_t|$. From the theory of normal variables, the average trading volume is given by $E(|d y_t|) = \sqrt{\frac{2}{\pi}} \sigma_y$. With our definition $TV = \sigma_y^2$, we observe that $TV$ is monotonic in $E(|d y_t|)$, and thus $TV$ can be used a measure of trading volume. Using (21), we compute the trading volume in our model by the formula

$$TV = \gamma^2 \sigma_w^2 + \mu^2 \sigma_{\tilde{w}}^2 + \sigma_u^2, \quad \text{with} \quad \sigma_w^2 = \frac{\text{Var}(\tilde{d} w_t)}{d t}. \quad (22)$$

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

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

$$\sigma_p^2 = \lambda^2 TV = \lambda^2 \left( \gamma^2 \sigma_w^2 + \mu^2 \sigma_{\tilde{w}}^2 + \sigma_u^2 \right). \quad (23)$$

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

Fifth, the \textit{speculator participation rate} is defined as the ratio of speculator trading volume over total trading volume, that is,

$$\text{SPR} = \frac{TV^s}{TV} = \frac{\gamma^2 \sigma^2_w + \hat{\mu}^2 \sigma^2_{\tilde{w}}}{\gamma^2 \sigma^2_w + \hat{\mu}^2 \sigma^2_{\tilde{w}} + \sigma^2_u}. \quad (24)$$

$\text{SPR}$ can also be interpreted as the fraction of price variance due to the speculators.

\textbf{Figure 2: Market Quality with Fast and Slow Traders.} This figure plots the following measures of market quality: (i) illiquidity $\lambda$; (ii) trading volume $TV$; (iii) price volatility $\sigma_p$; and (iv) speculator participation rate $\text{SPR}$. Panel A plots the dependence of the four market quality measures on the number of fast traders $N_F$, while taking the number of slow traders $N_L = 5$. Panel B plots the dependence of the four market quality measures on $N_L$, while taking $N_F = 5$. The other parameters are $\sigma_w = 1$, $\sigma_u = 1$.

We now give explicit formulas for our measures of market quality. As before, we use asymptotic notation when $N_F$ is large: $X \approx Y$ stands for $\lim_{N_F \to \infty} \frac{X}{Y} = 1$. 

23
**Proposition 2.** In the benchmark model with \( N_F \geq 1 \) fast traders and \( N_L \geq 0 \) slow traders, the price impact coefficient, trading volume, price volatility, and speculator participation rate satisfy:

\[
\lambda = \frac{\sigma_w}{\sigma_u} \sqrt{\frac{1 + b(a - b^2)}{N_F + 1}} \frac{N_F}{N_F - b}, \quad TV = \sigma_u^2(N_F + 1) \frac{a}{(1 + b)(a - b^2)}, \\
\sigma_p^2 = \frac{\sigma_w^2 N_F}{(N_F + 1)(N_F - b)}, \quad SPR = a + \frac{b^2(1 + b)}{N_F - b},
\]

(25)

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

Panel A of Figure 2 shows how the four measures of market quality vary with the number of fast traders \( N_F \), while holding the number of slow traders \( N_L \) constant. Panel B of Figure 2 shows how the four measures of market quality vary with \( N_L \), while holding \( N_F \) constant. We find that all four market quality measures vary in the same direction with respect to \( N_F \) and \( N_L \). Nevertheless, the number of fast traders has a much stronger effect on these measures than the number of slow traders.

To get more intuition about the effect of fast trading on market quality, we 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 (\( \bar{\mu} = 0 \)), hence the slow trading coefficient \( b \) is zero. Moreover, \( a = \frac{N_F - b}{N_F + 1} = \frac{N_F}{N_F + 1} \). Thus, we can solve the model \( \mathcal{M}_0 \) by simply using Proposition 2. Nevertheless, it is instructive to solve for the equilibrium of \( \mathcal{M}_0 \) independently.

**Proposition 3.** Consider the model \( \mathcal{M}_0 \), with \( N_F \) fast traders whose trading strategy is of the form \( dx_t = \gamma_t dw_t \). Then, the optimal coefficient \( \gamma \) 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,

\[
\lambda = \frac{\sigma_w}{\sigma_u N_F + 1}, \quad TV = \sigma_u^2(N_F + 1), \quad \sigma_p^2 = \frac{\sigma_w^2 N_F}{N_F + 1}, \quad SPR = \frac{N_F}{N_F + 1}.
\]

(26)

Using Proposition 3, we now discuss in more detail the effect of the number \( N_F \) of fast traders on the measures of market quality. First, we note by quickly inspecting the formulas in Proposition 3, that we obtain the same qualitative results as those displayed
in Figure 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 in our model the speculator participation rate can be made arbitrarily close to 1 if the number of fast traders is large. Thus, 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, Cao, and Willard (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 fast traders is large. Indeed, in the proof of Proposition 3 we show that the rate of change of the forecast error variance \( \Sigma' \) is constant and equal to \( \frac{\sigma^2 w}{N_F + 1} \). Since by assumption there is no initial informational asymmetry (\( \Sigma_0 = 0 \)), it follows that \( \Sigma_t \leq \frac{\sigma^2 w}{N_F + 1} \) for all \( t \). In other words, the price stays close to the fundamental value at all times. Thus, a larger number \( N_F \) of fast traders, rather than destabilizing the market, makes the market more efficient.

The trading volume \( TV \) strongly increases with the number of fast traders. This occurs because of the competition among FTs make them trade more aggressively. By trading more aggressively, FTs reveal more information, which as we see later lowers the traders’ price impact. This has an amplifier effect on the trading aggressiveness, such that the trading volume grows essentially linearly in the number of speculators (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 fast traders make the market more liquid, as more information is revealed when there are more competing speculators. This appears to be in contradiction with the fact that more informed trading should increase the amount of adverse selection. To understand the source of this apparent contradiction, note that illiquidity is measured by the price impact \( \lambda \) of one unit of volume. But, while the trading volume \( TV \) strongly increases in \( N_F \) in an unbounded way, its price impact is bounded by magnitude of the signal \( dw_t \).\(^{30}\) Thus, the price impact per unit of volume actually decreases, indicating that prices are more informative. This makes the market

\(^{30}\)In Internet Appendix Section 1, we make this intuition rigorous in the general case; see the discussion surrounding Proposition IA.4.
overall more liquid. This result is consistent with the empirical studies of Zhang (2010), Hendershott, Jones, and Menkveld (2011), and Boehmer, Fong, and Wu (2015).

To understand the effect of fast traders on the price volatility $\sigma_p$, consider the pricing formula $\Delta y_t = \lambda \Delta y_t$, which implies $\sigma_p^2 = \lambda^2 TV$. There are two effects of $N_F$ on the price volatility $\sigma_p$. First, the trading volume $TV$ strongly increases in $N_F$, which has a positive effect on $\sigma_p$. Second, price impact $\lambda$ decreases in $N_F$, which has a negative effect on $\sigma_p$. The first effect is slightly stronger than the second, hence the net effect is that price volatility $\sigma_p$ increases in $N_F$. This result is consistent with the empirical studies of Boehmer, Fong, and Wu (2015) and Zhang (2010).

A few caveats are in order. First, all these studies analyze the effects of HFT activity, where activity is proxied either by turnover or by intensity of order-related message traffic, and not by the number of HFTs present in the market. An answer to this concern is that, as we have seen, trading volume does increase in $N_F$. Second, in our paper we 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 (2012) document a negative effect of HFTs on volatility, possibly because they also consider passive HFTs, which by providing liquidity have a stabilizing effect on price volatility. Moreover, Chaboud, Chiquoine, Hjalmarsson, and Vega (2014) find essentially no relation. In our 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

We start by analyzing the autocorrelation of the components of the order flow. Since the dealer is competitive and risk neutral, the total order flow $\Delta y_t$ has zero autocorrelation. But because the dealer cannot identify the part of the order flow that comes from speculators, the speculator order flow can in principle be autocorrelated.

As in Section 3.2, the aggregate speculator order flow decomposes into its fast trading...
and slow trading components:
\[
\text{d} \bar{x}_t = \underbrace{\text{d} \bar{x}^F_t}_{\text{Fast Trading}} + \underbrace{\text{d} \bar{x}^S_t}_{\text{Slow Trading}}, \quad \text{with} \quad \text{d} \bar{x}^F_t = \bar{\gamma} \text{d} w_t, \quad \text{d} \bar{x}^S_t = \bar{\mu} \tilde{\text{d}} w_{t-1}, \tag{27}
\]
with \( \bar{\gamma} = N_F \gamma \) and \( \bar{\mu} = N_L \mu \). As before, we say that slow trading exists if \( b = \rho \bar{\mu} > 0 \), or equivalently \( N_L > 0 \).

We define *speculator order flow autocorrelation* by \( \text{Corr}(\text{d} \bar{x}_t, \text{d} \bar{x}_{t+1}) \). Because \( \text{d} \bar{x}^F_{t+1} \) is orthogonal to both components of \( \text{d} \bar{x}^F_t \), we obtain the decomposition:
\[
\rho_{\bar{x}} = \text{Corr}(\text{d} \bar{x}_t, \text{d} \bar{x}_{t+1}) = \frac{\text{Cov}(\text{d} \bar{x}^F_t, \text{d} \bar{x}^S_{t+1})}{\text{Var}(\text{d} \bar{x}_t)} + \frac{\text{Cov}(\text{d} \bar{x}^S_t, \text{d} \bar{x}^S_{t+1})}{\text{Var}(\text{d} \bar{x}_t)}. \tag{28}
\]

We denote the *anticipatory trading* part by \( \rho_{\text{AT}} \) and the *expectation adjustment* part by \( \rho_{\text{EA}} \). 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 \( t + 1 \) \( (\bar{\mu} \tilde{\text{d}} w_t) \). 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, we see below that the slow order flow is negatively autocorrelated. Formally, slow trading at \( t + 1 \) \( (\bar{\mu} \tilde{\text{d}} w_t) \) is proportional to the lagged signal minus dealer’s expectation, \( \tilde{\text{d}} w_t = \text{d} w_t - \rho \text{d} y_t \). But the dealer’s expectation is proportional on the total order flow at \( t \), which includes the previous slow trading \( (\text{d} y_t = \bar{\gamma} \text{d} w_t + \bar{\mu} \tilde{\text{d}} w_{t-1} + \text{d} u_t) \). We compute:
\[
\rho_{\bar{x}} = \rho_{\text{AT}} + \rho_{\text{EA}}, \quad \text{with} \quad \rho_{\text{AT}} = \frac{\bar{\mu} \bar{\gamma}}{\text{Var}(\text{d} \bar{x}_t)}, \quad \rho_{\text{EA}} = -\frac{\bar{\mu}^2}{\text{Var}(\text{d} \bar{x}_t)}. \tag{29}
\]

**Proposition 4.** Consider the benchmark model with \( N_F \geq 1 \) fast traders and \( N_L \geq 0 \) slow traders. Then, the speculator order flow autocorrelation and its components satisfy
\[
\rho_{\bar{x}} = \frac{b(b+1)(a-b^2)}{a^2 + b^2(1-a)} \frac{1}{N_F + 1}, \quad \rho_{\text{AT}} = \frac{a}{a-b^2}, \quad \rho_{\text{EA}} = -\frac{b^2}{a-b^2}, \tag{30}
\]
where \( a \) and \( b \) are as in Proposition 2. Moreover, \( \rho_{\bar{x}} \) is strictly positive if and only if there exists slow trading, i.e., \( N_L > 0 \).
Figure 3: Speculator Order Flow Autocorrelation. This figure plots the speculator order flow autocorrelation $\rho_x$ (solid line) and the anticipatory trading component $\rho_{AT}$ (dashed line) as a function of the number $N_F$ of fast traders. The four graphs correspond to four values of the number $N_L$ of speculators using their lagged signal: $N_L = 1, 3, 5, 20$.

One implication of Proposition 4 is that, as long as there exists slow trading, the speculator order flow autocorrelation $\rho_x$ is nonzero. To understand why, note that both the anticipatory trading component and the expectation adjustment component depend on the existence 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.

Figure 3 shows how the speculator order flow autocorrelation ($\rho_x$) and its anticipatory trading component ($\rho_{AT}$) depend on the number of fast traders ($N_F$) for four different values of the number of slow traders ($N_L = 1, 3, 5, 20$). We see that both $\rho_x$ and $\rho_{AT}$ are decreasing in $N_F$. Indeed, when the number of fast traders is large, there is only $\frac{1}{N_F+1}$ of the signal left in the next period for the slow traders. 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 $N_L = 5$, we see that the speculator order flow autocorrelation is 22.56% when there is one FT, but decreases to 2.84% when there are 20 FTs. Our 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.

The anticipatory trading component $\rho_{AT}$ is increasing in the number of slow traders $N_L$ (to see this, fix for instance $N_F = 10$ in each of the four graphs in Figure 3). The intuition is simple: when the number of slow traders is larger, fast trading in each period can better predict the slow trading the next period, hence the correlation $\rho_{AT}$ is larger. Using Nasdaq data on high-frequency traders, Hirschey (2017) finds that HFT order
flow anticipates non-HFT order flow. But Nasdaq defines HFTs along several criteria including the use of large trading volume and low inventories. In our model, these are indeed the characteristics of fast traders, but not those of slow traders (see the next section for a discussion about traders’ inventories). Thus, if in our model we classified fast traders as HFTs and slow traders as non-HFTs, our previous results would imply that HFT order flow anticipates non-HFT order flow.

4 Inventory Management

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

$$
U = \mathbb{E}\left(\int_0^T (v_T - p_t)dx_t\right) - C_I \mathbb{E}\left(\int_0^T x_t^2 dt\right),
$$

(31)

where $T = 1$, and $C_I > 0$ is a constant called the trader’s *inventory aversion coefficient*. We do not identify the exact source of inventory costs for this type of traders, but the costs can be thought to arise either from capital constraints or from risk aversion.\(^{31}\)

We call the resulting setup the *model with inventory management*. To get some intuition for this model, we 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 of this problem is provided in closed form. Then, we continue with a general equilibrium analysis and we see that the equilibrium remains qualitatively the same. We study the properties of the general equilibrium, and the effect of the inventory management on market quality.

---

\(^{31}\)Like inventory aversion, risk aversion generates a quadratic penalty on inventory, but it generates other terms as well; see Hendershott and Menkveld (2014). Therefore, solving the model with risk averse traders would be considerably more difficult.
4.1 Setup

We consider a setup as in the benchmark model with $N_F + 1$ fast traders (who trade only on their current signal) and $N_L$ slow traders, but we replace one risk neutral fast trader with an IFT with utility as in (31). Thus, there are $N_F$ fast traders, $N_L$ slow traders, and one IFT.

To simplify the presentation, we assume directly that the speculators’ strategies have constant coefficients, and that the dealer has pricing rules as in the benchmark model. Thus, fast trader $i = 1, \ldots, N_F$ has trading strategy of the form $dx^F_{i,t} = \gamma_i dw_t$, and slow trader $j = 1, \ldots, N_L$ has trading strategy of the form $dx^S_{j,t} = \mu_j \tilde{dw}_{t-1}$. The coefficient $\lambda$ is chosen so that the dealer breaks even, meaning that her expected profit is zero.

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

$$dx_t = -\Theta x_{t-1} + G dw_t, \tag{32}$$

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

If $\Theta > 0$, in each trading round the IFT removes a fraction $\Theta$ 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 $dt$) that the process needs to halve the distance
from its mean, i.e.,

\[ \text{Inventory Half Life} = \frac{\ln(1/2)}{\ln(\phi)} \, dt = \frac{\ln(1/2)}{\ln(1 - \Theta)} \, dt. \quad (33) \]

Hence, the inventory half life is of the order of \( dt \). This in practice can be short (minutes, seconds, milliseconds), which means that when \( \Theta > 0 \) the IFT does very quick, “real-time” inventory management.

We end this section with a brief discussion of the different types of inventory management. In Section 4.3 we see that there is a discontinuity between the cases \( \Theta = 0 \) and \( \Theta > 0 \). For this reason, we introduce a new case in which \( \Theta \) is infinitesimal and of the form \( \Theta = \theta dt \), with \( \theta \in (0, \infty) \). It turns out that this intermediate inventory management regime indeed connects continuously the other two. Thus, there are three different cases (regimes):

- \( \Theta = 0 \), the \textit{neutral regime}: IFT’s strategy is of the form \( dx_t = Gdw_t \), similar to the strategy of a (risk neutral) fast trader.

- \( \Theta > 0 \), the \textit{quick regime}: IFT’s strategy is of the form \( dx_t = -\Theta x_{t-1} + Gdw_t \). In this regime, the inventory half life is of the order of \( dt \).

- \( \Theta = \theta dt \), the \textit{smooth regime}: IFT’s strategy is of the form \( dx_t = -\theta x_{t-1} \, dt + Gdw_t \), with \( \theta \in (0, \infty) \).\(^{36}\) In this regime, the inventory half life \( \frac{\ln(1/2)}{\ln(1 - \theta dt)} \, dt = \frac{\ln(2)}{\theta} \), which is much larger than the inventory half life in the quick regime.

The smooth regime is discussed in detail in Internet Appendix Section 6. We find that indeed the smooth regime connects continuously 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.\(^{37}\) In general, we show that the smooth regime is never optimal for the IFT, and therefore we can focus only on the comparison between the neutral and the quick regimes.

\(^{36}\)This is called an Ornstein-Uhlenbeck process.

\(^{37}\)See also Figure 5 and the discussion after Theorem 2 in the next section.
4.2 Zero Inventories

When the IFT follows the quick inventory management, his inventory follows an autoregressive process: \( x_t = \phi x_{t-1} + Gdw_t \) with coefficient \( \phi \in (-1, 1) \). Thus, the variance of his inventory is \( \text{Var}(x_t) = \text{Var}(dw_t) / (1 - \phi^2) \). But the variance of the increment \( dw_t \) is equal to \( \sigma_w^2 dt \), hence it is infinitesimal, and therefore so is the inventory \( x_t \).\(^{38}\) Thus, in the continuous time limit the inventory is essentially zero at all times. This fact can be seen also from the formula (33), which shows that the inventory half life in the quick regime is a multiple of the infinitesimal time increment \( dt \).

In general, the expected profit of any speculator satisfies:

\[
\pi = E \int_0^T (v_T - p_t) dx_t = E \left( v_T (x_T - x_0) \right) + E \int_0^T (-p_t) dx_t. \tag{34}
\]

The \textit{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 \textit{cash component} is the expected profit that comes from changes in the cash account due to trading.

The next result 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 \( dx_t \) 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 his expected profit is

\[
\pi_c = E \int_0^T x_{t-1} dp_t. \tag{35}
\]

Thus, whenever inventory management results in zero inventories for the speculator, his trading strategy can be profitable only when his inventory \( (x_{t-1}) \) forecasts the subsequent change in price \( (dp_t) \). 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.

\(^{38}\)See also equation (A25) in the Appendix, where we show that \( E(x_t^2) \) is of the order of \( dt \).
We can 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 exists slower trading.

In the particular case of the IFT, we have already seen above that his inventory is zero at all times, so we can apply Proposition 5. We see 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 $Gd_{t-1}$, which is positively correlated with the aggregate order flow at $t$ via the orders of the ST, $\mu_j \tilde{d}_{t-1}$. Hence, it is possible for the IFT to make positive profits while keeping zero inventory at all times. In Section 4.3 we provide conditions under which the IFT actually behaves in this manner, and we learn in more detail the mechanism by which the IFT extracts profits from slower trading.

4.3 IFT and Inventory Management

Consider the inventory management model with $N_F$ fast traders, $N_L$ slow traders, and one IFT. In this section, we solve for the optimal strategy of the IFT in a partial equilibrium analysis, keeping the behavior of the other players fixed. This allows us to get insight about IFT’s behavior, without having to do a full equilibrium analysis. We leave this more general analysis to Section 4.4.

We thus fix the coefficients $\gamma$ and $\mu$ that describe the strategies of the FTs and STs, and the coefficients $\lambda$ and $\rho$ that describe the dealer’s pricing rules. Suppose the IFT has a trading strategy as in (32): $dx_t = -\Theta x_{t-1} + Gd_t$, which is not necessarily optimal. The expected profit of the IFT can be written then as $\pi = \mathbb{E} \int_0^T (w_t - p_t)dx_t$. Since $w_t = w_{t-1} + dw_t$ and $p_t = p_{t-1} + \lambda dy_t$, we have the following decomposition:

$$\pi = G \mathbb{E} \int_0^T (dw_t - \lambda dy_t)dw_t - \Theta \mathbb{E} \int_0^T (w_{t-1} - p_{t-1})x_{t-1}dt + \Theta \mathbb{E} \int_0^T x_{t-1}dp_t.$$  

(36)

The first term, denoted by $\pi_0$, is IFT’s expected profit when $\Theta = 0$, which reflects the profits that result from exploiting his signals ($dw_t$). The second term, denoted by $\ell_r$, is the informational loss that comes from inventory mean reversion: indeed, by
reducing inventory by $\Theta x_{t-1}$ each period, there is an expected loss coming from the correlation of $x_{t-1}$ with the remaining informational advantage $w_{t-1} - p_{t-1}$. Differently put, by managing inventory the IFT trades against his previous signals. The third term, denoted by $\pi_a$, is the profit that comes from anticipation of slow trading: at time $t$ the IFT reduces his inventory by $\Theta x_{t-1}$ exactly when the slow traders submit a market order in the opposition direction (which is part of the current aggregate order flow $d_y_t$). Note that the third term is equal to $\Theta \pi_c$, where $\pi_c$ is the expected profit of a speculator who keeps all his profits in cash: see equation (35).

When the IFT mean reverts his inventory ($\Theta > 0$), we have seen in Section 4.2 that the inventory of the IFT is zero, and from equation (35) 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 IFT’s trading on his signals. To understand the intuition for this result, suppose the IFT observes a new signal $d w_t$. Initially, the IFT trades on his signal ($G d w_t$), 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 ($\Theta x_{t-1}$), 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 $\pi_a$.) But this is only possible if there exist slow traders, whose lagged signals can be predicted by the IFT.

We formalize this last result by providing an explicit formula for IFT’s expected profit. For that, define additional model coefficients by:

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

(37)

**Proposition 6.** Let $d x_t = -\Theta x_{t-1} + G d w_t$ be 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 $\pi$ satisfies:

$$
\pi = \lambda \left( \bar{\mu} G \frac{1 - a^-}{1 + \phi b} - G^2 \frac{b + \frac{1}{1 + \phi}}{1 + \phi b} \right) \sigma^2_w.
$$

(38)
As a result of keeping all his profits in cash, the behavior of the IFT is very different than the behavior of risk neutral speculators such as 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 ($\bar{\mu} = 0$), the IFT makes negative expected profits.

Equation (38) explains also how IFT’s profit depends on the model coefficients $a^- = N_F \rho \gamma$ and $b = N_L \rho \mu$, which respectively measures the amount of fast trading and slow trading. When there is more fast trading ($a^-$ is higher), IFT’s profit is smaller because of increased competition from fast traders. 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 slow traders, but also a larger cost ($\sigma_w^2 \lambda G^2 \frac{b+1/\left(1+\phi\right)}{1+\phi b}$). This cost arises from the fact that when the IFT at $t$ provides liquidity to the slow traders, these do not trade on the lagged signal ($d_{w_t} - 1$) but rather on its unanticipated part ($\tilde{d}_{w_t} - 1 = d_{w_t} - 1 - \rho d_{y_{t-1}}$), which reduces IFT’s profits.  

We now describe the optimal strategy of the IFT. Recall that beside the expected profit, IFT’s utility also includes a penalty cost that is quadratic in the inventory: $C_I E \left( \int_0^T x_t^2 dt \right)$. 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 $C_I$. The next result describes IFT’s optimal strategy when the slow trading coefficient $b$ is above a threshold: $b > \sqrt{17} - 1 \approx 0.3904$. This condition is true if for instance there are $N_F \geq 1$ fast traders and $N_L \geq 2$ slow traders.  

**Theorem 2.** In the inventory management model, suppose the model coefficients satisfy $0 \leq a^- , b < 1$ and $\lambda , \rho > 0$. In addition, suppose $b > \frac{\sqrt{17} - 1}{8} \approx 0.3904$. Let $\bar{C}_I = \frac{\sqrt{17} - 1}{8}$. It turns out that, compared to the benefit, the cost is more strongly increasing in $b$, and hence, as we can see in equation (40), the optimal $G$ is actually decreasing in $b$.  

If instead $b < 0.3904$, we show Internet Appendix Section 5.1 that a similar analysis holds. The IFT still manages inventory but the optimal $\Theta$ is at its lowest possible value, which we denote by $0_+$. This value is the same as $\theta = \infty$ in the smooth regime.  

In equilibrium (section 4.4) we have 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 $N_F = 0$.  

39It turns out that, compared to the benefit, the cost is more strongly increasing in $b$, and hence, as we can see in equation (40), the optimal $G$ is actually decreasing in $b$.  

40If instead $b < 0.3904$, we show Internet Appendix Section 5.1 that a similar analysis holds. The IFT still manages inventory but the optimal $\Theta$ is at its lowest possible value, which we denote by $0_+$. This value is the same as $\theta = \infty$ in the smooth regime.  

41In equilibrium (section 4.4) we have 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 $N_F = 0$.  

35
Figure 4: Optimal IFT Inventory Management. This figure plots the coefficients of IFT’s optimal trading strategy \( dx_t = -\Theta x_{t-1} + G d w_t \) in the inventory management model with \( N_F = 5 \) fast traders and \( N_L = 5 \) slow traders. On the horizontal axis is IFT’s inventory aversion, \( C_I \). The parameter values are \( \sigma_w = 1, \sigma_u = 1 \). For the model coefficients, we use the equilibrium values from Section 4.4: \( a^- = 0.7088, b = 0.5424, \lambda = 0.3782, \rho = 0.3439 \). The formulas for \( G, \Theta, \) and \( \bar{C}_I \) are from Theorem 2.

\[
2\lambda \left( \frac{(1-Ra^-)^2(1+\sqrt{1-b})^2}{R^b(1-a^-)^2} - 1 \right). \quad \text{Then, if } C_I < \bar{C}_I, \text{ the optimal strategy of the IFT is to set} \\
\Theta = 0, \quad G = \frac{1 - Ra^-}{2\lambda + \bar{C}_I}. \tag{39}
\]

\[
\text{If } C_I > \bar{C}_I, \text{ the optimal strategy of the IFT is to set} \\
\Theta = 2 - \frac{\sqrt{1-b}}{b} \in (0,2), \quad G = \frac{1 - a^-}{2\rho \left( 1 + \frac{1}{\sqrt{1-b}} \right)}. \tag{40}
\]

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 (\( \bar{C}_I \)).

- **Neutral regime** If the inventory aversion coefficient is small (below \( \bar{C}_I \)), 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.
• \textit{(Quick regime)} If the inventory aversion is large (above $\bar{C}_I$), 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 $\Theta$ to maximize expected profit (more details below).

Thus, a small change in IFT’s inventory aversion can have a large effect on IFT’s behavior. Figure 4 plots the coefficients of the optimal strategy when there are $N_F = 5$ fast traders and $N_S = 5$ slow traders. We see that when IFT’s inventory aversion rises above the threshold $\bar{C}_I = 0.1021$, his optimal mean reversion coefficient jumps from $\Theta = 0$ to $\Theta = 0.7530$. Also, 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 IFT’s optimal trading strategy, we now discuss the smooth regime in connection with the neutral and the quick regimes. Recall that IFT’s trading strategy is of the form $dx_t = -\Theta x_{t-1} + Gdw_t$, where either (i) $\Theta = 0$ (neutral regime), (ii) $\Theta = \theta dt$ (smooth regime), or (iii) $\Theta > 0$ (quick regime). One then verifies that IFT’s expected utility, along with its components described above, varies continuously across the three regimes. More formally, if we denote by $U(\Theta)$ IFT’s expected utility in either of the three regimes, in Internet Appendix Section 6 we show that $\lim_{\theta \to 0} U(\theta dt) = U(0)$ and $\lim_{\theta \to \infty} U(\theta dt) = \lim_{\Theta \to 0} U(\Theta)$. Thus, the smooth regime indeed connects continuously the neutral regime with the quick regime.

Figure 5 plots the expected utility $U$ as a function of $\Theta$ across the smooth and quick regimes. To simplify the presentation, instead of considering $U = U(\Theta, G)$ as a function of both $\Theta$ and $G$, we only consider the value of $G$ that maximizes $U$ given $\Theta$.\footnote{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 Internet Appendix Section 6 implies that $\frac{U}{\sigma_w^2} = G \left( (1 - Ra^-) - F_\theta (1 - R a^{-b} \frac{\lambda}{1+b}) \right) - \frac{G^2}{2} \left( 2\lambda (1 - \frac{F_\theta^2}{1+b}) + F_\theta \frac{\lambda}{1+b} + \frac{C_I}{\theta} \right)$, where $F_\theta = 1 - \frac{1-e^{-\theta}}{\theta}$.
}

Formally, if $U(\Theta, G)$ indicates the dependence of $U$ on both $\Theta$ and $G$, in Figure 5 we plot $U = \max_G U(\Theta, G)$, relative to the value $\pi_0 = \max_G \pi_0(G)$. Figure 5 shows that IFT’s utility indeed changes continuously from the smooth regime to the quick regime. Moreover, when we let the inventory aversion coefficient $C_I$ vary, we see that there are two cases:
**Figure 5: IFT Inventory Management and Utility.** This figure plots the maximum normalized expected utility of the IFT for a fixed mean reversion rate, in the inventory management model with $N_F = 5$ fast traders and $N_L = 5$ slow traders. On the horizontal axis is IFT’s mean reversion rate given by (i) $\theta$ from IFT’s trading strategy, $dx_t = -\theta x_{t-1} dt + G dw_t$ (the smooth regime), or (ii) $\Theta$ from IFT’s trading strategy, $dx_t = -\Theta x_{t-1} + G dw_t$ (the quick regime). On the vertical axis is IFT’s maximum expected utility $U$ when $G$ varies and $\Theta$ (or $\theta$) is fixed, normalized by the maximum expected profit $\pi_0$ when $G$ varies and $\Theta = 0$ (the neutral regime). The other parameter values are $\sigma_w = 1, \sigma_u = 1$. For the model coefficients, we use the equilibrium values $a^- = 0.7088, b = 0.5424, \lambda = 0.3782, \rho = 0.3439$.

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

This result is indeed as stated in Theorem 2: when the inventory aversion $C_I$ crosses the threshold $C_I = 1.1021$, the optimal $\Theta$ jumps discontinuously from 0 to 0.7530. As observed in the Figure, the reason for this discontinuity is that in the smooth regime the optimum $\theta$ is either zero or infinity, but never in between.

To understand the intuition behind this last fact, consider again the we need to describe in more detail how IFT’s utility changes with $\Theta$. By definition, this utility is
equal to the expected profit minus the quadratic penalty on inventory. From (36),

\[ U = \pi_0 - E \int_0^T (w_{t-1} - p_{t-1})x_{t-1} \Theta + \Theta E \int_0^T x_{t-1} dp_t - C_I E \int_0^T x_{t-1}^2 dt. \]  

The term \( \pi_0 \) does not depend on \( \Theta \), and is the same for the smooth and quick regimes. The loss \( \ell_r \) 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 did not get yet incorporated into the price: \( w_t - p_t \)). The loss \( \ell_r \) is increasing in \( \Theta \) (or \( \theta \)): the more the IFT reverts his inventory, the larger the corresponding informational loss. The third term, \( \pi_a \), 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 slow traders only when he reverts a large enough part of his inventory, that is, when \( \Theta \) is not infinitesimal (the quick regime). The fourth term, the inventory penalty \( \ell_i \), is positive in the smooth regime, but starts decreasing fast in \( \theta \) when this coefficient is sufficiently large, and it approaches zero in the limit. Thus, in the quick regime \( \ell_i \) is zero, as IFT’s inventory is zero at all times.

We now explain why IFT’s maximum utility \( U \) in the smooth regime only occurs either at \( \theta = 0 \) or at \( \theta = \infty \) (see Figure 5). Initially, when the mean reversion coefficient \( \theta \) is small, an increase in \( \theta \) raises the informational loss \( \ell_r \) from trading against his own signals, while the associated reduction in inventory does not diminish much the penalty \( \ell_i \) (which is quadratic in inventory). But, when \( \theta \) is large, the inventory penalty is reduced more dramatically and contributes to a rise in utility as \( \theta \) approaches infinity. Because of the drop in utility in the middle range of \( \theta \), IFT’s maximum expected utility in the smooth regime can only occur at either the endpoints (0 or \( \infty \)).

We also explain why IFT’s maximum utility \( U \) in the quick regime is realized at an interior \( \Theta \) (equal to 0.7530 in Figure 5). Indeed, as explained above, when \( \Theta \) is in the quick regime, the inventory penalty \( \ell_i \) is zero, hence there are only two nonzero terms that depend on \( \Theta \): the informational loss from mean reversion \( \ell_r \), and the anticipatory gain from slow traders \( \pi_a \). As \( \Theta \) increases, the mean reversion loss \( \ell_r \) increases (it was already positive in the smooth regime), but the anticipatory gain increases as well (it
was zero in the smooth regime). When $\Theta$ is small, the term $\ell_r$ dominates and $U$ is increasing in $\Theta$. When $\Theta$ is large, the term $\pi_a$ dominates and $U$ is decreasing in $\Theta$. As a result, $U$ has an interior optimum in the quick regime.

Using our results, we argue that in practice fast speculators fall into two sharply different categories. (See also Figure 1 in the Introduction.) In both categories speculators generate large trading volume. But in one category 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. Our 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 minutes.

We finish this section with a brief discussion of how 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 (A39) in the Appendix). In the previous discussion, we have argued that this is possible only if the IFT trades in the opposite direction to the slow trading. We now prove this is indeed the case.

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

$$\text{Cov}(dx_t, d\tilde{x}_t^S) = -\Theta \text{Cov}(x_{t-1}, d\tilde{x}_t^S) < 0.$$  \hspace{1cm} (42)

We call this phenomenon the hot potato effect, or the intermediation chain effect. The intuition is that 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. Kirilenko et al. (2017) and Weller (2012) document such hot potato effects among high frequency traders.
4.4 Equilibrium Results

In this section, we solve for the full equilibrium of the inventory management model. For simplicity, we assume that the IFT is sufficiently averse, meaning that his inventory aversion is above a certain threshold (formally, above the threshold value $\bar C_I$ from Theorem 2). Then, 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, $N_F$ fast traders, and $N_L$ slow traders. Suppose there is an equilibrium in which the speculators’s strategies are: $dx_t = -\Theta x_{t-1} + Gdw_t$ (the IFT), $dx_t^F = \gamma dw_t$ (the FTs), $dx_t^S = \mu \tilde dw_{t-1}$ (the STs); and the dealer’s pricing rules are: $dp_t = \lambda dy_t$, $\tilde dw_t = dw_t - \rho dy_t$. Denote the model coefficients $R$, $a^-$, $b$ as in (37). Suppose $\sqrt{17} - 1 < b < 1$. Then, the equilibrium coefficients satisfy equations (A40)–(A42) from the Appendix.

Conversely, suppose that equations (A40)–(A42) have a real solution such that $\sqrt{17} - 1 < b < 1$, $a < 1$, $\lambda > 0$. 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, we start by providing asymptotical results when the number of FTs and STs is large. The advantage is that the asymptotic results can be expressed in closed form, and thus help provide a clearer intuition for the equilibrium. Let $\bar C_I$ be the threshold aversion from Theorem 2. Let $\pi$ be the expected profit of a sufficiently averse IFT ($C_I \geq \bar C_I$), and $\pi^{C_I = 0}$ 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 $\gamma_0$ be the benchmark FT weight, and $\pi^F_0 = \frac{\gamma_0}{N_F + 2} \sigma_w^2$ the benchmark profit of a FT, as in Proposition 1. We use the asymptotic notation: $X \approx X_\infty$ stands for $\lim_{N_F, N_L \to \infty} \frac{X}{X_\infty} = 1$.

**Proposition 7.** Consider (i) the inventory management model with one sufficiently averse IFT, $N_F$ fast traders, and $N_L$ slow traders, and (ii) the benchmark model with $N_F + 1$ fast traders and $N_L$ slow traders. Then, the equilibrium coefficients $\gamma$, $\mu$, $\lambda$, $\rho$ 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 we have the following asymptotic formulas:

\[ \Theta \approx 1, \quad \frac{G}{\gamma_0} \approx 1 - b_\infty = 0.3820, \quad \frac{\pi}{\pi_0} \approx 2b_\infty - 1 = 0.2361, \]

\[ \frac{\pi}{\pi^{C_I=0}} \approx \frac{2}{5} b_\infty = 49.44\%, \quad \bar{C}_I \approx \frac{1+5b_\infty}{2} \lambda_\infty \approx 2.0451 \frac{\sigma_w}{\sigma_u} \frac{1}{\sqrt{N_F + 1}}. \]  

(43)

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, IFT’s inventory mean reversion becomes extreme ($\Theta$ approaches 1). This means that 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 ($\Theta = 1$) the IFT can trade so that he only loses on average only about 50% of his maximum expected profits when he has zero inventory aversion.\(^{43}\)

The equation $\bar{C}_I \approx 2.0451 \frac{\sigma_w}{\sigma_u} \frac{1}{\sqrt{N_F + 1}}$ implies that the threshold inventory aversion above which the IFT chooses to mean revert his inventory becomes very small when the number of competing fast traders 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.

We now compare the IFT with the other speculators. For the IFT, we consider

\(^{43}\)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.”
the following variables: (i) IFT’s trading volume, measured by his order flow variance $TV_x = \text{Var}(dx_t)/dt$, as in Section 3.2, (ii) IFT’s order flow autocorrelation, $\rho_x = \text{Corr}(dx_t, dx_{t+1})$; and (iii) $\beta_{x,\bar{x}S} = \text{Cov}(dx_t, d\bar{x}_S^t)/\text{Var}(d\bar{x}_S^t)$, which is the regression coefficient of IFT’s strategy ($dx_t$) on the slow trading component ($d\bar{x}_S^t$). We are also interested in the individual FT volume, $TV_{xF}$; the aggregate FT volume, $TV_{\bar{x}F}$; and the aggregate ST volume, $TV_{xS}$; the aggregate FT order flow autocorrelation, $\rho_{x\bar{x}}$; and the aggregate ST order flow autocorrelation, $\rho_{\bar{x}S}$.

The next result computes all these quantities, and provides asymptotic results when both $N_F$ and $N_L$ are large. Some of these results provide new testable implications, regarding the relationship between trading volume, order flow covariance, and inventory.

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

$$
\frac{TV_x}{TV_{xF}} = \frac{2G^2}{(1+\phi)\gamma^2} \approx 4 - 6b_\infty = 0.2918, \quad \frac{TV_{\bar{x}S}}{TV_{\bar{x}F}} = \frac{b^2(1-a)}{(a^-)^2} \approx \frac{b_\infty}{N_F + 1},
$$

$$
\rho_x = -\frac{\Theta}{2} \approx -\frac{1}{2}, \quad \rho_{x\bar{x}} = 0, \quad \rho_{\bar{x}S} \approx -b_\infty = -0.6180,
$$

$$
\beta_{x,\bar{x}S} = -\frac{\Theta(1-a^-)}{2b(1+2\sqrt{1-b})} \approx -\frac{3 + b_\infty}{5(N_F + 1)} = -0.7236\frac{N_F}{N_F + 1}.
$$

The last result 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 fast traders, 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, IFT’s trading volume is about 30% of the individual FT trading volume. This implies that 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 our intuition that in an empirical analysis which 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 we compare order flow autocorrelations, we see 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. Finally, if we compare inventories, the IFT has infinitesimal inventory, while the variance of the other speculators’ inventory increases over time. Nevertheless, the STs’ inventories are smaller relative to FTs’ inventories, since the STs have smaller volume.

We now present some numerical results for the equilibrium coefficients. Figure 6 plots the equilibrium coefficients (\(\Theta, G, \gamma, \mu, \lambda, \rho\)). We normalize some variables \(X\) in the inventory management model by the corresponding variable \(X_0\) in the benchmark model. Panel A of Figure 6 plots the variables against \(N_F\), while holding \(N_L\) constant. Panel B plots the same variables against \(N_L\), while taking \(N_F = 5\). The other parameters are \(\sigma_w = 1, \sigma_u = 1\).

As expected, we find that the mean reversion coefficient \(\Theta\) is increasing in the number of slow traders \(N_L\). This is because the IFT needs slow traders in order to make profits.

44 Even if we allowed FTs to trade on lagged signals, one can see that the FTs would still have very small order flow autocorrelation (of the order of \(1/(N_F + 1)\)) because of their large trading volume.

45 For the IFT, \(\text{Var}(x_t) = \frac{G^2}{1-\phi^2} \sigma_w^2 dt\) (see equation (A24) in the Appendix); while for the FT, \(\text{Var}(x_t) = t\sigma_w^2\), as FT’s inventory follows a random walk.

46 We consider \(N_F, N_L \geq 2\). The reason is that in order to apply Theorem 2, we need to have \(b > \frac{\sqrt{17} - 1}{8}\). This is true in equilibrium if \(N_F, N_L \geq 2\).
The IFT’s weight $G$ is less than half the benchmark weight $\gamma_0$, 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 ($\gamma/\gamma_0$ and $\mu/\mu_0$ are both above one), despite the price impact $\lambda$ being larger than in the benchmark (we see that $\lambda/\lambda_0 > 1$). 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 $\gamma_0$), and therefore the informational efficiency is lower. To see directly that the market is less informationally efficient in the inventory management model, we use the fact that in our model price volatility is a proxy for informational efficiency (see the discussion in Section 3.2). Then, we 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 we discuss several model assumptions, and verify that our 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 we justify this by an information processing cost per signal. But by choosing strategies as in (11), we 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.\(^{47}\) Thus, we can add a Kyle term of the form $\beta_t(w_t - p_t)dt$ to the strategy in (11). In that case, we argue that 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 Internet Appendix Section 4 we introduce an orthogonal dimension of the fundamental value, and we show

---

\(^{47}\)Note that we 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 we work in continuous time, the price is an infinite combination of past signals.
that trading strategies that add 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, slow traders do not need to learn at \( t \) about the fast traders’ previous signal \( dw_{t-1} \), because at \( t \) they already learn \( dw_{t-1} \) perfectly. Nevertheless, in the model \( \mathcal{M}_2 \) (with fast, medium, and slow traders), 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 of the form \( \hat{\gamma} dw_{t-1} + du_t \), which is the aggregate order flow \( du_{t-1} \) minus the part \( \mu \tilde{dw}_{t-2} + \nu \tilde{dw}_{t-3} \) already observed by the slowest traders at \( t \).\(^{48}\) We 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 Internet Appendix Section 2.2 we consider an extension of \( \mathcal{M}_2 \) in which the slowest traders can learn from the order flow at no cost. In this extension, we observe that the main results of the benchmark model remain correct. 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 we have not been able to solve such an extension, we believe that our intuition would remain very similar. Indeed, as observed by Back, Cao, and Willard (2000), when signals are not perfectly correlated, initially speculators trade very aggressively on the common part of their signals (the “rat race”). As in our model signals have to be dropped after a short while, 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, Hombert, and Roșu (2016) show that the optimal strategy of a (fast) speculator must be of the form \( dx_t = \beta_t (w_t - p_t) dt + \gamma_0 dw_t + \cdots + \gamma_k \tilde{dw}_{t-k} \). Thus, the only difference between this strategy

\(^{48}\)Here we denote by \( \hat{\gamma}, \mu \) and \( \nu \) the aggregate coefficients on the signals of lag 0, 1 and 2, respectively.
and the strategy in (11) is the Kyle term $\beta_t(w_t - p_t)dt$. But, as we have seen before, this term can be ignored if we believe that speculators want strategies that protect against the price containing information about other components of the fundamental value.

In Internet Appendix Section 3, we consider an extension of the model $\mathcal{M}_1$ in which noisy version of the signals are made public with lag $k = 2$. For this extension, the precision of public news (which is measured by the ratio $\sigma_w/\sigma_v$) becomes a parameter that 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 ($\sigma_w/\sigma_v$ less than 0.8) the equilibrium behavior of speculators 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 our 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 $\lambda$ 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 Internet Appendix 7 we analyze an extension $\mathcal{D}_1$, which is a discrete version of $\mathcal{M}_1$ in which in addition we allow these covariances to depend on speculators’ strategies. Then, by taking the limit of $\mathcal{D}_1$ when the time interval approaches zero, we see 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 our 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 HFTs and LFTs 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 our benchmark model with fast and slow traders, if we assume that larger lags are not used. Intuitively, our 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.\(^{49}\)

We also describe several extensions of the model with inventory management. Our main results are robust to two extensions that allow that: First, in Internet Appendix Section 5.3, we consider a more general IFT strategy that has a component of trading on the lagged signal: \( dx_t = -\Theta x_{t-1} + Gdw_t + M\tilde{dw}_{t-1} \).\(^{50}\) Second, in Internet Appendix 5.4, 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 \( \lambda_t \) is determined by her expected profit being zero, but she correctly sets \( dp_t = \lambda_t \tilde{dy}_t \), where \( \tilde{dy}_t \) is the unanticipated part of the aggregate order flow at \( t \).

The main result of Section 4 is that by quick inventory management the IFT provides liquidity to the slow traders. In Internet Appendix Section 5.5 we show in an extension to multiple IFTs that this result holds true even if all fast traders become IFTs, and the only speculators remaining are slow. In that case, the fast traders (that are only IFTs) no longer speculate on the long term value, but just pass their inventory (the “hot potato”) to the slower traders.

Finally, in Internet Appendix Section 5.6 we consider an extension of the benchmark model \( \mathcal{M}_2 \) that has both an inventory-averse fast trader (IFT) and an inventory-averse medium trader (IMT). Recall that in \( \mathcal{M}_2 \) there are three types of speculator: (i) fast trader (FT) who at \( t \) observes \( dw_t \); (ii) medium trader (MT) who at \( t \) observes \( dw_{t-1} \),

\(^{49}\)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 IFT’s trading is correlated with aggregate trading next period, which can also come from the HFTs when they trade on their lagged signals).

\(^{50}\)In Internet Appendix Section 6.3 we also show that smooth strategies of the form \( dx_t = -\theta x_{t-1} dt + Gdw_t + M\tilde{dw}_{t-1} \) are never optimal when \( \theta \in (0, \infty) \).
and (iii) slow trader (ST) who at $t$ observes $dw_{t-2}$. Thus, in this extension, we consider an IFT with strategy $dx_t = -\Theta x_{t-1} + Gdw_t$, and an IMT with strategy $dz_t = -\Omega z_{t-1} + H\tilde{d}w_{t-1}$. This creates an intermediation chain with two links: the IFT and the IMT. Compared to the situation in which one chain 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

We 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). We have found 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. We have also considered an extension of the model in which a fast speculator, called the inventory-averse fast trader (IFT), has quadratic inventory costs. We 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

Notation Preliminaries

Recall that \( t - k \) is notation for \( t - k dt \), and

\[
T = 1. \quad (A1)
\]

For a process \( X_t \), we denote by \( \sigma_{X,t} \) the instantaneous volatility of \( X_t \), which is the limit

\[
\lim_{\Delta t \to 0} \frac{\text{Var}(\Delta X_t)}{\Delta t}, \quad \text{if this limit exists.}
\]

In general, a tilde above a symbol denotes normalization by \( \sigma_w \) or \( \sigma_w^2 \). For instance, if \( \sigma_u \) is the instantaneous volatility of the noise trader order flow, and \( \sigma_y \) the instantaneous volatility of the total order flow, we denote by

\[
\tilde{\sigma}_u = \frac{\sigma_u}{\sigma_w}, \quad \tilde{\sigma}_y = \frac{\sigma_y}{\sigma_w}. \quad (A2)
\]

If \( dx_t \) is a trading strategy, \( t \in (0, T] \), let \( \tilde{\pi} \) be the normalized expected profit at \( t = 0 \):

\[
\tilde{\pi} = \frac{1}{\sigma_w^2} \mathbb{E} \left( \int_0^T (w_t - p_t) dx_t \right). \quad (A3)
\]

For covariances, a tilde above a symbol means normalization by both \( \sigma_w^2 \) and \( dt \):

\[
\tilde{\text{Var}}(\tilde{dw}_t) = \frac{\text{Var}(\tilde{dw}_t)}{\sigma_w^2 dt} = A_t, \quad \tilde{\text{Cov}}(w_t, \tilde{dw}_t) = \frac{\text{Cov}(w_t, \tilde{dw}_t)}{\sigma_w^2 dt} = B_t. \quad (A4)
\]

**Proof of Theorem 1.** We 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 we take the dealer’s pricing functions as given, and solve for the optimal trading strategies for the FTs and STs. In the second step, we take the speculators’ trading strategies as given, and we compute the dealer’s pricing functions. In Section 2, we have assumed that the speculators take the signal covariance structure as given (see equation (13)). In the current context, this means that the speculators take the following covariances \( A_t \) and \( B_t \) from (A4) and fixed constants.
Thus, in the rest of the Appendix we consider that the dealer also sets $A$ and $B$, in addition to setting $\lambda$ and $\rho$.

**Speculators’ Optimal Strategy** ($\gamma$, $\mu$)

Since we search for an equilibrium with constant coefficients, we assume that the speculators take as given the dealer’s pricing rules $dp_t = \lambda dy_t$ and $z_{t-1,t} = \rho dy_{t-1}$, and also the covariances $A = \text{Var}(\tilde{d}w_t)$ and $B = \text{Cov}(w_t, \tilde{d}w_t)$.

Consider a FT, indexed by $i = 1, \ldots, N_F$. He chooses $dx^i_t = \gamma^i_t dw_t + \mu^i_t \tilde{d}w_{t-1}$, and assumes that at each $t \in (0, T]$, the price satisfies:

$$dp_t = \lambda dy_t, \quad \text{with} \quad dy_t = (\gamma^i_t + \gamma^{-i}_t) dw_t + (\mu^i_t + \mu^{-i}_t) \tilde{d}w_{t-1} + du_t, \quad (A5)$$

where the superscript “$-i$” indicates the aggregate quantity from the other speculators.

Since $dw_t$ and $\tilde{d}w_{t-1}$ are both orthogonal on the public information set $I_t$, and $p_{t-1} \in I_t$, it follows that $dx^i_t$ is orthogonal to $p_{t-1}$ as well. The normalized expected profit of FT $i$ at $\tau \in [0, T)$ satisfies:

$$\tilde{\pi}^F_{\tau} = \frac{1}{\sigma_w^2} E \int^T_\tau \left( w_t - p_{t-1} - \lambda \left( (\gamma^i_t + \gamma^{-i}_t) dw_t + (\mu^i_t + \mu^{-i}_t) \tilde{d}w_{t-1} + du_t \right) \right) dx^i_t$$

$$= \int^T_\tau \left( \gamma^i_t - \lambda \gamma^{-i}_t (\gamma^i_t + \gamma^{-i}_t) + \mu^i_t B - \lambda \mu^i_t (\mu^i_t + \mu^{-i}_t) A \right) dt. \quad (A6)$$

This is a pointwise optimization problem, hence it is enough to consider the profit at $\tau = 0$, and maximize the expression over $\gamma^i_t$ and $\mu^i_t$. The solution of this problem is $\lambda \gamma^i_t = \frac{1 - \lambda \gamma^{-i}_t}{2}$, and $\lambda \mu^i_t = \frac{B/A - \lambda \mu^{-i}_t}{2}$. The ST $j = 1, \ldots, N_S$ solves the same problem, only that his coefficient on $dw_t$ is $\gamma^j_t = 0$. Thus, all $\gamma$’s are equal for the FTs, and all $\mu$’s are equal for the FTs and STs. We also find that they are constant, and since $N_L = N_F + N_S$, we have

$$\gamma = \frac{1}{\lambda} \frac{1}{1 + N_F}, \quad \mu = \frac{B/A}{\lambda} \frac{1}{1 + N_L}. \quad (A7)$$

---

51By the assumption (13), the speculators take the covariance structure as computed by the dealer. By construction, the lagged signal $\tilde{d}w_{t-1}$ is orthogonal to the dealer’s information set at time $t$, which includes the price $p_{t-1}$, hence the covariance $\text{cov}(\tilde{d}w_{t-1}, p_{t-1})$ is set to zero.
Dealer’s Pricing Rules ($\lambda$, $\rho$, $A$, $B$)

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

$$dy_t = du_t + \gamma dw_t + \bar{\mu} \tilde{w}_{t-1}, \quad \text{with} \quad \gamma = N_F \gamma, \quad \bar{\mu} = N_L \mu. \quad (A8)$$

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

$$\tilde{w}_{t-1} = dw_{t-1} - \rho_* dy_{t-1}. \quad (A9)$$

Later we require that in equilibrium the dealer’s pricing coefficient $\rho$ coincides with the coefficient $\rho_*$ used by the speculators.

Since the order flow $dy_t$ is orthogonal to the dealer’s information set $\mathcal{I}_t$, the dealer sets $\lambda_t, \rho_t, A_t, B_t$ such that the following equations are satisfied:

$$\lambda_t = \frac{\text{Cov}(w_t, dy_t)}{\text{Var}(dy_t)} = \frac{\gamma + \bar{\mu} B_{t-1}}{\sigma_{y,t}^2}, \quad dp_t = \lambda_t dy_t,$$

$$\rho_t = \frac{\text{Cov}(dw_t, dy_t)}{\text{Var}(dy_t)} = \frac{\gamma}{\sigma_{y,t}^2}, \quad \tilde{w}_t = dw_t - \rho_t dy_t,$$

$$\sigma_{y,t}^2 = \text{Var}(dy_t^2) = \sigma_u^2 + \gamma^2 + \bar{\mu}^2 A_{t-1},$$

$$B_t = \text{Cov}(w_t, dw_t - \rho_* dy_t) = (1 - \rho_* \gamma) - \rho_* \bar{\mu} B_{t-1},$$

$$A_t = \text{Var}(dw_t - \rho_* dy_t) = 1 - 2 \rho_* \bar{\gamma} + \rho_*^2 \sigma_{y,t}^2$$

$$= 1 - 2 \rho_* \bar{\gamma} + \rho_*^2 (\bar{\sigma}_u^2 + \gamma^2) + \rho_*^2 \bar{\mu}^2 A_{t-1}. \quad (A10)$$

Consider the last equation in (A10), $A_t = 1 - 2 \rho_* \bar{\gamma} + \rho_*^2 (\bar{\sigma}_u^2 + \gamma^2) + \rho_*^2 \bar{\mu}^2 A_{t-1}$, which is a recursive equation in $A_t$. Then, Lemma A.1 (below) implies that $A$ does not depend on $t$, as long as $|\rho_* \bar{\mu}| < 1$. But, since the dealer takes the speculators’ strategies as given, we can use the equilibrium condition $\rho_* \bar{\mu} = b \in (0, 1)$. 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 $B$:

$$A = \frac{(1 - \rho_* \bar{\gamma})^2 + \rho_*^2 \bar{\sigma}_u^2}{1 - (\rho_* \bar{\mu})^2}, \quad B = \frac{1 - \rho_* \bar{\gamma}}{1 + \rho_* \bar{\mu}}. \quad (A11)$$
Then, equation (A10) shows that $\lambda, \rho, \tilde{\sigma}_y$ are independent on $t$ as well.

**Equilibrium Conditions**

We now use the equations derived above to solve for the equilibrium values of $\gamma, \mu, \lambda, \rho = \rho_*, A, B, \tilde{\sigma}_y$. Denote by

$$a = \rho \gamma, \quad b = \rho \bar{\mu}, \quad R = \frac{\lambda}{\rho}. \quad \text{(A12)}$$

From (A11) we have $A = \frac{(1-a)^2 + \rho^2 \tilde{\sigma}^2_\gamma}{1-b^2}$. Then, substitute $A$ in $\tilde{\sigma}^2_\gamma = \tilde{\sigma}^2_u + \gamma^2 + \tilde{\mu}^2 A$ from (A10), to obtain $\rho^2 \tilde{\sigma}^2_\gamma = \frac{\rho^2 \tilde{\sigma}^2_u + (a^2 + b^2 - 2ab^2)}{1-b^2}$. To summarize,

$$B = \frac{1-a}{1+b}, \quad A = \frac{(1-a)^2 + \rho^2 \tilde{\sigma}^2_u}{1-b^2}, \quad \rho^2 \tilde{\sigma}^2_\gamma = \frac{\rho^2 \tilde{\sigma}^2_u + (a^2 + b^2 - 2ab^2)}{1-b^2}. \quad \text{(A13)}$$

Using (A10), we get $R = \frac{\lambda}{\rho} = \frac{\gamma + \bar{\mu} B}{\gamma} = \frac{a + b}{a(1+b)}$. Also, the equation for $\rho$ implies $\rho = \frac{\gamma}{\tilde{\sigma}_\gamma} = \frac{\rho a}{\tilde{\sigma}_\gamma}$. Using the formula for $\rho^2 \tilde{\sigma}^2_\gamma$ in (A13), we compute $\rho^2 \tilde{\sigma}^2_u = (1-a)(a-b^2)$. Using this formula, we obtain $\rho^2 \tilde{\sigma}^2_y = a$ and $A = 1-a$. To summarize,

$$R = \frac{\lambda}{\rho} = \frac{a + b}{a(1+b)}, \quad \rho^2 \tilde{\sigma}^2_u = (1-a)(a-b^2), \quad \rho^2 \tilde{\sigma}^2_y = a, \quad A = 1-a. \quad \text{(A14)}$$

From (A7), we have $\frac{N_F}{N_F + 1} = \lambda \tilde{\gamma} = \frac{\lambda}{\rho} a = \frac{a+b}{1+b}$, and $B = \frac{1-a}{1+b} = \frac{B(1+b)}{N_F + 1} = \lambda \bar{\mu} = \frac{\lambda}{\rho} b = \frac{b(a+b)}{a(1+b)}$. Since $B = 1$, we have $\frac{N_L}{N_L + 1} = \frac{b(a+b)}{a(1+b)}$, 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,

$$a = \frac{N_F - b}{N_F + 1}, \quad B = \frac{1}{N_F + 1}, \quad b(1+b) \frac{N_F}{N_F + 1} = \frac{N_F - b}{N_F + 1} \frac{N_L}{N_L + 1}. \quad \text{(A15)}$$

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

From (A15), we obtain 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 $b = \omega^{1/2} = \sqrt{\frac{N_F}{N_F + 1}} \frac{N_L}{N_L + 1}$, which leads to a negative $\tilde{\sigma}^2_y$ (see (A13)). Thus, we must choose the other solution, $b = \frac{\omega^{1/2} - \omega^{1/2}}{1/2} \geq 0$. Let $b_\infty = \sqrt{\frac{N_F}{N_F + 1}} - 1$. Since $b_\infty^2 + b_\infty = 1$ and $\omega \geq 1$, we have $b_\infty^2 + b_\infty \omega \geq 1$. Moreover, since $b^2 + b\omega = \frac{N_L}{N_L + 1} < 1$, we get $b^2 + b\omega < b_\infty^2 + b_\infty \omega$. But the
function \( b^2 + b\omega \) is strictly increasing in \( b \) when \( b \geq 0 \), hence we obtain \( b < b_\infty \). Thus, \( b \in [0, b_\infty) \), as stated in the Theorem. We also obtain \( a = \frac{N_F - b}{N_F + 1} \in (0, 1) \). The proof of the exact formulas in (17) is now complete.

We now derive the asymptotic formulas in (17). When \( N_F \) is large, note that \( a = \frac{N_F - b}{N_F + 1} \approx a_\infty = 1 \), \( \omega = 1 + \frac{1}{N_F} \frac{N_L}{N_F + 1} \approx \omega_\infty = 1 \). Therefore, we also get \( b \approx b_\infty = \sqrt{\frac{b}{2}} \).

One can now verify that the formulas for \( \gamma_\infty, \mu_\infty, \lambda_\infty, \) and \( \rho_\infty \) are as stated in (17).

We now 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 \( F(\beta, \omega) = \sqrt{\omega^2 + 4\beta^2} - \omega \), and note that \( b = F(\beta, \omega)/2 \), with \( \beta = \frac{N_F}{N_L + 1} \) and \( \omega = 1 + \frac{\beta}{N_F} \). We compute \( \frac{\partial \beta}{\partial N_F} = \frac{\partial \beta}{\partial N_L} = \frac{1}{(N_L + 1)^2} \), \( \frac{\partial \omega}{\partial N_F} = -\frac{N_L(N_F + 1) - N_F}{N_F(N_F + 1)^2} < 0 \), \( \frac{\partial \omega}{\partial N_L} = \frac{1}{N_F(N_F + 1)^2} > 0 \). Also, \( \frac{\partial F}{\partial \beta} = \frac{2}{\sqrt{\omega^2 + 4\beta^2}} > 0 \), and \( \frac{\partial F}{\partial \omega} = \frac{\beta}{\sqrt{\omega^2 + 4\beta}} - 1 = -\frac{b}{\sqrt{\omega^2 + 4\beta}} < 0 \). Then, \( \frac{\partial(2b)}{\partial N_L} = \frac{\partial F}{\partial \beta} \cdot \frac{\partial \beta}{\partial N_L} + \frac{\partial F}{\partial \omega} \cdot \frac{\partial \omega}{\partial N_L} = \frac{1}{(N_L + 1)^2 \sqrt{\omega^2 + 4\beta}} (2 - \frac{b}{N_F}) > 0 \), where the last inequality follows from \( b \in (0, 1) \).

We end the analysis of the equilibrium conditions, by proving several more useful inequalities for \( a \) and \( b \). Denote by \( \beta_F = \frac{N_F}{N_F + 1} \) and recall that \( \beta = \frac{N_F}{N_L + 1} \). Then, \( b \) satisfies the quadratic equation \( b^2 + b\omega = \beta \), with \( \omega = 1 + \frac{\beta}{N_F} \). Now start with the straightforward inequality \( \beta < \beta_F + 1 \), and multiply it by \( \beta_F \). We get \( \beta \beta_F < \beta_F^2 + \beta_F \).

Since \( \beta_F = 1 - \frac{\beta_F}{N_F} \), we get \( \beta(1 - \frac{\beta_F}{N_F}) < \beta_F^2 + \beta_F \), or equivalently \( \beta < \beta_F^2 + \beta_F(1 + \frac{\beta}{N_F}) \).

Since \( b^2 + b\omega = \beta \) and \( \omega = 1 + \frac{\beta}{N_F} \), we get \( b^2 + b\omega < \beta_F^2 + \beta_F \omega \). Because the function \( f(x) = x^2 + x\omega \) is increasing in \( x \in (0, 1) \), we have \( b < \beta_F = \frac{N_F}{N_F + 1} \). This inequality is equivalent to \( N_F - b > N_F b \). Dividing by \( N_F + 1 \), we get \( a = \frac{N_F - b}{N_F + 1} > \frac{N_F b}{N_F + 1} = b\beta_F \). But we have already seen that \( \beta_F > b \), hence \( a > b\beta_F > b^2 \). To summarize,

\[
b < \frac{N_F}{N_F + 1}, \quad a > b^2. \tag{A16}
\]

Lemma A.1 can now be used to show that the coefficients \( A \) and \( B \) are constant. Indeed, in the proof of the Theorem, we have seen that 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 we work in continuous time, and \( t+1 \) actually stands for \( t+dt \), the convergence occurs in an infinitesimal amount of time. Thus, \( X_t \) is constant for all \( t \), and that constant is
equal to $\frac{\alpha}{1-\beta}$.

We 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:

$$X_t - \beta X_{t-1} = \alpha, \quad t \geq 2. \quad \text{(A17)}$$

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

**Proof.** First, note that $\bar{X}$ is well defined as long as $\beta \neq 1$. If we denote by $Y_t = X_t - \bar{X}$, the new sequence $Y_t$ satisfies the recursive equation $Y_t - \beta Y_{t-1} = 0$. We now show that $Y_t$ converges to 0 (and $\bar{X}$ is well defined) if and only if $\beta \in (-1, 1)$. Then, the difference equation $Y_t - \beta Y_{t-1} = 0$ has the following general solution:

$$Y_t = C\beta^t, \quad t \geq 1, \quad \text{with } C \in \mathbb{R}. \quad \text{(A18)}$$

Then, $Y_t$ is convergent for any values of $C$ if and only if all $\beta \in (-1, 1]$. But in the latter case, $1 - \beta = 0$, which makes $\bar{X}$ nondefined.

**Proof of Corollary 1.** In the proof of Theorem 1, equation (A7) implies $\lambda\bar{\gamma} = \frac{N_F}{N_F+1}$, $\lambda\bar{\mu} = \frac{B}{A} \frac{N_L}{N_L+1}$. But from (A13) and (A14), we have $\frac{B}{A} = \frac{1}{1+b}$, which proves the first row in (18). The second row in (18) just rewrites the formulas for $A$ and $B$ from equations (A13) and (A14).

**Proof of Proposition 1.** From Corollary 1, $\lambda\bar{\gamma} = \frac{N_F}{N_F+1}$ and $\lambda\bar{\mu} = \frac{B}{A} \frac{N_L}{N_L+1}$. From (A6), the equilibrium normalized expected profit of the FT is

$$\tilde{\pi}^F = \gamma - \lambda\bar{\gamma}\bar{\gamma} + \mu B - \lambda\bar{\mu}\bar{A} = \gamma \left(1 - \frac{N_F}{N_F+1}\right) + B\mu \left(1 - \frac{N_L}{N_L+1}\right) \quad \text{(A19)}$$

From (A15), $B = \frac{1}{N_F+1}$, which proves the desired formula for $\pi^F$. 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.

\[55\]
**Justification of Result 1.** According to Proposition 1, $\delta \, dt$ is the expected profit that speculators get per unit of time $dt$ from trading on their lagged signal $(dw_{t-1})$. 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 ($\delta$), but would bring a lower profit. For this last statement we use the results of Internet Appendix Section 1 (Proposition IA.3), where we show 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 $M_{0,1}$ model. The key step is to observe that in Theorem 1 the fast trader’s choice of $\mu$ is the same as the slow trader’s choice of $\mu$, 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 $1-a = \frac{1+b}{N_{F_{p+b}}}$, equation (17) implies that $\lambda = \rho \frac{N_F}{N_{F-p-b}} = \frac{\sigma_w}{\sigma_u} \sqrt{(1-a)(a-b^2)} \frac{N_{F-p}}{N_{F-p-b}}$, which proves the first equation in (25).

By definition, the trading volume is $TV = \sigma_y^2$. From (A14), $TV = \sigma_y^2 = \bar{a}^2 \sigma_w^2 = \frac{a^2 \rho^2}{\rho^2}$. From (17), $\rho^2 = \frac{\sigma_y^2}{\sigma_u} (1-a)(a-b^2)$, hence $TV = \sigma_u^2 \frac{a}{(1-a)(a-b^2)}$. Substituting $1-a = \frac{1+b}{N_{F+1}}$, we get $TV = \sigma_u^2 (N_F + 1) \frac{a}{(1+b)(a-b^2)}$, which proves the second equation in (25).

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

The speculator participation rate is $SPR = \frac{\sigma_w^2 \sigma_{\omega}^2 + \bar{\mu}^2 \sigma_w^2}{TV} = \frac{\rho^2 (\gamma^2 \bar{\sigma}_w^2 + \mu^2 \gamma^2 - \bar{w}^2 \gamma^2)}{\sigma_{\omega}^2}. \quad \text{Since } \rho \gamma = a, \rho \bar{\mu} = b, \text{ and } \sigma_{\omega}^2 = (1-a) \sigma_w^2, \text{ we get } SPR = \frac{a^2 + b^2 (1-a)}{a}. \quad \text{This proves the last equation in (25), since } \frac{1-a}{a} = \frac{1+b}{N_{F-p}}.$

**Proof of Proposition 3.** As in Theorem 1, we start with FT’s choice of optimal trading strategy. Each FT $i = 1, \ldots, N_F$ observes $d w_t$, and chooses $dx_t^i = \gamma_t^i dw_t$ to maximize the expected profit:

$$\pi_0 = E \left( \int_0^T \left( w_t - p_{t-1} - \lambda_t(dx_t^i + dx_{t-1}^i + dw_t) \right) dx_t^i \right) = \int_0^T \gamma_t^i \sigma_w^2 dt - \lambda_t \gamma_t^i (\gamma_t^i + \gamma_t^{-i}) \sigma_w^2 dt,$$

(A20)

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

The dealer takes FTs’ strategies as given, thus assumes that the aggregate order flow is of the form \( \text{d}y_t = \text{d}u_t + N_F \gamma_t \text{d}w_t \). To set \( \lambda_t \), the dealer sets \( p_t \) such that \( \text{d}p_t = \lambda_t \text{d}y_t \), with \( \lambda_t = \frac{\text{Cov}(w_t, \text{d}y_t)}{\text{Var}(\text{d}y_t)} = \frac{N_F \gamma_t \sigma_w^2}{\sigma_w^2 + N_F \gamma_t \sigma_w^2} \). This implies \( \lambda_t^2 \sigma_w^2 + (N_F \gamma_t \lambda_t)^2 \sigma_w^2 = N_F \gamma_t \lambda t \sigma_w^2 \).

But \( N_F \lambda_t \gamma_t = \frac{N_F}{N_F+1} \). Hence, \( \lambda_t^2 \sigma_w^2 + (N_F \lambda_t) \sigma_w^2 = N_F \gamma_t \lambda t \sigma_w^2 \), or \( \lambda_t^2 \sigma_w^2 = \frac{N_F}{(N_F+1)^2} \sigma_w^2 \), which implies the formula \( \lambda = \frac{\sigma_w}{\sigma_w} \sqrt{\frac{N}{N_F+1}} \). We then compute \( \gamma_t = \frac{1}{\lambda t} \frac{N_F}{N_F+1} = \frac{\sigma_w}{\sigma_w} \frac{1}{\sqrt{N_F+1}} \).

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

Finally, we compute \( \Sigma' \). From the formula above for \( \lambda \), we get \( \text{Var}(\text{d}p_t) = \lambda^2 \text{Var}(\text{d}y_t) = \lambda \text{Cov}(w_t, \text{d}y_t) = \text{Cov}(w_t, \text{d}p_t) \). Since \( \Sigma_t = \text{Var}(w_t - p_t-1) = E((w_t - p_t-1)^2) \), we compute

\[
\Sigma_t' = \frac{1}{dt} E(2(dw_{t+1} - dp_t)(w_t - p_t-1) + (dw_{t+1} - dp_t)^2) = -2 \frac{\text{Cov}(w_t, \text{d}p_t)}{dt} + \sigma_w^2 + \frac{\text{Var}(\text{d}p_t)}{dt} = \sigma_w^2 - \sigma_p^2 = \frac{\sigma_u^2}{N_F+1} \).
\]

**Proof of Proposition 4.** We use the formulas from the proof of Theorem 1. Since \( \tilde{d}w_t \) is orthogonal on \( \text{d}y_t \), we have \( \tilde{\text{Cov}}(\tilde{d}w_t, \text{d}y_t) = \tilde{\text{Cov}}(\tilde{d}w_t, \tilde{d}w_t) = A = 1 - a = \frac{1+b}{N_F+1} \).

Then, \( \tilde{\text{Cov}}(\tilde{d}w_t, \tilde{d}w_{t-1}) = \tilde{\text{Cov}}(\tilde{d}w_t - \rho \tilde{d}w_t - \rho \tilde{d}w_{t-1}, \tilde{d}w_{t-1}) = -\rho \tilde{\mu} A \). Therefore,

\[
\tilde{\text{Cov}}(\tilde{d}x_{t+1}, \tilde{d}x_t) = \tilde{\text{Cov}}(\tilde{d}w_{t+1} + \tilde{\mu} \tilde{d}w_t, \tilde{d}w_t + \tilde{\mu} \tilde{d}w_{t-1}) = \tilde{\mu} \tilde{\gamma} \tilde{A} + \tilde{\mu}^2 (-bA)
\]

\( \tilde{\text{Var}}(\tilde{d}x_t) = \tilde{\text{Var}}(\tilde{d}w_t + \tilde{\mu} \tilde{d}w_{t-1}) = \tilde{\gamma}^2 + \tilde{\mu}^2 A \).

By multiplying both the numerator and denominator by \( \rho^2 \), we compute

\[
\rho_x = \frac{\tilde{\mu} \tilde{\gamma} \tilde{A}}{\tilde{\gamma}^2 + \tilde{\mu}^2 A} - \frac{b \tilde{\mu}^2 A}{\tilde{\gamma}^2 + \tilde{\mu}^2 A} = \frac{ab(1-a)}{a^2 + b^2(1-a)} - \frac{b^3(1-a)}{a^2+b^2(1-a)}
\]

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

We now prove that \( \rho_x > 0 \) if and only if there exists slow trading. When there is no slow trading, \( b = \rho \tilde{\mu} = 0 \), hence \( \rho_x = 0 \). When there is slow trading, we show that \( \rho_x = \frac{b(1+b)-b^2}{a^2+b^2(1-a)} \frac{1}{N_F+1} > 0 \). Indeed, we have \( b > 0 \), \( a < 1 \), and from equation (A16), \( a - b^2 > 0 \).

**Proof of Proposition 5.** By definition, \( \text{d}(x_t p_t) = x_t p_t - x_{t-1} p_{t-1} = p_t \text{d}x_t + x_t \text{d}p_t \).

Integrating this equality, we get \( x_T p_T - x_0 p_0 = \int_0^T p_t \text{d}x_t + \int_0^T x_t \text{d}p_t \). But \( x_T = x_0 = 0 \)

57
(almost surely), hence \(-\int_0^T p_t \, dx_t = \int_0^T x_{t-1} \, dp_t\). We also have \(\int_0^T v_t \, dx_t = v_T(x_T - x_0) = 0\). Thus, \(\pi = \mathbb{E} \int_0^T (v_t - p_t) \, dx_t = -\mathbb{E} \int_0^T p_t \, dx_t = \mathbb{E} \int_0^T x_{t-1} \, dp_t\).

**Proof of Proposition 6.** If \(x_t\) is IFT’s inventory in the risky asset, denote by

\[
\Omega_t^{xx} = \frac{\mathbb{E}(x_t^2)}{\sigma_w^2 \mathbb{d}t}, \quad X_t = \frac{\mathbb{E}(x_t \, \mathbb{d}w_t)}{\sigma_w^2 \mathbb{d}t}, \quad Z_t = \frac{\mathbb{E}(x_{t-1} \, dy_t)}{\sigma_w^2 \mathbb{d}t}.
\]

(A23)

Since \(\Theta > 0\), we have \(\Theta \in (0, 2)\), or \(\phi = 1 - \Theta \in (-1, 1)\). From (32), \(x_t\) satisfies the recursive equation \(x_t = \phi x_{t-1} + G \, dw_t\). We compute \(\Omega_t^{xx} = \frac{\mathbb{E}(x_t^2)}{\sigma_w^2 \mathbb{d}t} = \mathbb{E}((\phi x_{t-1} + G \, dw_t)^2) = \phi^2 \Omega_{t-1}^{xx} + G^2\). Since \(\phi^2 \in (-1, 1)\), we apply Lemma A.1 to the recursive formula \(\Omega_t^{xx} = \phi^2 \Omega_{t-1}^{xx} + G^2\). Then, \(\Omega_t^{xx}\) is constant and equal to:

\[
\Omega^{xx} = \frac{G^2}{1 - \phi^2} = \frac{G^2}{\Theta(1 + \phi)},
\]

(A24)

which is the usual variance formula for the \(AR(1)\) process. From (A24) it follows that

\[
\mathbb{E}(x_t^2) = \Omega^{xx} \sigma_w^2 \mathbb{d}t,
\]

(A25)

which implies that the inventory is infinitesimal. It follows that the inventory costs are zero, and all the profits are in cash. Also, IFT’s expected utility is the same as his expected profit. As the initial inventory is \(x_0 = 0\), we have that \(x_T = 0\), and Proposition 5 implies that IFT’s expected profit is

\[
\pi_{\omega > 0} = \lambda \mathbb{E} \int_0^T x_{t-1} \, dy_t = \lambda \int_0^T Z_t \, dt.
\]

(A26)

The order flow at \(t\) is \(dy_t = -\Theta x_{t-1} + \gamma \, dw_t + \tilde{\mu} \, dw_{t-1} + du_t\), with \(\gamma = \gamma^- + G\). Then, \(Z_t\) is a function of \(X_{t-1}\):

\[
Z_t = \frac{\mathbb{E}(x_{t-1} \, dy_t)}{\sigma_w^2 \mathbb{d}t} = -\Theta \Omega_{t-1}^{xx} + \tilde{\mu} X_{t-1} = -\frac{G^2}{1 + \phi} + \tilde{\mu} X_{t-1}.
\]

(A27)

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

$$X = \frac{G(1 - a^-) - \frac{\rho G^2}{1 + \phi}}{1 + \phi b}. \quad (A28)$$

From (A27), $Z_t$ is also constant and satisfies:

$$Z = \bar{\mu}X - \frac{G^2}{1 + \phi} = \bar{\mu}G\frac{1 - a^-}{1 + \phi b} - G^2\frac{b + \frac{1}{1 + \phi}}{1 + \phi b}. \quad (A29)$$

From (A26), IFT’s expected profit is

$$\tilde{\pi}_{\Theta > 0} = \lambda Z = \lambda \left( \bar{\mu}G\frac{1 - a^-}{1 + \phi b} - G^2\frac{b + \frac{1}{1 + \phi}}{1 + \phi b} \right). \quad (A30)$$

This finishes the proof. \qed

**Proof of Theorem 2.** Let $\Theta = 0$. Then, IFT’s strategy is of the form $dx_t = Gdw_t$. We compute IFT’s expected profit $\pi_{\Theta = 0} = E \int_0^T (w_t - p_t)dx_t = E \int_0^1 (w_{t-1} - p_{t-1} + dw_t - \lambda dy_t)(Gdw_t) = E \int_0^1 (dw_t - \lambda dy_t)(Gdw_t) = E \int_0^1 (dw_t - \lambda \gamma dw_t)(Gdw_t) = G(1 - \lambda \gamma)\sigma_w^2$.

But $\lambda \gamma = \lambda G + \lambda \gamma^- = \lambda G + Ra^-$. The normalized IFT’s expected profit is:

$$\tilde{\pi}_{\Theta = 0} = G(1 - \lambda \gamma) = G(1 - Ra^-) - \lambda G^2. \quad (A31)$$

To compute IFT’s inventory costs, denote by $\Omega_{x^x}^t = \frac{E(x_t^2)}{\sigma_x^2}$. We compute $\frac{d\Omega_{x^x}^t}{dt} = \frac{1}{\sigma_x^2 dt} E(2x_{t-1}dx_t + (dx_t)^2) = \frac{1}{\sigma_x^2 dt} E(2Gx_{t-1}dw_t + G^2(dw_t)^2) = G^2$. Since $\Omega_{x^x}^0 = 0$, the solution of this first order ODE is $\Omega_{x^x}^t = tG^2$, for all $t \in [0, 1]$. Hence, the inventory costs are equal to

$$C_I E \int_0^1 x_t^2 dt = C_I G^2 \int_0^1 t dt = \frac{C_I}{2} G^2, \quad (A32)$$

From (A31) and (A32), IFT’s normalized expected utility when $\Theta = 0$ is:

$$\tilde{U}_{\Theta = 0} = G(1 - Ra^-) - G^2 \left( \lambda + \frac{C_I}{2} \right). \quad (A33)$$

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

\[
\bar{U}_{\Theta=0}^{\text{max}} = \frac{(1 - Ra^-)^2}{2(2\lambda + C_I)}.
\] (A34)

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

\[
G = \frac{b(1-a^-)}{2\rho \left( b + \frac{1}{1+\phi} \right)},
\] (A35)

which expresses the optimum \( G \) as a function of \( \phi \). The second order condition for a maximum is \( \lambda b(1+\phi)^2 > 0 \), 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:

\[
\bar{U}_{\Theta,0} = \frac{(Rb(1-a^-))^2}{4\lambda(1+\phi)(b + \frac{1}{1+\phi})}.
\] (A36)

We now analyze the function

\[
f(\phi) = (1+\phi b) \left( b + \frac{1}{1+\phi} \right) \implies f'(\phi) = \frac{b^2(1+\phi)^2 + b - 1}{(1+\phi)^2}.
\] (A37)

The polynomial in the numerator has two roots:

\[
\phi_1 = -1 + \frac{\sqrt{1-b}}{b}, \quad \phi_2 = -1 - \frac{\sqrt{1-b}}{b}.
\] (A38)

By assumption \( b < 1 \), hence both roots are real. Clearly, we have \( \phi_2 < -1 \). We show that \( \phi_1 \in (-1,1) \). First, note that \( \phi_1 \) is decreasing in \( b \). For \( b = 1 \) we have \( \phi_1 = -1 \); while for \( b = \frac{\sqrt{17}-1}{8} \) (which satisfies \( 4b^2 + b = 1 \)) we have \( \phi_1 = 1 \). 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 \( \bar{U}_{\Theta,0} \) from (A36) attains its maximum at \( \phi = \phi_1 \), or
\[ \Theta = 2 - \frac{\sqrt{\frac{1-a^-}{b}}}{b}, \] as stated in the Theorem. Also, if we substitute \( \phi = \phi_1 \) in (A35), we get
\[ G = \frac{1-a^-}{2\rho(1+\frac{1}{\sqrt{1-a^-}})}, \] as stated in the Theorem. The maximum value (over both \( G \) and \( \Theta \)) is:
\[ \tilde{U}_{\Theta > 0}^{\text{max}} = \frac{(Rb(1-a^-))^2}{4\lambda b(1+\sqrt{1-b})^2}. \] (A39)

To determine the cutoff value for the inventory aversion coefficient \( C_I \), we set \( \tilde{U}_{\Theta > 0}^{\text{max}} = \tilde{U}_{\Theta > 0}^{\text{max}} \). From (A34) and (A39), algebraic manipulation shows that the cutoff value is
\[ C_I = 2\lambda \left( \frac{(1-Ra^-)^2(1+\sqrt{1-b})^2}{1+\sqrt{1-b}} - 1 \right), \] as stated in the Theorem.

**Proof of Corollary 3.** Let \( \Theta > 0 \). We are in the context of Theorem 2, where \( b > \frac{\sqrt{\lambda}}{8} \) and \( \rho > 0 \), hence \( \bar{\mu} = b \rho > 0 \). The IFT’s strategy is \( dx_t = -\Theta x_{t-1} + Gdw_t \), while the slow trading component is \( \tilde{d}_t^S = \bar{\mu} \tilde{w}_{t-1} \). Since \( dw_t \) is orthogonal to \( \tilde{w}_{t-1} \), \( \text{Cov}(dx_t, \tilde{d}_t^S) = -\Theta \text{Cov}(x_{t-1}, \tilde{d}_t^S) = -\Theta \bar{\mu} \text{Cov}(x_{t-1}, \tilde{w}_{t-1}) \). This proves the equality in (42). Since \( \Theta > 0 \) and \( \bar{\mu} > 0 \), it remains to prove the inequality \( \text{Cov}(x_{t-1}, \tilde{w}_{t-1}) > 0 \). But \( \text{Cov}(x_{t-1}, \tilde{w}_{t-1}) = X\sigma_w^2 dt \) (see (A23)). From (A28), \( X = \frac{G(1-a^-) - \rho^2}{1+\rho} \). Substituting the optimal \( G \) and \( \phi = 1 - \Theta \) from Theorem 2, we obtain \( X = \frac{(1-a^-)^2}{4(1+\sqrt{1-b})} \). As in Theorem 2, \( a^- \in [0,1) \), hence \( X > 0 \) and the proof is complete.

**Proof of Theorem 3.** Consider the following implicit equation in \( b \)
\[ \frac{2b(1+b)(2B+1)}{n_L} = \frac{Q}{B^2(a^-+b)} + \frac{3bB+2b^2B-1-b}{b(1-a^-)} - 2, \] (A40)

where the following substitutions are made:\textsuperscript{52}
\[ n_F = \frac{N_F}{N_F + 1}, \quad n_L = \frac{N_L}{N_L + 1}, \quad B = \frac{1}{\sqrt{1-b}}, \]
\[ q = (B+1)\left(2(B^2-1) - n_F(3B^2-2)\right), \]
\[ a^- = \frac{-q \pm \sqrt{q^2 + n_F B^5\left((4-n_F)B + 2(2-n_F)\right)}}{B^2\left((4-n_F)B + 2(2-n_F)\right)}, \]
\[ Q = B^3(a^-)^2 + 2(3B^3 + 3B^2 - 2B - 1)a^- + (B^3 + 2B^2 - 2). \] (A41)

\textsuperscript{52}To be rigorous, we have included the case when \( a^- \) is negative. However, numerically this case never occurs in equilibrium, because it leads to \( \lambda < 0 \), which contradicts FT’s second order condition (IA.315) in Internet Appendix Section 5.
We write the equations for the other coefficients:

\[
R = \frac{4(B+1)B^2(a^-+b)}{Q}, \quad a = \frac{(2B+1)a^-+1}{2(B+1)}
\]

\[
\rho^2 = \left(\left(a-b^2\right)+\frac{2bB-1}{2B+1}(1-a)\right)\left(1-a\right)\frac{\sigma_w^2}{\sigma_u^2}, \quad \lambda = R\rho \quad \text{(A42)}
\]

\[
\Theta = 2 - \sqrt{1-b}, \quad G = \frac{1-a}{\rho(2B+1)}, \quad \gamma = \frac{a^-}{\rho N_F}, \quad \mu = \frac{b}{\rho N_L}.
\]

The proof is now left to Internet Appendix Sections 5.1 and 5.2.

\[\square\]

**Proof of Proposition 7.** See Internet Appendix Section 5.2.

\[\square\]

**Proof of Proposition 8.** See Internet Appendix Section 5.2.

\[\square\]

REFERENCES


