The Shadow Price of Latency: Improving Intraday Fill Ratios in Foreign Exchange Markets

Álvaro Cartea[a,b], Leandro Sánchez-Betancourt[a,c]

[a] *Mathematical Institute, University of Oxford, Oxford, UK*
[b] *Oxford-Man Institute of Quantitative Finance, Oxford, UK*
[c] *LMAX Exchange, London, UK*

Abstract

Latency is the time delay between an exchange streaming market data to a trader, the trader processing information and deciding to trade, and the exchange receiving the order from the trader. Liquidity takers face a moving target problem as a consequence of their latency in the marketplace. They send marketable orders that aim at a price and quantity they observed in the limit order book (LOB), and by the time their order is processed by the Exchange, prices could have worsened, so the order may not be filled, or prices could have improved, so the order is filled at a better price. We show how to choose the discretion of liquidity taking orders to walk the LOB to increase the chances of filling orders if, due to latency, prices or quantities in the LOB have worsened. The latency-optimal strategy balances the tradeoff between the costs of walking the LOB and targeting a desired percentage of filled orders over a period of time. We employ a proprietary data set of foreign exchange trades to analyze the performance of the strategy and build a function that maps latency of a trader to the percentage of orders filled by the trader's strategy. We employ this function to compute the shadow price of latency that a trader would be willing to pay for co-location and hardware to reduce their latency in the marketplace.

Keywords: Latency, fill ratio, high-frequency trading, algorithmic trading.

1. Introduction

With the advent of computerized trading, the speed at which agents operate in electronic markets is down to milliseconds or microseconds. While speed to process information and take decisions is paramount to the success of trading strategies, being faster than one's peers

[☆]We are grateful to Lane Hughston, José Penalva, and Andrew Stewart for useful discussions. We are also grateful to seminar participants at LMAX Exchange, King's College London, and the University of Oxford, for their comments and suggestions.

Email addresses: alvaro.cartea@maths.ox.ac.uk (Álvaro Cartea), leandro.sanchezbetancourt@maths.ox.ac.uk (Leandro Sánchez-Betancourt)

provides a valuable competitive edge. Traders require time to devise strategies and their instructions take time to arrive and to be processed by the Exchange. This time delay, known as latency, affects the efficacy of trading strategies because, in the meantime, the Exchange processes other instructions that update the limit order book (LOB) with information not known by the trader at the time she devised the strategy.

In foreign exchange (FX) markets, the supply of spot currency pairs displayed in the LOB might undergo thousands of updates over very short periods of time. When the best bid and best ask prices in the LOB are changing very rapidly, liquidity takers face a moving target problem as a consequence of their latency in the marketplace. Liquidity takers send marketable orders that aim at a price and quantity they have observed in the LOB, but by the time their order is processed by the Exchange, prices or quantities may have worsened, so the order cannot be filled; alternatively, prices or quantities may have improved, so the order is filled at better prices. Traders can buffer the adverse effects of missing a trade by including a price limit in their marketable orders to increase the probability of filling the order when it is processed by the Exchange. This price limit consists of the best price seen by the trader in the LOB plus a discretionary buffer that specifies the number of ticks the order can walk the LOB. This buffer does not preclude the order from being filled at better prices if the LOB is updated with more favorable prices or quantities.

In this paper we show how to choose the discretion of marketable orders of FX traders in an optimal way to improve fill ratios over a period of time (days, weeks, months), while keeping orders exposed to receiving a price improvement. Increasing fill ratios is costly. Everything else being equal, the chances of filling a marketable order increase if the order can walk the LOB. Thus, there is a tradeoff between ensuring high fill ratios and the execution costs borne by the trading strategy. In our approach, the dynamic optimization problem solved by the trader balances this tradeoff by minimizing the discretion specified in the marketable orders, while targeting a fill ratio over a trading horizon. The trader's optimal strategy specifies the discretion for each transaction depending on the proportion of orders that have been filled, how far is the strategy from the target fill ratio, the cost of walking the LOB, and the volatility of the exchange rate.

We employ a proprietary data set of FX trades to analyze the performance of the optimal strategy developed here. The data are provided by LMAX Exchange – an electronic multilateral trading facility in the FX market (www.lmax.com). We use transaction data for two FX traders, trader 1 (T1) and trader 2 (T2), to compare the fill ratios they have achieved to those attainable with the optimal strategy derived in this paper. We propose three measures of fill ratios. These are based on the number of trades sent to the Exchange, the number of misses due to latency, volume, and type of order, and the state of the LOB. The data spans a set of dates from December 2016 to March 2017. During this period both traders filled between approximately 80% and 90% of their liquidity taking orders in the currency pair US dollar and Japanese yen, which we refer to as the USD/JPY pair. Latency causes T1 and T2 to miss trades because by the time the Exchange processes their orders, the best bid and best ask prices in the LOB are updated. The effect of latency on trade fills is exacerbated during times of heightened volatility in the exchange rate. When volatility is arranged in quartiles, we find that 36.5% (resp. 40.0%) of T1's (resp. T2's) unfilled trades

occur in the top quartile of volatility.

Clearly, due to latency, trades are not filled because the market moves away from the prices and quantities that T1 and T2 are attempting to achieve. If T1 and T2 were to return to the market to complete unfilled trades, the prices they would receive would be highly likely to be worse than those at which they attempted in the original trade. The cost we impute to a missed trade is the mark-to-market price of a market order that walks the LOB until filled, and is sent 20ms or 100ms after the Exchange receives the instruction to execute the order that could not be filled. We find that the average mark-to-market cost[1] for T1's missed trades is 2.33 and 2.88 ticks and for T2 is 3.18 and 3.06 ticks, 20ms and 100ms later respectively (in the currency pair USD/JPY one tick is 10^{-3} JPY).[2]

We employ the optimal strategy developed here to show the tradeoff between increasing fill ratios and the costs incurred by the strategy. For example, for a particular choice of model parameters, we show that T1 and T2 can increase the percentage of filled trades, during the period 2 December 2016 to 30 March 2017, from 91% and 89%, respectively, to 99% in both cases. The increase in the fill ratios is due to the discretion included in the liquidity taking orders to walk the book, which come at a cost. In this example, the average cost incurred by T1 to fill missed trades is 1.76 ticks and for T2 is 1.24 ticks. On the other hand, the mark-to-market average cost of filling the missed trades (which were filled by the optimal strategy) employing market orders 20ms and 100ms later is 2.01 and 2.58 ticks respectively for T1, and 2.82 and 2.75 ticks respectively for T2.

The performance of the optimal strategy is more remarkable during times of heightened volatility of the exchange rate. In the top quartile of volatility, the average cost of filling missed trades is 1.88 ticks when T1 employs the optimal strategy, while the mark-to-market average cost of filling these missed trades with market orders 100ms later is 3.04 ticks. Similarly, for T2 the average cost of filling missed trades with the optimal strategy is 1.86 ticks, while the mark-to-market average cost of filling the missed trades with a market order 100ms later is 3.31 ticks.

Finally, we use the proprietary data to map different levels of latency to the corresponding percentage of filled orders. We use this mapping to calculate the shadow price of latency that T1 would be willing to pay to reduce latency in the marketplace. We show that T1 would be better off employing the latency-optimal strategy developed here, instead of investing in hardware and co-location services to reduce latency. The latency-optimal strategy is superior because it not only achieves the same fill ratios as those obtained with better hardware and co-location, but it scoops price improvements that stem from orders arriving with latency at the Exchange. Note that if latency is reduced with hardware and co-location, T1's liquidity taking strategy would be less affected by the moving target problem, so fill ratios are high, but would benefit little from changes in the LOB that, due to latency, provide price improvements.

[1]The average mark-to-market cost is calculated by taking the difference between the cost of filling all the volume missed in the order and the cost that the trader initially attempted to pay to fill the order, and dividing by the missed volume of the order.

[2]From February 2018, the tick size for the pair USD/JPY is 10^{-4} JPY in LMAX Exchange.

A number of authors have recently addressed various aspects of latency in electronic markets. Moallemi and Saĝlam (2013) look at the cost of latency in high-frequency trading. The authors employ market data to quantify the costs of latency in equity markets and show that the latency costs are of the same order of magnitude as other trading costs such as commissions and exchange fees. The data employed is publicly available from the NYSE, which does not contain trader identification to analyze latency for individual market participants. Our study, on the other hand, focuses on liquidity taking orders in the FX market and how to develop trading strategies that minimize the adverse effects of latency.

The work of Stoikov and Waeber (2016) shows how to execute a large order in electronic markets by employing the volume imbalance of the LOB to predict price changes and study the effect of latency in the efficacy of the execution strategy. The authors backtest their strategy on publicly available data for US Treasury bonds and show that the advantage of observing the LOB to compute volume imbalance dissipates quickly as latency increases. Lehalle and Mounjid (2017) employ data from Nasdaq-Omx and also find that as latency increases, the informational content in the volumes of the LOB diminishes.

Recent literature on high-frequency trading and algorithmic trading discusses various characteristics of trading and how the speed of traders is used to obtain informational advantages. Other strands of the literature discuss the relationship of market quality and the speed of market participants. See for example Bayraktar and Ludkovski (2010) and Guéant (2016) for trading in illiquid markets. Barger and Lorig (2018) model the rapid updates of the best quotes in the LOB to propose a model of stochastic price impact. Finally, Cartea and Penalva (2012), Hoffmann (2014), Biais et al. (2015), Foucault et al. (2016), and Fricke and Gerig (2018), look at issues of market quality and the speed of trading.

We believe that our paper is the first to employ proprietary data to analyze the effect of latency on the fill rates of liquidity taking strategies in FX markets and to compute the shadow price of latency for particular traders. Latency is specific to each trader and the effects of latency depend on the type of strategy employed by the trader. Access to proprietary data of FX transactions allows us to quantify the effects of latency on the efficacy of trading strategies and to develop and test latency-optimal strategies to compute the shadow price of latency for types of market participants.

The remainder of the paper is organized as follows. In Section 2 we define latency and propose various ways to measure the fill ratios of marketable orders. In Section 3 we describe the proprietary data set that we employ and we present empirical evidence of the relationship between fill ratios and the volatility of the exchange rate. In Section 4 we present the model of fill ratios and solve the trader's dynamic optimization problem, where we also assume that the trader makes her model of fill ratios robust to misspecification. In Section 5 we show how we estimate model parameters, and in Section 6 we show the performance of the trading strategy. Section 7 discusses the tradeoff between latency and fill ratios and computes the shadow price of latency for T1. Finally, Section 8 concludes, and we collect proofs in the Appendix.

2. Latency and Fill Ratios

2.1. Latency

Latency is defined as the time "delay between sending a message to the market and it being received and processed by the Exchange. Sometimes the time it takes for the Exchange to acknowledge receipt is also accounted for", see Cartea et al. (2015). This concept is useful if traders are only interested in measuring how long it takes a message to reach the Exchange once it has been sent. However, this definition is 'narrow' because it only takes into account hardware capacity and co-location to measure latency and does not account for other factors that affect how quickly the trader can process and react to market information before sending an instruction to the Exchange.

In this paper we employ a broader definition of latency. To the time delay considered in the narrow definition of latency we add the time it takes the trader to devise the trading strategy. Thus, latency consists of three layers of time delay associated to these events: (i) The Exchange streams quotes to the trader. (ii) The trader receives quotes from the Exchange and processes this and other relevant information to make a decision. (iii) Once a decision is made, the trader sends a message to the Exchange. The decision could be to trade using limit or marketable orders, or to amend or cancel a limit order already resting in the LOB, see Moallemi and Saĝlam (2013) and Stoikov and Waeber (2016) who employ a similar definition of latency.

We focus on two types of orders: Immediate-or-Cancel (IoC) and Fill-or-Kill (FoK). IoC and FoK are marketable orders in which the trader specifies the maximum (minimum) price they are willing to pay (receive) when buying (selling) currency pairs. IoC is an order to buy or sell currency pairs which must be executed immediately, obeying the price limit, and any portion of the volume of the order that cannot be filled at the desired price limit is cancelled. FoK is an order to buy or sell currency pairs which must be executed immediately in full or cancelled.

We provide an example of the steps in the life of a FoK order to illustrate the effect of latency on the fill of the trade.

Assume the Exchange streams a quote to a trader at time t_0. The trader receives the quote at time $t_1 > t_0$ and at time $t_2 > t_1$ the trader sends a buy FoK order for 100 lots of the currency pair USD/JPY. The limit price of the FoK order is the best ask streamed by the Exchange at time t_0. Finally, the Exchange receives and processes the order at time $t_3 > t_2$, thus our measure of latency is $t_3 - t_0$. The order will be filled in full if at time t_3 the best ask in the book is equal or lower (i.e., price improvement) than the best ask at time t_0 and there are at least 100 lots posted at that price, otherwise the order is not filled – in either case, the Exchange sends a message to the trader notifying the outcome of the trade.

The effect of latency on the outcome of strategies depends on a number of factors that affect the supply and demand of liquidity in the FX market, including: magnitude of latency, type of order sent by trader, volatility of the mid-exchange rate, and trend in the mid-exchange rate.

On the supply side of liquidity, the number of updates in the LOB will determine if the trade will be filled at the price, or a better price, than that specified in the FoK order, or

if the order is not filled because prices have worsened. Changes in the best quotes of the LOB are difficult to predict and depend on many factors such as news announcements and other idiosyncratic needs of the liquidity providers who send, update, and amend their limit orders in the LOB.

On the demand side of liquidity, the effect of latency will depend on the type of strategy employed by the trader. For example, latency will increase the chances of missing trades and decrease the chances of obtaining price improvements for momentum trading strategies, i.e., buy before rates increase or sell before rates decrease.

2.2. Fill ratios and post-trade analysis

Fill ratios of orders sent by a trader to an exchange can be computed in a number of ways. A trader who only sends FoK orders will receive fills of all volume or nothing, as opposed to IoC orders, which can be partially filled.

A measure of fill ratio is important because it summarizes how effective is an agent's strategy in completing trades. A simple measure is to compute the number of filled orders divided by the number of trade attempts. Here we propose new measures of fill ratios that include relevant post-trade information such as orders that were filled at better prices (i.e., price improvement), and orders that were not filled, but could have been filled at worse prices, had the order walked the LOB until filled in full. Below we provide more details about the post-trade information we include in the measures of fill ratio. Next, we describe the variables and notation we employ.

We denote by n the number of marketable orders employed to compute the trader's fill ratio. Let $\tau_1, \tau_2, \ldots, \tau_n$ be the times at which the trader sends liquidity taking orders to the Exchange and denote by $\ell_1, \ell_2, \ldots, \ell_n$ the times at which the orders are processed by the Exchange, $\ell_i > \tau_i$.

Marketable orders specify a limit price, denoted by L_i, which is the maximum (minimum) the trader is willing to pay (receive) per unit of bought (sold) currency pair. When the Exchange processes the trader's message, the order is filled if there is enough liquidity resting in the LOB up to, and including, the limit price L_i. A marketable buy (sell) order with limit price $L_i = \infty$ ($L_i = 0$) will walk all levels of the LOB until the order is filled in full at the average price B_i per unit of currency pair.

The quantity $L_i - B_i$ (which can be computed only after the Exchange receives and processes the order) is important because it gives post-trade information about different outcomes such as: (i) by how much did prices move to cause the liquidity taking order to miss its target, which we refer to as *potential slippage*, or (ii) by how much did the execution price improve, which we refer to as *price improvement*. For example, when the trader sends a buy order with a price limit of L_i per unit of currency pair, there is a price improvement if $L_i - B_i > 0$. On the other hand, if the average price required to fill the order is higher than limit price the trader is willing to pay, then the potential slippage is $L_i - B_i < 0$ – note that in this second case the trader does not incur the potential slippage because the order is not filled in full due to the limit price.

Moreover, the variable $I_i \in \{-1, 1\}$ is used to indicate the direction of the liquidity taking order, where 1 denotes a buy order and -1 denotes a sell order. Finally, V_i denotes

the volume of currency pair lots the trader attempts to fill, and V_i^f denotes the filled volume, clearly $V_i^f \leq V_i$.

We propose three ways to compute a fill measure of n consecutive trades. These measures are based on: (1) number of trades, (2) average volume of trades, and (3) volume of trades.

1.
$$F_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{I_i L_i \geq I_i B_i} + \mathrm{F}(n)\,, \tag{1}$$

2.
$$F_n^{AV} = \frac{1}{n}\sum_{i=1}^{n} \frac{V_i^f}{V_i} + \mathrm{F}(n)\,, \tag{2}$$

3.
$$F_n^{V} = \frac{\sum_{i=1}^{n} V^f_{\tau_i+\ell_i}}{\sum_{i=1}^{n} V_i} + \mathrm{F}(n)\,, \tag{3}$$

where $\mathbf{1}$ is the indicator function and

$$\mathrm{F}(n) = \frac{1}{n}\sum_{i=1}^{n} I_i\,(L_i - B_i)\left(\frac{1}{L_i} + \frac{1}{B_i}\right) \tag{4}$$

is a post-trade measure of price improvement and potential slippage.

Note that $\mathrm{F}(n) \in [-\infty, \infty]$, so the three measures of fill ratio we propose lie in $[-\infty, \infty]$. Equation (4) does not change if we include trade volumes.

3. Empirical Analysis

3.1. Data

We employ a set of proprietary data of transactions for T1 and T2 during the period 2 December 2016 to 30 March 2017 in the currency pair USD/JPY. Both traders are liquidity takers, i.e., do not send limit orders to provide liquidity in the LOB, during the period we study. We have the following information: side of order (buy or sell), time-stamp of when the Exchange processed the order, volume, currency pair, type of order (FoK, IoC), limit price L_i, and whether the order was filled and at what price(s). Finally, we have the LOB for the pair USD/JPY at millisecond intervals. A summary of the main variables is given in Table 1.

	Number of attempts N	**Successful filled trades** (incl. partial)	**Successful filled trades** (only full)	**Avg. vol. attempted** $\times 10^5$ USD	**Avg. filled vol.** $\times 10^5$ USD	**Order type**	F_N	F_N^{AV}	F_N^{V}
T1	116,912	106,878	106,878	4.5	4.1	FoK	90.78%	90.78%	90.54%
T2	68,129	61,440	59,303	1.2	1.0	IoC	87.04%	88.62%	79.00%

Table 1: Proprietary data and measures of fill ratios for period 2 December 2016 to 30 March 2017.

During this period, both traders send orders that aim at the best bid or best ask price that was streamed to them by the Exchange. T1 sent FoK orders with zero discretion and T2 sent IoC orders with zero discretion.

3.2. Fill ratios and volatility

The liquidity provided in the LOB is constantly being updated as a result of news, changes in liquidity providers' willingness to trade, and the arrival of liquidity taking orders, see for example Almgren (2012) for a discussion of trading with stochastic liquidity. Most of the updates of limit orders (cancellations, amendments, and arrival of new limit orders) take place in the best quotes and ticks closest to the prevailing mid-exchange rate. Traders with non-zero latency face a moving target problem when they attempt to hit bids and lift offers in the LOB. This problem is exacerbated during periods of high volatility of the exchange rate.

The top figure in the left-hand panel of Figure 1 shows the evolution of T1's fill ratio, using measure (1) with $n = 50$, during 26 January 2017. The bottom figure shows the volatility of the mid-exchange rate over the same period. Throughout this paper volatility is computed as the standard deviation of the log-returns of the micro-exchange rate (sampled every 500 milliseconds) over a 10-minute rolling window. The micro-exchange rate at time t is denoted by m_t and given by

$$m_t = \frac{P_t^a\, Q_t^b + P_t^b\, Q_t^a}{Q_t^a + Q_t^b}\,, \tag{5}$$

where Q_t^a and Q_t^b are the quantities available at the best ask price (denoted by P_t^a) and the best bid price (denoted by P_t^b), respectively.

From Figure 1, left-hand panel, we observe that there is an inverse relationship between volatility and the fill ratios achieved by T1's strategy, which always aims at the best bid or best offer price she observed before sending the marketable order. Clearly, volatility is high when the micro-exchange rate undergoes many changes and updates – this explains why the number of unfilled marketable orders (with zero discretion to walk the book) is higher when volatility is high.

The right-hand panel of Figure 1 shows a scatter plot of fill ratio and volatility. The solid line in the figure is obtained by performing an Ordinary Least Squares (OLS) regression of fill ratio on volatility.

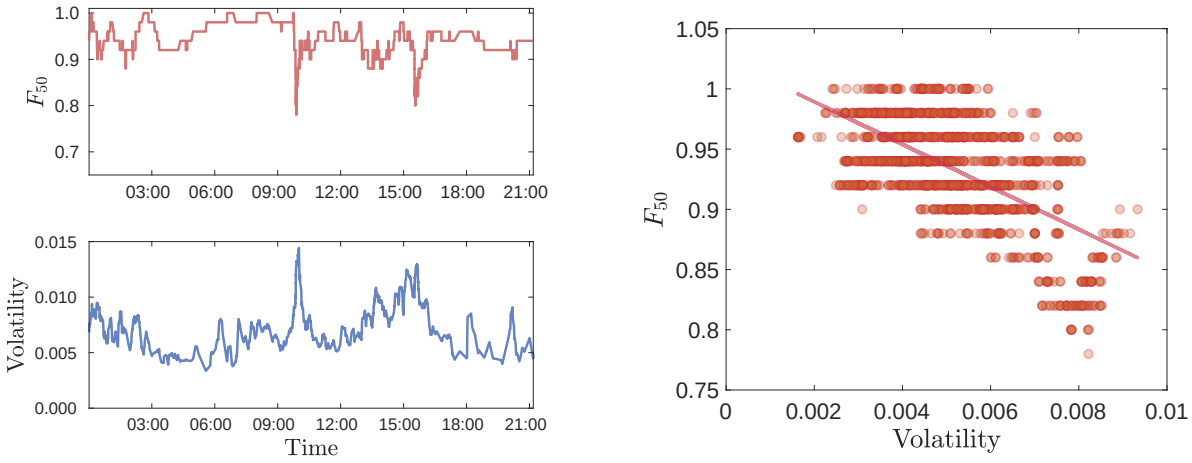

Figure 1: Left-hand top panel: fill ratio using measure (1) with $n = 50$. Left-hand bottom panel: volatility of micro-exchange rate. Right-hand panel: Scatter plot and OLS regression of volatility and fill ratio. Data are from 26 January 2017.

4. Model

In this section we describe the model of fill ratios and the dynamic optimization problem solved by the trader. We develop the model of fills in three steps. First, we present a model for the dynamics of volatility of the micro-exchange rate. Second, we propose a model of fill ratios that includes the effect of volatility of the micro-exchange rate on fills. Third, we specify how marketable orders with discretion to walk the LOB affect the dynamics of fill ratios.

Volatility of micro-exchange rate

We denote the volatility of the micro-exchange rate, see (5), by $v = (v_t)_{\{0\le t\le T\}}$ and assume it satisfies the stochastic differential equation (SDE)

$$\mathrm{d}v_t = \kappa\left(\bar{v} - v_t\right)\mathrm{d}t + \sigma_v\, v_t\, \mathrm{d}W_t^v\,, \tag{6}$$

where $\kappa > 0$ is the exponential speed at which volatility reverts to its long-term level, denoted by $\bar{v} > 0$, $\sigma_v > 0$ is a dispersion parameter (i.e., volatility of volatility), and $W^v = (W_t^v)_{\{0\le t\le T\}}$ is a standard Brownian motion.

Fill ratio: orders with no discretion to walk the LOB

The fill ratio of marketable orders (buy and sell) with no discretion to walk the LOB is denoted by $F = (F_t)_{\{0\le t\le T\}}$ and satisfies the SDE

$$\mathrm{d}F_t = \left(\lambda\left(\bar{F} - F_t^{\delta}\right) - h(v_t)\right)\mathrm{d}t + \sigma_F\,\mathrm{d}W_t^F\,. \tag{7}$$

Here, the function $h : \mathbb{R}_+ \mapsto \mathbb{R}_+$ describes the effect of the volatility of the micro-exchange rate on the fill ratio of the trader's strategy. The parameter $\lambda > 0$ is the exponential speed at which the fill rate reverts to the level $\bar{F} \in \mathbb{R}$, the dispersion parameter σ_F is a non-negative constant, and $W^F = (W_t^F)_{\{0 \le t \le T\}}$ is a standard Brownian motion uncorrelated to W^v. Recall that we propose three ways to compute fill ratios, see (1), (2), (3), so here F refers to one of those choices.

The parameters λ, $\bar{F}$, σ^F, and the function $h(v)$ are specific to each trader in the FX market. Latency and type of liquidity taking strategy are the key characteristics that make the fill ratio dynamics specific to each trader. For example, the dynamics of the fill ratio for a trader with zero latency are as in (7) with parameters $\sigma_F = 0$, $\lambda > 0$, $\bar{F} = 1$, and the function $h = 0$ (or one can also assume that $\lambda \to \infty$ and $\bar{F} = 1$, $\sigma^F > 0$).

The term $\sigma^F \, \mathrm{d}W^F$ represents the combined effect of two sources of uncertainty in the dynamics of the fill ratio, both of which depend on latency. One, the volatility of the micro-exchange rate will affect fill ratios via a deterministic component, given by $h(v)$ in (7), and an uncertain component. Two, between the time the trader is streamed quotes and the time the Exchange receives the trader's marketable order, the LOB will have processed instructions from other market participants, hence the fill of the order is uncertain.

Fill ratio: orders with discretion to walk the LOB

Liquidity takers acknowledge that however fast their computers are, or efficient their software is, or whether their hardware is co-located, the time to reach the market and execute a marketable order is not zero. A trading strategy where market orders aim at only the best bid or best ask prices in the LOB will hardly deliver fill ratios of 100%.

One alternative to compensate for not reaching the market in time is to include a discretion in the marketable order. This discretion specifies the maximum (minimum) price the trader is willing to pay (receive) when sending a marketable order with a price limit. In particular, the trader specifies a discretional amount δ, which is the additional slippage she is willing to incur relative to the best quote she observed at the time of making the decision to submit the order to the Exchange – the discretion to walk the LOB does not preclude the strategy from benefiting from price improvements. For example, if the trader sees the bid (ask) exchange rate at 1.000, then a marketable sell (buy) order with discretion $\delta = 0.001$ will be filled by the Exchange with limit buy (sell) orders resting in the LOB down (up) to a minimum (maximum) bid (offer) price of 0.999 (1.001).

When the trader includes discretion in her marketable orders she will exert upward pressure on the fill ratio of her strategy because the order can walk the LOB. Thus, we assume that the trader's fill ratio, denoted by the controlled process $F^\delta = (F_t^\delta)_{\{0 \le t \le T\}}$, follows the SDE

$$\mathrm{d}F_t^\delta = \left(\lambda\left(\bar{F} - F_t^\delta\right) + g(\delta_t) - h(v_t)\right)\mathrm{d}t + \sigma_F \, \mathrm{d}W_t^F \,, \tag{8}$$

where the function $g : \mathbb{R} \mapsto \mathbb{R}$ models the impact of discretion on the fill ratio – when $g = 0$ we obtain (7).

Assume the impact functions are of the form

$$g(\delta) = a\,\delta \qquad \text{and} \qquad h(v) = b\,(v - \bar{v})\,,$$

with a and b non-negative constants.

As usual, we work on a completed filtered probability space $(\Omega, \mathbb{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, with $\mathcal{F}_t$ the natural filtration generated by the 2−dimensional Brownian motion $\boldsymbol{W} = (W^F, W^v)$, and we refer to $\mathbb{P}$ as the trader's reference measure.

4.1. Model uncertainty

The trader acknowledges that the model of fill ratios (8) may be misspecified, so she considers alternative dynamics of the fill ratio to make the model robust to misspecification, see Cartea et al. (2017). This ambiguity about model choice, or model uncertainty, will affect the optimal strategy employed by the trader when choosing the discretion of the orders she sends to the Exchange. We incorporate the trader's ambiguity about model choice in two steps. First, we characterize alternative measures that describe the fill ratio dynamics considered by the trader, and then we determine how the trader chooses the reference measure $\mathbb{P}$ or one of the alternative measures.

The trader considers candidate measures $\mathbb{Q}(\boldsymbol{x})$ that are equivalent to $\mathbb{P}$ and characterized by the Radon-Nikodym derivative

$$\frac{\mathrm{d}\mathbb{Q}(\boldsymbol{x})}{\mathrm{d}\mathbb{P}}\Big|_t = \exp\left\{-\frac{1}{2}\int_0^t \boldsymbol{x}'_u\,\boldsymbol{x}_u\,\mathrm{d}u - \int_0^t \boldsymbol{x}_u\,\mathrm{d}\boldsymbol{W}_u\right\}, \tag{9}$$

where $\boldsymbol{x}_t = (x_t^F, 0)$ is a two-dimensional $\mathcal{F}_t$-adapted process such that $\left(\mathrm{d}\mathbb{Q}(\boldsymbol{x})/\mathrm{d}\mathbb{P}\big|_t\right)_{0\leq t\leq T}$ is a martingale. Thus, we denote by $\boldsymbol{\mathcal{Q}}$ the class of alternative measures

$$\boldsymbol{\mathcal{Q}} = \left\{\mathbb{Q}(\boldsymbol{x}) \mid \boldsymbol{x} \text{ is } \mathcal{F}-\text{adapted and } \left(\frac{\mathrm{d}\mathbb{Q}(\boldsymbol{x})}{\mathrm{d}\mathbb{P}}\Big|_t\right)_{0\leq t\leq T} \text{ is a martingale}\right\}.$$

Next, the trader penalizes deviations from the reference measure using the relative entropy from t to T, which is given by

$$\mathcal{H}_{t,T}\left(\mathbb{Q}\,|\,\mathbb{P}\right) = \frac{1}{\varphi}\log\left(\frac{\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}\,|_T}{\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}\,|_t}\right).$$

Here the parameter φ is a non-negative constant that represents the trader's degree of ambiguity aversion. If the trader is confident about the reference measure $\mathbb{P}$, then the ambiguity aversion parameter φ is small and any deviation from the reference model of fill dynamics is very costly. In the extreme $\varphi \to 0$, the trader is very confident about the reference measure, so she chooses $\mathbb{P}$ because the penalty that results from rejecting the reference measure is too high. On the other hand, if the trader is very ambiguous about the reference model, considering alternative models results in a very small penalty. In the extreme $\varphi \to \infty$, deviations from the reference model are costless, so the trader considers the worst case scenario.

4.2. Value function

The trader's value function is

$$V(t,\, f,\, v) = \sup_{\delta \in \mathcal{A}_{t,T}} \inf_{\mathbb{Q} \in \boldsymbol{\mathcal{Q}}} \mathbb{E}^{\mathbb{Q}}_{t,f,v}\Bigg[-\phi_\delta \int_t^T \left(\delta_s - \hat{\delta}\right)^2 \mathrm{d}s - \phi_F \int_t^T \left(F_s^\delta - \hat{F}\right)^2 \mathrm{d}s + \mathcal{H}_{t,T}\left(\mathbb{Q} \,|\, \mathbb{P}\right) \Bigg], \tag{10}$$

where $\hat{F}$ denotes the target fill ratio and $\hat{\delta}$ is a constant that denotes a fixed discretion to target.

Recall that the control variable δ denotes the discretion of the order to walk the LOB, which is measured from the best quote observed by the trader when receiving the quotes from the Exchange. The penalty parameters ϕ_δ and ϕ_F are non-negative constants. For large values of the parameter ϕ_F the optimal discretion will be larger the higher is the target ratio $\hat{F}$.

Finally, recall that $\boldsymbol{\mathcal{Q}}$ represents all the measures equivalent to the reference measure $\mathbb{P}$, and the set of admissible strategies is

$$\mathcal{A}_{t,T} = \Bigg\{ \delta = (\delta_t)_{\{0 \le t \le T\}} \;|\; \delta \text{ is } \mathcal{F}-\text{adapted, } \mathbb{E}\left[\int_0^T (\delta_s)^2 \,\mathrm{d}s\right] < \infty, \text{ and} \tag{11}$$

$$\delta_t = \mu\left(t,\, F_t^\delta,\, v_t\right), \text{ for } \mu \in \mathcal{M}_{0,T} \Bigg\}, \tag{12}$$

where

$$\mathcal{M}_{0,T} = \Bigg\{ \mu : [0,\, T] \times \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R} \;|\; \exists K \in \mathbb{R} \text{ s.t. } \forall \boldsymbol{x},\, \boldsymbol{y} \in \mathbb{R}^2 \text{ and } t \in [0,\, T]$$

$$|\mu(t,\boldsymbol{x}) - \mu(t,\boldsymbol{y})| \le K\,|\boldsymbol{x} - \boldsymbol{y}| \text{ and } |\mu(t,\boldsymbol{x})|^2 \le K^2\left(1 + |\boldsymbol{x}|^2\right) \Bigg\}. \tag{13}$$

We require the set of admissible strategies to satisfy the condition appearing in (12), so that the change of measure specified by (9) is valid.

By standard results, the Hamilton-Jacobi-Bellman-Isaacs (HJBI) equation associated to problem (10) is given by

$$V_t + \sup_\delta \inf_x \Bigg(\left(\lambda\left(\bar{F} - f\right) + a\,\delta - b\,(v - \bar{v}) - \sigma_F\, x\right) V_f + \frac{1}{2}\,\sigma_F^2\, V_{ff} + \frac{1}{2\,\varphi}\, x^2$$

$$+ \kappa\,(\bar{v} - v)\, V_v + \frac{1}{2}\, v^2\, \sigma_v^2\, V_{vv} - \phi_\delta\left(\delta - \hat{\delta}\right)^2 - \phi_F\left(f - \hat{F}\right)^2 \Bigg) = 0\,, \tag{14}$$

with terminal condition $V(T,\, f,\, v) = 0$. Here, subscripts in the function V represent partial derivatives, for example, $V_t = \partial V / \partial t$.

Proposition 1. *The supremum and infimum in HJBI (14) are achieved (in feedback form) at*

$$\delta^* = \hat{\delta} + \frac{a\, V_f}{2\, \phi_\delta}\,, \qquad \text{and} \qquad x^* = \varphi\, \sigma_F\, V_f\,,$$

respectively.

Proof. Apply first order conditions to the supremum and infimum terms in (14), and check second order conditions to verify that these are the maximizer and minimizer, respectively. □

Substituting the optimal controls δ^* and x^* in (14), the HJBI becomes

$$\begin{aligned} V_t + \big(\lambda(\bar{F} - f) - b\,(v - \bar{v})\big)\, V_f + \frac{1}{2}\,\sigma_F^2\, V_{ff} + \frac{a^2\, V_f^2}{2\,\phi_\delta} - \frac{1}{2}\,\varphi\, V_f^2\,\sigma_F^2 + a\, V_f\,\hat{\delta} \\ + \kappa\,(\bar{v} - v)\, V_v + \frac{1}{2}\, v^2\,\sigma_v^2\, V_{vv} - \phi_F\,\big(f - \hat{F}\big)^2 = 0\,, \qquad (15) \\ V(T,\, f,\, v) = 0\,. \end{aligned}$$

By inspection of PDE (15) and its terminal condition, we propose the ansatz

$$V(t,\, f,\, v) = h_0(t) + h_1(t)\, f + h_2(t)\, f^2 + h_3(t)\, v + h_4(t)\, v^2 + h_5(t)\, f\, v\,, \qquad (16)$$

which we substitute in equation (15) and obtain a coupled system of ordinary differential (ODEs). The system of ODEs can be solved explicitly to obtain closed-form solutions for the deterministic functions h_0, h_1, h_2, h_3, h_4, h_5, see Appendix A.1.

Theorem 1. ***(Verification)*** *Define the time-dependent deterministic functions* $g_1,\, g_2,\, g_3 : [0,\, T] \mapsto \mathbb{R}$ *as*

$$g_1(t) = h_1(t) + 2\,\hat{F}\, h_2(t) + \bar{v}\, h_5(t)\,, \qquad g_2(t) = -2\, h_2(t)\,, \qquad g_3(t) = h_5(t)\,.$$

Then, the optimal discretion

$$\delta_t^* = \hat{\delta} + \frac{a}{2\,\phi_\delta}\,\Big(g_1(t) + g_2(t)\,\big(\hat{F} - F_t^{\delta^*}\big) + g_3(t)\,\big(v_t - \bar{v}\big)\Big)\,, \qquad (17)$$

and

$$x_t^* = \varphi\,\sigma_F\,\big(h_1(t) + 2\, h_2(t)\, F_t^\delta + h_5(t)\, v_t\big)\,, \qquad (18)$$

are admissible controls and equation (16) is the trader's value function (10).

Proof. For the proof see Appendix A.2. □

We explain the intuition of the terms appearing in the optimal discretion (17). The first term on the right-hand side of (17) is the target discretion $\hat{\delta}$. For example, if the trader imposes an arbitrarily large penalty on deviations from this discretion target, i.e. penalty parameter $\phi_\delta \to \infty$, then the optimal discretion $\delta_t^* \to \hat{\delta}$.

The second term, $g_1(t)$, is the time-dependent baseline level of discretion the trader posts when her fill ratio is close to the target $\bar{F}$ and volatility is around its long-term level $\bar{v}$.

The third term, $g_2(t)\left(\hat{F} - F_t^\delta\right)$, adjusts the strategy if the fill ratio is not on target. The function $g_2(t)$ is positive for $t < T$, so when the fill ratio is below (above) its target $\hat{F}$ the discretion to walk the LOB is increased (decreased).

The last term of the optimal strategy shows how the optimal discretion depends on the volatility of the micro-exchange rate. The function $g_3(t)$ is positive for $t < T$, so when volatility is higher (lower) than its long-term level, the strategy increases (decreases) the discretion of the marketable orders.

5. Estimation of model parameters

In this section we show how to estimate the parameters of the models of volatility and fill ratio.

Volatility parameters

We discretize the continuous-time volatility model (6) and write

$$\tilde{v}_i = \tilde{v}_{i-1} + \kappa\left(\bar{v} - \tilde{v}_{i-1}\right)\Delta + \sigma_v\,\tilde{v}_{i-1}\sqrt{\Delta}\,\epsilon_i\,, \tag{19}$$

where $i \in \{1, 2, ..., N-1\}$, $0 = t_1 < t_2 < ... < t_N = T$, $T = 1$, $t_i = i\,T/N$, $\epsilon_i \sim \mathcal{N}(0, \sigma_v^2\,\Delta)$, and $\Delta = T/N$.

Assume $\{\tilde{v}_i\}_{i\in\{1,2,...,N-1\}}$ are positive and divide (19) by $\tilde{v}_{i-1}$, so that

$$\frac{\tilde{v}_i}{\tilde{v}_{i-1}} - 1 = -\kappa\,\Delta + \kappa\,\bar{v}\,\Delta\,\frac{1}{\tilde{v}_{i-1}} + \sigma_v\sqrt{\Delta}\,\epsilon_i\,,$$

which we write as the OLS problem

$$\boldsymbol{y} = \beta_0\,\mathbf{1} + \beta_1\,\boldsymbol{x} + \boldsymbol{\epsilon}\,. \tag{20}$$

Here $\boldsymbol{y}$ is a $(N-1)$–dimensional vector containing the dependent variables with entries $y_i = \tilde{v}_i/\tilde{v}_{i-1} - 1$, $\mathbf{1}$ is an $(N-1)$–dimensional vector of ones, β_0 and β_1 are scalar parameters, $\boldsymbol{x}$ is an $(N-1)$–dimensional vector containing the regressors $x_i = 1/\tilde{v}_{i-1}$, and $\boldsymbol{\epsilon}$ is the vector of residuals with i.i.d. entries $\epsilon_i \sim \mathcal{N}(0, \sigma_v^2\,\Delta)$.

To obtain parameter estimates for β_0 and β_1 we employ the volatility of the micro-exchange rate at 500 millisecond intervals (see subsection 3.2) and use a standard OLS routine to compute the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. Now, because $\beta_0 = -\kappa\,\Delta$ and $\beta_1 = \kappa\,\bar{v}\,\Delta$, we obtain the estimates

$$\hat{\kappa} = -\hat{\beta}_0/\Delta\,, \qquad \hat{\bar{v}} = -\hat{\beta}_1/\hat{\beta}_0\,, \qquad \text{and} \qquad \hat{\sigma}_v = \sigma_\epsilon/\sqrt{\Delta}\,, \tag{21}$$

where σ_ϵ is the standard deviation of the residuals in (20).

Fill ratio parameters

We discretize the continuous-time fill ratio model (8) and write

$$\tilde{F}_i - \tilde{F}_{i-1} = \lambda\,\bar{F}\,\Delta + a\,\delta_{i-1}\,\Delta - \lambda\,\Delta\,\tilde{F}_{i-1} - b\,(\tilde{v}_{i-1} - \bar{v})\,\Delta + \sigma_F\,\sqrt{\Delta}\,\eta_i\,, \tag{22}$$

where, as above, $i \in \{1, 2, ..., N-1\}$, $0 = t_1 < t_2 < ... < t_N = T$, $T = 1$, $t_i = i\,T/N$, $\eta_i \sim \mathcal{N}(0,\, \sigma_\eta^2\,\Delta)$, and $\Delta = T/N$.

We write (22) as the OLS problem

$$\boldsymbol{z} = \gamma_0\,\mathbf{1} + \gamma_1\,\boldsymbol{u} + \gamma_2\,\boldsymbol{w} + \gamma_3\,\boldsymbol{r} + \boldsymbol{\eta}\,, \tag{23}$$

where the vector $\boldsymbol{z}$ contains the dependent variables, γ_0, γ_1, γ_2, and γ_3 are scalar parameters, the vectors $\boldsymbol{u}$, $\boldsymbol{w}$, $\boldsymbol{r}$ are regressors, and $\boldsymbol{\eta}$ is the vector of residuals with i.i.d. entries $\eta_i \sim \mathcal{N}(0,\, \sigma_F^2\,\Delta)$.

To estimate the impact parameter a in (22) we compute the effect of the discretion δ on fill ratios for the set of values $\delta_t = 0, 1, 2, 3, 4, 5$, as follows. Throughout the day we assume that all FoK trades sent by T1 included a fixed discretion of $\delta = 0$, and then repeat the analysis assuming $\delta = 1$, and $\delta = 2$, and so on. This provides us with six scenarios with fixed discretion for the whole trading day. We employ LOB data to determine if the orders were filled and create six $(N-1)$–dimensional vectors $\boldsymbol{z_0}$, $\boldsymbol{z_1}$, $\boldsymbol{z_2}$, $\boldsymbol{z_3}$, $\boldsymbol{z_4}$, $\boldsymbol{z_5}$, where each vector corresponds to each scenario with fixed discretion. Next, we pool all the data so that the dependent variable on the left-hand side of the regression and the regressors on the right-hand side are $6 \times (N-1)$–dimensional vectors. We run a standard OLS routine to compute the parameter estimates $\hat{\gamma}_0$, $\hat{\gamma}_1$, $\hat{\gamma}_2$, and $\hat{\gamma}_3$. Then, the parameters of the fill ratio dynamics are

$$\hat{\lambda} = -\hat{\gamma}_1/\Delta\,, \qquad \hat{\bar{F}} = -\hat{\gamma}_1/\hat{\gamma}_0\,, \qquad \hat{b} = -\hat{\gamma}_2/\Delta\,, \qquad \hat{a} = \hat{\gamma}_3/\Delta\,, \qquad \hat{\sigma}_F = \sigma_\eta/\sqrt{\Delta}\,, \tag{24}$$

where σ_η is the standard deviation of the errors in (23). To estimate model parameters for T2 we proceed in the same way and use the definition of fill ratio (2) because T2 sends IoC orders to the Exchange.

Table 2 shows parameter estimates for T1 for the trading day starting at 22:05 of 3 January 2017 and ending at 22:00 on 4 January 2017, which for simplicity we refer to as the trading day of 4 January 2017.

Parameter	Estimate	p-value
$\hat{\bar{v}}$	0.0073	0.00
$\hat{\kappa}$	12.3887	0.00
$\hat{\sigma}_v$	1.3894	–
$\hat{\lambda}$	58.7206	0.00
$\hat{\bar{F}}$	0.9331	0.00
$\hat{\sigma}_F$	0.1944	–
$\hat{a}$	1.0027	0.00
$\hat{b}$	527.44	0.01

Table 2: Parameter estimates for T1's model, 4 January 2017.

6. Performance of strategy

In this section we show the performance of the optimal discretion strategy developed above. In subsection 6.1 we employ two days of trading data for T1 to discuss features of the strategy. In subsection 6.2 we employ data during the period December 2016 to March 2017 to study the performance of the strategy for T1 and T2. Finally, in Section 7 we use trade data to infer the latency of T1 and calculate the shadow price of latency that T1 would be willing to pay to reduce her latency in the marketplace.

6.1. Stylized features of latency-optimal discretion strategy

We showcase features of the strategy for T1. We employ parameter estimates for 4 January 2017 to compute the optimal discretion for the trading day 5 January 2007. During 5 January 2017, T1 sent 3,043 FoK orders (1,601 buys and 1,442 sells) to the Exchange. All orders were sent with a discretion of $\delta = 0$, and due to latency, T1 missed 279 trades, thus the fill ratio for the whole day was $F_{3403} = 90.83\%$.

Had T1 employed the strategy developed here, the optimal discretion in the marketable orders would have been as depicted in the top panel of Figure 2 and would have attained a fill ratio of $F^{\delta^*}_{3043} = 99.17\%$, i.e., miss 25 out of the 3,043 trade attempts. Compared to the fill ratio of the suboptimal strategy $\delta = 0$, which we refer to as the naive strategy, the latency-optimal strategy would have filled 254 more trades at an average cost of 1.74 ticks – recall that one tick in the USD/JPY pair is 10^{-3} JPY and that the naive and latency-optimal strategy receive the same price improvements. The average cost is calculated as the ratio of the cost of filling the 254 trades, in JPY, to the total volume of currency pairs of the 254 trades. The bottom panel of the same figure shows the volatility of the micro-exchange rate. In the next subsection we analyze in more detail the costs incurred by the optimal strategy to fill missed trades.

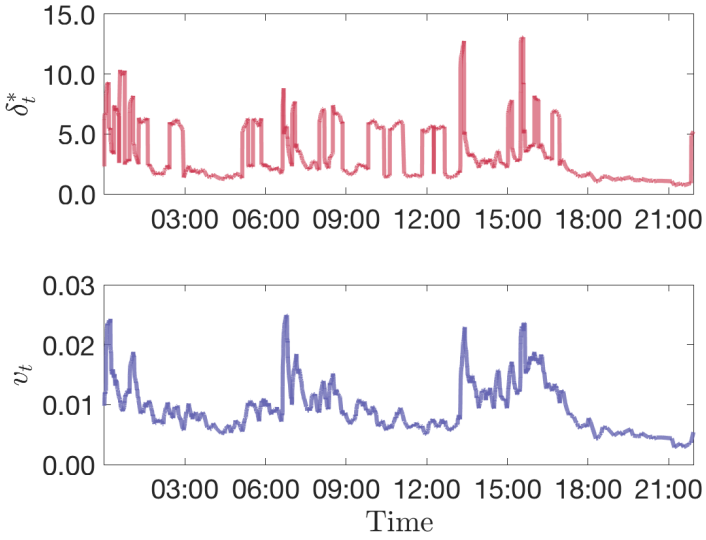

Figure 2: Parameters are $\hat{F} = 100\%$, $\phi_\delta = 5\times10^{-5}$, $\phi_F = 5$, $\varphi = 10^{-3}$, $T = 5$ days, and other model parameters are in Table 2.

The left-hand panel of Figure 3 shows a heatmap of the fill ratio $F^{\delta^*}_{3043}$ (i.e., 5 January 2017) for a range of parameters ϕ_F and ϕ_δ. We observe that when the penalty parameter ϕ_δ is low, the strategy obtains the highest fill ratios because the trader is more willing to bear the costs of walking the LOB. In the heatmap we also show, with a star, the pair $(\phi_\delta, \phi_F) = (5 \times 10^{-5}, 5)$, which are the parameters we use in the analysis of subsection 6.2.

The right-hand panel of the figure shows a heatmap of the average cost of the fill ratios attainable by the optimal strategy. The cost reported is the slippage of the order relative to T1's naive strategy.

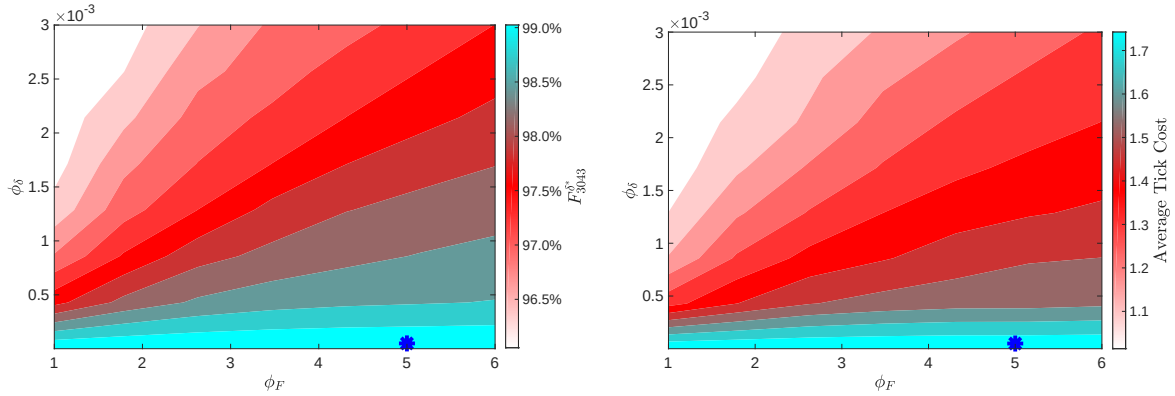

Figure 3: Fill ratio $F^{\delta^*}_{3043}$ for 5 January 2017. Parameters are $\varphi = 10^{-3}$, $\hat{\delta} = 0$, $\hat{F} = 100\%$, $T = 5$ days, and other model parameters in Table 2.

Finally, Figure 4 shows the fill ratio $F^{\delta^*}_{3043}$ for the range $\varphi \in [0, a^2/\phi_\delta\,\sigma_F^2)$ of the ambiguity aversion parameter when the trader employs the optimal discretion strategy. Recall that the

higher is φ, the less confident is the trader about the dynamics of F^{δ}. When $\varphi \to a^2/\phi_\delta\,\sigma_F^2$ the trader has very little confidence in her model of fill ratios, so the optimal strategy is to send orders with higher discretion to the Exchange – the ambiguity averse trader is more willing to bear slippage to ensure high fill ratios.

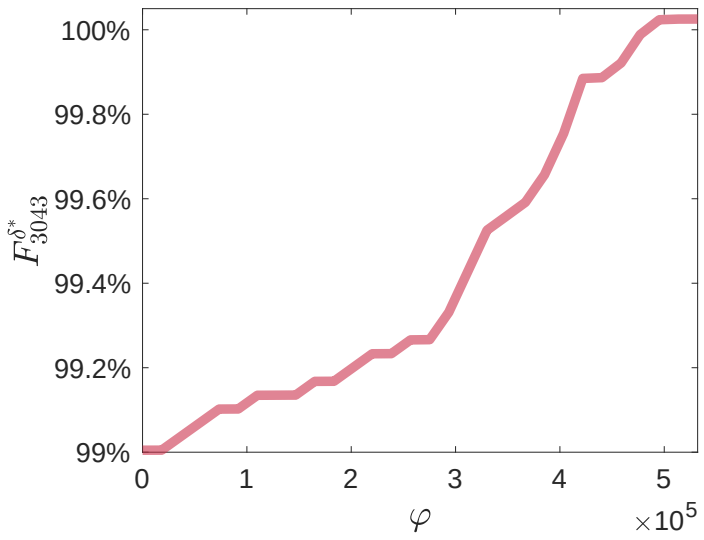

Figure 4: Fill ratio for optimal strategy. Parameters are $\phi_\delta = 5 \times 10^{-5}$, $\phi_F = 5$, $\hat{F} = 100\%$, $T = 5$ days, and other model parameters are in Table 2.

6.2. Performance of strategy for traders T1 and T2

We employ trade data for T1 and T2 between 5 December 2016 and 31 March 2017 to analyze the performance of the latency-optimal strategy. For each trading day we compute the optimal discretion (17) in Proposition 1, using the parameter estimates from the previous day and the volatility of the exchange rate (recall that the volatility is calculated over a ten-minute window). We use LOB information to determine if the orders sent with the optimal discretion δ^* would have been filled, and if so, at what price(s).

For each day we employ the first 10 minutes of the trading day (22:05 to 22:15) to compute the first observation of volatility of the micro-exchange rate. The first fill ratio observation (F_{50} for T1 and F^{AV}_{50} for T2) is computed the first time there are 50 trades after 22:15. Throughout the period we study, 4,115 trades for T1 and 4,089 for T2 are employed to compute the first observations of volatility and fill ratio for each trading day, and to calibrate the parameters of the model for the first day we employed 2,339 trades for T1 and 772 for T2. Thus, the number of trades we employ to analyze the performance of the strategy is reduced to 110,458 for T1 and 63,268 for T2.

6.2.1. Performance of latency-optimal strategy for T1

Tables 3 and 4 provide information of the costs that T1 would have incurred had she employed the latency-optimal strategy and shows the mark-to-market costs of going back to the Exchange to complete missed trades using marketable orders with discretion $\delta = \infty$.

During the period 5 December to 31 March 2017, T1 attempted 110,458 trades, all with discretion $\delta = 0$, and, due to latency, missed 9,483 trades, and T1 received 3,938,559 JPY in price improvements. Had T1 employed marketable orders with discretion $\delta = \infty$, the cost of filling the 9,483 trades would have been 8,914,046 JPY, see row (a) in Table 3. The last two columns of the same row show the mark-to-market cost, employing marketable orders with discretion $\delta = \infty$ to walk the LOB 20ms and 100ms later, of filling the 9,483 trades missed by the naive strategy – time is measured from the time-stamp recorded by the Exchange when T1's order, which could not be filled, was received. For example, if T1 were to return to the market (20ms and 100ms later) to retry the missed trade using a market order that can walk the LOB until the order is filled in full, then the costs to fill the 9,483 missed trades would have been 9,841,755 JPY and 12,158,945 JPY, respectively.

In rows (b) to (e) we show the performance of the latency-optimal strategy for discretion targets $\hat{\delta} \in \{0, 1, 2, 3\}$, respectively. For example, row (b) shows that a strategy with target discretion $\hat{\delta} = 0$ would have filled 8,318 more trades than the naive strategy. The cost of filling the 8,318 trades is 6,528,965 JPY and the mark-to-market total costs of these trades 20ms and 100ms later, using marketable orders with discretion $\delta = \infty$, are 7,459,198 JPY and 9,601,222 JPY, respectively.

Table 4 presents the costs of filling missed trades in ticks (10^{-3} JPY). Column 3 in row (a) shows that the average mark-to-market cost to fill the 9,483 trades missed by the naive strategy, employing marketable orders with $\delta = \infty$, is 2.11 ticks (i.e., $2.11 = 8,914,046/4,222 \times 10^6$). Row (b) shows that the average cost of filling 8,318 trades missed by the naive strategy, which were filled by the optimal strategy with discretion target $\hat{\delta} = 0$, would have been 1.76 ticks, and the mark-to-market cost of the 8,318 trades employing marketable orders with discretion $\delta = \infty$, which are submitted 20ms or 100ms later, is 2.01 or 2.58 ticks.

The last two columns in Table 4 show the 'excess cost' (in percentage terms relative to optimal strategy's attainable cost) of returning to the market, 20ms and 100ms later, to complete unfilled trades employing orders with discretion $\delta = \infty$. For example, the costs of filling missed trades with market orders 20ms (100ms) later are between 13% and 14% (44% and 47%) more than the cost of filling them with the optimal strategy.

The costs of returning to the market to fill missed trades employing orders with discretion $\delta = \infty$ is, on average, greater than the costs of completing the trades, on the first attempt, employing the latency-optimal strategy. Thus, not only does the optimal strategy attain a very high fill ratio, it does it at a cost which is lower than the mark-to-market value of completing the trades on a second attempt. These savings justify the extra cost of optimally walking the LOB in the first trade attempt to achieve higher fill ratios than the fill ratios achieved by the naive strategy. Note that we are not imputing other costs that the trader may incur if a trade is not filled. For example, some trades could be one leg of a trade, which if not completed, will affect other trades or strategies, see Cartea et al. (2018).

		Missed by naive, filled by alternative strategies				
		Extra Fills	**Extra Volume $\times 10^6$**	**Cost in JPY**	**Cost 20ms later in JPY**	**Cost 100ms later in JPY**
(a)	$\delta_t = \infty\,, \forall t$	9,483	4,222	8,914,046	9,841,755	12,158,945
	δ^* **Target**				$\delta_t = \infty\,, \forall t$	$\delta_t = \infty\,, \forall t$
(b)	$\hat{\delta} = 0$	8,318	3,718	6,528,965	7,459,198	9,601,222
(c)	$\hat{\delta} = 1$	8,603	3,839	6,807,416	7,770,984	9,959,191
(d)	$\hat{\delta} = 2$	8,758	3,906	6,999,804	7,950,543	10,185,240
(e)	$\hat{\delta} = 3$	8,938	3,982	7,232,545	8,181,285	10,447,450

Table 3: Performance of optimal strategy for T1. The total number of trade attempts is 110,458. Cost-related quantities are in JPY. Parameters: $\phi_\delta = 5 \times 10^{-5}$, $\phi_F = 5$, $\varphi = 10^{-3}$, $\hat{F} = 100\%$, and $T = 5$ days. Period is from 5 December 2016 to 31 March 2017, currency pair USD/JPY.

		Missed by naive, filled by alternative strategies					
		Extra Fills	**Avg. cost in ticks**	**Avg. cost 20ms later in ticks**	**Avg. cost 100ms later in ticks**	**Cost increase 20ms later %**	**Cost increase 100ms later %**
(a)	$\delta_t = \infty\,, \forall t$	9,483	2.11	2.33	2.88	10%	36%
	δ^* **Target**					$\delta_t = \infty\,, \forall t$	$\delta_t = \infty\,, \forall t$
(b)	$\hat{\delta} = 0$	8,318	1.76	2.01	2.58	14%	47%
(c)	$\hat{\delta} = 1$	8,603	1.77	2.02	2.59	14%	46%
(d)	$\hat{\delta} = 2$	8,758	1.79	2.04	2.61	14%	46%
(e)	$\hat{\delta} = 3$	8,938	1.82	2.05	2.62	13%	44%

Table 4: Performance of optimal strategy for T1. The total number of trade attempts is 110,458. Cost-related quantities are weighted by volume and expressed in ticks, 10^{-3} JPY. The metrics in column 6 (resp. 7) are the proportional increase of the values in columns 5 and 6 (resp. 5 and 7) in Table 3. Parameters: $\phi_\delta = 5 \times 10^{-5}$, $\phi_F = 5$, $\varphi = 10^{-3}$, $\hat{F} = 100\%$, and $T = 5$ days. Period is from 5 December 2016 to 31 March 2017, currency pair USD/JPY.

Table 5 shows the results of the optimal strategy arranged in volatility quartiles when the discretion target is $\hat{\delta} = 0$. As discussed above, when the micro-exchange rate is more volatile, the chances of missing trades are greater than when volatility is low because the LOB is undergoing many more updates. The optimal strategy developed here counterbalances the adverse effects of high volatility on fill ratios, so that in periods of high volatility, everything else being equal, the trader sends orders with more discretion to walk the book. The table shows that as volatility increases, the strategy maintains a stable level of fills, while T1's naive strategy misses more trades as volatility increases.

The last three columns in Table 5 show the cost of filling trades missed by T1's naive strategy, but filled by the optimal strategy with $\hat{\delta} = 0$. The first of the last three columns shows the cost incurred by the optimal strategy to execute missed trades for each volatility

quartile. The penultimate and last columns of the table show the cost of returning to the market to fill the missed trades with market orders that walk the LOB until they are filled. As discussed earlier, the cost incurred by the latency-optimal strategy to fill trades missed by the naive strategy are lower than the costs of returning to the market 20ms and 100ms later. The gap between the costs of returning to the market 20ms and 100ms later and the costs incurred by the optimal strategy widen as volatility increases.

Volatility Quartiles %	**Trades sent to Exchange**	**Fills** $\delta_t = 0, \forall t$	**%**	**Fills** $\delta_t^*, \hat{\delta} = 0$	**%**	**Missed by naive, filled by δ^* with $\hat{\delta} = 0$** **Avg. cost in ticks**	**Avg. cost 20ms later in ticks**	**Avg. cost 100ms later in ticks**
0–25	27,615	26,151	94.7%	27,274	98.8%	1.61	1.77	2.02
25–50	27,614	25,588	92.7%	27,303	98.9%	1.73	1.92	2.20
50 –75	27,615	25,086	90.8%	27,339	99.0%	1.67	1.91	2.49
75 –100	27,614	24,150	87.5%	27,377	99.1%	1.88	2.19	3.04

Table 5: Performance of strategy in volatility quartiles for T1. Parameters: $\phi_\delta = 5 \times 10^{-5}$, $\phi_F = 5$, $\varphi = 10^{-3}$, $\hat{F} = 100\%$, and $T = 5$ days. Ticks are 10^{-3} JPY. Period is from 5 December 2016 to 31 March 2017, currency pair USD/JPY.

6.2.2. Performance of optimal strategy: T2

We repeat the analysis of the performance of the strategy using T2's trades. Tables 6 and 7 are presented in the same format as Tables 3 and 4 respectively.

During the period 5 December 2016 to 31 March 2017, T2 attempted 63,268 IoC trades, all with discretion $\delta = 0$, and, due to latency, missed 7,038 trades, including partial misses. The cost of achieving better fill ratios using the optimal strategy (for various discretion targets $\hat{\delta}$) is justified by the cost of returning to the Exchange to fill missed trades 20ms or 100ms later with market orders. For example, had T2 employed the optimal strategy with discretion target $\hat{\delta} = 0$, only 755 trades would have been missed, as opposed to the 7,038 misses by the naive strategy. The average cost of filling the 6,283 trades missed by the naive strategy is 1.24 ticks, as opposed to 2.82 and 2.75 ticks, which are the mark-to-market costs of returning to the market 20ms and 100ms later.

		Missed by naive, filled by alternative strategies					
		Extra Fills (incl. partial)	**Extra Fills** (only full)	**Extra Volume ×10⁶ in JPY**	**Cost in JPY**	**Cost 20ms later in JPY**	**Cost 100ms later in JPY**
(a)	$\delta_t = \infty\,, \forall t$	7,038	7,038	1,455	2,294,999	4,628,993	4,455,516
	δ^* **Target**					$\delta_t = \infty\,, \forall t$	$\delta_t = \infty\,, \forall t$
(b)	$\hat{\delta} = 0$	6,283	6,238	1,320	1,634,888	3,727,171	3,631,491
(c)	$\hat{\delta} = 1$	6,446	6,411	1,340	1,665,638	3,742,734	3,649,011
(d)	$\hat{\delta} = 2$	6,548	6,517	1,354	1,667,995	3,801,813	3,690,360
(e)	$\hat{\delta} = 3$	6,648	6,627	1,373	1,733,090	3,903,934	3,795,341

Table 6: Performance of optimal strategy for T2. The total number of trade attempts is 63,268. Cost-related quantities in ticks (10^{-3} JPY). Parameters: $\phi_\delta = 5 \times 10^{-5}$, $\phi_F = 5$, $\varphi = 10^{-3}$, $\hat{F} = 100\%$, and $T = 5$ days. Period is from 5 December 2016 to 31 March 2017, currency pair USD/JPY.

		Missed by naive, filled by alternative strategies					
		Extra Fills (incl. partial)	**Avg. cost in ticks**	**Avg. cost 20ms later in ticks**	**Avg. cost 100ms later in ticks**	**Growth 20ms later %**	**Growth 100ms later %**
(a)	$\delta_t = \infty\,, \forall t$	7,038	1.58	3.18	3.06	102%	94%
	δ^* **Target**					$\delta_t = \infty\,, \forall t$	$\delta_t = \infty\,, \forall t$
(b)	$\hat{\delta} = 0$	6,283	1.24	2.82	2.75	128%	122%
(c)	$\hat{\delta} = 1$	6,446	1.24	2.79	2.72	125%	119%
(d)	$\hat{\delta} = 2$	6,548	1.23	2.81	2.73	128%	121%
(e)	$\hat{\delta} = 3$	6,648	1.26	2.84	2.76	125%	119%

Table 7: Performance of optimal strategy for T2. The total number of trade attempts is 63,268. Cost-related quantities are weighted by volume and expressed in ticks (10^{-3} JPY). The metrics in column 6 (resp. 7) are the growth of the values in columns 5 and 6 (resp. 5 and 7) in Table 3. Parameters: $\phi_\delta = 5 \times 10^{-5}$, $\phi_F = 5$, $\varphi = 10^{-3}$, $\hat{F} = 100\%$, and $T = 5$ days. Period is from 5 December 2016 to 31 March 2017, currency pair USD/JPY.

Table 8 shows the results of the optimal strategy arranged in volatility quartiles when the discretion target is $\hat{\delta} = 0$. The interpretation is the same as that of Table 5 for T1. The cost incurred by the latency-optimal strategy to fill the trades missed by the naive strategy are lower than the costs of returning to the market 20ms and 100ms later. The gap between the costs of returning to the market 20ms and 100ms later and the costs incurred by the optimal strategy widen as volatility increases.

Volatility Quartiles %	Trades sent to Exchange	Fills $\delta_t = 0, \forall t$ (incl. partial)	%	Fills $\delta_t^*, \hat{\delta} = 0$ (incl. partial)	%	Missed by naive, filled by δ^* with $\hat{\delta} = 0$ Avg. cost in ticks	Avg. cost 20ms later in ticks	Avg. cost 100ms later in ticks
0%–25%	15,817	14,950	94.5%	15,658	99.0%	0.94	1.90	1.87
25%–50%	15,817	14,545	92.0%	15,631	98.8%	1.12	2.72	2.68
50%–75%	15,817	14,226	89.9%	15,629	98.8%	1.25	2.75	2.92
75%–100%	15,817	13,329	84.3%	15,610	98.7%	1.86	3.51	3.31

Table 8: Performance of strategy in volatility quartiles for T2. Parameters: $\phi_\delta = 5 \times 10^{-5}$, $\phi_F = 5$, $\varphi = 10^{-3}$, $\hat{F} = 100\%$, and $T = 5$ days. Period is from 5 December 2016 to 31 March 2017, currency pair USD/JPY.

7. The cost of latency

One approach to mitigate the effects of latency on the efficacy of trading strategies is to invest on hardware and co-location services to reduce the time delays associated to the different components of the life cycle of a trade, see definition of latency in Section 2. As latency shortens, traders that aim at the best prices they observe in the LOB will be on target more often. This improvement in the marksmanship of the trader increases the fill ratio of the strategy, but also reduces the chances of receiving price improvements.

Another approach is to devise latency-optimal strategies that lessen the adverse effects of time delays, e.g., failing to complete trades. Although these strategies do not reduce time delays in the life cycle of trade, they are a substitute for cutting down latency when the objective is to increase the percentage of marketable orders that are filled.

Neither approach is free. The first approach requires the trader to pay fixed costs to operate in the marketplace with lower latency, which increases the efficacy of the marketable orders but reduces opportunities of obtaining price improvements. In the second approach, costs accrue to the strategy when marketable orders walk the LOB, while benefits accrue because the latency-optimal approach scoops price improvements when the market, due to latency, moves in the trader's interest.

In this section we compute the shadow price that T1 would be willing to pay to reduce her latency in the marketplace. We proceed as follows. i) Consider hypothetical traders who employ the same naive strategy (i.e., zero discretion to walk LOB) as that of T1, but each trader acts in the market with a different fixed latency – for simplicity we assume latency is not stochastic. ii) Compute the fill ratio achieved by each hypothetical trader's naive strategy. iii) Map levels of latency to fill ratios. iv) Find the hypothetical trader whose fill ratio is the same as that achieved by T1's latency-optimal strategy developed above and compute the price improvements of the hypothetical trader's strategy. v) Compute the difference of the latency-optimal costs of walking the LOB and its price improvements. Thus, the difference between the costs in (iv) and (v) is the shadow price of latency that T1 would be willing to pay to employ a naive strategy (no discretion to walk LOB), which obtains the same fill ratio as that of the latency-optimal strategy.

7.1. Latency and fill ratios

Recall that the attributes we have for each order include: type of trade, volume of the trade, price limit, and direction (buy or sell) of the trade. We also have the time-stamp of when the order is processed by the Exchange (this time is denoted by t_3 in our discussion of latency in Section 2 above). However, we do not know the time it took the trader to process information and make a decision before instructing the Exchange, or the time when the trader sent the orders to the Exchange – all of which affect the latency of the trader in the marketplace.

To map latency to fill ratios we devise fifty hypothetical traders who are clones of T1. The only difference between T1 and the hypothetical traders is that each clone has a different latency in the market. We denote each hypothetical trader by T_j with $j = 0, 1, 2, \cdots, 50$, and assume that the latency of T_j is j ms, i.e. $t_3 - t_0 = j$ ms. Note that although impossible for a trader to operate in the marketplace with a latency of 0ms, we include it in our list of hypothetical traders for comparison purposes.

Note that in our approach we assume that the trader employs the same strategy regardless of her latency. However, this may not be the case because if T1's latency was lower than her actual latency, her trading strategy could be different from the one she employed during the period we analyze. For example, T1 may not be attempting potentially profitable trades because she acknowledges that her latency is too long to successfully complete trades that require more speed.

We use T_{10} to illustrate how we map a latency of 10ms to a measure of fill ratio – the fill ratio of the other hypothetical traders is calculated in the same way. The fill ratio is computed as the percentage of filled orders in one trading day – this coincides with definition F_n, provided in (1), where n is the number of trades in the day. We assume that T_{10} replaces T1 in the market. Thus, T_{10} sends orders that have the same direction, volume, type (i.e., FoK), and discretion to walk the LOB, as those of T1, and assume the orders from T_{10} reach the LOB an instant before T1's trade was filled (for which we have a time-stamp), so the hypothetical trader and T1 face the same market conditions. Therefore, orders sent by T_{10} and T1 only differ in that T_{10} attempts to fill the trade at the best quote streamed by the Exchange 10ms before the Exchange receives the order.

For example, to compute the fill ratio that T_{10} would have obtained on 5 January 2017, we employ the time-stamps of the 3,043 orders sent by T1 on that day and compute the fill ratio that T_{10} would have achieved. We repeat this approach for each hypothetical trader on the same day and depict the results in the left-hand panel of Figure 5. As expected, there is a (monotonic) decreasing relationship between latency and fill ratio. In one extreme, when latency is zero, T_0 would have obtained a fill ratio of 100%. At the other extreme T_{50}'s fill ratio is 89%.

In the figure, the square marker shows the point where latency of 27ms is mapped to 90%, which is the fill ratio obtained by T1's naive strategy. The diamond marker in the upper-left corner of the figure shows the latency of a hypothetical trader who achieved a fill ratio of 99%. Note that this fill ratio corresponds (approximately) to the fill ratio T1 would have obtained ($F^{\delta^*}_{3043} = 98.73\%$) had she employed the latency-optimal strategy with targets

$\hat{\delta} = 0$ and $\hat{F} = 100\%$. Thus, we say that, on 5 January 2017, T1's implied latency is 27ms and the implied latency of the optimal strategy is 3ms.

We repeat the analysis separately for each day between 2 December 2016 and 30 March 2017. The right-hand panel of Figure 5 shows the implied latencies of T1's naive strategy (solid line) and that of the latency-optimal strategy with target $\hat{\delta} = 0$ (dotted line). The variations in 'implied' latency are due to a number of factors, including the volatility of the best quotes which, as discussed above, has a considerable effect on fill ratios. For ease of presentation the $y-$axis is capped at 50ms – there were two days when T1's implied latency was over 50ms.

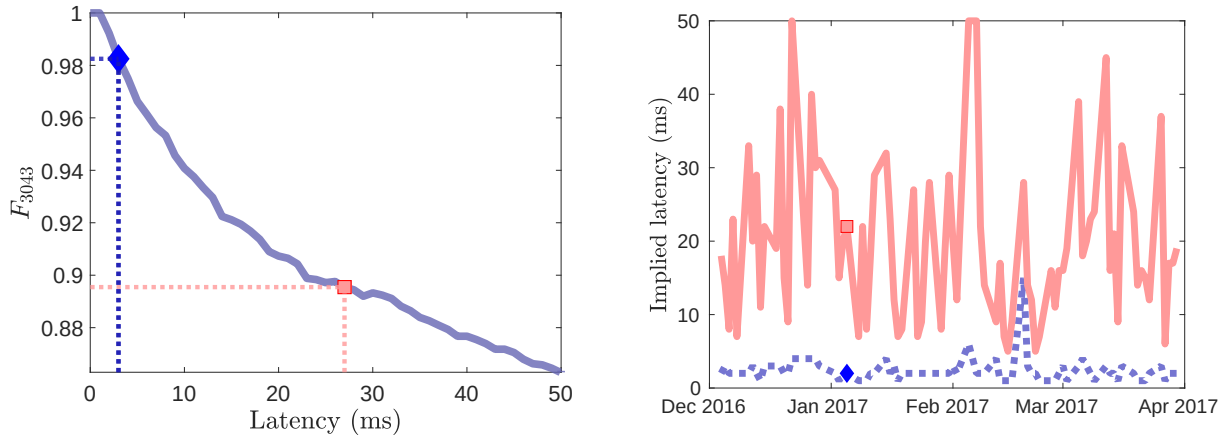

Figure 5: Left-hand panel: Mapping of latency and fill ratio employing T1's trade time-stamps and LOB information on 5 January 2017. Square marker is T1's implied latency. Diamond marker is latency-optimal implied latency. Right-hand panel: implied latency for T1 and implied latency for latency-optimal strategy employing T1's trade time-stamps and LOB information during the period 2 December 2016 to 31 March 2017.

Table 9 summarizes the results. The top panel of the tables show the mean, standard deviation, and median of the latency implied by the naive strategy and by the latency-optimal strategies with various discretion targets $\hat{\delta}$. The bottom panel provides summary statistics of the daily fill ratio obtained by the strategies. Here, the fill ratio F_N refers to the percentage of orders that were filled each day, and the table shows the mean, standard deviation, and median of the daily percentage of fills. For the latency-optimal strategy the mean fill is between 98.9% and 99.5% and for T1's naive strategy the mean fill is 91.0%. We observe that the standard deviation of the fill ratio of the optimal strategy is also considerably lower than that of the naive strategy. Therefore, the optimal strategy achieves much higher fill ratios than the naive strategy, and produces more stable results because the optimal strategy depends on the volatility of the micro-exchange rate.

	Naive strategy	Optimal Strategy $\boldsymbol{\delta}^*$ **Target** $\hat{\delta} = 0$	$\hat{\delta} = 1$	$\hat{\delta} = 2$	$\hat{\delta} = 3$	$\boldsymbol{\delta_t = \infty}$ $\forall t$
Latency						
mean	20.02	2.34	2.01	1.70	1.53	0
std	11.31	1.72	1.57	0.89	0.80	0
median	18.00	2.00	2.00	1.00	1.00	0
$\boldsymbol{F_N}$						
mean	91.9%	98.9%	99.2%	99.3%	99.5%	100%
std	2.9%	0.4%	0.4%	0.4%	0.3%	0%
median	92.2%	98.9%	99.2%	99.4%	99.5%	100%

Table 9: Shadow price of latency for T1. Period if 2 December 2016 to 30 March 2017, currency pair USD/JPY.

7.2. Shadow price of latency

The shadow price of latency is the price that T1 would be willing to pay to reduce her latency, so that the naive and the latency-optimal strategy obtain the same fill ratio. To compute this price we need: Latency-optimal cost of walking the LOB and price improvements, (ii) price improvements received by the hypothetical trader with latency equal to that implied by the latency-optimal strategy.

For example, on 5 January 2017, the latency-optimal strategy incurs 92,771 JPY in costs from walking the LOB and receives 69,305 JPY in price improvements. On the other hand, the naive strategy of the hypothetical trader with latency 3ms, receives 7,592 JPY in price improvements. Thus, on 5 January 2017, T1 would be willing to pay 31,058 JPY to reduce her latency from 22ms to 3ms. We repeat the analysis for all trading days between 2 December 2016 and 30 March 2017 and compute T1's shadow price of latency for each day. Table 10 contains summary statistics of the results.

The table shows that T1 would be willing to pay approximately $3{\cdot}10^3$ JPY (around 30,000 USD) to reduce her latency from 20.02ms to 2.34ms. The upper bound of approximately $5 \cdot 10^3$ JPY (around 50,000 USD) for the shadow price of latency is given by the last column and last row of the table, where we show the costs of always employing a marketable order with infinite discretion. These shadow prices are, according to market participants, lower than co-location, hardware, and other costs that traders would have to incur to reduce latency in the life cycle of a trade.

	Naive strategy	Optimal Strategy $\boldsymbol{\delta}^*$ **Target** $\hat{\delta} = 0$	$\hat{\delta} = 1$	$\hat{\delta} = 2$	$\hat{\delta} = 3$	$\boldsymbol{\delta_t = \infty}$ $\forall t$
Implied Latency (ms)	**20.02**	**2.34**	**2.01**	**1.70**	**1.53**	**0**
Walk LOB Cost	0	6,528,965	6,807,416	6,999,804	7,232,545	8,914,046
Price Improvement	3,938,559	3,938,559	3,938,559	3,938,559	3,938,559	3,938,559
Price Improvement with Reduced Latency	3,938,559	384,035	291,070	179,875	135,744	0
Shadow Price to Reduce Latency	0	2,974,441	3,159,927	3,241,120	3,429,730	4,975,487

Table 10: Shadow price to reduce latency reported in JPY. Period if 2 December 2016 to 30 March 2017, currency pair USD/JPY.

8. Conclusions

We employed a set of proprietary data of FX transactions to show how latency affects the percentage of liquidity taking orders filled by a trader's strategy. We analyzed the trades of two liquidity takers over a four-month period and showed that approximately 10% of their orders were not filled due to latency.

To mitigate the adverse effects of missing trades due to latency of liquidity takers, we proposed a model where the trader chooses the discretion of the orders to walk the LOB to increase the percentage of marketable orders that are filled. The optimal discretion for each liquidity taking order balances the tradeoff between the costs from walking the LOB and how far is the fill ratio from the trader's desired target. The trader's optimal strategy specifies the discretion for each transaction depending on the proportion of orders that have been filled, how far is the strategy from the target fill ratio, the cost of walking the LOB, and the volatility of the exchange rate. For example, everything else being equal, the optimal discretion to walk the LOB is higher (resp. lower) when the volatility of the exchange rate is above (resp. below) its long-term level.

We showed the performance of the latency-optimal strategy using trade data for two FX traders. We computed the costs incurred by the latency-optimal strategy to increase fill ratios and showed that the costs of the strategy are lower than the costs of returning to the market 20ms or 100ms later to fill a missed trade. Moreover, we showed that the latency-optimal strategy obtains similar fill ratios for different levels of volatility. This is in contrast to the performance of the naive strategy where the majority of the missed trades occur when the volatility of the exchange rate is high.

We employed the proprietary data set to build a function that maps latency of traders to fill ratios. We use this mapping to compute the shadow price of latency that traders are willing to pay to reduce latency in the marketplace. We computed the shadow price a particular trader is willing to pay to reduce her latency by investing in hardware, co-location, and other services. Our results showed that the trader would be better off employing the

latency-optimal strategy developed here, instead of investing in hardware and co-location services to reduce latency.

References

Almgren, R., 2012. Optimal trading with stochastic liquidity and volatility. SIAM Journal on Financial Mathematics 3 (1), 163–181.

Barger, W., Lorig, M., 2018. Optimal liquidation under stochastic price impact. arXiv preprint arXiv:1804.04170.

Bayraktar, E., Ludkovski, M., 2010. Optimal trade execution in illiquid markets. Mathematical Finance 21 (4), 681–701.

Biais, B., Foucault, T., Moinas, S., 2015. Equilibrium fast trading. Journal of Financial Economics 116 (2), 292 – 313.
URL http://www.sciencedirect.com/science/article/pii/S0304405X15000288

Cartea, Á., Donnelly, R., Jaimungal, S., 2017. Algorithmic trading with model uncertainty. SIAM Journal on Financial Mathematics 8 (1), 635–671.
URL https://doi.org/10.1137/16M106282X

Cartea, Á., Jaimungal, S., Penalva, J., 2015. Algorithmic and High-Frequency Trading. Cambridge University Press.

Cartea, Á., Jaimungal, S., Walton, J., 2018. Foreign exchange markets with Last Look. Mathematics and Financial Economics (forthcoming).

Cartea, Á., Penalva, J., 2012. Where is the value in high frequency trading? Quarterly Journal of Finance 02 (03), 1250014.

Cheridito, P., Filipović, D., Kimmel, R. L., 2007. Market price of risk specifications for affine models: Theory and evidence. Journal of Financial Economics 83 (1), 123 – 170.
URL http://www.sciencedirect.com/science/article/pii/S0304405X06001152

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

Fricke, D., Gerig, A., 2018. Too fast or too slow? determining the optimal speed of financial markets. Quantitative Finance 18 (4), 519–532.

Guéant, O., 2016. The financial mathematics of market liquidity: From optimal execution to market making. Vol. 33. CRC Press.

Hoffmann, P., 2014. A dynamic limit order market with fast and slow traders. Journal of Financial Economics 113 (1), 156–169.

Lehalle, C.-A., Mounjid, O., 2017. Limit order strategic placement with adverse selection risk and the role of latency. Market Microstructure and Liquidity 03 (01), 1750009.

Moallemi, C. C., Saĝlam, M., 2013. The cost of latency in high-frequency trading. Operations Research 61 (5), 1070–1086.

Stoikov, S., Waeber, R., 2016. Reducing transaction costs with low-latency trading algorithms. Quantitative Finance 16 (9), 1445–1451.

Appendix A. Proofs

Appendix A.1. System of ODEs satisfied by ansatz

Insert ansatz (16) into HJBI (15) and collect terms to obtain the following coupled system of six ODEs:

$$h_0'(t) + (b\,\bar{v} + a\,\hat{\delta} + \lambda\,\bar{F})\,h_1(t) + \sigma_F^2\,h_2(t) + \kappa\,\bar{v}\,h_3(t) + \frac{1}{2}\left(\frac{a^2}{\phi_\delta} - \varphi\,\sigma_F^2\right)(h_1(t))^2 - \phi_F\,\hat{F}^2 = 0\,, \tag{A.1}$$

$$h_1'(t) + (2\,b\,\bar{v} + 2\,\lambda\,\bar{F} + 2\,a\,\hat{\delta})\,h_2(t) - \lambda\,h_1(t) + \kappa\,\bar{v}\,h_5(t) + 2\left(\frac{a^2}{\phi_\delta} - \varphi\sigma_F^2\right)h_1(t)\,h_2(t) + 2\,\phi_F\,\hat{F} = 0\,, \tag{A.2}$$

$$h_2'(t) - 2\,\lambda\,h_2(t) + 2\left(\frac{a^2}{\phi_\delta} - \varphi\,\sigma_F^2\right)(h_2(t))^2 - \phi_F = 0\,, \tag{A.3}$$

$$h_3'(t) - b\,h_1(t) + (b\,\bar{v} + a\,\hat{\delta} + \lambda\,\bar{F})\,h_5(t) + 2\,\kappa\,\bar{v}\,h_4(t) - \kappa\,h_3(t) + \left(\frac{a^2}{\phi_\delta} - \varphi\,\sigma_F^2\right)h_1(t)\,h_5(t) = 0\,, \tag{A.4}$$

$$h_4'(t) - b\,h_5(t) + (\sigma_v^2 - 2\,\kappa)\,h_4(t) + \frac{1}{2}\left(\frac{a^2}{\phi_\delta} - \varphi\,\sigma_F^2\right)(h_5(t))^2 = 0\,, \tag{A.5}$$

$$h_5'(t) - 2\,b\,h_2(t) - (\lambda + \kappa)\,h_5(t) + 2\left(\frac{a^2}{\phi_\delta} - \varphi\,\sigma_F^2\right)h_2(t)\,h_5(t) = 0\,, \tag{A.6}$$

with terminal conditions

$$h_0(T) = 0\,,\; h_1(T) = 0\,,\; h_2(T) = 0\,,\; h_3(T) = 0\,,\; h_4(T) = 0\,,\; h_5(T) = 0\,.$$

Here the notation $h'(t)$ is short-hand for $\mathrm{d}h(t)/\mathrm{d}t$. Next we provide solutions for h_2, h_5, and h_1, which are the functions required to write the optimal control in closed-form.

Explicit solution for h_2.

The function h_2 satisfies the Riccati ODE in (A.3) with terminal condition $h_2(T) = 0$ and its solution depends on the sign of the coefficient $\varphi\,\sigma_F^2 - a^2/\phi_\delta$. Therefore:

- Case 1: $\varphi < a^2/\sigma_F^2\,\phi_\delta$, then

$$h_2(t) = c_1\,\tanh\left(c_2\,t + c_3\right) + c_4\,,$$

with

$$\begin{aligned} \alpha_1 &= \frac{a^2}{\phi_\delta} - \varphi\,\sigma_F^2\,, \quad c_1 = \frac{c_2}{2\,\alpha_1}\,, \quad c_2 = -\sqrt{2\,\alpha_1\,\phi_F + \lambda^2}\,, \\ c_3 &= \operatorname{atanh}\left(-\frac{\lambda}{c_2}\right) + T\,\sqrt{2\,\alpha_1\,\phi_F + \lambda^2}\,, \quad c_4 = \frac{\lambda}{2\,\alpha_1}\,. \end{aligned}$$

- Case 2: $\varphi = a^2/\sigma_F^2\,\phi_\delta$, then

$$h_2(t) = \frac{\phi_F}{2\,\lambda}\left(1 - e^{-2\,\lambda\,(T-t)}\right).$$

- Case 3: $\varphi > a^2/\sigma_F^2\,\phi_\delta$, then

$$h_2(t) = \tilde{c}_1\,\tan(\tilde{c}_2\,t + \tilde{c}_3) + \tilde{c}_4\,,$$

with

$$\begin{aligned}\tilde{\alpha}_1 &= \varphi\,\sigma_F^2 - \frac{a^2}{\phi_\delta}\,, \quad \tilde{c}_1 = \frac{c_2}{2\,\tilde{\alpha}_1}\,, \quad \tilde{c}_2 = -\sqrt{2\,\tilde{\alpha}_1\,\phi_F - \lambda^2}\,,\\ \tilde{c}_3 &= -\operatorname{atanh}\left(-\frac{\lambda}{\tilde{c}_2}\right) + T\,\sqrt{2\,\tilde{\alpha}_1\,\phi_F - \lambda^2}\,, \quad \tilde{c}_4 = \frac{\lambda}{2\,\tilde{\alpha}_1}\,.\end{aligned}$$

From now on we focus on Case 1 to obtain closed-form formulae for h_5 and h_1; Cases 2 and 3 are similar.

Explicit solution for h_5.

We use the integrating-factor technique to write

$$h_5(t) = -e^{\int_t^T (2\,\alpha_1\,h_2(u) - \kappa - \lambda)\,\mathrm{d}u} \int_t^T 2\,b\,h_2(s)\,e^{-\int_s^T (2\,\alpha_1\,h_2(u) - \kappa - \lambda)\,\mathrm{d}u}\,\mathrm{d}s\,,$$

which can be explicitly written as

$$\begin{aligned}h_5(t) &= -\frac{e^{-\kappa\,(T-t)}}{\cosh\left(t\,c_2 + c_3\right)}\\ &\quad\times\left[\left(\frac{e^{d_2(T-t)+d_3} - e^{d_3}}{d_2}\right) b\,(c_1 + c_4) + \left(\frac{e^{d_1(T-t)-d_3} - e^{-d_3}}{d_1}\right) b\,(c_4 - c_1)\right],\end{aligned}$$

where the constants $d_1 = \kappa + c_2$, $d_2 = \kappa - c_2$, and $d_3 = c_3 + T\,c_2$.

Explicit solution for h_1.

Similarly, we write

$$h_1(t) = e^{\int_t^T 2\,\alpha_1\,h_2(u) - \lambda\,\mathrm{d}u} \int_t^T \left(\kappa\,\bar{v}\,h_5(s) + e_1\,h_2(s) + e_2\right) e^{-\int_s^T 2\,\alpha_1\,h_2(u) - \lambda\,\mathrm{d}u}\,\mathrm{d}s\,,$$

where the constants $e_1 = 2\,b\,\bar{v} + 2\,\lambda\,\bar{F} + 2\,a\,\hat{\delta}$ and $e_2 = 2\,\phi_F\,\hat{F}$. The integrals can be solved explicitly to obtain

$$\begin{aligned}h_1(t) &= \frac{-\kappa\,\bar{v}}{\cosh(c_2\,t + c_3)}\\ &\quad\times\left[\frac{b\,(c_1 + c_4)\,e^{d_3}}{d_2}\left(\frac{e^{-\kappa\,(T-t)} - 1}{\kappa} - \frac{e^{-c_2\,(T-t)} - 1}{c_2}\right)\right.\\ &\qquad\left. + \frac{b\,(c_4 - c_1)\,e^{-d_3}}{d_1}\left(\frac{e^{c_2\,(T-t)} - 1}{c_2} + \frac{e^{-\kappa\,(T-t)} - 1}{\kappa}\right)\right]\\ &\quad + \frac{e_1\,c_1}{c_2\,\cosh(c_2\,t + c_3)}\left(\cosh(c_2\,T + c_3) - \cosh(c_2\,t + c_3)\right)\\ &\quad + \frac{e_1\,c_4 + e_2}{c_2\,\cosh(c_2\,t + c_3)}\left(\sinh(c_2\,T + c_3) - \sinh(c_2\,t + c_3)\right).\end{aligned}$$

Finally, the functions h_0 and h_3 are obtained explicitly in a similar way. Note that to obtain the optimal controls explicitly we only require h_1, h_2, h_5.

Appendix A.2. Proof of Theorem 1.

We first present the reader with a class of controls that produce strong solutions for F^δ and v, which is explicitly written in Lemma 1. We use Lemma 1 to prove admissibility of δ^*, which is presented in Corollary 1, then we prove that the control x^* is also admissible, and finally, conclude with the verification of the value function.

Lemma 1. *Let $\delta_t = \mu(t, F_t, v_t)$ with $\mu \in \mathcal{M}_{0,T}$. Then, for all $t \in [0, T]$ there exist strong solutions to the SDEs of F^δ and v.*

Proof. Let $\delta_t = \mu(t, F_t, v_t)$, for $\mu \in \mathcal{M}_{0,T}$. Thus, there exists $K \in \mathbb{R}$ such that for all $t \in [0,T]$, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2$

$$|\mu(t,\boldsymbol{x}) - \mu(t,\boldsymbol{y})| \le K\,|\boldsymbol{x}-\boldsymbol{y}|\,, \qquad \text{and} \qquad |\mu(t,\boldsymbol{x})|^2 \le K^2\left(1+|\boldsymbol{x}|^2\right). \tag{A.7}$$

Let $\boldsymbol{Z}_t^\delta = (F_t^\delta, v_t)'$, thus

$$\mathrm{d}\boldsymbol{Z}_t^\delta = \left(A(t) + B(t)\,\boldsymbol{Z}_t^\delta + \boldsymbol{\mu}(t,\boldsymbol{Z}_t^\delta)\right)\mathrm{d}t + C_t\,\mathrm{d}\boldsymbol{W}_t\,,$$

with $A(t)$, $B(t)$, C_t, and $\boldsymbol{\mu}$ given by

$$A(t) = \begin{bmatrix} \lambda\,\bar{F} + b\,\bar{v} \\ \kappa\,\bar{v} \end{bmatrix}, \quad B(t) = \begin{bmatrix} -\lambda & -b \\ 0 & -\kappa \end{bmatrix}, \quad C_t = \begin{bmatrix} \sigma_F & 0 \\ 0 & \sigma_v\,v_t \end{bmatrix}, \quad \boldsymbol{\mu}(t,\boldsymbol{Z}_t^\delta) = \begin{bmatrix} a\,\mu(t,\boldsymbol{Z}_t^\delta) \\ 0 \end{bmatrix}.$$

Define the functions $c : [0,T]\times\mathbb{R}^2 \to \mathbb{R}^2$, and $\sigma : [0,T]\times\mathbb{R}^2 \to \mathbb{R}^{2\times 2}$, which are given by

$$c(t,\boldsymbol{Z}) = \left(A(t) + B(t)\,\boldsymbol{Z} + \mu(t,\boldsymbol{Z})\right), \qquad \sigma(t,\boldsymbol{Z}) = \begin{bmatrix} \sigma_F & 0 \\ 0 & \sigma_v\,v_t \end{bmatrix}.$$

Now, let $t \in [0,T]$, $\boldsymbol{Z_1}$, $\boldsymbol{Z_2} \in \mathbb{R}^2$. We first verify the Lipschitz condition on c,

$$\begin{aligned} |c(t,\boldsymbol{Z_1}) - c(t,\boldsymbol{Z_2})| &= |B(t)\,(\boldsymbol{Z_1}-\boldsymbol{Z_2}) + \mu(t,\boldsymbol{Z_1}) - \mu(t,\boldsymbol{Z_2})| \\ &\le |B(t)|\,|\boldsymbol{Z_1}-\boldsymbol{Z_2}| + |\mu(t,\boldsymbol{Z_1}) - \mu(t,\boldsymbol{Z_2}| \\ &\le \bar{B}\,|\boldsymbol{Z_1}-\boldsymbol{Z_2}| + K\,|\boldsymbol{Z_1}-\boldsymbol{Z_2}| = \hat{K}\,|\boldsymbol{Z_1}-\boldsymbol{Z_2}|\,, \end{aligned}$$

for $\hat{K} = K + \bar{B}$, and $\bar{B} = \max_{t\in[0,T]} B(t)$, which is clearly attained. For σ, we have

$$|\sigma(t,\boldsymbol{Z_1}) - \sigma(t,\boldsymbol{Z_2})| \le \sigma_v\,|\boldsymbol{Z_1}-\boldsymbol{Z_2}|\,.$$

We now test the linear growth condition on c, we use triangle inequalities and simple bounds to obtain

$$\begin{aligned} |c(t,\boldsymbol{Z_1})|^2 &= |A(t) + B(t)\,\boldsymbol{Z_1} + \mu(t,\boldsymbol{Z_1})|^2 \le \left(|A(t)| + |B(t)\,\boldsymbol{Z_1}| + |\mu(t,\boldsymbol{Z_1})|\right)^2 \\ &\le 3\,|A(t)|^2 + 3\,|B(t)|^2\,|\boldsymbol{Z_1}|^2 + 3\,|\mu(t,\boldsymbol{Z_1})|^2 \\ &\le 3\,\bar{A}^2 + 3\,\bar{B}^2\,|\boldsymbol{Z_1}|^2 + 3\,K^2\left(1+|\boldsymbol{Z_1}|^2\right) \\ &\le \bar{D}^2 + \bar{D}^2\,|\boldsymbol{Z_1}|^2 + 3\,K^2\left(1+|\boldsymbol{Z_1}|^2\right) \\ &= \left(\bar{D}^2 + 3\,K^2\right)\left(1+|\boldsymbol{Z_1}|^2\right), \end{aligned}$$

for $\bar{A} = \max_{t\in[0,T]} A(t)$, which is clearly attained. Finally, we take $\bar{D}^2 = \max\{\bar{A}^2, \bar{B}^2\}$.

On the other hand, for σ, we have

$$\begin{aligned}|\sigma(t,\, \boldsymbol{Z_1})|^2 &\leq 2\,\sigma_F^2 + 2\,\sigma_v^2\,|\boldsymbol{Z_1}|^2\\ &\leq \bar{E}^2\,\left(1 + |\boldsymbol{Z_1}|^2\right),\end{aligned}$$

for $\bar{E} = \max\{\sigma_F^2, \sigma_v^2\}$. Lipschitz and linear growth conditions follow by choosing

$$\bar{K} = \max\left\{\bar{E},\, \sqrt{\bar{D}^2 + 3\,K^2},\, \sigma_v\,, \hat{K}\right\} < \infty\,.$$

It is then proved that the Lipschitz and linear growth conditions are satisfied, therefore, for all $t \in [0, T]$, there exits a strong solution to the SDE satisfied by $\boldsymbol{Z}^\delta$, and subsequently F^δ and v_t. □

Corollary 1. *Let $\delta \in \mathcal{A}_{t,T}$. Then, for all $t \in [0, T]$ there exist strong solutions to the SDEs of F^δ and v.*

This straightforward implication proves that as long as the control δ is admissible, F^δ and v have strong solutions.

Proof. Let $\delta \in \mathcal{A}_{t,T}$, this implies that there exist $\mu \in \mathcal{M}_{0,T}$ such that $\delta_t = \mu\,(t,\, F_t,\, v_t)$. By Lemma 1, for all $t \in [0, T]$ there exist strong solutions to the SDEs of F^δ and v. □

Proof. **The control δ^* is admissible .**

Recall that δ^* is given by

$$\delta_t^* = \hat{\delta} + \frac{a}{2\,\phi_\delta}\,\left(h_1(t) + 2\,h_2(t)\,F_t^{\delta^*} + h_5(t)\,v_t\right).$$

Define $f : [0, T] \times \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ as

$$f(t, x, y) = \hat{\delta} + \frac{a}{2\,\phi_\delta}\,h_1(t) + \frac{a\,h_2(t)}{\phi_\delta}\,x + \frac{a\,h_5(t)}{2\,\phi_\delta}\,y\,,$$

we see that $\delta_t^* = f\,(t, F_t, v_t)$, and, since f is linear in x and y, and a continuous function of t, then $f \in \mathcal{M}_{0,T}$. The control δ^* is $\mathcal{F}$-adapted, because it is a continuous function of two $\mathcal{F}$-adapted processes. Then, we only need to prove that

$$\mathbb{E}\left[\int_0^T (\delta_s^*)^2\,\mathrm{d}s\right] < \infty\,.$$

Based on Young's inequality we find the following bound for the optimal control:

$$(\delta_t^*)^2 \;\leq\; 2\left(\hat{\delta} + \frac{a}{2\,\phi_\delta}\,h_1(t)\right)^2 + \frac{2\,a^2}{\phi_\delta^2}\,(h_2(t))^2\left(F_t^{\delta^*}\right)^2 + \frac{a^2}{2\,\phi_\delta^2}\,(h_5(t))^2\,(v_t)^2\,,$$

and because h_1, h_2 and h_5 are continuous functions in the closed interval $[0, T]$, they are bounded. Let U be an upper bound for $|h_1|$, $|h_2|$, and $|h_5|$ on $[0, T]$.

As $\delta_t^* = f(t, F_t, v_t)$ for $f \in \mathcal{M}_{0,T}$, by Lemma 1, we have that F^{δ^*} and v have strong solutions for all $t \in [0, T]$, thus

$$\mathbb{E}\left[\sup_{0\le t\le T}\left(F_t^{\delta^*}\right)^2\right] < \infty\,, \qquad \text{and} \qquad \mathbb{E}\left[\sup_{0\le t\le T}(v_t)^2\right] < \infty\,. \tag{A.8}$$

Finally, using the bound U and the inequalities in (A.8), we conclude that

$$\begin{aligned}
\mathbb{E}\left[\int_0^T (\delta_s^*)^2\,\mathrm{d}s\right] &\le \mathbb{E}\left[\int_0^T 2\left(\hat{\delta} + \frac{a}{2\,\phi_\delta}\,h_1(t)\right)^2 + \frac{2\,a^2}{\phi_\delta^2}\,(h_2(t))^2\left(F_t^{\delta^*}\right)^2 + \frac{a^2}{2\,\phi_\delta^2}\,(h_5(t))^2\,(v_t)^2\right] \\
&\le \mathbb{E}\left[2\,T\left(\hat{\delta} + \frac{a}{2\,\phi_\delta}\,U\right)^2 + \frac{2\,a^2}{\phi_\delta^2}\,(U)^2 \sup_{0\le t\le T}\left(F_t^{\delta^*}\right)^2 + T\,\frac{a^2}{2\,\phi_\delta^2}\,(U)^2 \sup_{0\le t\le T}(v_t)^2\right] \\
&< \infty\,.
\end{aligned}$$

□

Proof. **The control x^* is admissible.**

This proof is similar to that of Theorem 1 in Cheridito et al. (2007).

Let $\delta \in \mathcal{A}_{t,T}$, then $\exists\,\mu \in \mathcal{M}_{t,T}$, such that $\delta_t = \mu(t, F_t, v_t)$. Recall that the optimal control x^* is given by

$$\begin{aligned}
x_t^* &= \varphi\,\sigma_F\left(h_1(t) + 2\,h_2(t)\,F_t^\delta + h_5(t)\,v_t\right) \\
&= \eta(t, F_t, v_t)\,,
\end{aligned}$$

which is adapted because it is a continuous function of adapted processes, so only remains to show that

$$Z_t = \exp\left\{-\frac{1}{2}\int_0^t (x_u^*)^2\,\mathrm{d}u - \int_0^t x_u^*\,\mathrm{d}W_u^F\right\},$$

is a martingale. As the control x^* is a well-defined continuous process, Z is a well-defined local martingale with respect to $\mathbb{P}$, and subsequently, because Z is non-negative, it follows that it is a super-martingale, and for it to be a martingale, we only need to prove

$$\mathbb{E}^{\mathbb{P}}[Z_T] = 1\,.$$

Consider the function $\mu^{\mathbb{Q}}$ given by

$$\mu^{\mathbb{Q}}(t, F_t, v_t) = \mu(t, F_t, v_t) - \frac{\sigma_F}{a}\,\eta(t, F_t, v_t)\,,$$

and since the term $\eta(t, F_t, v_t)$ is continuous in time and linear in F and v, it belongs to $\mathcal{M}_{t,T}$, therefore $\mu^{\mathbb{Q}} \in \mathcal{M}_{t,T}$.

Given that $\mu,\ \mu^{\mathbb{Q}} \in \mathcal{M}_{t,T}$, by Lemma 1, we know there is a strong solution for the SDEs given by

$$\mathrm{d}F_t = \left(\lambda\left(\bar{F} - F_t\right) + a\,\mu(t, F_t, v_t) - b\,(v_t - \bar{v})\right)\mathrm{d}t + \sigma_F\,\mathrm{d}W_t^F\,, \tag{A.9}$$

$$\mathrm{d}\tilde{F}_t = \left(\lambda\left(\bar{F} - \tilde{F}_t\right) + a\,\mu^{\mathbb{Q}}(t, \tilde{F}_t, v_t) - b\,(v_t - \bar{v})\right)\mathrm{d}t + \sigma_F\,\mathrm{d}W_t^F\,. \tag{A.10}$$

Now define

$$\tilde{x}_t^* = \varphi\,\sigma_F\left(h_1(t) + 2\,h_2(t)\,\tilde{F}_t^{\delta} + h_5(t)\,v_t\right) \tag{A.11}$$

$$= \eta(t, \tilde{F}_t, v_t)\,, \tag{A.12}$$

and, for $n \in \mathbb{N}$, consider the stopping times τ_n, and $\tilde{\tau}_n$

$$\tau_n = \inf\{t > 0 \mid x_t^* \geq n\} \wedge T\,, \tag{A.13}$$

$$\tilde{\tau}_n = \inf\{t > 0 \mid \tilde{x}_t^* \geq n\} \wedge T\,, \tag{A.14}$$

and since x^* and $\tilde{x}^*$ are $\mathbb{P}$-a.s. finite, then

$$\lim_{n\to\infty} \mathbb{P}\left(\tau_n = T\right) = \lim_{n\to\infty} \mathbb{P}\left(\tilde{\tau}_n = T\right) = 1\,. \tag{A.15}$$

Define for each $n \in \mathbb{N}$ the process

$$x_t^{(n)} = x_t^*\,\mathbf{1}_{\{t\leq\tau_n\}}\,,$$

which satisfies the Novikov's condition

$$\mathbb{E}^{\mathbb{P}}\left[\exp\left\{\tfrac{1}{2}\int_0^T (x_u^n)^2\,du\right\}\right] \leq \exp\left\{\frac{n^2\,T}{2}\right\} < \infty\,.$$

Then it follows that

$$\begin{aligned}\frac{\mathrm{d}\mathbb{Q}\left(x^{(n)}\right)}{\mathrm{d}\mathbb{P}}\Big|_t &= \exp\left\{-\frac{1}{2}\int_0^t \left(x_u^{(n)}\right)^2 \mathrm{d}u - \int_0^t x_u^{(n)}\,\mathrm{d}W_u^n\right\}\\ &= Z_t^{(n)}\,,\end{aligned}$$

defines an equivalent measure to $\mathbb{P}$, and by Girsanov's theorem, $W_t^n = W_t^F + \int_0^t x_s^{(n)}\,\mathrm{d}s$ is a Brownian motion under the measure $\mathbb{Q}\left(x^{(n)}\right)$, abbreviated as $\mathbb{Q}^n$. Finally,

$$\begin{aligned}\mathbb{E}^{\mathbb{P}}\left[Z_T\right] &= \mathbb{E}^{\mathbb{P}}\left[\lim_{n\to\infty} Z_T\,\mathbf{1}_{\{\tau_n=T\}}\right]\\ &= \mathbb{E}^{\mathbb{P}}\left[\lim_{n\to\infty} Z_T^n\,\mathbf{1}_{\{\tau_n=T\}}\right]\\ &= \lim_{n\to\infty} \mathbb{E}^{\mathbb{P}}\left[Z_T^n\,\mathbf{1}_{\{\tau_n=T\}}\right]\\ &= \lim_{n\to\infty} \mathbb{E}^{\mathbb{Q}^n}\left[\mathbf{1}_{\{\tau_n=T\}}\right]\\ &= \lim_{n\to\infty} \mathbb{Q}^n\left(\tau_n = T\right)\,,\end{aligned}$$

where the first two equalities hold because the control x^* is $\mathbb{P}$-a.s. finite and from definition (A.13). The third equality follows from the monotone convergence theorem. The fourth equality is the change of measure imposed by Z_T^n.

Finally, starting at F_0, the stopped process $F_{t\wedge\tau_n}$, under $\mathbb{Q}^n$ is given by

$$\begin{aligned}F_{t\wedge\tau_n} &= F_0 + \int_0^{t\wedge\tau_n}\left(\lambda\left(\bar{F} - F_s\right) + a\,\mu(s, F_s, v_s) - \sigma_F\,x_s^{(n)} - b\left(v_s - \bar{v}\right)\right)\mathrm{d}s + \int_0^{t\wedge\tau_n}\sigma_F\,\mathrm{d}W_s^n\\ &= F_0 + \int_0^{t\wedge\tau_n}\left(\lambda\left(\bar{F} - F_s\right) + a\,\mu(s, F_s, v_s) - \sigma_F\,x_s^{*} - b\left(v_s - \bar{v}\right)\right)\mathrm{d}s + \int_0^{t\wedge\tau_n}\sigma_F\,\mathrm{d}W_s^n\\ &= F_0 + \int_0^{t\wedge\tau_n}\left(\lambda\left(\bar{F} - F_s\right) + a\,\mu^{\mathbb{Q}}(s, F_s, v_s) - b\left(v_s - \bar{v}\right)\right)\mathrm{d}s + \int_0^{t\wedge\tau_n}\sigma_F\,\mathrm{d}W_s^n\,,\end{aligned}$$

and similarly, the stopped process $\tilde{F}_{t\wedge\tilde{\tau}_n}$, under $\mathbb{P}$ is given by

$$\tilde{F}_{t\wedge\tilde{\tau}_n} = F_0 + \int_0^{t\wedge\tilde{\tau}_n} \left(\lambda\left(\bar{F} - \tilde{F}_s\right) + a\,\mu^{\mathbb{Q}}(s, \tilde{F}_s, v_s) - b\,(v_s - \bar{v})\right) \mathrm{d}s + \int_0^{t\wedge\tilde{\tau}_n} \sigma_F \,\mathrm{d}W_s^F\,.$$

Therefore $\left(\tilde{F}_{t\wedge\tilde{\tau}_n}\right)_{0\le t\le T}$, and $(F_{t\wedge\tau_n})_{0\le t\le T}$ have the same distribution under $\mathbb{Q}^n$ and $\mathbb{P}$ respectively, thus

$$\lim_{n\to\infty} \mathbb{Q}^n\left(\tau_n = T\right) = \lim_{n\to\infty} \mathbb{P}\left(\tilde{\tau}_n = T\right) = 1\,,$$

and

$$\begin{aligned}\mathbb{E}^{\mathbb{P}}\left[Z_T\right] &= \lim_{n\to\infty} \mathbb{Q}^n\left(\tau_n = T\right)\\ &= \lim_{n\to\infty} \mathbb{P}\left(\tilde{\tau}_n = T\right)\\ &= 1\,,\end{aligned}$$

which concludes the proof.

□

Finally, we present the last proof of Theorem 1. We prove that (16) is the trader's value function (10).

Proof. V **is the value function of the agent's control problem in** (10)**.**

Let $\mathbb{Q}^* = \mathbb{Q}^*(\delta)$ be the measure inferred by $x_t^*(\delta) = \varphi\,\sigma_F\left(h_1(t) + 2\,h_2(t)\,F_t^\delta + h_5(t)\,v_t\right)$, in other words,

$$\frac{\mathrm{d}\mathbb{Q}^*}{\mathrm{d}\mathbb{P}}\Big|_T = \exp\left\{-\frac{1}{2}\int_0^T \boldsymbol{x}^{*\prime}_u\,\boldsymbol{x}^*_u\,\mathrm{d}u - \int_0^T \boldsymbol{x}^*_u\,\mathrm{d}\boldsymbol{W}_u\right\}, \tag{A.16}$$

where $\boldsymbol{x}_t^* = (x_t^*, 0)$. Clearly $x_t^*(\delta)$ represents the infimum measure of (10) for an admissible δ.

The dynamics of v under $\mathbb{Q}^*$ do not change, but that of $F^{\delta,\mathbb{Q}^*}$ does, specifically

$$\mathrm{d}F_t^{\delta,\mathbb{Q}^*} = \left(\lambda\,\bar{F} - \left(\lambda + 2\,\varphi\,\sigma_F^2\,h_2(t)\right)\,F^{\delta,\mathbb{Q}^*} + a\,\delta_t + b\,\bar{v} - \left(b + \varphi\,\sigma_F^2\,h_5(t)\right)\,v_t - \varphi\,\sigma_F^2\,h_1(t)\right)\mathrm{d}t + \sigma_F\,\mathrm{d}W^{F,\mathbb{Q}^*}$$

.

As $V \in \mathcal{C}^{1,2,2}$, applying Itô's Lemma under $\mathbb{Q}^*$, and considering a $\delta \in \mathcal{A}_{t,T}$, we have

$$V(T, F_T^\delta, v_T) = V(t,f,v) + \int_t^T \partial_s V + \mathcal{L}^{\mathbb{Q}^*} V\,\mathrm{d}s + \int_t^T \partial_v V\,\sigma_v\,v_s\,\mathrm{d}W_s^{v\,\mathbb{Q}^*} + \int_t^T \partial_f V\,\sigma_F\,\mathrm{d}W_s^{F,\mathbb{Q}^*}\,, \tag{A.17}$$

where $W_s^{F\star}$ and $W_s^{v\star}$ are standard $\mathbb{Q}^*$-Brownian motions. The function V is a solution of (14), and δ is an arbitrary control, thus

$$\partial_s V + \mathcal{L}^{\mathbb{Q}^*} V \le \phi_\delta\left(\delta - \hat{\delta}\right)^2 + \phi_F\left(f - \hat{F}\right)^2 - \frac{1}{2\,\varphi}\left(x^*\right)^2\,.$$

Now, take expectations on both sides of (A.17) to obtain

$$\begin{aligned}0 &= \mathbb{E}^{\mathbb{Q}^*}_{t,f,v}\left[V(T, F_T^\delta, v_T)\right]\,,\\ &= V(t,f,v) + \mathbb{E}^{\mathbb{Q}^*}_{t,f,v}\left[\int_t^T \partial_s V + \mathcal{L}_s^{\mathbb{Q}^*} V\,\mathrm{d}s\right]\,,\\ &\le V(t,f,v) + \mathbb{E}^{\mathbb{Q}^*}_{t,f,v}\left[\phi_\delta\int_t^T \left(\delta_s - \hat{\delta}\right)^2\mathrm{d}s + \phi_F\int_t^T \left(F_s^\delta - \hat{F}\right)^2\mathrm{d}s - \frac{1}{2\,\varphi}\int_t^T \left(x_s^*\right)^2\mathrm{d}s\right]\,,\\ &= V(t,f,v) + \mathbb{E}^{\mathbb{Q}^*}_{t,f,v}\left[\phi_\delta\int_t^T \left(\delta_s - \hat{\delta}\right)^2\mathrm{d}s + \phi_F\int_t^T \left(F_s^\delta - \hat{F}\right)^2\mathrm{d}s - \frac{1}{2\,\varphi}\int_t^T \left(x_s^*\right)^2\mathrm{d}s + \frac{1}{\varphi}\int_t^T x_s^*\mathrm{d}W_s^{\mathbb{Q}^*}\right]\,,\end{aligned}$$

therefore

$$\mathbb{E}^{\mathbb{Q}^*}_{t,f,v}\left[-\phi_\delta\int_t^T\left(\delta_s-\hat{\delta}\right)^2\mathrm{d}s-\phi_F\int_t^T\left(F^\delta_s-\hat{F}\right)^2\mathrm{d}s+\frac{1}{2\,\varphi}\int_t^T\left(x^*_s\right)^2\mathrm{d}s-\frac{1}{\varphi}\int_t^T x^*_s\mathrm{d}W^{\mathbb{Q}^*}_s\right]\leq V(t,f,v)\,,$$

or alternatively,

$$\mathbb{E}^{\mathbb{Q}^*}_{t,f,v}\left[-\phi_\delta\int_t^T\left(\delta_s-\hat{\delta}\right)^2\mathrm{d}s-\phi_F\int_t^T\left(F^\delta_s-\hat{F}\right)^2\mathrm{d}s-\frac{1}{2\,\varphi}\int_t^T\left(x^*_s\right)^2\mathrm{d}s-\frac{1}{\varphi}\int_t^T x^*_s\mathrm{d}W_s\right]\leq V(t,f,v)\,,$$

$$\mathbb{E}^{\mathbb{Q}^*}_{t,f,v}\left[-\phi_\delta\int_t^T\left(\delta_s-\hat{\delta}\right)^2\mathrm{d}s-\phi_F\int_t^T\left(F^\delta_s-\hat{F}\right)^2\mathrm{d}s+\mathcal{H}_{t,T}\left(\mathbb{Q}\,|\,\mathbb{P}\right)\right]\leq V(t,f,v)\,,$$

and because $V(t,f,v)$ is a bound for any $\delta\in\mathcal{A}_{t,T}$, we have that

$$\sup_{\delta\in\mathcal{A}_{t,T}}\mathbb{E}^{\mathbb{Q}^*}_{t,f,v}\left[-\phi_\delta\int_t^T\left(\delta_s-\hat{\delta}\right)^2\mathrm{d}s-\phi_F\int_t^T\left(F^\delta_s-\hat{F}\right)^2\mathrm{d}s+\mathcal{H}_{t,T}\left(\mathbb{Q}^*\,|\,\mathbb{P}\right)\right]\leq V(t,f,v)$$

$$\sup_{\delta\in\mathcal{A}_{t,T}}\inf_{\mathbb{Q}\in\boldsymbol{\mathcal{Q}}}\mathbb{E}^{\mathbb{Q}}_{t,f,v}\left[-\phi_\delta\int_t^T\left(\delta_s-\hat{\delta}\right)^2\mathrm{d}s-\phi_F\int_t^T\left(F^\delta_s-\hat{F}\right)^2\mathrm{d}s+\mathcal{H}_{t,T}\left(\mathbb{Q}\,|\,\mathbb{P}\right)\right]\leq V(t,f,v)\,,\quad\text{(A.18)}$$

which completes the first part of the proof.

Now, fix the strategy δ^*, then for any $\mathbb{Q}\in\boldsymbol{\mathcal{Q}}$, because V is a solution of (14), and as in (14) the inf and sup can be interchanged, we have that

$$\partial_s V+\mathcal{L}^{\mathbb{Q}}V\geq\phi_\delta\left(\delta^*-\hat{\delta}\right)^2+\phi_F\left(f-\hat{F}\right)^2-\frac{1}{2\,\varphi}\left(x\right)^2\,,\quad\text{(A.19)}$$

and, following the same steps as before, but now using (A.19) and working under $\mathbb{Q}$, we have

$$\mathbb{E}^{\mathbb{Q}}_{t,f,v}\left[-\phi_\delta\int_t^T\left(\delta^*_s-\hat{\delta}\right)^2\mathrm{d}s-\phi_F\int_t^T\left(F^{\delta^*}_s-\hat{F}\right)^2\mathrm{d}s+\mathcal{H}_{t,T}\left(\mathbb{Q}\,|\,\mathbb{P}\right)\right]\geq V(t,f,v)\,,$$

and because this is true for all $\mathbb{Q}\in\boldsymbol{\mathcal{Q}}$, we have

$$\inf_{\mathbb{Q}\in\boldsymbol{\mathcal{Q}}}\mathbb{E}^{\mathbb{Q}}_{t,f,v}\left[-\phi_\delta\int_t^T\left(\delta^*_s-\hat{\delta}\right)^2\mathrm{d}s-\phi_F\int_t^T\left(F^{\delta^*}_s-\hat{F}\right)^2\mathrm{d}s+\mathcal{H}_{t,T}\left(\mathbb{Q}\,|\,\mathbb{P}\right)\right]\geq V(t,f,v)$$

$$\sup_{\delta\in\mathcal{A}_{t,T}}\inf_{\mathbb{Q}\in\boldsymbol{\mathcal{Q}}}\mathbb{E}^{\mathbb{Q}}_{t,f,v}\left[-\phi_\delta\int_t^T\left(\delta_s-\hat{\delta}\right)^2\mathrm{d}s-\phi_F\int_t^T\left(F^\delta_s-\hat{F}\right)^2\mathrm{d}s+\mathcal{H}_{t,T}\left(\mathbb{Q}\,|\,\mathbb{P}\right)\right]\geq V(t,f,v)\,,\quad\text{(A.20)}$$

which completes the second part of the proof.

After combining (A.18) and (A.20) we obtain the desired result.

□