A Model of Price Impact and Market Maker Latency

Preliminary Draft

Jakub Rojček [*]

January 25, 2017

Abstract

Price impact measures the difference between the best quoted price and the realized price as a function of order size. This paper analyzes how price impact depends on the latency that a market maker is subject to. I propose a tractable model which allows incorporating both order size and latency effects as determinants of price impact. The model is solved analytically and is novel in the theoretical microstructure literature. Larger latency increases adverse selection costs to the market maker and reduces his probability of trading with a slow investor. A larger order size decreases the slow trader's outside option, making him susceptible to accept a worse price for his trade. It is shown that the first-order effect of increased latency and increased order size is to increase price impact. Their joint impact is also positive. When the probability of trading is taken into consideration, the utility of the slow institutional investor decreases with increasing latency.

JEL classification: G14, G28, C73.

Keywords: Price Impact, High-Frequency Trading, Trade Size, Latency, Market Quality, Welfare

[*]Department of Banking and Finance, University of Zurich and Swiss Finance Institute. E-mail: jakub.rojcek@bf.uzh.ch. Address: Plattenstrasse 22, 8032 Zurich, Switzerland. I am grateful to Ramazan Gençay, Michel Habib, Thorsten Hens, Boyan Jovanovic, Felix Kübler, Albert Menkveld, Per Östberg, Nikola Vasiljević ,and Alexandre Ziegler for providing me with comments, discussion and suggestions on this paper, as well as participants of the Belgrade Young Economist Conference 2016 and University of Zurich Brown Bag Seminar.

I. Introduction

The joint influence of trading speed and order size has been largely neglected in recent high-frequency trading research. Most analyses have focused on understanding the impact of high-frequency trading on market quality and its influence on execution quality for traditional institutional investors in terms of the bid-ask spreads valid for unit size orders.[1] While theoretical models have shown in this setting that algorithmic traders in continuous limit order markets derive a competitive advantage from faster analysis of order book evolution, processing of news and higher monitoring speed,[2] the empirical literature has focused on the impact of algorithmic trading on traditional measures of market quality.[3] The majority of the empirical literature confirms the positive first-order effects of HFT, especially lower bid-ask spreads and faster price discovery. However, Hendershott et al. (2011) report that the presence of HFTs also decreases the depth of the order book and increases the costs of executing large orders. Taking execution costs into consideration, Tong (2015) measures the execution shortfall of institutional investors and finds that HFTs increase transaction costs for them. It is this overall effect of increased price impact which large institutional traders suffer that this paper focuses on. I present a model which incorporates the joint effect of latency and order size in a setting with a high-frequency market maker, high-frequency snipers, and slow investors. The model predicts that higher market maker latency and larger order size lead to higher price impact. Taking the probability of trading into consideration, slow institutional investors' utility is strictly deteriorating in larger order size and latency. The following sections present a review of the related literature, the model, comparative statics of the quoted price, and a welfare analysis.

II. Related Literature

The existing theoretical literature on HFT does not provide a model of price impact as a function of latency. There are two related streams of theoretical models. Repeated double auction models, which represent synchronous trading; and asynchronous trading models, which represent

[1] For an overview of the literature on high-frequency trading, see the surveys by Jones (2013), O'Hara (2015) and Menkveld (2016).
[2] Foucault et al. (2015), Aït-Sahalia and Saglam (2014), Biais et al. (2015) and Hoffmann (2014).
[3] See Jovanovic and Menkveld (2011), Brogaard (2010), Hasbrouck and Saar (2013), Riordan and Storkenmaier (2012), Brogaard et al. (2014), Tong (2015), and Hendershott et al. (2011).

trading in continuous time. This section reviews the theoretical predictions of models in these two categories.

The double auction models of Roşu (2016), Rostek and Weretka (2015), Du and Zhu (2016), and Foucault et al. (2015) build on the models of Kyle (1985), Vayanos (1999), and Vives (2011). Rostek and Weretka (2015) show that for traders, maximizing welfare and stabilizing liquidity through disclosure of information about fundamentals at the same time represents a trade-off. The traders in their model balance the present execution value against future price impact. In a rational expectations equilibrium, traders split their orders optimally. The point from which I try to depart in my model is the synchronicity of traders arrivals. Du and Zhu (2016) depart from the synchronicity of arrivals in the double auction framework. They introduce one fast trader who is in the market every time step, while the rest of the traders arrives only every certain number of time steps. This leads to interesting results where the fast trader prefers higher frequency of trading, whereas the slow traders prefer slower trading. However, because this model falls into the category of double auction markets, traders do not suffer the latency effect per se. Their supplies change based on their frequency of trading, but not based on the risk they bear due to latency, during which they cannot change their orders. In the later version of the same paper, Du and Zhu (2016) introduce a model, where fast traders intermediate trades among asynchronously arriving slow traders, this extension is solved numerically. The present paper models this latency effect directly as a parameter of the price impact function. The closest to the modelling goal of the current paper is the paper by Foucault et al. (2015), where the speculator can receive the news about the fundamental value with a time advantage. This can be considered latency in the extension, where the time advantage is not instantaneous, but represents a time interval of certain length. In the repeated double auction setting they consider, the authors cannot solve for the equilibrium analytically and resolve to numerical solutions. Roşu (2016) models fast traders as those receiving information about the fundamental value instantly and slow traders as those receiving the information with a lag. Both categories are speculators. The comparative statics of price impact are derived as a function of the number of fast and slow traders. The variable lag size in the generalized model can serve as a good model for latency in double auction market. The current paper solves for price impact as a function of latency and size analytically in asynchronous arrivals modelling

framework.

Departing from the double auctions modelling framework are the asynchronous trading models of Menkveld and Zoican (2016), Budish et al. (2015), and Chacko et al. (2008).
Using Poisson arrivals, Menkveld and Zoican (2016) model liquidity traders (submitting market orders), HF Bandits (also submitting market orders), HF market makers (submitting limit orders), and good or bad news about the fundamental value. The net effect of latency on the spread depends on the news to liquidity traders ratio. What this model is lacking is the effect of trade size on prices. Budish et al. (2015) predict that in a limit order market with competing fast traders, the quoted size will always be one unit. An increasing presence of these fast snipers means that the liquidity provider can update his order only with diminishing probability in case it becomes mispriced. This results in a partial equilibrium, where the book contains a single unit limit order on each side of the market. However, apart from this prediction, their model does not provide the price impact function as a function of size, because their equlibrium size is always one, nor as a function of latency. Chacko et al. (2008) model sell limit orders as writing a perpetual American call option, requiring delivery of the underlying block of shares upon execution. Similarly, a limit order to buy is like a short position in an American put option. What is specific about this model is that the limit orders have to be executed immediately in order to be able to use option pricing techniques. To ensure immediate execution, the initiator of a transaction offer (the option writer) must offer a price at which it is currently optimal for the receiver of the transaction offer (the option owner) to exercise the option early. In effect the market order is modelled as a limit order, which is submitted at the best opposite quote in order to ensure immediate execution. This structure does not permit an analysis of the effect of latency on price impact. What it allows, however, is to consider the impact of size, which is derived from the market maker having to resell the inventory back to the market. As a result, because the traders arrive at frequency which decreases with order size, this directly translates into positive price impact, with its shape coming from the optimal execution rules for perpetual American options.

This paper in its core also borrows from the seminal model of Grossman and Miller (1988), where liquidity trader shares risk with market makers, but the impact of latency is not considered.

The goal of this paper is to obtain a tractable model of price impact as a function of order size and the market maker's latency. To achieve this goal, I propose a continuous time, asynchronous arrival model that consists of a risk-neutral monopolist market maker, a risk-averse buyer and seller, and high-frequency bandits. The buyer and the seller arrive in a stylized fashion according to a Poisson process and look for immediate trading opportunities. They submit a market order if the utility of doing so is at least their reservation utility. The reservation utility is the utility of submitting a limit order and waiting for a counterparty with opposite trading needs. In the general case, the fundamental value is modelled as a Brownian motion and it is possible to solve for the ask price numerically. I also provide a closed-form solution for price impact as a function of latency for the special case in which changes in the fundamental value over very short discrete intervals follow a uniform distribution. The paper provides novel insights on the role of high-frequency market makers and their contribution to liquidity and welfare in modern financial markets. It introduces a novel modelling framework in the theoretical microstructure literature providing scope for many surmisable extensions.

III. General Modelling Framework of Price Impact with Latency

This section presents a general framework that allows solving for the equilibrium ask price numerically. Further assumptions are used in the next section to derive an analytical expression for the ask price in order to facilitate the comparative statics and welfare analysis.

There is one financial asset, whose fundamental value at time t is denoted v_t. The fundamental value's dynamics is represented by a continuous-time stochastic process with zero drift and variance proportional to the time passed, $\sigma^2 t$, where the parameter σ is the volatility of the fundamental value. Let's denote the distribution of the increment $x := v_T - v_t$ as $F(x; \mu = 0, \sigma^2(T-t))$.

There is one buyer, one seller, and N risk-neutral high-frequency traders (HFT) in the market. One of the HFTs is currently the market maker (HFM), while the remaining $N-1$ HFTs are high-frequency bandits (HFB), who wait for mispriced limit orders and in that case, they snipe them and realize a profit.[4] The buyer and the seller are risk-averse.

[4] This setting generalizes to a dynamic arrivals setting where buyers and sellers arrive according to the stylized fashion described in detail below.

The flow of events in the model is the following.[5] At time t the market maker submits a quote to sell, a, which is called the *ask price*. The ask price is expressed as a deviation from the current fundamental value v_t, so that the overall price the market maker sets for the asset is $v_t + a$. The market maker suffers a *latency* Δ and can return and update his ask price only after time Δ passes.[6] Two events can happen. Either the buyer enters according to a point process with intensity λ_B at a random time T_B before time Δ passes, evaluates whether he will demand immediacy and trade on the current ask, or one of the HFTs arrives after time Δ passes, in which case the market maker adjusts the ask if he arrives before one of the HFBs. If an HFB arrives before the market maker, the HFB snipes the ask if it is mispriced.[7]

The buyer has a private valuation of π_B in addition to the asset's current fundamental value. He will demand immediacy if his utility of holding the asset minus the ask price he pays is at least his reservation utility from submitting a limit order and waiting for the seller. If the buyer submits a market order, he will disappear from the market and the market maker will post a bid price at which he is willing to buy back the asset from the arriving seller and eliminate his inventory.

Then the seller arrives[8] according to a point process with intensity λ_S. The same reasoning as for the buyer applies. The seller decides whether to directly trade with the market maker or to submit a limit order and wait for the arrival of a new buyer, who will arrive at a random time T_{B2}. The seller submitting a limit order faces the risk that the fundamental value moves in an adverse direction. In that case, he might suffer a loss. On the other hand, if the fundamental value moves too much in a favourable direction, the buyer will not trade with the seller's limit order. The seller takes these possibilities into consideration when setting his ask price.

I assume that the traders trade Q shares at once and there is no uncertainty about this quantity.[9] As in Chacko et al. (2008), the arrival rate of traders on the opposite side of the market, in this case the seller, is assumed to be a decreasing function of the trade's size.

[5] The corresponding flow of events for the closed-form example is summarized in Figure 1.

[6] Latency in a strict sense would mean that an ask set at time $t\Delta$ would be valid in the time interval $[(t+1)\Delta, (t+2)\Delta)$ based on information at time $(t-1)\Delta$. This is because processing information and quote submission are also subject to latency. For this model, I assume that the impact of latency can be simplified to the starting specification with the market maker's ask set at time $t\Delta$ being valid in interval $[(t)\Delta, (t+1)\Delta)$ based on the information at time $t\Delta$.

[7] The slow traders arrive in continuous time and the HFTs arrive at discrete points in time, Δ time units apart from each other.

[8] I assume that buyer and seller arrivals alternate in a deterministic fashion.

[9] This is a classical assumption in the microstructure literature (Ho and Stoll (1981)), which is relaxed in Budish et al. (2015) and particularly in the optimal execution literature.

By submitting a limit order to sell at price a, the market maker issues an option to the fundamental buyer. The fundamental buyer executes on this limit order if $v_{T_B} - v_t - a + a' \geq 0$, where a' is the reservation ask. The reservation ask is the maximum price the buyer is willing to pay to trade the asset immediately and depends on the reservation value derived from submitting a buy limit order and waiting for a seller. The reservation ask price depends on future trading opportunities; it is computed in closed-form for the case of a uniform distribution in the next section. If the slow buyer does not arrive before time Δ passes and the market maker's limit order becomes mispriced, meaning that $v_{t+\Delta} > v_t + a$, the remaining $N-1$ HFBs will try to snipe it. Each one will be successful with equal probability $\frac{1}{N}$. With probability $\frac{1}{N}$, the market maker will be successful at cancelling this mispriced order.[10]

Were the market maker only facing the slow buyer and no HFBs, his payoff could be decomposed into a short position in an American call option plus owning a cash-or-nothing digital call option, both with strike price of $a - a'$ and the payoff of the cash-or-nothing call option equal to a'. The buyer on the other hand owns an American call option with the same strike price, $a - a'$. Formally, the market maker's payoff is

$$\mathbb{E}[LP(a)] = \mathbb{E}[(a - (v_{T_B} - v_t))\, \mathbb{1}_{v_{T_B} - v_t \geq a - a'} \mathbb{1}_{T_B \leq \Delta}]. \tag{1}$$

Assuming that the buyer arrives according to a Poisson process with intensity λ_B, this expectation can be written as

$$\mathbb{E}[LP(a)] = \int_0^{\Delta} \lambda_B e^{-\lambda_B y} \int_{a-a'}^{\infty} (a - x) F(\mathrm{d}x; 0, \sigma^2 y) \mathrm{d}y. \tag{2}$$

In the case with HFBs present in the market, the payoff of the market maker has three parts, one arising from trading with the slow buyer, the other two from trading with the HFBs. The market maker trades with the HFBs if the ask price at time $t + \Delta$ is mispriced ($v_{t+\Delta} > v_t + a$), provided that the slow trader does not arrive before $t + \Delta$ or if he does, he decides not to trade with the market maker. The market maker's payoff from trading with HFBs is thus equivalent to selling

[10] The corresponding payoff of the market maker in the closed-form example is depicted in Figure 2.

$\frac{N-1}{N}$ call options with strike price a. Equation (1) generalizes to

$$\begin{aligned}\mathbb{E}[LP(a)] = &\mathbb{E}[(a-(v_{T_B}-v_t))\,\mathbb{1}_{v_{T_B}-v_t\geq a-a'}\mathbb{1}_{T_B\leq\Delta}]\\ &-\frac{N-1}{N}\mathbb{E}[(v_{t+\Delta}-v_t-a)\,\mathbb{1}_{v_{t+\Delta}-v_t\geq a}\mathbb{1}_{T_B>\Delta}]\\ &-\frac{N-1}{N}\mathbb{E}[(v_{t+\Delta}-v_t-a)\,\mathbb{1}_{v_{t+\Delta}-v_t\geq a}\mathbb{1}_{T_B\leq\Delta}\mathbb{1}_{v_{T_B}-v_t<a-a'}],\end{aligned} \tag{3}$$

and the expectations can be computed as

$$\begin{aligned}\mathbb{E}[LP(a)] = &\int_0^\Delta \lambda_B e^{-\lambda_B y}\int_{a-a'}^\infty (a-x)F(\mathrm{d}x;0,\sigma^2 y)\mathrm{d}y\\ &-\frac{N-1}{N}\left(\int_\Delta^\infty \lambda_B e^{-\lambda_B y}\mathrm{d}y\right)\int_a^\infty (x-a)F(\mathrm{d}x;0,\sigma^2\Delta)\\ &-\frac{N-1}{N}\int_0^\Delta \lambda_B e^{-\lambda_B y}\int_{-\infty}^{a-a'}\int_{a-x}^\infty (z+x-a)F(\mathrm{d}z;0,\sigma^2(\Delta-y))F(\mathrm{d}x;0,\sigma^2 y)\mathrm{d}y,\end{aligned} \tag{4}$$

where the first line represents the possible profit by trading with the slow trader, the second line the loss due to the mispriced limit order being sniped by one of the HFBs at the end of the latency period if the slow buyer only arrives after Δ time passes, and the third line the case where the slow trader arrives before Δ time passes, decides not to trade because $v_{T_B}-v_t < a-a'$ and subsequently the fundamental value rises to $v_{t+\Delta}-v_t > a$.[11]

The reservation ask a' is the highest price the buyer is willing to pay to obtain the asset from the market maker. It is the price that equates the immediately available utility with the expected utility the buyer could obtain by submitting a buy limit order at price b and waiting for a seller.[12] The seller's private valuation for the asset is $-\pi_S$, where $\pi_S > 0$. The payoff structure is depicted in Figure 3, where the buyer's payoff is increasing in the fundamental value increment, but shrinks to zero once the fundamental value increment exceeds the seller's reservation value.[13] The expected

[11] The last term might be neglected if the probability $\int_0^\Delta \lambda_B e^{-\lambda_B y}\int_{-\infty}^{a-a'}\int_{a-x}^\infty F(\mathrm{d}z;0,\sigma^2(\Delta-y))F(\mathrm{d}x;0,\sigma^2 y)\mathrm{d}y$ is small enough.

[12] The bid price, b, is again expressed as a deviation from the fundamental value at time T_B and the overall bid price is $v_{T_B}+b$. The buyer sets b in order to maximize $V_{B,LO}(b)$.

[13] We assume that the seller accepts the limit order if his payoff is non-negative. However, he could also optimize and submit a sell limit order, in which case we assume that the previous buyer exits the game. The number of optimizing agents is driven by computational considerations and trades off the precision with computation time.

payoff from submitting the limit order is the following

$$V_{B,LO}(b) = \mathbb{E}[u(\pi_B - v_{T_B} - b + v_{T_S})|v_{T_B} + b \geq v_{T_S} - \pi_S]. \tag{5}$$

The reservation ask price is then the price which solves the following equation at time T_B

$$u(\pi_B + v_{T_B} - a') = V_{B,LO}(b). \tag{6}$$

Given the reservation ask price, the market maker then uses Equation (1) to set a such that he fulfills an equilibrium condition. He might either set a in order to maximize his payoff or such that his expected payoff equals the expected HFB's payoff in case the competition from HFBs prevents profit maximizing behavior. One can solve for the equilibrium ask price a numerically in the general case. In the next section, I will solve an example which allows for a closed-form solution.

IV. Closed-form Example

Generally, it is not possible to solve for the ask price in the above problem in closed-form. This section provides an example of a closed-form solution that allows investigating the role of latency in the price impact function.

The fundamental value's dynamics is now represented by a discrete-time stochastic process. As before, we let Δ denote the market maker's latency. We assume that the change in the fundamental value over the short interval Δ is distributed according to a uniform distribution $\mathcal{U}(-\sqrt{3}\sigma\sqrt{\Delta}, \sqrt{3}\sigma\sqrt{\Delta})$.[14] The expected change in the fundamental value is thus zero and the variance of the fundamental value change is proportional to the latency, $\sigma^2\Delta$.

The flow of events is summarized in Figure 1. A fundamental buyer arrives according to a Poisson point process with intensity λ. The probability of n buyers arriving by time Δ is $\mathbb{P}[B(0,\Delta] = n] = \frac{(\lambda\Delta)^n}{n!}e^{-\lambda\Delta}$. We assume that the latency is sufficiently small that two and more arrivals of fundamental traders are very unlikely during the Δ interval. The probability that exactly one

[14] Empirically, high-frequency returns are not normally distributed and have heavy tails with large spikes around zero.

fundamental buyer arrives by time Δ is $\lambda\Delta + \mathcal{O}(\Delta^2)$, which comes from applying a Taylor approximation to the probability of Poisson arrivals. The probability of no buyer arriving by time Δ is then $1 - \lambda\Delta + \mathcal{O}(\Delta^2)$.

In order to be able to obtain a closed-form solution to the market maker's problem, we also suppose that the fundamental buyer faces the fundamental value at the end of the interval Δ. This means that if the buyer arrives during $(t, t+\Delta]$, the fundamental value he takes into account is $v_{t+\Delta}$.[15] When the buyer arrives, he can execute on the current ask price set by the market maker, a, or submit his own limit order and wait for the potential fundamental seller. In addition to the fundamental value $v_{t+\Delta}$, the buyer derives a private value π_B from holding the asset. Let a' denote the buyer's *reservation ask*. The reservation ask is the maximum price the buyer is willing to pay to trade the asset immediately and depends on the reservation value derived from submitting a buy limit order and waiting for a seller. The buyer's decision whether to execute or not based on the current ask price, a, is depicted in Figure 2 and can be summarized as follows

$$v_{t+\Delta} - v_t \begin{cases} \geq a - a' & \text{buyer executes,} \\ < a - a' & \text{buyer does not execute.} \end{cases} \tag{7}$$

Market Maker's Problem

Because the stylized limit order book can only hold one ask, only one HFT can become a market maker (HFM). By contrast with Section III, I assume that the slow buyer is never successful at picking off a mispriced ask, so his payoff lies between 0 and a'. If the ask becomes mispriced, it is cancelled by the HFM with probability $\frac{1}{N}$ and picked off by one of the HFBs with probability $\frac{N-1}{N}$. The expected payoff to the market maker is given by

$$\mathbb{E}[LP(a)] = \lambda\Delta \int_{a-a'}^{a} (a-x) \frac{1}{2\sigma\sqrt{\Delta}\sqrt{3}} \mathrm{d}x - \frac{N-1}{N} \int_{a}^{\sigma\sqrt{\Delta}\sqrt{3}} (x-a) \frac{1}{2\sigma\sqrt{\Delta}\sqrt{3}} \mathrm{d}x \tag{8}$$

$$= \lambda\Delta \frac{a'^2}{4\sqrt{3}\sqrt{\Delta}\sigma} - \frac{N-1}{N} \left(\frac{a^2}{4\sqrt{3}\sqrt{\Delta}\sigma} - \frac{a}{2} + \frac{1}{4}\sqrt{3}\sqrt{\Delta}\sigma \right). \tag{9}$$

[15] Weighting the market maker's payoff by the arrival time leads to fixed point problems from which it is not possible to back out the ask price as a function of parameters and basic functions in closed form.

A HFB has a chance of $\frac{1}{N}$ that he would successfully snipe a mispriced ask. His expected payoff is given by

$$\mathbb{E}[SP(a)] = \frac{1}{N}\int_a^{\sigma\sqrt{\Delta}\sqrt{3}} (x-a)\frac{1}{2\sigma\sqrt{\Delta}\sqrt{3}}\mathrm{d}x \tag{10}$$

$$= \frac{1}{N}\left(\frac{a^2}{4\sqrt{3}\sqrt{\Delta}\sigma} - \frac{a}{2} + \frac{1}{4}\sqrt{3}\sqrt{\Delta}\sigma\right). \tag{11}$$

As in Menkveld and Zoican (2016), our equilibrium condition states that the expected payoff of the HFM and HFBs must be equal

$$\mathbb{E}[LP(a)] = \mathbb{E}[SP(a)]. \tag{12}$$

By applying this condition and rearranging terms, we obtain the following quadratic equation in a, which must hold

$$-\frac{a^2}{4\sqrt{3}\sqrt{\Delta}\sigma} + \frac{a}{2} - \frac{1}{4}\sqrt{3}\sqrt{\Delta}\sigma + \lambda\Delta\frac{a'^2}{4\sqrt{3}\sqrt{\Delta}\sigma} = 0. \tag{13}$$

This equation has two solutions

$$a^*_{1,2} = \sqrt{3}\sqrt{\Delta}\sigma \pm \sqrt{\lambda\Delta}a'. \tag{14}$$

Economically meaningful is the solution which increases in the reservation ask price, $a^* = \sqrt{3}\sqrt{\Delta}\sigma + a'\sqrt{\lambda\Delta}$. The first term in the equilibrium ask price comes from the snipe off part. It represents the ask price which equates the expected loss of the market maker with the HFB's expected profit in the case that the probability of the slow buyer's arrival is zero. The second term represents an adjustment for the expected profit from trading with the slow buyer.

Proposition 1. Market maker's ask price. *Let there be N high-frequency traders in the market. Given the latency Δ, the buyer's arrival rate λ, and assuming $v_\Delta - v_0 \sim \mathcal{U}(-\sqrt{3}\sqrt{\Delta}\sigma, \sqrt{3}\sqrt{\Delta}\sigma)$, the equilibrium ask price is*

$$a^* = \sqrt{3}\sqrt{\Delta}\sigma + \sqrt{\lambda\Delta}a'. \tag{15}$$

Proof. Follows from the steps above. □

We next solve for the highest reservation price a' that a buyer is willing to pay.

Buyer's Reservation Value

I assume that both the buyer and the seller are risk averse and have exponential constant absolute risk-aversion utility functions $u(x) = 1 - e^{-\alpha x}$, where α is the risk-aversion coefficient.

The buyer arrives at time $t+\Delta$ and observes the fundamental value at that time, $v_{t+\Delta}$. He executes if the value of submitting a market order $1 - e^{-\alpha(\pi_B + v_{t+\Delta} - v_t - a)}$ is higher than the value of submitting a limit order, which we compute below. Otherwise, the buyer submits a limit order and waits for a seller, who arrives according to a Poisson process with intensity $\lambda_S(Q)$. The arrival intensity is a decreasing function of the order size Q, meaning that the buyer would in expectation have to wait longer for an opposite side trader if his order is larger. Although this paper does not depend in its findings on the precise functional form, for illustrating figures, we use the functional form proposed by Chacko et al. (2008), where the intensity is inversely related to the quantity traded, $\lambda_S(Q) = \frac{\Lambda_S}{Q}$. The parameter Λ_S represents the arrival intensity of unit size order seller. This is equivalent to assuming that the demand for trading is stationary per unit of time as in Garman (1976). I will use λ_S and $\lambda_S(Q)$ interchangeably.

It is assumed that the mean arrival time of the seller, $\frac{1}{\lambda_S}$, is much larger than the market maker's latency, Δ. Because the innovations to the fundamental value are uniformly distributed with mean zero and variance $\sigma^2\Delta$, it follows from the central limit theorem, that the sum of such innovations, $\sum_{j=1}^{J}(v_{t+j\Delta} - v_{t+(j-1)\Delta})$ is normally distributed with mean zero and variance $\sigma^2 J\Delta$, for large values of J. The change in the fundamental value between the buyer's arrival (reset to 0 for convenience) and the seller's arrival time T_S is approximately normally distributed with mean 0 and variance $\sigma^2 T_S$, $v_{T_S} \sim N(0, \sigma^2 T_S)$. This simplifies the calculations for the buyer's reservation value and enables closed-form solution for the buyer's reservation ask price.

Suppose that the buyer submits a limit order priced at the fundamental value v_t.[16] It is assumed that the seller will execute on the buyer's order in case his payoff is not negative. His utility from the trade is $1 - e^{-\alpha(\pi_S - v_{T_s})}$, where $-\pi_S$ is his private valuation for the asset in addition to its current fundamental value. It thus follows that the seller executes on the buyer's order in case the fundamental value is below π_S, as this still leaves the seller with a positive trading surplus. The buyer's payoff increases up until this point and is zero once the fundamental value is larger than π_S. The seller's and buyer's payoffs from trading are depicted in Figure 3. The seller's decision can be summarized as follows

$$v_{T_S} \begin{cases} \leq \pi_S & \text{seller executes,} \\ > \pi_S & \text{seller does not execute.} \end{cases} \tag{16}$$

Taking the seller's decision into consideration, the following lemma states the reservation value of the buyer.

Lemma IV.1. Buyer's reservation value. *Given the seller's arrival rate λ_S and private valuation π_S, the buyer's risk-aversion coefficient α and private valuation π_B, the fundamental value's volatility σ and $\lambda_S > \frac{\alpha^2\sigma^2}{2}$, the buyer's reservation value, $V_{B,LO}$, is the following:*

$$V_{B,LO} = 1 - \frac{\lambda_S e^{-\alpha\pi_B}}{\lambda_S - \frac{\alpha^2\sigma^2}{2}} \left(1 + e^{-\alpha\pi_S} e^{-\sqrt{\frac{2\lambda_S\pi_S}{\sigma^2}}} \left[\frac{\alpha\sigma}{2}\sqrt{\frac{\pi}{\lambda_S}} - \sqrt{\frac{\pi}{2}}\sqrt{\pi_S}\right]\right). \tag{17}$$

Proof. The proof is given in Section A. □

In case the seller's private valuation π_S is much larger than the variance $\sigma^2 T_S$, the above expression is approximately equal to

$$V_{B,LO} \approx 1 - \frac{\lambda_S e^{-\alpha\pi_B}}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}. \tag{18}$$

In the following, the buyer's reservation utility is transformed to the maximum ask price a' at which he is willing to buy the asset from the market maker.

[16] v_t serves here as a reference point, the analysis and conclusions do not change due to this choice. It is equivalent to setting the bid, b, in Equation (5) to 0. This paper also does not aim at explicitly modelling the limit order submission decision of the buyer, who is being considered here as a liquidity trader with a binary choice.

The Buyer's Reservation Ask Price

The buyer will be indifferent between trading at the market maker's ask price and submitting a limit order if his utility of submitting the market order, $V_{B,MO}(a')$, equals his utility from submitting a limit order, $V_{B,LO}$, which we derived above as the buyer's reservation utility. The *reservation ask price* a' is the ask price that equates these two utilities. Using Equation (18) and $V_{B,MO}(a') = 1 - e^{-\alpha(\pi_B - a')}$ yields

$$1 - \frac{\lambda_S e^{-\alpha\pi_B}}{\lambda_S - \frac{\alpha^2\sigma^2}{2}} = 1 - e^{-\alpha(\pi_B - a')}. \tag{19}$$

Simplifying yields

$$e^{-\alpha\pi_B}\left(e^{\alpha a'} - \frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right) = 0. \tag{20}$$

Solving this equation for a' results in the following lemma:

Lemma IV.2. Reservation ask price. *Given the seller's arrival rate λ_S, the buyer's risk-aversion coefficient α and the fundamental value's volatility σ, the highest price at which the buyer is willing to buy the asset, a', is:*

$$a' = \frac{1}{\alpha}\log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right) \tag{21}$$

Proof. The proof follows from the steps above.[17]

□

The Market Maker's Ask Price

The last step to compute the market maker's ask price is to insert the expression for the reservation ask price from lemma IV.2 into the general solution for the ask price given in proposition 1. This yields

[17] The $\log(\cdot)$ represents the natural logarithm function.

Proposition 2. Market maker's ask price with reservation ask. *Let there be N high-frequency traders in the market. Given the latency Δ, the buyer's arrival rate λ, the seller's arrival rate λ_S and assuming $v_\Delta - v_0 \sim \mathcal{U}(-\sqrt{3}\sqrt{\Delta}\sigma, \sqrt{3}\sqrt{\Delta}\sigma)$ and that $\frac{1}{\lambda_S} >> \Delta$, the equilibrium ask price is*

$$a^* = \sqrt{3}\sqrt{\Delta}\sigma + \sqrt{\lambda\Delta}\frac{1}{\alpha}\log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right). \tag{22}$$

Proof. Follows from the steps above. □

In the following section we use this result to investigate the relationship market maker latency and price impact.

V. Comparative Statics

This paper's main result is stated in proposition 3. It provides the comparative statics analysis of the ask price with respect to the model primitives.

Proposition 3. Price impact and latency. *Given the market maker's ask price a^* as stated in Equation (22), the ask price*

1. *increases in the market maker's latency Δ,*
2. *decreases in the seller's arrival rate λ_S,*
3. *increases in the trade size Q,*
4. *increases in the asset's volatility σ.*

Proof. The proof is outlined below.

□

Let us first analyze the impact of latency Δ, starting from Equation (22) by taking derivatives with respect to Δ:

$$\frac{\partial a^*}{\partial \Delta} = \frac{\partial}{\partial \Delta}\left(\sqrt{3}\sqrt{\Delta}\sigma + \sqrt{\lambda\Delta}\frac{1}{\alpha}\log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right)\right) \tag{23}$$

$$= \frac{\sigma}{2}\sqrt{\frac{3}{\Delta}} + \frac{1}{2}\sqrt{\frac{\lambda}{\Delta}}\frac{1}{\alpha}\log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right) > 0. \tag{24}$$

The first term comes from the increased adverse selection the market maker is facing and is positive. Thus, larger latency leads to higher adverse selection costs, which are compensated by a higher spread. The second term represents the mark up the market maker is able to charge the buyer on average due to the buyer's preference for immediacy. That term is also positive as the reservation ask price is positive.

The impact of the intensity of seller arrivals is determined in a similar fashion:

$$\frac{\partial a^*}{\partial \lambda_S} = \frac{\partial}{\partial \lambda_S}\left(\sqrt{3}\sqrt{\Delta}\sigma + \sqrt{\lambda\Delta}\frac{1}{\alpha}\log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right)\right) \tag{25}$$

$$= \sqrt{\lambda\Delta}\frac{1}{\alpha}\frac{\lambda_S - \frac{\alpha^2\sigma^2}{2}}{\lambda_S}\frac{\lambda_S - \frac{\alpha^2\sigma^2}{2} - \lambda_S}{(\lambda_S - \frac{\alpha^2\sigma^2}{2})^2} \tag{26}$$

$$= -\sqrt{\lambda\Delta}\frac{1}{\alpha}\frac{\frac{\alpha^2\sigma^2}{2}}{\lambda_S(\lambda_S - \frac{\alpha^2\sigma^2}{2})} < 0. \tag{27}$$

The negative sign confirms the intuition that if the buyer's outside option is more valuable, because his chance of meeting a seller sooner is higher, he will be less willing to pay a high ask price.
The impact of trade size on the ask price can be obtained by recalling that $\lambda_S(Q) = \frac{\Lambda_S}{Q}$ and applying the chain rule,

$$\frac{\partial a^*(\lambda_S(Q))}{\partial Q} = \frac{\partial a^*}{\partial \lambda_S}\frac{\partial \lambda_S}{\partial Q} \tag{28}$$

$$= -\sqrt{\lambda\Delta}\frac{1}{\alpha}\frac{\frac{\alpha^2\sigma^2}{2}}{\lambda_S(\lambda_S - \frac{\alpha^2\sigma^2}{2})}\left(-\frac{\Lambda_S}{Q^2}\right) \tag{29}$$

$$= \sqrt{\lambda\Delta}\frac{1}{\alpha}\frac{\frac{\alpha^2\sigma^2}{2}}{\lambda_S(\lambda_S - \frac{\alpha^2\sigma^2}{2})}\left(\frac{\Lambda_S}{Q^2}\right) > 0. \tag{30}$$

The larger the quantity the buyer wants to trade, the longer he would need to wait for a potential seller, the lower is the value of his outside option and thus the higher is the price the market maker is able to charge.
The cross-derivative of the ask price with respect to latency Δ and the order size Q is given by

$$\frac{\partial a^*}{\partial Q \partial \Delta} = \frac{1}{2}\sqrt{\frac{\lambda}{\Delta}}\frac{1}{\alpha}\frac{\frac{\alpha^2\sigma^2}{2}}{\lambda_S(\lambda_S - \frac{\alpha^2\sigma^2}{2})}\left(\frac{\Lambda_S}{Q^2}\right) > 0. \tag{31}$$

The higher is the probability that a buyer with a lower valued outside option will come, the higher the ask price that the market maker can charge.

The effect of an increase in the volatility of the fundamental value on the ask price is:

$$\frac{\partial a^*}{\partial \sigma} = \frac{\partial}{\partial \sigma}\left(\sqrt{3}\sqrt{\Delta}\sigma + \sqrt{\lambda\Delta}\frac{1}{\alpha}\log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right)\right) \tag{32}$$

$$= \sqrt{3}\sqrt{\Delta} + \sqrt{\lambda\Delta}\frac{1}{\alpha}\frac{\lambda_S - \frac{\alpha^2\sigma^2}{2}}{\lambda_S}\frac{\lambda_S\alpha^2\sigma}{(\lambda_S - \frac{\alpha^2\sigma^2}{2})^2} \tag{33}$$

$$= \sqrt{3}\sqrt{\Delta} + \sqrt{\lambda\Delta}\frac{\alpha\sigma}{\lambda_S - \frac{\alpha^2\sigma^2}{2}} > 0. \tag{34}$$

Volatility increases the ask price through two channels. First, a higher volatility leads to higher adverse selection costs for the market maker. Second, it reduces the value of the risk-averse buyer's outside option, allowing the market maker to charge a higher ask price.

Illustrative examples are provided in Figure 4. This figure represents the sensitivity of the ask price a^* to the latency Δ, order size Q, and volatility σ. The base parameters chosen are $\sigma = 0.2$, $\lambda = 1$, $\Lambda_S = 10$, and $Q = 1$. The left panel shows the impact of latency and order size. The middle panel shows the impact of latency and volatility and the right panel the impact of order size and volatility. The ask price is increasing in latency and this increase is more prominent, the larger the order size. The ask price also increases in the volatility of the fundamental value.

VI. Welfare Analysis

This section analyzes comparative statics of slow traders' and high-frequency traders' expected profits as well as the probability of trading.

High-frequency Traders' Profit

The profit of the HFM has to be equal in expectation to the profit of the HFBs. HFTs' profit is obtained by plugging the solution for the ask price from Equation (22) into the HFBs' profit

Equation (10). The sensitivity of HFBs' profit to the model parameters is given in the following lemma.

Lemma VI.1. Equilibrium HFT profits. *Let there be N HFTs in the market. Given the seller's arrival rate λ_S, the buyer's risk-aversion coefficient α, the market maket's latency Δ and the fundamental value's volatility σ, the HFT's expected profit is*

$$\mathbb{E}[SP(a^*)] = \frac{1}{N}\left(\frac{\sqrt{\Delta}\lambda a'^2}{4\sqrt{3}\sigma}\right) \tag{35}$$

$$= \frac{1}{N}\left(\frac{\sqrt{\Delta}\lambda\left[\log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right)\right]^2}{4\sqrt{3}\sigma\alpha^2}\right). \tag{36}$$

HFTs' profit

1. *decreases in the number of HFTs N,*
2. *increases in the buyer's arrival rate λ,*
3. *decreases in latency Δ,*
4. *decreases in the seller's arrival rate λ_S,*
5. *increases in the asset's volatility σ.*

Proof. The proof is outlined below. □

The effect of an increase in latency can be directly observed in Equation (35), where the increase in latency leads to an increase in HFT profits in proportion to $\sqrt{\Delta}$. This is due to the increased probability of the arrival of the slow buyer by time Δ; thus, it is more natural to normalize the profit by time. Doing so yields $\frac{1}{N}\left(\frac{\lambda a'^2}{4\sqrt{\Delta}\sqrt{3}\sigma}\right)$, which is decreasing in Δ. This is because as Δ rises, the dispersion of the fundamental value increases, lowering the chance that the fundamental value will lie in the execution interval $[a^* - a', a^*]$.

HFTs' profit increases in the arrival rate of the slow buyer by increasing the chance of a trade in the next Δ time interval. Profits decrease in the number of HFTs N, as they are divided by a

larger number of possible liquidity providers. HFTs' profit also decreases in the seller's arrival rate. The reason is that a higher λ_S increases the buyer's reservation value:

$$\frac{\partial \mathbb{E}[SP(a^*)]}{\partial \lambda_S} = -\frac{1}{N}\left(\frac{\sqrt{\Delta}\lambda\sigma \log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right)}{\sqrt{3}\lambda_S(\lambda_S - \frac{\alpha^2\sigma^2}{2})}\right) < 0. \tag{37}$$

The impact of volatility on HFTs' profit is positive. The profit increases if

$$\frac{\partial \mathbb{E}[SP(a^*)]}{\partial \sigma} = \frac{1}{N}\left(\frac{\sqrt{\Delta}\lambda \log\left(\frac{\lambda_S}{\lambda S - \frac{\alpha^2\sigma^2}{2}}\right)}{2\sqrt{3}(\lambda_S - \frac{\alpha^2\sigma^2}{2})} - \frac{\sqrt{\Delta}\lambda\left[\log\left(\frac{\lambda_S}{\lambda S - \frac{\alpha^2\sigma^2}{2}}\right)\right]^2}{4\sqrt{3}\alpha^2\sigma^2}\right), \tag{38}$$

which is positive provided that

$$\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}} - 1 > \frac{1}{4}\log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right). \tag{39}$$

Remembering that $\lambda_S > \frac{\alpha^2\sigma^2}{2}$ and setting $y = \frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}} > 1$, the last inequality is always true, because the function $y - 1 - \frac{1}{4}\log y$ is always positive for $y > 1$.

The effects of latency, order size, and volatility on HFTs' profit are illustrated in Figure 5. The left panel shows that HFTs' profit decreases in latency. The middle panel shows the positive effect of the volatility and the right panel of the order size.

Slow Buyer's Utility and the Probability of Trading

In the following, I compute the expected utility of the slow buyer, V_B. This is in general given by the buyer's utility from submitting a market order in case he executes on the market maker's ask quote and by the utility he derives from submitting a limit order in case (i) the ask price was higher than his reservation ask, or (ii) the high fundamental value created an arbitrage opportunity for the HFBs from which the slow buyer cannot profit:

$$V_B = \mathbb{E}\left[\left(1 - e^{-\alpha(\pi_B + v_\Delta - a^*)}\right) \mathbb{1}_{a^* - a' \leq v_\Delta \leq a^*}\right] + V_{B,LO}\mathbb{P}[a^* - a' > v_\Delta \vee v_\Delta > a^*]. \tag{40}$$

Before computing the buyer's expected utility, it is useful to investigate the drivers of the probability that the trade will happen, $\mathbb{P}[a^* - a' \leq v_\Delta \leq a^*]$.

Lemma VI.2. Equilibrium Probability of Trading. *Given the seller's arrival rate λ_S, the buyer's risk-aversion coefficient α and arrival rate λ, latency Δ and the fundamental value's volatility σ, the probability that a trade between the buyer and the market maker will happen is*

$$\begin{aligned}
\mathbb{P}[a^* - a' \leq v_\Delta \leq a^*] &= \lambda\Delta \int_{a^*-a'}^{a^*} \frac{1}{2\sqrt{3}\sqrt{\Delta}\sigma}\, dx && (41)\\
&= \lambda\sqrt{\Delta}\frac{a'}{2\sqrt{3}\sigma} && (42)\\
&= \frac{\lambda\sqrt{\Delta}\log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right)}{2\sqrt{3}\sigma\alpha}. && (43)
\end{aligned}$$

The probability of trading

1. *decreases in the seller's arrival rate λ_S,*
2. *decreases in latency Δ,*
3. *increases in the buyer's arrival rate λ,*
4. *increases in the asset's volatility σ.*

Proof. The remainder of the proof is outlined below.

□

From the calculation above it can be seen that the probability of trading between the market maker and the buyer does not directly depend on the market maker's ask price a^*. However, it depends on the reservation ask a'. The higher the seller's arrival rate, the lower the reservation ask price and the lower therefore the probability of trading with the market maker:

$$\frac{\partial\mathbb{P}[\cdot]}{\partial\lambda_S} = -\frac{\alpha\sigma\lambda\sqrt{\Delta}}{4\sqrt{3}\lambda_S(\lambda_S - \frac{\alpha^2\sigma^2}{2})} < 0. \tag{44}$$

Increasing the latency or the buyer's arrival rate increases the chance that the buyer will arrive by time Δ. On the other hand, increasing latency leads to a higher dispersion of the fundamental value, reducing the chance that it will fall in the acceptable trading range $[a^* - a', a^*]$, lowering the trading probability. We are interested in the effect per unit of time. Overall, this will be negative. Indeed,

$$\frac{\partial(\mathbb{P}[\cdot]/\Delta)}{\partial\Delta} = -\frac{a'}{4\sqrt{3}\Delta^{\frac{3}{2}}\sigma} < 0. \tag{45}$$

The impact of asset price volatility on the probability of trading is given by

$$\frac{\partial\mathbb{P}[\cdot]}{\partial\sigma} = \lambda\sqrt{\Delta}\frac{\frac{\frac{\alpha^2\sigma^2}{2}}{\lambda_S - \frac{\alpha^2\sigma^2}{2}} - \frac{1}{2}\log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right)}{\sqrt{3}\alpha}. \tag{46}$$

This expression is positive provided that

$$\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}} - 1 > \frac{1}{2}\log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right) \tag{47}$$

Setting $y = \frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}} > 1$, this condition is always met, because the function $y - 1 - \frac{1}{2}\log y$ is always positive for $y > 1$.

The effects of latency, order size, and volatility on the probability of trading are illustrated in Figure 6. The left panel shows that higher latency reduces the probability of trading per unit of time. The middle panel shows the positive impact of the volatility of the fundamental value and the right the positive impact of the order size.

We are now in a position to compute the expected utility of the slow buyer. The expected utility can be divided into two components. The first is the payoff obtained from executing on the market maker's ask price. The second arises from submitting a limit order in case trading with the market maker is no longer the best option for the buyer.

Proposition 4. Equilibrium Slow Trader's Utility. *Given the seller's arrival rate λ_S, the*

buyer's arrival rate λ, his risk-aversion coefficient α and private valuation π_B, latency Δ and the fundamental value's volatility σ, the slow buyer's expected utility derived from a limit order is

$$V_{B,LO}\mathbb{P}[a^* - a' > v_\Delta \vee v_\Delta > a^*] = \left(1 - \frac{\lambda_S e^{-\alpha\pi_B}}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right)\lambda\Delta\left(1 - \frac{\log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right)}{2\sqrt{3}\sqrt{\Delta}\sigma\alpha}\right). \tag{48}$$

The limit order payoff

1. *increases in the seller's arrival rate λ_S if $V_{B,LO}$ is sufficiently large,*
2. *increases in the latency Δ,*
3. *increases in the buyer's arrival rate λ.*

The slow buyer's expected utility derived from submitting a market order $V_{B,MO}$ is

$$\mathbb{E}\left[\left(1 - e^{-\alpha(\pi_B + v_\Delta - a^*)}\right)\mathbb{1}_{a^* - a' \leq v_\Delta \leq a^*}\right] = \lambda\sqrt{\Delta}\frac{\log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right) + e^{-\alpha\pi_B}\left(1 - \frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right)}{2\sqrt{3}\alpha\sigma}. \tag{49}$$

The market order payoff

1. *increases in the buyer's arrival rate λ,*
2. *decreases in the seller's arrival rate λ_S if $V_{B,LO} > 0$,*
3. *decreases in the latency Δ if $e^{\alpha\pi_B} > \frac{\frac{\partial a'}{\partial\sigma^2}}{\frac{a'}{\sigma^2}}$.*

The slow buyer's overall payoff $V_{B,LO}\mathbb{P}[a^* - a' > v_\Delta \vee v_\Delta > a^*] + V_{B,MO}$

1. *increases in the seller's arrival rate λ_S if $2\sqrt{3}\sigma\sqrt{\Delta} > a'$,*
2. *increases in the buyer's arrival rate λ,*
3. *decreases in latency Δ.*

Proof. The proof of Equation (48) and Equation (49) is given in Section B and the comparative statics are derived below.

□

The expected limit order surplus of the slow buyer increases in the arrival rate of the seller as it increases the value of the limit order and decreases the probability of trading with the market maker at the same time, provided that the value of the limit order is sufficiently high. Indeed, one has

$$\frac{\partial}{\partial \lambda_S}\left(V_{B,LO}\mathbb{P}[a^* - a' > v_\Delta \vee v_\Delta > a^*]\right) = \underbrace{\frac{\partial V_{B,LO}}{\partial \lambda_S}}_{>0}\underbrace{\mathbb{P}[\cdot]}_{>0} + \underbrace{\frac{\partial \mathbb{P}[\cdot]}{\partial \lambda_S}}_{>0} V_{B,LO}. \tag{50}$$

$\frac{\partial V_{B,LO}}{\partial \lambda_S}$ equals $e^{-\alpha\pi_B}\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}} - 1\right)$, which is positive since $\lambda_S - \frac{\alpha^2\sigma^2}{2} > 0$. The trade probability is decreasing in λ_S by Equation (44), so the no-trade probability $\mathbb{P}[a^* - a' > v_\Delta \vee v_\Delta > a^*]$ increases in λ_S. The term $V_{B,LO}$ may in general take negative values. It will remain positive if $e^{\alpha\pi_B} > \frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}$, which holds for a broad range of parameters. Thus the buyer's expected limit order surplus is typically increasing in λ_S.

The expected limit order surplus of the slow buyer, $V_{B,LO}\mathbb{P}[a^* - a' > v_\Delta \vee v_\Delta > a^*]$, increases in latency Δ as a whole, because the probability of slow buyer's arrival increases. As in the case of the probability of trading, we are interested in the effects per unit of time. As shown above, because the dispersion of the fundamental value rises with Δ, the probability per unit of time Δ that the fundamental value will fall within the execution range, falls. Thus, the time-normalized effect of an increase in Δ on the expected limit order profit of the slow buyer is positive if the outside utility of the slow buyer is positive, $V_{B,LO} > 0$. Formally,

$$\frac{\partial}{\partial \Delta}\left(V_{B,LO}\frac{\mathbb{P}[a^* - a' > v_\Delta \vee v_\Delta > a^*]}{\Delta}\right) = \log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right)\left(1 - \frac{\lambda_S e^{-\alpha\pi_B}}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right)\frac{\lambda}{4\sqrt{3}\Delta^{\frac{3}{2}}\sigma\alpha} > 0. \tag{51}$$

Furthermore, the slow buyer's expected profit increases in λ as trading between the buyer and the market maker becomes more likely. Formally,

$$\frac{\partial}{\partial \lambda}\left(V_{B,LO}\frac{\mathbb{P}[a^* - a' > v_\Delta \vee v_\Delta > a^*]}{\Delta}\right) = \left(1 - \frac{\lambda_S e^{-\alpha\pi_B}}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right)\left(1 - \frac{\log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right)}{2\sqrt{3}\sqrt{\Delta}\sigma\alpha}\right) > 0. \tag{52}$$

Let us now consider the sensitivity of the market order surplus before analyzing the overall welfare of the slow buyer. The expected market order surplus of the slow buyer is decreasing in λ_S if the outside utility of the slow buyer is positive, $V_{B,LO} > 0$.

$$\frac{\partial}{\partial \lambda_S}\mathbb{E}\left[\left(1-e^{-\alpha(\pi_B+v_\Delta-a^*)}\right)\mathbb{1}_{a^*-a'\leq v_\Delta\leq a^*}\right] = \frac{\alpha\sigma\sqrt{\Delta}e^{-\alpha\pi_B}\left(\alpha^2\sigma^2e^{\alpha\pi_B}-2\lambda_S\left(e^{\alpha\pi_B}-1\right)\right)}{2\sqrt{3}\lambda_S\left(\alpha^2\sigma^2-2\lambda_S\right)^2} \tag{53}$$

This expression is negative provided that

$$1-\frac{e^{-\alpha\pi_B}\lambda_S}{\lambda_S-\frac{\alpha^2\sigma^2}{2}} > 0 \tag{54}$$

i.e. if

$$V_{B,LO} > 0. \tag{55}$$

To analyze the sensitivity of the market order surplus with respect to Δ, we use again the profit per unit of time by dividing the expression by Δ. One has

$$\frac{\partial}{\partial\Delta}\left(\frac{\mathbb{E}\left[\left(1-e^{-\alpha(\pi_B+v_\Delta-a^*)}\right)\mathbb{1}_{a^*-a'\leq v_\Delta\leq a^*}\right]}{\Delta}\right) = -\frac{e^{-\alpha\pi_B\left(e^{\alpha\pi_B}\log\left(\frac{\lambda_S}{\lambda_S-\frac{\alpha^2\sigma^2}{2}}\right)-\frac{\lambda_S}{\lambda_S-\frac{\alpha^2\sigma^2}{2}}+1\right)}}{4\sqrt{3}\alpha\sqrt{\Delta^3}\sigma}. \tag{56}$$

This expression is negative provided that

$$e^{\alpha\pi_B} > \frac{\frac{\lambda_S}{\lambda_S-\frac{\alpha^2\sigma^2}{2}}-1}{\log\left(\frac{\lambda_S}{\lambda_S-\frac{\alpha^2\sigma^2}{2}}\right)} = \frac{\frac{\partial a'}{\partial\sigma^2}}{\frac{a'}{\sigma^2}}. \tag{57}$$

The expected market order surplus of the slow buyer is decreasing in Δ if $e^{\alpha\pi_B} > \frac{\frac{\lambda_S}{\lambda_S-\frac{\alpha^2\sigma^2}{2}}-1}{\log\left(\frac{\lambda_S}{\lambda_S-\frac{\alpha^2\sigma^2}{2}}\right)}$. In words, the expected payoff to a market order will decrease in latency if the private valuation is high enough compared to the elasticity of the reservation ask price with respect to the variance of the fundamental value σ^2.

Turning now to the impact of the model parameters on the slow buyer's overall welfare $V_{B,LO}\mathbb{P}[a^*-$

$a' > v_\Delta \vee v_\Delta > a^*] + V_{B,MO}$, the effect of the seller's arrival rate can be computed as

$$\frac{\partial}{\partial \lambda_S}\left(V_{B,LO}\mathbb{P}[a^* - a' > v_\Delta \vee v_\Delta > a^*] + V_{B,MO}\right) = \frac{\alpha\sigma\sqrt{\Delta}\lambda e^{-\alpha\pi_B}\left(2\sqrt{3}\alpha\sqrt{\Delta}\sigma - \log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right)\right)}{\sqrt{3}\left(\lambda_S - \frac{\alpha^2\sigma^2}{2}\right)^2}. \tag{58}$$

This expression is positive if

$$2\sqrt{3}\sigma\sqrt{\Delta} > a'. \tag{59}$$

Thus, the buyer's overall utility increases in λ_S if the reservation ask price is lower than the dispersion of the fundamental value $2\sqrt{3}\sigma\sqrt{\Delta} > a'$, which can be interpreted as a feasibility condition. The overall expected payoff of the slow buyer also increases in his arrival rate λ as both his expected payoff due to market and limit order increase in λ.

Turning now to the effect of latency Δ, we are again interested in welfare per unit of time. One has:

$$\frac{\partial}{\partial \Delta}\left(\frac{V_{B,LO}\mathbb{P}[a^* - a' > v_\Delta \vee v_\Delta > a^*] + V_{B,MO}}{\Delta}\right) = \frac{\lambda e^{-\alpha\pi_B}\left(\frac{\alpha^2\sigma^2}{2} - \lambda_S \log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right)\right)}{2\sqrt{3}\sqrt{\Delta^3}\alpha\sigma\left(\lambda_S - \frac{\alpha^2\sigma^2}{2}\right)}. \tag{60}$$

This expression is negative if

$$\frac{\alpha^2\sigma^2}{2} < \lambda_S \alpha a'. \tag{61}$$

The overall expected payoff of the slow buyer decreases in latency if the risk-aversion adjustment is lower than a term proportional to his reservation ask price $\frac{\alpha^2\sigma^2}{2} < \lambda_S \alpha a'$. Inserting the value of a' yields the condition

$$\frac{\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}} - 1}{\log\left(\frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}\right)} < \frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}. \tag{62}$$

Setting $y = \frac{\lambda_S}{\lambda_S - \frac{\alpha^2\sigma^2}{2}}$, Equation (62) can be written as $y - 1 - y \log y < 0$. Because $\lambda_S - \frac{\alpha^2\sigma^2}{2} > 0$, y varies between 1 and ∞. At $y = 1$, the above expression equals zero. Futhermore, the derivative of the expression is equal to $-\log y$, which is negative for $y > 1$. Taken together this means that the slow buyer's overall welfare is decreasing in latency.

The effects of latency, order size, and volatility are illustrated in Figure 7. The left panel shows the negative impact of latency on welfare for different order sizes Q. The middle panel shows the impact of latency and volatility, and the right panel the impact of order size and volatility, where we illustrate the effect of Equation (59) by choosing $\Lambda_S = 10$ for the top panels and $\Lambda_S = 100$ for the bottom panels.

VII. Conclusion

Are faster markets better for institutional investors? The current paper presents a model considering the impact of both order size and latency on the price impact of trades. A model is proposed, which is solved analytically and is novel in the theoretical microstructure literature. Larger latency increases adverse selection costs to the market maker and reduces his probability of trading with a slow investor. A larger order size decreases the slow trader's outside option, making him susceptible to accept a worse price for his trade. It is shown that the first order effect of increased latency and increased order size is to increase price impact. Their joint impact is also positive. When the probability of trading is taken into consideration, the utility of the slow institutional investor decreases with latency. Furthermore, this model is surmisable to possible extensions. Natural extensions of this work include taking the order size as an endogenous variable in a dynamic model, creating scope for trading influencing prices every round and the market maker learning from such price changes. It would also be interesting to observe how such a model could be extended to include the market maker's inventory management or competition among different high-frequency traders that are possibly subject to different latencies.

References

Aït-Sahalia, Y. and M. Saglam (2014). High frequency traders: Taking advantage of speed. *NBER Working Paper*.

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

Brogaard, J. (2010). High frequency trading and its impact on market quality. *Northwestern University Working Paper*.

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

Budish, E. B., P. Cramton, and J. J. Shim (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics 130*, 1547––1621.

Chacko, G. C., J. W. Jurek, and E. Stafford (2008). The price of immediacy. *The Journal of Finance 63*(3), 1253–1290.

Du, S. and H. Zhu (2016). Welfare and optimal trading frequency in dynamic double auctions. *National Bureau of Economic Research*.

Foucault, T., J. Hombert, and I. Roşu (2015). News trading and speed. *The Journal of Finance*.

Garman, M. B. (1976). Market microstructure. *Journal of Financial Economics 3*(3), 257 – 275.

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

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

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

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

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

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

Jovanovic, B. and A. J. Menkveld (2011). Middlemen in limit-order markets. *New York University Working Paper*.

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

Lebedev, N. N., R. A. Silverman, and D. Livhtenberg (1972). Special functions and their applications. *Physics Today 18*, 70.

Menkveld, A. J. (2016). The economics of high-frequency trading: Taking stock. *Working paper*.

Menkveld, A. J. and M. A. Zoican (2016). Need for speed? exchange latency and liquidity.

O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics 116*, 257–270.

Riordan, R. and A. Storkenmaier (2012). Latency, liquidity and price discovery. *Journal of Financial Markets 15*, 416–437.

Rostek, M. and M. Weretka (2015). Dynamic thin markets†. *Review of Financial Studies*, hhv027.

Roşu, I. (2016). Fast and slow informed trading. *AFA 2013 San Diego Meetings Paper*.

Tong, L. (2015). A blessing or a curse? The impact of high frequency trading on institutional investors. *Fordham University Working Paper*.

Vayanos, D. (1999). Strategic trading and welfare in a dynamic market. *The Review of Economic Studies 66*(2), 219–254.

Vives, X. (2011). Strategic supply function competition with private information. *Econometrica 79*(6), 1919–1966.

Appendices

A. Price Impact as a Function of Latency

The following proves lemma IV.1 by computing the buyer's reservation utility.

Proof. The buyer's payoff is depicted in Figure 3. The fundamental value, v_{T_S}, is normally distributed with mean zero and variance $\sigma^2 T_S$. The seller, who comes at time T_S will only execute if $v_{T_S} \leq \pi_S$. Put altogether, the buyer's utility from submitting a limit order is given by the expectation of the exponential of a truncated normal variable weighted by the random arrival time T_S, which is exponentially distributed with intensity λ_S per unit of time. Formally, it is given by the following double integral

$$V_{B,LO} = 1 - \int_0^\infty \lambda_S e^{-\lambda_S y} \int_{-\infty}^{\pi_S} \frac{e^{-\alpha(\pi_B + x)}}{\sqrt{2\pi\sigma^2 y}} e^{-\frac{1}{2}\frac{x^2}{\sigma^2 y}} \mathrm{d}x\mathrm{d}y \tag{63}$$

$$= 1 - \lambda_S e^{-\alpha\pi_B} \int_0^\infty e^{-\lambda_S y + \frac{\alpha^2\sigma^2}{2} y} \int_{-\infty}^{\pi_S} \frac{1}{\sqrt{2\pi\sigma^2 y}} e^{-\frac{1}{2}\frac{(x+\alpha\sigma^2 y)^2}{\sigma^2 y}} \mathrm{d}x\mathrm{d}y \tag{64}$$

$$= 1 - \lambda_S e^{-\alpha\pi_B} \int_0^\infty e^{-\lambda_S y + \frac{\alpha^2\sigma^2}{2} y} \Phi\left(\frac{\pi_S + \alpha\sigma^2 y}{\sigma\sqrt{y}}\right) \mathrm{d}y, \tag{65}$$

where $\Phi(\cdot)$ represents the normal CDF. The second equation was obtained by expanding the x^2 by $\alpha\sigma^2 y$. The third equality was obtained by the change of variables $z = x + \alpha\sigma^2 y$. Let's focus on the integral term, which we denote I and solve by parts.

$$I := \int_0^\infty e^{-\lambda_S y + \frac{\alpha^2\sigma^2}{2} y} \Phi\left(\frac{\pi_S + \alpha\sigma^2 y}{\sigma\sqrt{y}}\right) \mathrm{d}y \tag{66}$$

$$= -\frac{1}{\lambda_S - \frac{\alpha^2\sigma^2}{2}} \left[e^{-y(\lambda_S - \frac{\alpha^2\sigma^2}{2})} \Phi\left(\frac{\pi_S + \alpha\sigma^2 y}{\sigma\sqrt{y}}\right) \right]_0^\infty - \frac{(-1)1}{\lambda_S - \frac{\alpha^2\sigma^2}{2}} \int_0^\infty e^{-\lambda_S y + \frac{\alpha^2\sigma^2}{2} y} \frac{\partial}{\partial y} \Phi\left(\frac{\pi_S + \alpha\sigma^2 y}{\sigma\sqrt{y}}\right) \mathrm{d}y \tag{67}$$

$$= \frac{1}{\lambda_S - \frac{\alpha^2\sigma^2}{2}} + \frac{1}{\lambda_S - \frac{\alpha^2\sigma^2}{2}} \int_0^\infty e^{-y(\lambda_S - \frac{\alpha^2\sigma^2}{2}) - \frac{1}{2}(\frac{\pi_S + \alpha\sigma^2 y}{\sigma\sqrt{y}})^2} \left(\frac{\alpha\sigma}{2\sqrt{y}} - \frac{\pi_S}{2\sqrt{y^3}\sigma} \right) \mathrm{d}y \tag{68}$$

$$= \frac{1}{\lambda_S - \frac{\alpha^2\sigma^2}{2}} \left(1 + e^{-\alpha\pi_S} \int_0^\infty e^{-y\lambda_S - \frac{\pi_S}{2\sigma^2 y}} \left(\frac{\alpha\sigma}{2\sqrt{y}} - \frac{\pi_S}{2\sqrt{y^3}\sigma} \right) \mathrm{d}y \right). \tag{69}$$

Here, the second equality comes from solving the I integral by parts. Computing the limits of the left term and the partial derivative for the right term yields the third equality. The fourth results

from collecting terms. Let us again compute the integral part separately.

$$J := \int_0^\infty e^{-y\lambda_S - \frac{\pi_S}{2\sigma^2 y}} \left(\frac{\alpha\sigma}{2\sqrt{y}} - \frac{\pi_S}{2\sqrt{y^3}\sigma} \right) \mathrm{d}y \tag{70}$$

$$= \int_0^\infty e^{-y\lambda_S - \frac{\pi_S}{2\sigma^2 y}} \left(\frac{\alpha\sigma}{2\sqrt{y}} \right) \mathrm{d}y - \int_0^\infty e^{-y\lambda_S - \frac{\pi_S}{2\sigma^2 y}} \left(\frac{\pi_S}{2\sqrt{y^3}\sigma} \right) \mathrm{d}y \tag{71}$$

$$= \left(\frac{\alpha\sigma}{2} \right) \int_0^\infty \frac{1}{\sqrt{y}} e^{-\frac{1}{2}(2\lambda_S y + \frac{\pi_S}{\sigma^2 y})} \mathrm{d}y - \left(\frac{\pi_S}{2\sigma} \right) \int_0^\infty \frac{1}{\sqrt{y^3}} e^{-\frac{1}{2}(\lambda_S y + \frac{\pi_S}{\sigma^2 y})} \mathrm{d}y \tag{72}$$

$$= J_1 - J_2. \tag{73}$$

The above separation is useful, because the two integrals are special Bessel functions found in Lebedev et al. (1972) Sects. 8.432 6 p. 959, and 8.469 3 p. 967:

$$\int_0^\infty \frac{1}{\sqrt{t}} e^{-\frac{\delta}{2}(t + \frac{1}{t})} \mathrm{d}t = 2\sqrt{\frac{\pi}{2\delta}} e^{-\delta}, \tag{74}$$

$$\int_0^\infty \frac{1}{\sqrt{t^3}} e^{-\frac{1}{2}(\beta t + \frac{\gamma}{t})} \mathrm{d}t = \sqrt{\frac{2\pi}{\gamma}} e^{-\sqrt{\beta\gamma}}, \tag{75}$$

$$\tag{76}$$

where $\beta = 2\lambda_S$, $\gamma = \frac{\pi_S}{\sigma^2}$ and $\delta = \sqrt{\beta\gamma}$. Hence,

$$J_1 = \frac{\alpha\sigma}{2} \sqrt{\frac{\pi}{\lambda_S}} e^{\sqrt{-2\lambda_S \frac{\pi_S}{\sigma^2}}}, \tag{77}$$

$$J_2 = \sqrt{\frac{\pi}{2}} \sqrt{\pi_S} e^{-\sqrt{2\lambda_S \frac{\pi_S}{\sigma^2}}}. \tag{78}$$

Inserting $J = J_1 - J_2$ back into the integral $I := \frac{1}{\lambda_S - \frac{\alpha^2\sigma^2}{2}} \left(1 + e^{-\alpha\pi_S} J\right)$ and I back into the utility of submitting a limit order, $V_{B,LO} = 1 - \lambda_S e^{-\alpha\pi_B} I$ yields the following

$$V_{B,LO} = 1 - \frac{\lambda_S e^{-\alpha\pi_B}}{\lambda_S - \frac{\alpha^2\sigma^2}{2}} \left(1 + e^{-\alpha\pi_S} e^{-\sqrt{\frac{2\lambda_S\pi_S}{\sigma^2}}} \left[\frac{\alpha\sigma}{2} \sqrt{\frac{\pi}{\lambda_S}} - \sqrt{\frac{\pi}{2}} \sqrt{\pi_S} \right] \right), \tag{79}$$

which concludes the proof of the buyer's reservation utility from submitting a limit order and waiting for a seller. □

B. Welfare Analysis of Market Maker's Latency

Proof. The following proves the expected utility of the slow buyer from a market order, Equation (49):

$$\mathbb{E}\left[1-e^{-\alpha(\pi_B+v_\Delta-a^*)}|a^*-a'\le v_\Delta\le a^*\right]=\lambda\Delta\int_{a^*-a'}^{a^*}\frac{1-e^{-\alpha(-a^*+\pi_B+x)}}{2\sqrt{3\Delta}\sigma}\,\mathrm{d}x \tag{80}$$

$$=\lambda\Delta\frac{e^{-\alpha\pi_B}\left(\alpha a' e^{\alpha\pi_B}-e^{\alpha a'}+1\right)}{2\sqrt{3}\alpha\sqrt{\Delta}\sigma} \tag{81}$$

$$=\lambda\sqrt{\Delta}\frac{\log\left(\frac{\lambda_S}{\lambda_S-\frac{\alpha^2\sigma^2}{2}}\right)+e^{-\alpha\pi_B}\left(1-\frac{\lambda_S}{\lambda_S-\frac{\alpha^2\sigma^2}{2}}\right)}{2\sqrt{3}\alpha\sigma}. \tag{82}$$

The first and second equalities follow directly from computing the expectation. The third equality is obtained by plugging in the value for the reservation ask price a' from lemma IV.2. □

Proof. The following proves the expected utility of the slow buyer from a limit order Equation (48). Let us first compute the probability of not executing upon slow buyer's arrival:

$$\mathbb{P}[a^*-a'>v_\Delta\vee v_\Delta>a^*]=\lambda\Delta\left(\int_{-\sqrt{3\Delta}\sigma}^{a^*-a'}\frac{1}{2\sqrt{3\Delta}\sigma}\,\mathrm{d}x+\int_{a}^{\sqrt{3\Delta}\sigma}\frac{1}{2\sqrt{3\Delta}\sigma}\,\mathrm{d}x\right) \tag{83}$$

$$=\lambda\Delta\left(1-\frac{a'}{2\sqrt{3}\sqrt{\Delta}\sigma}\right) \tag{84}$$

$$=\lambda\Delta\left(1-\frac{\log\left(\frac{\lambda_S}{\lambda_S-\frac{\alpha^2\sigma^2}{2}}\right)}{2\sqrt{3}\alpha\sqrt{\Delta}\sigma}\right). \tag{85}$$

The first and second equalities follow directly from computing the probability. The third equality is obtained by plugging in the value for the reservation ask price a' from lemma IV.2.
Multiplying this probability by the expected payoff from limit order $V_{B,LO}$ given in lemma IV.1 yields

$$V_{B,LO}\mathbb{P}[a^*-a'>v_\Delta\vee v_\Delta>a^*]=\lambda\Delta\left(1-\frac{\log\left(\frac{\lambda_S}{\lambda_S-\frac{\alpha^2\sigma^2}{2}}\right)}{2\sqrt{3}\alpha\sqrt{\Delta}\sigma}\right)\left(1-\frac{\lambda_S e^{-\alpha\pi_B}}{\lambda_S-\frac{\alpha^2\sigma^2}{2}}\right). \tag{86}$$

□

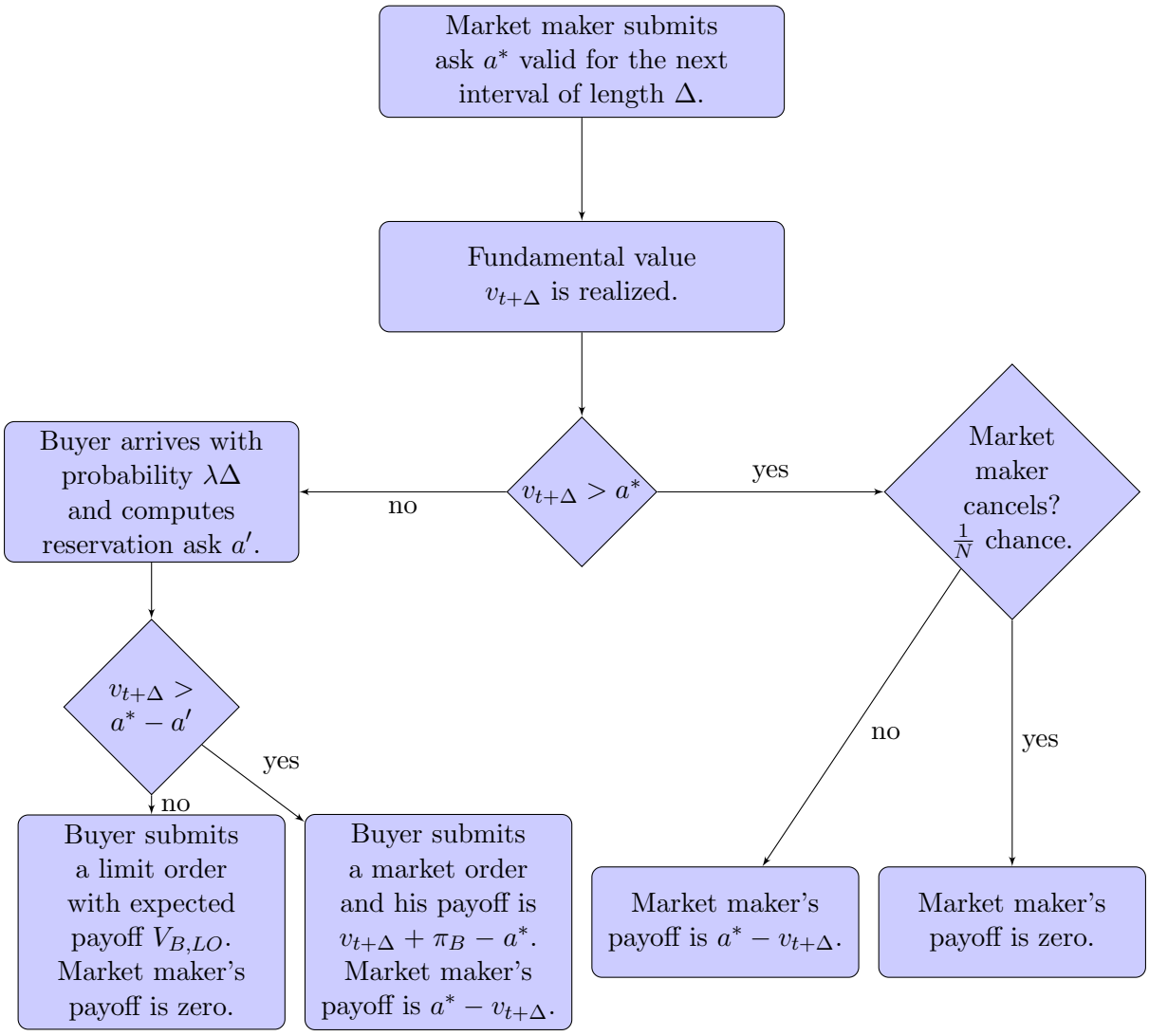

Figure 1: Flow of events. For simplicity, the time t fundamental value is normalized to zero, $v_t = 0$.

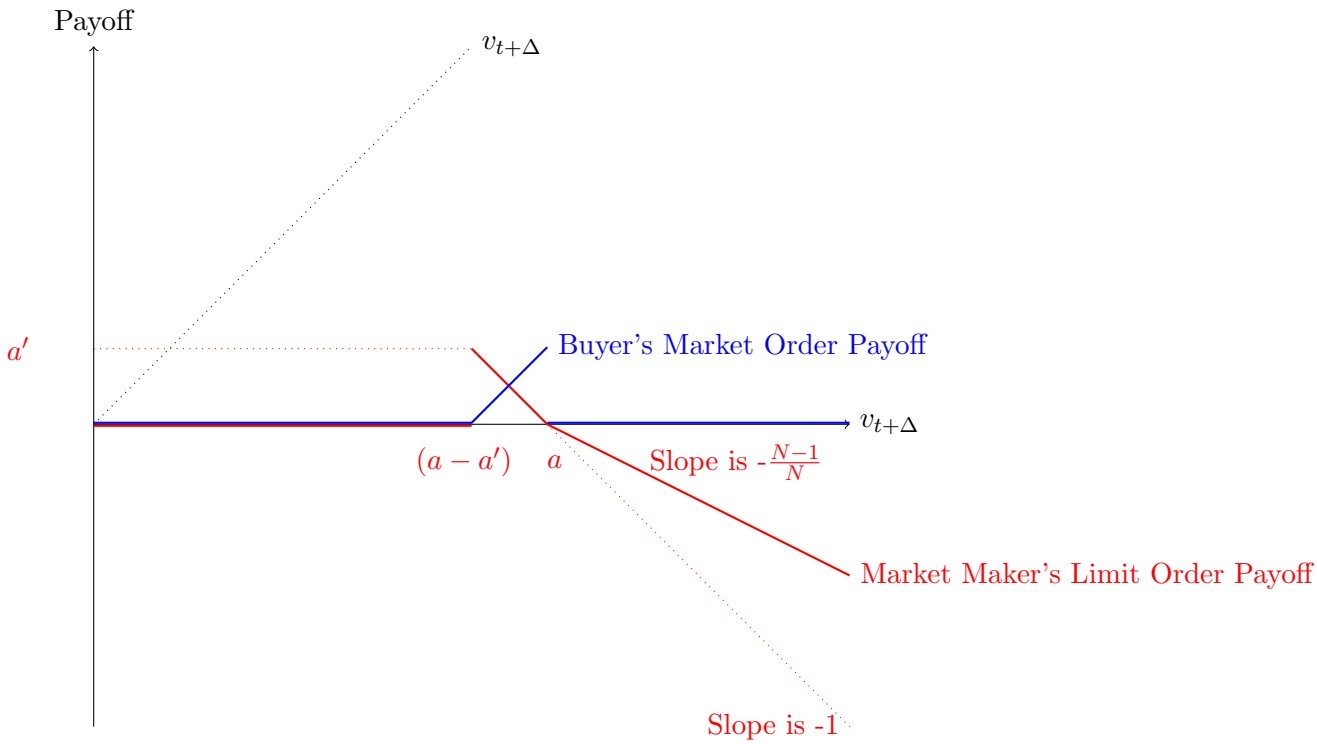

Figure 2: Market maker's and buyer's payoff structure as a function of the fundamental value $v_{t+\Delta}$. There are N high-frequency traders in the market. Market maker submits ask at price a and the maximum price at which the buyer is willing to buy is the reservation ask, a'.

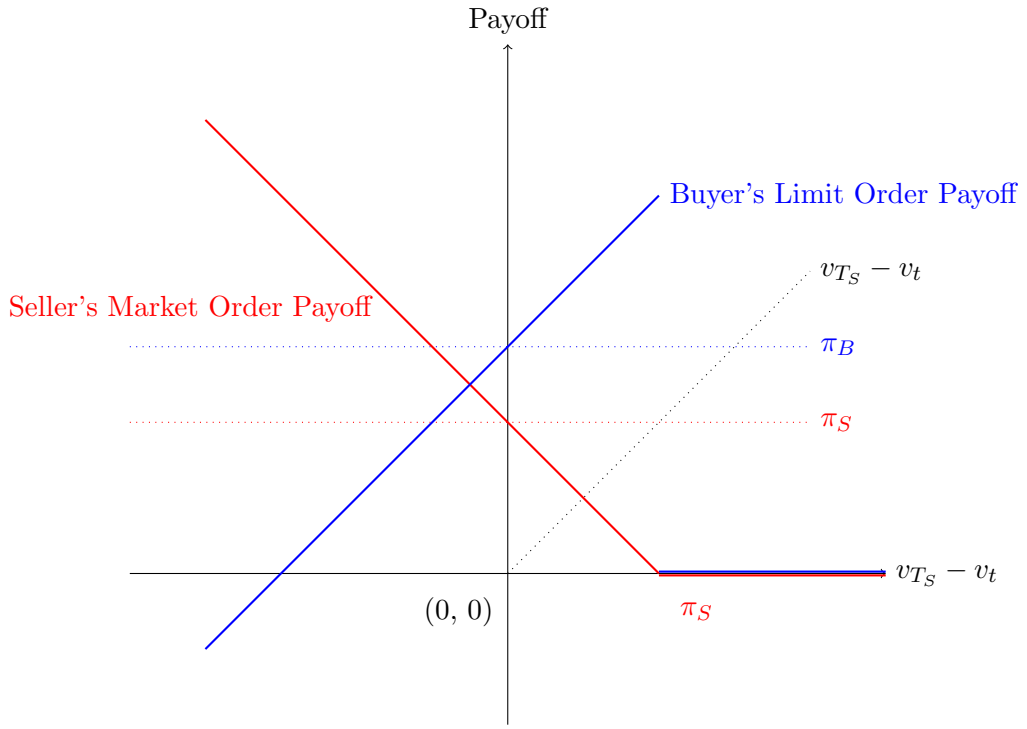

Figure 3: Buyer's and seller's payoff structure as a function of the change in the fundamental value $v_{T_S} - v_t$. The buyer submits a bid at price v_t and the minimum price at which the seller is willing to sell is his private valuation π_S.

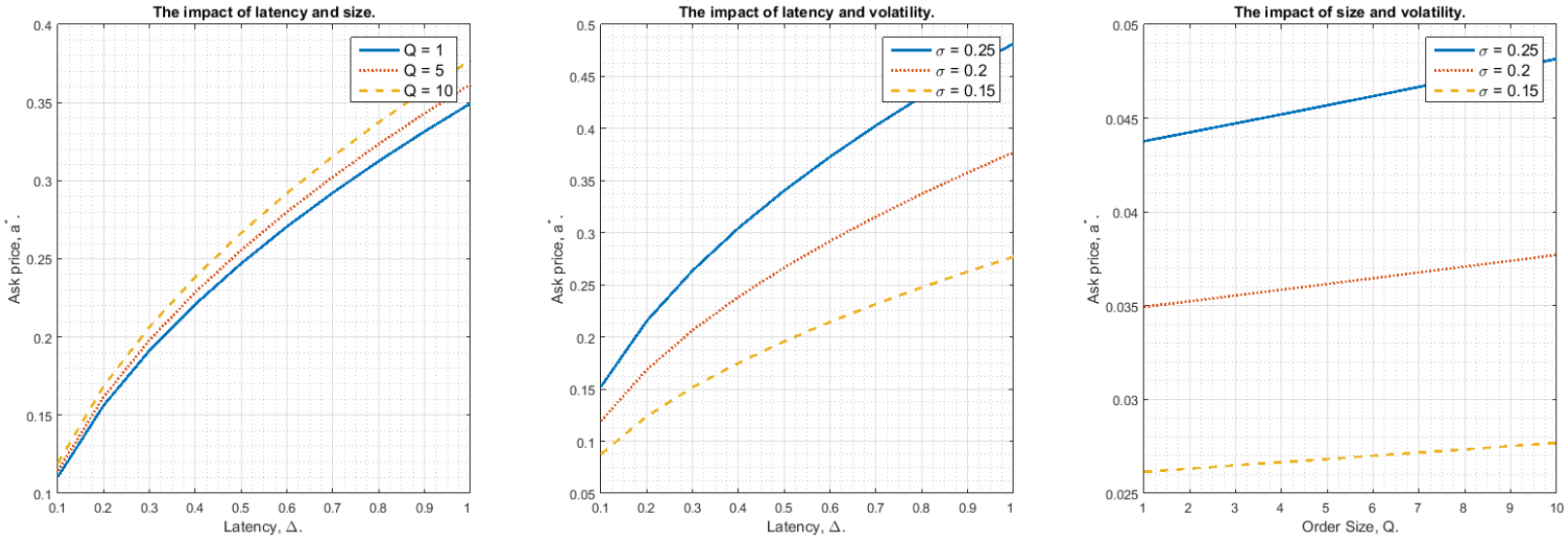

Figure 4: **Market maker's ask price as a function of latency Δ and order size Q.** This figure illustrates the sensitivity of the ask price a^* to the latency Δ, order size Q, and volatility σ. The base parameters chosen are $\sigma = 0.2$, $\lambda = 1$, $\Lambda_S = 10$, and $Q = 1$. The left panel shows the impact of latency and order size. The middle panel shows the impact of latency and volatility, and the right panel the impact of order size and volatility.

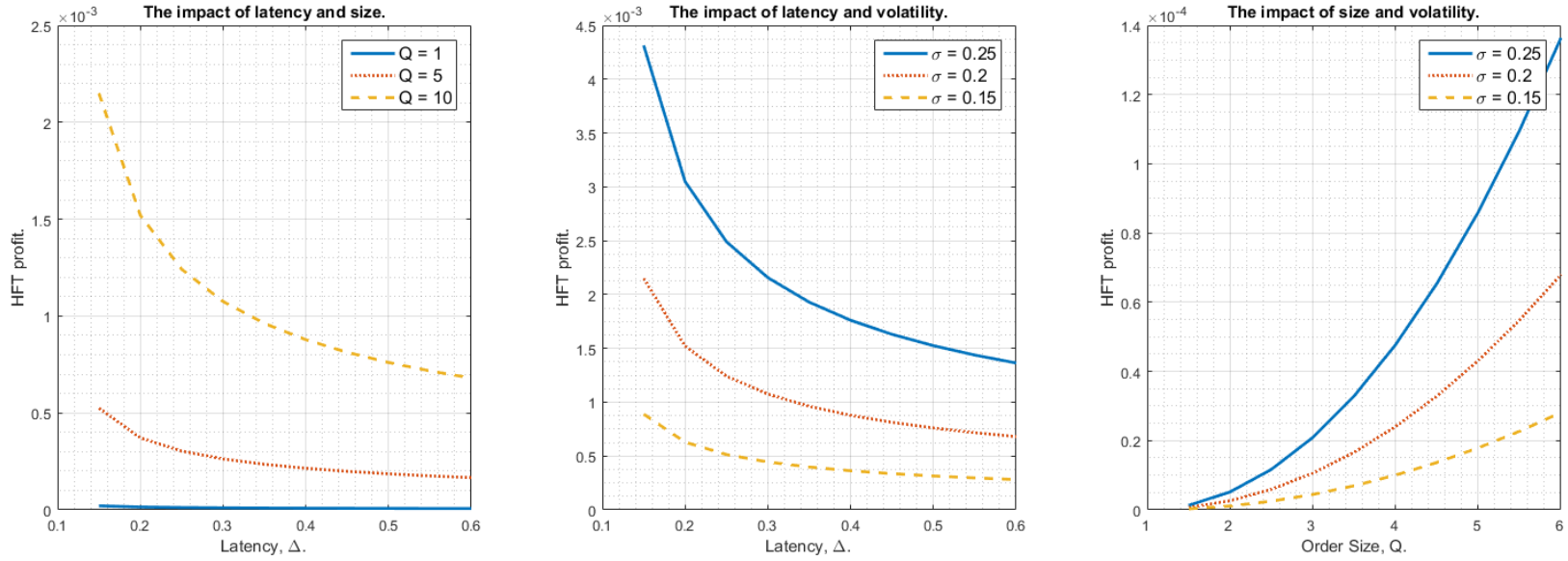

Figure 5: **Expected market maker's profit as a function of latency Δ and order size Q.** This figure illustrates the sensitivity of the market maker's profit to the latency Δ, order size Q, and volatility σ. The base parameters chosen are $\sigma = 0.2$, $\Lambda_S = 10$, and $Q = 1$. The left panel shows the impact of latency and order size. The middle panel shows the impact of latency and volatility, and the right panel the impact of order size and volatility.

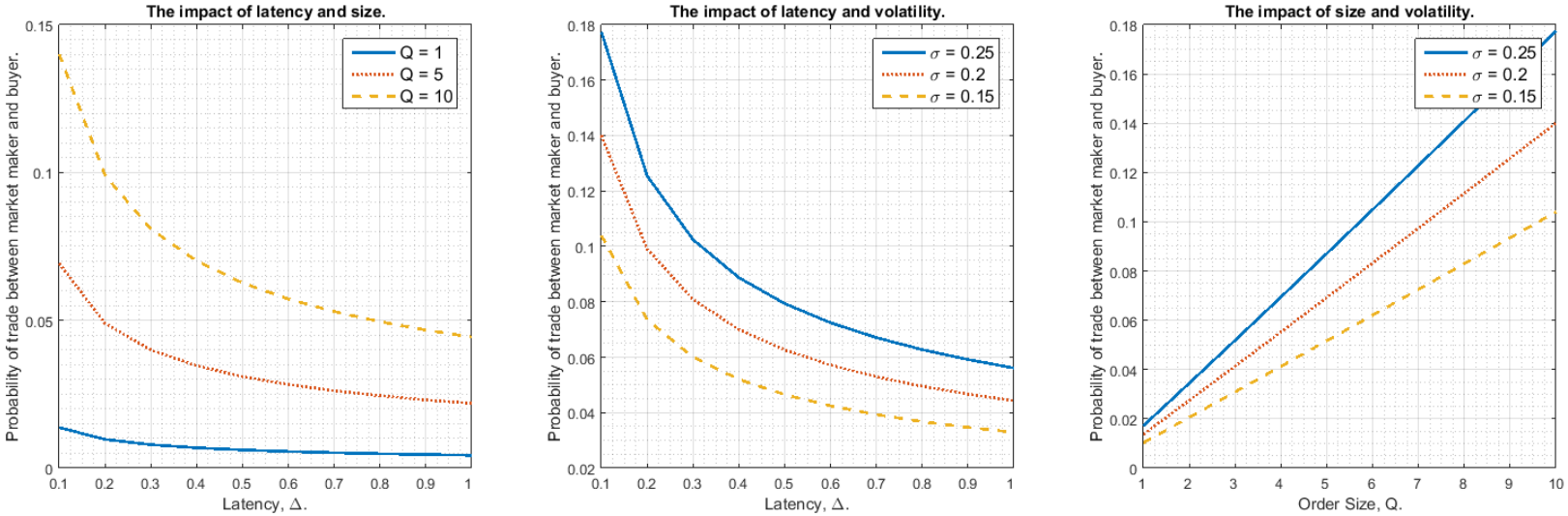

Figure 6: **Probability of trade between the market maker and the buyer as a function of latency Δ and order size Q.** This figure illustrates the sensitivity of the probability of trading to the latency Δ, order size Q, and volatility σ. The base parameters chosen are $\sigma = 0.2$, $\Lambda_S = 10$, $Q = 1$, and $\lambda = 1$. The left panel shows the impact of latency and order size. The middle panel shows the impact of latency and volatility, and the right panel the impact of order size and volatility.

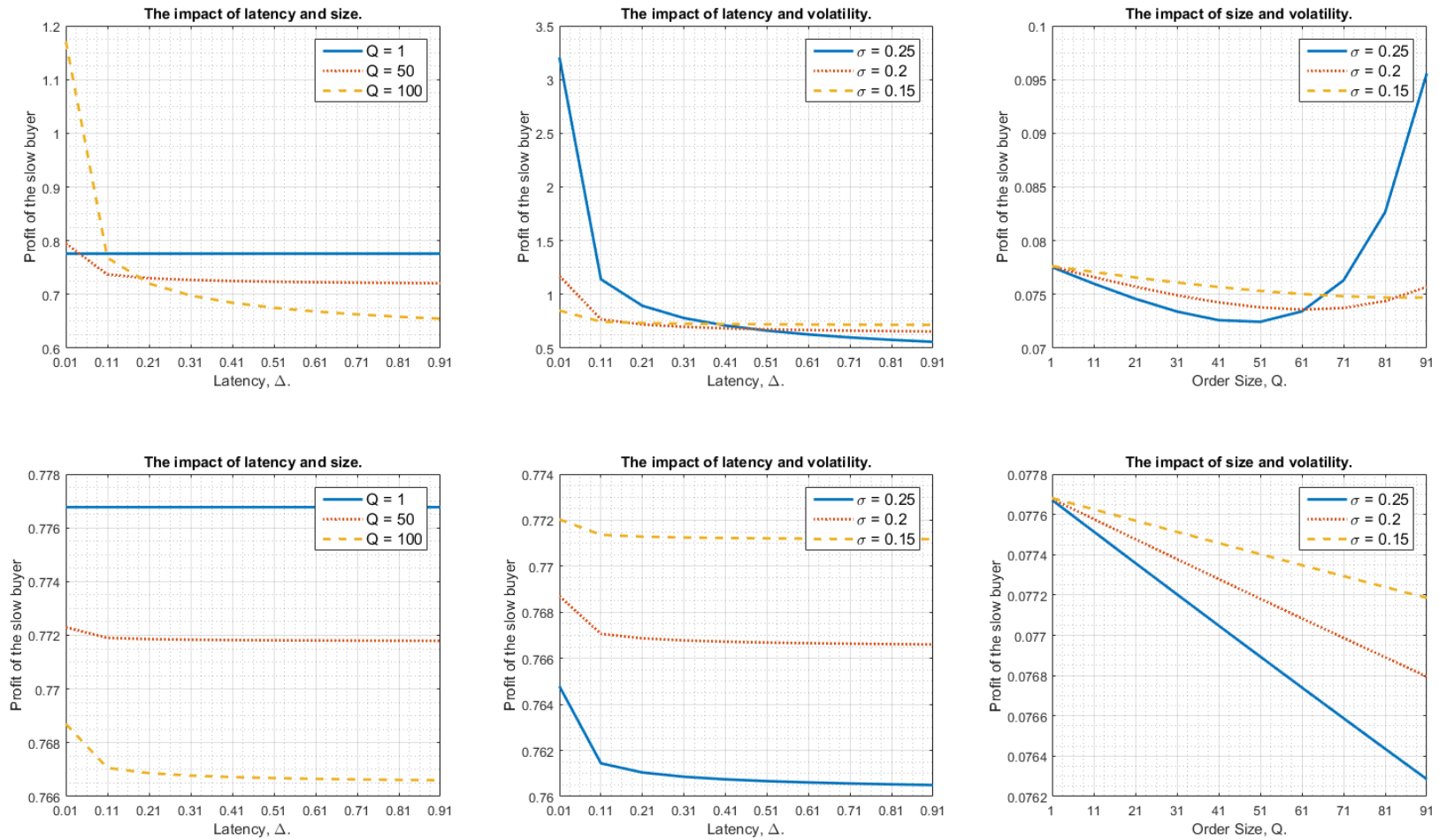

Figure 7: **Slow buyer's expected profit as a function of latency Δ and order size Q.** This figure illustrates the sensitivity of the slow buyer's profit to latency Δ, order size Q, and volatility σ. The base parameters chosen are $\sigma = 0.2$, $\Lambda_S = 10$, and $Q = 1$. The left panel shows the impact of latency and order size. The middle panel shows the impact of latency and volatility, and the right panel the impact of order size and volatility. The top three panels are computed with $\Lambda_S = 10$ and the bottom three panels with $\Lambda_S = 100$.