The Cost of Latency in High-Frequency Trading*

Ciamac C. Moallemi
Graduate School of Business
Columbia University
email: ciamac@gsb.columbia.edu

Mehmet Sağlam
Bendheim Center for Finance
Princeton University
email: msaglam@princeton.edu

Current Revision: February 5, 2013

Abstract

Modern electronic markets have been characterized by a relentless drive towards faster decision making. Significant technological investments have led to dramatic improvements in latency, the delay between a trading decision and the resulting trade execution. We describe a theoretical model for the quantitative valuation of latency. Our model measures the trading frictions created by the presence of latency, by considering the optimal execution problem of a representative investor. Via a dynamic programming analysis, our model provides a closed-form expression for the cost of latency in terms of well-known parameters of the underlying asset. We implement our model by estimating the latency cost incurred by trading on a human time scale. Examining NYSE common stocks from 1995 to 2005 shows that median latency cost across our sample roughly tripled during this time period. Furthermore, using the same data set, we compute a measure of implied latency, and conclude that the median implied latency decreased by approximately two orders of magnitude. Empirically calibrated, our model suggests that the reduction in cost achieved by going from trading on a human time scale to a low latency time scale is comparable with other execution costs faced by the most cost efficient institutional investors, and is consistent with the rents that are extracted by ultra low latency agents, such as providers of automated execution services or high frequency traders.

1. Introduction

In the past decade, electronic markets have become pervasive. Technological advances in these markets have led to dramatic improvements in latency, or, the delay between a trading decision and the resulting trade execution. In the past 30 years, the time scale over which a trade is processed has gone from minutes[1] to milliseconds[2] — "low latency" in a contemporary electronic market would be qualified as under 10 milliseconds, "ultra low latency" as under 1 millisecond. This change represents a dramatic reduction by *five orders of magnitude.* To put this in perspective, human reaction time is thought to be in the hundreds of milliseconds.

One factor behind this trend has been competition between exchanges, as one mechanism for differentiation between exchanges is latency. This competition is driven by a significant demand

*The first author would like to thank Jim Gatheral for a helpful discussion that motivated this work. The authors would like to thank Albert Menkveld, Larry Glosten, and Jialin Yu for their helpful comments.

[1]NYSE, pre-1980 upgrade (Easley et al., 2008).

[2]"The value of a millisecond: Finding the optimal speed of a trading infrastructure," TABB Group, April 2008.

amongst a class of investors, sometimes called "high frequency" traders, for low latency trade execution. High frequency traders are thought to account for more than half of all US equity trades.[3] They expend significant resources in order to develop algorithms and systems that are able to trade quickly. For example, on the time scale of milliseconds, the speed of light can become a binding constraint on the delay in communications. Hence, traders seeking low latency will "co-locate", or house their computers in the same facility as the exchange, in order eliminate delays due to a lack of physical proximity. This co-location comes at a significant expense, however it has been stated that a 1 millisecond advantage can be worth $100 million to a major brokerage firm.[4]

There has been much discussion of the importance of latency among various market participants, regulators, and academics. Despite the significant amount of recent interest, however, latency remains poorly understood from a theoretical perspective. For example, how does latency relate to transaction costs? Is latency only relevant to investors with short time horizons, such as high frequency traders, or does latency also affect long term investors such as pension funds and mutual funds? Many of these important questions have been considered in anecdotal or ad hoc discussions. Our goal here is to provide a framework for quantitative analysis of these issues.

In particular, we wish to understand the benefit to a single trader in the marketplace of lowering their latency, while holding everything else fixed. This is a different question than understanding the social costs of latency, i.e., whether in equilibrium the collective marketplace is better or worse off given lower latency. One might imagine, for example, that the benefit to a individual agents of lower latency may diminish in an equilibrium setting. Equilibrium or welfare analysis of low latency trading is a complex question with important policy and regulatory implications. We believe that understanding the single-agent effects of low latency trading, however, is an important first step which will inform our ultimate understanding of collective effects.

The cost that a trader bears due to latency can take many different forms, depending on the precise trading strategy. However, we can identify a number of broad themes,[5] sometimes overlapping, as to why the ability to trade with low latency might be valuable to an investor:

1. **Contemporaneous decision making.** A trader with significant latency will be making trading decisions based on information that is stale.

   For example, consider an automated trader implementing a market-making strategy in an electronic limit order book. The trader will maintain active limit orders to buy and sell. The prices at which the trader is willing to buy or sell will naturally depend on, say, the limit orders submitted by other investors, the price of the asset on other exchanges, the price of related assets, overall market factors, etc. If the trader cannot update his orders in a timely fashion in response to new information, he may end up trading at disadvantageous prices.

2. **Comparative advantage/disadvantage.** The ability to trade with low latency in absolute

[3] "Stock traders find speed pays, in milliseconds," *New York Times*, July 23, 2009.
[4] "Wall Street's quest to process data at the speed of light," *Information Week*, April 21, 2007.
[5] See Cespa and Foucault (2008) for a related discussion.

terms may not be as important as the ability to trade with low *relative* latency, that is, as compared to competitors.

For example, consider a program trader implementing an index arbitrage strategy, seeking to profit on the difference between an index and its underlying components. There may be many market participants pursuing such strategies and identifying the same discrepancies. The challenge for the trader is to be able to act in the marketplace to exploit a discrepancy *before* a price correction takes place, i.e., before competitors are able to act. The means having a low relative latency.

3. **Time priority rules.** Many modern markets treat orders differentially based on the time of arrival, and favor earlier orders.

   For example, in an electronic limit order book, the limit orders on each side of the market are prioritized in a particular way. When a market order to buy arrives, it is matched against the limit orders to sell according to their priorities. Priority is first determined by price, i.e., limit orders with more lower prices receive higher priority. In many markets, however, prices are mandated to be discrete with a minimum tick size. In these markets, there may be multiple limit orders at the same price, which are then prioritized according to the time of their arrival. While a trader can always increase the priority of his orders by decreasing price, this comes at an obvious cost. If a trader can submit orders in a faster fashion, however, he can increase priority while maintaining the *same* price. Higher priority can be valuable for two reasons: first, higher priority orders have a higher likelihood of execution over any given time horizon. To the extent that investors submitting limit orders have a desire to trade, and to trade sooner rather than later, this is desirable. Second, higher priority orders at the same price level experience less adverse selection (see, e.g., Glosten, 1994; Sandås, 2001). Hence, all things being equal, an investor who submits orders with lower latency will benefit from higher priority than if that investor had higher latency. This can be particularly important (in that a small improvement in latency can result in a significant difference in priority) when an existing quote is about to change. For example, consider the situation where a stock price is about to move up because of trades or cancellations at the best offered price. One might expect the bid price to rise as well, there will be a race among traders reacting to the same order book events to establish time priority at the new bid.

In this paper, we will quantify the cost of latency due to the first effect, a lack of contemporaneous decision making. We do not consider effects of latency that arise from strategic considerations, or from time priority rules or price discreteness. It is an open question as to whether the other effects are more or less significant than the first, and their relative importance may depend on the particular investor and their trading strategy. Our analysis does not speak to this point. However, in what follows we will demonstrate that, by itself, the lack of contemporaneous decision making can induce trading costs that are of the same order of magnitude as other execution costs faced by large investors, and hence cannot be neglected.

Further, the importance of contemporaneous decision making will certainly vary from investor to investor. We will focus on an aspect of this that is universal, however, which is the importance of timely information for the execution of *contingent orders*. A contingent order, such as a limit order in an electronic limit order book or a resting order in a dark pool, presents the possibility of uncertain execution over an interval of time in exchange for price improvement relative to a market order, which executes immediately and with certainty. Specifically, when an investor employs a contingent order, the investor may be exposed to the realization of new information (for example, in the form of price movements, news, etc.) over the lifespan of the order. Latency, which prevents the investor from continuously and instantaneously accessing the market so as to update the order, can thus adversely impact the investor.

As a broad proxy for understanding the importance of latency in contingent order execution, we consider the effects of latency in an extremely simple yet fundamental trade execution problem: that of a risk-neutral investor who wishes to sell 1 share of stock (i.e., an atomic unit) over a fixed, short time horizon (i.e., seconds) in a limit order book, and must decide between market orders and limit orders. Our problem formulation is reminiscent of barrier-diffusion models for limit order execution (e.g., Harris, 1998). It captures the fundamental *cost of immediacy* of trading (e.g., Grossman and Miller, 1988; Chacko et al., 2008), that is, the premium due to a patient liquidity supplier (who submits limit orders) relative to an impatient demander of liquidity (who submits market orders). While this problem is quite stylized, we will argue that it is broadly relevant since, at some level, *all* investors make such a choice of immediacy. For example, it may not seem at first glance that our execution problem is relevant for a pension fund that trades large blocks of stock over multiple days. However, the execution of a block trade via algorithmic trading involves the division of a large "parent" order into many atomic orders over the course of a day, each of these atomic "child" orders can be executed as limit orders or as market orders.

In our problem, in the absence of latency, the optimal strategy of the seller is a "pegging" strategy: the seller maintains a limit order at a constant spread above the bid price at any instant in time. We consider this case as a benchmark. In the presence of latency, the seller can no longer maintain continuous contact with the market so as to track the bid price in the market. The seller is forced to deviate from the benchmark policy in order to take into account the uncertainty introduced by the latency delay by incorporating a safety margin and lowering his limit order prices. The friction introduced by latency thus results in a loss of value to the seller. We will establish the difference in value to the seller between the case with latency and the benchmark case via dynamic programming arguments, and thus provide a quantification of the effects of latency.

The contributions of this paper are as follows:

- *We mathematically quantify the cost of latency.*

  The trading problem we consider (deciding between limit and market orders) is faced by all large investors in modern equity markets, either directly (e.g., high frequency traders) or indirectly (e.g., pension funds who execute large trades via providers of automated execution services). Our analysis suggests that latency impacts all of these market participants, and that, all else being equal, the ability to trade with low latency results in quantifiably lower transaction costs. Further, when calibrated with market data, the latency cost we measure can be significant. It is of the same order of magnitude as other trading costs (e.g., commissions, exchange fees, etc.) faced by the most cost efficient large investors. Moreover, it is consistent with the rents that are extracted by agents who have made the requisite technological investments to trade with ultra low latency. For example, the latency cost of our model is comparable to the execution commissions charged by providers that offer algorithmic trade execution services on an agency basis. It is also comparable to the reported profits of high frequency traders.

To our knowledge, our model is the first to provide a quantification of the costs of latency in trade execution.

- *We provide a closed-form expression for the cost of latency as a function of well-known parameters of the asset price process.*

  The cost of latency in our model can be computed numerically via dynamic programming. However, in the regime of greatest interest, where the latency is close to zero, we provide a closed-form asymptotic expression. In particular, define the *latency cost* associated with an asset as the costs incurred due to latency as a fraction of the overall cost of immediacy (the premium paid to a patient liquidity supplier by an impatient demander of liquidity). Given a latency of Δt, a price volatility of σ, and a bid-offer spread of δ, the latency cost takes the form

$$\frac{\sigma\sqrt{\Delta t}}{\delta}\sqrt{\log\frac{\delta^2}{2\pi\sigma^2\Delta t}} \tag{1}$$

  as $\Delta t \to 0$.

- *Our method can provide qualitative insight into the importance of latency.*

  From (1), it is clear that the latency cost is an increasing function of the ratio of the standard deviation of prices over the latency interval (i.e., $\sigma\sqrt{\Delta t}$) to the bid-offer spread. Latency has a more important role when trading assets that are either more volatile (σ large) or, alternatively, more liquid (δ small). Further, as the latency approaches 0, the marginal benefit of latency reduction is increasing.

- *We empirically demonstrate that latency cost incurred by trading on a human time scale has dramatically increased for U.S. equities and the implied latency of a representative trader in this market decreased by approximately two orders of magnitude.*

We consider the cost due to the latency of trading on the time scale of human interaction.Using the data-set of Aït-Sahalia and Yu (2009), we estimate the latency cost of NYSE common stocks over the 1995–2005 period. We show that the median latency cost roughly tripled in this time. This coincides with a period of decreasing tick sizes and increasing algorithmic and high frequency trading activity (Hendershott et al., 2010).

An alternative perspective is to consider a hypothetical investor who fixes a target level of cost due to latency, relative to the overall cost-of-immediacy. The representative trader maintains this target over time through continual technological upgrades to lower levels of latency. We determine the requisite level of *implied latency* for such a trader, over time and across the aggregate market. Using the same data-set, we observe that the median implied latency decreased by approximately two orders of magnitude over this time frame.

The rest of this paper is organized as follows: In Section 1.1, we review the related literature. In Section 2, as a starting point, we present a stylized, continuous-time trade execution problem in the absence of latency. We develop a variation of the model with latency in Section 3. In Section 4, we provide a mathematical analysis of the optimal policy for our problem. By contrasting the results in the presence and absence of latency, we are able to quantitatively assess the cost of latency. In Section 5, we consider some empirical applications of the model. Finally, in Section 6 we conclude and discuss some future directions.

1.1. Related Literature

There has been a significant empirical literature studying, broadly speaking, the effects of improvements in trading technology. Closest to the aspect we consider is the work of Easley et al. (2008). They empirically test the hypothesis that latency affects asset prices and liquidity by examining the time period around an upgrade to the New York Stock Exchange technological infrastructure that reduced latency. Hendershott et al. (2010) explore the more general, overall effects of algorithmic and high frequency trading. Hasbrouck and Saar (2009) provide different evidence of changes in investor trading strategies that may be a result of improved technology. In subsequent work, they further consider the impact of measurements of low latency on market quality (Hasbrouck and Saar, 2010). Hendershott and Riordan (2009) analyze the impact of algorithmic trading on the price formation process using a data set from Deutsche Börse and conclude that algorithmic trading assists in the efficient price discovery without increasing the volatility. Kirilenko et al. (2010) consider the impact of high frequency trading on the 'flash crash' of 2010, while Brogaard (2010) more broadly examines the impact of high frequency traders on market quality.

On the theoretical front, Cespa and Foucault (2008) consider a rational expectations equilibrium between investors with different access to past transaction data. Some investors observe transactions in real-time, while others only observe transactions with a delay. This model of latency focuses on latency of the price ticker of past transactions, as opposed to latency in execution, which we consider here. Moreover, the goals of the two models differ significantly: Cespa and Foucault (2008) seek to build intuition regarding the equilibrium welfare implications of differential access to information via a structural model. We, on the other hand, seek a reduced form model that can be used to directly estimate the value of execution latency in a particular real world instance, given readily available data. Also related is the work of Ready (1999) and Stoll and Schenzler (2006), who consider the ability of intermediaries (e.g., specialists or dealers) to delay customer orders for their own benefit, thus creating a "free option" in the presence of execution latency. Cohen and Szpruch (2011) show that latency arbitrage exists between two traders with different speeds of trading in the presence of a limit order book. Finally, Cvitanić and Kirilenko (2010) and Jarrow and Protter (2011) consider the effect of high frequency traders on asset prices.

The trade execution problem we consider is that of an investor who wishes to sell a single share of and must decide between market and limit orders. This problem has been considered by many others (e.g., Angel, 1994; Harris, 1998; Lo et al., 2002). Our formulation is similar to the class of barrier-diffusion models considered by these authors; Hasbrouck (2007) provides a good account of this line of work. For a broad survey on limit order markets, see Parlour and Seppi (2008). In our model, the inability to trade continuously gives a limit order an option-like quality that relates execution cost, order duration, and asset volatility. This idea goes as far back as the work of Copeland and Galai (1983). Closely related is the concept of the cost of immediacy, or, the premium paid by a liquidity demander via a market order to a liquidity supplier who posts a limit order. Grossman and Miller (1988) and Chacko et al. (2008) develop theoretical explanations of the cost of immediacy. For empirical evidence of the demand for immediacy in capital markets, see Bacidore et al. (2003) and Werner (2003).

Finally, also related is work on the discrete-time hedging of contingent claims with or without transaction costs (e.g., Boyle and Emanuel, 1980; Leland, 1985; Bertsimas et al., 2000). This literature addresses a different problem and draws different conclusions than our paper, however both relate to implications of a lack of continuous access to the market.

2. A Stylized Execution Model without Latency

Our goal is to understand the impact on the trade execution of latency. To this end, we will first describe a trade execution problem in the absence of latency. In Section 3, we will revisit this model in the presence of latency, so as to understand the resulting trade friction that is introduced. The spirit of our model it to consider an investor who wants to trade, but at a price that depends on an informational process that evolves stochastically and must be monitored continuously. We could directly consider such an abstract model of investor behavior. Instead, however, we will motivate the informational dependence of the trader through a specific optimal execution problem.

Consider the following stylized execution problem of an uninformed trader who must sell exactly one share[6] of a stock over a time horizon $[0, T]$. At any time $t \in [0, T)$, the trader can take one of two actions:

1. The trader can submit a market order to sell. This order will execute at the best bid price at time t, denoted by S_t. We assume that the bid price evolves according to

$$S_t = S_0 + \sigma B_t, \tag{2}$$

where the process $(B_t)_{t \in [0,T]}$ is a standard Brownian motion and $\sigma > 0$ is an (additive) volatility parameter. Here, the choice of Brownian motion is made for simplicity; our model can be extended to the more general class of Markovian martingales, as discussed in Section 4.4.

2. The trader can choose to submit a limit order to sell. In this case, the trader must also decide the limit price associated with the order, which we denoted by L_t.

Once the trader sells one share, he exits the market. If the trader is not able to sell 1 share before time T, however, we assume that he is forced sell via a market order at time T, and therefore receives S_T. Here, we imagine the time horizon T to be small, on the order of the typical trade execution time (i.e., seconds).

2.1. Limit Order Execution

It remains to describe the execution of limit orders. In our setting, a limit order can execute in one of the following two ways:

1. We assume that there are impatient buyers who arrive to the market according to a Poisson process with rate μ. Denote by $(N_t)_{t \in [0,T)}$ the cumulative arrival process for impatient buyers. Each impatient buyer seeks to buy a single share. An arriving impatient buyer arriving at time t has a reservation price $S_t + z_t$, expressed as a premium $z_t \geq 0$ above the bid price S_t that the buyer is willing to forgo in order to achieve immediate execution. We assume that the premium z_t is independent and identically distributed with cumulative distribution function $F\colon \mathbb{R}_+ \to [0, 1]$. In this setting, the instantaneous arrival rate of impatient buyers at time t willing to pay a limit order price of L_t is given by

$$\lambda(u_t) \triangleq \mu(1 - F(u_t)), \tag{3}$$

where $u_t \triangleq L_t - S_t$ is the instantaneous price premium of the limit order. In what follows,

[6]Note that the trade quantity of a single share is meant to represent an atomic unit of the asset, or the smallest commonly traded lot size. The underlying assumption is that the desired trade execution will ultimately be accomplished by a single transaction. In typical U.S. equity markets, for example, this atomic unit might be a block of 100 shares.

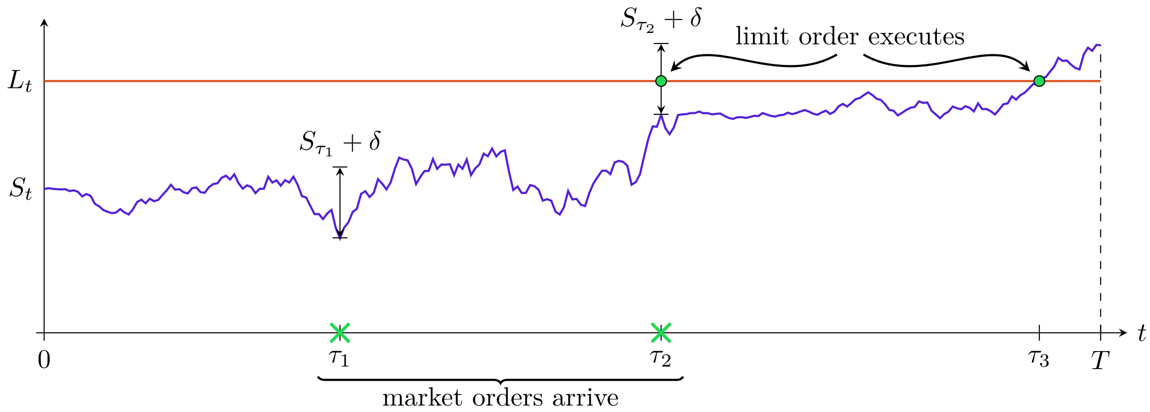

Figure 1: An illustration of the limit order execution in the stylized model over the time horizon $[0, T]$. Here, we assume the trader leaves a limit order with the (constant) price L_t and S_t is the bid price process. If market orders arrive at times τ_1 and τ_2, the limit order would execute at time τ_2 but not time τ_1, since the limit order price is in excess of δ to the best bid price. The limit order would also execute at time τ_3 in the absence of a market order arrival, since the bid price crosses the limit order price at this time.

we will be particularly interested in the special case where

$$\lambda(u_t) \triangleq \begin{cases} \mu & \text{if } u_t \leq \delta, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Here, we assume that every impatient buyer is willing to pay a price premium of at most $\delta > 0$. We assume that δ will be specific to the security and fixed for the trading horizon. We will discuss the extension to the general case (3) in Section 4.4.

Given (4), an impatient buyer is willing to buy 1 share at a fixed premium $\delta > 0$ to the bid price at the time of their arrival. Hence, if a buyer arrives at time $\tau \in [0, T)$, and the trader has placed a limit order with price L_τ, the limit order will execute if $L_\tau \leq S_\tau + \delta$.

2. Alternatively, a limit order will also execute at time τ if the bid price crosses the limit order price, i.e., $S_\tau \geq L_\tau$.

The execution of limit orders in the model is illustrated in Figure 1.

The limit order execution dynamics above can also be economically interpreted in the spirit of the non-informational trade model of Roll (1984). In particular, imagine that the asset has a fundamental value V_t at time t, and that V_t evolves exogenously according to the additive random walk

$$V_t = V_0 + \sigma B_t.$$

If all investors observe this underlying value process and are symmetrically informed, competitive market makers will always be willing to sell shares at a price of $\delta/2$ above the fundamental value or buy shares at a spread of $\delta/2$ below the fundamental value. Here, the quantity δ captures the per share operating costs of trade to the market markers. The liquidating trader can thus sell at the bid price $S_t = V_t - \delta/2$ at any time t. We assume that all other traders in the market are impatient, and that these traders arrive according to the Poisson dynamics described above. An arriving impatient buyer will choose to purchase from the liquidating trader only at a price lower than that provided by the market makers, i.e., only below the price of $V_t + \delta/2 = S_t + \delta$. In this way, we can interpret the parameter δ as the *prevailing bid-offer* spread, that is, the bid-offer spread in the absence of the liquidating trader.

2.2. Optimal Solution

Let P denote the random variable associated with the sale price. We assume the trader is risk-neutral and seeks to maximize the expected sale price. Equivalently, we assume the trader seeks to solve the optimization problem

$$\bar{h}_0 \triangleq \text{maximize } \mathsf{E}[P] - S_0. \tag{5}$$

Here, the maximization is over policies of market orders and limit orders which are non-anticipating, i.e., policies adapted to the filtration generated by the underlying market primitives, $(B_t, N_t)_{t\in[0,T]}$. This objective is equivalent to minimizing implementation shortfall (Perold, 1988).

Note that, while this stylized problem may seem quite simplified, it seeks to answer a fundamental question: at the level of an atomic unit of stock and over a short time horizon, how should a risk-neutral investor choose between limit orders and market orders? This problem is a central ingredient in more sophisticated optimal execution problems involving risk averse investors selling large quantities over longer time horizons.[7] This is because, in a typical algorithmic trading setting, a large "parent" order will be scheduled across time into many very small "child" orders. Each of these "child" orders need to be executed optimally. Since each child order is small and since there are many such child orders, it is reasonable to view the investor as risk-neutral with respect to each child order.

The following lemma characterizes a simple strategy that is optimal for the execution problem we have described:

Lemma 1. *An optimal strategy is to employ only limit orders at times $t \in [0, T)$, with limit price $L_t = S_t + \delta$. In other words, the limit order price is "pegged" at a constant premium δ above the bid price. This pegging strategy achieves the optimal value*

$$\bar{h}_0 = \delta\left(1 - e^{-\mu T}\right). \tag{6}$$

[7] For example, see Bertsimas and Lo (1998) or Almgren and Chriss (2001). These questions have also recently been addressed by Back and Baruch (2007) and Pagnotta (2010) in equilibrium settings.

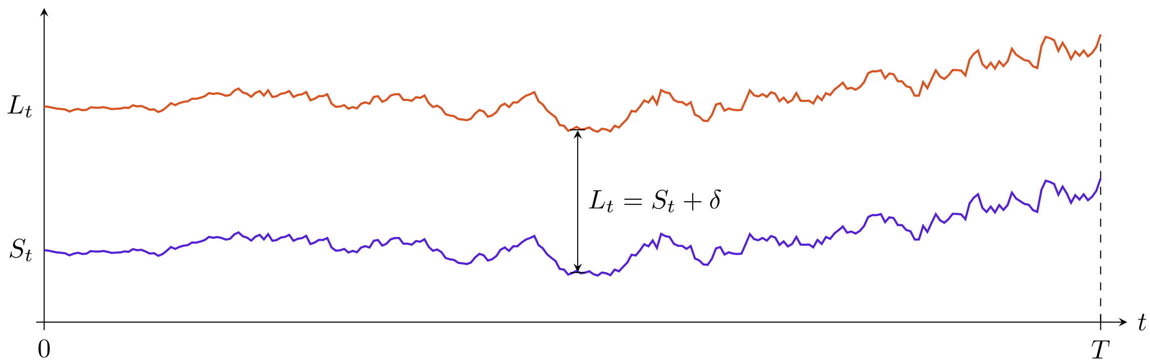

Figure 2: An illustration of an optimal strategy with no latency, over the time horizon $[0, T]$. The trader uses only limit orders prior to end of the time T. The limit order price L_t is pegged to the bid price S_t, with an additional premium corresponding to the bid-offer spread δ.

Proof. Consider a trader using an arbitrary strategy, and denote by $\tau \in [0, T]$ the (random) time at which the trader sells the share, and by $\tau_1 \in [0, \infty)$ the time at which the first impatient buyer arriving to the market. Let $\mathcal{E}$ be the event that the trader sells via a limit order to an impatient buyer at the price L_τ. Then, under the event $\mathcal{E}^c$, the trader sells at the bid price S_τ. Then, the sale price P can be written as[8]

$$P = S_\tau \mathbb{I}_{\mathcal{E}^c} + L_\tau \mathbb{I}_{\mathcal{E}} \leq S_\tau \mathbb{I}_{\mathcal{E}^c} + (S_\tau + \delta)\mathbb{I}_{\mathcal{E}} \leq S_\tau + \delta \mathbb{I}_{\{\tau_1 < T\}}. \tag{7}$$

Here, for the first inequality, we used the fact that an impatient buyer will only buy at time τ is $L_\tau \leq S_\tau + \delta$, and, for the second inequality, we used the fact that the event $\mathcal{E}$ can only occur if an impatient buyer arrives in the time interval $[0, \tau)$. Denote by $\bar{h}_0$ the value under an optimal strategy. Using the fact that τ is a bounded stopping time and the fact that S_t is a martingale, by the optional sampling theorem,

$$\bar{h}_0 \leq \mathsf{E}[P] - S_0 \leq \mathsf{E}[S_\tau + \delta \mathbb{I}_{\{\tau_1 < T\}}] - S_0 = \delta \mathsf{P}(\tau_1 < T) = \delta \left(1 - e^{-\mu T}\right).$$

On the other hand, the hypothesized strategy results in equality in (7). Thus, the result follows. ■

The optimal pegging strategy suggested by Lemma 1 is illustrated in Figure 2. This policy can be interpreted intuitively as follows: since the trader is risk-neutral and the bid price process is a martingale, the trader is indifferent between trading at time 0 at the bid price or trading at any other time at the bid price. Via a limit order, however, the trader can receive a price which is in excess of the bid price. The excess premium is limited to δ, since an impatient buyer will not pay more than this. Hence, the trader maintains a single limit order in the book, and continuously updates the price to track bid price, plus an additional premium of δ.

Note that our stylized execution model captures only the behavior of a single agent. Our model

[8]We denote by $\mathbb{I}_{\mathcal{E}}$ the indicator function of the event $\mathcal{E}$.

does not capture the strategic response of other agents, either competing agents submitting limit orders to sell, or contra-side impatient buyers. Both of these types of agents might be expected to react to the activity of the limit order trader, and may diminish the gains of the limit order trader. Separately, our model also exaggerates the gains to be earned by placing limit orders rather than market orders, due to the fact we do not include adverse selection costs incurred by limit orders.

However, at a high level, a trader in our model with a mandate to trade over a fixed time horizon but with no private information as to the asset value prefers limit orders to market orders. We believe this is representative of the situation of algorithmic traders executing large "parent" orders in practice. When executing a "child" order over a short time horizon, such traders typically first submit limit orders, and then "clean up" with market orders as time runs short. Hence, despite omissions of strategic considerations and other significant simplifications, the resulting policies do capture representative features of real world trading, if only at a stylized level. Moreover, our simplified single-agent mode enables us to address the dynamic nature of trade execution and obtain a closed-form expression highlighting the exact drivers of the latency cost.

3. A Model for Latency

The optimal policy for the stylized execution problem of Section 2 relied on the ability of a trader to continuously track an informational process, namely, the bid price in the market, and to update his order as the process evolves. Here, we will consider a variation of that problem where the trader is unable to continuously participate in the market, but faces a fixed latency $\Delta t > 0$.[9] We are interested in quantifying the cost of this latency by comparing the expected payoff in this model to that in the stylized model without latency. Note that the model at hand is quite basic with regards to some of primitives (e.g., the stochastic process describing the evolution of bid prices), we will discuss a number of tractable extensions in Section 4.4, including more complicated models of the bid price process and of limit order execution.

In general, latency that a trader experiences can take many forms. Minimally, for example, there is the delay of the data feeds that deliver market price information to the trader. There is the delay of the trader's own decision making. Finally, there is the delay of the trader's resulting order reaching the marketplace. We assume that the trader makes decisions instantaneously — we will see that this is reasonable since the optimal decision rule for the trader will take a very simple form. Further, from the trader's perspective, the roundtrip delay (the total delay for an order to be processed by an exchange and reflected in the data feeds observed by the trader) cannot be decomposed into a delay to the exchange and a delay from the exchange. Hence, without loss of

[9]Note that many modern exchanges explicitly allow for pegged orders; these orders obviate the need for the trader to continually track the bid price in the manner we describe. However, more generally, when tracking an alternative informational process such as the price on a different exchange, the fundamental value (see Section 2), etc., a trader would still need to continuously monitor the market relative to the informational process, and latency would be important.

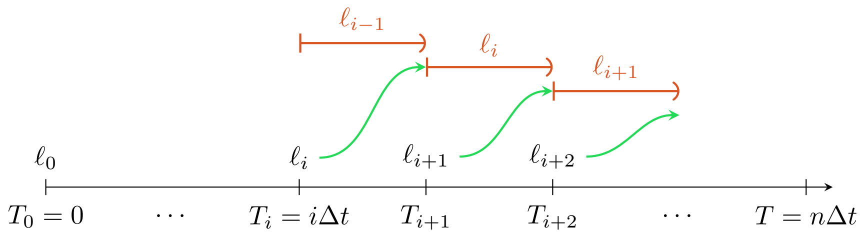

Figure 3: An illustration of the model of latency. Here, the time horizon $[0, T]$ is divided into n slots, each of duration equal to the latency Δt. The limit order price ℓ_i is decided at the start of the ith time slot, i.e., at time T_i. This price only takes effect Δt units of time later, and is active during the subsequent time interval $[T_{i+1}, T_{i+2})$.

generality, we will assume that the trader is able to observe market price information with no delay or latency,[10] but that the trader's orders experience a latency Δt before they are processed by the exchange. This latency is meant to capture, for example, networking or routing delays that are specific to the trader, and that might be reduced through colocation or additional investment in networking technology.

In our latency model, we consider an investor who maintains a limit order to sell one share over the time horizon $[0, T]$ (the possibility of market orders will be discussed shortly), so that once the limit order is executed, the investor immediately exits the market. The time horizon $[0, T]$ is divided into n slots each of length Δt, i.e., $T = n\Delta t$. For each $i \in \{0, 1, \ldots, n\}$, define $T_i \triangleq i\Delta t$.

At each time T_i, based on all information observed thus far, we assume that the trader can instantaneously decide to update the limit order with a new price ℓ_i. Due to a latency of Δt, the updated price does not reach the market and take effect until the beginning of the next time slot, i.e., T_{i+1}. This limit order price remains active until time T_{i+2}, at which point it is superseded[11] by the next price ℓ_{i+1}. This sequence of events is illustrated in Figure 3. Between the time T_i, when the price ℓ_i is decided, and the time T_{i+1}, when the updated order reaches the market, the following events can occur:

- $\mathcal{E}_i^{(1)}$: An impatient buyer arrives in the time interval (T_i, T_{i+1}) and $\ell_{i-1} \leq S_{T_i} + \delta$, i.e., the *prior* limit price ℓ_{i-1}, which is active at that time, is within a margin δ of the bid price at the start of the interval. In this case, the limit order executes at the price ℓ_{i-1}, and the investor leaves the market. Note that the updated limit price ℓ_i never takes effect.

  We assume that the probability that an impatient buyer arrives in any given time slot is $\mu\Delta t$, and that these arrivals occur independently of everything else.[12] We assume that $\Delta t < 1/\mu$

[10] Equivalently, we can assume that our definition of time corresponds to the trader's clock.

[11] In practice, this ordering scheme might be achieved by a sequence of cancel-and-replace limit orders, each of which cancels the prior limit order, and inserts a new limit order with the updated price. If the prior limit order has already been filled when a subsequent cancel-and-replace order arrives, the new order will fail. Hence, the investor is guaranteed to sell at most one share.

[12] Note that this is simply a discrete-time Bernoulli arrival process that is analogous to the the Poisson arrival process of Section 2.

so that this probability is well-defined. The bid price process evolves according to the random walk (2).

- $\mathcal{E}_i^{(2)}$: Otherwise, if $S_{T_{i+1}} \geq \ell_i$, i.e., the bid price has crossed the order price ℓ_i at the instant the order reaches the market, then the order immediately executes at price $S_{T_{i+1}}$.
- $\mathcal{E}_i^{(3)}$: Otherwise, the limit order price ℓ_i is active over the time interval $[T_{i+1}, T_{i+2})$.

In order to consider the possibility of market orders, we allow the limit price $\ell_i = -\infty$. By picking this price, the trader can guarantee that the bid price at time T_{i+1} will cross the order price, i.e., $S_{T_{i+1}} \geq \ell_i$ with probability 1. Thus, the choice of $\ell_i = -\infty$ corresponds to a certain execution at the bid price $S_{T_{i+1}}$, i.e., a market order. Similarly, the trader can make the decision at time T_i not to trade by setting $\ell_i = \infty$. As in the model of Section 2, if the investor has been unable to sell the share by the end of the time horizon T, the investor is forced to sell via a 'clean-up' trade, i.e., a market order at time T. This is accomplished by enforcing the constraint that $\ell_{n-1} = -\infty$, which we will assume implicitly in what follows.

As before, if P is the random variable associated with the sale price, the trader is risk-neutral and seeks to solve the optimization problem

$$h_0(\Delta t) \triangleq \underset{\ell_0,\ldots,\ell_{n-1}}{\text{maximize}} \ \mathsf{E}[P] - S_0. \tag{8}$$

Here, the maximization is over the choice of limit order prices $(\ell_0, \ell_1, \ldots, \ell_{n-1})$. We assume that the price decisions are non-anticipating, i.e., each ℓ_i is adapted to the filtration generated by the bid price process and the arrival of impatient buyers up to and including time T_i. Our goal is to analyze $h_0(\Delta t)$, which is the value under an optimal trading strategy when the latency is Δt.

Note that, as compared to the model of Section 2, our present model with latency differs in two ways: First, the trader makes decisions at the beginning of discrete-time intervals of length Δt, as opposed to continuously. Second, the orders of the trader incur a latency or delay of length Δt before they reach the marketplace. We are interested in studying the impact of the latter feature, latency, and we adopt the former feature, discrete-time decision making, so as to admit a tractable dynamic programming analysis. In Section 4.3, however, we will see that in the low latency regime in which we are most interested, the discrete-time nature of our model has a negligible impact.

4. Analysis

In this section, we solve for the optimal policy for the trader in the latency model of Section 3. This problem can be solved via a dynamic programming decomposition that is presented in Section 4.1. While the exact dynamic programming solution can be computed numerically, in Section 4.2 we will present an asymptotic analysis that provides a closed-form analytic expression for the cost of latency in the low latency regime, where $\Delta t \to 0$. In Section 4.3, we will consider the implications of the discrete-time nature of our latency model. Finally, in Section 4.4, we will discuss a number of extensions of our latency model.

4.1. Dynamic Programming Decomposition

The standard approach to solving the optimal control problem (8) is to employ dynamic programming arguments. In Appendix A of the electronic companion, we formally derive the optimal control policy using these methods. In order to focus on the high level picture, however, for the moment we will be content with summarizing those results.

In particular, assume a fixed latency of Δt. For each decision time T_i with $0 \leq i < n$, define $\mathcal{U}_i$ to be the event that the trader's limit order remains unfulfilled prior to time T_{i+1}, i.e., none of the orders submitted at prices $\ell_0, \ldots, \ell_{i-1}$ are executed. Note that if the event $\mathcal{U}_i$ *does not* hold, then the limit order price ℓ_i to be decided at time T_i is irrelevant. This is because, by the time that order arrives to the market, the trader would have already sold a share. Define the quantity

$$h_i \triangleq \underset{\ell_i,\ldots,\ell_{n-1}}{\text{maximize}} \ \mathsf{E}\left[P \mid S_{T_i}, \mathcal{U}_i\right] - S_{T_i}. \tag{9}$$

Note that $h_0 = h_0(\Delta t)$, where $h_0(\Delta t)$ is defined in (8), and thus our notation is consistent. More generally, for $i > 0$, we can interpret h_i to be the trader's expected payoff at time T_i relative to the current bid price S_{T_i} under the optimal policy, the order does not get filled prior to time T_{i+1}. Thus, h_i can be interpreted as a *continuation value* in the dynamic programming context.

The continuation values $\{h_i\}$ quantify the remaining value for a trader at each time period if his order remains unfulfilled. Given the continuation values, at each time T_i, the investor can make an optimal decision as to the limit order price ℓ_i by balancing the benefits of execution in the time slot $[T_{t+1}, T_{i+2})$ with the value h_{i+1} that will be obtained if the order is not executed. Moreover, the optimal decisions and continuation values can be jointly computed via backward induction of a Bellman equation. This result is captured in the following theorem. The proof, which is provided in Appendix A of the electronic companion, follows from formal dynamic programming arguments.

Theorem 1. *Suppose $\{h_i\}$ satisfy, for $0 \leq i < n-1$,*

$$\begin{aligned} h_i = \max_{u_i} \ \Bigg\{ & \mu \Delta t \left[u_i \left(\Phi\left(\frac{u_i}{\sigma\sqrt{\Delta t}}\right) - \Phi\left(\frac{u_i - \delta}{\sigma\sqrt{\Delta t}}\right)\right) + \sigma\sqrt{\Delta t}\left(\phi\left(\frac{u_i}{\sigma\sqrt{\Delta t}}\right) - \phi\left(\frac{u_i-\delta}{\sigma\sqrt{\Delta t}}\right)\right)\right] \\ & + h_{i+1}\left[(1-\mu\Delta t)\Phi\left(\frac{u_i}{\sigma\sqrt{\Delta t}}\right) + \mu\Delta t \Phi\left(\frac{u_i-\delta}{\sigma\sqrt{\Delta t}}\right)\right]\Bigg\}, \end{aligned} \tag{10}$$

and

$$h_{n-1} = 0. \tag{11}$$

Here, ϕ and Φ are, respectively, the p.d.f. and c.d.f. of the standard normal distribution. Then, $\{h_i\}$ correspond to the continuation values under the optimal policy.

Suppose further that, for $0 \leq i < n-1$, u_i^* *is a maximizer of* (10). *Then, a policy which chooses limit order prices which are pegged to the bid prices according to the premia defined by* $\{u_i^*\}$, *i.e.,*

$$\ell_i^* = S_{T_i} + u_i^*, \quad \forall\ 0 \leq i < n-1,$$

is optimal.

Theorem 1 suggests a computational strategy for determining continuation values and an optimal policy. Starting with the terminal condition $h_{n-1} = 0$, one proceeds via backward induction, solving the single variable optimization problem (10) over the decision variable u_i once per time slot. So long as optimal solutions exist, they will determine the continuation values and optimal policy. Moreover, the optimal policy is a pegging strategy. That is, the limit order price is pegged at a deterministic (but time varying) premium above the current bid price. These limit order premia are given by the maximizers $\{u_i^*\}$.

In the following theorem, whose proof is provided in Appendix B of the electronic companion, we establish the existence and uniqueness of the optimal solutions to (10) and provide upper and lower bounds for the resulting limit price premia, for small values of latency Δt.

Theorem 2. *Fix* $\alpha > 1$. *If* Δt *is sufficiently small, then there exists a unique optimal solution* $\{h_i\}$ *to the dynamic programming equations* (10)–(11). *Moreover, the corresponding optimal policy* $\{u_i^*\}$ *is unique. For* $0 \leq i < n-1$, *this strategy chooses limit prices in the range*

$$\ell_i^* \in \left(S_i + \delta - \sigma\sqrt{\Delta t \log \frac{\alpha L}{\Delta t}},\ S_i + \delta - \sigma\sqrt{\Delta t \log \frac{R(\Delta t)}{\Delta t}} \right),$$

where

$$L \triangleq \frac{\delta^2}{2\pi\sigma^2}, \qquad\qquad R(\Delta t) \triangleq \frac{\delta^2(1-\mu\Delta t)^{2n}}{2\pi\sigma^2}.$$

Figure 4 illustrates the intuition behind Theorem 2, by considering the situation of a trader at time $t = 0$, when the bid price is S_0. In the absence of latency, the trader would peg the limit order price at a fixed premium of δ, i.e., $\ell_0 = S_0 + \delta$. This would result in a trade with the next impatient buyer with probability 1. If there is latency present, however, this limit price is not optimal. To see this, suppose that an impatient trader will arrive at time $\tau_1 \in (\Delta t, 2\Delta t)$. If the limit order price is set at ℓ_0, the probability that the trade *does not* get executed is

$$\mathsf{P}\left(\ell_0 \geq S_{\Delta t} + \delta\right) = \mathsf{P}\left(S_0 \geq S_{\Delta t}\right) = 1/2.$$

When Δt is small, the probability of missing an execution can be significantly lowered at a small cost by lowering ℓ_0 by an additional safety margin. If we set this safety margin to be C standard deviations of the one-period price change, i.e., $\ell_0' = S_0 + \delta - C\sigma\sqrt{\Delta t}$, then the probability of missing

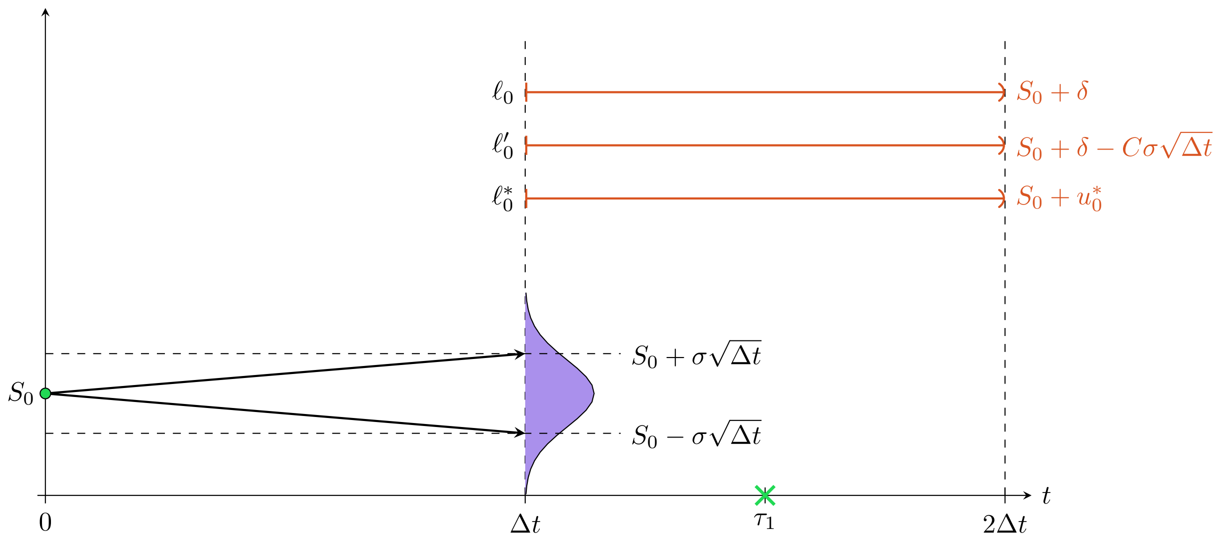

Figure 4: An illustration of the optimal policy of Theorem 2. In the absence of latency, at time $t = 0$, the trader would set the limit price at a premium of δ, i.e., $\ell_0 = S_0 + \delta$. In an environment with latency, the trader might set the limit price to be ℓ_0', which lowers ℓ_0 by an additional safety margin of C standard deviations. This serves to increase the likelihood of trade execution in the interval $(\Delta t, 2\Delta t)$. The optimal limit price ℓ_0^* utilizes a safety margin that is slightly larger.

execution becomes

$$\mathsf{P}\left(\ell_0' \geq S_{\Delta t} + \delta\right) = \mathsf{P}\left(S_0 - C\sigma\sqrt{\Delta t} \geq S_{\Delta t}\right) = \Phi(-C).$$

This probability can be made close to 0 by the choice of C. However, given a fixed choice of C independent of Δt, the probability remains constant (i.e., independent of Δt) and non-zero. The additional safety margin corresponding to the log term in Theorem 2 is a second order adjustment. This is introduced so that, given the optimal limit price ℓ_0^*, the probability of execution tends to 1 as $\Delta t \to 0$.

4.2. Asymptotic Analysis

The dynamic programming decomposition developed in Section 4.1 allows the exact numerical computation of the value $h_0(\Delta t)$, the value under an optimal policy of the latency model introduced in Section 3, when the latency is Δt. As discussed earlier, the latency observed in modern electronic markets is extremely small, often on the time scale of milliseconds. Thus, we are most interested in the qualitative behavior of $h_0(\Delta t)$ in the asymptotic regime where $\Delta t \to 0$. The main result of this section is the following theorem, whose proof is provided in Appendix C of the electronic companion. It provides a closed-form expression for $h_0(\Delta t)$, which holds asymptotically[13] as $\Delta t \to 0$.

[13] In what follows, given arbitrary functions f and g, and a positive function q, we will say that $f(\Delta t) = g(\Delta t) + O(q(\Delta t))$ if $\limsup_{\Delta t \to 0} |f(\Delta t) - g(\Delta t)|/q(\Delta t) < \infty$, i.e., if the difference between f and g, as $\Delta t \to 0$, is asymptotically bounded above by *some* positive multiple of q. Similarly, we will say that $f(\Delta t) = g(\Delta t) + o(q(\Delta t))$

Theorem 3. *As* $\Delta t \to 0$,

$$h_0(\Delta t) = \bar{h}_0 \left(1 - \frac{\sigma}{\delta} \sqrt{\Delta t \log \frac{\delta^2}{2\pi\sigma^2 \Delta t}} \right) + o\left(\sqrt{\Delta t}\right),$$

where

$$\bar{h}_0 = \delta \left(1 - e^{-\mu T}\right)$$

is the optimal value for the stylized model without latency, i.e., the value defined by (5).

Theorem 3 is not surprising when considered in the context of Theorem 2. In the stylized model without latency, the optimal strategy is to peg the limit order price at a premium of δ, and this yields a value of $\bar{h}_0$. On the other hand, Theorem 2 suggests a trader facing latency Δt will lower this limit price premium by a factor of, approximately,

$$\frac{\sigma}{\delta} \sqrt{\Delta t \log \frac{\delta^2}{2\pi\sigma^2 \Delta t}} + o\left(\sqrt{\Delta t}\right).$$

If this lowers the ultimate value proportionally, then the value of the optimal policy in the presence of latency Δt should approximately be

$$\bar{h}_0 \left(1 - \frac{\sigma}{\delta} \sqrt{\Delta t \log \frac{\delta^2}{2\pi\sigma^2 \Delta t}} \right) + o\left(\sqrt{\Delta t}\right).$$

The proof of Theorem 3, provided in Appendix C of the electronic companion, makes this intuition precise.

One implication of Theorem 3 is that $h_0(\Delta t) \to \bar{h}_0$ as $\Delta t \to 0$, i.e., the value of the latency model converges to that of the stylized model without latency of Section 2. This suggests the following definition:

Definition 1. *Define the* ***latency cost*** *associated with latency* Δt *by*

$$\mathsf{LC}(\Delta t) \triangleq \frac{\bar{h}_0 - h_0^*(\Delta t)}{\bar{h}_0}. \tag{12}$$

Latency cost has an easy interpretation. Using $\bar{h}_0$, the value obtained in the stylized model without latency as a benchmark, the numerator of (12) is the *lost revenue* incurred due the the presence of latency. On the other hand, we can regard the denominator as the *cost of immediacy* for an impatient investor in a time horizon of length T. This is because, in the stylized model without

if $\lim_{\Delta t \to 0} |f(\Delta t) - g(\Delta t)|/q(\Delta t) = 0$, i.e., if the difference between f and g, as $\Delta t \to 0$, is asymptotically dominated by *every* positive multiple of q. Finally, we will say that $f(\Delta t) = g(\Delta t) + \Theta(q(\Delta t))$ if $0 < \liminf_{\Delta t \to 0} |f(\Delta t) - g(\Delta t)|/q(\Delta t) \leq \limsup_{\Delta t \to 0} |f(\Delta t) - g(\Delta t)|/q(\Delta t) < \infty$, i.e., if the difference between f and g is asymptotically bounded above and below by positive multiples of q.

latency, it is the difference in revenue obtained by a risk-neutral investor willing to patiently provide liquidity by employing limit orders over the length of the time horizon, and an impatient investor who demands immediate liquidity and sells at the bid price at time $t = 0$, cf. (5). Therefore, we can describe the latency cost as the amount a trader forgoes due to latency, as a percentage of the cost of immediacy.

The following corollary restates the asymptotic approximation of Theorem 3 in terms of latency cost.

Corollary 1. *As* $\Delta t \to 0$*,*

$$\mathsf{LC}(\Delta t) = \frac{\sigma\sqrt{\Delta t}}{\delta}\sqrt{\log\frac{\delta^2}{2\pi\sigma^2\Delta t}} + o\left(\sqrt{\Delta t}\right).$$

There are a number of interesting observations that can be made regarding the asymptotic approximation of Corollary 1. First of all, asymptotically, latency cost *does not* depend on the length of the time horizon T or the arrival rate of impatient traders μ. As a function of the remaining parameters, the asymptotic latency cost depends only on a composite parameter that is the ratio the one-period standard deviation of price changes $\sigma\sqrt{\Delta t}$ to the bid-offer spread δ. Both of these quantities are readily measurable empirically. Corollary 1 suggests that the latency cost increasing in this ratio. Thus, at the same level of latency, the latency cost is most significant for assets which are very volatile or very liquid. Further, Corollary 1 suggests that, when latency is low, there are increasing marginal benefits to further reductions in latency, i.e., $\mathsf{LC}''(\Delta t) < 0$. In Section 5.1, we illustrate some of facts numerically, as well as considering the accuracy of our approximation, as compared to the exact latency cost.

4.3. Discreteness of Time vs. Latency

The latency model introduced in Section 3 differs from the the stylized model without latency of Section 2 in two principal ways: (i) the trader faces a delay or latency between the time that trading decisions are made and when they reach the marketplace, and (ii) the latency model is formulated in discrete-time rather than continuous time. The latter point refers to the facts that, in the model with latency, a trader is only able to update his limit order at discrete intervals of time rather than continuously, impatient buyers arrive according to a Bernoulli process rather than a Poisson process, etc. In order to disentangle these two effects, in this section we will briefly describe a trading model that is formulated in discrete time but *without* latency. By considering this model, we will demonstrate that the asymptotic latency cost derived in Section 4.2 is indeed due to latency effects and not due to the discreteness of time.

To this end, consider a model in the discrete-time setting of Section 3 but with no latency. Here, at each time $T_i \triangleq i\Delta t$, for $i = 0, 1, \ldots, n$, the investor sets a limit order price ℓ_i. This limit order price takes effect immediately. Between time T_i and time T_{i+1} the following events can occur:

- If $S_{T_i} \leq \ell_i$, i.e., the bid price is less than the limit order price, the limit order immediately executes at the price S_{T_i}.
- Otherwise, suppose that an impatient buyer arrives in the time interval (T_i, T_{i+1}) and $\ell_i \leq S_{T_i} + \delta$, i.e., the limit price ℓ_i is within a margin δ of the bid price at the start of the interval. In this case, the limit order executes at the price ℓ_i. We assume that an impatient buyer arrives with probability $\mu \Delta t$, independent of everything else.

As before, if the investor is unable to sell the share by the end of the time interval, he is forced to sell via a market order, i.e., $\ell_n = -\infty$. If P is the sale price, the optimal value for the trader in this discrete model is given by

$$h_0^{\mathsf{D}}(\Delta t) \triangleq \underset{\ell_0,\ldots,\ell_n}{\text{maximize}} \ \mathsf{E}[P] - S_0.$$

We have the following result, whose proof is identical to the martingale argument used to establish Lemma 1.

Lemma 2. *An optimal strategy for the discrete model is to place limit orders at the price $\ell_i = S_{T_i} + \delta$, for $i = 0, 1, \ldots, n-1$. This strategy achieves the value*

$$h_0^{\mathsf{D}}(\Delta t) \triangleq \delta\big(1 - (1 - \mu \Delta t)^n\big).$$

Now, note that, for all $0 < \Delta t < 1/\mu$,

$$e^{-\mu T - \frac{1}{2}\mu^2 T \Delta t} \leq (1 - \mu \Delta t)^{T/\Delta t} \leq e^{-\mu T}.$$

Therefore, the difference in value between the continuous model of Section 2 and the discrete model considered here is at most

$$\left|h_0^{\mathsf{D}}(\Delta t) - \bar{h}_0\right| \leq \delta e^{-\mu T}\left(1 - e^{-\frac{1}{2}\mu^2 T \Delta t}\right) \leq \tfrac{1}{2}\delta \mu^2 T e^{-\mu T} \Delta t.$$

In other words, this difference is asymptotically $O(\Delta t)$. By Theorem 3, however, the difference between the continuous model and the latency model is asymptotically

$$\Theta\big(\sqrt{\Delta t \log(1/\Delta t)}\big).$$

Hence, the asymptotic effect of latency dominates the asymptotic effect of the discreteness of time.

4.4. Extensions

The analysis of the latency model that we have presented proceeded according to two high level steps:

(i) First, in Section 4.1, a simplified dynamic programming decomposition was developed. In this decomposition, at each time, the trader's value function is parameterized by a single scalar, rather than being an arbitrary function of state. This allows the Bellman equation to be solved through a system of n equations in n unknowns, given by (10)–(11).

(ii) Second, in Section 4.2, an asymptotic analysis of the simplified dynamic programming equations (10)–(11) was performed. This gave rise to the asymptotic latency cost expression of Corollary 1.

The dynamic programming decomposition step (i) that is at the heart of our analysis can be extended to a much broader set of stochastic primitives than the present setting. In each of these cases, a different set of simplified dynamic programming equations, analogous to (10)–(11) would arise, and would require a customized variation of asymptotic analysis step (ii). In particular, consider the following tractable generalizations:

- **Price process.** In our model, the price process S_t is a Brownian motion. Our dynamic programming decomposition only requires that the S_t be a Markov process and a martingale. It would be straightforward to extend the dynamic programming step (i) and consider other Markovian martingales, for example, allowing for non-Gaussian processes, time-inhomogeneous volatility, or for jump processes.

  On the other hand, the asymptotic analysis step (ii) we have presented is quite sensitive to distributional assumptions of the price process, and would require specialized analysis for any such generalization. In Appendix D of the electronic companion, we consider one generalization of particular interest, where the price dynamics also contain a jump component.

- **Limit order execution.** In our model, the execution of a limit order in the time slot (T_i, T_{i+1}) required that the limit order price ℓ_{i-1} be within a spread δ of the bid price S_{T_i}, and that an impatient trader arrive. More generally, our dynamic programming decomposition only requires that the execution of this limit order, conditional on the price difference $\ell_{i-1} - S_{T_i}$, be independent of everything else. This can accommodate a number of generalizations, for example, the arrival rate of impatient buyers can be time-varying. Further, the maximum premium above the bid price S_t that an impatient buyer is willing to pay can be randomly distributed, as in (3). This would allow models where a limit order that is priced aggressively low has a much higher probability of execution. Such models could alternatively be interpreted, as discussed in Section 2, as cases where the prevailing bid-offer spread is not constant, but is independent and identically distributed, varying from period to period.

5. Empirical Estimation of Latency Cost

In this section, we will consider empirical applications of our model. First, we will illustrate the optimal trading policy and the corresponding value function when the model parameters are estimated from high frequency market data for a single stock. We will also compare the exact latency cost (numerically computed via dynamic programming) to the approximation provided by Corollary 1 in order to access the quality of our approximation. Subsequently, we show the historical evolution of latency cost and implied latency across a range of U.S. equities using cross-sectional data on volatilities and bid-offer spreads during the 1995–2005 period.

Our empirical analysis should be regarded as a first-order study to obtain a rough calibration of our model. It will allow us to analyze the model in relevant parameter regimes, as well as gaining a broad understanding the implications of our model for the trading of U.S. equities. Under our modeling assumptions (e.g., Browian motion price processes, Poisson arrivals of impatient traders, constant bid-offer spread, etc.), our empirical measurement of latency cost requires estimates of the high frequency price volatility σ and the prevailing bid-offer spread δ. Here, we make a number of simplifications and rely on the recent empirical work of Aït-Sahalia and Yu (2009) to obtain these quantities:

- We estimate price volatility σ using the maximum likelihood estimates of the volatility of returns provided by Aït-Sahalia and Yu (2009). Note that this estimation of high frequency volatility aims to filter out the impact of microstructure noise and obtain an unbiased estimate of daily volatility. However, for an investor with a trading horizon of 1 second, microstructure noise needs to be incorporated as well. Therefore, the high frequency volatility estimate that is used in our empirical analysis underestimates the actual volatility faced by a high frequency trader with a very short trading horizon.

- Recall that the prevailing bid-offer spread, δ, equals the bid-offer spread in the absence of the liquidating trader. In the empirical data, it is impossible to disentangle the presence of liquidating traders. Moreover, the bid-offer spread will not be constant, but will vary over the course of the trading day. As a proxy for δ, we use the average bid-offer spread over the trading day.

Despite these shortcomings, we believe that our empirical analysis can shed light on the importance of latency in the trading of U.S. equities.

5.1. The Optimal Policy and the Approximation Quality

In what follows, we will numerically evaluate the optimal policy in our model, the corresponding value function, and the latency cost approximation. These numerical experiments are meant to be illustrative of our model. We will use realistic model parameters estimated from recent market data for a single stock. Our methodology here is not meant to be authoritative — there are many subtleties in the analysis of high frequency data; these are beyond the scope of the work at hand. However, we do seek to demonstrate that our model parameters can be readily derived from commonly available data.

Specifically, the model parameters herein are estimated from trade-and-quote (TAQ) data for a stock that is a representative example of a liquid name, Goldman Sachs Group, Inc. (NYSE: GS), on the trading day of January 4, 2010. This data was obtained from the Wharton Research and Data Services (WRDS) consolidated TAQ database. Only trades and quotes originating from the primary exchange (NYSE) during regular trading hours were considered. The model parameters were estimated as follows:

- Initial bid price: $S_0^{\mathsf{GS}} = \$170.00$. This was chosen to be the first transaction price on the trading day.
- Bid-offer spread: $\delta^{\mathsf{GS}} = \$0.058$, i.e., equivalently, 3.4 basis points relative to the initial price S_0^{GS}. This was estimated by computing the average spread between bid and offer quotes over the course of the trading day and rounding to the nearest cent.
- Arrival rate of market orders: $\mu^{\mathsf{GS}} = 12.03$ (per minute). This was estimated by dividing the total number of NYSE trades by the length of the trading day.
- Price volatility: $\sigma^{\mathsf{GS}} = \$1.92$ (daily), i.e., approximately equivalent to an annualized volatility of returns of 17.9%. These were estimated from the time series of transaction prices over the course of the trading day, using maximum likelihood estimation as described inAït-Sahalia and Yu (2009).
- Trading horizon: $T = 10$ (seconds).

Figure 5 illustrates the optimal limit order policy for GS under different values of latency. If there is no latency, the limit orders are submitted at a constant premium of δ. When there is latency, the optimal order policy is obtained using the exact dynamic programming solution of (10)–(11). As the latency increases, the limit order premium is reduced below δ so as to account for the increasing uncertainty of price movements over the latency interval. Theorem 2 suggests that this reduction is approximately equal to

$$\sigma\sqrt{\Delta t \log \frac{\delta^2}{2\pi\sigma^2\Delta t}}. \tag{13}$$

In Figure 5, we see that, with a latency of 500 ms, this adjustment is up to approximately $1.4\sigma\sqrt{\Delta t}$, i.e., 1.4 times the standard deviation of prices over the latency interval. When the latency is reduced to 250 ms and to 50 ms, the adjustment increases to 1.6 and 2.1 standard deviations, respectively. The fact that this adjustment, when measured as a multiple of the uncertainty over the latency period, increases as the latency decreases is consistent with (13).

In Figure 5, we also observe that as t increases and the trading deadline approaches, the limit order premium u_t^* becomes lower. This makes intuitive sense: the trader faced with a terminal value of 0 since he is required to sell using market order at the end of the period. As the deadline

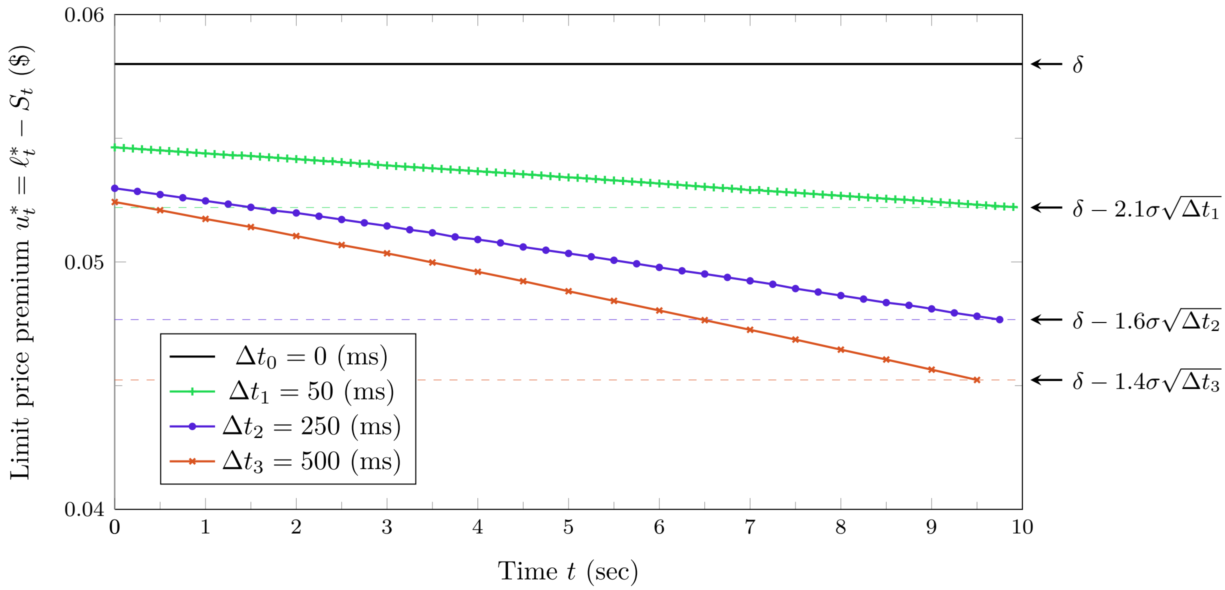

Figure 5: An illustration of the optimal strategy for GS, expressed in terms of limit price premium over the course of the time, for different choices of latency. In each case, the dashed line illustrates the relative distance below the bid-offer spread δ of the price premium of the final limit order, as a multiple of the standard deviation of prices over the latency interval.

approaches, the trader is more willing to sacrifice the potential profits of a limit order in order to increase the probability of execution.

Figure 6 illustrates the corresponding continuation value under the optimal policy for GS, for different values of latency. Clearly, the trader's expected payoff decreases as latency increases or the end of the trading horizon approaches.

Finally, Figure 7 illustrates the latency cost as a function of latency. Both the exact value of the latency cost, computed numerically via the dynamic programming decomposition (10)–(11), and the asymptotic latency cost approximation provided by Corollary 1 are shown. The latency costs decrease from approximately 20% of the cost of immediacy to 5% of the cost of immediacy, as the latency decreases from 500 ms to 5 ms. Further, the marginal benefit of reducing latency increases as the latency approaches zero. Finally, we note that the approximate and exact latency costs are quite close across the entire range of latency values. This suggests that the approximation is of very high quality in this case.

5.2. Historical Evolution of Latency Cost

In this section, we will examine the historical evolution of latency cost in U.S. equities. Here, we consider the situation of a hypothetical investor with a fixed latency of 500 milliseconds. This choice of latency is made approximately to reflect the reaction time of a very fast human trader. We will use this as a proxy for the fastest possible trading on a "human time scale". By analyzing

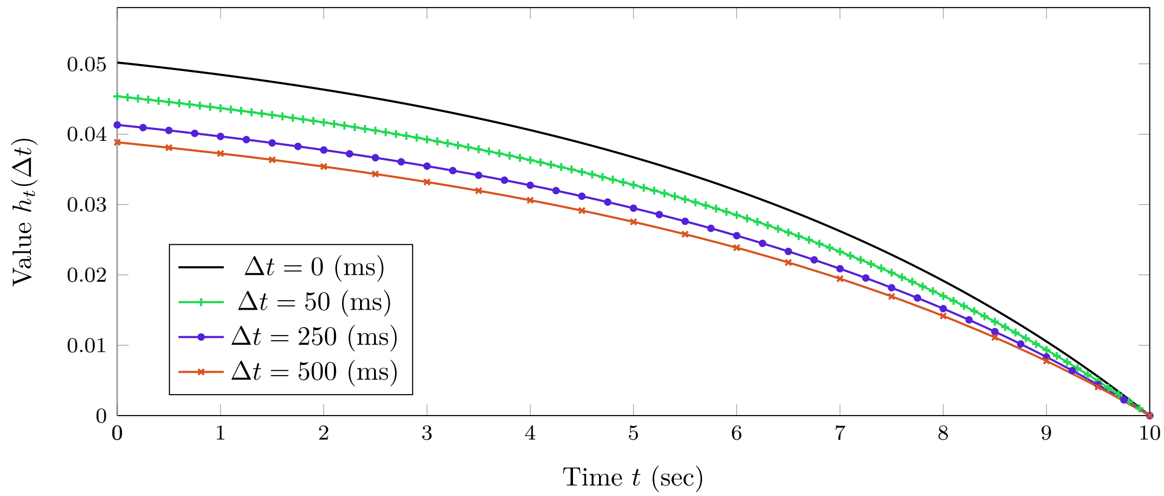

Figure 6: An illustration for the evolution of the continuation value of the optimal policy over time for GS, for different choices of latency. The expected value of the trader decreases as latency increases or as the end of the trading horizon approaches. As the latency increases from 0 ms to 500 ms, the trader loses more than 0.01 of the 0.05 cent spread, i.e., more than 20% of the spread.

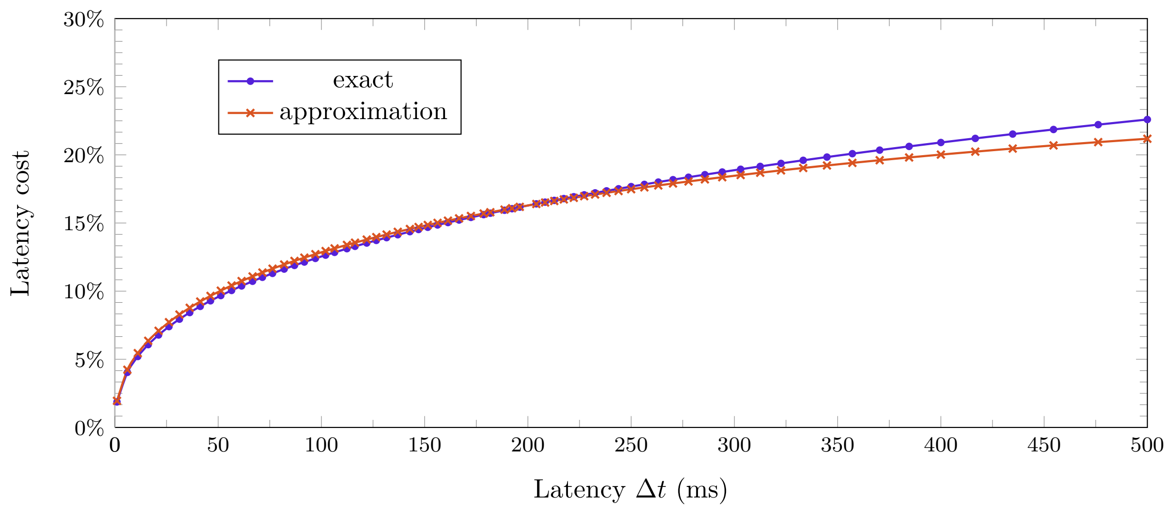

Figure 7: An illustration of the latency cost as a function of the latency. both the exact latency cost and the asymptotic approximation are shown. The approximate latency cost closely aligns with the exact latency cost across the entire range of latency values. This illustrates that our closed-form formula can accurately approximate the exact latency cost for low values of latency.

the evolution of the associated latency cost, we will get a sense of the importance of latency over time.

Our empirical analysis relies on the data set of Aït-Sahalia and Yu (2009). Their data set contains estimates for various liquidity measures for all NYSE common stocks on a daily basis during the sample period of June 1, 1995 to December 31, 2005. The estimates are derived from intraday transaction prices and quotes from the NYSE TAQ database. We utilized only the volatility and bid-offer spread data as we have seen both analytically (Corollary 1) and numerically (Figure 7) that, under our modeling assumptions, latency cost can be approximated accurately for low values of latency using only these two measures.

The data set contain volatility and bid-offer spread estimates for given stock on a particular day if the number of transactions on that day exceeds 200. The minimum, average, and maximum number of stocks in the sample on any day are 61, 653, and 1,278, respectively. In particular, earlier periods in the data set contain fewer stocks due to a smaller number of firms and a lower volume of transactions. In this data set, the bid-offer spread is estimated using only NYSE quotes in the regular trading hours. The volatility estimate is obtained using maximum likelihood estimation in the presence of market microstructure noise. Maximum likelihood estimation is preferred over other nonparametric estimation methods (e.g., "Two Scales Realized Volatility") as a simulation study shows that maximum likelihood estimation provides robust estimators under reasonable stochastic volatility and jump models in the underlying asset. The reader is urged to consult to Section 2.1 of Aït-Sahalia and Yu (2009) for full details of their estimation procedure.

For each stock in the data set, on a daily basis, we compute the latency cost facing an investor with a fixed latency of 500 ms using the asymptotic approximation of Corollary 1. These daily latency costs are then averaged over each month. Figure 8 displays percentiles of the monthly averages of latency cost over all of the stocks in the sample, as a function of time. As a representative example of a liquid name, we also report the monthly averages of latency cost of Goldman Sachs Group, Inc. (NYSE: GS). Note that the time series for GS begins from its initial public offering in 1999. For reference, we have added an additional point to this time series based on our estimation in Section 5.1 of the latency cost for GS on January 4, 2010.

Figure 8 illustrates that latency costs have had an increasing trend over the 1995–2005 period. In particular, we observe that the median latency cost incurred by trading on a human time scale roughly tripled, by increasing from approximately 5% to approximately 14%. One important factor in this increase has been the reduction of bid-offer spreads over this time period. Instances during the period when the NYSE reduced the tick size (from $1/8 to $1/16 in June 1997, and from $1/16 to $0.01 in January 2001) coincide with spikes in latency cost. This is consistent with bid-offer spreads decreasing significantly and volatility maintaining the same level at these times. This suggests that any future reduction in tick sizes will result in increased latency costs.

Using a data set in a similar time-frame, from February 2001 to December 2005, Hendershott et al. (2010) conclude that in the post-decimalization era, the increase in algorithmic trading activity

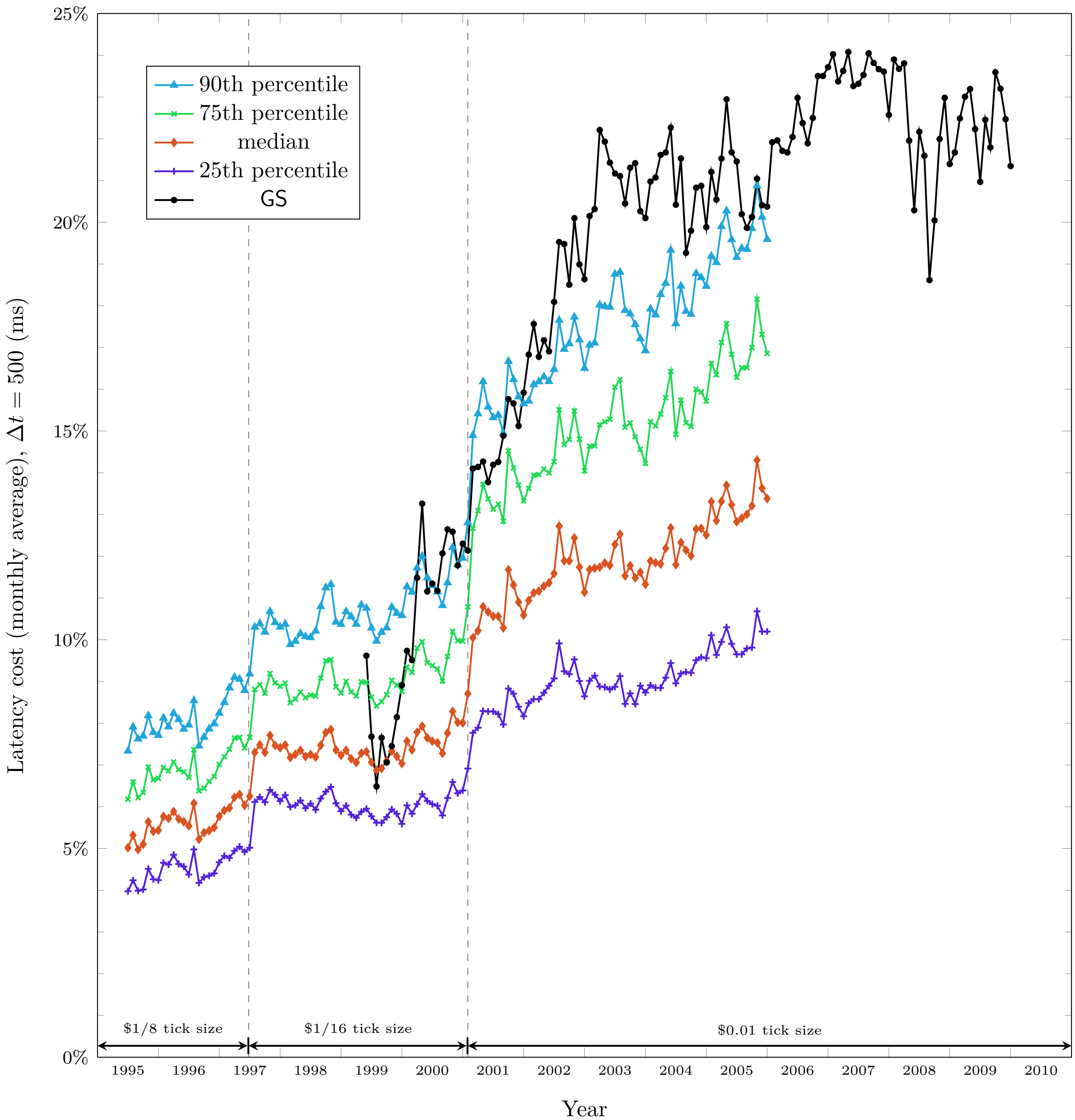

Figure 8: An illustration of the historical evolution of latency cost over the 1995–2005 time period. Here, we consider a hypothetical "human time scale" investor with a fixed latency of $\Delta t = 500$ (ms). Percentiles for the resulting latency cost are reported across NYSE common stocks. The latency costs are computed from data set of Aït-Sahalia and Yu (2009). The latency cost for GS is also reported, beginning from its IPO. The dashed lines correspond to dates where the NYSE tick size was reduced. We observe that latency cost had a consistent increasing trend over the 1995–2005 period. Specifically, the median latency cost approximately increased three-fold by reaching roughly to 14% from 5%.

had a positive impact on the level of liquidity. This result suggests that the increase in algorithmic trading in and of itself elevated the importance of low latency trading and increased the cost of latency.

5.3. Historical Evolution of Implied Latency

An alternative perspective on the historical importance of latency comes from considering a hypothetical investor with a target level for the cost of latency, relative to the overall cost-of-immediacy. The representative trader maintains this target over time through continual technological upgrades to lower levels of latency. We determine the requisite level of latency for such a trader, over time and across the aggregate market. In other words, fixing the latency cost percentage LC to the target level, we can solve the asymptotic approximation (12) for the level of latency required at each time to achieve latency cost LC. We call this the *implied latency*.

Figure 9 illustrates the implied latency values over the 1995–2005 period assuming that the target level $\mathsf{LC} = 10\%$ of overall transaction costs result from latency. We observe that the median implied latency decreased by approximately two orders of magnitude over this time frame. The 90th percentile of U.S. equities, for example, went from an implied latency on the scale of seconds to an implied latency on the scale of tens of milliseconds.

5.4. Empirical Importance of Latency

Our model captures the cost of latency due to a lack of contemporaneous information. Figure 8 suggests that, when our model is calibrated to the topmost quartile of U.S. equities, a investor with latency on the human time scale faces a latency cost of at 15% to 25%. In order to assess the significance of this, we can compare it to other trading costs. Suppose we normalize the cost of immediacy to $0.01, which is the typical bid-offer spread for a liquid U.S. equity. Then, our model suggests that the benefit of reducing latency from a human time scale of 500 ms to an ultra low latency time scale of less than 1 ms is approximately $0.0015–$0.0025 per share traded.

While this might seem very small as an absolute number, note that is of the same order of magnitude as other trading costs faced by the most cost efficient institutional investors. For example, a hedge fund would pay an average commission of $0.0007 per share for market access.[14] Furthermore, investors may pay an SEC fee of $0.0005 per share traded,[15] and exchange fees or rebates of $0.0020–$0.0030 per share traded. To the extent that a sophisticated institutional investor is cost sensitive and wishes to optimize these other execution costs, they should also be concerned with latency. This isn't to suggest that latency cost is important to all investors. A typical retail

[14] "U.S. Equity Trading: Low Touch Trends," TABB Group, July 2010.

[15] As of January 21, 2011, the SEC fee is a fraction $0.0000192 of the proceeds of an equity sale. If we assume a typical stock price of $50, this is approximately $0.0010 per share sold. Amortizing this cost equally between buys and sells results in $0.0005 per share traded.

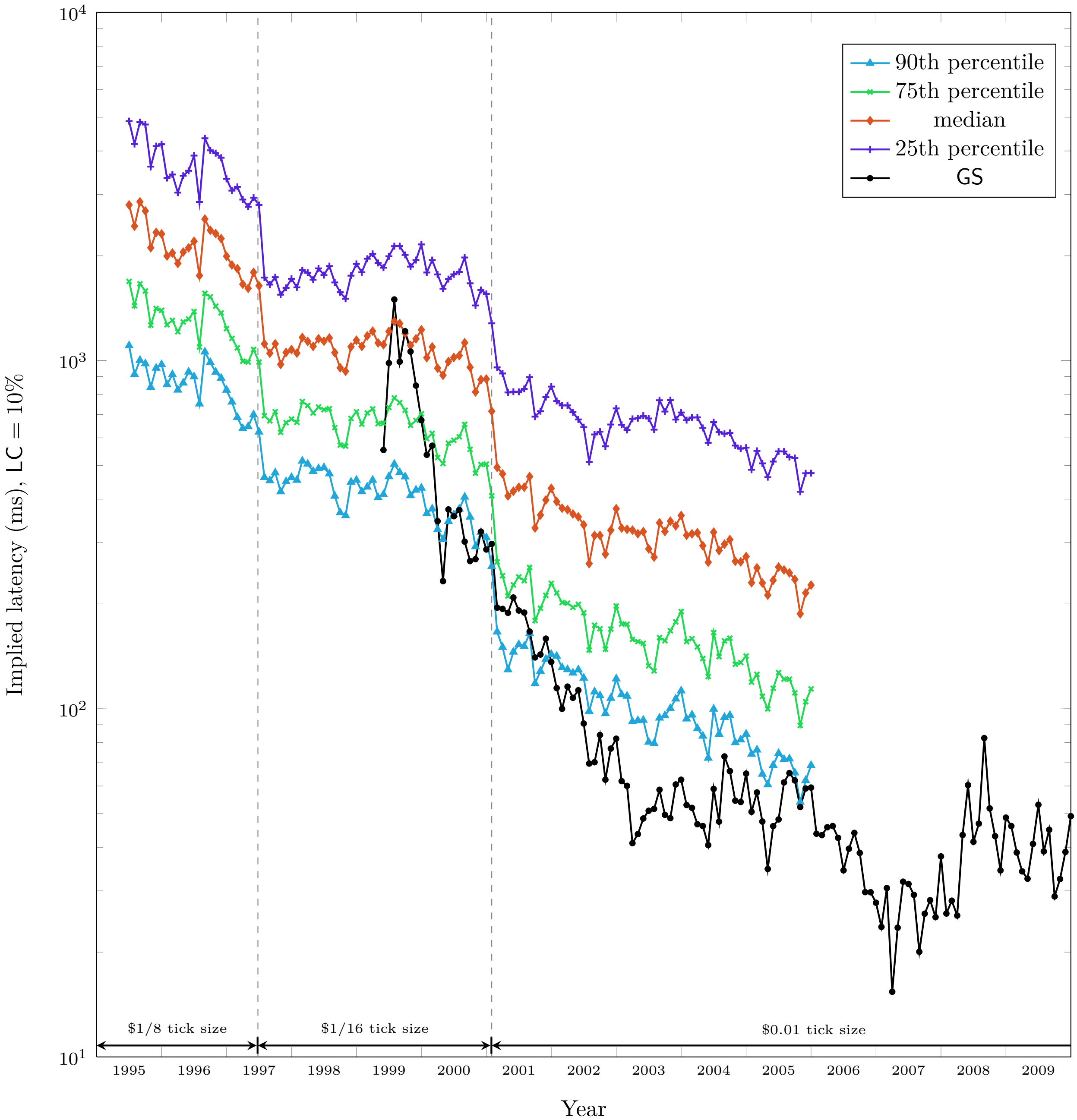

Figure 9: An illustration of the historical evolution of implied latency over the 1995–2005 time period. Here, we consider a hypothetical investor who makes sufficient technological investments to ensure a constant latency cost of 10%. The implied latency is the level of latency required to achieve this latency cost. Percentiles for the implied latency are reported across NYSE common stocks. The implied latencies are computed from data set of Aït-Sahalia and Yu (2009). The implied latency for GS is also reported, beginning from its IPO. We observe that implied latency has had a decreasing trend over the 1995–2005 period. Specifically, the median implied latency decreased by approximately two orders of magnitude over this time frame.

investor, for example, may pay a brokerage fee that is up to \$0.10 per share traded.[16] For this latter type of investor, the cost of latency as described here is not a significant component of overall trading costs.

Alternatively, we can compare the \$0.0015–\$0.0025 per share traded latency cost to the rents extracted by agents that have made the required technological investments to trade on an ultra low latency time scale. For example, providers of automated algorithmic trade execution services charge an average commission of \$0.0033 per share traded for their execution services, which leverage sophisticated low latency technological infrastructure.[17] Note that this cost is comparable to the latency cost. Another class of agents with ultra low latency trading capabilities are high frequency traders. Reported net profit numbers for high frequency traders are in the range of \$0.0010–\$0.0020 per share traded.[18] This is of the same order of magnitude as the latency cost.

6. Conclusion and Future Directions

This paper provides a model to quantify the cost of latency on transaction costs. We consider a stylized execution problem, where a trader must sell an atomic unit of stock over a fixed time horizon. We consider this model in the absence of latency as a benchmark, and we incorporate latency by not allowing the trader to continuously participate in the market. Orders submitted by the trader reach the market with a fixed latency, and the trader is forced to deviate from the benchmark policy in order to take into account the uncertainty introduced by this delay. We quantify the cost of latency as the normalized difference in expected payoffs between this model and the stylized model without latency.

Since the latency values observed in modern electronic markets are on the order of milliseconds, we provide an asymptotic analysis for the low latency regime, in which we obtain an explicit closed-form solution. In order to compute this asymptotic latency cost empirically, we only need to estimate the volatility and the average bid-offer spread of the stock. This is an elegant and practical result as data sets and estimation procedures for these quantities are readily abundant in the literature. Indeed, using an existing data set, we show that the cost of latency incurred by trading on a human time scale (500 ms) increased three-fold over the 1995–2005 time-frame. In addition, using the alternative approach of keeping a fixed level of latency cost through continuous technological improvements, we compute the various percentiles of the implied latency over this time frame. Using the same data set, we observe that the median implied latency decreased by approximately two orders of magnitude.

[16] For example, at the time of writing, the brokerage firm E-TRADE charges \$10 per trade. Assuming a typical trade of 100 shares, this cost is \$0.10 per share traded.

[17] "U.S. Equity Trading: Low Touch Trends," TABB Group, July 2010. Note that some institutional investors pay significantly larger commissions for trade execution in order to compensate their brokers for trading ideas or research services. The commission we quote here is for "non-idea driven" services that relate purely to trade execution using the algorithms and technological platform of the broker.

[18] "Tradeworx, Inc. Public Commentary on SEC Market Structure Concept Release," Tradeworx, Inc., April 2010.

Our empirical analysis can also be utilized to compare the magnitude of latency cost to other trading costs incurred by institutional investors. Our results suggest that the difference in payoff between trading with a human time scale (500 ms) and an automated trading platform with ultra low latency (1 ms) is approximately of the same order of magnitude as other trading costs faced by institutional investors. This observation certainly underlines the significance of latency for such investors. In conclusion, our model is the first theoretical approach in the literature to concretely quantify the impact of latency on the optimal order submission policy and its resulting cost to the trader.

There are a number of interesting future directions for research. First, as discussed in Section 4.4, there are a number of tractable extensions to the present model that can be analyzed. More generally, in the introduction, we identified a number of broad themes to the costs that arise from latency. The model we have presented captures mainly costs due to a lack of contemporaneous decision making. It does not capture the latency costs due to strategic effects (i.e., comparative advantage/disadvantage relative to other investors) or due to time priority rules. These remain important questions for future research.

References

Y. Aït-Sahalia and J. Jacod. Is Brownian motion necessary to model high-frequency data? *The Annals of Statistics*, 38(5):3093–3128, 2010.

Y. Aït-Sahalia and J. Yu. High frequency market microstructure noise estimates and liquidity measures. *The Annals of Applied Statistics*, 3(1):422–457, 2009.

R. Almgren and N. Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3:5–40, 2001.

J. Angel. Limit versus market orders. Working Paper No. FINC-1377-01-293, School of Business Administration, Georgetown University, Washington, D.C., 1994.

J. Bacidore, R. Battalio, and R. Jennings. Order submission strategies, liquidity supply, and trading in pennies on the New York Stock Exchange. *Journal of Financial Markets*, 6:337–362, 2003.

K. Back and S. Baruch. Working orders in limit order markets and floor exchanges. *The Journal of Finance*, 62(4):1589–1621, 2007.

O. E. Barndorff-Nielsen, D. G. Pollard, and N. Shephard. Discrete-valued Lévy processes and low latency financial econometrics. Working paper, 2010.

D. P. Bertsekas and S. E. Shreve. *Stochastic Optimal Control: The Discrete Time Case*. Academic Press, Orlando, 1978.

D. Bertsimas and A. Lo. Optimal control of execution costs. *Journal of Financial Markets*, 1(1):1–50, 1998.

D. Bertsimas, L. Kogan, and A. W. Lo. When is time continuous? *Journal of Financial Economics*, 55: 173–204, 2000.

P. P. Boyle and D. Emanuel. Discretely adjusted option hedges. *Journal of Financial Economics*, 8:259–282, 1980.

J. Brogaard. High frequency trading and its impact on market quality. Working paper, 2010.

G. Cespa and T. Foucault. Insiders-outsiders, transparency, and the value of the ticker. Working Paper No. 628, Department of Economics, Queen Mary, University of London, April 2008.

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

S.N. Cohen and L. Szpruch. A limit order book model for latency arbitrage. *Arxiv preprint arXiv:1110.4811*, 2011.

T. Copeland and D. Galai. Information effects and the bid-ask spread. *Journal of Finance*, 38(5):1457–1469, 1983.

J. Cvitanić and A. Kirilenko. High frequency traders and asset prices. Working paper, March 2010.

R. Durrett. *Probability: Theory and Examples*. Duxbury Press, 3rd edition, 2004.

D. Easley, T. Hendershott, and T. Ramadorai. The price of latency. Working paper, May 2008.

L. R. Glosten. Is the electronic open limit order book inevitable? *Journal of Finance*, 49(4):1127–1161, 1994.

S. Grossman and M. Miller. Liquidity and market structure. *Journal of Finance*, 43:617–633, 1988.

L. Harris. Optimal dynamic order submission strategies in some stylized trading problems. *Financial Markets, Institutions & Instruments*, 7(2):1–76, April 1998.

J. Hasbrouck. *Empirical Market Microstructure*. Oxford University Press, 2007.

J. Hasbrouck and G. Saar. Technology and liquidity provision: The blurring of traditional definitions. *Journal of Financial Markets*, 12:143–172, 2009.

J. Hasbrouck and G. Saar. Low-latency trading. Working paper, 2010.

T. Hendershott and R. Riordan. Algorithmic trading and information. Working paper, September 2009.

T. Hendershott, C. M. Jones, and A. Menkveld. Does algorithmic trading improve liquidity? To appear in *Journal of Finance*, April 2010.

R. A. Jarrow and P. Protter. A dysfunctional role of high frequency trading in electronic markets. Working paper, 2011.

A. Kirilenko, A. Kyle, M. Samadi, and T. Tuzun. The flash crash: The impact of high frequency trading on an electronic market. Working paper, 2010.

H. Leland. Option pricing and replication with transactions costs. *Journal of Finance*, 40:1283–1301, 1985.

A. W. Lo, A. C. MacKinlay, and J. Zhang. Econometric models of limit order execution. *Journal of Financial Economics*, 65:31–71, 2002.

E. Pagnotta. Information and liquidity trading at optimal frequencies. Technical report, Working paper, NYU Stern, 2010.

C.A. Parlour and D.J. Seppi. Limit order markets: A survey. *Handbook of Financial Intermediation and Banking*, 5, 2008.

A. F. Perold. The implementation shortfall: Paper versus reality. *Journal of Portfolio Management*, 14(3): 4–9, 1988.

M. J. Ready. The specialist's discretion: Stopped orders and price improvement. *Review of Financial Studies*, 12(5):1075–112, Winter 1999.

R. Roll. A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance*, 39:1127–1139, 1984.

P. Sandås. Adverse selection and competitive market making: Empirical evidence from a limit order market. *Review of Financial Studies*, 14(3):705–734, 2001.

H. R. Stoll and C. Schenzler. Trades outside the quotes: Reporting delay, trading option or trade size? *Journal of Financial Economics*, 79:615–653, 2006.

I. Werner. NYSE order flow, spreads, and information. *Journal of Financial Markets*, 6:309–335, 2003.

Ciamac C. Moallemi is the Barbara and Meyer Feldberg Associate Professor of Business in the Decision, Risk, and Operations Division of the Graduate School of Business at Columbia University. His research interests are in the area of the optimization and control of large-scale stochastic systems, with an emphasis on applications in financial engineering and the design of financial markets.

Mehmet Sağlam is currently a Postdoctoral Research Associate at Bendheim Center for Finance, Princeton University. His research is broadly focused on optimal (or near-optimal) dynamic decision-making in high-dimensional stochastic systems. He is particularly interested in financial applications with market frictions arising from transaction costs, illiquidity, and trading infrastructure.

Electronic Companion to "The Cost of Latency in High-Frequency Trading"

Ciamac C. Moallemi
Graduate School of Business
Columbia University
email: ciamac@gsb.columbia.edu

Mehmet Sağlam
Bendheim Center for Finance
Princeton University
email: msaglam@princeton.edu

Current Revision: February 5, 2013

A. Dynamic Programming Decomposition

In order to solve the optimal control problem (8) via dynamic programming, note that we can equivalently consider the objective of maximizing the sale price P. Consider a decision time T_i with $0 \leq i < n$, and assume that the trader's limit order remains unfilled at time T_i. The state of the system consists of the current price, S_{T_i} as well as the previously chosen limit price,[1] ℓ_{i-1}, since this price will become active at time T_i. We can define an optimal value function $J_i(S_{T_i}, \ell_{i-1})$, as a function of this state, by optimizing the eventual sale price over all future decisions. In other words,

$$J_i(S_{T_i}, \ell_{i-1}) \triangleq \underset{\ell_i,\ldots,\ell_{n-1}}{\text{maximize}} \ \mathsf{E}\left[P \mid S_{T_i}, \ell_{i-1}\right]. \tag{A.1}$$

At time $T = T_n$, the trader must sell via a market order, hence

$$J_n(S_{T_n}, \ell_{n-1}) = S_{T_n}. \tag{A.2}$$

Now, for $0 \leq i < n$, there are three mutually exclusive events one of which must occur between time T_i and time T_{i+1}. These are the events $\mathcal{E}_i^{(1)}$, $\mathcal{E}_i^{(2)}$, and $\mathcal{E}_i^{(3)}$ described in Section 3. By considering cases corresponding to these events, we have the Bellman equation

$$J_i(S_{T_i}, \ell_{i-1}) \triangleq \max_{\ell_i} \ \mathsf{E}\left[\mathbb{I}_{\mathcal{E}_i^{(1)}} \ell_{i-1} + \mathbb{I}_{\mathcal{E}_i^{(2)}} S_{T_{i+1}} + \mathbb{I}_{\mathcal{E}_i^{(3)}} J_{i+1}(S_{T_{i+1}}, \ell_i) \ \middle|\ S_{T_i}, \ell_{i-1}\right]. \tag{A.3}$$

Here, the first term corresponds to an execution at the prior price ℓ_{i-1}, the second term corresponds to the price ℓ_i being crossed by the bid price upon arrival to the market, and the third term corresponds to all other cases.

Define the function Q_i, for $0 \leq i \leq n$, by

$$Q_i(S_{T_i}, v_{i-1}) \triangleq J_i(S_{T_i}, S_{T_i} + v_{i-1}) - S_{T_i}.$$

[1] We will assume that $\ell_{-1} = \infty$, i.e., there is no limit order active at the beginning of the time horizon.

The function Q_i is the premium of the value at time T_i, *relative* to the current bid price S_{T_i}. Similarly, $v_{i-1} \triangleq \ell_{i-1} - S_{T_i}$ is the premium of limit price decided at time T_{i-1} *relative* to the current bid price at time T_i. Then, applying (A.3), we have for $0 \le i < n$,

$$\begin{aligned} & Q_i(S_{T_i}, v_{i-1}) \\ &= \max_{\ell_i} \ \mathsf{E}\left[\mathbb{I}_{\mathcal{E}_i^{(1)}}(S_{T_i} + v_{i-1}) + \mathbb{I}_{\mathcal{E}_i^{(2)}} S_{T_{i+1}} + \mathbb{I}_{\mathcal{E}_i^{(3)}} J_{i+1}(S_{T_{i+1}}, \ell_i) \ \middle| \ S_{T_i}, v_{i-1} \right] - S_{T_i} \\ &= \max_{u_i} \ \mathsf{E}\left[\mathbb{I}_{\mathcal{E}_i^{(1)}} v_{i-1} + \mathbb{I}_{\mathcal{E}_i^{(2)} \cup \mathcal{E}_i^{(3)}} X_{i+1} + \mathbb{I}_{\mathcal{E}_i^{(3)}} Q_{i+1}(S_{T_i} + X_{i+1}, u_i - X_{i+1}) \ \middle| \ S_{T_i}, v_{i-1} \right]. \end{aligned}$$

Here, $X_{i+1} \triangleq S_{T_{i+1}} - S_{T_i} \sim N(0, \sigma^2 \Delta t)$ is the change in bid price from time T_i to time T_{i+1}. We define $u_i \triangleq \ell_i - S_{T_i}$ as the premium of the limit price at time T_i (i.e., the decision variable) *relative* to the current bid price S_{T_i}. Note that the price change X_{i+1} is zero mean under the event

$$\mathcal{E}_i^{(2)} \cup \mathcal{E}_i^{(3)} = \left(\mathcal{E}_i^{(1)} \right)^c,$$

by our assumption that the arrival of impatient buyers is independent of the bid price process, hence

$$Q_i(S_{T_i}, v_{i-1}) = \max_{u_i} \ \mathsf{E}\left[\mathbb{I}_{\mathcal{E}_i^{(1)}} v_{i-1} + \mathbb{I}_{\mathcal{E}_i^{(3)}} Q_{i+1}(S_{T_i} + X_{i+1}, u_i - X_{i+1}) \ \middle| \ S_{T_i}, v_{i-1} \right]. \tag{A.4}$$

Finally, by (A.2),

$$Q_n(S_{T_n}, v_{n-1}) = 0. \tag{A.5}$$

As should be clear from the above discussion, the Bellman equation (A.3) with terminal condition (A.2) and the backward recursion (A.4) with terminal condition (A.5) are completely equivalent, up to a change in variables. Expressing these equations in the latter form, however, brings significant simplifications, as the following lemma shows.

Lemma 3. *Suppose a collection of functions $\{Q_i\}$ satisfies the dynamic programming equations (A.4)–(A.5). Then, for each $0 \le i < n$, Q_i does not depend on the price S_{T_i}, and takes the form*

$$Q_i(v_{i-1}) = \mathbb{I}_{\{v_{i-1} \le \delta\}} \left[\mu \Delta t v_{i-1} + (1 - \mu \Delta t) h_i \right] + \mathbb{I}_{\{v_{i-1} > \delta\}} h_i, \tag{A.6}$$

where the scalar h_i satisfies

$$h_i = \max_{u_i} \ \mathsf{P}(X_{i+1} < u_i) \mathsf{E}\left[Q_{i+1}(u_i - X_{i+1}) \ | \ X_{i+1} < u_i \right]. \tag{A.7}$$

Proof. Observe that, for $0 \leq i < n$, (A.4) can be simplified according to

$$\begin{aligned} & Q_i(S_{T_i}, v_{i-1}) \\ \text{(A.8)} \qquad & = \max_{u_i} \; \mu\Delta t v_{i-1}\mathbb{I}_{\{v_{i-1}\leq\delta\}} \\ & \qquad + \Big(1-\mu\Delta t v_{i-1}\mathbb{I}_{\{v_{i-1}\leq\delta\}}\Big)\,\mathsf{E}\Big[\mathbb{I}_{\{X_{i+1}<u_i\}}Q_{i+1}(S_{T_i}+X_{i+1}, u_i - X_{i+1}) \;\Big|\; S_{T_i}\Big], \end{aligned}$$

where we have used the definitions of the events $\mathcal{E}_i^{(1)}$ and and $\mathcal{E}_i^{(3)}$.

Now, we proceed by backward induction. For the terminal case $i = n-1$, from (A.8) and the fact that $Q_n = 0$ and $u_{n-1} = -\infty$ (i.e., the trader must use a market order at the last time slot), we have that

$$Q_{n-1}(S_{T_{n-1}}, v_{n-2}) = \mu\Delta t v_{n-2}\mathbb{I}_{\{v_{n-2}\leq\delta\}}.$$

In other words, Q_{n-1} satisfies the hypotheses of the lemma, with $h_{n-1} = 0$.

Now, suppose that the result holds for some $0 \leq i+1 < n$. By (A.8), and since Q_{i+1} does not depend on $S_{T_{i+1}}$,

$$\begin{aligned} & Q_i(S_{T_i}, v_{i-1}) \\ & = \max_{u_i} \; \mu\Delta t v_{i-1}\mathbb{I}_{\{v_{i-1}\leq\delta\}} + \Big(1-\mu\Delta t v_{i-1}\mathbb{I}_{\{v_{i-1}\leq\delta\}}\Big)\,\mathsf{E}\Big[\mathbb{I}_{\{X_{i+1}<u_i\}}Q_{i+1}(u_i - X_{i+1})\Big] \\ & = \mu\Delta t v_{i-1}\mathbb{I}_{\{v_{i-1}\leq\delta\}} + \Big(1-\mu\Delta t v_{i-1}\mathbb{I}_{\{v_{i-1}\leq\delta\}}\Big)\, h_i \\ & = \mathbb{I}_{\{v_{i-1}\leq\delta\}}\big[\mu\Delta t v_{i-1} + (1-\mu\Delta t v_{i-1})h_i\big] + \mathbb{I}_{\{v_{i-1}>\delta\}}h_i. \end{aligned}$$

Here, in the second equality, we define h_i through (A.7). The result then follows. ■

Notice that, at the beginning of the trading horizon, there is no active limit order, i.e., $u_{-1} = \infty$. From Lemma 3, we have that

$$h_0 = Q_0(\infty) = \underset{\ell_0,\ldots,\ell_{n-1}}{\text{maximize}} \; \mathsf{E}\,[P \mid S_0] - S_0.$$

In other words, $h_0 = h_0(\Delta t)$, as defined in (8), and our notation is consistent. More generally, for $i > 0$, from (A.7), we can interpret h_i to be the trader's expected payoff at time T_i relative to the current bid price under the optimal policy, assuming that the limit order does not get executed in that time slot. Thus, h_i can be interpreted as a *continuation value* in the dynamic programming context, as in (9).

The continuation values $\{h_i\}$ allow for a compact representation of the value function, since they consist of only a single real number for each time slot, rather than a function of the entire state space. Theorem 1 directly expresses the dynamic programming equations (A.4)–(A.5) in terms of this representation. The proof follows by explicitly computing the expectations in Lemma 3.

Theorem 1. *Suppose $\{h_i\}$ satisfy, for $0 \leq i < n-1$,*

$$\begin{aligned} h_i = \max_{u_i} \; \Bigg\{ & \mu \Delta t \left[u_i \left(\Phi\left(\frac{u_i}{\sigma\sqrt{\Delta t}} \right) - \Phi\left(\frac{u_i - \delta}{\sigma\sqrt{\Delta t}} \right) \right) + \sigma\sqrt{\Delta t} \left(\phi\left(\frac{u_i}{\sigma\sqrt{\Delta t}} \right) - \phi\left(\frac{u_i - \delta}{\sigma\sqrt{\Delta t}} \right) \right) \right] \\ & + h_{i+1} \left[(1 - \mu\Delta t)\Phi\left(\frac{u_i}{\sigma\sqrt{\Delta t}} \right) + \mu\Delta t \Phi\left(\frac{u_i - \delta}{\sigma\sqrt{\Delta t}} \right) \right] \Bigg\}, \end{aligned} \tag{A.9}$$

and

$$h_{n-1} = 0. \tag{A.10}$$

Here, ϕ and Φ are, respectively, the p.d.f. and c.d.f. of the standard normal distribution. Then, $\{h_i\}$ correspond to the continuation values under the optimal policy. In other words, the value functions $\{Q_i\}$ defined by $\{h_i\}$ via (A.6) *solve the dynamic programming equations* (A.4)*–*(A.5)*.*

Suppose further that, for $0 \leq i < n-1$, u_i^* is a maximizer of (A.9)*. Then, a policy which chooses limit prices according to the premia defined by $\{u_i^*\}$, i.e.,*

$$\ell_i^* = S_{T_i} + u_i^*, \quad \forall \; 0 \leq i < n-1,$$

is optimal.

Proof. Suppose that we are given $\{h_i\}$ that satisfy the hypotheses of the theorem. Define $\{Q_i\}$ by setting, for $0 \leq i \leq n-1$,

$$Q_i(v_{i-1}) \triangleq \mathbb{I}_{\{v_{i-1} \leq \delta\}} [\mu\Delta t v_{i-1} + (1 - \mu\Delta t) h_i] + \mathbb{I}_{\{v_{i-1} > \delta\}} h_i, \tag{A.11}$$

and $Q_n \triangleq 0$. We wish to show that $\{Q_i\}$ solve the dynamic programming equations (A.4)–(A.5).

Note that (A.5) holds by definition. For $0 \leq i < n$, we have that (A.4) is equivalent to (A.8). Define $\hat{Q}_i$ to be the right side of (A.8), i.e.,

$$\hat{Q}_i(v_{i-1}) \triangleq \mu\Delta t v_{i-1} \mathbb{I}_{\{v_{i-1} \leq \delta\}} + \left(1 - \mu\Delta t v_{i-1} \mathbb{I}_{\{v_{i-1} \leq \delta\}}\right) \max_{u_i} \; \mathsf{E}\left[\mathbb{I}_{\{X_{i+1} < u_i\}} Q_{i+1}(u_i - X_{i+1}) \right].$$

Comparing with (A.11), in order that the dynamic programming equation (A.8) hold (i.e., that $\hat{Q}_i = Q_i$), we must have that

$$h_i = \max_{u_i} \; \mathsf{E}\left[\mathbb{I}_{\{X_{i+1} < u_i\}} Q_{i+1}(u_i - X_{i+1}) \right] \tag{A.12}$$

Using the definition of Q_{i+1} from (A.11), this is equivalent to

$$\begin{aligned} & h_i \\ & = \max_{u_i} \; \mathsf{E}\left[\mathbb{I}_{\{X_{i+1} < u_i\}} \left(\mathbb{I}_{\{u_i - X_{i+1} \leq \delta\}} [\mu\Delta t (u_i - X_{i+1}) + (1 - \mu\Delta t) h_{i+1}] + \mathbb{I}_{\{u_i - X_{i+1} > \delta\}} h_{i+1} \right) \right] \\ & = \max_{u_i} \; \mathsf{E}\left[\mathbb{I}_{\{0 < u_i - X_{i+1} \leq \delta\}} \mu\Delta t (u_i - X_{i+1}) + \mathbb{I}_{\{X_{i+1} < u_i\}} (1 - \mu\Delta t) h_{i+1} + \mathbb{I}_{\{u_i - X_{i+1} > \delta\}} \mu\Delta t h_{i+1} \right]. \end{aligned}$$

For the first term in the expectation, we have

$$
\begin{aligned}
&\mathsf{E}\left[\mathbb{I}_{\{0<u_i-X_{i+1}\leq\delta\}}\mu\Delta t(u_i-X_{i+1})\right]\\
&=\mu\Delta t\int_{-\infty}^{u_i}(u_i-x)\mathbb{I}_{\{u_i-x\leq\delta\}}\frac{1}{\sigma\sqrt{\Delta t}}\phi\left(\frac{x}{\sigma\sqrt{\Delta t}}\right)dx\\
&=\mu\Delta t\int_{u_i-\delta}^{u_i}(u_i-x)\frac{1}{\sigma\sqrt{\Delta t}}\phi\left(\frac{x}{\sigma\sqrt{\Delta t}}\right)dx\\
&=\mu\Delta t\left[u_i\left(\Phi\left(\frac{u_i}{\sigma\sqrt{\Delta t}}\right)-\Phi\left(\frac{u_i-\delta}{\sigma\sqrt{\Delta t}}\right)\right)+\int_{u_i-\delta}^{u_i}\frac{-x}{\sigma\sqrt{\Delta t}}\phi\left(\frac{x}{\sigma\sqrt{\Delta t}}\right)dx\right]\\
&=\mu\Delta t\left[u_i\left(\Phi\left(\frac{u_i}{\sigma\sqrt{\Delta t}}\right)-\Phi\left(\frac{u_i-\delta}{\sigma\sqrt{\Delta t}}\right)\right)+\sigma\sqrt{\Delta t}\left(\phi\left(\frac{u_i}{\sigma\sqrt{\Delta t}}\right)-\phi\left(\frac{u_i-\delta}{\sigma\sqrt{\Delta t}}\right)\right)\right].
\end{aligned}
$$

For the second term in the expectation, we have

$$
\mathsf{E}\left[\mathbb{I}_{\{X_{i+1}<u_i\}}(1-\mu\Delta t)h_{i+1}\right]=(1-\mu\Delta t)h_{i+1}\Phi\left(\frac{u_i}{\sigma\sqrt{\Delta t}}\right).
$$

finally, for the last term in the expectation, we have

$$
\mathsf{E}\left[\mathbb{I}_{\{u_i-X_{i+1}>\delta\}}\mu\Delta t h_{i+1}\right]=\mu\Delta t h_{i+1}\Phi\left(\frac{u_i-\delta}{\sigma\sqrt{\Delta t}}\right)
$$

Combining all the terms, we obtain the desired recursion for h_i.

The balance of the theorem involves establishing the optimality of the $\{u_i^*\}$ policy. This follows from standard dynamic programming arguments (see, e.g., Chapter 3, Bertsekas and Shreve, 1978).

■

B. Proof of Theorem 2

We begin with a preliminary lemma.

Lemma 4. *Suppose that* $\{h_i : 0 \leq i < n\}$ *solves the dynamic programming recursion* (10)–(11). *Then, for* $0 \leq i < n$,

(B.1)
$$
0 \leq h_i \leq \delta\big(1-(1-\mu\Delta t)^n\big) < \delta.
$$

Proof. First, note that the result is trivially true for $i = n-1$, since $h_{n-1} = 0$. Now, if $0 \leq i < n-1$, we can always choose $u_i = -\infty$, i.e., a market order, and this results in a continuation value of 0. Thus, $h_i \geq 0$.

For the upper bound, consider the discrete model without latency described in Section 4.3. Any strategy for the latency model is also feasible for the discrete model, since the trader can simply delay the implementation of trading decisions by one period. Therefore, at time T_i (with $0 \leq i < n-1$), a policy with latency cannot achieve more value than the optimal policy for the discrete model without latency. At time T_i, there are $n-i-1$ trading decisions remaining. This corresponds to the initial time of a discrete model with a total time horizon of $(n-i-1)\Delta t$. Then, with reference to Lemma 2, we have that

$$h_i \leq \delta(1-(1-\mu\Delta t)^{n-i-1}).$$

The result immediately follows. ■

Theorem 2. *Fix $\alpha > 1$. If Δt is sufficiently small, then there exists a unique optimal solution $\{h_i\}$ to the dynamic programming equations (10)–(11). Moreover, the corresponding optimal policy $\{u_i^*\}$ is unique. For $0 \leq i < n-1$, this strategy chooses limit prices in the range*

$$\ell_i^* \in \left(S_i + \delta - \sigma\sqrt{\Delta t \log \frac{\alpha L}{\Delta t}},\ S_i + \delta - \sigma\sqrt{\Delta t \log \frac{R(\Delta t)}{\Delta t}}, \right),$$

where

$$L \triangleq \frac{\delta^2}{2\pi\sigma^2}, \qquad R(\Delta t) \triangleq \frac{\delta^2(1-\mu\Delta t)^{2n}}{2\pi\sigma^2}.$$

Proof. Assume that, for some $0 \leq i < n-1$, a solution $\{h_j \ : \ i+1 \leq j < n\}$ exists to (10)–(11). We will establish that, for Δt sufficiently small (and not dependent on i), a solution h_i also exists and satisfies the conditions of the theorem. The result will follow by backward induction. Note that the base case of our induction (i.e., the existence of h_{n-1}) is trivial.

To this end, define the auxiliary function f by

$$\begin{aligned} f(u,h) \triangleq{}& \mu\Delta t\Big[u(\Phi(A_u)-\Phi(B_u)) + \sigma\sqrt{\Delta t}(\phi(A_u)-\phi(B_u))\Big] \\ &+ h\Big[(1-\mu\Delta t)\Phi(A_u) + \mu\Delta t\Phi(B_u)\Big], \end{aligned} \tag{B.2}$$

where

$$A_u \triangleq \frac{u}{\sigma\sqrt{\Delta t}}, \qquad B_u \triangleq \frac{u-\delta}{\sigma\sqrt{\Delta t}}. \tag{B.3}$$

Then, from Theorem 1, for $0 \leq i < n-1$, the dynamic programming recursion is given by

$$h_i = \max_{u_i} f(u_i, h_{i+1}), \tag{B.4}$$

and we can establish the present theorem by proving that, for Δt sufficiently small, (B.4) has a unique maximizer $u_i^* \in (\hat{u}_L, \hat{u}_R)$, where

$$\hat{u}_L \triangleq \delta - \sigma\sqrt{\Delta t \log \frac{\alpha L}{\Delta t}}, \qquad \hat{u}_R \triangleq \delta - \sigma\sqrt{\Delta t \log \frac{R(\Delta t)}{\Delta t}}. \tag{B.5}$$

Note that

$$R(0) \triangleq \lim_{\Delta t \to 0} R(\Delta t) = \lim_{\Delta t \to 0} L(1-\mu\Delta t)^{2T/dt} = Le^{-2\mu T} < \alpha L. \tag{B.6}$$

Hence, there exists some $\overline{\Delta t} > 0$ so that if $0 < \Delta t < \overline{\Delta t}$, then

$$\delta/2 < \hat{u}_L < \hat{u}_R < \delta, \quad \text{and} \quad 0 < 1 - \mu\Delta t < 1.$$

For the balance of the theorem, we will assume that $0 < \Delta t < \overline{\Delta t}$, in addition to whatever other assumptions are made regarding the magnitude of Δt.

The first and second derivatives of $f(\cdot, h)$ are given by

$$\begin{aligned} f_u(u,h) &= \mu\Delta t \left[\Phi(A_u) - \Phi(B_u) + A_u(\phi(A_u) - \phi(B_u)) - \frac{u\phi(A_u) + (\delta - u)\phi(B_u)}{\sigma\sqrt{\Delta t}}\right] \\ &\quad + \frac{h}{\sigma\sqrt{\Delta t}}\Big[(1-\mu\Delta t)\phi(A_u) + \mu\Delta t\phi(B_u)\Big] \\ &= \frac{(1-\mu\Delta t)h}{\sigma\sqrt{\Delta t}}\phi(A_u) + \mu\Delta t\left[\Phi(A_u) - \Phi(B_u) - \frac{\delta}{\sigma\sqrt{\Delta t}}\phi(B_u)\right] + \frac{h\mu\sqrt{\Delta t}}{\sigma}\phi(B_u) \\ &= \frac{(1-\mu\Delta t)h}{\sigma\sqrt{\Delta t}}\phi(A_u) + \mu\Delta t\big[\Phi(A_u) - \Phi(B_u)\big] + \frac{\mu\sqrt{\Delta t}}{\sigma}\phi(B_u)(h-\delta), \end{aligned}$$

$$\begin{aligned} f_{uu}(u,h) &= \frac{-u(1-\mu\Delta t)h}{\sigma^3\Delta t\sqrt{\Delta t}}\phi(A_u) + \frac{\mu\sqrt{\Delta t}}{\sigma}\big[\phi(A_u) - \phi(B_u)\big] + \frac{\mu(\delta-u)}{\sigma^3\sqrt{\Delta t}}\phi(B_u)(h-\delta) \\ &= \phi(A_u)\left[\frac{\mu\sqrt{\Delta t}}{\sigma} - \frac{u(1-\mu\Delta t)h}{\sigma^3\Delta t^{3/2}}\right] + \phi(B_u)\left[\frac{\mu(\delta-u)}{\sigma^3\sqrt{\Delta t}}(h-\delta) - \frac{\mu\sqrt{\Delta t}}{\sigma}\right]. \end{aligned} \tag{B.7}$$

First, we will show that, for Δt sufficiently small, $f(\cdot, h_{i+1})$ has a *local* maximum u_i^* in the interval $(\hat{u}_L, \hat{u}_R)$, and that this is the unique maximizer over the larger interval $(\delta/2, \delta)$. That is, $u \in (\delta/2, \delta)$ and $u \neq u_i^*$, then

$$f(u, h_{i+1}) < f(u_i^*, h_{i+1}), \quad \text{for all } u \in (\delta/2, \delta),\ u \neq u_i^*. \tag{B.8}$$

This is implied by the following claims, which we will demonstrate hold for Δt sufficiently small:

(i) $f_u(\hat{u}_L, h_{i+1}) > 0$.

(ii) $f_u(\hat{u}_R, h_{i+1}) < 0$.

(iii) $f_{uu}(u, h_{i+1}) < 0$, for all $u \in (\delta/2, \delta)$.

Claim (i): Note that

$$
\text{(B.9)} \qquad \begin{aligned}
f_u(\hat{u}_L, h_{i+1}) &= \frac{(1-\mu\Delta t)h_{i+1}}{\sigma\sqrt{\Delta t}}\phi(A_{\hat{u}_L}) + \mu\Delta t\big[\Phi(A_{\hat{u}_L}) - \Phi(B_{\hat{u}_L})\big] + \frac{\mu\Delta t}{\delta\sqrt{\alpha}}(h_{i+1}-\delta) \\
&\geq \mu\Delta t\big[\Phi(A_{\hat{u}_L}) - \Phi(B_{\hat{u}_L})\big] - \frac{\mu\Delta t}{\sqrt{\alpha}},
\end{aligned}
$$

where we use the fact that $h_{i+1} \geq 0$ (cf. Lemma 4). In order to calculate a lower bound for $\Phi(A_{\hat{u}_L}) - \Phi(B_{\hat{u}_L})$, we need the following standard bound on the tail probabilities of the normal distribution (see, e.g., Durrett, 2004). Define Q to be the tail probability of a standard normal distribution, i.e.,

$$
Q(x) \triangleq 1 - \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-\frac{1}{2}u^2}\, du.
$$

Then, for all $x > 0$,

$$
\text{(B.10)} \qquad \frac{x^2-1}{x^3\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \leq Q(x) \leq \frac{1}{x\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.
$$

Applying this to (B.9),

$$
\begin{aligned}
f_u(\hat{u}_L, h_{i+1}) &\geq \mu\Delta t\Big[1 - Q(A_{\hat{u}_L}) - Q(-B_{\hat{u}_L})\Big] + \frac{\mu\Delta t}{\sqrt{\alpha}} \\
&= \mu\Delta t\left(1 - \frac{1}{\sqrt{\alpha}}\right) - \mu\Delta t\big[Q(A_{\hat{u}_L}) + Q(-B_{\hat{u}_L})\big] \\
&> \mu\Delta t\left(1 - \frac{1}{\sqrt{\alpha}}\right) - 2\mu\Delta t Q(-B_{\hat{u}_L}) \\
&\geq \mu\Delta t\left(1 - \frac{1}{\sqrt{\alpha}}\right) - \frac{2\mu\Delta t^{3/2}}{\sqrt{2\pi\alpha L\log\frac{\alpha L}{\Delta t}}} \\
&> 0,
\end{aligned}
$$

for sufficiently small Δt. Here, we have used the fact that $Q(-B_{\hat{u}_L}) > Q(A_{\hat{u}_L})$.

Claim (ii): Similarly, for the other endpoint of the interval, we have

$$\begin{aligned} f_u(\hat{u}_R, h_{i+1}) &= \frac{(1-\mu\Delta t)h_{i+1}}{\sigma\sqrt{\Delta t}}\phi(A_{\hat{u}_R}) + \mu\Delta t[1 - Q(A_{\hat{u}_R}) - Q(-B_{\hat{u}_R})] \\ &\quad + \frac{\mu\Delta t}{\delta(1-\mu\Delta t)^n}(h_{i+1} - \delta) \\ &\leq \frac{(1-\mu\Delta t)h_{i+1}}{\sigma\sqrt{\Delta t}}\phi(A_{\hat{u}_R}) + \mu\Delta t[1 - Q(A_{\hat{u}_R}) - Q(-B_{\hat{u}_R})] \\ &\quad + \frac{\mu\Delta t}{\delta(1-\mu\Delta t)^n}\left[\delta\left(1 - (1-\mu\Delta t)^n\right) - \delta\right] \\ &= \frac{(1-\mu\Delta t)h_{i+1}}{\sigma\sqrt{\Delta t}}\phi(A_{\hat{u}_R}) + \mu\Delta t[1 - Q(A_{\hat{u}_R}) - Q(-B_{\hat{u}_R})] - \mu\Delta t \\ &= \frac{(1-\mu\Delta t)h_{i+1}}{\sigma\sqrt{\Delta t}}\phi(A_{\hat{u}_R}) - \mu\Delta t[Q(A_{\hat{u}_R}) + Q(-B_{\hat{u}_R})] \\ &\leq \frac{\delta}{\sigma\sqrt{\Delta t}}\phi(A_{\hat{u}_R}) - \mu\Delta t Q(-B_{\hat{u}_R}), \end{aligned}$$

where we have used the upper bound on h_{i+1} from Lemma 4. Using (B.6), for sufficiently small Δt, we have

$$\sqrt{\log\frac{R(\Delta t)}{\Delta t}} < \frac{\delta}{2\sigma\sqrt{\Delta t}},$$

and thus

$$A_{\hat{u}_R} \geq \frac{\delta}{2\sigma\sqrt{\Delta t}}.$$

On the other hand, using (B.10),

$$Q(-B_{\hat{u}_R}) \geq \left[\left(\log\frac{R(\Delta t)}{\Delta t}\right)^{-1/2} - \left(\log\frac{R(\Delta t)}{\Delta t}\right)^{-3/2}\right]\sqrt{\frac{\Delta t}{2\pi R(\Delta t)}}.$$

Thus,

$$\begin{aligned} f_u(\hat{u}_R, h_{i+1}) &\leq \frac{\delta}{\sigma\sqrt{2\pi\Delta t}}\exp\left(\frac{-\delta^2}{8\sigma^2\Delta t}\right) \\ &\quad - \mu\Delta t\left[\left(\log\frac{R(\Delta t)}{\Delta t}\right)^{-1/2} - \left(\log\frac{R(\Delta t)}{\Delta t}\right)^{-3/2}\right]\sqrt{\frac{\Delta t}{2\pi R(\Delta t)}} \\ &< 0, \end{aligned}$$

for sufficiently small Δt.

Claim (iii): Note that, for $u \in (\delta/2, \delta)$,

$$\phi\left(\frac{\delta}{\sigma\sqrt{\Delta t}}\right) < \phi(A_u) < \phi\left(\frac{\delta}{2\sigma\sqrt{\Delta t}}\right) < \phi(B_u) < \phi(0).$$

Then, from (B.7), and using the fact that $0 \leq h_{i+1} < \delta$ (cf. Lemma 4), we have for Δt sufficiently small,

$$\begin{aligned} f_{uu}(u, h_{i+1}) &\leq \phi(A_u)\frac{\mu\sqrt{\Delta t}}{\sigma} + \phi(B_u)\left[\frac{\mu(\delta - u)}{\sigma^3\sqrt{\Delta t}}(h_{i+1} - \delta) - \frac{\mu\sqrt{\Delta t}}{\sigma}\right] \\ &\leq \phi(B_u)\frac{\mu(\delta - u)}{\sigma^3\sqrt{\Delta t}}(h_{i+1} - \delta) < 0. \end{aligned}$$

In order to complete the proof, it suffices to demonstrate that the *local* maximum $u_i^* \in (\hat{u}_L, \hat{u}_R)$ is the unique *global* maximum. Since u_i^* achieves a higher value than any other $u \in (\delta/2, \delta)$, we will analyze cases where $u \notin (\delta/2, \delta)$ as follows:

- $u \in [0, \delta/2]$.

  Here,
  $$\phi\left(\frac{\delta}{\sigma\sqrt{\Delta t}}\right) \leq \phi(B_u) \leq \phi\left(\frac{\delta}{2\sigma\sqrt{\Delta t}}\right) \leq \phi(A_u) \leq \phi(0).$$
  Further, for Δt sufficiently small,
  $$\Phi(A_u) - \Phi(B_u) \geq \Phi(0) - \Phi\left(\frac{-\delta}{2\sigma\sqrt{\Delta t}}\right) \geq \frac{1}{4}.$$
  Then, for Δt sufficiently small,
  $$\begin{aligned} f_u(u, h_{i+1}) &\geq \frac{(1-\mu\Delta t)h_{i+1}}{\sigma\sqrt{\Delta t}}\phi\left(\frac{\delta}{2\sigma\sqrt{\Delta t}}\right) + \mu\Delta t\left[\Phi(0) - \Phi\left(\frac{-\delta}{2\sigma\sqrt{\Delta t}}\right)\right] \\ &\quad + \frac{\mu\sqrt{\Delta t}}{\sigma}\phi\left(\frac{\delta}{2\sigma\sqrt{\Delta t}}\right)(h_{i+1} - \delta) \\ &\geq \frac{h_{i+1}}{\sigma\sqrt{\Delta t}}\phi\left(\frac{\delta}{2\sigma\sqrt{\Delta t}}\right) + \mu\Delta t\left[\Phi(0) - \Phi\left(\frac{-\delta}{2\sigma\sqrt{\Delta t}}\right)\right] \\ &\quad - \frac{\delta\mu\sqrt{\Delta t}}{\sigma}\phi\left(\frac{\delta}{2\sigma\sqrt{\Delta t}}\right) \\ &\geq \frac{\mu\Delta t}{4} - \frac{\mu\delta\sqrt{\Delta t}}{\sqrt{2\pi}\sigma}\exp\left(\frac{-\delta^2}{8\sigma^2\Delta t}\right) > 0. \end{aligned} \tag{B.11}$$
  Here, we have used the fact that $h_{i+1} \geq 0$. Using (B.8) and the fact that $f(\cdot, h_{i+1})$ is continuous, this implies that
  $$\sup_{u\in[0,\delta/2]} f(u, h_{i+1}) \leq f(\delta/2, h_{i+1}) < f(u_i^*, h_{i+1}). \tag{B.12}$$

- $u \in (-\infty, 0)$.

In this case, since $h_{i+1} \geq 0$ and $B_u < A_u < 0$,

$$\begin{aligned} f_u(u, h_{i+1}) &\geq \mu\Delta t \left[\Phi(A_u) - \Phi(B_u)\right] - \frac{\delta\mu\sqrt{\Delta t}}{\sigma}\phi(B_u) \\ &= \mu\Delta t \int_{B_u}^{A_u} \phi(z)\, dz - \frac{\delta\mu\sqrt{\Delta t}}{\sigma}\phi(B_u) \\ &> \mu\Delta t(A_u - B_u)\phi(B_u) - \frac{\delta\mu\sqrt{\Delta t}}{\sigma}\phi(B_u) = 0. \end{aligned}$$

In conjunction with (B.12), this implies that

$$\text{(B.13)} \qquad \sup_{u\in(-\infty,0)} f(u, h_{i+1}) \leq f(0, h_{i+1}) < f(u_i^*, h_{i+1}).$$

- $u \in [\delta, \infty)$.

  In this case, using the upper bound on h_{i+1} from Lemma 4,

$$\text{(B.14)} \qquad \begin{aligned} f_u(u, h_{i+1}) \leq \frac{\delta}{\sigma\sqrt{\Delta t}}\phi(A_u) + \mu\Delta t \left[\Phi(A_u) - \Phi(B_u)\right] \\ - \frac{\mu\delta(1-\mu\Delta t)^n\sqrt{\Delta t}}{\sigma}\phi(B_u). \end{aligned}$$

  Consider two cases. First, assume that $u > \delta + \sqrt{\Delta t}$. Then, applying (B.10),

$$\begin{aligned} f_u(u, h_{i+1}) &\leq \frac{\delta}{\sigma\sqrt{\Delta t}}\phi(A_u) + \mu\Delta t Q(B_u) - \frac{\mu\delta(1-\mu\Delta t)^n\sqrt{\Delta t}}{\sigma}\phi(B_u) \\ &\leq \frac{\delta}{\sigma\sqrt{\Delta t}}\phi(A_u) + \frac{\mu\sigma\Delta t^{3/2}}{u-\delta}\phi(B_u) - \frac{\mu\delta(1-\mu\Delta t)^n\sqrt{\Delta t}}{\sigma}\phi(B_u) \\ &\leq \phi(B_u)\left[\frac{\delta}{\sigma\sqrt{\Delta t}}\exp\left(\frac{-\delta^2}{2\sigma^2\Delta t}\right) + \mu\sigma\Delta t - \frac{\mu\delta(1-\mu\Delta t)^n\sqrt{\Delta t}}{\sigma}\right]. \end{aligned}$$

  Note that $(1-\mu\Delta t)^n \to e^{-\mu T}$ as $\Delta t \to 0$. Then, for Δt sufficiently small,

$$\text{(B.15)} \qquad \tfrac{1}{2}e^{-\mu T} < (1-\mu\Delta t)^n.$$

  Hence, for Δt sufficiently small,

$$f_u(u, h_{i+1}) \leq \phi(B_u)\left[\frac{\delta}{\sigma\sqrt{\Delta t}}\exp\left(\frac{-\delta^2}{2\sigma^2\Delta t}\right) + \mu\sigma\Delta t - \frac{\mu\delta e^{-\mu T}\sqrt{\Delta t}}{2\sigma}\right] < 0.$$

  On the other hand, suppose that $u \in [\delta, \delta + \sqrt{\Delta t}]$. Then, from (B.14), (B.15), and since

$0 < B_u < A_u$, we have for Δt sufficiently small,

$$\begin{aligned}
f_u(u, h_{i+1}) &\leq \frac{\delta}{\sigma\sqrt{\Delta t}}\phi(A_u) + \frac{\mu\Delta t}{2} - \frac{\mu\delta(1-\mu\Delta t)^n\sqrt{\Delta t}}{\sigma}\phi(B_u) \\
&\leq \frac{\delta}{\sigma\sqrt{\Delta t}}\phi(A_u) + \frac{\mu\Delta t}{2} - \frac{\mu\delta e^{-\mu T}\sqrt{\Delta t}}{2\sigma}\phi(B_u) \\
&\leq \frac{\delta}{\sigma\sqrt{2\pi\Delta t}}\exp\left(-\frac{\delta^2}{2\sigma^2\Delta t}\right) + \frac{\mu\Delta t}{2} - \frac{\mu\delta e^{-\mu T}\sqrt{\Delta t}}{2\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}\right) < 0.
\end{aligned}$$

The above discussion, combined with (B.8) and the fact that $f(\cdot, h_{i+1})$ is continuous, implies that

$$\sup_{u\in[\delta,\infty)} f(u, h_{i+1}) \leq f(\delta, h_{i+1}) < f(u_i^*, h_{i+1}). \tag{B.16}$$

■

C. Proof of Theorem 3

We will establish Theorem 3 via a sequence of lemmas. First, recall the function $f(u, h)$ defined in (B.2) and the quantities $\hat{u}_L$ and $\hat{u}_R$ defined in (B.5).

Lemma 5. *(i) As $\Delta t \to 0$,*

$$\max_{\substack{u\in[\hat{u}_L,\hat{u}_R] \\ 0\leq i<n-1}} |f_{uu}(u, h_{i+1})| = O\left(\sqrt{\Delta t \log\frac{1}{\Delta t}}\right).$$

(ii) For all $h \in \mathbb{R}$ and Δt sufficiently small,

$$0 \leq f_h(\hat{u}_R, h) \leq 1.$$

Proof. We begin with (i). Recall A_u and B_u from (B.3). Let u be in the interval $[\hat{u}_L, \hat{u}_R]$. Then, for $0 \leq i < n-1$, from (B.7),

$$\begin{aligned}
|f_{uu}(u, h_{i+1})| &\leq \left|\phi(A_u)\left[\frac{\mu\sqrt{\Delta t}}{\sigma} - \frac{u(1-\mu\Delta t)h_{i+1}}{\sigma^3\Delta t^{3/2}}\right] + \phi(B_u)\left[\frac{\mu(\delta-u)}{\sigma^3\sqrt{\Delta t}}(h_{i+1}-\delta) - \frac{\mu\sqrt{\Delta t}}{\sigma}\right]\right| \\
&\leq \phi(A_u)\frac{\mu\sqrt{\Delta t}}{\sigma} + \phi(A_u)\frac{\delta^2}{\sigma^3\Delta t^{3/2}} + \phi\left(\frac{\delta-u}{\sigma\sqrt{\Delta t}}\right)\left[\frac{\delta\mu(\delta-u)}{\sigma^3\sqrt{\Delta t}} + \frac{\mu\sqrt{\Delta t}}{\sigma}\right].
\end{aligned}$$

Here, we have used the fact that $0 \leq u \leq \delta$ and $0 \leq h_{i+1} < \delta$ (cf. Lemma 4). Note that, for Δt

sufficiently small, $\hat{u}_L \geq \delta/2$. Then,

$$\max_{u\in[\hat{u}_L,\hat{u}_R]} \phi(A_u) \leq \phi(A_{\delta/2}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\delta^2}{8\sigma^2\Delta t}\right) \leq c_0\Delta t^2,$$

for an appropriately chosen constant c_0. Thus,

$$\begin{aligned}
|f_{uu}(u,h_{i+1})| &\leq \frac{c_0\mu}{\sigma}\Delta t^{5/4} + \frac{c_0\delta^2}{\sigma^3}\sqrt{\Delta t} + \phi\left(\frac{\delta-u}{\sigma\sqrt{\Delta t}}\right)\left[\frac{\delta\mu(\delta-u)}{\sigma^3\sqrt{\Delta t}} + \frac{\mu\sqrt{\Delta t}}{\sigma}\right] \\
&\leq \frac{c_0\mu}{\sigma}\Delta t^{5/4} + \frac{c_0\delta^2}{\sigma^3}\sqrt{\Delta t} + \phi\left(\frac{\delta-\hat{u}_R}{\sigma\sqrt{\Delta t}}\right)\left[\frac{\delta\mu(\delta-\hat{u}_L)}{\sigma^3\sqrt{\Delta t}} + \frac{\mu\sqrt{\Delta t}}{\sigma}\right] \\
&= \frac{c_0\mu}{\sigma}\Delta t^{5/4} + \frac{c_0\delta^2}{\sigma^3}\sqrt{\Delta t} + \sqrt{\frac{\Delta t}{2\pi R(\Delta t)}}\left[\frac{\delta\mu}{\sigma^2}\sqrt{\log\frac{\alpha L}{\Delta t}} + \frac{\mu\sqrt{\Delta t}}{\sigma}\right] \\
&\leq \frac{c_0\mu}{\sigma}\Delta t^{5/4} + \frac{c_0\delta^2}{\sigma^3}\sqrt{\Delta t} + \frac{\mu\Delta t}{\sigma\sqrt{2\pi R(\Delta t)}} + \frac{\delta\mu}{\sigma^2\sqrt{2\pi R(\Delta t)}}\sqrt{\Delta t\log\frac{\alpha L}{\Delta t}}.
\end{aligned}$$

Since $R(\Delta t) \to Le^{-2\mu T}$ as $\Delta t \to 0$, the last term asymptotically dominates and (i) follows.

For (ii), note that $\Phi(A_{\hat{u}_R}), \Phi(B_{\hat{u}_R}) \in (0,1)$, so if $\Delta t < 1/\mu$, then for all h,

$$f_h(\hat{u}_R, h) = (1-\mu\Delta t)\Phi(A_{\hat{u}_R}) + \mu\Delta t\Phi(B_{\hat{u}_R}) \in (0,1).$$

■

Lemma 6. *As* $\Delta t \to 0$,

$$\hat{u}_R - \hat{u}_L = O\left(\sqrt{\frac{\Delta t}{\log\frac{1}{\Delta t}}}\right).$$

Proof. Note that

$$\hat{u}_R - \hat{u}_L = \sigma\sqrt{\Delta t}\left(\sqrt{\log\frac{\alpha L}{\Delta t}} - \sqrt{\log\frac{R(\Delta t)}{\Delta t}}\right) = \sigma\sqrt{\Delta t}\left[g(\alpha L) - g(R(\Delta t))\right],$$

where $g(x) \triangleq \sqrt{\log\frac{x}{\Delta t}}$. Then, by mean value theorem, for some $z \in [R(\Delta t), \alpha L]$,

$$\hat{u}_R - \hat{u}_L = \sigma\sqrt{\Delta t}g'(z)[\alpha L - R(\Delta t)] = \frac{\sigma}{2z}[\alpha L - R(\Delta t)]\sqrt{\frac{\Delta t}{\log\frac{z}{\Delta t}}} \leq \frac{\sigma\alpha L}{2R(\Delta t)}\sqrt{\frac{\Delta t}{\log\frac{R(\Delta t)}{\Delta t}}}.$$

The result follows since $R(\Delta t) \to R(0) \triangleq Le^{-2\mu T}$ as $\Delta t \to 0$. ■

Let $\{h_i : 0 \leq i < n-1\}$ be the optimal solution to the dynamic programming recursion (10)–(11), and let $\{u_i^* : 0 \leq i < n-1\}$ define the corresponding optimal policy. Define $\{\hat{h}_i : 0 \leq$

$i \leq n-1\}$ by the recursion

$$\hat{h}_i \triangleq \begin{cases} f(\hat{u}_R, \hat{h}_{i+1}) & \text{if } 0 \leq i < n-1, \\ 0 & \text{if } i = n-1. \end{cases}$$

Note that $\hat{h}_i$ is the continuation value of the suboptimal policy that always chooses $u_i = \hat{u}_R$, for $0 \leq i < n-1$. We are interested in quantifying its difference to the optimal continuation value.

Lemma 7. *As* $\Delta t \to 0$,

$$0 \leq h_0 - \hat{h}_0 = O\left(\sqrt{\frac{\Delta t}{\log \frac{1}{\Delta t}}}\right).$$

Proof. For $0 \leq i < n-1$, define $\Delta_i \triangleq h_i - \hat{h}_i$. Clearly, $\Delta_i \geq 0$.

Using the mean value theorem,

$$\begin{aligned} \Delta_i &= f(u_i^*, h_{i+1}) - f(\hat{u}_R, \hat{h}_{i+1}) \\ &= \Big[f(u_i^*, h_{i+1}) - f(\hat{u}_R, h_{i+1})\Big] + \Big[f(\hat{u}_R, h_{i+1}) - f(\hat{u}_R, \hat{h}_{i+1})\Big] \\ &= -\tfrac{1}{2} f_{uu}(\bar{u}, h_{i+1})(\hat{u}_R - u_i^*)^2 + f_h(\hat{u}_R, \bar{h})\Delta_{i+1}. \end{aligned}$$

where $\bar{u}$ is some point on the interval $(u_i^*, \hat{u}_R)$ and $\bar{h}$ is some point on the interval $(\hat{h}_{i+1}, h_{i+1})$. Here, we have used the fact that the optimal solution u_i^* satisfies the first order condition $f_u(u_i^*, h_{i+1}) = 0$.

Using Lemmas 5 and 6, for Δt sufficiently small, there exist constants c_1 and c_2 so that

$$\max_{\substack{u \in [\hat{u}_L, \hat{u}_R] \\ 0 \leq i < n-1}} |f_{uu}(u, h_{i+1})| \leq c_1 \sqrt{\Delta t \log \frac{1}{\Delta t}}, \qquad \hat{u}_R - \hat{u}_L \leq c_2 \sqrt{\frac{\Delta t}{\log \frac{1}{\Delta t}}}.$$

Also, from Lemma 5, note that $0 \leq f_h(\hat{u}_R, \bar{h}) \leq 1$. Then, we obtain that, for Δt sufficiently small,

$$\Delta_i \leq \frac{c_1 (\hat{u}_R - u_i^*)^2}{2} \sqrt{\Delta t \log \frac{1}{\Delta t}} + \Delta_{i+1} \leq \frac{c_1 c_2}{2} \frac{\Delta t^{3/2}}{\sqrt{\log \frac{1}{\Delta t}}} + \Delta_{i+1}.$$

Then, since $\Delta_{n-1} = 0$, we have that

$$\Delta_0 \leq \left(\frac{T}{\Delta t}\right) \frac{c_1 c_2}{2} \frac{\Delta t^{3/2}}{\sqrt{\log \frac{1}{\Delta t}}} = \frac{c_1 c_2 T}{2} \sqrt{\frac{\Delta t}{\log \frac{1}{\Delta t}}}.$$

■

Define the sequence $\{\hat{\beta}_i \ : \ 0 \leq i \leq n-1\}$ by the linear recursion

$$\hat{\beta}_i \triangleq \begin{cases} \mu \Delta t (\hat{u}_R - \hat{\beta}_{i+1}) + \hat{\beta}_{i+1} & \text{if } 0 \leq i < n-1, \\ 0 & \text{if } i = n-1. \end{cases} \tag{C.1}$$

Here, $\hat{\beta}_i$ is an approximation to the value $\hat{h}_i$. The next lemma bounds the approximation error.

Lemma 8. *As $\Delta t \to 0$,*

$$|\hat{h}_0 - \hat{\beta}_0| = O\left(\sqrt{\frac{\Delta t}{\log \frac{1}{\Delta t}}}\right).$$

Proof. For $0 \leq i < n-1$, define $\epsilon_i \triangleq \hat{h}_i - \hat{\beta}_i$. Recall the following definition from the proof of Theorem 2,

$$A_{\hat{u}_R} \triangleq \frac{\hat{u}_R}{\sigma\sqrt{\Delta t}}, \qquad B_{\hat{u}_R} \triangleq \frac{\hat{u}_R - \delta}{\sigma\sqrt{\Delta t}} = -\sqrt{\log \frac{R(\Delta t)}{\Delta t}}.$$

Then, by the recursive definitions of $\hat{h}_i$ and $\hat{\beta}_i$, $0 \leq i < n-1$,

$$\begin{aligned} \epsilon_i = (1-\mu\Delta t)\epsilon_{i+1} - \mu\Delta t\hat{u}_R\big[1 - \Phi(A_{\hat{u}_R}) + \Phi(B_{\hat{u}_R})\big] + \mu\sigma\Delta t^{3/2}\big[\phi(A_{\hat{u}_R}) - \phi(B_{\hat{u}_R})\big] \\ - (1-\mu\Delta t)\hat{h}_{i+1}\left[1 - \Phi(A_{\hat{u}_R}) - \frac{\mu\Delta t}{1-\mu\Delta t}\Phi(B_{\hat{u}_R})\right]. \end{aligned}$$

Since $\hat{\mu}_R$ is not the optimal policy, we have $\hat{h}_{i+1} \leq h_{i+1} < \delta$ (cf. Lemma 4). Further, for Δt sufficiently small, $0 < \phi(A_{\hat{u}_R}) \leq \phi(B_{\hat{u}_R})$. This implies that

$$\begin{aligned} |\epsilon_i| \leq (1-\mu\Delta t)|\epsilon_{i+1}| + \delta\mu\Delta t\big[1 - \Phi(A_{\hat{u}_R}) + \Phi(B_{\hat{u}_R})\big] + \mu\sigma\Delta t^{3/2}\phi(B_{\hat{u}_R}) \\ + \delta\left[1 - \Phi(A_{\hat{u}_R}) + \frac{\mu\Delta t}{1-\mu\Delta t}\Phi(B_{\hat{u}_R})\right]. \end{aligned}$$

Note that, except for the first term, there is no dependence on i in the right side of this equality. Then, we can define

$$\begin{aligned} C(\Delta t) \triangleq \delta\mu\Delta t\big[1 - \Phi(A_{\hat{u}_R}) + \Phi(B_{\hat{u}_R})\big] + \mu\sigma\Delta t^{3/2}\phi(B_{\hat{u}_R}) \\ + \delta\left[1 - \Phi(A_{\hat{u}_R}) + \frac{\mu\Delta t}{1-\mu\Delta t}\Phi(B_{\hat{u}_R})\right], \end{aligned}$$

and we have that

$$|\epsilon_i| \leq (1-\mu\Delta t)|\epsilon_{i+1}| + C(\Delta t).$$

Since $\epsilon_{n-1} = 0$, it is easy to verify by backward induction on i that

$$|\epsilon_i| \leq \frac{1 - (1-\mu\Delta t)^{n-i-1}}{\mu\Delta t}\, C(\Delta t).$$

Therefore,

$$
\begin{aligned}
|\epsilon_0| &\leq \frac{1-(1-\mu\Delta t)^{n-1}}{\mu\Delta t} C(\Delta t) \leq \frac{C(\Delta t)}{\mu\Delta t} \\
&= \delta\big[1-\Phi(A_{\hat{u}_R}) + \Phi(B_{\hat{u}_R})\big] + \mu\sigma\sqrt{\Delta t}\phi(B_{\hat{u}_R}) + \frac{\delta}{\mu\Delta t}\left[1-\Phi(A_{\hat{u}_R}) + \frac{\mu\Delta t}{1-\mu\Delta t}\Phi(B_{\hat{u}_R})\right] \\
&= \delta\big[Q(A_{\hat{u}_R}) + Q(-B_{\hat{u}_R})\big] + \mu\sigma\sqrt{\Delta t}\phi(B_{\hat{u}_R}) + \frac{\delta}{\mu\Delta t}\left[Q(A_{\hat{u}_R}) + \frac{\mu\Delta t}{1-\mu\Delta t}Q(-B_{\hat{u}_R})\right].
\end{aligned} \tag{C.2}
$$

From (B.10), however,

$$
Q(A_{\hat{u}_R}) \leq \frac{\sigma}{\hat{u}_R}\sqrt{\frac{\Delta t}{2\pi}} \exp\left(-\frac{\hat{u}_R^2}{2\sigma\Delta t}\right).
$$

Since $\hat{u}_R \to \delta$ as $\Delta t \to 0$, for Δt sufficiently small, there exists constants a_1 and a_2, with $0 < a_2 < \delta^2/2\sigma$, so that

$$
Q(A_{\hat{u}_R}) \leq a_1\sqrt{\Delta t} \exp\left(-\frac{a_2}{\Delta t}\right).
$$

Also by (B.10),

$$
Q(-B_{\hat{u}_R}) \leq \sqrt{\frac{\Delta t}{2\pi R(\Delta t)\log\frac{R(\Delta t)}{\Delta t}}}.
$$

Since and $R(\Delta t) \to R(0) \triangleq Le^{-2\mu T}$ as $\Delta t \to 0$, for Δt sufficiently small, there exists a constant a_3 so that

$$
Q(-B_{\hat{u}_R}) \leq a_3\sqrt{\frac{\Delta t}{\log\frac{1}{\Delta t}}}.
$$

Finally,

$$
\phi(B_{\hat{u}_R}) = \sqrt{\frac{\Delta t}{2\pi R(\Delta t)}},
$$

so for Δt sufficiently small, there exists a constant a_4 with

$$
\phi(B_{\hat{u}_R}) \leq a_4\sqrt{\Delta t}.
$$

Applying these bounds to (C.2), the result follows. ■

Lemma 9. *As $\Delta t \to 0$,*

$$
\hat{\beta}_0 = \hat{u}_R\left(1-e^{-\mu T}\right) + O(\Delta t).
$$

Proof. Note that the recurrence (C.1) can be explicitly solved to obtain

$$
\hat{\beta}_0 = \sum_{i=0}^{n-2}(1-\mu\Delta t)^i \mu\Delta t\, \hat{u}_R = \hat{u}_R\left(1-(1-\mu\Delta t)^{n-1}\right) = \hat{u}_R\left(1-(1-\mu\Delta t)^{T/\Delta t-1}\right).
$$

The result follows since $(1-\mu\Delta t)^{T/\Delta t} = e^{-\mu T} + O(\Delta t)$ as $\Delta t \to 0$. ■

We are now ready to prove Theorem 3.

Theorem 3. *As* $\Delta t \to 0$,

$$h_0(\Delta t) = \bar{h}_0 \left(1 - \frac{\sigma}{\delta}\sqrt{\Delta t \log \frac{\delta^2}{2\pi\sigma^2\Delta t}}\right) + o\left(\sqrt{\Delta t}\right),$$

where

$$\bar{h}_0 = \delta\left(1 - e^{-\mu T}\right)$$

is the optimal value for the stylized model without latency, i.e., the value defined by (5).

Proof. First, define

$$\hat{\gamma}_0 \triangleq \left(1 - e^{-\mu T}\right)\left(\delta - \sigma\sqrt{\Delta t \log \frac{\delta^2}{2\pi\sigma^2\Delta t}}\right).$$

Then,

$$\text{(C.3)} \qquad \left| h_0 - \bar{h}_0 \left(1 - \frac{\sigma}{\delta}\sqrt{\Delta t \log \frac{\delta^2}{2\pi\sigma^2\Delta t}}\right)\right| = |h_0 - \hat{\gamma}_0| \\ \leq |h_0 - \hat{h}_0| + |\hat{h}_0 - \hat{\beta}_0| + |\hat{\beta}_0 - \hat{\gamma}_0|.$$

We will bound each of the terms in the right side of (C.3). First, by Lemma 7,

$$\text{(C.4)} \qquad |h_0 - \hat{h}_0| = O\left(\sqrt{\frac{\Delta t}{\log \frac{1}{\Delta t}}}\right).$$

Next, by Lemma 8,

$$\text{(C.5)} \qquad |\hat{h}_0 - \hat{\beta}_0| = O\left(\sqrt{\frac{\Delta t}{\log \frac{1}{\Delta t}}}\right).$$

Finally, by Lemma 9, for Δt sufficiently small, there exists a constant c_1 so that

$$\begin{aligned} |\hat{\beta}_0 - \hat{\gamma}_0| &\leq \sigma\left(1 - e^{-\mu T}\right)\left(\hat{u}_R - \delta + \sqrt{\Delta t \log \frac{L}{\Delta t}}\right) + c_1\Delta t \\ &\leq \sigma\left(1 - e^{-\mu T}\right)\left(\hat{u}_R - \delta + \sqrt{\Delta t \log \frac{\alpha L}{\Delta t}}\right) + c_1\Delta t, \\ &= \sigma\left(1 - e^{-\mu T}\right)(\hat{u}_R - \hat{u}_L) + c_1\Delta t, \end{aligned}$$

where $\alpha > 1$ and L are defined by Theorem 2. Applying Lemma 6, we have that

$$|\hat{\beta}_0 - \hat{\gamma}_0| = O\left(\sqrt{\frac{\Delta t}{\log \frac{1}{\Delta t}}}\right). \tag{C.6}$$

By applying (C.4)–(C.6) to (C.3), we have that

$$\left| h_0 - \bar{h}_0 \left(1 - \frac{\sigma}{\delta}\sqrt{\Delta t \log \frac{\delta^2}{2\pi\sigma^2 \Delta t}}\right)\right| = O\left(\sqrt{\frac{\Delta t}{\log \frac{1}{\Delta t}}}\right),$$

which implies the desired result. ■

D. Price Dynamics with Jumps

At a high level, our goal has been to understand and build intuition as to the impact of a latency friction introduced by the lack of contemporaneous information. The spirit of our model it to consider an investor who wants to trade, but at a price that depends on an informational process that evolves stochastically and must be monitored continuously. While we have principally interpreted the informational process to be the bid price process, our model can alternatively be interpreted (as discussed in Section 2.1) in terms of a fundamental value process.

Thus far, we have employed a diffusive model to describe informational innovations over a short time horizon. There is significant empirical evidence that this is insufficient, particularly when modeling price processes, and that it is important to also allow for the instantaneous arrival of information, i.e., jumps. For example, Barndorff-Nielsen et al. (2010) propose the following compound Poisson process for high frequency price dynamics:

$$S_t = S_0 + \sum_{i=1}^{M_t} Y_i,$$

where N_t is a Poisson process counting the number of trades up to time t and Y_i is the potential jump movement at the ith trade time, having a distribution G.

On a short time horizon, innovations to fundamental value can be both instantaneous or diffusive.[2] In a recent empirical study, Aït-Sahalia and Jacod (2010) construct two formal statistical tests to deduce whether there is a need for a Brownian motion in modeling high-frequency data. Using individual high-frequency stock data, they conclude that both tests suggest the necessity of including a continuous component driven by Brownian motion.

Motivated by these studies, we will generalize the price dynamics of Section 2 by including both

[2] As an example, note that an instantaneous innovation may result from a news event. On the other hand, the value of a stock will have a component that is driven by the market factor, i.e., an average of returns across all stocks. Innovations to the market factor can have a diffusive component even if all individual stock prices are discrete, by virtue of cross-sectional averaging.

a continuous component (Brownian motion) and a jump component (governed by a compound Poisson process). In particular, consider a price process that evolves according to

$$S_t = S_0 + \sigma B_t + \sum_{i=1}^{M_t} Y_i, \tag{D.1}$$

where the process $(B_t)_{t\in[0,T]}$ is a standard Brownian motion, $\sigma > 0$ is an (additive) volatility parameter, and $(M_t)_{t\in[0,T]}$ is a Poisson process with intensity λ. For now, we will further assume that each jump Y_i has an i.i.d. Gaussian distribution with zero mean and variance ν^2 — we revisit the assumption of Gaussian jump sizes at the end of this section.

In the context of the latency model of Section 3, we define the price increment $X_{i+1} \triangleq S_{T_{i+1}} - S_{T_i}$ by the discrete time analog of (D.1),

$$X_{i+1} \sim \begin{cases} N(0, \sigma^2\Delta t) & \text{with probability } (1-\lambda\Delta t), \\ N(0, \sigma^2\Delta t + \nu^2) & \text{with probability } \lambda\Delta t. \end{cases} \tag{D.2}$$

With this definition, the dynamic programming decomposition outlined in Lemma 3 holds exactly as before. Incorporating jumps, we then obtain the following analog of Theorem 1, that expresses dynamic programming equations (A.4)–(A.5) in terms of the continuation values $\{h_i\}$. The proof of this theorem follows steps identical to the proof of Theorem 1, and is omitted.

Theorem 4. *Define $v^2(\Delta t) \triangleq \sigma^2\Delta t + \nu^2$. Suppose $\{h_i\}$ satisfy, for $0 \le i < n-1$,*

(D.3)

$$\begin{aligned} h_i = \max_{u_i} \ \Bigg\{ & (1-\lambda\Delta t)\Bigg(\mu\Delta t\Bigg[u_i\left(\Phi\left(\frac{u_i}{\sigma\sqrt{\Delta t}}\right) - \Phi\left(\frac{u_i-\delta}{\sigma\sqrt{\Delta t}}\right)\right) \\ & \qquad + \sigma\sqrt{\Delta t}\left(\phi\left(\frac{u_i}{\sigma\sqrt{\Delta t}}\right) - \phi\left(\frac{u_i-\delta}{\sigma\sqrt{\Delta t}}\right)\right)\Bigg] \\ & + h_{i+1}\left[(1-\mu\Delta t)\Phi\left(\frac{u_i}{\sigma\sqrt{\Delta t}}\right) + \mu\Delta t\Phi\left(\frac{u_i-\delta}{\sigma\sqrt{\Delta t}}\right)\right]\Bigg) \\ & + \lambda\Delta t\Bigg(\mu\Delta t\left[u_i\left(\Phi\left(\frac{u_i}{v(\Delta t)}\right) - \Phi\left(\frac{u_i-\delta}{v(\Delta t)}\right)\right) + v(\Delta t)\left(\phi\left(\frac{u_i}{v(\Delta t)}\right) - \phi\left(\frac{u_i-\delta}{v(\Delta t)}\right)\right)\right] \\ & + h_{i+1}\left[(1-\mu\Delta t)\Phi\left(\frac{u_i}{v(\Delta t)}\right) + \mu\Delta t\Phi\left(\frac{u_i-\delta}{v(\Delta t)}\right)\right]\Bigg)\Bigg\}, \end{aligned}$$

and

$$h_{n-1} = 0. \tag{D.4}$$

Suppose further that, for $0 \le i < n-1$, u_i^* is a maximizer of (D.3)*. Then, a policy which* *chooses limit prices according to the premia defined by* $\{u_i^*\}$, *i.e.,*

$$\ell_i^* = S_{T_i} + u_i^*, \quad \forall\ 0 \le i < n-1,$$

is optimal.

The following theorem provides an analog of Theorem 2 that characterizes the optimal solution for the dynamic programming equation in the low latency regime, with the presence of jumps. The proof is similar to that of Theorem 2, and is again omitted.

Theorem 5. *Fix* $\alpha > 1$ *and define*

$$\kappa \triangleq 1 + \frac{\lambda\delta}{\nu\mu\sqrt{2\pi}}.$$

If Δt *is sufficiently small, then there exists a unique optimal solution* $\{h_i\}$ *to the dynamic programming equations* (D.3)–(D.4). *Moreover, the corresponding optimal policy* $\{u_i^*\}$ *is unique. For* $0 \le i < n-1$, *this strategy chooses limit prices in the range*

$$\ell_i^* \in \left(S_i + \delta - \sigma\sqrt{\Delta t \log \frac{\alpha L}{\Delta t}},\ S_i + \delta - \sigma\sqrt{\Delta t \log \frac{R(\Delta t)}{\Delta t}} \right),$$

where

$$L \triangleq \frac{\delta^2}{2\pi\sigma^2}, \qquad\qquad R(\Delta t) \triangleq \frac{\delta^2(1-\mu\Delta t)^{2n}}{2\pi\sigma^2\kappa^2}.$$

Note that, when compared to Theorem 2, the addition of jump component in Theorem 5 causes $R(\Delta t)$ to decrease by a constant multiple. Thus, the range containing the optimal solution is gets larger. However, the upper bound of the range is of the same order asymptotically (as $\Delta t \to 0$) as before. Hence, we can again provide a asymptotic closed-form expression for $h_0(\Delta t)$, as is done by the following theorem, which is an analog of Theorem 3 and Corollary 1. (As before, we omit the proof.)

Theorem 6. *As* $\Delta t \to 0$,

$$h_0(\Delta t) = \bar{h}_0 \left(1 - \frac{\sigma}{\delta}\sqrt{\Delta t \log \frac{\delta^2}{2\pi\sigma^2\Delta t}} \right) + o\left(\sqrt{\Delta t}\right),$$

where

$$\bar{h}_0 \triangleq \frac{\delta\mu}{\mu + \lambda p}\left(1 - e^{-(\mu+\lambda p)T}\right), \tag{D.5}$$

is the zero latency limit of $h_0(\Delta t)$, and

$$p \triangleq 1 - \Phi\left(\frac{\delta}{\nu}\right),$$

the probability of a jump size greater than δ.

Furthermore, latency cost is unchanged with the introduction of the jump components in the bid price dynamics, i.e., as $\Delta t \to 0$,

$$\mathsf{LC}(\Delta t) = \frac{\sigma\sqrt{\Delta t}}{\delta}\sqrt{\log\frac{\delta^2}{2\pi\sigma^2\Delta t}} + o\left(\sqrt{\Delta t}\right).$$

Our analysis with the jump-diffusion model can be interpreted as follows. Theorem 6 illustrates that, when there is a jump component (i.e., $\lambda > 0$), the zero latency limit $\bar{h}_0$ has a lower value than in the absence of jumps, (i.e., $\lambda = 0$), all else being equal. In other words, the presence of jumps is detrimental even in the absence of latency. To see why, note that jumps are zero mean innovations in the price process. In our model, an investor only earns excess value by waiting for an impatient buyer. Jumps may cause the bid price to cross the investor's limit order price and execute his share without giving him the chance to revise his order. Thus, jumps reduce the probability of trading with an impatient buyer.

This intuition can be made precise by interpreting the zero latency limit in (D.5). Observe that $\mu + \lambda p$ is the combined arrival rate of impatient buyers asking for an immediate execution, or positive jumps in the price of the stock that are larger than the bid-offer spread and would result in trade execution. The quantity

$$\frac{\mu}{\mu + \lambda p}\left(1 - e^{-(\mu+\lambda p)T}\right)$$

is the probability that there at least one such arrival, and that the first such arrival is that of an impatient buyer. In this case, the trader earns a relative spread of δ. In all other cases (i.e., no arrivals, or the case where the first arrival is a large positive jump), the trade occurs at the bid price and the trader earns no spread. These two cases yield the expression for $\bar{h}_0$.

Now, comparing with our earlier results, jumps also negatively impact the investor in the presence of latency, for similar reasons as in the zero latency case. However, when measured *relative* to the zero latency case, i.e., in term of latency cost, jumps create no additional impact. That is, the latency cost expressions in Theorem 6 and Corollary 1 are identical. Intuitively, in our model, jumps are instantaneous, and the investor cannot react to them even in the absence of latency. Hence, latency cost, measured relatively, only depends on the diffusive innovations.

Note that we have thus far assumed Gaussian jump sizes. In Theorem 6, the only place that this distribution or its parameter ν arises explicitly is the quantity p. This is the probability that the jump will be larger than the prevailing bid-offer spread, δ, and hence will cross with the limit order places by the investor. This leads us to conjecture (without proof) the result in the non-Gaussian case: if the jump size Y_i in (D.1) is an i.i.d. zero mean random variable that has a cumulative distribution function G, then Theorem 6 holds with $p \triangleq 1 - G(\delta)$.