Latency in Algorithmic Trading

Pankaj Kumar
Perm State University,Russia

Perm Winter School, 2012

Outline

Outline

Algorithmic Arms Race

- In any gun fight, it’s not enough just to shoot fast or to shoot straight. Survival depends on being able to do both. And a single shot isn’t always enough either – you also need to be quick to load and fire again.
- Similarly, the financial market’s lone gun-slinger of the open-outcry trading floors are almost being replaced by ultra-fast, computerised trading systems which are more akin to robots with machine guns.

Algorithmic Arms Race

- In any gun fight, it's not enough just to shoot fast or to shoot straight. Survival depends on being able to do both. And a single shot isn't always enough either – you also need to be quick to load and fire again.
- Similarly, the financial market's lone gun-slinger of the open-outcry trading floors are almost being replaced by ultra-fast, computerised trading systems which are more akin to robots with machine guns.
- Survival rules:
  - **Shoot straight** – the ability to define a trading strategy that adapts dynamically not only to changes in the market, but also to the impact of other firm's trading strategies.
  - **Shoot fast** – the ability to reduce latency (the time it takes to react to changes in the market and execute a trade) to an absolute minimum.
  - **Shoot often** – the ability to process massive volumes of trades.
- Indeed the need for speed is now so great, that there is buzz about latency arbitrage.

Outline

Latency: Drivers

Definition

the processing delay measured from the entry of the order (at the vendor‘s computer) to the transmission of an acknowledgement (from the vendor‘s computer).

Speed is important because:

1. The inherent fundamental volatility of financial securities means that rebalancing positions faster could result in higher utility.
2. Irrespective of the absolute speed, being faster than other traders can create profit opportunities by enabling a prompt response to news or market-generated events.
3. Competition between exchanges driven by a significant demand amongst a class of investors, sometimes called "high frequency" traders, for low latency trade execution.
4. 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”.

[1] “Wall Street’s quest to process data at the speed of light,” Information Week, April 21, 2007.

Latency: Drivers

Definition

the processing delay measured from the entry of the order (at the vendor‘s computer) to the transmission of an acknowledgement (from the vendor‘s computer).

Speed is important because:

1. The inherent fundamental volatility of financial securities means that rebalancing positions faster could result in higher utility.
2. Irrespective of the absolute speed, being faster than other traders can create profit opportunities by enabling a prompt response to news or market-generated events.
3. Competition between exchanges driven by a significant demand amongst a class of investors, sometimes called "high frequency" traders, for low latency trade execution.
4. 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”.
5. 1 millisecond advantage can be worth $ 100 million to a major brokerage firm.[1]

[1] “Wall Street’s quest to process data at the speed of light,” Information Week, April 21, 2007.

Latency Buzz

There have been much discussions on importance of latency among various market participants, regulators, and academics.

- 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?
- Collective marketplace is better or worse off given lower latency?
- Optimal latency (as latency cannot be expressed per se but must be compared to the latency of competitors) reduces the implementation risk or not?
- The potential impact of co-location on the price formations mechanism.
- Is co-location is suitable for market efficiency?

Associated Cost

The precise trading strategy defines cost that a trader bears due to latency. They can be:

- A trader with significant latency will be making trading decisions based on information that is stale.
- Low relative latency.
- Time priority.
- Execution of contingent orders.

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.

Outline

Cost Quantification

- To provide a quantification of the costs of latency in trade execution. 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; Hasbrouck (2007)).
- Numerical computation by using stochastic dynamic programming to quantify costs. However, in the regime of greatest interest, where the latency is close to zero, we provide a closed-form asymptotic expression.
- Quantitative insight to behaviour of latency.

Outline

Empirical and Numerical

1. Numerical evaluation's of the optimal policy in our model, the corresponding value function, and the latency cost approximation.
2. Use realistic model parameters estimated from recent market data for a single stock

Outline

Execution Model without Latency

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.
An uninformed trader who must sell exactly one block of a stock over a time horizon $[0, T]$. At any time $t\varepsilon[0, T]$, 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:

$$S_t = S_0 + \sigma B_t \tag{1}$$

   Where, $(B_t)_{t\varepsilon[0,T]}$ is standard Brownian motion and σ is volatility.
2. The trader can choose to submit a limit order to sell, and limit price is L_t.

Limit order execution is done using following ways:

1. Impatient buyers arrives in market by Poisson process with rate μ, $(N_t)_{t\varepsilon[0,T)}$ the cumulative arrival process for impatient buyers.
   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.
   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:

$$\lambda(u_t) \triangleq \mu(1 - F_t(u_t)) \tag{2}$$

   where $u_t \triangleq L_t - S_t$ is the instantaneous price premium of the limit order.
2. Alternatively, a limit order will also execute at time τ if the bid price crosses the limit order price, i.e., $S_t \geq L_t$

Limit Order Execution

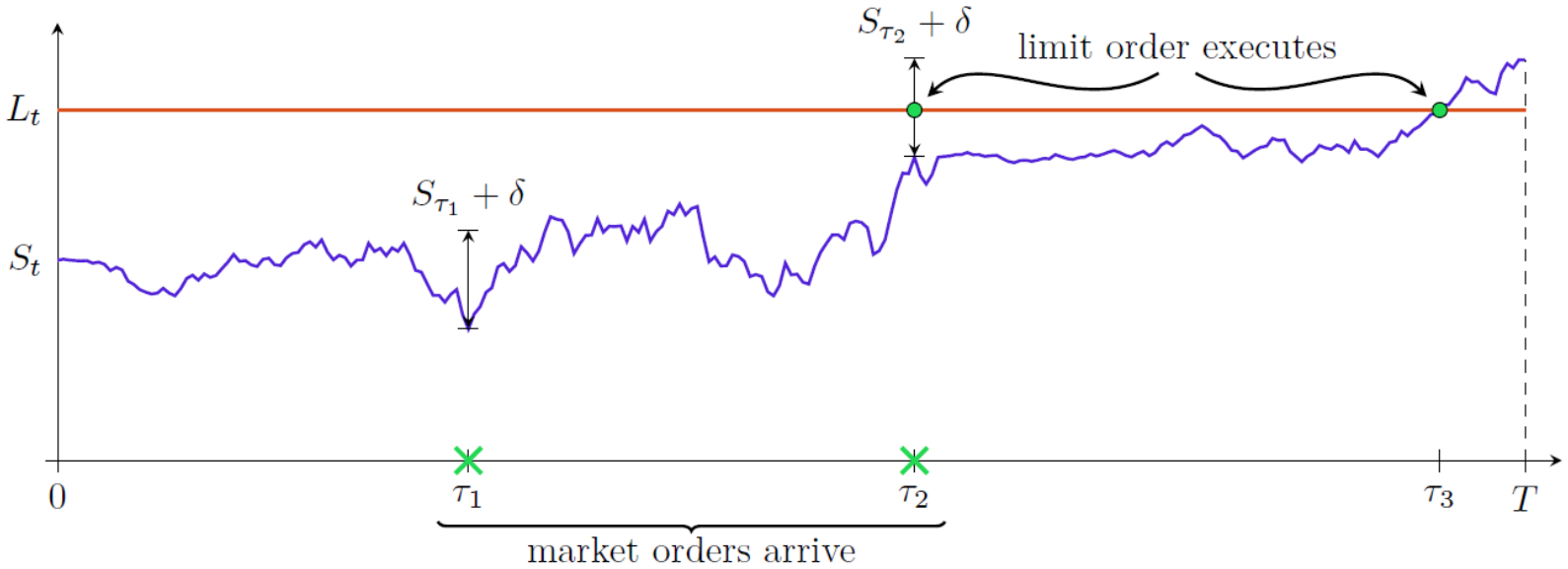


Figure: **Limit Order Execution**: An illustration of the limit order execution in the stylized model over the time horizon $[0, T]$. Here, we assume the trader leaves a limit order with the (constant) price L_t and S_t is the bid price process. If market orders arrive at times τ_1 and τ_2, the limit 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.

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

$$\bar{h_0} \triangleq \textit{maximize} \;\; E(P) - S_0. \tag{3}$$

Lemma

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

$$\bar{h_0} = \delta(1 - e^{-\mu T}).$$

Outline

Latency Model

- Trader is unable to continuously participate in the market, but faces a fixed latency $\Delta t > 0$.
- Trader is able to observe market price information with no delay. or latency

It can be well explained by following pictures:

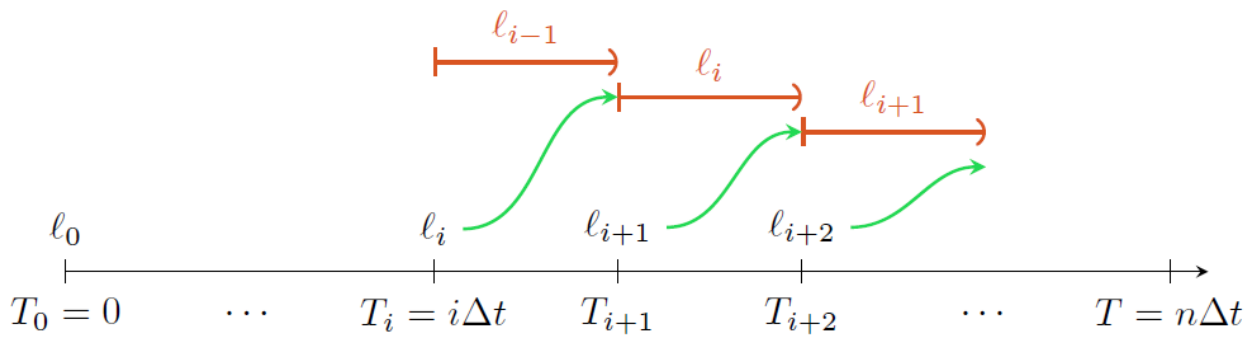


Figure: **Latency Model**:Here, the time horizon $[0, T]$ is divided into n slots, each of duration equal to the latency Δt. The limit order price l_i is decided at the start of the ith time slot, i.e., at time T_i. This price only takes effect Δt units of time later, and is active during the subsequent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

Optimal policy in presence of latency

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

$$h_0(\Delta t) \triangleq maximize_{l_0,\ldots,l_n-1}\ E[P] - S_0 \tag{4}$$

Here, the maximization is over the choice of limit order prices $(l_0, l_1, \ldots, l_n - 1)$. 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.

Outline

The standard approach to solving the optimal control problems is to use stochastic dynamic programming arguments.
Let U_i is event when trader's limit orders remains unfulfilled prior to time T_{i+1}. Also,

$$h_i \triangleq maximize_{l_i,\ldots,l_n-1}\ E[P \mid S_{Ti}, U_i] - S_{Ti}. \tag{5}$$

Suppose h_i satisfy, for $0 \leq i < n-1$,

$$h_i = \max_{u_i} \left\{ \mu\Delta t \left[u_i \left(\Phi\left(\frac{u_i}{\sigma\sqrt{\Delta t}}\right) - \Phi\left(\frac{u_i-\delta}{\sigma\sqrt{\Delta t}}\right)\right) + \sigma\sqrt{\Delta t}\left(\phi\left(\frac{u_i}{\sigma\sqrt{\Delta t}}\right) - \phi\left(\frac{u_i-\delta}{\sigma\sqrt{\Delta t}}\right)\right)\right] \right.$$
$$\left. + h_{i+1}\left[(1-\mu\Delta t)\Phi\left(\frac{u_i}{\sigma\sqrt{\Delta t}}\right) + \mu\Delta t\Phi\left(\frac{u_i-\delta}{\sigma\sqrt{\Delta t}}\right)\right]\right\},$$

and $h_{n-1} = 0$. Here, ϕ and Φ are, respectively, the p.d.f. and c.d.f. of the standard normal distribution.
Suppose futher that, for $0 \leq i < n-1$, u^* maximize the above equation. Then, a policy which chooses limiy order prices which are pegged to the bid prices according to the premia defined by $u_i{}^*$.

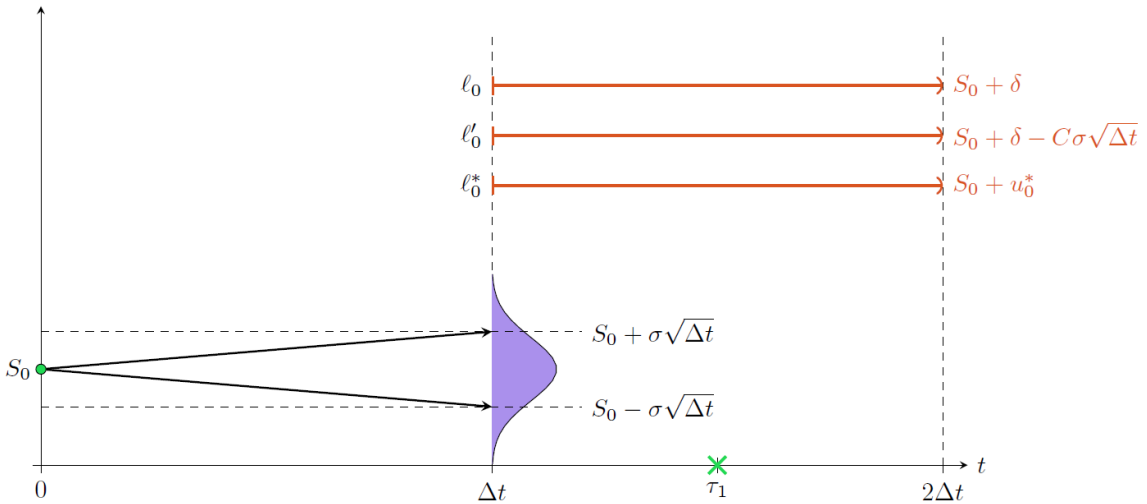
For $0 \leq i < n-1$ and corresponding to the optimal policy, this strategy chooses limit prices in the range

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

where

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

It can be represented by :



Outline

Asymptotic Analysis

The stochastic dynamic programming allows exact numerical computation of the value $h_0(\Delta t)$, the value under an optimal policy of the latency model. As, latency are now running on the time scale of milliseconds, we would be more interested in the qualitative behaviour of $h_0(\Delta t)$ in the asymptotic regime where $\Delta t \to 0$.

$$As\ \Delta t \to 0,$$

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

$$where$$

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

Asymptotically, latency cost does not depend on the length of the time horizon T or the arrival rate of the impatient traders μ. It is represented in form of:

$$As\ \Delta t \to 0,$$

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

Outline

Data for Latency Estimation

- We use NASDAQ data, TotalView-ITCH, free from (LOBSTER) are identical to those supplied to subscribers, providing real-time information about orders and executions on the NASDAQ system.
- These data are comprised of time-sequenced messages that describe the history of trade and book activity.
- Each message is time-stamped to the millisecond (i.e., one-thousand of a second), and hence these data provide a detailed picture of the trading process and the state of the NASDAQ book.
- We are able to observe four different types of messages in the TotalView-ITCH dataset:
  1. the addition of a displayed order to the book,
  2. the cancellation of a displayed order,
  3. the execution of a displayed order, and
  4. the execution of a non-displayed order.

Outline

Optimal Policy and Approximation Quality

- 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 stochastic dynamic programming) to the approximation to access the quality of our approximation.
- Our methodology here is not meant to be authoritative — there are many subtleties in the analysis of high frequency data, We use **Joel Hasbrouck and Gideon Saar, 2011** methodology used in "Low Latency Trading"
- Representative example of a liquid name, CITIBANK, on the trading day of January 4, 2010 is taken.

Optimal Policy and Approximation Quality

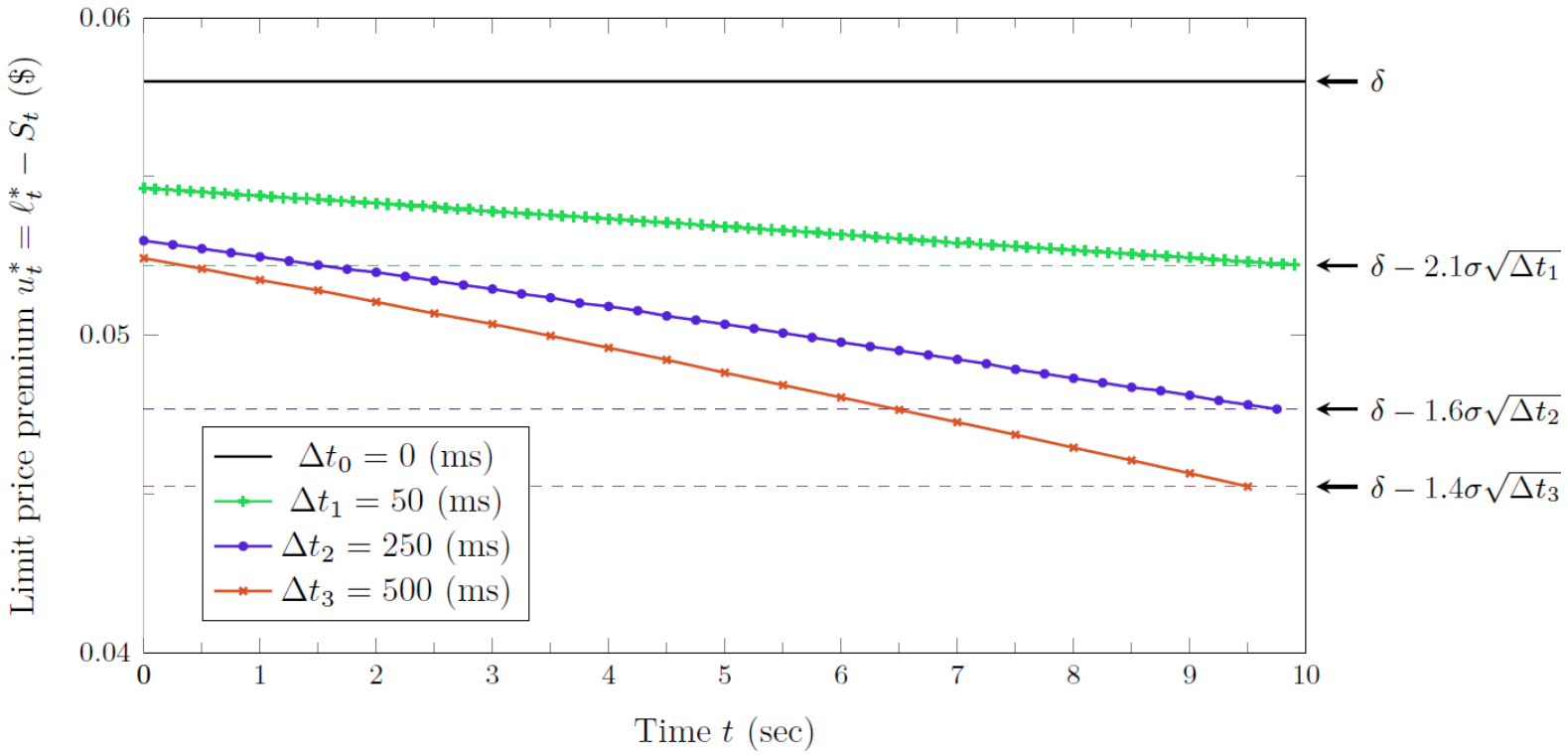
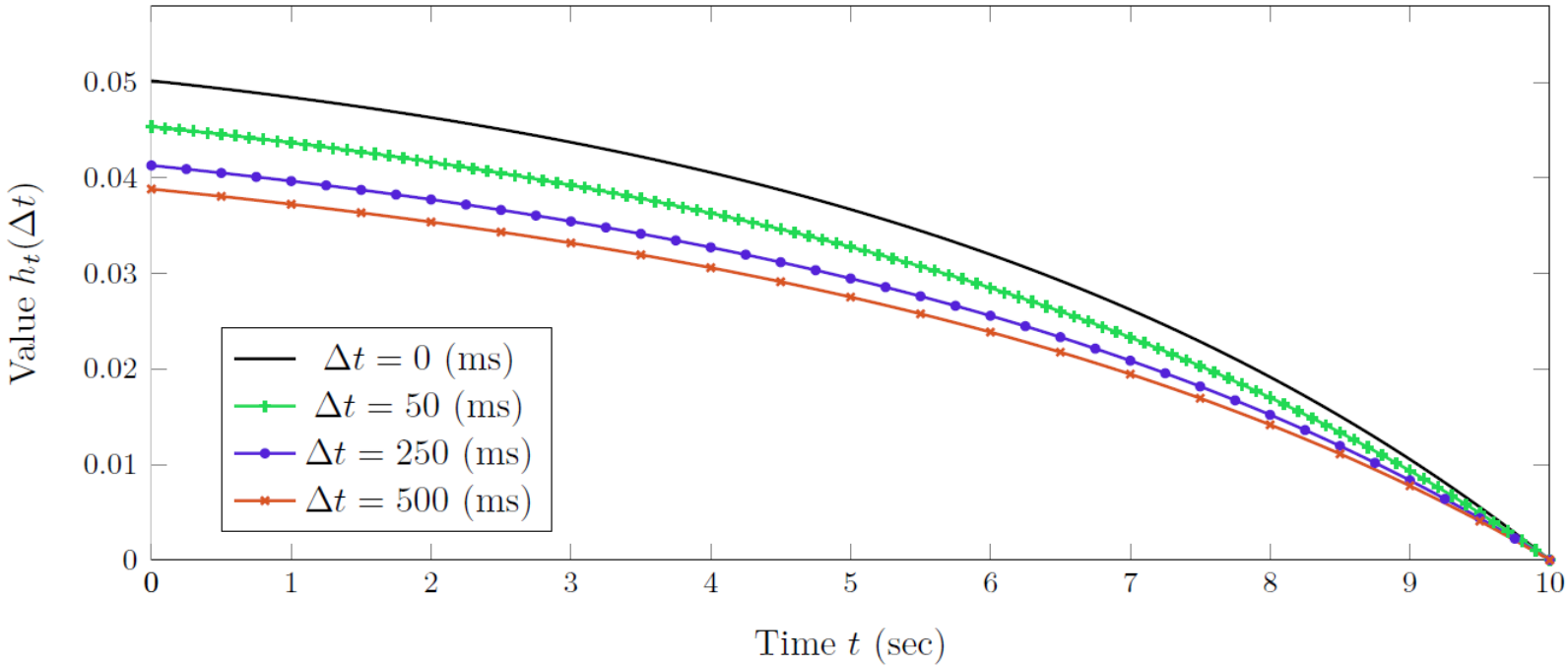


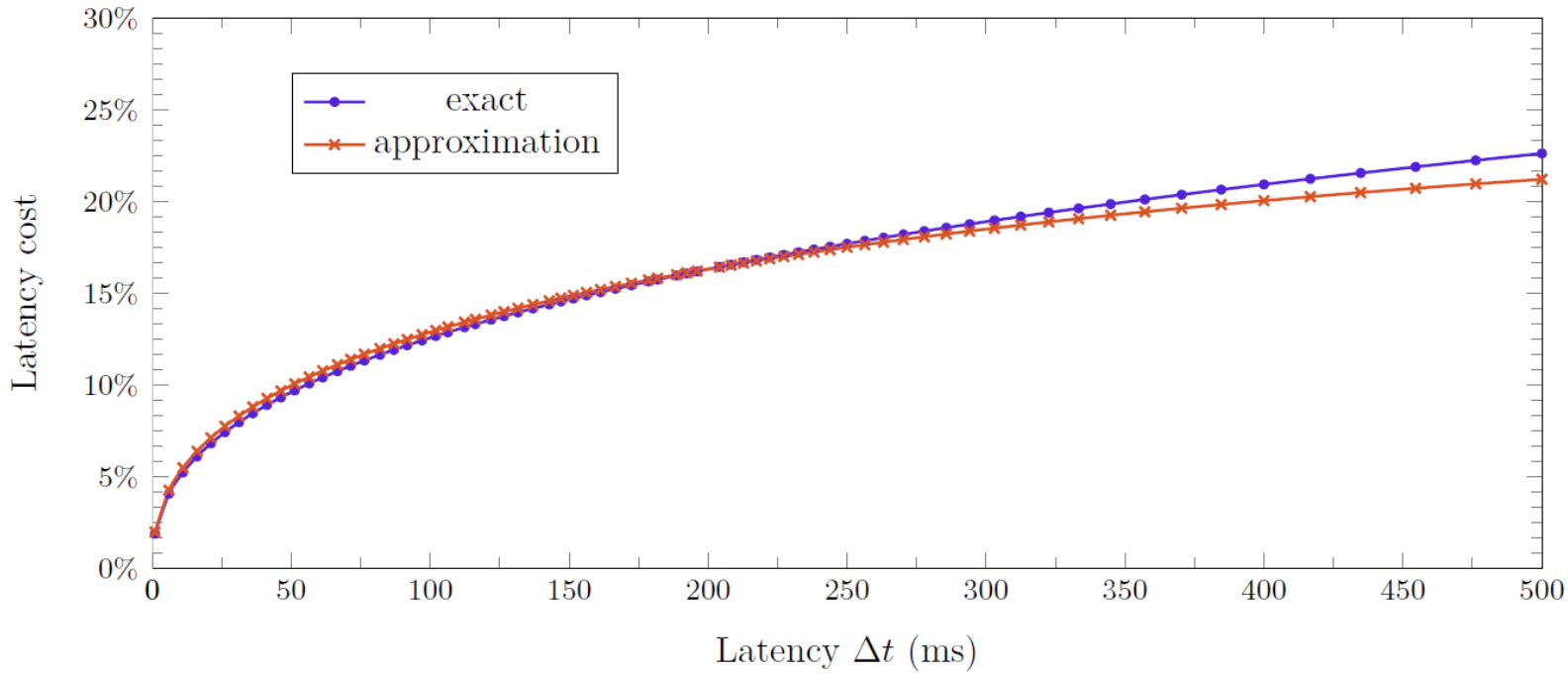
Figure: An illustration of the optimal strategy for CITIBANK, expressed in terms of limit price premium over the course of the time, for different choices of latency. In each case, the dashed line illustrates the relative distance below the bid-offer spread δ of the price premium of the final limit order,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

Optimal Policy and Approximation Quality



An illustration for the evolution of the continuation value of the optimal policy over time for CITIBANK, 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.

Optimal Policy and Approximation Quality



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.

Conclusions

- In our model, the bid price process S_t is a Brownian motion. It would be straightforward to consider other Markovian martingales, for example, allowing for non-Gaussian processes, time-inhomogeneous volatility, or for jump processes.
- The prevailing bid-offer spread is not constant, but is independent and identically distributed, varying from period to period while modelling LOB.
- 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.

Thank you!