Comments on The Cost of Latency by Moallemi and Saglam
Terrence Hendershott

Technological change has revolutionized the way financial assets are traded. Investors place orders via computer rather than speaking to a broker on the phone. Trading floors have largely been replaced by electronic trading platforms. The nature of order execution has changed dramatically as well, as virtually all large market participants now employ algorithmic trading, commonly defined as the use of computer algorithms to manage the trading process. Moallemi and Saglam study an important topic in this area: the cost of latency to investors attempting to avoid paying for immediacy by using limit orders. A limit order is a quantity to buy or sell of a security at a price at which no other investor currently wants to trade. The limit order submitter hopes that a later arriving investor will choose to execute a marketable order against the previously entered limit order.

Before algorithmic trading, investors with a large order might have hired a broker-dealer to search for a counterparty to execute the entire quantity at once in a block trade. Alternatively, that investors might have hired a broker to quietly “work” the order (possibly on the floor of the New York Stock Exchange), using his judgment and discretion to buy a little bit here and there over the course of the trading day to keep from driving the IBM share price up too far. As trading became more electronic, it became easier and cheaper to replicate the human trader with a computer program doing algorithmic trading. Now virtually every large broker-dealer offers a suite of algorithms to its institutional customers to help them execute orders in a single stock, in pairs of stocks, or in baskets of stocks. Algorithms typically determine the timing, price, and quantity of orders, dynamically monitoring market conditions across different securities and trading venues, reducing market impact by optimally (and possibly randomly) breaking large orders into smaller pieces, and closely tracking benchmarks such as the volume-weighted average price (VWAP) over the execution interval.

How to execute trades over a trading horizon involves complex dynamic optimization problems to determine order size, order frequency, and order type. Significant work had been done on how to optimize the trading intensity/frequency given the constraint of completing the entire transaction by a fixed date. Typically orders are broken into pieces so as to minimize cost. Many features have been added to the basic problem: portfolio considerations, modeling the information that motivates the trade, assuming that liquidity does not replenish immediately after it is taken but only gradually over time, and how trading should be sped up or slowed down conditional on beliefs about prices changes mean reverting or continuing.

Moallemi and Saglam examine an significant aspect of optimally implementing an investor’s trading decision: the order type choice problem for how each individual piece of a larger order should be executed. Prior work typically assumes an investor must pay a transitory price impact generated by their trading, which includes the bid-ask spread. An ad hoc transitory price impact function is assumed and optimization proceeds. One way to think of Moallemi and Saglam’s contribution is as studying the details of how to minimize that transitory impact by using limit orders. The paper examines the optimal control problem for an investor wanting to capture the bid-ask spread by placing a limit order rather than paying the spread by placing a market order. The main challenge in using limit orders is that as the
underlying stock price varies, the original limit order price becomes stale (no longer optimal). The selling
investor would like to keep the order at the best ask price. As the stock price moves up the limit order
should be revised upwards to capture as much of the spread as possible. As the stock price moves down
the limit order should be revised downwards to remain at the best price to allow execution. Using their
model the authors quantify the benefits of being able to revise/reprice limit orders more quickly.

The authors make numerous assumptions to generate an optimal policy which has nice properties as the
frequency at which order can be repriced becomes continuous. Allowing for infrequent adjustment of
the limit order’s price is naturally interpreted as latency. The optimal policy and its asymptotic
parsimony are the paper’s main contributions, which are quite nice.

Models of trading in financial markets are typically stylized to focus on particular aspects of the
problem. Subsequent work relaxes these assumptions, so future research opportunities are evident in
the assumptions a paper makes. For this paper, some obvious assumptions are that trading contains no
information (is uncorrelated with the martingale price process), the trading horizon and quantity are
given, the bid-ask spread is exogenous, all trading occurs in one market, and prices are continuous.
Relaxing any of these assumptions would likely require detailed modeling of strategic interactions
between market participants. For example, how does an exogenous bid-ask spread arise? Typically via
an assumption of perfectly competitive markets makers making zero profits. However, the optimal
solution in Theorem 2 shows that the limit order investor places their order at a price that narrows the
bid-ask spread. Hence, a valuable extension would be to examine how competition between multiple
such limit order investors narrows the bid-ask spread.

The paper lists three main contributions in the introduction with the final one being “We empirically
demonstrate that latency cost incurred by trading on a human time scale has dramatically increased for
U.S. equities...” This statement requires proper qualification. The authors are correct in a relative sense,
but incorrect in an absolute sense. Using data on a stock the authors’ chose, Goldman Sachs, in the
context of the authors’ model below I show how the absolute cost of latency has fallen.

The paper’s relative definition latency is clear in Definition 1 where the latency costs are defined as the
percentage difference between the latency free value of the optimal policy and the value of the optimal
policy with latency. Figure 8 shows that this percentage difference increases over time. What is not clear
from Figure 8 is whether the increase is coming from an increase in the numerator or a decrease in the
denominator. If the rise in Figure 8 is from an increase in the numerator then the cost of latency shown
is increasing in both absolute and relative terms. If the trend in Figure 8 is from a fall in the denominator
then absolute latency costs (the numerator) could be falling even though relative latency costs are
rising.

The latency free value of the optimal policy is given in Theorem 3 and is proportional to the bid-ask
spread. Therefore, I calculate the bid-ask spread for Goldman Sachs from 1999 to 2005. The below figure
shows the bid-ask spread in both dollar terms and as a percentage of stock price. Goldman’s share
remains close to $100 throughout the period so the two measures track each other closely. The figure
shows the spread measures falling roughly seven fold. This decline is substantially greater than the
percentage increase shown in Figure 8. Combining my figure and Figure 8 in the paper suggests that latency costs at the beginning of the sample (pre-2001) represent roughly two cents per share, 10% cost of latency times a 20 cent bid-ask spread. At the end of the sample spreads have fallen to roughly three cents per share. Multiplying this times a 20% cost of latency from Figure 8 gives an absolute cost of latency of approximately 0.6 cents per share. This is a decline in of roughly two thirds. Calculations done in basis points are similar. Hence, while I agree that latency costs as a portion of the costs of immediacy have increased, the absolute importance of latency appears to decline. Thus, in the paper’s context, while latency is more important in the trading process, latency may be less important for the investing process.