Computational Finance 14

Time is Money: Estimating the Cost of Latency in Trading

Sasha Stoikov (joint work with Rolf Waeber)

Cornell University

October 28, 2013

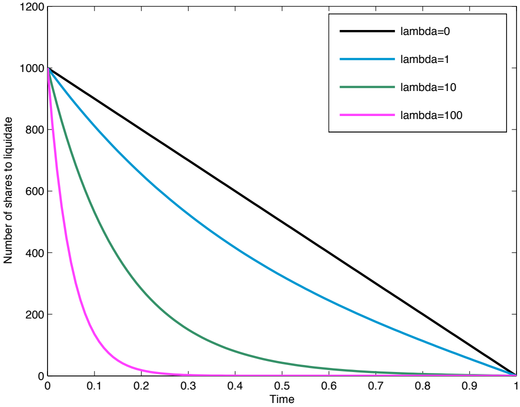
Optimal Liquidation

How to liquidate X shares of an asset?

1. **Macroscopic** time scale:
   - Horizon $\bar{T} > 0$ over which the shares X need to be liquidated.
   - Depends on *long term* variables: average daily volume, strategic considerations, news events, ...
2. **Mesoscopic** time scale:
   - Trade schedule $0 \leq t_0 \leq t_1 \ldots \leq t_i \leq \ldots \leq t_n = \bar{T}$ for the "child" trades.
   - Depends on *medium term* variables: volatility of the stock, risk aversion of the trader, price impact considerations, ...
3. **Microscopic** time scale:
   - Within a time interval $(t_i, t_{i+1}]$, what is the *timing* and the *type of order* used to liquidate the "child" trade?
   - Depends on *short term* variables: limit order book information.

Mesoscopic Time Scale

The trade schedule (Almgren and Chriss (1998)):



Microscopic Time Scale

- We assume that the trade schedule is **given**.
- The goal is then to liquidate one lot (the shares x_t) in the time window $(t_i, t_{i+1}]$, i.e., what is the optimal time τ in $[0, T]$ to sell the lot, where $T = t_{i+1} - t_i > 0$.
- T is typically short, e.g., 1 minute.
- For such short time periods, observing the limit order book can be very advantageous in identifying good liquidation times.
- However, **latency** in the trade execution can diminish this advantage!

Latency

- Latency arises in every trade execution:
  1. Time of datafeed to travel from exchange to execution machine;
  2. The algorithm making a decision;
  3. The order being sent back to the market.
- Latency has no effect on deterministic trade schedules.
- In our model the algorithm will take into account that if a market order is sent at time t it will actually be executed at the best price available at time $t + l$, for latency $l > 0$.
- This worsen the performance of our optimal liquidation algorithm, thus allowing us to quantify the **cost of latency**.

Outline

1. Optimal liquidation:
   - The top-of-book imbalance process.
   - Optimal stopping problem.
   - The trade and no-trade regions.
2. Trading with latency.
3. Dynamic programming.
4. Backtesting strategy on TAQ data.
5. Conclusions.

The Imbalance Process

- The **imbalance process**:

$$I(t) = \frac{B(t)}{A(t) + B(t)}$$

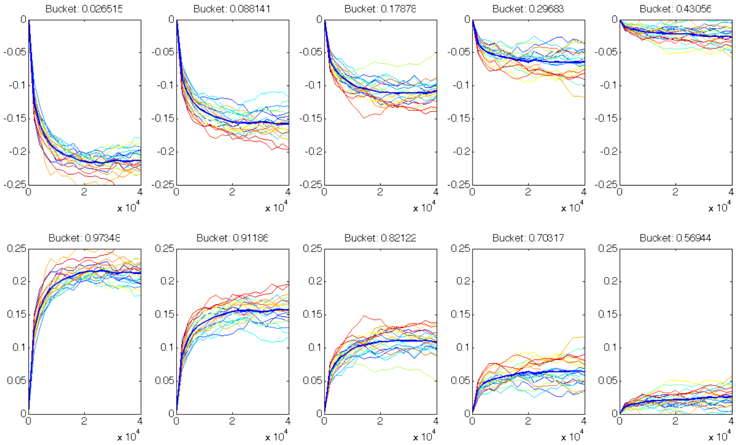
  $B(t)$ is the bid size, $A(t)$ is the ask size.
- We assume $I(t)$ is a Markov process.
- Imbalance is a predictor of short term price moves
  - As a consequence of a zero-intelligence model: Cont, Stoikov and Talreja (2010)
  - Empirically: Avellaneda, Reed and Stoikov (2011)

Motivation

There is empirical evidence that selling on small imbalances can be profitable:

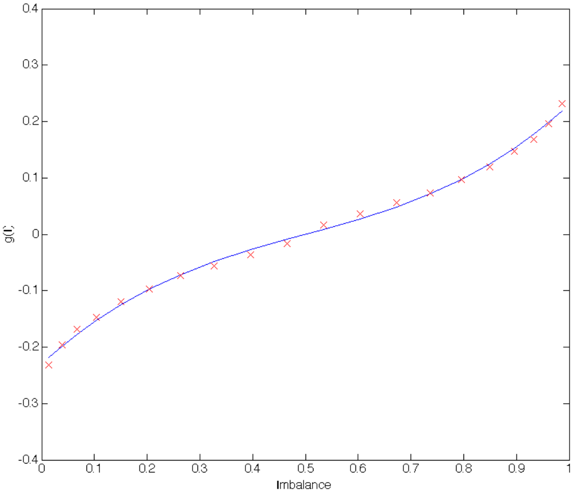
- On each quote i, record the imbalance I_i and the mid price S_i^m
- At a later quote in the future j, record the mid price S_j^m
- Take averages of $(S_j^m - S_i^m)$ for I_i in different buckets

Cost as a fraction of the spread



x axis is time, y axis is cost

Cost of trading on a given imbalance, for dt=20 seconds



The Optimal Liquidation Problem

- Goal: Identify an optimal time τ in $[0, T]$ to sell the share, i.e.,

$$V(t,x) = \inf_{t \leq \tau \leq T} E[I_\tau | I_t = x],$$

for $x \in [0,1]$ and $t \in [0, T]$, and $\tau \in \mathcal{T}$, where $\mathcal{T}$ is the set of stopping times with respect to $\sigma(I(t))_{t \geq 0}$.

- In general we may solve

$$V(t,x) = \inf_{t \leq \tau \leq T} E[g(I_\tau) | I_t = x],$$

Optimal Liquidation based on Minimizing Imbalance

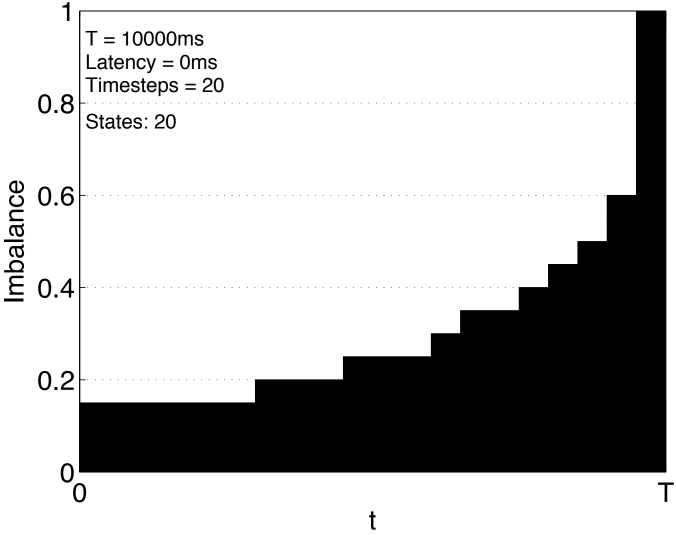
Define

$$
\begin{aligned}
D &= \{(t,x) \in [0,T] \times [0,1) : V(t,x) = x\}\,, \\
C &= \{(t,x) \in [0,T] \times [0,1) : V(t,x) < x\}\,.
\end{aligned}
$$

Proposition

There exists a non-decreasing function $w^* : [0,T] \to [0,1]$ with $w^*(T) = 1$, such that $D = \{(x,t) \in [0,1) \times [0,T] : x \le w^*(t)\}$.

Trade/no Trade Regions



Trading with Latency

- A trade triggered at time t is executed at time $t+L$ for $L>0$.
- Consider

$$V^L(t,x) = \inf_{t \le \tau^L \le T-L} \mathbb{E}[I(\tau^L + L) | I(t) = x],$$

where $\tau^L \in \mathcal{T}$.

- This is equivalent to:

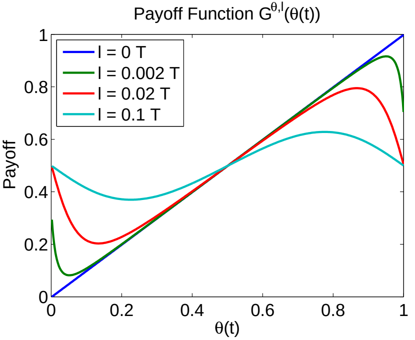
$$V^L(t,x) = \inf_{t \le \tau^L \le T-L} \mathbb{E}[G^L(I(\tau^L)) | I(t) = x].$$

The function G

$V^L(t,x)$ is equivalent to

$$V^L(t,x) = \inf_{t \le \tau^L \le T-l} \mathbb{E}[G^L(I(\tau^L)) | I(t) = x].$$

where $G^L(u) = \mathbb{E}[I(L)|I(0) = u]$.



Latency is Costly

Proposition

Fix $t \in [0, T], s \in \mathbb{R}$, then $V^L(t, x)$ is increasing in L for $L \in [0, T]$.

Trade/No-Trade Regions with Latency

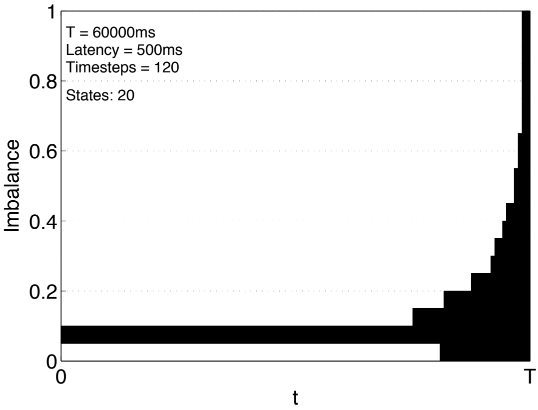
The "trade region" is still connected, but the "no-trade" region does not need to be connected anymore:

Proposition

There exists a non-decreasing function $w_L^* : [0, T] \to [0, 1]$ and a non-increasing function $v_L^* : [0, T] \to [0, 1]$, with $v_L^* \leq w_L^*$, $w_L^*(t) = 1$ for $t \in [T - L, T]$ and $v_L^* = 0$ for $t \in [T - L, T]$, such that

$$D^L = \{(t, u) \in [0, T] \times [0, 1) : v_L^*(t) \leq u \leq w_L^*(t)\} .$$

Trade/No-Trade Regions with Latency cont.



The no-trade region is split in two.

Discretization Approximation

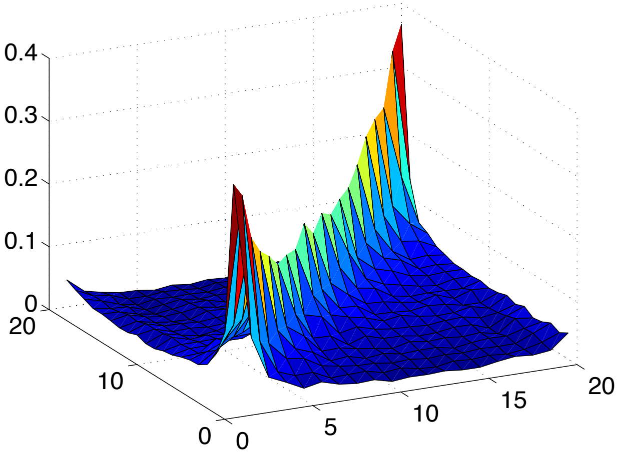
- Knowing $V^L(t,x)$, is enough to identify good liquidation times.
- Let $N, E \in \mathbb{N}$. Define,

$$\begin{aligned} k : [0, T] &\to K = \{0, \dots, N\} \\ t &\mapsto k(t) = \sup \{ n \in \{0, \dots, N\} \, | nT/N \leq t \} , \\ h : [0, 1) &\to H = \{1, \dots, E\} \\ x &\mapsto h(x) = \lfloor Ex \rfloor + 1. \end{aligned}$$

- These mappings transform the original state space $[0, T] \times [0, 1)$ into a *discrete state space* with $(N+1)E$ states.

The transition matrix

The probability p_{ij} that the imbalance will transitions from state i to state j in 500ms



Dynamic Program

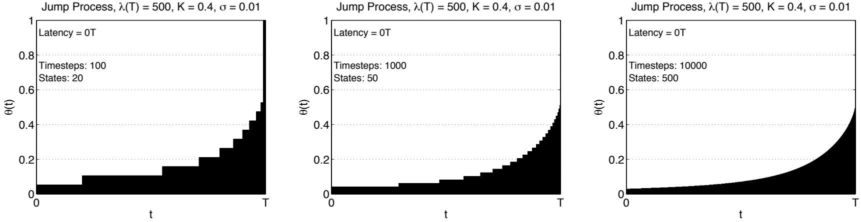
- Bellman's recursion:

$$V_{E,N}^{L}(n,i) = \max\left\{G^{L}(i), \mathbb{E}[V_{E,N}^{L}(n+1, I(n+1))|I(n)=i]\right\},$$

- Conditional probability:

$$\mathbb{E}[V_{E,N}^{L}(n+1, I(n+1))|I(n)=i] = \sum_{k=1}^{E} p_{ik} V_{E,N}^{L}(n+1,k).$$

Discretization Convergence



As $N \to \infty$ and $E \to \infty$ the boundary between trade and no-trade region converges to a smooth curve.

Overview

Backtesting on TAQ data for 5-years US treasury bonds for 21 days (July 2010).

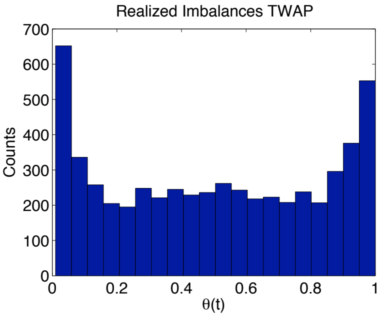
1. The time-weighted average price (TWAP) strategy liquidates one share per minute **independently** of the state of the limit order book.
2. Our imbalance-based algorithm will have T equal to 1 minute. For each day we backtest,
   - we compute the optimal execution region, using the empirical transition matrix from the previous day's data
   - we walk through each quote, decide whether we are in the trade region or not
   - if we are in the trade region submit a sell order which will be executed at the bid L milliseconds later

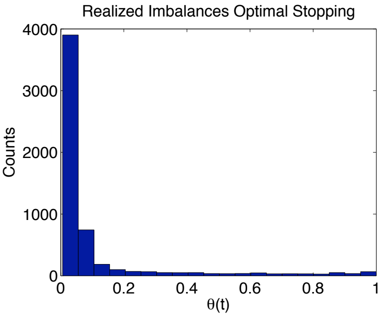
Optimal Stopping vs. TWAP Strategy

- Consider residuals $\hat{R} = S^b_\tau - S^b_T$, where τ is the stopping time from the optimal stopping problem $V(t,x)$.
- Compare 5,649 intervals of length 1 minute.
- **Without latency** the optimal liquidation strategy saves on average 31 $ per share, i.e., **1/3 of the spread** (Spread is 78$ for 5 yrs US-treasury bonds):

	$\mathbb{E}[\hat{R}]$	$\sigma(\hat{R})$
Optimal policy vs. TWAP	31.26 $	49.14 $

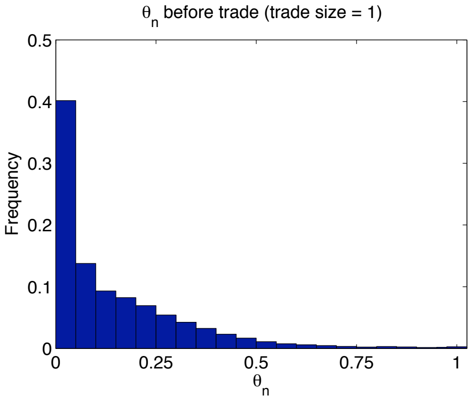
Realized Imbalances





Empirical evidence for trading on low imbalances

Empirical observed imbalance $I(t)$ **conditioned a trade occurs on the next quote**.



Cost of Latency

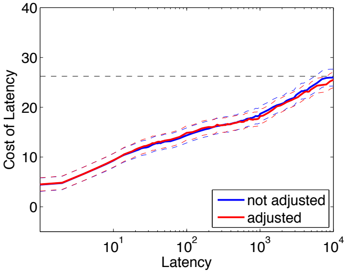
- **Cost of latency**:

$$COL = \mathbb{E}[S_b(\tau) - S_b(\tau + L)],$$

where τ is the stopping time induced by $V(t, x)$.

- Note, we calculate the COL with respect to the **optimal strategy with no latency**.

The Cost of Latency cont.



- 10ms latency $\approx$ 10$ per share.
- For latencies $\geq$ 2000ms (i.e., 2 secs) the advantage of observing the limit order book diminishes (performance becomes similar to TWAP).
- Adjusting the liquidation policy brings only minor improvement in the performance.

Conclusions

- We consider an optimal stopping problem that depends on:
  - Information found in the order book;
  - Latency;
  - The time left to catch up with the TWAP algorithm.
- The solution comes in the form of a trade/ no-trade regions in the imbalance process.
- We estimate model parameters with level-I trades and quotes data.
- We find that our optimal liquidation algorithm significantly outperforms a TWAP algorithm.
- We quantify the cost of latency.
- Reference: Optimal Asset Liquidation Using Limit Order Book Information