Optimal execution in a limit order book and a microstructure model of market impact

Costis Maglaras

Joint work with Ciamac Moallemi, Hua Zheng
Acknowledgment: Praveen Sharma

Columbia Business School

Sept 2015| Cornell FE Seminar Series
Outline

1. Overview of algorithmic trading
   - quick overview of market impact modeling

2. Limit order book as a queueing system

3. Optimal execution in a LOB

4. Microstructure model of market impact
Simplified view of trading

portfolio manager (buy-side)

algorithmic trading engine (buy- or sell-side)

market makers / high-frequency traders

market centers

ARCA  NASDAQ  BATS  …  Dark Pool #1  …

C. Maglaras, 09/2015 – 3/73
Modern U.S. Equity Markets

- Electronic

- Decentralized/Fragmented
  NYSE, NASDAQ, ARCA, BATS, Direct Edge, …

- Exchanges (≈ 70%)
  electronic limit order books (LOBs)

- Alternative venues (≈ 30%)
  ECNs, dark pools, internalization, OTC market makers, etc.

- Participants increasingly automated
  - institutional investors: “algorithmic trading”
  - market makers: “high-frequency trading” (≈ 60% ADV)
  - opportunistic/active (price sensitive) investors: “aggressive/electronic”
  - retail: “manual” (≈ 5% ADV; small order sizes)
Institutional traders (broad strokes)

- Investment decisions & trade execution are often separate processes.
- Institutional order flow typically has “mandate” to execute.
- Trader selects broker, algorithms, block venue, ... (algorithm \( \approx \) trading constraints).
- Main considerations:
  - “Best execution”
  - Access to liquidity (larger orders)
  - Short-term alpha (discretionary investors)
  - Information leakage (large orders the spread over hrs, days, weeks)
  - Commissions (soft dollar agreements)
  - Incentives (portfolio manager & trading desk; buy side & sell side).
- Execution costs feedback into portfolio selection decisions & fund perf.
- S&P500:
  - \( \text{ADV} \approx <1\% \ MktCap \ (0.1\% - 2\%) \)
  - Depth (displayed, top of book) \( \approx .1\% \ \text{ADV} \)
  - Depth (displayed, top of book) \( \approx 10^{-6} - 10^{-5} \ of \ MktCap \)
  \( \Rightarrow \) orders need to be spread out over time.
Market Makers & HFT participants (broad strokes)

- supply short-term liquidity and capture bid-ask spread capture mostly intraday flow; limited overnight exposure

- small order sizes $\sim$ depth; short trade horizons / holding periods

- profit $\approx (\text{captured spread}) - (\text{adverse selection}) - (\text{TC})$
  - critical to model adverse selection: short term price change conditional on a trade

- important to model short term future prices (“alpha”):
  - microstructure signals (limit order book)
  - time series modeling of prices (momentum; reversion)
  - cross-asset signals (statistical arbitrage, ETF against underlying, . . .)
  - news (NLP)
  - detailed microstructure of market mechanisms
    …

- risks: adverse price movements; flow toxicity; accumulation of inventory & aggregate market exposure
Examples of Algorithmic Trading Strategies (90+% of institutional flow)

- **VWAP (Volume-Weighted-Average-Price):** trades according to forecasted volume profile to achieve (or beat) the market’s volume weighted average price
  - Passive strategy; subject to significant market risk

- **TWAP (Time-Weighted-Average-Price):** trades uniformly over time to achieve (or beat) TWAP benchmark
  - Passive strategy; market risk; not very popular in practice

- **POV (Percent-of-Volume):** Executes while tracking the realized volume profile at a target participation rate, e.g., buy IBM at 15% part. rate
  - Controls behavior during volume spikes to avoid excessive cost
  - Popular in practice ~ 5%-30% part.rates; (part.rate ~ cost)

- **IS (Implementation Shortfall):** schedules trade so as to optimally tradeoff expected shortfall (cost) against execution risk
  - variable execution speed; adapts wrt changes in mkt conditions
  - Popular, especially with portfolios where cost/risk tradeoff is intricate
Algorithmic Trading Systems: typically decomposed into three steps

- **Trade scheduling**: splits parent order into ~5 min “slices”
  - relevant time-scale: minutes-hours
  - schedule follows user selected “strategy” (VWAP, POV, IS, ...)
  - reflects urgency, “alpha,” risk/return tradeoff
  - schedule updated during execution to reflect price/liquidity/…

- **Optimal execution of a slice (“micro-trader”)**: further divides slice into child orders
  - relevant time-scale: seconds–minutes
  - strategy optimizes pricing and placing of orders in the limit order book
  - execution adjusts to speed of LOB dynamics, price momentum, ...

- **Order routing**: decides where to send each child order
  - relevant time-scale: ~1–50 ms
  - optimizes fee/rebate tradeoff, liquidity/price, latency, etc.

separation of these components mostly technological/historical artifact
(should not be treated separately)
Algorithmic Trading Systems: basic building blocks

- forecasts for intraday trading patterns
  - volume
  - volatility
  - bid-ask spread
  - market depth
  - ...

- real-time market data analytics

- market impact model

- risk model
  - “of the shelf” risk models calibrated using EOD closing price data do not incorporate intraday correlation structure
  - intraday data? (tractable for liquid securities, e.g., S&P500 universe)
  - cross-asset liquidity model & market impact model
Intraday volume profile: cross-sectional average of S&P500

S&P500 cross-sectional, smoothed intraday trading volume profile (min-by-min).
Volatility Profile

Cross-sectional, smoothed intraday volatility profiles
Spread Profile

Cross-sectional, smoothed intraday spread profiles

Liquid: daily volume 1-100

Iliquid: daily volume 1901-2000
Intraday depth profile

US equities, top 100 securities wrt ADV, cross-sectional, intraday depth profile, in units of $10^{-4} \cdot \text{ADV}$.
Log-depth as a function of spread (top 100 names ranked by ADV)

“Large tick” stocks:

- liquid & low priced stocks, spread $\approx$ $0.01$, but 1 spread = 5 – 15 bps
- depth $\nearrow$ as spread (in bps) $\uparrow$
  ...capturing spread yields significant return
Algorithmic Trading Systems: trade scheduling

- **Trade scheduling:** splits parent order into $\sim$ 5 min “slices”
  - relevant time-scale: minutes-hours
  - schedule follows user selected “strategy” (VWAP, POV, IS, ...)
  - reflects urgency, “alpha,” risk/return tradeoff
  - schedule updated during execution to reflect price/liquidity/...

- **Optimal execution of a slice (“micro-trader”):** further divides slice into child orders
  - relevant time-scale: seconds–minutes
  - strategy optimizes pricing and placing of orders in the limit order book
  - execution adjusts to speed of LOB dynamics, price momentum, ...

- **Order routing:** decides where to send each child order
  - relevant time-scale: $\sim$ 1–50 ms
  - optimizes fee/rebate tradeoff, liquidity/price, latency, etc.

separation of these components mostly technological/historical artifact (should not be treated separately)
Canonical example for trade scheduling: Implementation Shortfall strategy

- Shortfall \( S := \bar{p} - p_{\text{arrival}} \)

- Fundamental tradeoff:
  - quick execution \( \Rightarrow \) adverse price movement (market impact)
  - slow execution \( \Rightarrow \) subject to price risk due to market movement

- Control problem (one possible variation): choose \( X(t) \) to

\[
\min \ E[S] + \lambda \text{Var}[S]
\]

where \( \lambda > 0 \) is a risk aversion parameter

- Alternate formulations:

\[
\min \{ V[S] : E[S] \leq \kappa \} \quad \text{or} \quad \min \{ E[S] : \text{Var}[S] \leq \eta \} \]
Essential building block: market impact model

- Optimizing the trade schedule, i.e., how to split a large trade over smaller waves to be executed over time, requires a cost function for:
  - immediate costs due to current trading decisions (e.g., next 3 min)
  - impact of current decisions on future prices (and future trades)

- Key considerations:
  - transient costs: impact of current trading decisions on price
  - decay of transient costs: instantaneous? impact decays over time?
  - permanent costs: is there a “permanent” cost (information content)?
  - time-scales: interpretation of “transient”, “decay”, “permanent”

- Calibration
  - how to model? functional forms? (depends on relevant time-scale)
  - what data is needed
  - stock segmentation
Market impact models commonly used in practice - 1

- Instantaneous (:= \( f(x(t)) \)) + Permanent (:= \( h(x(t)) \))

\[
\tilde{p}(t) = p(t) + f(x(t)) \quad \text{and} \quad p(t + 1) = p(t) + h(x(t)) + \sigma(t)N(0, 1)
\]

- \( x(t) = q(t)/v(t) = (\text{trade quantity})/(\text{interval volume}) = \text{part. rate} \)
- permanent impact = linear (no-arbitrage argument)
- instantaneous (candidate form):

\[
f(x(t)) = \alpha_{1,t} \cdot s_t + \alpha_{2,t}x(t)^p, \quad p = 1 \text{ or } p = \frac{1}{2}, \frac{2}{3}, \ldots
\]

\( \alpha_{i,t} \) depend on tick size, volume, volatility, . . .

- total expected cost (\( q = [q(1), \ldots, q(T)] \)):

\[
\sum_t q(t)f(x(t)) + \sum_t q(t) \left( \sum_{s < t} h(x(s)) \right) \sim \sum_t q(t)^{1+p} + q^THq
\]
Market impact models commonly used in practice - 2

- Transient (\(f(x(t))\)) with decay (\(g(t - s), \ t > s\))

\[
p(t + 1) = p(0) + \left( \sum_{s \leq t} \frac{1}{(t - s)^b} f(x(s)) \right) + \sigma B(t)
\]

- transient cost rate:

\[
f(x(t)) = \alpha_{1,t} \cdot s_t + \alpha_{2,t} x(t)^p, \ p = 1 \text{ or } p = \frac{1}{2}, \frac{2}{3}, \ldots
\]

- decay rate:

\[
g(t - s) = \frac{1}{(t - s)^b}
\]

- total expected cost:

\[
\sum_t q(t) \left( \sum_{s < t} \frac{1}{(t - s)^b} f(x(s)) \right)
\]
Comments related to market impact modeling

- relevant time scales?
  - instantaneous
  - transient
  - permanent

- short-term costs:
  - intraday characteristics of volume, spread, tick size, volatility
  - wrt participation rate?
  - wrt microstructure variables?
  - do short-term execution costs vary intraday? (say for a 5 minute execution of 5,000 shares of IBM)

- tactical trading decisions (sec? min? hr? days?)
Realized Shortfall for a sample of POV and VWAP orders

- model is more estimable for aggressive executions ($\geq 10\%$ part rate)
Realized Shortfall for POV and VWAP orders (cont.)

- model is more estimable for slow duration orders ($\leq 30$ min)
Algorithmic Trading Systems: short horizon execution in LOB

- **Trade scheduling**: splits parent order into $\sim$ 5 min “slices”
  - relevant time-scale: minutes-hours
  - schedule follows user selected “strategy” (VWAP, POV, IS, ...)
  - reflects urgency, “alpha,” risk/return tradeoff
  - schedule updated during execution to reflect price, liquidity/...

- **Optimal execution of a slice (‘micro-trader’)**: further divides slice into child orders
  - relevant time-scale: seconds–minutes
  - strategy optimizes pricing and placing of orders in the limit order book
  - execution adjusts to speed of LOB dynamics, price momentum, ...

- **Order routing**: decides where to send each child order
  - relevant time-scale: $\sim$ 1–50 ms
  - optimizes fee/rebate tradeoff, liquidity/price, latency, etc.
The Limit Order Book (LOB)

- Buy limit order arrivals
- Cancellations
- Market sell orders
- Market buy orders
- Sell limit order arrivals
- Cancellations

Price

BID

ASK
Multiple Limit Order Books

Price levels are coupled through protection mechanisms (Reg NMS)

national best bid/ask (NBBO)
What are the key considerations & decisions?
Execution in LOB: key modeling and trading decisions

- real-time measurements and forecasts for event rates (arrivals, trades, cancellations on each side of the LOB)
- heterogenous flows wrt arrivals, executions, cancellations (tomorrow)
- time/price queue priority:
  - estimate queueing delay & $P$ (fill in $T$ time units)
  - limit order placement ... depends on queueing effects at each exchange
  - maintain / estimate queue position (& residual queueing delay)
  - adverse selection as a function of exchange, depth, queue position, ...
  - transaction cost models
- microstructure, short-term alpha signals
- optimize execution price by tactically controlling
  - when to post limit orders, and to which exchanges
  - when to cancel orders
  - when & how to execute using market orders
Outline

1. Overview of algorithmic trading
   – quick overview of market impact modeling

2. Limit order book as a queueing system

3. Optimal execution in a LOB

4. Microstructure model of market impact
LOB: event driven (short-term) view

- **Buy limit order arrival rates**: \( \lambda^b_1, \ldots, \lambda^b_{bt-1}, \lambda^b_{bt} \)
- **Price**: \( p_{at}, p_{at+1}, \ldots, p_N \)
- **Sell limit order arrival rates**: \( \lambda^s_{at}, \lambda^s_{at+1}, \ldots, \lambda^s_N \)
- **Buy market order rate**: \( \mu^b_{at} \)
- **Sell market order rate**: \( \mu^s_{bt} \)
- **Cancellation rate**: \( \gamma \)
LOB re-drawn as a multi-class queueing network

\[ \lambda_{bt}, \gamma \rightarrow \mu_{bt}^s \rightarrow \mu_{bt} \rightarrow \text{market sell orders} \]

\[ \lambda_{N}, \gamma \rightarrow \mu_{at} \rightarrow \text{market buy orders} \]

\[ \lambda_{at}, \gamma \rightarrow \text{limit sell orders} \]

\[ \lambda_{1}, \gamma \rightarrow \text{limit buy orders} \]
Limit order arrivals

- Poisson?

- rate fcn’s $\lambda$ (limit order submissions), $\mu$ (trades = service completions)
  - time-of-day
  - price level, distance from best bid / best ask, spread
  - depth, certainly at top of book
  - effective tick size
  - rates of other flows; large blocks; …

other possible considerations:

- model “strategies” that generate flow, e.g.,
  - POV responds to (filtered) volume
  - HFT participants respond “quickly” to queue depletion events
    …

  structurally estimate state-dependent rate fcn
  (complex / over fitting? / depends on intended use)

- bursts?
Order sizes

- distinguish trades that happen on exchanges (as opposed to dark pools)
- most trades in increments of round lots: 100, 200, ...

<table>
<thead>
<tr>
<th></th>
<th>top 500 names (ADV)</th>
<th>top 1000 names (ADV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1 (# shares)</td>
<td>87</td>
<td>84</td>
</tr>
<tr>
<td>Q2 (# shares)</td>
<td>101</td>
<td>101</td>
</tr>
<tr>
<td>Q3 (# shares)</td>
<td>151</td>
<td>139</td>
</tr>
</tbody>
</table>

- odd lots (mostly < 100 share trades – non-negligible)
- roll up trades over $\delta t$ to account for “simultaneous” prints triggered by same parent
- think in $ or in shares (or in depth multiples)?
- trade sizes are heavy-tailed (lognormal gives reasonable fit)
Cancellations

1. disregard cancellations

2. timer-based cancellations:
   – each limit order has a patience $\xi$
   
     e.g., $\xi \sim \exp(\gamma) \Rightarrow$ cancellation outflow $\approx -\gamma Q(t) \delta t$
   
   – state-dependent cancellation flow “stabilizes” queues
   – reasonable model for child orders generated by algorithmic strategies

3. constant cancellation outflow $\approx -\eta \delta t$

   – no feedback stabilization, i.e., as $Q(t) \uparrow$ cancellation flow constant
   – state independent (not good), but more tractable

4. capture heterogeneous trading behaviors:
   i. timer-based algo orders
   ii. state-dependent (event-driven) cancelations for MM orders

   – appropriate model to use depends on the context
Heterogenous trading behaviors

- different market participants exhibit significantly different behavior wrt
  - limit order submission
  - cancellations
  - trade sizes & trade submission triggers

- should we model flow through one order generating process? (single type model)
  - e.g., Poisson ($\lambda(t, \text{state vars})$), sizes $\sim G$, patience $\sim F$

- or model heterogenous behavior and use a mixture model, e.g.,
  - algo: Poisson ($\lambda(t, \text{state vars})$), sizes $\sim \text{Geo}(1/s)$, patience $\sim \exp(\theta)$
  - MM: event driven arrivals, cancellations, trades (typically as a fcn of state and signals)
  - blocks: Poisson($\eta(t, \text{state vars})$), sizes $\sim \text{lognormal}$
Event rates (top of book)
Normalized event rates (top of book)

- Cancellations: $C/\mu$
- Arrival: $\lambda/\mu$

- Cancellation volume (at top of book) $\gg$ trade volume
- Arrival volume (limit orders at top of book) $\gg$ traded volume
Interarrival times (top of book)
Interarrival times (log scale) (top of book)

- liquid stocks: # trades, # cancellations, # limit order arrivals are large
- # trades ≈ 1 order of magnitude less frequent than cancels or order arrivals
Tick period / queueing delay against # trade events

Tick period versus queueing delay: ratio against # trade events. (liquid names)

- tick period = avg time between changes in the mid-price
- tick period is on same (or smaller) order magnitude as queueing delay
Variability of order arrival rates

<table>
<thead>
<tr>
<th>Time</th>
<th>% obs. in ±2σₜ</th>
<th>% obs. in ±3σₜ</th>
<th>% obs. outside ±3σₜ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 min</td>
<td>63.33%</td>
<td>79.23%</td>
<td>20.77%</td>
</tr>
<tr>
<td>3 min</td>
<td>32.56%</td>
<td>50.39%</td>
<td>49.61%</td>
</tr>
<tr>
<td>5 min</td>
<td>27.27%</td>
<td>35.06%</td>
<td>64.94%</td>
</tr>
<tr>
<td>10 min</td>
<td>13.16%</td>
<td>31.58%</td>
<td>68.42%</td>
</tr>
</tbody>
</table>

- Table checks if \( \mu_{t+1} \in \text{intervals } \mu_t \pm k\sigma_t \text{ for } k = 2, 3 \)
- \((\lambda, \mu)\) exhibit significant differences in the time scale of 3 - 5 minutes
- cf. top 100 names (by ADV): average queueing delay = 61 sec (more on this later on)
Some observations

- Event data:
  \[ \lambda \gg \mu \quad \text{and} \quad \text{cancellation flow} \gg \mu \]

- Significant cancellation volume to balance order flow at top of book

- Price changes on the same time-scale as queueing delays

- Event arrival rates fluctuate at slightly slower time scale than queueing delays

- Heterogeneous trading behavior may impact order flow dynamics

- Fragmentation affects delay estimates and cancellation behavior
Suggest a microstructure model of short-term market impact

Motivating questions:

- microstructure variables relevant to short-term execution costs? (min)
- how much do exec costs vary in response to variations of microstructure variables?
- cost variation significant to affect trade execution decisions?
LOB: event driven (short-term) view

- Buy limit order arrival rates: $\lambda^b_1, \ldots, \lambda^b_{bt-1}, \lambda^b_{bt}$
- Sell limit order arrival rates: $\lambda^s_{at}, \ldots, \lambda^s_{at+1}, \ldots, \lambda^s_N$
- Buy market order rate: $\mu^b_{at}$
- Sell market order rate: $\mu^s_{bt}$
- Cancellation rate: $\gamma$
Stylized optimal execution in a LOB

**objective:** how to buy $C$ shares within time $T$ at the lowest price

**controls:** how much, when, at what prices to trade

- trade with limit orders / market orders
- trade with block trades / continuously submitted trades (rate upper bounded by $\kappa_i$)

- $T$ is same order of magnitude as the queueing delays ($\approx 1$ - $5$ min)
- microstructure of the LOB impact execution policy and resulting costs
- we focus on a stylized execution problem (tractable)
  ...to generate insight on impact cost drivers
LOB fluid model dynamics

\[ \dot{Q}_i^b(t) = \lambda_i^b \cdot 1 \{ i \leq b_t \} - \mu_i^s \cdot 1 \{ i = b_t \} - \gamma Q_i^b(t), \]
\[ \dot{Q}_i^s(t) = \lambda_i^s \cdot 1 \{ i \geq a_t \} - \mu_i^b \cdot 1 \{ i = a_t \} - \gamma Q_i^b(t). \]

Main assumptions

- \( \lambda_i^b > \mu_i^s \) (motivated from earlier data analysis)
- constant bid-ask spread, no limit orders inside spread
- if price moves, limit orders slide, queue positions maintained
LOB behavior

\[ b_t = \text{best bid at time } t; \quad a_t = \text{best ask at time } t \]

- (best bid & best ask do not change) \( b_t = b_0, a_t = a_0 \), for all \( t \geq 0 \),

- \( Q(t) \rightarrow q^* \) as \( t \rightarrow \infty \)

\[
q^{*,b}_i := \begin{cases} 
\frac{\lambda_i^b}{\gamma} & \text{if } 1 \leq i < b_0, \\
\frac{\lambda_i^b - \mu_i^s}{\gamma} & \text{if } i = b_0, \\
0 & \text{if } b_0 < i \leq N,
\end{cases}
\]

\[
q^{*,s}_i := \begin{cases} 
0 & \text{if } 1 \leq i < a_0, \\
\frac{\lambda_i^s - \mu_i^b}{\gamma} & \text{if } i = a_0, \\
\frac{\lambda_i^s}{\gamma} & \text{if } a_0 < i \leq N,
\end{cases}
\]

- top of book queues equilibriate to balance arrivals with trades + cancellations
- other queues balance arrivals with cancellations
- cf. Gao, Dai, Dieker, Deng
Optimal execution policy

- **Limit orders**
  - at $t = 0$ submit $C_L$ limit orders to the best bid $b_0$
    \[
    C_L = \min \left\{ \mu_{b_0}^s \left( T - \frac{1}{\gamma} \log \left( 1 + \frac{\gamma}{\mu_{b_0}} Q_0^0(0) \right) \right)^+, C \right\};
    \]

- **Market orders** ($\kappa_i = \kappa$, $\forall i$)
  - at $t = 0$ submit block trade $\min\{C - C_L, Q_{a_0}^s(0)\}$ to the best ask $a_0$;
  - for time $0 < t < T$ continuously submit market orders at the best ask $a_0$ at rate $\kappa$, or until $C$ is filled;
  - at $t = T$, clean up with block trade, may deplete multiple queues at higher price levels.
Practical considerations

- avoid clean up trade, especially if this is a slice of a longer trade
- often times micro-trader does not have to complete \( C \) by \( T \)
- account of multiple exchanges in deciding how much and where to post
- do not post all limit order quantity in one block to avoid information leakage

- policy predicated on the following assumption:
  - trader can execute continuously with market orders at rate \( \kappa \) (presumably low)
  - \( \kappa_i = \kappa \) for all price levels \( i \)
    one may expect supply to increase at higher price levels (more later)

- ...
Outline

1. Overview of algorithmic trading
   - quick overview of market impact modeling

2. Limit order book as a queueing system

3. Optimal execution in a LOB

4. Microstructure model of market impact
   - use cost fcn of control problem as candidate market impact model
   - calibrate model on proprietary trade data set
Drivers of short-term execution cost

- % order traded via limit orders:
  - queue length (near-side)
  - trading volume (near-side)
  - cancellation behavior
  - short-term alpha signal

- % order traded via market orders at top-of-book:
  - queue length (far-side)
  - trading volume (far-side)
  - arrival rate of limit orders at top-of-book (far-side)
  - quote resiliency
  - short-term alpha signal

- % order traded via market orders at higher (when buying) price levels:
  - tick size
  - depth of book
Execution cost

\[ \bar{IS} := \frac{\text{Total cost}}{C} - p = \frac{s}{2} - s \cdot \frac{C_L}{C} + \sum_{k=1}^{N-a_0} k \delta \cdot \frac{C_{a_0+k}}{C}. \]

Simplifications:

- disregard cancellations on the near side (limit order term)
  \[ - C_L = \min \left\{ \mu_{b_0}^s (T - w^0)^+, C \right\}, \text{ where } w^0 = Q_b(0)/\mu^s \]

- clean up cost: the number of price levels needed to complete the trade is

\[ n := \frac{(C - C_L - Q_{a_0}^s(0) - \kappa T)^+}{Q^s} \approx \frac{(C - Q_{a_0}^s(0) - \kappa T)^+}{Q^s} \]

For large \( C \): Total cost \( \sim \alpha_0 + \alpha_1 \cdot C + \alpha_2 \cdot C^2 \)

- \( \alpha_0 \) captures limit order offset, expect to be (-ve)
- \( \alpha_2 \) captures effect of the additional price levels needed, expect to be (+ve)
Microstructure market impact model

- implementation shortfall of a buying order
  - benchmark on the aggressive side: $p_{0,\text{mid}} + s/2 = a_0$
  $$\overline{IS} := \bar{p} - p_{0,\text{mid}} = s/2 + (\text{limit order benefit}) + (\text{higher price level adjustment})$$

- keep insightful structure, simplify the functional form
  $$\overline{IS} = s/2 - \frac{\min \left\{ \left( \mu_{b0}^s T - Q_{b0}^b(0) \right)^+, C \right\}}{C} \cdot s + \frac{\delta}{2} \cdot \frac{(C - Q_{a0}^s(0) - \kappa T)^+}{Q^s} + \frac{\delta}{2}.$$

  \begin{align*}
  \text{limit order benefit} & \quad \frac{\min \left\{ \left( \mu_{b0}^s T - Q_{b0}^b(0) \right)^+, C \right\}}{C} \cdot s \\
  \text{higher price level adjustment} & \quad \frac{\delta}{2} \cdot \frac{(C - Q_{a0}^s(0) - \kappa T)^+}{Q^s} + \frac{\delta}{2}.
  \end{align*}
Microstructure market impact model

- implementation shortfall of a buying order benchmark on aggressive side: \( p_{0,\text{mid}} + \frac{s}{2} = a_0 \)

\[
\overline{IS} = \frac{s}{2} - \min \left\{ \left( \mu_{b_0}^s T - Q_{b_0}^b(0) \right)^+, C \right\} \cdot s + \frac{\delta}{2} \cdot \left( C - Q_{a_0}^s(0) - \kappa T \right)^+ + \frac{\delta}{2}
\]

- (Effect of limit orders) decreasing in \( \mu_{b_0}^s \), \( T \), increasing in \( Q_{b_0}^b(0) \)
- (Effect of top-of-book market orders) decreasing in \( \kappa \) and \( Q_{a_0}^s(0) \)
- (Effect of higher price market orders) decreasing in \( \bar{Q}^s \)
- increasing in \( s, \delta \)
- if \( \kappa_i \) increase as \( p_i \) grows, then cost exhibits sub-linear growth
  - if \( \kappa_i \) grow linearly in \( p_i \), then cost grows like \( \sqrt{C} \) for large \( C \)
realized trade stats: 5min slices for 2013/7-2013/9, > 1,800 securities traded

<table>
<thead>
<tr>
<th></th>
<th>JUL 2013</th>
<th>AUG 2013</th>
<th>SEP 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sample Size</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5min Slices</td>
<td>27,760</td>
<td>30,054</td>
<td>29,226</td>
</tr>
<tr>
<td>Parent Orders</td>
<td>3,396</td>
<td>3,607</td>
<td>3,882</td>
</tr>
<tr>
<td>Distinct Securities</td>
<td>988</td>
<td>896</td>
<td>885</td>
</tr>
<tr>
<td><strong>Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Daily Volume</td>
<td>3,014,000</td>
<td>2,595,000</td>
<td>2,509,000</td>
</tr>
<tr>
<td>Size of 5min Slices</td>
<td>1,294</td>
<td>1,043</td>
<td>849</td>
</tr>
<tr>
<td>Average Queue Length</td>
<td>10,280</td>
<td>21,730</td>
<td>17,750</td>
</tr>
<tr>
<td>Realized Participation Rate</td>
<td>9.60%</td>
<td>9.40%</td>
<td>8.39%</td>
</tr>
<tr>
<td>Price ($)</td>
<td>46.80</td>
<td>38.16</td>
<td>41.41</td>
</tr>
<tr>
<td>Spread ($)</td>
<td>0.031</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td>Daily Volatility</td>
<td>2.23%</td>
<td>1.90%</td>
<td>1.94%</td>
</tr>
<tr>
<td>Implementation Shortfall (bps)</td>
<td>3.04</td>
<td>3.09</td>
<td>3.48</td>
</tr>
</tbody>
</table>
Calibration of auxiliary model parameters

Three quantities not directly observable from data: continuous trading rate $\kappa$, equilibrium queue length $\bar{Q}^s$, effective tick size $\delta$

- **calibration of $\kappa$:**
  1. postulate $\kappa = \theta \cdot \mu$, assume $\theta$ is the same on the bid and ask side
  2. identify slices that: a) queue length at far side less than $1/3$ average length; b) no price change
  3. generate forecast for nominal trading rate $\mu$
  4. $\theta$ estimated as average ratio of executed quantity to the nominal trading rate

- $\bar{Q}^s$ approximated by time-averaged queue length at top of the book over each 5min interval

- $\sigma$ as a proxy for $\delta$
Microstructure market impact model

\[
\overline{IS} = s/2 - \min \left\{ \left( \frac{\mu_{b0}^s T - Q_{b0}^b (0)}{C} \right)^+, C \right\} \cdot s + \frac{\delta}{2} \cdot \frac{\left( C - Q_{a0}^s (0) - \kappa T \right)^+}{\overline{Q}^s} + \frac{\delta}{2}
\]

limit order benefit

higher price level adjustment

\[
\overline{IS} = \beta_0 + \beta_1 \cdot s^* + \beta_2 \cdot (R^L s^*) + \beta_3 \cdot (R^M \delta^*) + \beta_4 \cdot \delta^*
\]

linear regression:

- \( R^L := \frac{\min \left\{ C, \left( \frac{\mu_{b0}^s T - Q_{b0}^b (0)}{C} \right)^+ \right\}}{C} \)
- \( R^M := \frac{\left( C - Q_{a0}^s (0) - \kappa T \right)^+}{\overline{Q}^s} \)
In-sample regressions (ADV $\geq 300,000$ shares; POV $\in (1\%, 30\%)$)

Monthly linear regression results for microstructure market impact model

<table>
<thead>
<tr>
<th></th>
<th>JUL 2013</th>
<th>AUG 2013</th>
<th>SEP 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>(intercept)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coefficient</td>
<td>-0.6888***</td>
<td>-0.6941***</td>
<td>-0.5832**</td>
</tr>
<tr>
<td>std. error</td>
<td>0.1232</td>
<td>0.1140</td>
<td>0.1076</td>
</tr>
<tr>
<td>spread (bps): $s^*$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coefficient</td>
<td>0.3187***</td>
<td>0.3905***</td>
<td>0.3950***</td>
</tr>
<tr>
<td>std. error</td>
<td>0.0069</td>
<td>0.0077</td>
<td>0.0070</td>
</tr>
<tr>
<td>limit order: $R^L s^*$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coefficient</td>
<td>-0.3027***</td>
<td>-0.3415***</td>
<td>-0.3658***</td>
</tr>
<tr>
<td>std. error</td>
<td>0.0107</td>
<td>0.0100</td>
<td>0.0099</td>
</tr>
<tr>
<td>add. tick to pay: $R^M \sigma^*$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coefficients</td>
<td>0.0991***</td>
<td>0.1480***</td>
<td>0.1486***</td>
</tr>
<tr>
<td>std. error</td>
<td>0.0234</td>
<td>0.0225</td>
<td>0.0348</td>
</tr>
<tr>
<td>tick size: $\sigma^*$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coefficients</td>
<td>2.3238***</td>
<td>1.8508***</td>
<td>2.4290***</td>
</tr>
<tr>
<td>std. error</td>
<td>0.1098</td>
<td>0.0997</td>
<td>0.0996</td>
</tr>
<tr>
<td>R-squared</td>
<td>9.91%</td>
<td>10.62%</td>
<td>13.48%</td>
</tr>
</tbody>
</table>

Significance: *** p<0.001, ** p<0.01, * p<0.05
In-sample regressions

Monthly in-sample linear regression results for microstructure market impact model

- consistently good performance of our model, represented by high $R^2$ values

- coefficients of “micro-level” book variables are statistically significant

- signs of the coefficients are intuitive: limit order $-$, higher price order $+$
Cross-validation

- **Cross-Validation**
  (our “micro” model)

\[
\nu^* = \beta_0 + \beta_1 \cdot s^* + \beta_2 \cdot R^L s^* + \beta_3 \cdot R^M \delta^* + \beta_4 \cdot \delta^*
\]

(benchmark “macro” model)

\[
\nu^* = \beta_0 + \beta_1 \cdot (\text{Percent of Market Vol.})^\alpha \sigma^* + \beta_2 \cdot \sigma^*
\]

out-of-sample $R^2$: our model 11% vs. benchmark models 3%

<table>
<thead>
<tr>
<th></th>
<th>Our Model</th>
<th>Linear</th>
<th>Square Root</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg. out-of-sample $R^2$</td>
<td>11.03%</td>
<td>3.11%</td>
<td>3.12%</td>
</tr>
<tr>
<td>relative improvement</td>
<td>0.00%</td>
<td>255%</td>
<td>254%</td>
</tr>
</tbody>
</table>

- $\sigma(t)$ above; using daily $\sigma$ reduces explanatory power by 1-2%
- serial correlation: including 1 or 2 lagged residuals improves performance coefficients are stat. significant and have right signs
Simulated costs as microstructure variables are varied ($C = 3 \times \text{Depth}$)

- randomly generated 4-tuples for ($Q_b, Q_a, \mu_b, \mu_s$)
- variables varied by a random multiplier in $(1/3, 1)$ w.p. .5 and $(1, 3)$ w.p. .5
- cost estimates vary by $\pm 60\%$ around “nominal” values (significant for trade decisions)
Robustness - order & security segmentation

**Segmentation:** by market participation rate

<table>
<thead>
<tr>
<th>Percent of market vol.</th>
<th>micro model</th>
<th>macro model</th>
<th>sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>linear</td>
<td>square root</td>
<td></td>
</tr>
<tr>
<td>[1%,10%]</td>
<td>8.82%</td>
<td>1.87%</td>
<td>1.89%</td>
</tr>
<tr>
<td>(10%,20%)</td>
<td>14.10%</td>
<td>5.34%</td>
<td>5.21%</td>
</tr>
<tr>
<td>(20%,30%)</td>
<td>15.08%</td>
<td>4.23%</td>
<td>4.24%</td>
</tr>
<tr>
<td>overall: [1%,30%]</td>
<td>11.03%</td>
<td>3.11%</td>
<td>3.12%</td>
</tr>
</tbody>
</table>

- *micro* model outperforms the *macro* benchmark models for all groups
- all models improve as the participation rate increases
Robustness - order & security segmentation

Segmentation: by (average daily volume, average queue length)

<table>
<thead>
<tr>
<th></th>
<th>Low Depth</th>
<th>Mid Depth</th>
<th>High Depth</th>
<th>Ultra Deep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low ADV</td>
<td>6.26%</td>
<td>10.23%</td>
<td>17.14%</td>
<td>N/A</td>
</tr>
<tr>
<td>Mid ADV</td>
<td>5.38%</td>
<td>8.12%</td>
<td>12.62%</td>
<td>N/A</td>
</tr>
<tr>
<td>High ADV</td>
<td>N/A</td>
<td>5.56%</td>
<td>10.32%</td>
<td>24.84%</td>
</tr>
</tbody>
</table>

- *micro* model outperforms the *macro* benchmark models for all groups
- model accuracy improves with queue length
- similar results when segmenting queue lengths in shares and dollars
Robustness - effect of nonlinearity

**Simplification:** remove the non-linearities

\[
\nu^* = \beta_0 + \beta_1 \cdot s^* + \beta_2 \cdot \frac{\left(\mu_{b_0}^s T - Q_{b_0}^b (0)\right)}{C} \cdot s^* + \beta_3 \cdot \frac{\left(C - Q_{a_0}^s (0) - \kappa T\right)}{\bar{Q}^s} \cdot \delta^* + \beta_4 \cdot \delta^*
\]

<table>
<thead>
<tr>
<th></th>
<th>micro w/o nonlinearity</th>
<th>macro linear</th>
<th>macro square root</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg. out-of-sample $R^2$</td>
<td>8.19%</td>
<td>3.11%</td>
<td>3.12%</td>
</tr>
<tr>
<td>relative improvement</td>
<td>0.00%</td>
<td>163%</td>
<td>163%</td>
</tr>
</tbody>
</table>

- may affect computational tractability in context of optimization e.g., stock selection, trade scheduling
- still significantly outperforms benchmark models
Robustness - effect of time horizon

**Time horizon**: 5min vs. 1min slices

Model accuracy depends on the time horizon of the trade slices, *micro* model has even better statistical fit for shorter-horizon slices

<table>
<thead>
<tr>
<th></th>
<th>Our Model</th>
<th>Linear</th>
<th>Square Root</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg. out-of-sample $R^2$</td>
<td>16.57%</td>
<td>2.67%</td>
<td>2.81%</td>
</tr>
<tr>
<td>relative improvement</td>
<td>0.00%</td>
<td>521%</td>
<td>490%</td>
</tr>
</tbody>
</table>
## 1-min horizon: order / stock segmentation

### Percent of market vol.

<table>
<thead>
<tr>
<th>Percent of market vol.</th>
<th>Our Model</th>
<th>Linear</th>
<th>Square Root</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1%,10%]</td>
<td>13.53%</td>
<td>0.94%</td>
<td>0.96%</td>
<td>73,166</td>
</tr>
<tr>
<td>(10%,20%]</td>
<td>19.24%</td>
<td>2.26%</td>
<td>2.26%</td>
<td>40,631</td>
</tr>
<tr>
<td>(20%,30%]</td>
<td>21.51%</td>
<td>3.59%</td>
<td>3.59%</td>
<td>19,830</td>
</tr>
<tr>
<td>overall: [1%,30%]</td>
<td>16.57%</td>
<td>2.67%</td>
<td>2.81%</td>
<td>133,627</td>
</tr>
</tbody>
</table>

### Our Model

<table>
<thead>
<tr>
<th>Our Model</th>
<th>Low depth</th>
<th>Mid depth</th>
<th>High depth</th>
<th>Ultra deep</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low ADV</td>
<td>12.18%</td>
<td>13.81%</td>
<td>23.12%</td>
<td>too few obs.</td>
<td></td>
</tr>
<tr>
<td>Mid ADV</td>
<td>9.41%</td>
<td>10.84%</td>
<td>18.78%</td>
<td>too few obs.</td>
<td></td>
</tr>
<tr>
<td>High ADV</td>
<td>too few obs.</td>
<td>3.91%</td>
<td>20.74%</td>
<td>28.98%</td>
<td>16.57%</td>
</tr>
</tbody>
</table>
Robustness - prediction vs. attribution

**Prediction:** pre-trade cost estimates use information available at the beginning of the trading slice

<table>
<thead>
<tr>
<th></th>
<th>Our Model</th>
<th>Linear</th>
<th>Square Root</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>predictive</td>
<td>predictive</td>
<td>predictive</td>
</tr>
<tr>
<td>8.20%</td>
<td>11.07%</td>
<td>2.26%</td>
<td>2.25%</td>
</tr>
<tr>
<td></td>
<td>2.82%</td>
<td>2.84%</td>
<td></td>
</tr>
</tbody>
</table>

- the drop in explanatory power is more significant in *micro* model
- *micro* model still significantly outperforms the two benchmark models
- similar comparison when using historical forecasts (monthly averages)
Low ADV securities (ADV ∈ (50K, 300K) shares, POV ∈ (1%, 30%))

<table>
<thead>
<tr>
<th></th>
<th>JUL 2013</th>
<th>AUG 2013</th>
<th>SEP 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>(intercept)</td>
<td>-0.1240</td>
<td>-0.0611</td>
<td>1.4911***</td>
</tr>
<tr>
<td>std. error</td>
<td>0.3093</td>
<td>0.2442</td>
<td>0.2147</td>
</tr>
<tr>
<td>spread (bps): $s^*$</td>
<td>0.3576***</td>
<td>0.3958***</td>
<td>0.2826***</td>
</tr>
<tr>
<td>std. error</td>
<td>0.0078</td>
<td>0.0072</td>
<td>0.0049</td>
</tr>
<tr>
<td>limit order: $R^Ls^*$</td>
<td>-0.2829***</td>
<td>-0.2582***</td>
<td>-0.1753***</td>
</tr>
<tr>
<td>std. error</td>
<td>0.0123</td>
<td>0.0110</td>
<td>0.0093</td>
</tr>
<tr>
<td>add. tick to pay: $R^Mσ^*$</td>
<td>0.7137***</td>
<td>0.5796***</td>
<td>0.5271***</td>
</tr>
<tr>
<td>std. error</td>
<td>0.1326</td>
<td>0.1499</td>
<td>0.1242</td>
</tr>
<tr>
<td>tick size: $σ^*$</td>
<td>1.1214***</td>
<td>0.6174***</td>
<td>1.2526***</td>
</tr>
<tr>
<td>std. error</td>
<td>0.2267</td>
<td>0.1972</td>
<td>0.1791</td>
</tr>
<tr>
<td>R-squared</td>
<td>25.02%</td>
<td>27.20%</td>
<td>21.56%</td>
</tr>
</tbody>
</table>

Significance: *** p<0.001, ** p<0.01, * p<0.05

<table>
<thead>
<tr>
<th></th>
<th>Micro model</th>
<th>Benchmark macro model</th>
<th>Mean predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α = 1</td>
<td>α = 0.5</td>
<td></td>
</tr>
<tr>
<td>Avg. out-of-sample $R^2$ (vs. predicted mean)</td>
<td>23.26%</td>
<td>4.72%</td>
<td>4.91%</td>
</tr>
<tr>
<td>relative improvement</td>
<td>0.00%</td>
<td>393%</td>
<td>374%</td>
</tr>
</tbody>
</table>
Microstructure market impact model ... closing comments

- Microstructure information helps in cost prediction & attribution
- Short-term price momentum predictions further improve model statistical accuracy
- Microstructure model could serve as “canonical” model of short-term costs &
  - Estimate associated decay
  - Estimate permanent cost
- Fluctuation of microstructure variables (that are common) may lead to fluctuations in estimated costs of ± 60% ... significant in optimizing trading decisions
References: some background on limit order book markets


- E. Smith, J. D. Farmer, L. Gillemot, S. Krishnamoorthy, Statistical theory of the continuous double auction, Quantitative Finance, 2003, 3, 481-514. (good starting point for the “process flow” view of LOB; queueing per se is not in that paper)


References on queueing models of limit order books

  
  (first paper to explicitly study the queueing dynamics of the LOB and make some tactical predictions of its queueing behavior; used exact analysis)


  
  (this paper derived and studied a diffusion model (in the conventional heavy-traffic regime) (cf. slides 78-79, 82, 209))


Background references on market impact


Some references on modeling market impact with microstructure variables

