Optimal Order Execution with Empirical Cost Models

Michael G Sotiropoulos
Algorithmic Trading Quantitative Research
Bank of America Merrill Lynch

Princeton Quant Trading Conference
25-Mar-2011
All statements in this presentation are the author’s personal views and not necessarily those of Bank of America Merrill Lynch.
Contents

Interpreting Trading Cost
  Point of View
  Setup and Notations
  Marking to Market

Empirical Cost Models
  Use Cases
  Impact Kernels
  Calibration

Optimal Pre-trade Trajectories
  Meaning and Usage
  Two Period Trading
  $M$ Period Trading
  Real Life Trading

Summary

References
Trading cost optimization is discussed from the point of view of a large broker/dealer, such as Global Execution Services (GES) at Bank of America Merrill Lynch.

A full service broker provides

- Coverage of all publicly traded asset classes (equities, futures, options, commodities, FX)
- Worldwide access to exchanges, ATS/MTF, crossing networks
- Access to its own dark aggregator
- Algorithmic trading (VWAP, TWAP, POV, IS, other proprietary)
- Smart Order Routing (SORT) for accessing liquidity in fragmented markets
- Client support with pre-trade consultancy and post-trade analysis

A large agency trading business services a variety of clients

- Institutional
- Retail
- Quant funds
- Internal desk flow
Interpreting Trading Cost: Point of View (II)

The point of view determines the aspects and even the language used for understanding trading cost.

For an agency broker

- Order execution is asset, quantity and time constrained by the client.
- A high fill ratio is important, the client wants the order executed.
- The arrival price is a benchmark favored by many clients.
- Sourcing liquidity from various venues in a fragmented market place and internal pools is vital.
- The risk of entering the position belongs to the client.
- The long-term view (alpha) is almost never revealed by the client to the broker.

For a market maker

- Quantity and execution horizon is open-ended (within risk bounds).
- The business model is about capturing spread and exchange rebates.
- All execution risk belongs to the market maker.
- There is no long term view on the traded asset.

But is trading cost really point of view dependent?
Interpreting Trading Cost: Setup and Notations

Consider an order to buy ($\epsilon = 1$) or sell ($\epsilon = -1$) a quantity of $X$ shares of an asset.

There is a continuous double auction market, and at each event time $t_i$ some of the observables are

- $p_i$ trade price
- $\Delta V_i$ traded market volume
- $\Delta x_i$ our executed quantity, $\Delta x_i = \epsilon |\Delta x_i|$
- $p_i^{(b)}, p_i^{(a)}$ best bid and ask prices, top of the book
- $q_i^{(b)}, q_i^{(a)}$ best bid and ask quantities, top of the book

Common computed quantities are

- $s_i = p_i^{(a)} - p_i^{(b)}$ the spread
- $m_i = \frac{1}{2} \left( p_i^{(a)} + p_i^{(b)} \right)$ the mid-quote price

Let the time horizon for trading all the $X$ shares be $T$. Within time $T$ the market prints volume $V_T = \sum_i \Delta V_i$. Our realized participation rate is $\rho := X / V_T$.

Orders fall into two basic categories:

1. Time constrained (VWAP) defined by ($\epsilon, X, T$)
2. Participation rate constrained (POV, in-line), defined by ($\epsilon, X, \rho_{\text{target}}$)
Interpreting Trading Cost: Marking to Market (I)

Typically, trading cost is understood as coming from the price impact generated by our order's sequence of executions of $\Delta x_i$ shares.

- But this rationale can be applied to all orders of all agents in the market. Is our order special?

Following the standard mark-to-market approach, let's set up a portfolio $\Pi$ that contains the risky asset and cash as follows:

- For a buy (sell) order at time $t = 0$, the portfolio is short (long) $X$ shares and long (short) cash equal to the arrival value of the shares.

$$\text{MTM value: } \Pi_0 = -\epsilon X p_0 + \epsilon X p_0 = 0$$

- At time $t = T$ the portfolio will contain $0$ shares.

$$\text{MTM value: } \Pi_T \text{ (all cash)}$$

The cost of trading or implementation shortfall is the portfolio loss $C_T := -\Pi_T$.

Mark-to-Market:

- Call $t_i^+$ the time just after the $i$-th trade has been executed.
- Our order has traded quantity $\Delta x_i := x_i - x_{i-1}$, which may be zero if we did not participate.
- The execution price is $p_i$. This is the new reference price, effective from $t_i^+$ to $t_{i+1}^+$. 
- We mark our share position $x_i$ at $t_i^+$ and update our cash account for debits/credits.
## Interpreting Trading Cost: Marking to Market (II)

<table>
<thead>
<tr>
<th>$t$</th>
<th>$N_{SHR}$</th>
<th>$\Delta V_{SHR}$</th>
<th>$\Delta V_{CSH}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_i^+$</td>
<td>$x_i$</td>
<td>$x_i p_i - x_{i-1} p_{i-1}$</td>
<td>$-\Delta x_i p_i$</td>
</tr>
<tr>
<td>$T^+$</td>
<td>0</td>
<td>$-x_{T-1} p_{T-1}$</td>
<td>$-\Delta x_T p_T$</td>
</tr>
</tbody>
</table>

The net value change at time $t_i^+$ is

$$\Delta \Pi_i = (x_i p_i - x_{i-1} p_{i-1}) - (x_i - x_{i-1}) p_i. \tag{1}$$

The cost of trading at completion time $T$ is (using $x_T = 0$, $x_0 = -\epsilon X$)

$$C_T := -\Pi_T = - \sum_{i=1}^{T} \Delta \Pi_i = \sum_{i=1}^{T} \Delta x_i p_i - \epsilon X p_0. \tag{2}$$

Equivalently, from eq (1),

$$\Delta \Pi_i = x_{i-1} \Delta p_i, \tag{3}$$

and the cost of trading becomes

$$C_T := -\Pi_T = - \sum_{i=1}^{T} \Delta \Pi_i = - \sum_{i=1}^{t} x_{i-1} \Delta p_i. \tag{4}$$

In continuous time eq (1) becomes

$$C_T = - \int_0^T d(x_t p_t) + \int_0^T d x_t p_t = \int_0^T d x_t p_t + x_0 p_0 = - \int_0^T x_t dp_t.$$

---

Implementation shortfall.  
Short term alpha.
Conclusion:

The cost of trading relative to arrival price, or implementation shortfall, can be interpreted as:

1. Sum of costs for trading $dx_i$ shares at price $p_i$ and moving the price in the process (impact story).

2. Sum of P&L from marking to market the holding or “leaves” quantity $x_{i-1}$ (alpha story).

The broker’s and the market maker’s interpretations are linked by an integration by parts, eq (5).

Consequently:

- Trading cost is affected both by our demand for liquidity as well as the collective buy/sell behavior of the other agents.

- Trading cost and short term alpha are dual concepts.

In what follows we will adopt the broker’s interpretation of trading cost (impact story).
Empirical Cost Models: Use Cases

Macroscopic models of the trading cost are difficult to estimate and they explain a very small fraction of the execution variance. Yet, we do maintain them because:

- They are useful for giving rough pre-trade estimates of cost. In this way they act as classifiers of the difficulty to trade an order.
- They can be used for deriving target execution schedules to be followed by the algorithms. In other words, they set the scales for trading a particular asset over a certain period of time.
- Typically they have simple parametric forms and are easy to use in more complex applications, such as constrained portfolio optimization.

Macroscopic cost models are effective theories, where all the microstructure degrees of freedom have been averaged over. Here we briefly review their formulation in continuous time.

Some useful quantities are:

\[ \dot{x}_t \] our speed of trading at time \( t \), in number of shares per unit time

\[ \dot{V}_t \] the market’s speed of trading, in number of shares per unit time

\[ \rho_t \] instantaneous participation rate, \( \rho_t = \frac{\dot{x}_t}{\dot{V}_t} \), dimensionless

\[ \sigma_t \] volatility of the fundamental price, in units of currency over square-root time.

The fundamental price is not observable but should be cointegrated with the mid-quote price \( m_t \).
Empirical Cost Models: Impact Kernels (I)

The expected total cost has two parts:

1. Slippage from the microstrategy of spread crossing
2. Price impact from our speed of trading (liquidity demand)

\[ C_{\text{tot}} = c\bar{s} + \int_{0}^{T} \dot{x} dt \int_{0}^{t} h(\dot{x}_s, s, t) ds, \]  \hspace{1cm} (6)

where \( \bar{s} \) is the average spread within \([0, T]\) and \( h(\dot{x}_s, s, t) \) the impact kernel, with units ccy/(share x time).

ASSUMPTION 1: The impact kernel can be factorized into a price impact function and a time-homogeneous, non-increasing decay kernel as

\[ h(\dot{x}_s, s, t) = f(\dot{x}_s) G(t - s). \]  \hspace{1cm} (7)

NOTE: The factorization assumption is for phenomenological convenience, it does not necessarily arise from the underlying limit order book dynamics.

ASSUMPTION 2: The impact function is a power law of the participation rate

\[ f(\dot{x}_t) = \eta \epsilon_t \rho_t^\alpha, \]  \hspace{1cm} (8)

with the asset specific coefficient \( \eta \) having units ccy/share.

From now on we focus on buy-only orders, \( \epsilon_t = 1 \).

Identical results follow from sell-only orders, \( \epsilon_t = -1 \).
Empirical Cost Models: Impact Kernels (II)

With the above assumptions the impact cost becomes

$$C [\dot{x}_t] = \eta \int_0^T \dot{x}_t dt \int_0^t \rho_t^\alpha G (t - s) ds.$$  \hfill (9)

Enter the market volume:

$V_t$ is a non-decreasing, stochastic function of $t$, i.e. a clock.

- $V_t$ is another source of execution uncertainty for volume tracking algorithms
- Forecasting $V_t | V_{t-1}, V_{t-2} \ldots$ is done via econometric models, typically auto-regressive with intraday adjustments.

Integrals over clock time can be transformed into integrals over volume time by the change of variables $v = V_t$ as

$$\int_0^T \dot{x}_t dt = \int_0^T \rho_t \dot{V}_t dt = \int_0^{V(T)} \bar{\rho}_v dv.$$  \hfill (10)

ASSUMPTION 3: The market speed $\dot{V}_t$ is deterministic.

- Volume uncertainty is a higher order effect for macroscopic cost modeling.
- Without loss of generality we will further assume that the market speed $\dot{V}$ is constant.
- A POV order of target rate $\bar{\rho}$ and a VWAP of duration $T$ are related as

$$\bar{\rho} = \frac{X}{\dot{V} T}.$$  \hfill (11)
Empirical Cost Models: Impact Kernels (III)

**Instantaneous Decay:**
If $G(t - s) = \delta(t - s)$ then, from eq (9),

$$ C[\dot{x}_t] = \eta \int_0^T \dot{x}_t \rho_t^\alpha dt = \eta \dot{V} \int_0^T \rho_t^{1+\alpha} dt. $$

The per-share cost of a constant POV (or VWAP) strategy becomes

$$ \frac{C(\bar{\rho})}{\bar{X}} = \eta \tilde{\rho}^\alpha. $$

- $\alpha = 1$: linear instantaneous impact; $\alpha = 0$: linear permanent impact

**Square-root Decay:**
If $G(t - s) \propto 1/\sqrt{(t - s)}$ then, from eq (9),

$$ C[\dot{x}_t] = \eta \bar{\sigma} \int_0^T \dot{x}_t dt \int_0^t \frac{\rho_s^\alpha}{\sqrt{t - s}} ds, $$

with $\bar{\sigma}$ the average volatility within $[0, T]$, conveniently providing the extra $1/\sqrt{T}$ time scale. Setting $\eta = bp_0$, the cost of a constant POV (or VWAP) strategy in bps becomes

$$ \frac{C(\bar{\rho})}{\bar{X}p_0} = \frac{4}{3} b \bar{\sigma} \sqrt{T} \tilde{\rho}^\alpha. $$

- For $\alpha = 1/2$ we get the “famous” square-root model with normalized cost

$$ \frac{C(\bar{\rho})}{\bar{X}p_0 \bar{\sigma} \sqrt{T}} = \frac{4}{3} b \sqrt{\frac{X}{VT}} \propto \sqrt{\text{PctADV}}. $$
Empirical Cost Models: Calibration

Model parameters are estimated by fitting the cost of a large number of orders, typically $\mathcal{O}(10^5)$.

ML: square-root decay kernel and power law impact.
Assume we have adopted and fitted a macro cost model like the ones described earlier. How does this inform our trading?

- Derive minimum pre-trade cost trajectories.
- Set these trajectories as targets in the algorithmic trading engine.
- Allow opportunistic or forced deviations from the target trajectory, if a trading signal becomes strong, or if we must deliver a predetermined amount of shares within a time slice.

In the following:

1. We work with the square-root decay model.
2. We set the risk aversion to zero (risk neutral) since we are more interested in how time decay affects the minimum pre-trade cost trajectory.
Optimal Pre-trade Trajectories: Two Period Trading (I)

• **PROBLEM** (Two period trading): Given an order to buy $X$ shares within time $T$, while the market is printing trades at constant speed $\dot{V}$, if we are allowed to change the speed of trading at some time $T_1 = \theta T$, $0 \leq \theta \leq 1$, what are the constant participation rates $\rho_1$ through $[0, \theta T)$ and $\rho_2$ through $[\theta T, T]$ that minimize the cost of trading?

Let’s use the cost of VWAP $C(\tilde{\rho})$ as a reference. In the first (second) period we trade at speed $\rho_{1(2)} = z_{1(2)} \tilde{\rho}$, with $z_{1,2} \geq 0$. Given $\theta$, what are the $z_{1,2}$ that minimize $C(\rho_1, \rho_2) / C(\tilde{\rho})$?

Using eq (14), we can write

$$
\frac{C(\rho_1, \rho_2)}{C(\tilde{\rho}) (4/3) \tilde{\rho}^{1+\alpha} T^{3/2}} = \int_0^{\theta T} \rho_1 dt \int_0^t \rho_1^\alpha (t-s)^{-1/2} ds
$$

$$
+ \int_{\theta T}^T \rho_2 dt \int_{\theta T}^t \rho_2^\alpha (t-s)^{-1/2} ds
$$

$$
+ \int_{\theta T}^T \rho_2 dt \int_0^{\theta T} \rho_1^\alpha (t-s)^{-1/2} ds.
$$

(17)

For a two period trading we have two direct terms and one cross term.
Optimal Pre-trade Trajectories: Two Period Trading (II)

- Define $\tilde{\theta} := 1 - \theta$. The cost is calculated to be

$$\frac{C(\rho_1, \rho_2)}{C(\tilde{\rho})} = \rho_1^{1+\alpha} \theta^{3/2} + \rho_2^{1+\alpha} \tilde{\theta}^{3/2} + \rho_1^\alpha \rho_2 \left(1 - \theta^{3/2} - \tilde{\theta}^{3/2}\right). \quad (18)$$

Sanity check: for $\rho_1 = \rho_2 = \tilde{\rho}$ and for every $\theta$ the above gives a relative cost of 1.

- In terms of the acceleration factors $z_{1,2}$ we have the optimization problem:

$$\min_{z_{1,2}} \frac{C(\rho_1, \rho_2)}{C(\tilde{\rho})} = z_1^{1+\alpha} \theta^{3/2} + z_2^{1+\alpha} \tilde{\theta}^{3/2} + z_1^\alpha z_2 \left(1 - \theta^{3/2} - \tilde{\theta}^{3/2}\right) \quad (19)$$

subject to

$$z_1 \theta + z_2 \tilde{\theta} = 1 \quad (20)$$

$$z_1 \geq 0, \quad z_2 \geq 0. \quad (21)$$

- Substituting $z_2 = (1 - z_1 \theta) / \tilde{\theta}$, the first order condition w.r.t. $z \equiv z_1$ becomes

$$\left(1 + \alpha \right) z^{\alpha} \theta^{3/2} - \left(1 + \alpha \right) (1 - z \theta)^\alpha \tilde{\theta}^{3/2 - \alpha}$$

$$+ \left(\alpha z^{\alpha-1} - (1 + \alpha) z^{\alpha} \theta \right) \left(1 - \theta^{3/2} - \tilde{\theta}^{3/2}\right) / \tilde{\theta} = 0. \quad (22)$$

- NOTE: The VWAP solution $z = 1$ is not a root of the above for general $\theta \in (0, 1)$. If we can find other minima, we can execute cheaper than VWAP.
Optimal Pre-trade Trajectories: Two Period Trading (III)

The problem can be solved analytically for $\theta = 1/2$ and $\alpha = 1/2$ (the extrema are roots of a quadratic equation). The minimum cost solution is

$$
\rho_1 = 1.57 \tilde{\rho}, \quad \rho_2 = 0.43 \tilde{\rho}, \quad C(\rho_1, \rho_2) = 95.3\% C(\tilde{\rho}).
$$

(24)

For general $\alpha$ the minimum can be found via root searching. We can also optimize for both the acceleration $z$ and the reval time $\theta$.

![Cost relative to VWAP vs. first period acceleration (z) for trading in two equal periods ($\theta = 1/2$).](image)

Cost relative to VWAP vs. first period acceleration ($z$) for trading in two equal periods ($\theta = 1/2$).
Optimal Pre-trade Trajectories: \( M \) Period Trading (I)

**PROBLEM** \((M\) period trading\): Given an order to buy \( X \) shares within time \( T \), while the market is printing trades at constant speed \( \dot{V} \), if we are allowed to change the speed of trading at any intermediate time \( t_i = iT/M, \ i = 1, \ldots, M \), what are the constant participation rates \( \rho_i \) within the corresponding periods \([t_{i-1}, t_i)\) that minimize the cost of trading?

The cost of trading relative to VWAP is again a sum of self-interaction and cross interaction terms

\[
\frac{C[\rho_i]}{C(\bar{\rho})(4/3)\bar{\rho}^{1+\alpha}T^{3/2}} = \sum_{i=1}^{M} \int_{t_{i-1}}^{t_i} \rho_i dt \int_{t_{i-1}}^{t} \rho_i^\alpha (t - s)^{-1/2} ds
\]

\[+ \sum_{i=1}^{M} \sum_{j=1}^{i-1} \int_{t_{i-1}}^{t_i} \rho_i dt \int_{t_{i-1}}^{t_i} \rho_j^\alpha (t - s)^{-1/2}. \tag{25}\]

In terms of the acceleration factors \( z_i = \rho_i/\bar{\rho} \) the relative cost \( C = C[\rho_i]/C(\bar{\rho}) \) becomes

\[
C[z_i] = \left( \frac{1}{M} \right)^{3/2} \sum_{i=1}^{M} z_i^{1+\alpha}
\]

\[
+ \left( \frac{1}{M} \right)^{3/2} \sum_{i=2}^{M} \sum_{j=1}^{i-1} z_i z_j^\alpha \left[ (i - j + 1)^{3/2} + (i - j - 1)^{3/2} - 2(i - j)^{3/2} \right]. \tag{27}\]

Sanity check: \( C[z_i = 1] = 1 \) (shew).
Optimal Pre-trade Trajectories: $M$ Period Trading (II)

The optimization problem becomes

$$\min \{ z_i \} \quad C[z_i] \quad \text{(28)}$$

subject to

$$\sum_{i=1}^{M} z_i = M \quad \text{(29)}$$

and

$$z_i \geq 0. \quad \text{(30)}$$

The problem can be solved with a numerical optimizer. We are interested in the solution as $M$ increases (frequent revals) with fixed $T$. This is plotted in the next figure for $M = 2, 4, 16, 64$.

- Note the front-loading of the trading trajectory. This comes from the decay of the impact, not from risk aversion.
- The strategy prefers to trade a lot early on, then slow down and then clean up the trade.
- Caveat: the analysis assumes that our speed of trading always stays small enough relative to the market (corrections for our size can be added).
- Similar front loading has been derived by Almgren and Chriss (2000) in a linear cost model with risk aversion, and Obizhaeva and Wang (2005) using linear price impact that decays exponentially.
- See also the recent work of Gatheral, Schied and Slynko (2010) for an alternative characterization of temporary impact.
Optimal Pre-trade Trajectories: $M$ Period Trading (III)

Figure 3: Optimal acceleration w.r.t. VWAP for each trading period. The model is square-root decay, square-root impact ($\alpha = 1/2$).

Optimal acceleration relative to VWAP for each trading period.

The model is square-root decay, square-root impact ($\alpha = 1/2$).
Optimal Pre-trade Trajectories: Real Life Trading (I)

Coarse-grained view of an algorithmic trading stack

- Order
  - AlgoChooser
    - Historical Statistics
      - Impact Model
    - Scheduler
      - Historical and Real Time Statistics
        - Impact Model
      - Slicer
        - Trading Signals (Imbalance, Autocorrelation, ...)
      - Router
        - Order Book Statistics
  - Markets
Optimal Pre-trade Trajectories: Real Life Trading (II)

REMARKS:

1. The cost model is used at the AlgoChooser layer as a classifier. Based on the expected cost, we can make decisions on the appropriate algorithmic parameters to use with the order.

2. The cost model is also used by the Scheduler to create a “target” execution path. A risk term and the corresponding risk aversion parameter can be added at this level.

3. The Slicer is allowed to opportunistically deviate from the schedule, based on current short-term alpha forecasts (signals).

4. The Router handles the microstrategy of staying posted or crossing the spread, based on the execution constraints that the Slicer has set for the current period.

5. The feedback loop between the Router and the Slicer is important for dynamically adjusting the liquidity demand, conditional on the current state of the market and the realized cost so far.

Note: As we go down the stack, the frequency of the inputs increases, and the decision time horizon decreases.

Questions:

1. How well did we do? What is the best way to measure post-trade the quality of execution?

2. Is implementation shortfall (arrival price benchmark) the best way to assess performance or should we use a strategy as a benchmark (VWAP, POV at a fixed target rate)?

Standardizing TCA (Transaction Cost Analysis) is an actively debated topic in the industry.
Macro or effective cost models have a role to play in today's electronic trading world.

- They provide convenient high level summaries of execution cost.
- They can be used in parts of the algorithmic trading stack.
- They are calibrated on large samples of orders, so that the net price drift (alpha) averages out.

However,

- By optimizing trading trajectories using such models we are not solving the stochastic control problem of order execution.
- The models are not applicable to small samples or to specific client orders.

We still carry the burden of continually improving the algorithms and understanding postmortem what happened to individual orders.
References