The execution puzzle: How and when to trade to minimize cost

Jim Gatheral

Baruch College
The City University of New York

5th International Financial and Capital Markets Conference
Campos de Jordão, Brazil, August 27th, 2011

Overview of this talk

- What is the execution puzzle?
  - Components of an algorithmic execution
- How much does it cost to trade?
- How to reduce cost
  - Optimal scheduling
  - Choice of market or limit order
  - Optimal order routing
- How much can trading costs be reduced?

Overview of the execution puzzle

Typically, an execution algorithm has three layers:

- The macrotrader
  - This highest level layer decides how to slice the meta-order: when the algorithm should trade, in what size and for roughly how long.
- The microtrader
  - Given a slice of the meta-order to trade (a child order), this layer of the algorithm decides whether to place market or limit orders and at what price level(s).
- The smart order router
  - Given a limit or market order, to which venue should the order be sent?

Our approach to the puzzle

First we will present some stylized facts.

Then, for each of the pieces of the execution puzzle, we will

- Present conventional techniques and the theory underlying them
- Compare theoretical assumptions with empirical evidence
- Suggest modified approaches whenever theory is inconsistent with observation

We will attempt to present a state-of-the-art understanding of order execution, emphasizing intuition and leaving out most of the mathematical details.

References

[Alfonsi, Fruth and Schied] Aurélien Alfonsi, Antje Fruth and Alexander Schied, Optimal execution strategies in limit order books with general shape functions, *Quantitative Finance* **10**(2) 143–157 (2010).

[Almgren and Chriss] Robert Almgren and Neil Chriss, Optimal execution of portfolio transactions, *Journal of Risk* **3** 5–40 (2001).

[Almgren and Harts] Robert Almgren and Bill Harts, A dynamic algorithm for smart order routing, *White paper StreamBase* (2008).

[Bouchaud, Farmer, Lillo] Jean-Philippe Bouchaud, J. Doyne Farmer, and Fabrizio Lillo, How Markets Slowly Digest Changes in Supply and Demand, in *Handbook of Financial Markets: Dynamics and Evolution* 57-156. (2009).

[Gatheral] Jim Gatheral, No-dynamic-arbitrage and market impact, *Quantitative Finance* **10**(7) 749–759 (2010).

[Kearns et al.] Kuzman Ganchev, Michael Kearns, Yuriy Nevmyvaka and Jennifer Wortman Vaughan, Censored exploration and the dark pool problem, *Communications of the ACM* **53**(5) (2010).

[Moro et al.] Esteban Moro, Javier Vicente, Luis G Moyano, Austin Gerig, J. Doyne Farmer, Gabriella Vaglica, Fabrizio Lillo, and Rosario N Mantegna, Market impact and trading prole of hidden orders in stock markets, *Physical Review E* **80**(6) 066102 (2009).

[Obizhaeva and Wang] Anna Obizhaeva and Jiang Wang, Optimal trading strategy and supply/demand dynamics *MIT working paper* (2005).

[Tóth et al.] Bence Tóth, Yves Lempérière, Cyril Deremble, Joachim de Lataillade, Julien Kockelkoren, and Jean-Philippe Bouchaud, Anomalous price impact and the critical nature of liquidity in financial markets, `http://arxiv.org/abs/1105.1694v1` (2011).

Market structure

We will consider a limit order book such as the BOVESPA or the NYSE and two basic order types:

- Market orders
  - Always executed if there is sufficient quantity available.
- Limit orders
  - Are executed only if the limit price is reached.
  - On the BOVESPA, the NYSE and most other limit order book markets, priorities are price first, then time.

Market impact

- The *market impact* function relates expected price change to the volume of a transaction.
- However, there is no reason *a priori* to expect that market impact should be a function of volume only.
- Market impact could be a function of:
  - Market capitalization
  - Bid-ask spread
  - The timescale over which the trade is executed.

- Note that it is not only market orders that impact the market price; limit orders and cancelations should also have market impact.

The square-root formula for market impact

- For many years, traders have used the simple sigma-root-liquidity model described for example by Grinold and Kahn in 1994.
- Software incorporating this model includes:
  - Salomon Brothers, StockFacts Pro since around 1991
  - Barra, Market Impact Model since around 1998
  - Bloomberg, TCA function since 2005
- The model is always of the rough form

$$\Delta P = \text{Spread cost} + \alpha\,\sigma\,\sqrt{\frac{Q}{V}}$$

where σ is daily volatility, V is daily volume, Q is the number of shares to be traded and α is a constant pre-factor of order one.

Heuristic explanation of the square-root formula

- Each trade impacts the stock price and variance adds so impact should be proportional to volatility.
- The amount of each individual impact is proportional to the square root of trade size because risk capital should be proportional to the square-root of the holding period.

Empirical question

We have a simple formula with a heuristic derivation. Does the formula work in practice?

Impact of proprietary metaorders (from Tóth et al.)

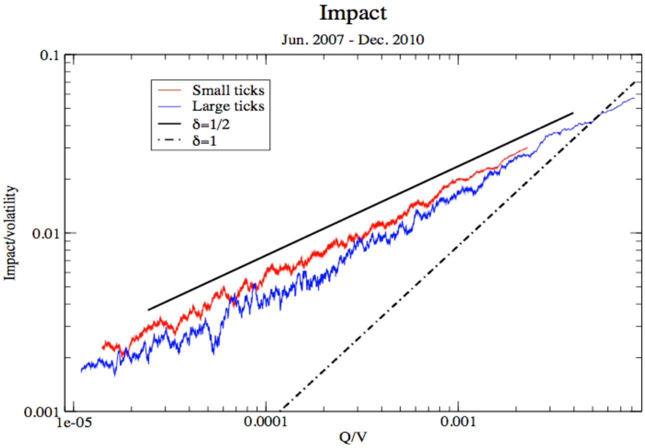


Figure 1: Log-log plot of the volatility-adjusted price impact vs the ratio Q/V

Notes on Figure 1

- In Figure 1 which is taken from [Tóth et al.], we see the impact of metaorders for CFM[1] proprietary trades on futures markets, in the period June 2007 to December 2010.
  - Impact is measured as the average execution shortfall of a meta-order of size Q.
  - The sample studied contained nearly 500,000 trades.

- We see that the square-root market impact formula is verified empirically for meta-orders with a range of sizes spanning two to three orders of magnitude!

[1] Capital Fund Management (CFM) is a large Paris-based hedge fund.

Another explanation for the square-root formula

- In [Tóth et al.], the authors present an argument which says that if latent supply and demand is linear in price over some reasonable range of prices, market impact should be square-root.
- The condition for linearity of supply and demand over a range of prices is simply that submitters of buy and sell meta-orders should be insensitive to price over this range.
  - That seems like an innocuous assumption!

Some implications of the square-root formula

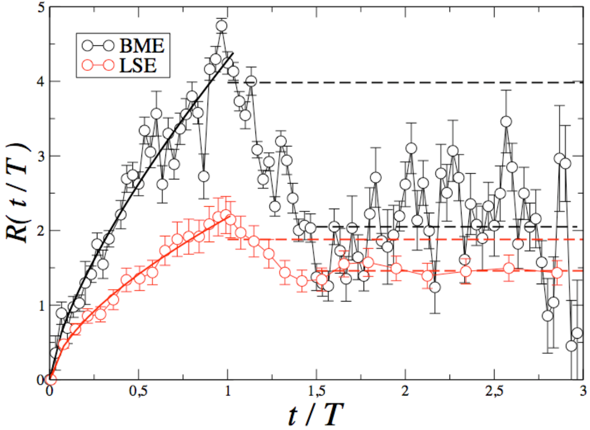
- The square-root formula refers only to the size of the trade relative to daily volume.
- It does not refer to for example:
  - The rate of trading
  - How the trade is executed
  - The capitalization of the stock
- Surely impact must be higher if trading is very aggressive?
  - The database of trades only contains sensible trades with reasonable volume fractions.
  - Were we to look at very aggressive trades, we would indeed find that the square-root formula breaks down.

Empirical question

What happens on average to the stock price while a metaorder is being executed? And what happens to the stock price after completion?

Path of the stock price during execution (from Moro et al.)

Figure 2: Average path of the stock price during execution of a metaorder on two exchanges



Empirically observed stock price path

From Figure 2, we see that

- There is reversion of the stock price after completion of the meta-order.
- Some component of the market impact of the meta-order appears to be permanent.
- The path of the price prior to completion looks like a power law.
  - From [Moro et al.]

$$
\begin{aligned}
m_t - m_0 &\approx (4.28 \pm 0.21)\left(\frac{t}{T}\right)^{0.71\pm0.03} && \text{(BME)} \\
m_t - m_0 &\approx (2.13 \pm 0.05)\left(\frac{t}{T}\right)^{0.62\pm0.02} && \text{(LSE)}
\end{aligned}
$$

where T is the duration of the meta-order.

Summary of empirical observations

- The square-root formula gives an amazingly accurate rough estimate of the cost of executing an order.
- During execution of a meta-order, the price moves on average roughly according to $(t/T)^{2/3}$.
- Immediately after completion of a meta-order, the price begins to revert.

Summary of empirical observations

- According to a literal reading of the square-root formula, the cost of trading doesn't depend on trading strategy.
- Does this mean that there is nothing that we can do to reduce trading costs?

No! We can reduce the size of the prefactor α by breaking down the execution puzzle into its components and attacking each one in turn.

- We now turn our attention to optimal scheduling.

Statement of the optimal scheduling problem

- Given a model for the evolution of the stock price, we would like to find an optimal strategy for trading stock, the strategy that minimizes some cost function over all permissible strategies.
- A *static* strategy is one determined in advance of trading.
- A *dynamic* strategy is one that depends on the state of the market during execution of the order, *i.e.* on the stock price.
  - Delta-hedging is an example of a dynamic strategy. VWAP is an example of a static strategy.
- It turns out, surprisingly, that in many models, a statically optimal strategy is also dynamically optimal.
- For all the models we will describe, a static strategy set in advance of trading is optimal.

Almgren and Chriss

- The seminal paper of [Almgren and Chriss] treats the execution of a hidden order as a tradeoff between risk and execution cost.
- According to their formulation:
  - The faster an order is executed, the higher the execution cost
  - The faster an order is executed, the lower the risk (which is increasing in position size).
- The Almgren-Chriss model can be considered the conventional market standard.

Almgren and Chriss

In the Almgren and Chriss model, the stock price S_t evolves as

$$dS_t = \gamma\, dx_t + \sigma\, dZ_t$$

and the price $\tilde{S}_t$ at which transactions occur is given by

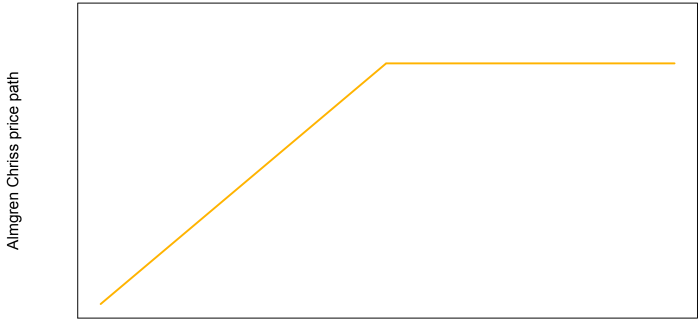
$$\tilde{S}_t = S_t + \eta\, v_t$$

where $v_t := -\dot{x}_t$ is the rate of trading.

- Temporary market impact is proportional to the rate of trading v_t.
  - Temporary market impact decays instantaneously and has no effect on the market price S_t.
- The expectation $\mathbb{E}[S_t] = \gamma\, x_t$ of the market price during execution is linear in the position x_t. It does not revert after completion.

Price path in the Almgren and Chriss model

Figure 3: The Almgren and Chriss average price path is plotted in orange.



- The optimal strategy (with no price of risk) is just VWAP – constant trading in volume time.

Obizhaeva and Wang 2005

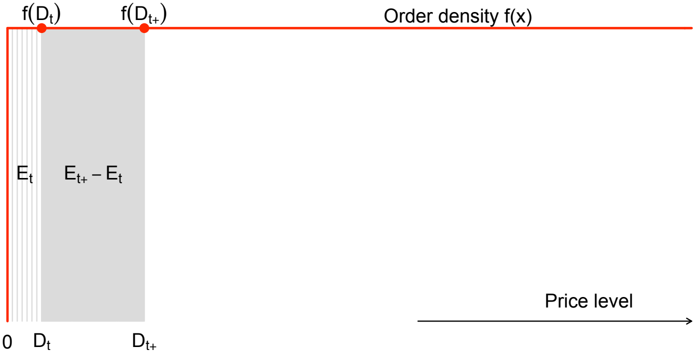
In the Obizhaeva and Wang model,

$$S_t = S_0 + \eta \int_0^t v_s \, e^{-\rho\,(t-s)} \, ds + \int_0^t \sigma \, dZ_s \tag{1}$$

with $v_t = -\dot{x}_t$.

- Market impact is linear in the rate of trading but in contrast to Almgren and Chriss, market impact decays exponentially with some non-zero half-life.

Obizhaeva Wang order book process



When a trade of size ξ is placed at time t,

$$\text{the volume impact process } E_t \quad \mapsto \quad E_{t+} = E_t + \xi$$

$$\text{the spread } D_t = \eta\, E_t \quad \mapsto \quad D_{t+} = \eta\, E_{t+} = \eta\, (E_t + \xi)$$

Exponential resiliency

- The volume impact process E_t reverts exponentially to zero. Thus, if there were no trades in the interval $(t, t + \Delta]$, we would have

$$E_{t+\Delta} = E_t \, e^{-\rho \, \Delta}$$

  - Alternatively, the spread D_t reverts to zero.
- This is referred to as *exponential resiliency of the order book.*

Optimal strategy in the Obizhaeva Wang model

The optimal strategy in the OW model is

$$v_s = \delta(s) + \rho + \delta(s - T)$$

where $\delta(\cdot)$ is the Dirac delta function.

- The optimal strategy consists of a block trade at time $t = 0$, continuous trading at the rate ρ over the interval $(0, T)$ and another block trade at time $t = T$.

Example of OW optimal strategy

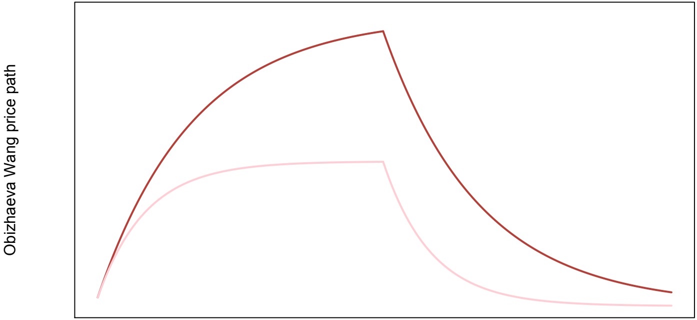
- Consider a Brazilian stock with 14,000 trades per day and a liquidation whose horizon is 1 hour.
  - A rule of thumb is that the order book refreshes after 10-15 trades. So we take the half-life of the order book resilience process to be $20 \times \log 2 \approx 14$ trades.
  - $\log 2/\rho = \log 2 \times 20$ trades so $\rho = 1/20$ in trade time. One hour has 2,000 trades so $\rho\, T = 100$.
- Recall that the optimal strategy is $u_s = \delta(s) + \rho + \delta(s - T)$. Thus,

$$X = \int_0^T v_s\, ds = 2 + \rho\, T = 102$$

- The optimal strategy thus consists of a block trade of relative size one at the beginning, another trade of size one at the end and an interval VWAP of relative size 100.
  - The optimal strategy in this case is very close to VWAP.

Price path in the Obizhaeva-Wang model

Figure 4: The OW average price path is plotted for two different values of ρ.



- The optimal strategy is bucket-like.

The model of Alfonsi, Fruth and Schied

[Alfonsi, Fruth and Schied] consider the following (AS) generalization of the OW model:

- There is a continuous (in general nonlinear) density of orders $f(x)$ above some martingale ask price A_t. The cumulative density of orders up to price level x is given by

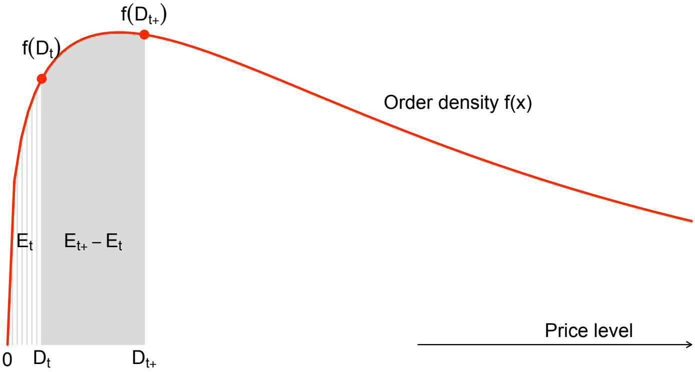
$$F(x) := \int_0^x f(y)\, dy$$

- Executions eat into the order book.
- A purchase of ξ shares at time t causes the ask price to increase from $A_t + D_t$ to $A_t + D_{t+}$ with

$$\xi = \int_{D_t}^{D_{t+}} f(x)\, dx = F(D_{t+}) - F(D_t)$$

- The order book has exponential resiliency; either the volume impact process E_t or the spread D_t revert exponentially.

Schematic of the model



When a trade of size ξ is placed at time t,

$$\begin{array}{rcl} E_t & \mapsto & E_{t+} = E_t + \xi \\ D_t = F^{-1}(E_t) & \mapsto & D_{t+} = F^{-1}(E_{t+}) = F^{-1}(E_t + \xi) \end{array}$$

Optimal liquidation strategy in the AS model

The optimal strategy in the AS model is

$$v_t = \xi_0 \, \delta(t) + \xi_0 \, \rho + \xi_T \delta(T - t).$$

- Just as in the OW model, the optimal strategy consists of a block trade at time $t = 0$, continuous trading at the rate ρ over the interval $(0, T)$ and another block trade at time $t = T$.
- The only difference is that in the AS model, the final block is not the same size as the initial block.

Generalization

- It can be shown that the bucket-shaped strategy is optimal under more general conditions than exponential resiliency.
  - Specifically, if resiliency is a function of the volume impact process E_t (or equivalently the spread D_t) only, the optimal strategy has block trades at inception and completion and continuous trading at a constant rate in-between.
- These conditions may appear quite general but in fact, there are many other models that do not satisfy them.

A transient market impact model

The price process assumed in [Gatheral] is

$$S_t = S_0 + \int_0^t f(v_s)\, G(t-s)\, ds + \text{ noise} \tag{2}$$

- The instantaneous impact of a trade at time s is given by $f(v_s)$ – some function of the rate of trading.
- A proportion $G(t-s)$ of this initial impact is still felt at time $t > s$.

The square-root model

Consider the following special case of (2) with $f(v) = \frac{3}{4}\sigma\sqrt{v/V}$ and $G(\tau) = 1/\sqrt{\tau}$:

$$S_t = S_0 + \frac{3}{4}\,\sigma \int_0^t \sqrt{\frac{v_s}{V}}\,\frac{ds}{\sqrt{t-s}} + \text{ noise} \tag{3}$$

which we will call the *square-root process*.

It turns out that the square-root process is consistent with the square-root formula for market impact:

$$\frac{\mathcal{C}}{X} = \sigma\sqrt{\frac{X}{V}} \tag{4}$$

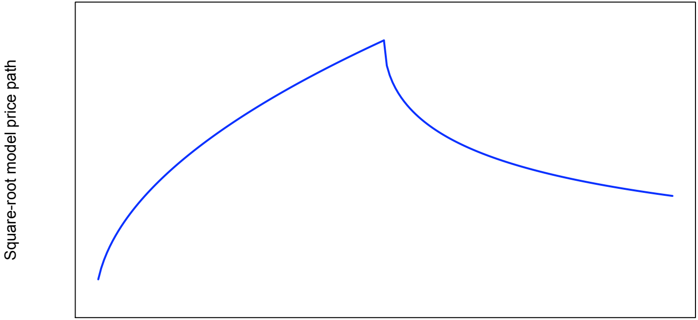
- Of course, that doesn't mean that the square-root process is the true underlying process!

The optimal strategy under the square-root process

- Because $f(\cdot)$ is concave, an optimal strategy does not exist in this case.
  - It is possible to drive the expected cost of trading to zero by increasing the number of slices and decreasing the duration of each slice.
  - To be more realistic, $f(v)$ must be convex for large v and in this case, an optimal strategy does exist that involves trading in bursts, usually more than two.

Price path in the Square-root model

Figure 5: The square-root model average price path.



- The optimal strategy does not exist in this model.

Intuition

- The optimal strategy depends on modeling assumptions.
- In the Almgren-Chriss model where there is no price reversion after completion of the meta-order, the optimal strategy is a simple VWAP.
- In other models where there is reversion, the optimal strategy is to make big trades separated in time, perhaps with some small component of continuous trading.

The intuition is easy to see:

The price reversion idea

If the price is expected to revert after completion, stop trading early and start again later after the price has reverted!

The market or limit order decision

- Having decided how to slice the meta-order, should we send market or limit orders?
- Many market participants believe that market orders should only be sent when absolutely necessary – for example when time has run out.
- Conventional wisdom has it that the more aggressive an algorithm is, the more costly it should be.
  - This cannot be true on average. Traders are continuously monitoring whether to send market or limit orders so in equilibrium, market and limit orders must have the same expected cost.
    - Market orders incur an immediate cost of the half-spread but limit orders suffer from adverse selection.

Adverse selection

- Limit orders are subject to adverse selection:
  - If the price is moving towards us, we get filled. We would rather that our order had not been filled. Had we not got the fill, we could have got a better price.
  - If the price is moving away from us, we don't get filled. We need to resubmit at a worse price.
- In general, we regret sending a market order because we have to pay the half-spread.
- In general, we regret sending a limit order because of adverse selection.

The order book signal

- If we know the price is going against us, we should send a market order. Otherwise we should send a limit order.
- In practice, we cannot predict the future; we can compute relative probabilities of future events.
  - One simple idea is to look at the shape of the order book. If there are more bids than offers, the price is more likely to increase than decrease.

The SFGK zero-intelligence model

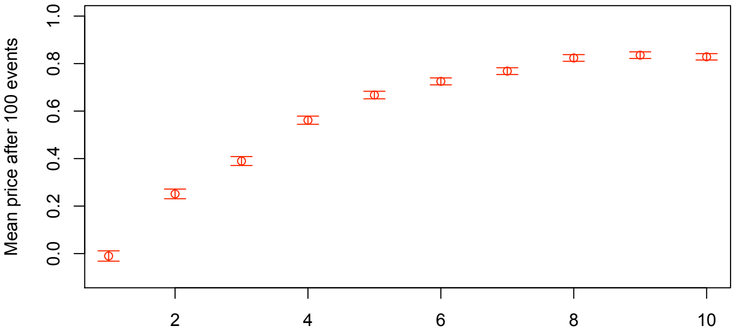
In this model due to Smith, Farmer, Gillemot and Krishnamurthy:

- Limit orders can be placed at any integer price level p where $-\infty < p < \infty$.
    - If worried about negative prices, think of these as being logarithms of the actual price.
- Limit sell orders may be placed at any level greater than the best bid $B(t)$ at time t and limit buy orders at any level less than the best offer $A(t)$.
- Market orders arrive randomly at rate μ.
- Limit orders (per price level) arrive at rate α.
- A proportion δ of existing limit orders is canceled.

Price signal in the ZI simulation

- Even in the ZI model, the shape of the order book allows prediction of price movements.
  - Traders really would need to have zero intelligence not to condition on book shape!

Figure 6: With one share at best offer, future price change vs size at best bid.



Microprice

- The relationship between the imbalance in the order book and future price movements is sometimes described in terms of the *microprice*.
  - This can be thought of as a fair price, usually between the bid and the ask.
- As an example, in the context of the zero-intelligence model, Cont and Larrard derived the following asymptotic expression:

Proposition 2 of Cont & Larrard

The probability $\phi(n,p)$ that the next price is an increase, conditioned on having n orders at the bid and p orders at the ask is:

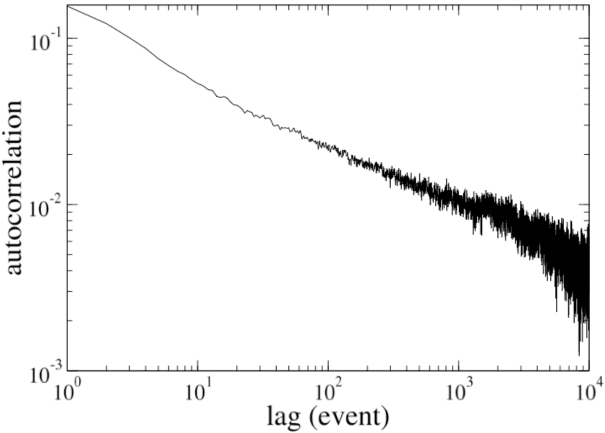
$$\phi(n,p) = \frac{1}{\pi}\int_0^{\pi} dt \left(2 - \cos t - \sqrt{(2-\cos t)^2 - 1}\right)^p \frac{\sin n\,t\,\cos\frac{t}{2}}{\sin\frac{t}{2}}$$

Order splitting

- The typical size of a meta-order is a large multiple of the quantity typically available at the best quote.
- Consequently, meta-orders need to be split into smaller child orders.
- This gives rise to a long-memory autocorrelation function which is significant at all lags.
- Order sign (or order flow) is thus highly predictable.

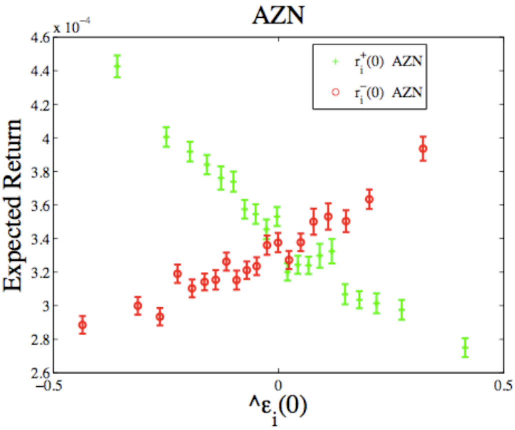
Autocorrelation of order signs

Figure 7: Autocorrelation function of the time series of signs of Vodafone market orders in the period May 2000 – December 2002, a total of 580,000 events (from [Bouchaud, Farmer, Lillo])



Market impact relates to unexpected order flow

Figure 8: Average impact of AZN market orders vs expected order sign. Buy orders in green; sell orders in red (from [Bouchaud, Farmer, Lillo])



An explanation for price reversion due to Lillo et al.

- The longer a meta-order has been active, the more likely it is that it will continue.
    - If the meta-order does continue, the marginal market impact of child orders will decrease.
    - If the meta-order stops (because it completes), market impact will be large.
    - On average, price moves are unpredictable.

The market/ limit order decision

- We have two (related) signals:
  - From the current state of the order book, we can predict the sign of the next price change.
  - From recent order flow history, we can judge whether active meta-orders are more on the buy side or on the sell side.
- The practical recipe is therefore to:
  - Send market orders when the market is going against us, limit orders otherwise.
  - Trade more when others want to trade with us, less when there are fewer counterparties.

Order routing

- Having optimally scheduled child orders and for each such child order, having decided whether to send a market or a limit order, where should the order be sent?
- In Brazil, the answer is straightforward; there is currently only one market – the BOVESPA.
- In the US, there are currently approximately 13 lit venues and over 40 dark venues.
  - On the one hand, it would be prohibitively complicated and expensive to route to all of them.
  - On the other hand, by not routing to a particular venue, the trader misses out on potential liquidity, and all things being equal, will cause incur greater market impact.
- Most traders have to use a smart order routing (SOR) algorithm provided by a dealer.

Smart order routing (SOR)

- The goal of an SOR algorithm is to buy (or sell) as many shares as possible in the shortest time by optimally allocating orders across both lit and dark venues.
  - In the case of lit venues, there are hidden orders so there is typically more liquidity available than is displayed.
  - In dark venues, by definition, all liquidity is hidden.
- We will briefly describe
  - A heuristic algorithm due to Almgren and Harts
  - An algorithm based on machine learning techniques due to Michael Kearns and his collaborators.

The Almgren and Harts (AH) algorithm

- The idea of this algorithm is that the more hidden quantity is detected in a given venue, the more hidden quantity there is likely to be.
  - This is a characteristic of distributions with fatter tails than exponential.
  - Empirically, we find that order sizes are power-law distributed in which case this assumption would definitely be justified.
- For simplicity, let's focus on the sale of stock.
- If hidden quantity w is detected (by selling more than the visible quantity) on a particular venue, the current estimate of hidden liquidity is increased by w.
- If no hidden quantity quantity is detected on a venue, the existing estimate is decremented by a factor ρ.

Conditional distribution of quantity: Power-law case

Suppose that the distribution of order sizes Q is power-law so that

$$\Pr(Q > n) = \frac{C}{n^{\alpha}}$$

Assuming the conditional probability that hidden quantity is greater than n given that n slices have already been observed is given by

$$\begin{aligned}
\Pr(Q \geq (n+1) | Q \geq n) &= \frac{\Pr(Q \geq (n+1))}{\Pr(Q \geq n)} \\
&= \left(\frac{n}{n+1}\right)^{\alpha} \\
&\to 1 \text{ as } n \to \infty
\end{aligned}$$

If the distribution of Q is power-law, the more quantity you observe, the more likely it is that there is more quantity remaining.

A simplified version of the Almgren-Harts (AH) algorithm

Our goal is to execute a sell order as quickly as possible by optimally allocating quantity to all N venues.

- We allocate quantity quasi-greedily sequentially to the venue with the highest estimated quantity, visible and hidden.
- If we see a fill of size n_j when the displayed quantity is q_j on the jth venue, the pre-existing liquidity estimate R_j is decayed by a factor ρ and incremented by the detected hidden liquidity:

$$R_j \mapsto \rho \, R_j + (n_j - q_j)^+$$

- Repeat until our quantity is exhausted and the order is completed.

The Kearns et al. (GKNW) algorithm

- The idea behind the GNKW is not dissimilar to the idea behind the AH algorithm although, as written, it is applied only to dark pools.
- In the *allocation* phase, orders are allocated greedily to the venue with the greatest estimated liquidity.
- In the *re-estimation* phase, parametric order-size distributions are updated.
  - They find that the most practical approach is to estimate separately the probability that the quantity is zero and the exponent of a power-law for the probabilities of nonzero quantities.
- Allocation and re-estimation are performed in a continuous loop.

Simulation results

- An algorithm can only be tested by experiment or simulation.
  - The data used for model estimation comes from particular choices of algorithm and we can't predict what would have been if these algorithms had chosen to act differently.
- In simulations, the GKNW algorithm outperformed two other obvious choices of algorithm:
  - Equal allocation across venues
  - A *bandit* algorithm that begins with equal weights. If there is any execution at a particular venue, that venues weight is increased by a factor $\alpha = 1.05$.

Scope for further improvement

- We have presented two conceptually similar smart order routing algorithms.
    - Both algorithms implicitly assume that all trading venues offer the same quality of execution.
    - But different venues have different latencies, at least one respect in which not all venues can offer the same overall cost.
    - We need to think about incorporating execution quality into routing decisions.

A philosophical question

Question

Why should it be so complicated to trade stock?

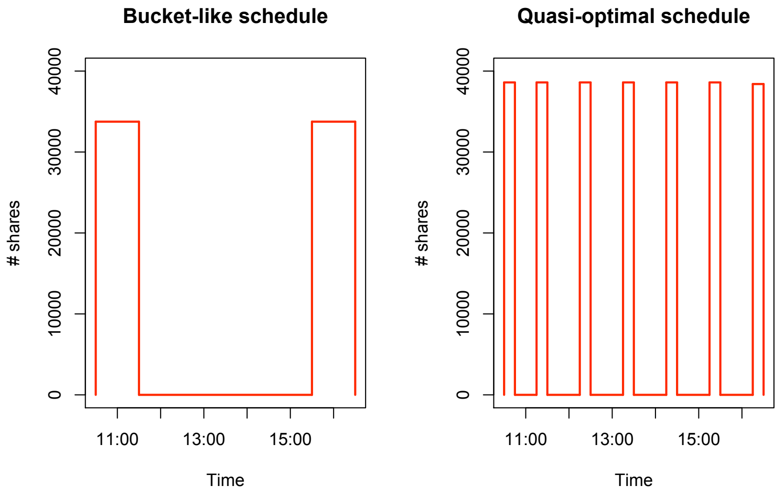
One answer goes something like this:

- Changes in market structure together with technological innovation have massively reduced trading costs.
- Nevertheless, some market participants achieve significantly lower costs.
  - This requires either substantial investment in technology and trading expertise or
  - careful selection of broker algorithms.
- Note however that algorithm performance is very hard to assess ex-post. Ideally, randomized experiments are required.

Potential cost savings from optimal scheduling

- To estimate potential savings from optimal scheduling, assume that the square-root process (3) is correct and consider a one-day order to sell 270,000 shares of Vale SA.
  - Daily volatility is assumed to be 2% and daily volume to be 3 million shares.
  - We consider liquidation starting at 10:30 and ending at 16:30 with child orders lasting 15 minutes.
- Because we are not confident in the square-root model for high volume fractions, we constrain volume fraction to be no greater than 25%.
- We compare the costs of VWAP, a two-slice bucket-like strategy and a quasi-optimal strategy that consists of seven roughly equal slices.

Stock trading schedules



Comparison of results

In the square-root model (3), the cost of a VWAP execution is given exactly by the square-root formula:

$$\sigma \sqrt{\frac{Q}{V}} = 0.02 \times \sqrt{\frac{270}{3000}} = 0.02 \times 0.3 = 60 \text{ bp}$$

Table 1: Cost comparison

Strategy	Cost	Saving
VWAP	60.0 bp	
Bucket-like	49.6 bp	17%
Quasi-optimal	40.8 bp	32%

Potential cost savings from microtrader improvements

- Were we just to blindly send market orders, we would estimate that the cost of each child order would be around a half-spread.
- Practical experience shows that it is not possible to reduce this cost much below one third of a spread.
- We conclude that we could potentially save up to one sixth of the spread.

Potential cost savings from smart order routing

- A naïve estimate would be to use the square-root market impact formula, changing the denominator V to reflect the potential increase in liquidity from routing to extra venues.
  - If the potential liquidity accessed is doubled, costs should be decreased by $1 - 1/\sqrt{2} \approx 30\%$ according to this simple computation!
  - This could be one explanation for the multiplicity of trading venues in the US.
- We expect actual savings to be much less than this because the different venues are all connected and information leaks from one venue to the other.

Final conclusion

- Recent empirical and theoretical work on market microstructure has led to a much improved understanding of how to trade optimally.
- Potential savings from careful order execution relative to VWAP using a simple first-generation algorithm are substantial.
- Cost savings of 25% for reasonably sized orders are not unreasonable.