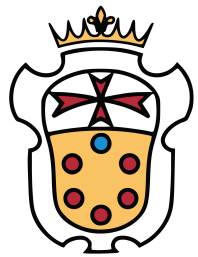
Market Impact, Slippage Costs, and Optimal Execution of Large Trades

Fabrizio Lillo

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



SCUOLA
NORMALE
SUPERIORE

Understanding the Diversity of Financial Risk - Budapest, November 26, 2021

What is market impact?

- Market impact refers to the "correlation" between an incoming order (to buy or to sell) and the contemporaneous and subsequent price change.
- Market impact induces extra costs. Indeed, large volumes must typically be fragmented and executed incrementally (Kyle 1985). The total cost of this large trade is quickly dominated, as sizes become large, by the average price impact.
- Monitoring and controlling impact has therefore become one of the most active domains of research in quantitative finance since the mid-nineties.
- Volume dependence of impact (By how much do larger trades impact prices more than smaller trades?), and temporal behavior of impact (is the impact of a trade immediate and permanent, or does the impact decay after one stops trading?).
- Impact is a dynamical quantity since it depends on the available liquidity, but also on the recent history of my trades.

Market impact of metaorders

- We are not interested here in market impact of individual market events (e.g. a market order or a limit order), but rather to a **metaorder**, i.e. a sequence of orders and trades following a single trading decision.
- This is the main quantity of interest in reality.
- Public market data do not allow to identify metaorders.
- The empirical analysis of market impact of metaorders is a guide to construct realistic models of market impact on which optimal execution problem can be built.
- I first review some old and new statistical regularities of market impact of metaorders and then I discuss the optimal execution problem.

A very recent review

Quantitative Finance > Trading and Market Microstructure

[Submitted on 2 May 2021]

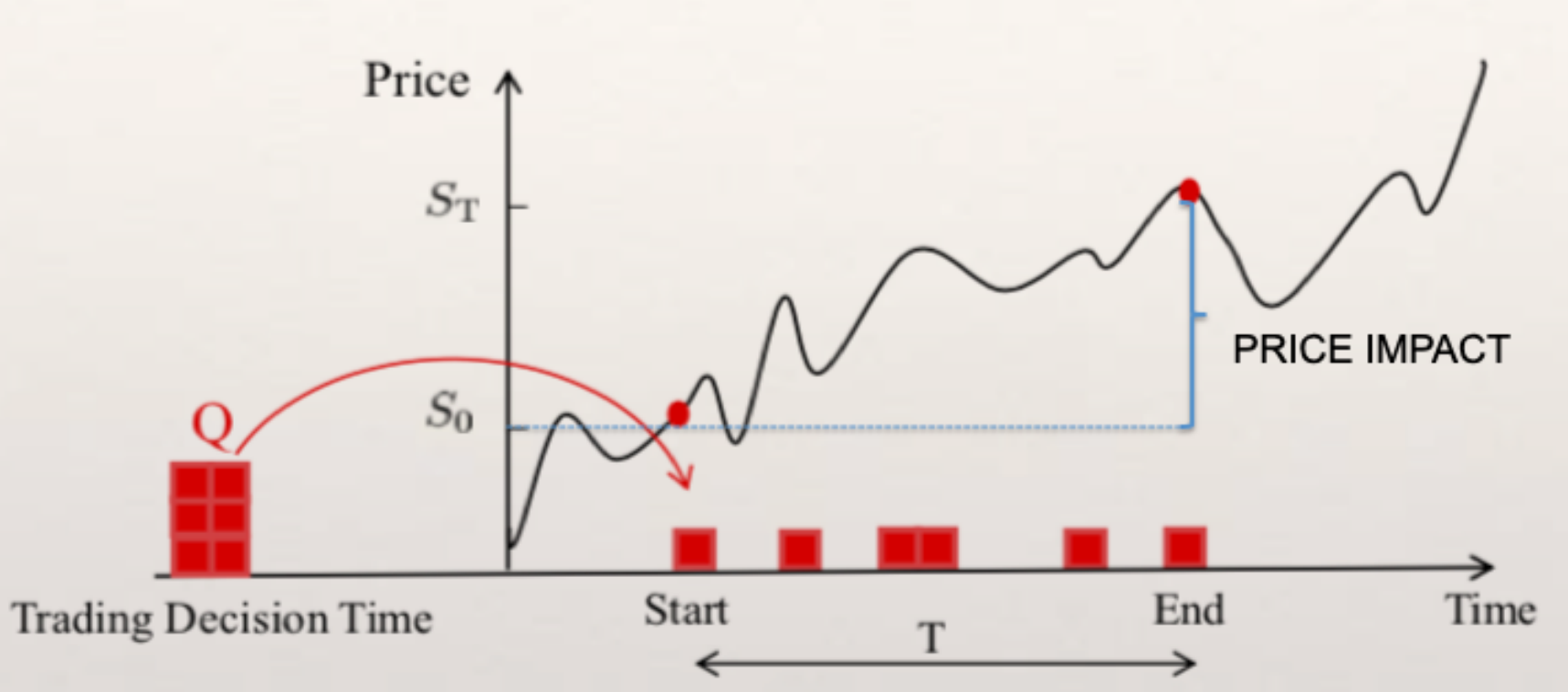
Order flow and price formation

Fabrizio Lillo

I present an overview of some recent advancements on the empirical analysis and theoretical modeling of the process of price formation in financial markets as the result of the arrival of orders in a limit order book exchange. After discussing critically the possible modeling approaches and the observed stylized facts of order flow, I consider in detail market impact and transaction cost of trades executed incrementally over an extended period of time, by comparing model predictions and recent extensive empirical results. I also discuss how the simultaneous presence of many algorithmic trading executions affects the quality and cost of trading.

Comments: 24 pages. To appear in "Machine Learning in Financial Markets" (A. Capponi and C.A Lehalle, editors), Cambridge University Press
Subjects: **Trading and Market Microstructure (q-fin.TR)**
Cite as: arXiv:2105.00521 **[q-fin.TR]**
(or arXiv:2105.00521v1 **[q-fin.TR]** for this version)

$$IS(\$) = \underbrace{Q(S_D - S_0)}_{\text{Delay Cost}} + \underbrace{\sum x_j S_j - \sum x_j S_0}_{\text{Trading Cost}} + \underbrace{(Q - \sum x_j)(S_T - S_0)}_{\text{Opportunity Cost}} + \text{Visible Cost}$$



Definitions and relation with cost

The main quantity of interest is the *metaorder impact*

$$\mathcal{I}(Q, T) \equiv \mathbb{E}[\epsilon \Delta \log p | Q, T]$$

where

- $\Delta \log p$ is the logprice change between the start and the end of the metaorder,
- Q is the size of the metaorder (in shares),
- T is the metaorder duration (in seconds or in volume time)
- ϵ is the sign of the metaorder (i.e $\epsilon = +1$ for a buy and $\epsilon = -1$ for a sell metaorder).

$\mathcal{I}(Q, T)$ is directly related to the average impact cost of a metaorder execution.The expected implementation shortfall cost, i.e. the difference between the expected cost and the theoretical cost obtained by marking to market the trade with the initial price, is

$$Cost = \int_0^T \dot{x}_t \mathcal{I}(x_t, t)\ dt$$

where $\dot{x}_t$ is the time derivative of the asset position x_t at time t.

The square-root impact law

- Remarkably, many empirical studies seem to agree on the validity of the **square-root impact law**, obtained when conditioning on the volume fraction of the metaorder.
- Setting $\phi = Q/V_d$ with σ_d and V_d daily volatility and volume

$$\mathcal{I}(Q, T) \approx Y \sigma_d \sqrt{\phi}$$

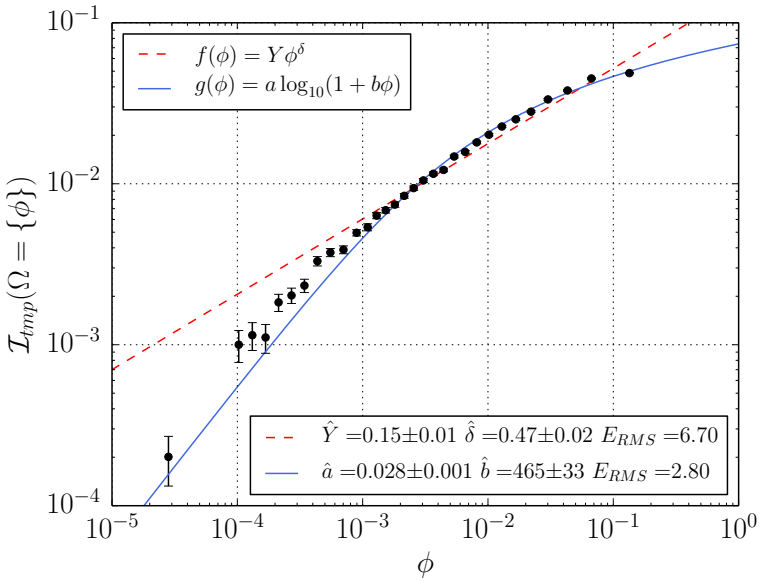
where $Y \simeq 1$ is a numerical constant. This relation has been empirically shown also for disparate asset classes as options and Bitcoin.

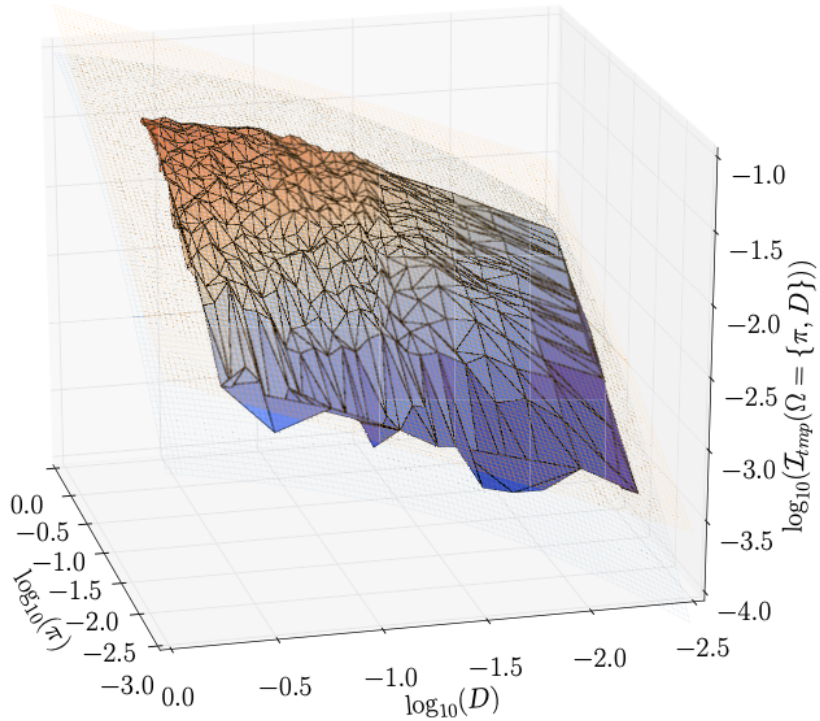
- This empirical relation is at first sight surprising: it indicates that the style of trading (for example using limit orders or market orders), the duration T of the execution, the trading speed (i.e. the number of shares traded per unit time), etc, are not relevant!
- These observations indicate that there must be some limitations to the validity of this 'law'. For example, the prefactor Y might depend on the trading algorithm.

The Ancerno database

- ANcerno Ltd (formerly the Abel Noser Corporation) is a widely recognized consulting firm that works with institutional investors to monitor their equity trading costs.
- Each metaorder is characterized by a broker and an investor label, a stock symbol, the total volume $|Q|$ (in number of shares) and the times at the start and at the end of its execution with sign $\epsilon = \pm 1$.
- Data from $\sim$ 3000 US equities in 2007-2010
- Around 8 million metaorders distributed quite uniformly in time and across market capitalizations representing around 5-10% of the total market volume.
- No information on the execution style or trading profile during execution.

Empirical results[1]





Market impact of metaorders under different conditioning: $\phi = Q/V_d$, D is the (volume time) duration of the metaorder, and $\pi = \phi/D$ is the participation rate.

[1] E. Zarinelli, M.Treccani, J. D. Farmer, F. Lillo, Beyond the square root: Evidence for logarithmic dependence of market impact on size and participation rate, *Market Microstructure and Liquidity* (2015)

- By considering $\mathcal{I}(Q, T)$ as a function only of $\phi = Q/V$, it is clear that a logarithmic function fits the data better than a power law function; this indicates a linear behavior of impact for small volumes and an extra concavity (likely due to a selection bias) for very large volumes. Below we will present two possible explanations for the linear behavior of the impact for small ϕ.

- By considering $\mathcal{I}(Q, T)$ as a function only of $\phi = Q/V$, it is clear that a logarithmic function fits the data better than a power law function; this indicates a linear behavior of impact for small volumes and an extra concavity (likely due to a selection bias) for very large volumes. Below we will present two possible explanations for the linear behavior of the impact for small ϕ.
- By considering $\mathcal{I}(Q, T)$ as a function of both variables, Zarinelli *et al.* introduces the **market impact surface** and showed that a double logarithmic function outperforms the power law form of Eq. 1.

- By considering $\mathcal{I}(Q, T)$ as a function only of $\phi = Q/V$, it is clear that a logarithmic function fits the data better than a power law function; this indicates a linear behavior of impact for small volumes and an extra concavity (likely due to a selection bias) for very large volumes. Below we will present two possible explanations for the linear behavior of the impact for small ϕ.
- By considering $\mathcal{I}(Q, T)$ as a function of both variables, Zarinelli *et al.* introduces the **market impact surface** and showed that a double logarithmic function outperforms the power law form of Eq. 1.
- Considering a power law dependence on T and the participation rate η, Zarinelli *et al.* investigates the regression

$$\mathcal{I}(Q, T) = A \; T^{\delta_T} \eta^{\delta_\eta} \cdot \textit{noise} \tag{1}$$

The fit gives $\delta_T = 0.54 \pm 0.01$ and $\delta_\eta = 0.52 \pm 0.01$, and $A = 0.207 \pm 0.005$. The fact that both exponents are very close to $1/2$ indicates that $\mathcal{I}(Q, T) \approx \sqrt{T\eta} = \sqrt{\phi}$, even when considering **separately** the effect of participation rate and duration.

Important comments

Important comments

- Measuring the impact implies measuring the **drift**, thus the inclusion of the sign ϵ in the definition is critical. By neglecting the sign, for example taking the absolute value of the impact, one measures the correlation between volume and (a proxy of) volatility.

Important comments

- Measuring the impact implies measuring the **drift**, thus the inclusion of the sign ϵ in the definition is critical. By neglecting the sign, for example taking the absolute value of the impact, one measures the correlation between volume and (a proxy of) volatility.
- Fluctuations are very large. What is plotted in the previous figure is the *average* market impact across a very large number of metaorders.

Important comments

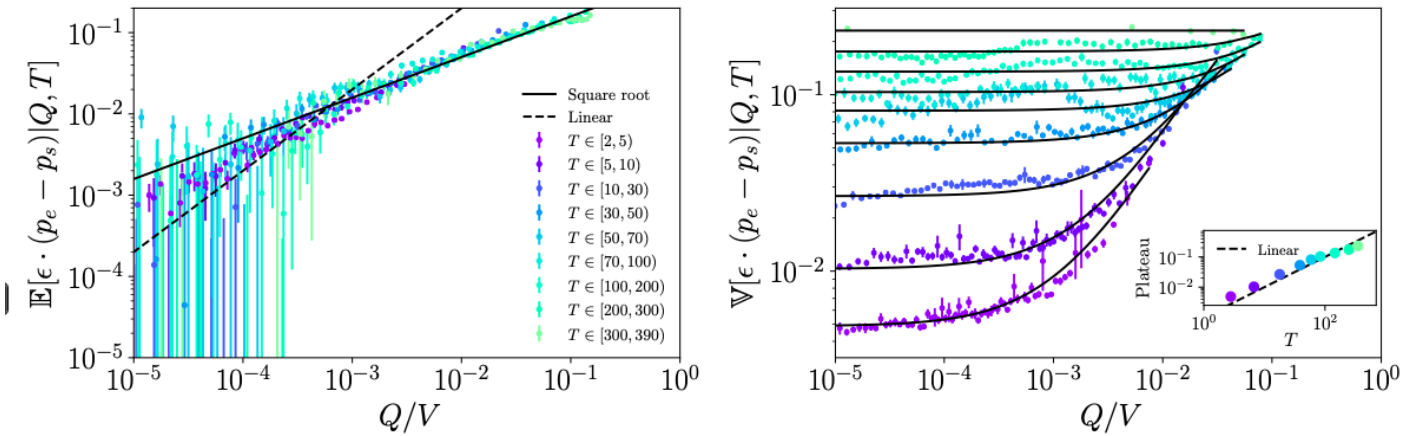
- Measuring the impact implies measuring the **drift**, thus the inclusion of the sign ϵ in the definition is critical. By neglecting the sign, for example taking the absolute value of the impact, one measures the correlation between volume and (a proxy of) volatility.
- Fluctuations are very large. What is plotted in the previous figure is the *average* market impact across a very large number of metaorders.
- The error bars are standard errors, which are small because of the sample size. However the average impact is always significantly different from zero.

Important comments

- Measuring the impact implies measuring the **drift**, thus the inclusion of the sign ϵ in the definition is critical. By neglecting the sign, for example taking the absolute value of the impact, one measures the correlation between volume and (a proxy of) volatility.
- Fluctuations are very large. What is plotted in the previous figure is the *average* market impact across a very large number of metaorders.
- The error bars are standard errors, which are small because of the sample size. However the average impact is always significantly different from zero.
- Conditioning on the metaorder is critical: other types of market impact are obtained conditioning on the (anonymous) flow of orders in the market. Obviously in these latter conditions one can obtain much higher R^2 (even 1, since the order flow completely determines price).

Impact is not just volatility[2]

The square root impact law is not related to the fact that volatility scales as the square root of (execution) time, which, for a fixed participation rate, is proportional to metaorder size.

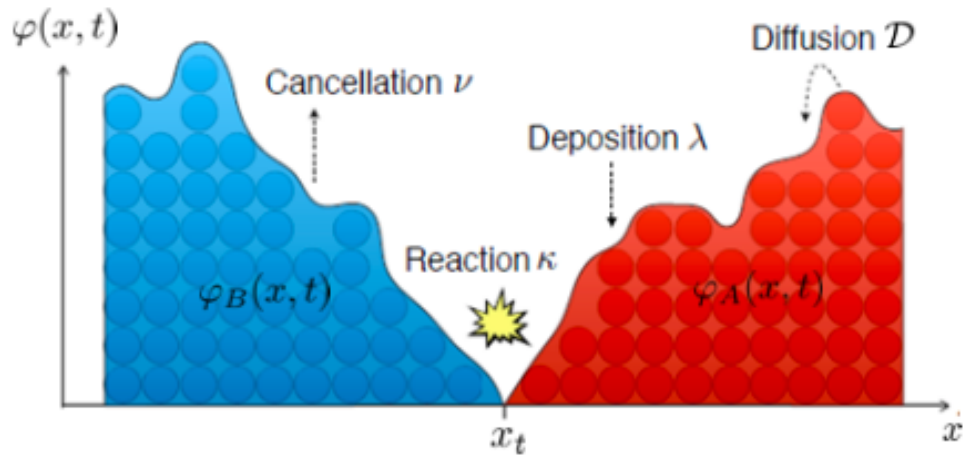


Market impact curves of metaorders with $\phi \gtrsim 5 \cdot 10^{-4}$ (roughly 80% of those in the ANcerno database) are independent on T and consistent with a square root dependence on ϕ.
Impact of the remaining small metaorders are better described by a linear relation.
The variance of impact depends linearly on T, as expected by the diffusivity of price, and this price uncertainty largely exceeds the average reaction impact contribution (which in turn explains why the R^2 in the market impact estimation is typically very small).

[2] Bucci et al. Impact is not just volatility. Quantitative Finance, 19(11):1763-1766 (2019)

Explaining the impact law: the Locally Linear Order Book (LLOB)[3]

Limit order book model: $\varphi(x,t)$ is the density of orders at price x at time t



$$\partial_t \varphi_A(x,t) = \mathcal{D}\partial_{xx}\varphi_A(x,t) - \nu\varphi_A(x,t) + \lambda\Theta(x_t - x) - R_{A,B}(x,t)$$
$$\partial_t \varphi_B(x,t) = \mathcal{D}\partial_{xx}\varphi_B(x,t) - \nu\varphi_B(x,t) + \lambda\Theta(x - x_t) - R_{A,B}(x,t)$$

where $R_{A,B}(x,t) = \kappa\varphi_A(x,t)\varphi_B(x,t)$.

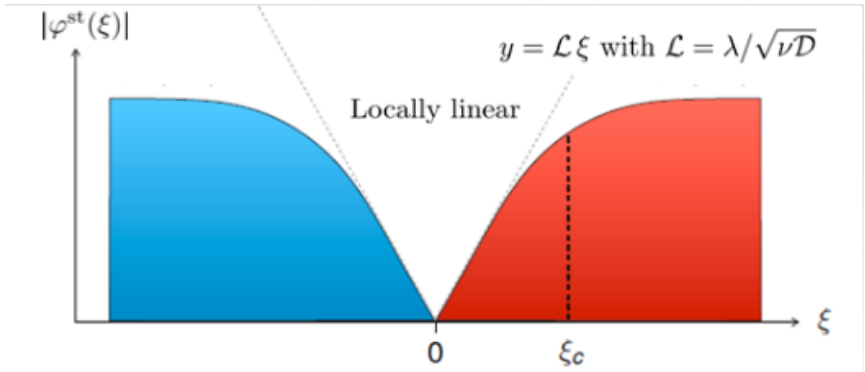
Price equation for the transaction price x_t

$$\varphi_A(x_t,t) = \varphi_B(x_t,t)$$

[3] Donier et al. (2015)

Stationary solution and metaorder solution

The stationary solution in the price reference frame is linear when x is small, i.e. $\varphi_{st}(x) = \mathcal{L}x$ where $\mathcal{L} = \lambda/\sqrt{\mathcal{D}\nu}$ is a measure of liquidity.



The total transaction rate J is the flux of orders through the origin, i.e. $J \equiv \mathcal{D}\partial_x \varphi_{st}|_{x=0} = \mathcal{D}\mathcal{L}$.

In the limit of a slow order book (i.e. $\nu T \ll 1$), the price trajectory $p_m(t)$ during the execution of the metaorder with trading velocity $m = Q/T$ is given by the self-consistent expression

$$p_m(t) = p_0(t) + y(t), \tag{2}$$

$$y(t) = \frac{m}{\mathcal{L}} \int_0^t \frac{\mathrm{d}s}{\sqrt{4\pi\mathcal{D}(t-s)}} \exp\left[-\frac{(y(t)-y(s))^2}{4\mathcal{D}(t-s)}\right], \tag{3}$$

where $p_0(t)$ is the price trajectory in the absence of the metaorder in $t \in [0, T]$.

Impact of a metaorder according to LLOB

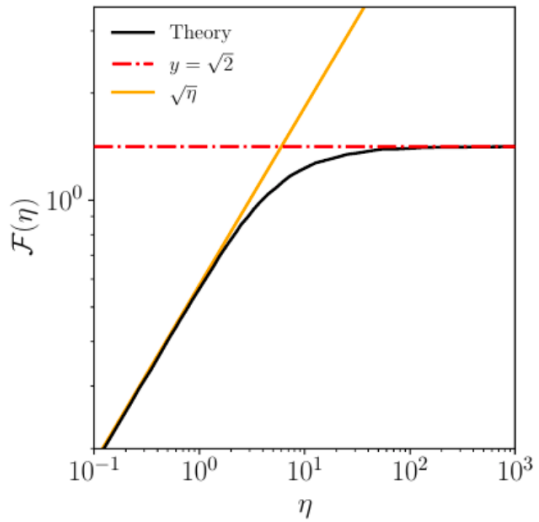
Price impact of a metaorder is $\mathcal{I}(Q,T) = y(T)$ and equal to

$$\mathcal{I}(Q,T) = \sqrt{\frac{\mathcal{D}Q}{J}}\,\mathcal{F}(\eta)\ , \quad \text{with} \quad \eta \equiv \frac{Q}{JT}\ , \tag{4}$$

where η is the participation rate

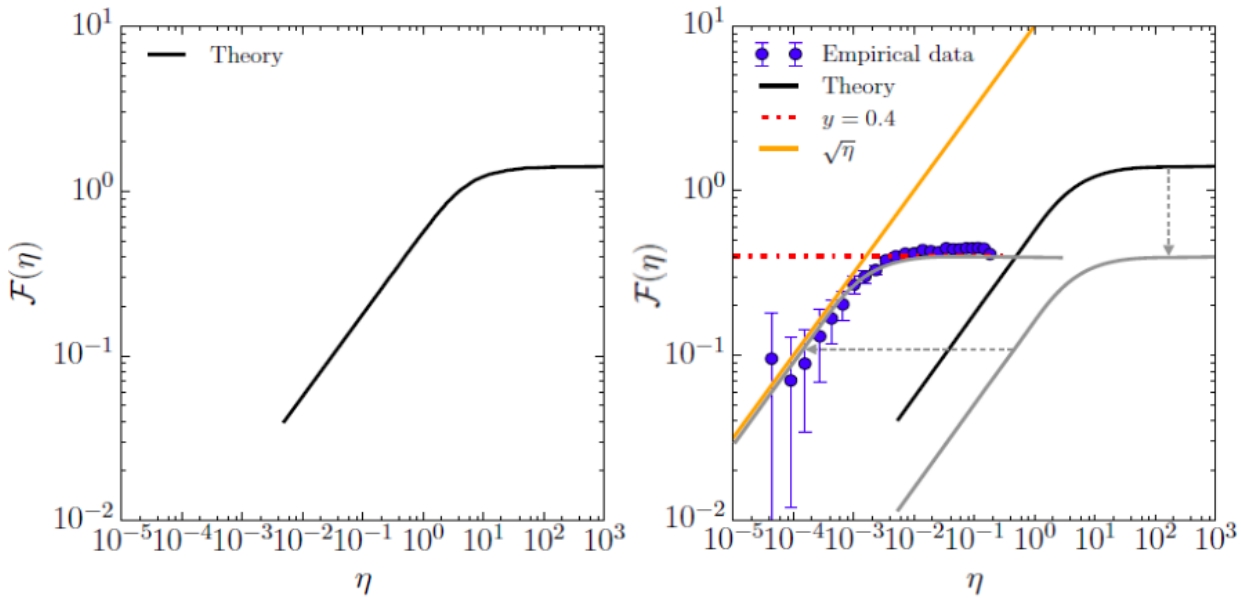
$$\mathcal{F}(\eta) = \begin{cases} \sqrt{\eta/\pi} & \text{for} \quad \eta \ll 1 \\ \sqrt{2} & \text{for} \quad \eta \gg 1 \end{cases}$$

$\mathcal{I}(Q,T)$ is linear in Q for small Q at fixed T, and crosses over to a square-root for large Q. In the square-root regime, impact is predicted to be independent of the execution time T, as observed empirically



LLOB theory vs Empirical Data

Data from ANcerno database: 8 million metaorders in US equity markets (2007-2010)



Good qualitative agreement but not quantitative...

Intuition: The total market turnover J is actually dominated by HFTs/market makers, while resistance to slow metaorders can only be provided by slow participants

Solution: Introduce the LLOB mode with two time-scales of market participants, i.e. *fast* and *slow*.

Market impact with fast and slow traders[4]

Two contributions to the order flow, with $\nu_s \ll \nu_f$:

$$\begin{aligned}
\partial_t \phi_s(x,t) &= \mathcal{D}_s \partial_{xx} \varphi_s(x,t) - \nu_s \varphi_s(x,t) + \lambda_s \Theta(x - x_t) + m_{s,t}\delta(x - x_t) \\
\partial_t \phi_f(x,t) &= \mathcal{D}_f \partial_{xx} \varphi_f(x,t) - \nu_f \varphi_f(x,t) + \lambda_f \Theta(x - x_t) + m_{f,t}\delta(x - x_t)
\end{aligned}$$

and $m_{f,t} + m_{s,t} = m_0$. The model can be exactly solved for $T > T^\dagger$ where

$$\begin{aligned}
T^\dagger &= \nu_f^{-1} \eta^{*-2} \mathcal{D}_s / \mathcal{D}_f \\
\eta^* &= J_s / J_f
\end{aligned}$$

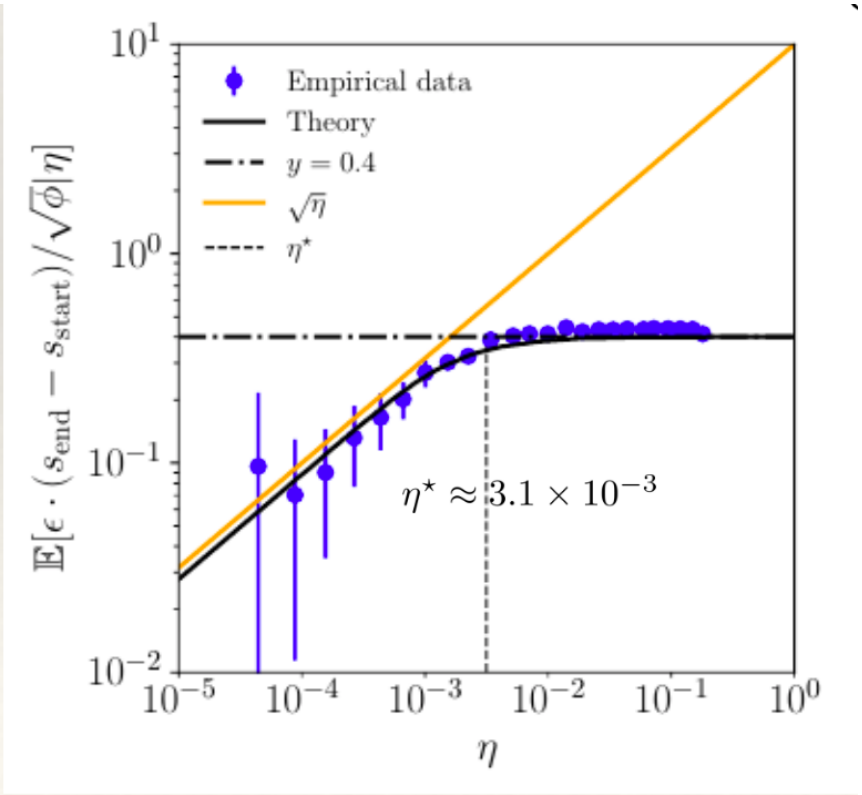
and in this case

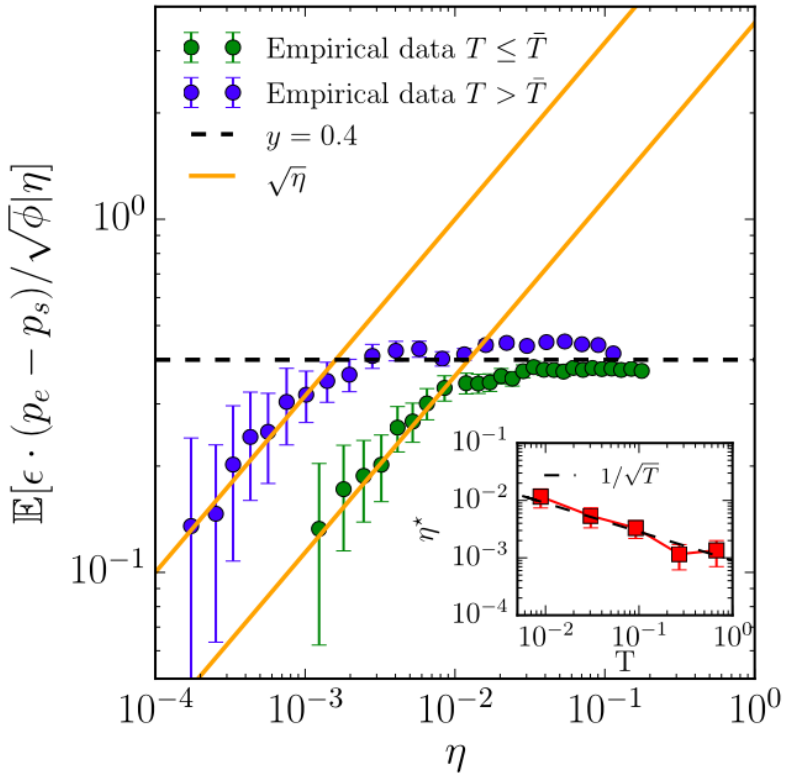
$$\mathcal{I}(Q,T) = \sqrt{\frac{\mathcal{D}_s Q}{J_s}} \mathcal{F}\left(\frac{\eta}{\eta^*}\right)$$

Empirically we estimate $T^\dagger \sim 45$ seconds (for $\nu_f = 1$ second). Since the median execution time of the metaorders in our sample is 35 minutes, it follows that $T > T^\dagger$.

[4] Benzaquen and Bouchaud (2018)

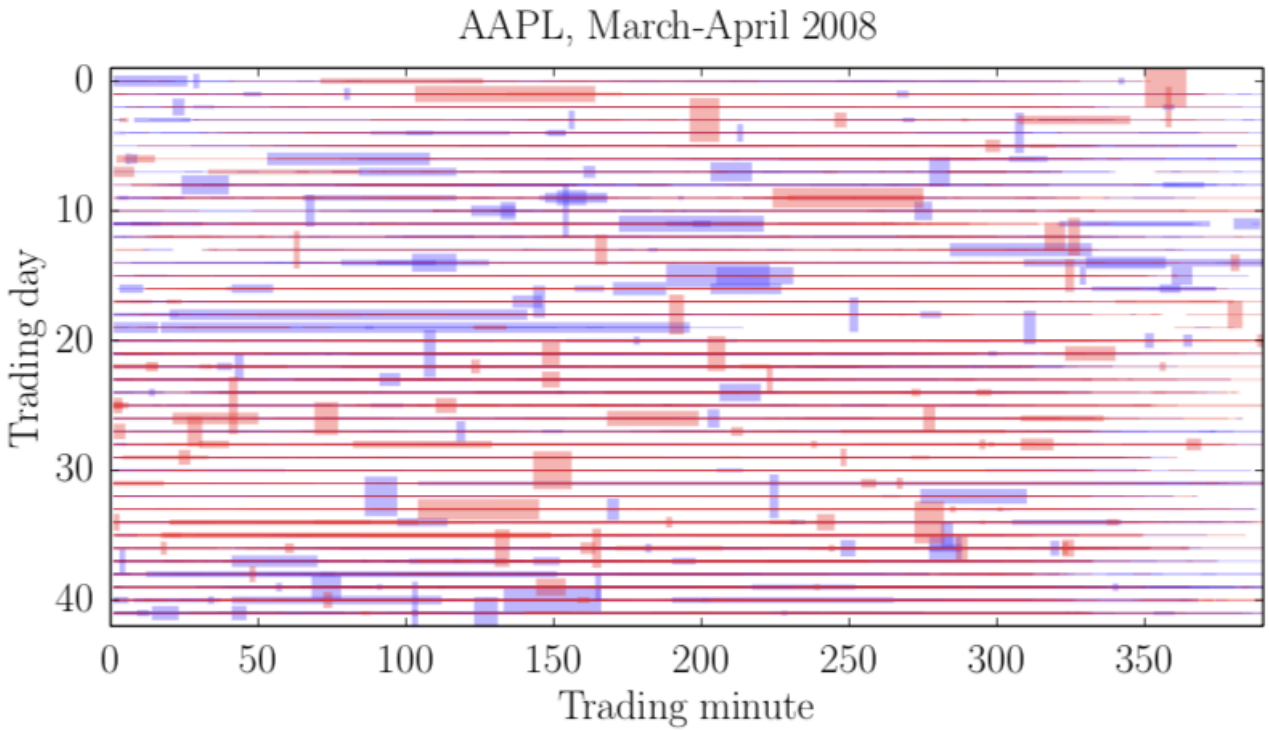
Empirical results





Trading is crowded....

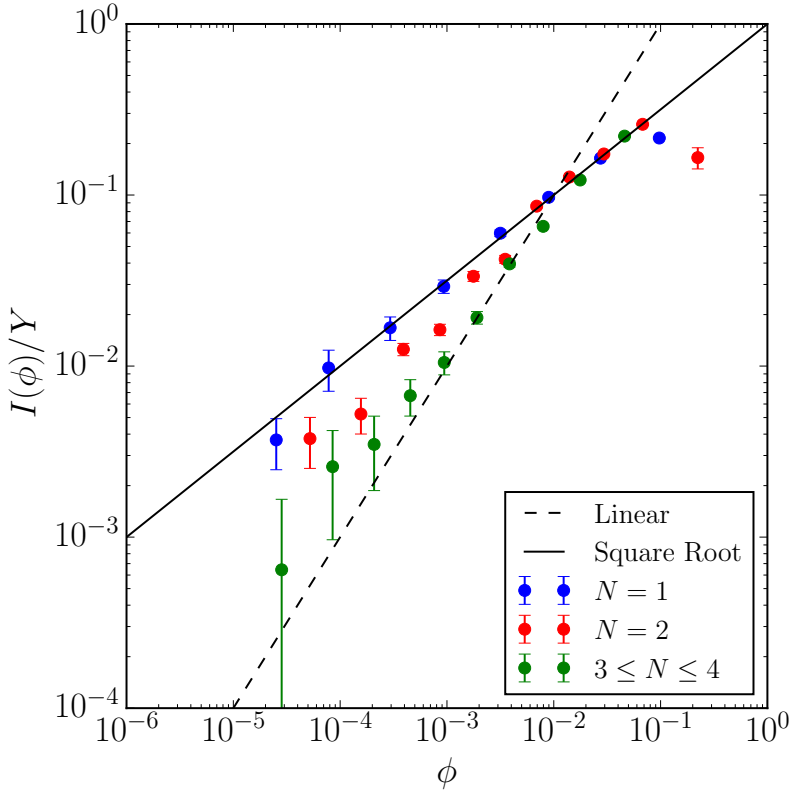
A snapshot of the market

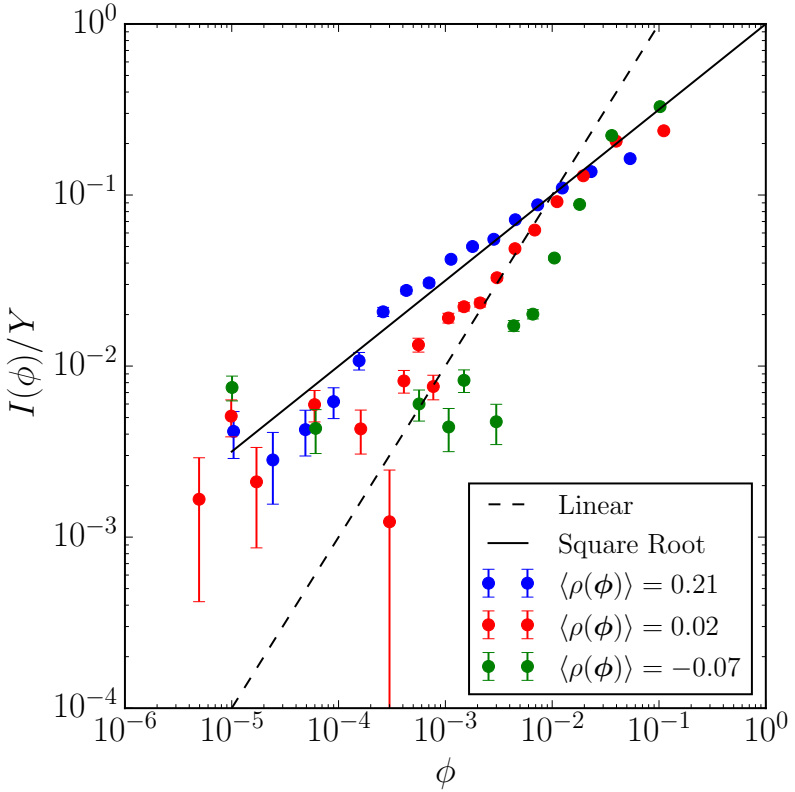


Time series of metaorders active on the market for AAPL in the period March-April 2008. Buy (Sell) metaorders are depicted in blue (red). The thickness of the line is proportional to the metaorder participation rate. More metaorders in the same instant of time give rise to darker colours. Each horizontal line is a trading day. We observe very few blanks, meaning that there is almost always an active metaorder from our database, which is of course only a subset of the number of orders that are active in the market.

Co-impact: the role of investors crowding

Empirical data from Ancerno database





Co-impact: the role of investors crowding

What is the role of the other metaorders being executed in the same day? Data shows[5] that the market reacts to the aggregated order flow

$$\mathcal{I}(\phi_1, ..., \phi_N) \equiv \mathbb{E}[s_{cl} - s_{op} | \phi_1, ..., \phi_N] = Y \text{sign}(\Phi) \sqrt{|\Phi|}$$

where $\Phi = \phi_1 + ... + \phi_N$

This is somewhat puzzling since, conditioning on one metaorder ϕ while the rest of the market trades ϕ_m

$$I(\phi + \phi_m) = Y \times \sqrt{|\phi + \phi_m|}. \tag{5}$$

This tends to $Y\sqrt{|\phi_m|}$ when $\phi \to 0$, behaves linearly when $|\phi| \ll |\phi_m|$ and as a square root when $|\phi| \gg |\phi_m|$.

[5] F. Bucci, I. Mastromatteo, Z. Eisler, F. Lillo, J.-P. Bouchaud, C.-A. Lehalle, Co-impact: Crowding effects in institutional trading activity, Quantitative Finance (2020)

Suppose the manager k wants to execute a volume fraction $\phi_k = \phi$ and there are N metaorders today. Since the other $N-1$ metaorders are not known, her best estimate of the average impact given N is given by conditional expectation

$$I_N(\phi) = \mathbb{E}[\mathcal{I}(\Phi)|\phi_k = \phi] = \mathbb{E}\Big[\mathcal{I}\Big(\phi_k + \sum_{i\neq k}^{N} \phi_i\Big)\Big|\phi_k = \phi\Big] \tag{6}$$

over the conditional distribution $P(\boldsymbol{\varphi}_N|\phi_k = \phi)$ of the metaorders.

Since the number of metaorders is in general not known either, the expected individual market impact is given by

$$I(\phi) = Y \times \sum_N p(N) \int \mathrm{d}\phi_1 \ldots \mathrm{d}\phi_N P(\boldsymbol{\varphi}_N|\phi_k = \phi)\mathrm{sign}\Big(\phi_k + \sum_{i\neq k}^{N} \phi_i\Big)\Big(\phi_k + \sum_{i\neq k}^{N} \phi_i\Big)^{1/2} \tag{7}$$

In order to compute $I_N(\phi)$ and $I(\phi)$ we need to know the joint probability density function $P(\boldsymbol{\varphi}_N) := P(\phi_1, \ldots, \phi_N)$

Example: $P(\varphi_N)$ is a multivariate Gaussian

- For independent Gaussian metaorders with with zero mean and variance Σ_N^2
  - For small metaorders the noise term dominates, leading to

$$I_N(\phi) \propto \phi \quad \text{when} \quad \phi \ll \phi_N^* := \Sigma_N \sqrt{N-1}.$$

  - For large metaorders the $N-1$ other simultaneous metaorders can be neglected and thus

$$I_N(\phi) \propto \sqrt{\phi} \quad \text{when} \quad \phi \gg \phi_N^*.$$

- For Gaussian metaorders with zero mean, variance Σ_N^2, and correlation ρ_N, the average impact $I_N(\phi)$ can be obtained by making the substitution

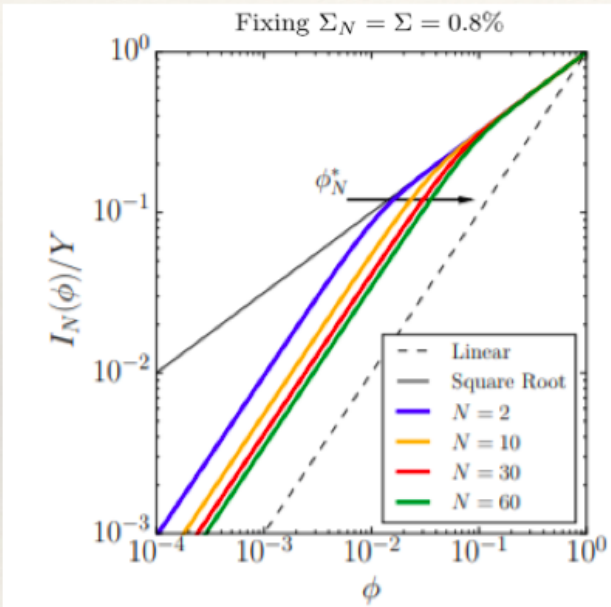
$$\phi \to \phi(1 + (N-1)\rho_N). \tag{8}$$

in the expression of $I_N(\phi)$ for independent Gaussians. This is because $(N-1)\rho_N\phi$ is the effective number of additional volume-weighted metaorders correlated to the original one.

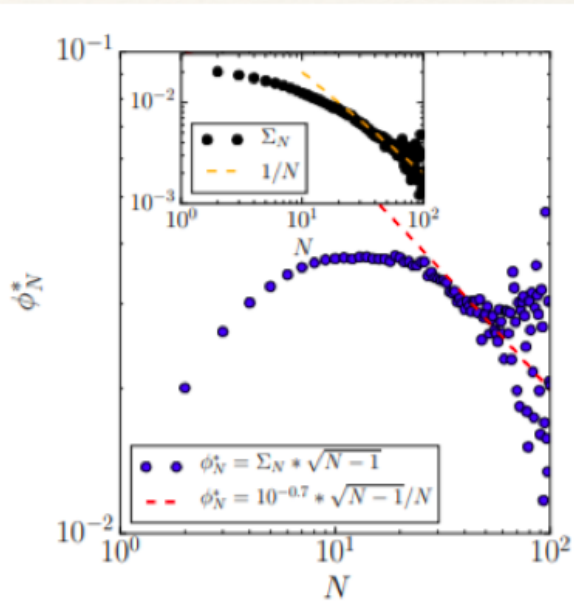
From linear to square root

Left. full analytical solutions for different N, but fixed $\Sigma_N = \Sigma$.
However, interestingly, one expects Σ_N to decrease with N, simply because as the number of metaorders increases, the volume fraction represented by each of them must decrease. (Inset right)

Analitical solutions for different N:



From empirical data:



For $N \to \infty$ we find that $\phi^\star_N = \Sigma_N \sqrt{N-1} \propto \sqrt{N-1}/N \to 0$, i.e. Square Root Law is recovered.

$$I_N(\phi) \simeq \begin{cases} \phi^{\bullet 1/2} & \text{for} \quad \phi \gg \phi^\star_N \\ \phi & \text{for} \quad \phi \ll \phi^\star_N \end{cases}$$

A calibrated model

- Empirically metaorder sizes are almost independent, while signs are correlated. Moreover there is not a huge disparity between metaorder sizes.
- We thus assume that the joint distribution of the ϕ_i's can be written as

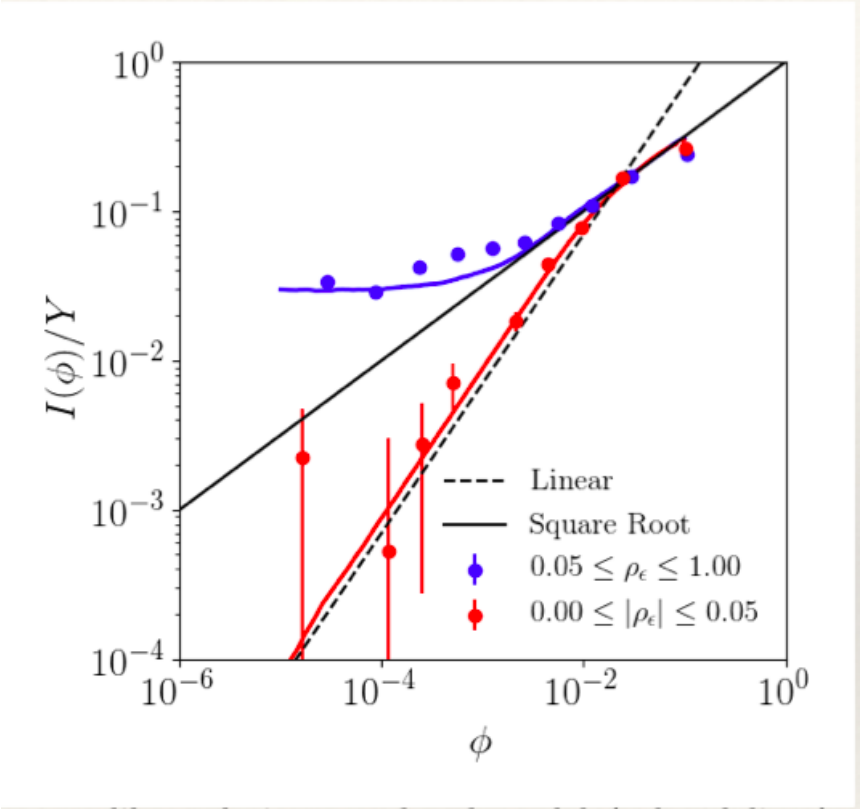
$$P(\boldsymbol{\varphi}_N) = \mathcal{P}_N(\boldsymbol{\epsilon}) \prod_{i=1}^{N} p(|\phi_i|), \tag{9}$$

- We further assume that there is a unique common factor determining the sign of the metaorders:

$$\mathbb{P}(\epsilon_i = +1|\tilde{\epsilon}) = \frac{1}{2}(1 + \gamma_\epsilon \tilde{\epsilon}); \qquad \mathbb{P}(\epsilon_i = -1|\tilde{\epsilon}) = \frac{1}{2}(1 - \gamma_\epsilon \tilde{\epsilon}), \tag{10}$$

where $\tilde{\epsilon}$ is the hidden sign factor, such that $\mathbb{P}(\tilde{\epsilon} = \pm 1) = 1/2$, and γ_ϵ is the sign correlation between each sign ϵ_i and the hidden sign factor $\tilde{\epsilon}$.

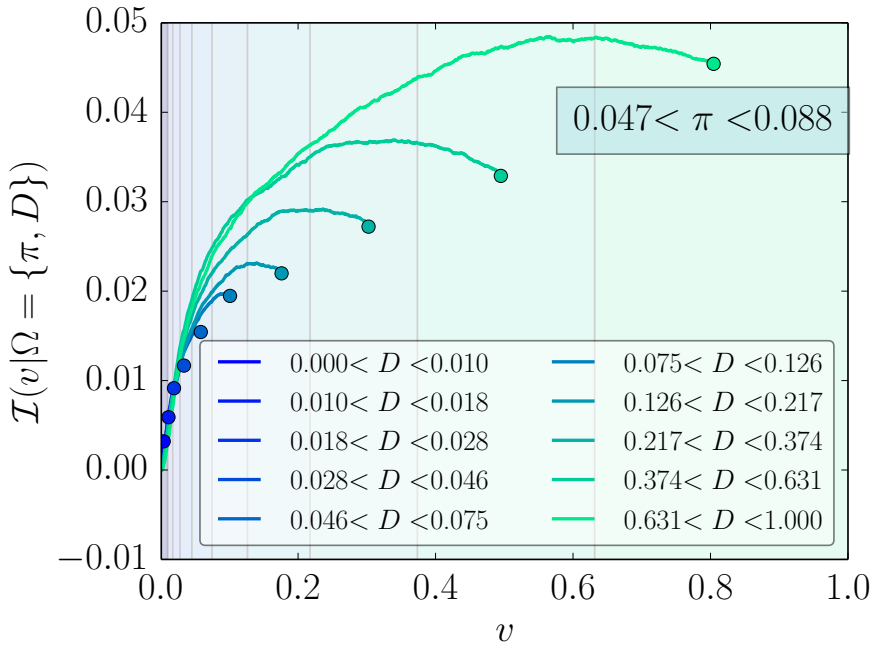
Comparison with data



Comparison between calibrated sign-correlated model (colored lines) and empirical data (circles): The sample is split into two sub-samples depending a realized sign correlation

$$\rho(\boldsymbol{\epsilon}) := \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} \epsilon_i \epsilon_j$$

The price dynamics during the metaorder execution

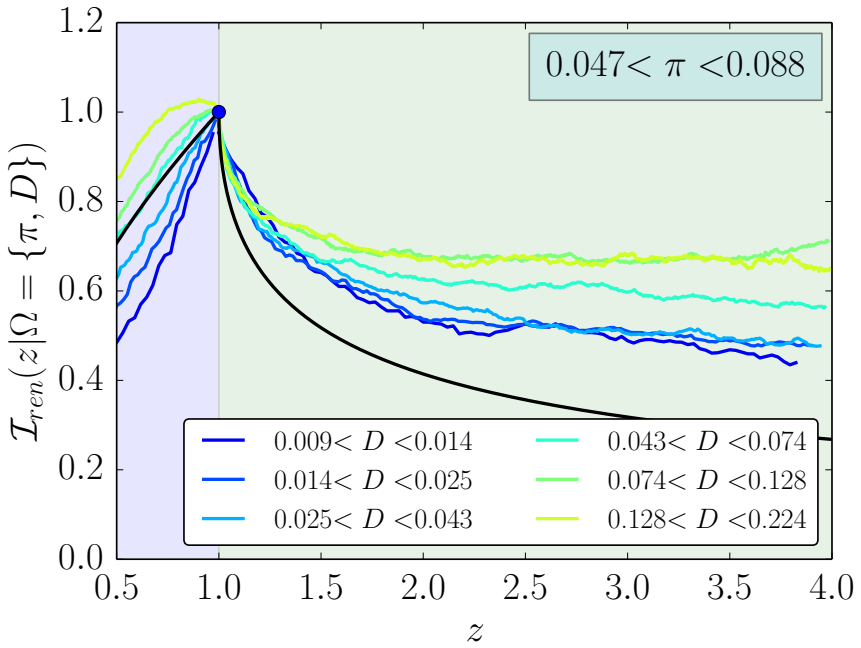


Average price trajectory during the execution of a metaorder[6]. v is volume time, D is the metaorder duration, and π the participation rate.

The price starts reverting before the end of the metaorder execution

[6] E. Zarinelli, M.Treccani, J. D. Farmer, F. Lillo, Beyond the square root: Evidence for logarithmic dependence of market impact on size and participation rate, *Market Microstructure and Liquidity* (2015)

The price dynamics after the metaorder execution



Average price trajectory after the execution of a metaorder[7]. D is the metaorder duration and π the participation rate. z is time rescaled to the metaorder duration (thus the metaorder ends at $z = 1$).

Price reverts significantly after the end of the metaorder execution

[7] E. Zarinelli, M.Treccani, J. D. Farmer, F. Lillo, Beyond the square root: Evidence for logarithmic dependence of market impact on size and participation rate, *Market Microstructure and Liquidity* (2015)

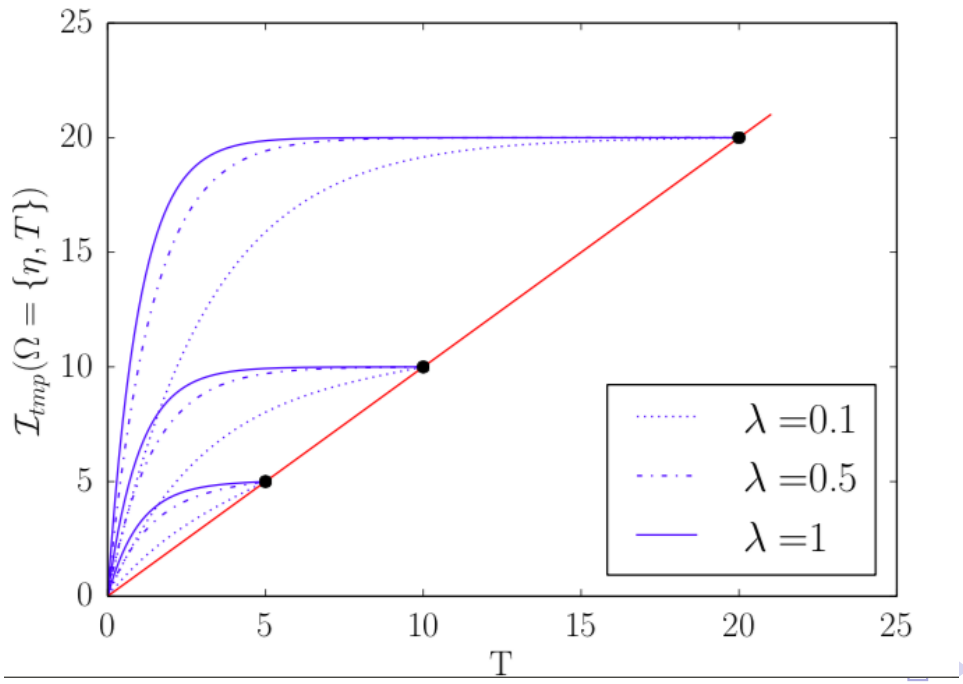
Which model can explain these phenomena?

Under Almgren-Chriss type model

$$S_t = S_0 + k \int_0^t \dot{x}_s ds + \sigma \int_0^t dW_s$$

where $\dot{x}_t \ dt > 0$ is the amount of shares purchased by the considered execution in $[t, t + dt]$, the optimal solution for a metaorder of size x_0 of an agent with risk aversion λ is

$$\dot{x}_t = x_0 \frac{\sinh b(T - t)}{\sinh bT} \qquad b = \sqrt{\lambda \sigma^2 / k}$$



The Transient Impact Model (TIM)

- Considering a time interval $[0, T]$, the price S_t at time t is

$$S_t = S_0 + \int_0^t f(\dot{x}_s) G(t-s) ds + \int_0^t \sigma_s dW_s \tag{11}$$

where $\dot{x}_t \ dt > 0$ is the amount of shares purchased by the considered execution in $[t, t+dt]$, W_s is a Wiener process in a suitable probability space, and volatility σ_s is a deterministic function.

- The function f describes the instantaneous impact of the executed trades on price and in the linear case, it is

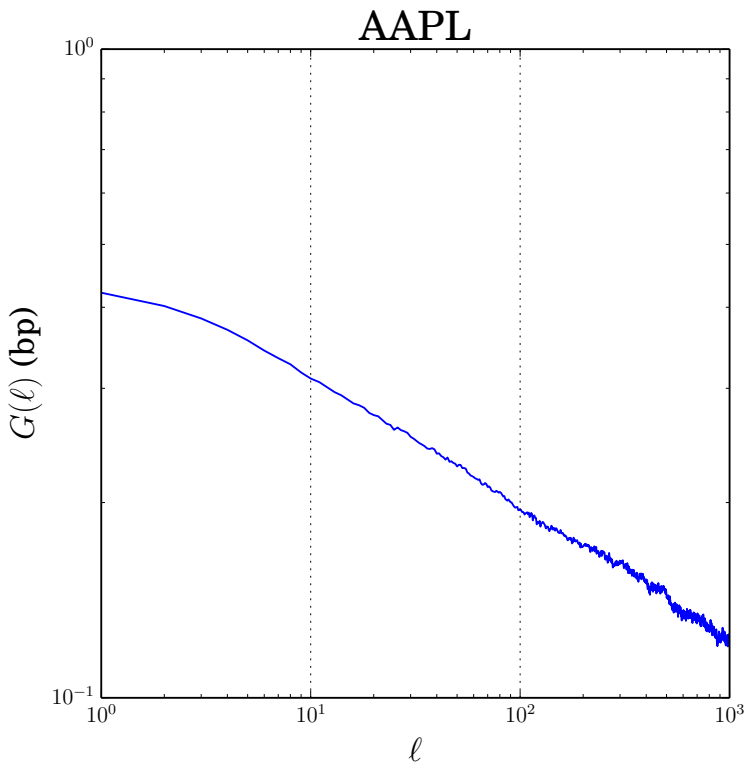
$$f(\dot{x}_t) = k\dot{x}_t \tag{12}$$

- The function $G(t)$, termed the kernel or propagator of the model, describes the delayed effect of trading on price and $G(t-s)$ characterizes how a trade at time s affects the price at time t
- When impact is small, the price dynamics of the LLOB coincides with that of the TIM with $f(z) = kz$ and $G(t) = 1/\sqrt{t}$.

The kernel: empirical evidences

Empirical evidences[8] obtained from transaction data points unambiguously towards a power law kernel (Bouchaud et al, 2004)

$$G(t) = t^{-\kappa} \qquad \kappa < 1$$



Typical values of κ are in the interval $[0.2, 0.5]$.

[8] From D. E. Taranto, G. Bormetti, J.-P. Bouchaud, F. Lillo, B. Toth, *Quantitative Finance* 2018.

Different benchmarks for optimal execution

Different benchmarks for optimal execution

- Implementation Shortfall (IS). Fix the execution time interval $[0, T]$ and minimize

$$\int_0^T S_t dx_t - x_0 S_0$$

or a risk adjusted version of it. Typically used to benefit from a price opportunity.

Different benchmarks for optimal execution

- Implementation Shortfall (IS). Fix the execution time interval $[0, T]$ and minimize

$$\int_0^T S_t dx_t - x_0 S_0$$

or a risk adjusted version of it. Typically used to benefit from a price opportunity.

- Target Close (TC). Fix the execution time interval $[0, T]$ and minimize

$$\int_0^T S_t dx_t - x_0 S_{close}$$

Used mainly by fund managers whose Net Asset Value is computed using closing price.

Different benchmarks for optimal execution

- Implementation Shortfall (IS). Fix the execution time interval $[0, T]$ and minimize

$$\int_0^T S_t dx_t - x_0 S_0$$

or a risk adjusted version of it. Typically used to benefit from a price opportunity.

- Target Close (TC). Fix the execution time interval $[0, T]$ and minimize

$$\int_0^T S_t dx_t - x_0 S_{close}$$

Used mainly by fund managers whose Net Asset Value is computed using closing price.

- Volume-Weighted Average Price (VWAP) orders. Fix the execution time interval $[0, T]$. Let $V_t dt$ the volume traded by the market in $[t, t + dt]$. Minimize

$$\int_0^T S_t dx_t - x_0 VWAP_0^T \qquad with \qquad VWAP_0^T = \frac{\int_0^T S_t V_t dt}{\int_0^T V_t dt}$$

Benchmark for traders who buy or sell shares in line with their global investment strategies or to hedge a risky position.

Problem setting: the generalized VWAP

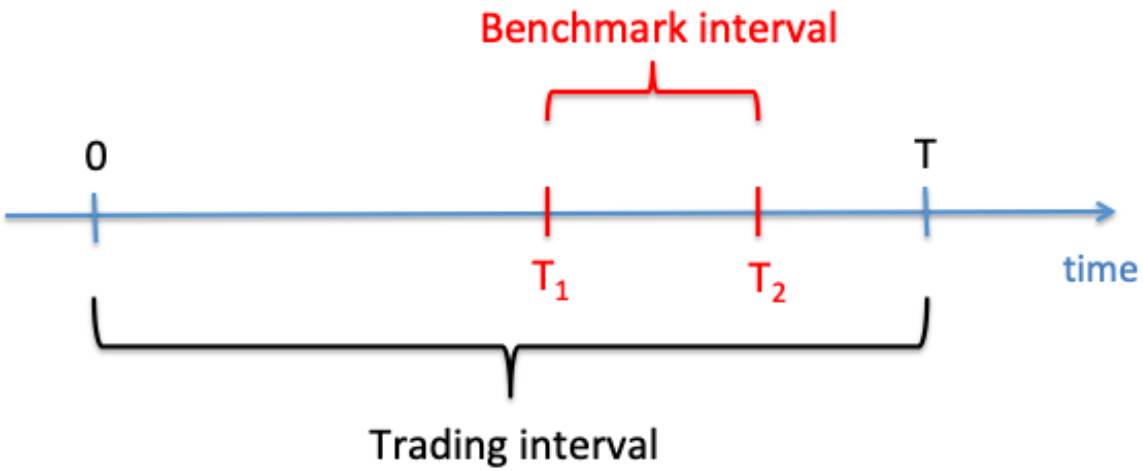
- What is the optimal execution under general benchmarks if price follows the TIM?

Problem setting: the generalized VWAP

- What is the optimal execution under general benchmarks if price follows the TIM?
- A broker wants to sell a quantity of $x_0 > 0$ shares in a time window $[0, T]$: the *trading interval*

Problem setting: the generalized VWAP

- What is the optimal execution under general benchmarks if price follows the TIM?
- A broker wants to sell a quantity of $x_0 > 0$ shares in a time window $[0, T]$: the *trading interval*
- The broker is benchmarked against the market Volume Weighted Average Price in a time window $[T_1, T_2] \subseteq [0, T]$, termed the *benchmark interval*



Problem setting: the generalized VWAP

- Special cases:
  - Implementation Shortfall with $T_1 = T_2 = 0^-$
  - Target Close with $T_1 = T_2 = T^+$
  - Standard VWAP with $T_1 = 0,\ T_2 = T$
- General case: Industry point in time benchmarks are being replaced with interval benchmarks

Let $V_t dt$ be the deterministic market volume traded in $[t, t + dt]$. The VWAP benchmark is given by

$$VWAP_{T_1}^{T_2} = \frac{\int_{T_1}^{T_2} S_t V_t dt}{\int_0^T V_t dt} = \int_0^T \eta_t S_t dt \tag{13}$$

where

$$\eta_t = \frac{V_t}{\int_{T_1}^{T_2} V_s ds} I_{t \in [T_1, T_2]} \tag{14}$$

Problem setting: the generalized VWAP

- The objective function of the broker is the difference between the cash she is able to obtain from the proceeds in the trading interval and the cash she will give back to the client, equal to the random variable $x_0 VWAP_{T_1}^{T_2}$.
- Let us define the cash process

$$dX_t = \dot{x}_t S_t dt \qquad X_0 = 0. \tag{15}$$

- Assuming a CARA risk averse agent, the objective function for a strategy $\mathbf{x} \equiv \{x_t\}_0^T$ is

$$U[\mathbf{x}] = \mathbb{E}_0[-\exp(-2\gamma(X_T - x_0 VWAP_{T_1}^{T_2}))] \tag{16}$$

where 2γ is the risk aversion parameter.

Transforming the optimization

Proposition

Under linear impact, $f(z) = -kz$ with $k > 0$, the maximization of the utility function (16) is equivalent to the minimization of the functional

$$C[\mathbf{x}] \equiv \frac{1}{2}\int_0^T \int_0^T \dot{x}_t \dot{x}_s G(|t-s|)ds\ dt - x_0 \int_0^T \eta_t dt \int_0^t G(t-s)\dot{x}_s ds \qquad (17)$$
$$+\frac{\gamma}{k}\int_0^T \int_0^T dt\ dt'(\dot{x}_t - x_0\eta_t)(\dot{x}_{t'} - x_0\eta_{t'}) \int_0^{t\wedge t'} \sigma_s^2 ds$$

Transforming the optimization

Proposition

Under linear impact, $f(z) = -kz$ with $k > 0$, the maximization of the utility function (16) is equivalent to the minimization of the functional

$$C[\mathbf{x}] \equiv \frac{1}{2}\int_0^T \int_0^T \dot{x}_t \dot{x}_s G(|t-s|)ds\ dt - x_0 \int_0^T \eta_t dt \int_0^t G(t-s)\dot{x}_s ds \qquad (17)$$
$$+\frac{\gamma}{k}\int_0^T \int_0^T dt\ dt'(\dot{x}_t - x_0\eta_t)(\dot{x}_{t'} - x_0\eta_{t'}) \int_0^{t\wedge t'} \sigma_s^2 ds$$

Proposition

The strategy $\{x_t^*\}_0^T$ minimizing the functional (17) with $\gamma = 0$ satisfies the integral equation

$$\int_0^T G(|t-s|)dx_s^* - x_0 \int_t^T \eta_s G(s-t)ds = \lambda \qquad (18)$$

where λ is a constant set by the normalization of the total volume traded

$$\int_0^T dx_s^* = x_0 \qquad (19)$$

Special cases

- **Implementation Shortfall.** When $T_1 = T_2 = 0$, it is $\eta_t = 2\delta(t)$ thus the integral equation becomes

$$\int_0^T G(|t-s|)dx_s^* = \lambda \tag{20}$$

as derived by Gatheral, Schied, & Slynko (*Mathematical Finance* 2012).

Special cases

- **Implementation Shortfall.** When $T_1 = T_2 = 0$, it is $\eta_t = 2\delta(t)$ thus the integral equation becomes

$$\int_0^T G(|t-s|)dx_s^* = \lambda \tag{20}$$

as derived by Gatheral, Schied, & Slynko (*Mathematical Finance* 2012).

- **Target Close.** When $T_1 = T_2 = T$ the integral equation becomes

$$\int_0^T G(|t-s|)dx_s^* = \lambda + x_0 G(T-t)$$

with solution $\dot{x}_s^* = w_s^{(1)} + x_0\delta(T-t)$ (the sum of $x_0/2$ shares traded as in the IS case and the remaining $x_0/2$ shares traded at $t = T$).

Standard VWAP with power law kernel

- We consider here the case when the benchmark VWAP interval $[T_1, T_2]$ coincides with the trading interval $[0, T]$ and $\eta_t = 1/T$, $\forall t \in [0, T]$, i.e. the market volume is constant in the interval.
- Consistently with data we choose a power law kernel $G(t) = t^{-\kappa}$ with $\kappa < 1$
- The optimal schedule is

$$\dot{x}_t = \frac{x_0}{T} \frac{2^{\kappa-2}\sqrt{\pi}\csc\left(\frac{\kappa\pi}{2}\right)}{\Gamma\left(1-\frac{\kappa}{2}\right)\Gamma\left(\frac{1+\kappa}{2}\right)} \frac{\left[\kappa + {}_2F_1\left(1, -1+\kappa, \frac{1+\kappa}{2}; \frac{t}{T}\right)\right]}{\left[\frac{t}{T}\left(1-\frac{t}{T}\right)\right]^{(1-\kappa)/2}} \tag{21}$$

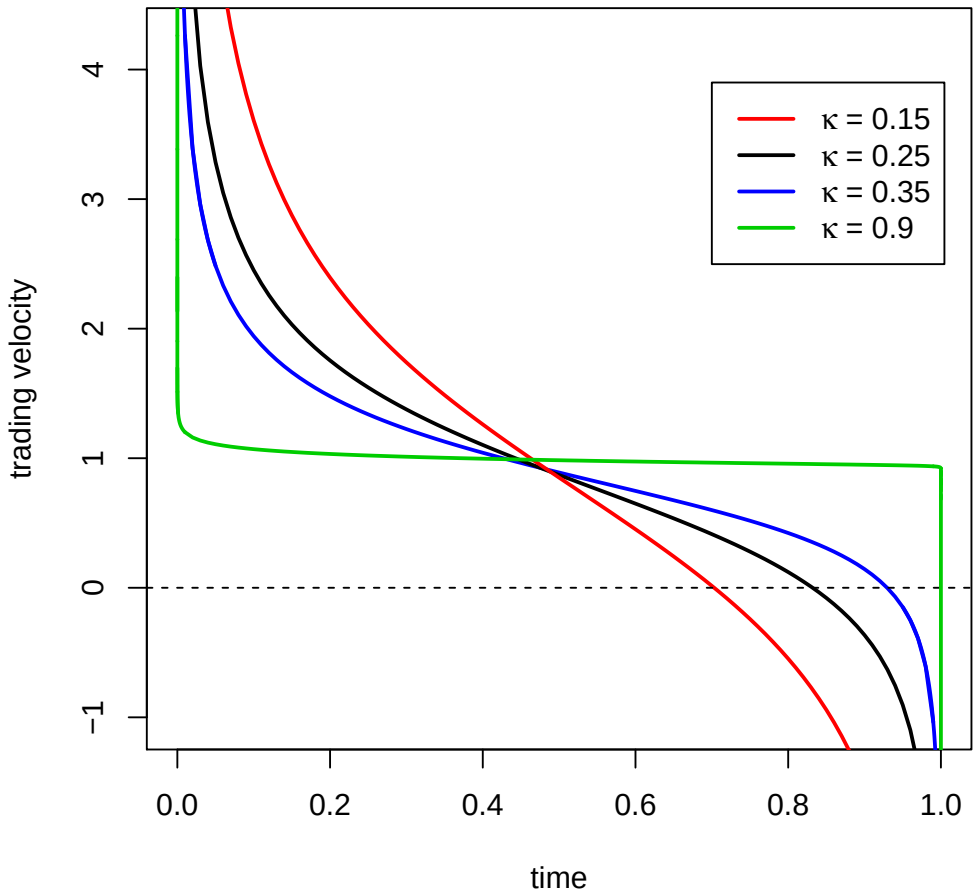


Figure: Optimal trading schedule for a VWAP of $x_0 = 1$ with benchmark interval coincident with the trading interval $[0, 1]$. The price is without drift and the broker is risk neutral. Four schedules for different values of the exponent κ of the kernel.

Comments

- In a VWAP sell execution it is optimal to buy toward the end of the trading period and thus that this strategy allows for transaction triggered price manipulation even when G is convex.

Comments

- In a VWAP sell execution it is optimal to buy toward the end of the trading period and thus that this strategy allows for transaction triggered price manipulation even when G is convex.
- When κ is small, the region of negative trading velocity becomes larger.

Comments

- In a VWAP sell execution it is optimal to buy toward the end of the trading period and thus that this strategy allows for transaction triggered price manipulation even when G is convex.
- When κ is small, the region of negative trading velocity becomes larger.
- In the limit $\kappa \to 1$, the optimal schedule is $v_t = x_0/T$, i.e. to trade at constant speed.

Comments

- In a VWAP sell execution it is optimal to buy toward the end of the trading period and thus that this strategy allows for transaction triggered price manipulation even when G is convex.
- When κ is small, the region of negative trading velocity becomes larger.
- In the limit $\kappa \to 1$, the optimal schedule is $v_t = x_0/T$, i.e. to trade at constant speed.
- For comparison, in the Almgren-Chriss framework, $\dot{x}_t$ is a straight line with negative slope depending on the temporary impact (Gueant and Royer, *SIAM Journal of Financial Mathematics*, 2014).

Comments

- In a VWAP sell execution it is optimal to buy toward the end of the trading period and thus that this strategy allows for transaction triggered price manipulation even when G is convex.
- When κ is small, the region of negative trading velocity becomes larger.
- In the limit $\kappa \to 1$, the optimal schedule is $v_t = x_0/T$, i.e. to trade at constant speed.
- For comparison, in the Almgren-Chriss framework, $\dot{x}_t$ is a straight line with negative slope depending on the temporary impact (Gueant and Royer, *SIAM Journal of Financial Mathematics*, 2014).
- With exponential kernel $G(t) = e^{-\rho t}$ the optimal strategy is

$$v_t = \frac{x_0}{\rho T(2+\rho T)} \left[2(1+\rho T)\delta(t) + \rho(1+\rho T) - 2\delta(t-T)\right]$$

i.e. to sell a finite amount at time $t = 0$, then selling at a constant rate for the whole interval $(0, T)$ and finally *buying* a finite amount at time $t = T$.

Solution in discrete time

Whenever more constraints are added to the optimal execution problem, it is convenient to frame the problem in discrete time. This can be done at three different levels

1. express the cost function in discrete time and solve the optimization;
2. use discrete time to obtain a quadrature of the integral equation;
3. write the Transient Impact Model in discrete time, derive the corresponding cost, and then minimize it.

It is worth noticing that the three procedures do not give exactly the same result, however if the time intervals used in the discretization are sufficiently small, the differences become negligible. In the following we will consider approach (3).

Transient Impact Model in discrete time

- Let us divide the interval $[0, T]$ in N equal intervals and define $\tau = T/N$.

Transient Impact Model in discrete time

- Let us divide the interval $[0, T]$ in N equal intervals and define $\tau = T/N$.
- The strategy is now a vector $\mathbf{x} = (x_1, ..., x_N)'$, where x_i is the amount of shares traded in interval i, i.e. for $t \in [(i-1)\tau, i\tau]$.

Transient Impact Model in discrete time

- Let us divide the interval $[0, T]$ in N equal intervals and define $\tau = T/N$.
- The strategy is now a vector $\mathbf{x} = (x_1, ..., x_N)'$, where x_i is the amount of shares traded in interval i, i.e. for $t \in [(i-1)\tau, i\tau]$.
- The price dynamics of a sell execution in discrete time is

$$S_\ell = S_0 - k \sum_{i=1}^{\ell} G(\ell - i) x_i + \tau^{1/2} \sum_{i=1}^{\ell} \epsilon_\ell \qquad \ell = \{0, ..., N\} \tag{22}$$

which can be rewritten in vector form as

$$\mathbf{S} = S_0 \mathbf{1} - kG\mathbf{x} + \tau^{1/2} L\boldsymbol{\epsilon} \tag{23}$$

where $\mathbf{S} = (S_1, ..., S_N)'$, $\mathbf{1} = (1, ..., 1)'$, L is the lower triangular matrix of ones (i.e. $L_{ij} = 1$ if $i \geq j$, zero otherwise), and G is the lower triangular matrix such that $G_{ij} = G[\tau(i-j)]$ if $i \geq j$ and zero otherwise.

Transient Impact Model in discrete time

- Let us divide the interval $[0, T]$ in N equal intervals and define $\tau = T/N$.
- The strategy is now a vector $\mathbf{x} = (x_1, ..., x_N)'$, where x_i is the amount of shares traded in interval i, i.e. for $t \in [(i-1)\tau, i\tau]$.
- The price dynamics of a sell execution in discrete time is

$$S_\ell = S_0 - k \sum_{i=1}^{\ell} G(\ell - i) x_i + \tau^{1/2} \sum_{i=1}^{\ell} \epsilon_\ell \qquad \ell = \{0, ..., N\} \tag{22}$$

  which can be rewritten in vector form as

$$\mathbf{S} = S_0 \mathbf{1} - kG\mathbf{x} + \tau^{1/2} L\boldsymbol{\epsilon} \tag{23}$$

  where $\mathbf{S} = (S_1, ..., S_N)'$, $\mathbf{1} = (1, ..., 1)'$, L is the lower triangular matrix of ones (i.e. $L_{ij} = 1$ if $i \geq j$, zero otherwise), and G is the lower triangular matrix such that $G_{ij} = G[\tau(i-j)]$ if $i \geq j$ and zero otherwise.
- Finally $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ is a Gaussian random vector describing the price dynamics without execution.

General VWAP

- In full generality, we consider a VWAP benchmark between $t = T_1$ and $t = T_2$, corresponding to $\ell_1 = \lfloor NT_1/T \rceil$ $\ell_2 = \lfloor NT_2/T \rceil$ are the rounding to the nearest integer giving the initial and final trading intervals.

General VWAP

- In full generality, we consider a VWAP benchmark between $t = T_1$ and $t = T_2$, corresponding to $\ell_1 = \lfloor NT_1/T \rceil$ $\ell_2 = \lfloor NT_2/T \rceil$ are the rounding to the nearest integer giving the initial and final trading intervals.
- We introduce $B = \{\ell \in \mathbb{N} : \ell_1 \leq \ell \leq \ell_2\}$ and a vector $\boldsymbol{\eta}$ with components

$$\eta_\ell = \frac{V_\ell}{||\boldsymbol{\eta}||_1} I_{\ell \in B} \tag{24}$$

where V_ℓ is the market volume traded in interval ℓ.

General VWAP

- In full generality, we consider a VWAP benchmark between $t = T_1$ and $t = T_2$, corresponding to $\ell_1 = \lfloor NT_1/T \rceil$ $\ell_2 = \lfloor NT_2/T \rceil$ are the rounding to the nearest integer giving the initial and final trading intervals.
- We introduce $B = \{\ell \in \mathbb{N} : \ell_1 \leq \ell \leq \ell_2\}$ and a vector $\boldsymbol{\eta}$ with components

$$\eta_\ell = \frac{V_\ell}{||\boldsymbol{\eta}||_1} I_{\ell \in B} \tag{24}$$

  where V_ℓ is the market volume traded in interval ℓ.
- The benchmark is $x_0 \boldsymbol{\eta}' \mathbf{S}$ and the normalization ensures that $\mathbf{1}' \boldsymbol{\eta} = 1$.

General VWAP

- In full generality, we consider a VWAP benchmark between $t = T_1$ and $t = T_2$, corresponding to $\ell_1 = \lfloor NT_1/T \rceil$ $\ell_2 = \lfloor NT_2/T \rceil$ are the rounding to the nearest integer giving the initial and final trading intervals.
- We introduce $B = \{\ell \in \mathbb{N} : \ell_1 \leq \ell \leq \ell_2\}$ and a vector $\boldsymbol{\eta}$ with components

$$\eta_\ell = \frac{V_\ell}{||\boldsymbol{\eta}||_1} I_{\ell \in B} \tag{24}$$

  where V_ℓ is the market volume traded in interval ℓ.
- The benchmark is $x_0 \boldsymbol{\eta}' \mathbf{S}$ and the normalization ensures that $\mathbf{1}'\boldsymbol{\eta} = 1$.
- The utility function is $\mathcal{U}[(\mathbf{x} - x_0\boldsymbol{\eta})'\mathbf{S}]$ and, using the Gaussian assumption under CARA utility function with risk aversion 2γ, the expected utility is

$$U[\mathbf{x}] = \mathbb{E}_0[(\mathbf{x} - x_0\boldsymbol{\eta})'\mathbf{S}] - \gamma \mathbb{V}_0[(\mathbf{x} - x_0\boldsymbol{\eta})'\mathbf{S}] \tag{25}$$

Proposition

Under CARA utility function with risk aversion 2γ, the optimal VWAP execution, which maximizes the expected utility, is the solution of the quadratic optimization

$$\min_{\mathbf{x}} \left[\mathbf{x}' A \mathbf{x} - \mathbf{b}' \mathbf{x}\right] \qquad s.t. \quad \mathbf{1}' \mathbf{x} = x_0$$

where

$$A = kG + \gamma\tau L \Sigma L' \tag{26}$$

$$\mathbf{b}' = k x_0 \boldsymbol{\eta}' G + 2\gamma\tau x_0 \boldsymbol{\eta}' L \Sigma L' + \tau^{1/2} \boldsymbol{\mu}' L' \tag{27}$$

Moreover, the matrix A is positive definite if Σ is positive definite. Thus the solution of the quadratic optimization exists and is unique.

Since the problem can be recast in a quadratic optimization form, several additional constraints can be added without affecting the difficulty of the problem.
For example, it is possible to add the constraint that all the trades have the same sign, e.g. no buys in a sell execution ($x_i \geq 0, \forall i$), or a constraint on the maximal trading speed ($|x_i| \leq x_{max}, \forall i$).

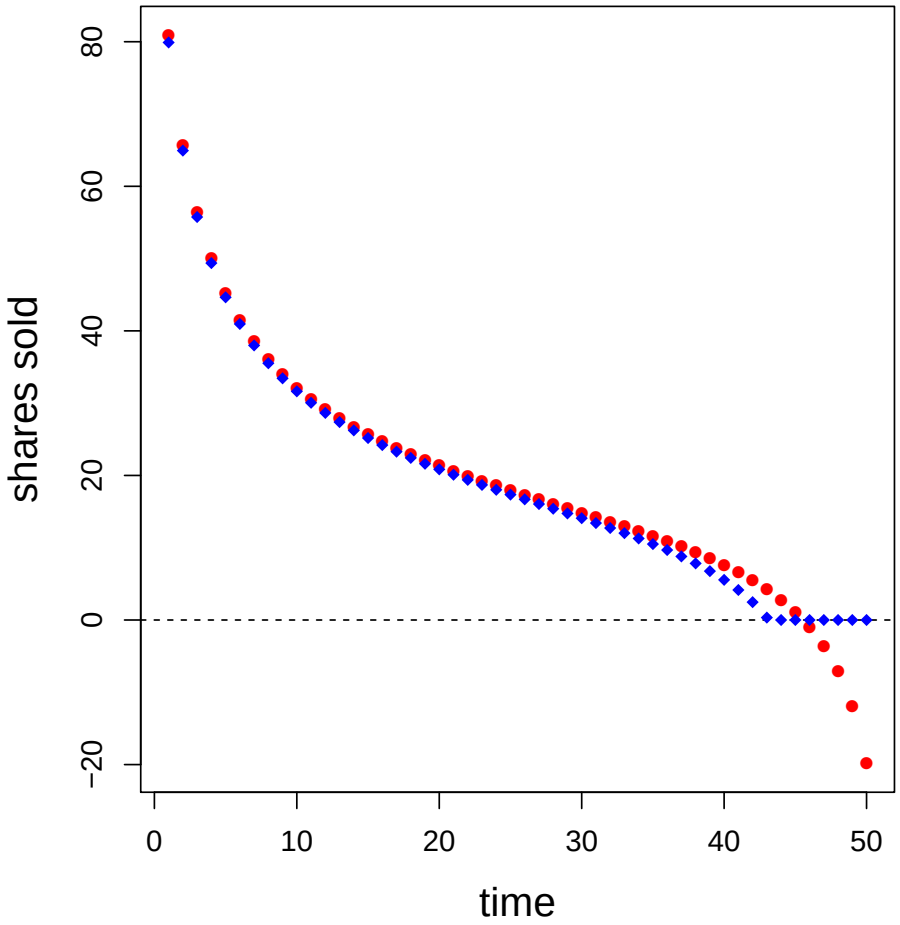


Figure: Optimal trading schedule for a VWAP with benchmark interval coincident with the trading interval. The price is without drift and the broker is risk neutral. The red dots refer to the unconstrained case, while the blue ones to the case with a constraint on the non-negativity of trades (no buys for a sell execution).

Role of drift and risk aversion

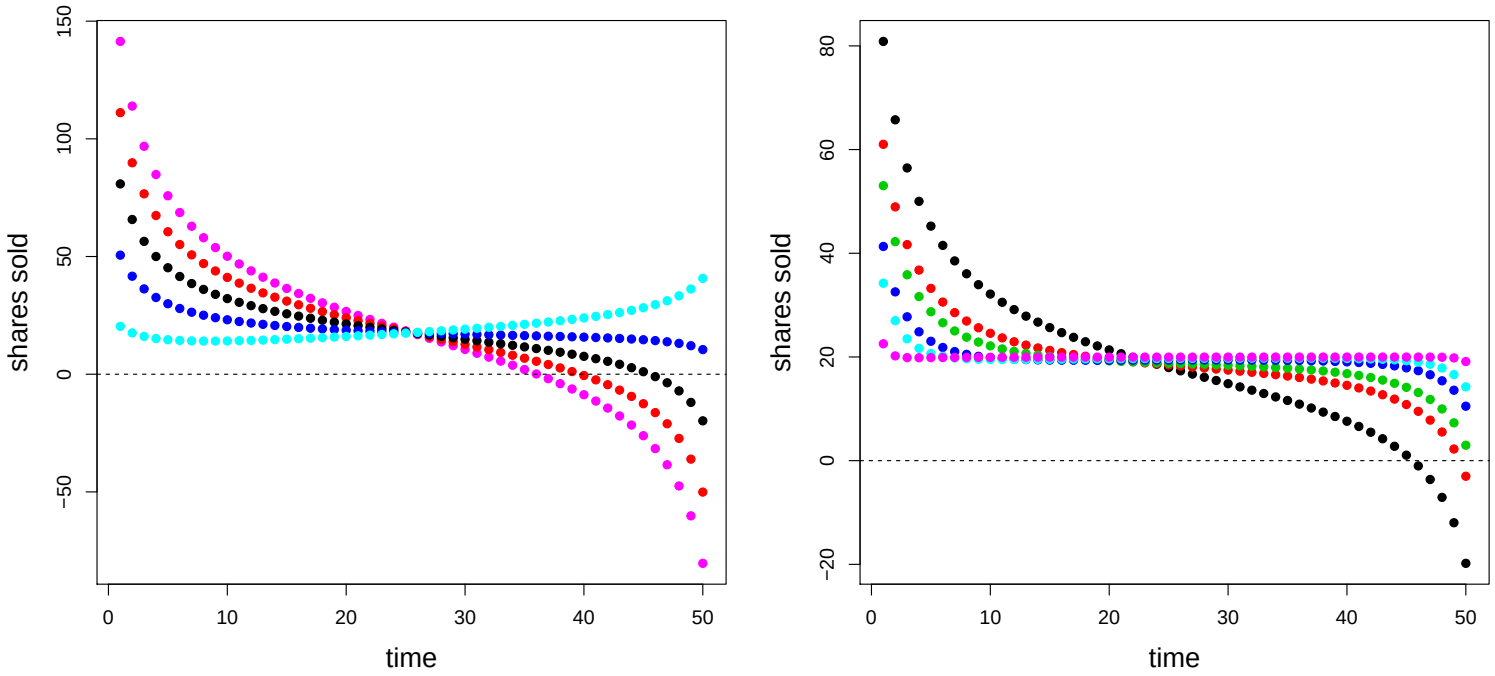


Figure: Left. Optimal VWAP schedule for a sell order by a risk neutral broker for different values of the price drift $\mu_i = 4$ (cyan), $\mu_i = 2$ (blue) $\mu_i = -2$ (red), and $\mu_i = -4$ (magenta). Black dots refer to the driftless benchmark case. Right. Optimal VWAP schedule for a risk averse broker under driftless price. The values of the risk aversion parameter γ are 0 (black), 0.5 (red), 1 (green), 3 (blue), 7 (cyan), 100 (magenta). In both panels the benchmark interval is coincident with the trading interval.

When the benchmark interval does not coincides with the trading interval

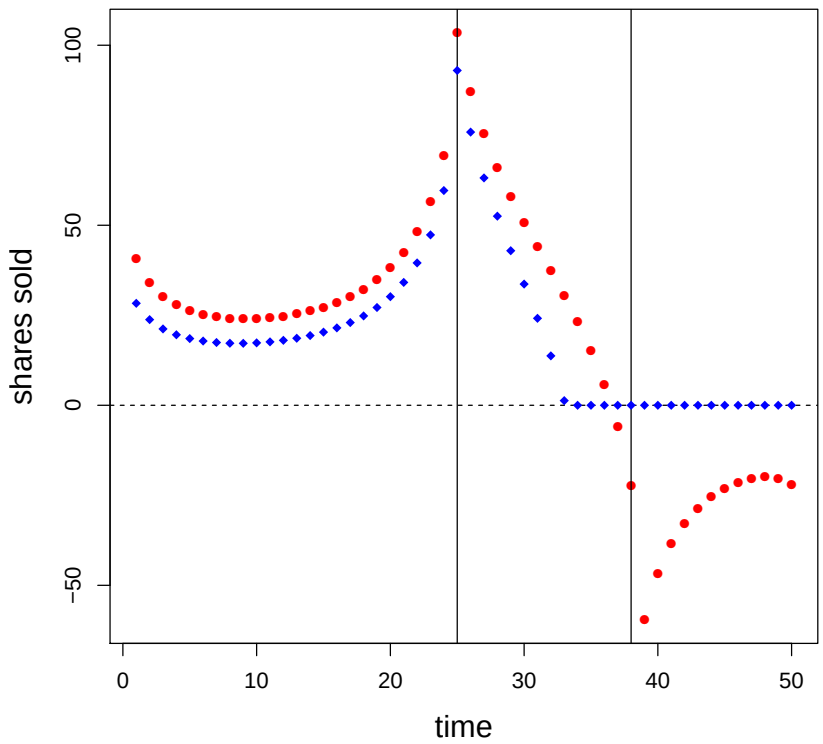


Figure: Optimal schedule without (red) and with (blue) constraint on trade sign for a VWAP with benchmark interval $T_1 = 25$ and $T_2 = 38$ (vertical lines).

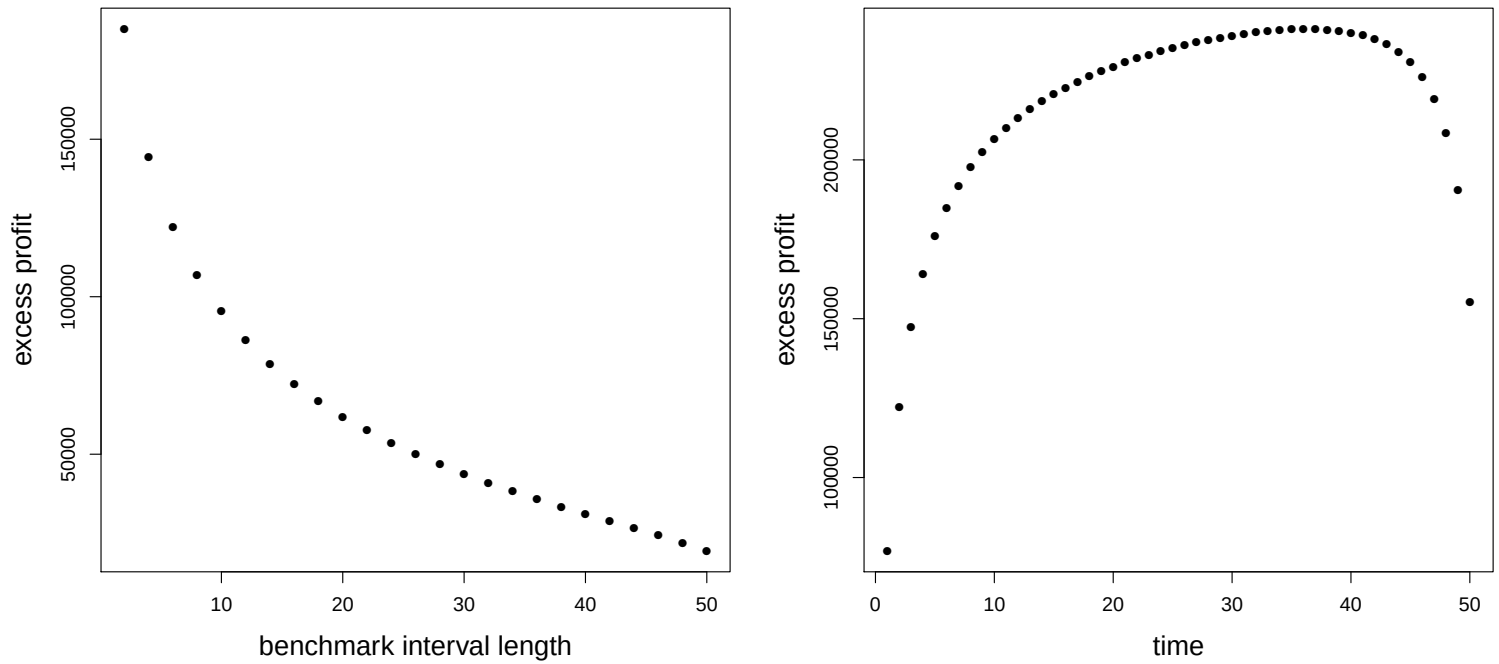


Figure: Excess profit of the broker for a VWAP execution with benchmark interval different from trading interval. The left panel shows the profit as a function of the length of the benchmark period when it is centered in $T/2$. The right panel shows the profit as a function of the time within the trading period when the benchmark period has unit length.

The excess profit (the difference between the cash at the end of the trading period and the VWAP in the benchmark period) is maximal for a benchmark period of length one and located in the second half of the trading interval (assuming the impact model continues to hold also for the very large trading intensities).

Bibliography

- Elia Zarinelli, Michele Treccani, J. Doyne Farmer, Fabrizio Lillo, Beyond the square root: Evidence for logarithmic dependence of market impact on size and participation rate, Market Microstructure and Liquidity (2015)
- Frederic Bucci, Michael Benzaquen, Fabrizio Lillo, Jean-Philippe Bouchaud, Slow decay of impact in equity markets: insights from the ANcerno database, Market Microstructure and Liquidity (2019)
- Frederic Bucci, Michael Benzaquen, Fabrizio Lillo, Jean-Philippe Bouchaud, Crossover from linear to square-root market impact, Physical Review Letters (2019)
- Frederic Bucci, Iacopo Mastromatteo, Zoltan Eisler, Fabrizio Lillo, Jean-Philippe Bouchaud, Charles-Albert Lehalle, Co-impact: Crowding effects in institutional trading activity, Quantitative Finance (2020)
- Frederic Bucci, Fabrizio Lillo, Jean-Philippe Bouchaud, Michael Benzaquen, Are trading invariants really invariant? Trading costs matter, Quantitative Finance (2020)
- Alexander Barzykin, Fabrizio Lillo, Optimal VWAP execution under transient price impact, (2019)