Market Microstructure and Algorithmic Trading

University of Oxford

Fayçal Drissi
faycal.drissi@eng.ox.ac.uk
http://www.faycaldrissi.com

lecture notes

January 15, 2024

Overview of **Algorithmic Trading** models in **high frequency** markets.

- Algorithmic Trading is the use of computerized algorithms that make trading decisions.
- **high frequency** trading is characterised by the reliance on speed because trading decisions are frequent.

Electronic markets design and mechanisms.

The **role** of different market participants.

The **methodology** in algorithmic trading models:

1. (i) Formulate the decision problem of an agent based on a specific need.
2. (ii) Propose a parsimonious (dynamic) model of the environment.
3. (iii) Frame the decision problem as an optimisation problem that can be solved using classical mathematical tools.
4. (iv) Study/discuss the solution (simulations / backtest).

What you **need** to know:

- Basic convex analysis.
  - Legendre-Fenchel transforms.
  - Bolza problems and Hamiltonian systems.
- Basics of Stochastic Optimal Control.
  - Control for diffusion processes
  - Control for jump processes.

All you need to know in 6 pages: Section 1.5 of the lecture notes.

- (1) Almgren-Chriss model for optimal execution of large orders.
- (2) The Cartea-Jaimungal framework for optimal execution.
- (3) Optimal execution and statistical arbitrage with predictive signals.
- (4) Optimal trading of portfolios (multi-asset execution).
- (5) Optimal execution with transient impact.
- (6) Optimal trading with limit orders.
- (7) Optimal market making.
- (8) (if we have time) optimal trading in decentralised finance.

Algorithmic trading:

- (Cartea et al. 2015) Cartea, Álvaro, Sebastian Jaimungal, and José Penalva. Algorithmic and high-frequency trading. Cambridge University Press, 2015.
- (Guéant 2016) Guéant, Olivier. The Financial Mathematics of Market Liquidity: From optimal execution to market making. Vol. 33. CRC Press, 2016.
- (Donnelly 2022) Donnelly, R. (2022). Optimal execution: A review. Applied Mathematical Finance, 29(3), 181-212.

Convex Optimization:

- (Rockafellar 1997) Rockafellar, R.T., 1997. Convex analysis. volume 11. Princeton university press.

Dynamic programming

- (Pham 2009) Pham, Huyên. Continuous-time stochastic control and optimization with financial applications. Vol. 61. Springer Science & Business Media, 2009.

Finance and microstructure

- (O'hara 1998) O'hara, M. (1998). Market microstructure theory. John Wiley Sons.

Financial markets after the 2008 crisis.

- No more appetite for sophisticated products.
- Financial system moved from a **bespoke** market to a **mass** market.[1]
- Logistics are optimized: listing, standardization, central clearing, price formation, etc.
- Regulators pushed for this change because it prevents accumulation of inventory risk.

[1] Bespoke means products that are diverse and personalised: no economy of scale but high margins. Mass market means many similar products, so logistics are optimized.

Why algorithmic trading ?

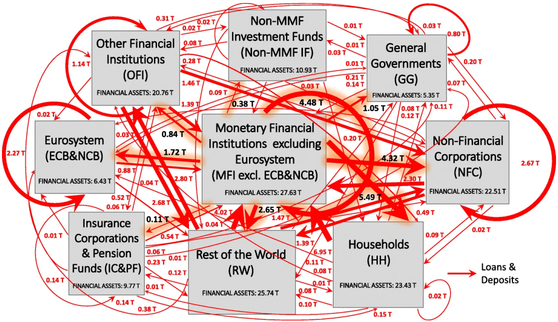
Three main reasons:

- Intermediation
- Electronification of markets
- Market fragmentation

Role of the financial system: **risk transformation**.

Need for **intermediation** (Scholtens and Van Wensveen 2003):

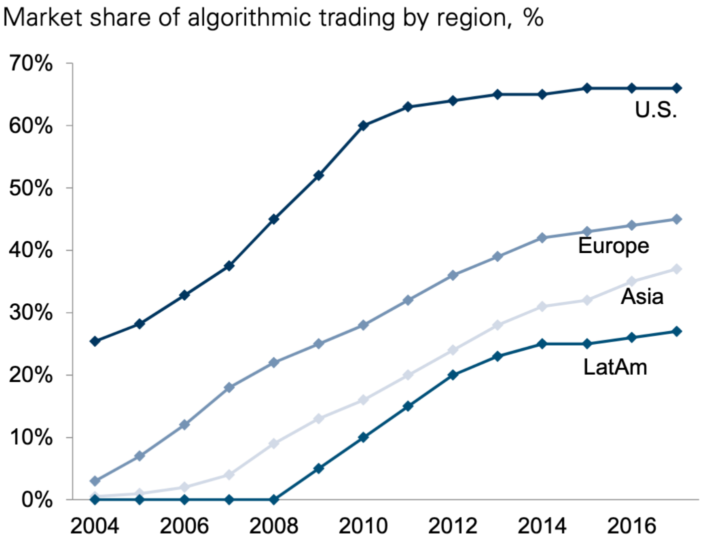
- Concentrate the flow
- Build a market place (Find counterparts)
- Provide neutral information
- No inventory risk
- Surveillance (laundering, manipulation, frauds)



The euro area macro-network of financial exposures via loans and deposits; June 2017. Source: (Perillo and Battiston 2018).

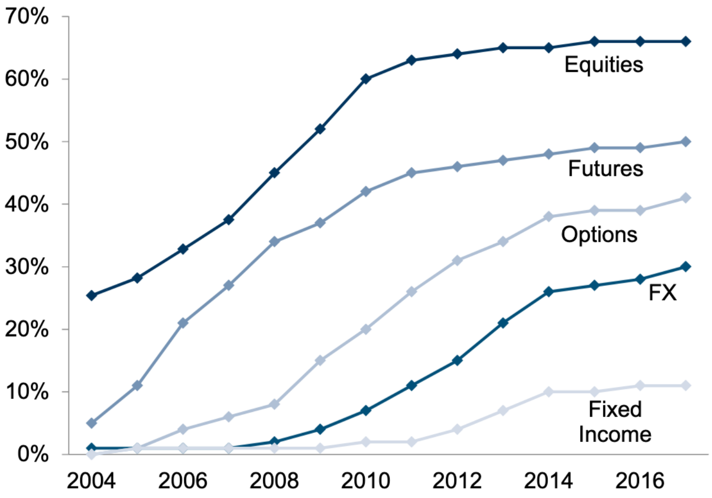
Banks: Match borrowers and lenders, match buyers and sellers, structured prod. / hedgeing.
Others: central banks, IBs, mutual funds, brokers, dealers, central counterparty clearing houses, insurances, etc.

- Policy makers push towards electronification:
  - improves traceability
  - less information asymmetry (price discovery)
  - efficient logistics (CCP)
- Growth in electronic trading platforms (ETPs) and electronic communication networks (ECNs)
  - helped to pool liquidity by enabling the multilateral and cross-border interaction between buyers and sellers.
- Main consequences of electronification:
  - reduce the costs of trading
  - high-frequency trading (HFT) that reduce the delay (latency) in execution and increase the speed at which market participants can access markets.



Source: Aite Group, Goldman Sachs Global Investment Research.

Market share of algorithmic trading by asset class, %



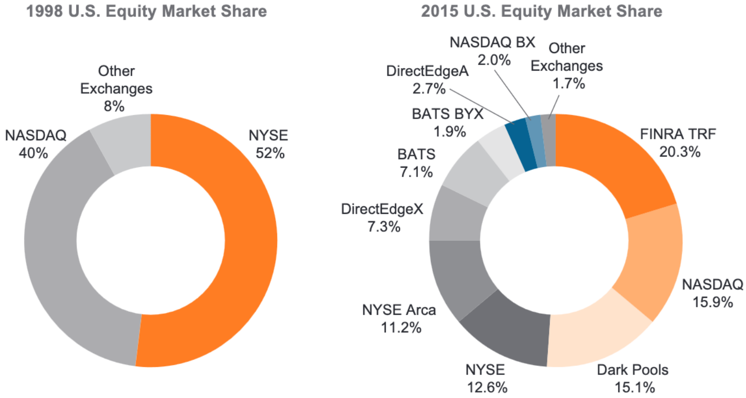
Source: Aite Group, Goldman Sachs Global Investment Research.

- Policy makers push towards competition between trading venues (fragmentation).

  2005 Regulation National Market System (Reg NMS): decentralise exchanges who compete mainly on price.

- Main consequences
  - Orders may be spread among competing venues.
  - New players have proliferated, further fragmenting the US equity markets (and other markets)
  - lowered the barriers to entry with lower prices and higher speed.
  - near-instantaneous liquidity

Fragmentation Evident from Comparison of Market Structure in 1998 vs. 2015



Source: Credit Suisse Trading Strategy
Note: "Other TRF" includes broker capital commitments and internalizations. "Other Exchanges" includes NYSE Amex, CBSX, Chicago, National and PSX.

What is it ?

- Multiple agents operate in a large fragmented and electronic market.
- Each agent has
  - A specific goal / objective.
  - A space of decisions: buy, sell, quantity, speed, location, time, ...
  - An urgency (or not) to complete the objective.
  - An appetite for risk / risk constraints.

  $\Longrightarrow$ They need to act optimally

- Each agent needs
  - A model of the environment.
  - An optimisation problem to solve.
  - Optimal trading is about optimizing a trading process.

The **ingredients** of an optimal trading problem

- The **model** of the environment
  - What is deterministic ? what is stochastic ?
  - The simpler the model, the easier it is to obtain closed-form formulas (fast to compute in high-frequency markets).
- The performance **criterion**:
  - What is important for the agent: generally (expected) terminal wealth, but can include risk preferences (utility).
- The **controls**:
  - What are the possible actions of the agent ?

Optimal execution $\subset$ optimal trading.

Market operators with **large orders** regularly come to the market.

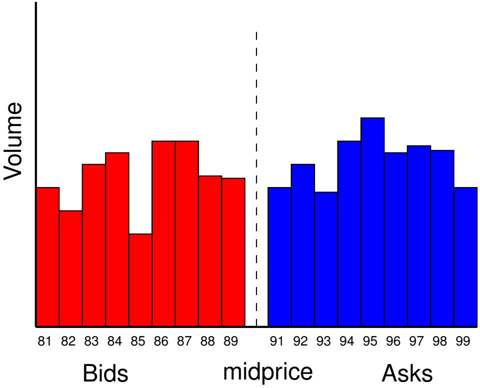
Examples:

- Asset managers delegate trading to dealing desks.
- Banks manage their liquidity risk through central risk books (CRB).
- Brokers act on behalf of pension funds, hedge funds, mutual funds..
- HFTs or fast hedge funds take decisions because of liquidity signals.

When the orders are a significant portion of the overall volume: The market operator (agent) must **slice** the **parent order** (metaorder) into **child orders**.

Slow execution exposes the agent to **adverse price fluctuations**. **Fast execution** exposes the agent to **high execution costs**.

The agent must formulate a model to decide how to execute a large order optimally.



- The LOB allows to match buyers with sellers.
- Buyers sit at the **bid** side of the book.
- Sellers sit at the **ask** side of the book.
- The **midprice** is the average of the **best bid** and **best ask**.

- Buyers and sellers send **limit orders** (LOs): composed of a **price** level, a **volume**, and an **indicator** to buy or to sell the asset.
- LOs are called **passive** orders because they do not consume liquidity immediately.

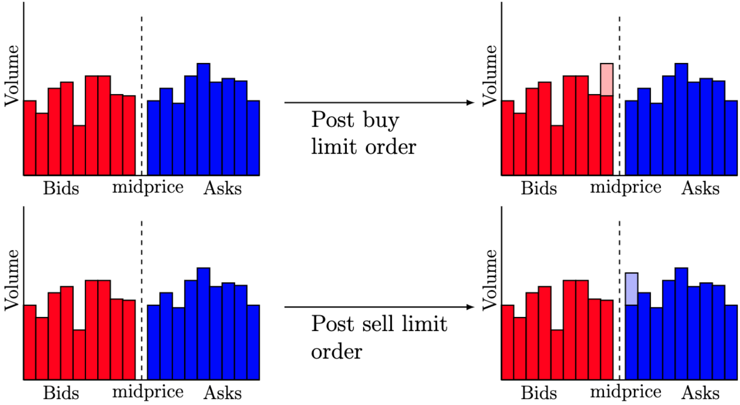


Figure 1: Changes in the LOB after a buy / sell LO.

- Traders can also **cancel** (fully or partially) LOs.
- Traders can also post LOs at different price levels.

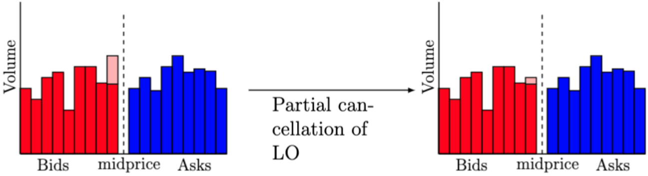


Figure 2: Changes in the LOB after a cancellation.

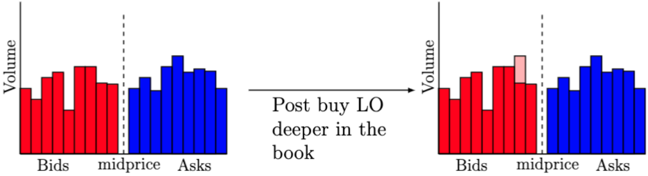


Figure 3: Changes in the LOB after a buy LO at the second best level.

- Traders can also send **market orders** (MOs).
- MOs are called **agressive** orders because they consume liquidity immediately.
- MOs pay **the bid-ask spread**.
- Market makers earn **the bid-ask spread** for providing liquidity.

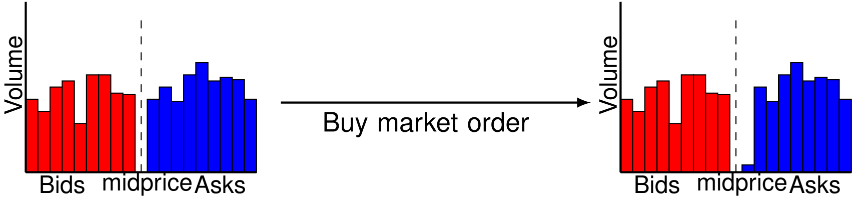


Figure: Changes in the LOB after a buy MO.

Bid-ask spread and volatility/risk

Universal relation between b-a spread η and volatility σ (Wyart et al. 2008).

$$\eta \propto \frac{\sigma}{\sqrt{M}}, \quad M\text{: number of trades per day.}$$

- Higher trading activity $\Longrightarrow$ liquidity is cheap.
- Higher volatility $\Longrightarrow$ liquidity is expensive.

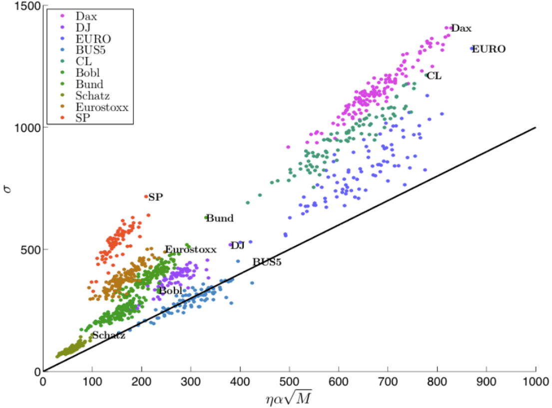


Figure: Source: (Lehalle and Laruelle 2018).

Volatility seasonality

- Volatility has seasonality: less intense at the end than at the start of the day.

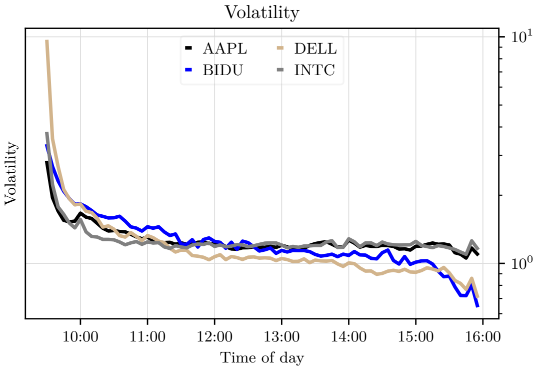


Figure: Average volatility throughout the trading day. Data is between January and March 2023

B.A-spread and **volume on the book**

- BA-spread is large at the start of the day, but finishes small because of market maker running to get rid of their inventory passively.
- VOB $\left(\left(Q^{A}+Q^{B}\right)/2\right)$ seasonality is the inverse of that of the B.A-spread.

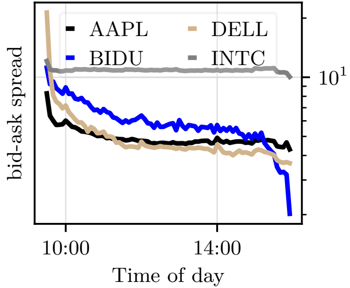


Figure: Average bid-ask spread and book volume throughout the trading day. Data is between January and March 2023

Market impact

Market impact $=$ the effect that MOs have on prices.

The literature generally splits the impact of MOs into a **temporary** and a **permanent** impact.

- **Temporary** price impact, or **execution costs**: difference between the midprice and transaction price per share.**Temporary** price impact.
- **Permanent** price impact: refers to the relationship between the volume of MOs and the midprice at future times.

Temporary price impact / execution costs: difference between the midprice and transaction price per share.

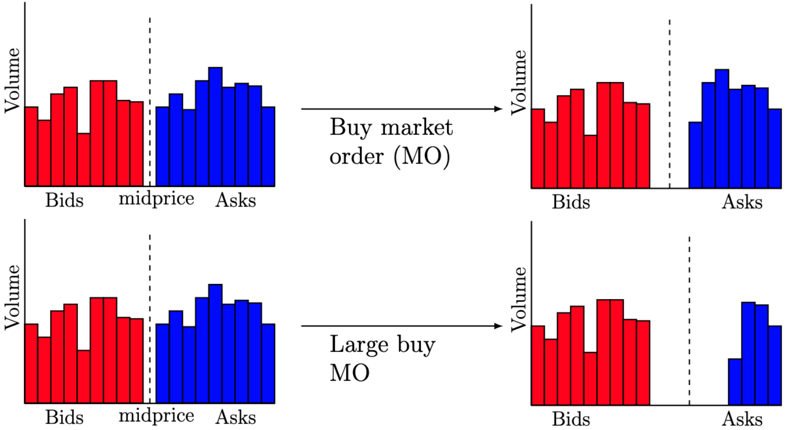


Figure 5: Limit order book.

Provided with a snapshot of the LOB, traders can estimate the execution costs.

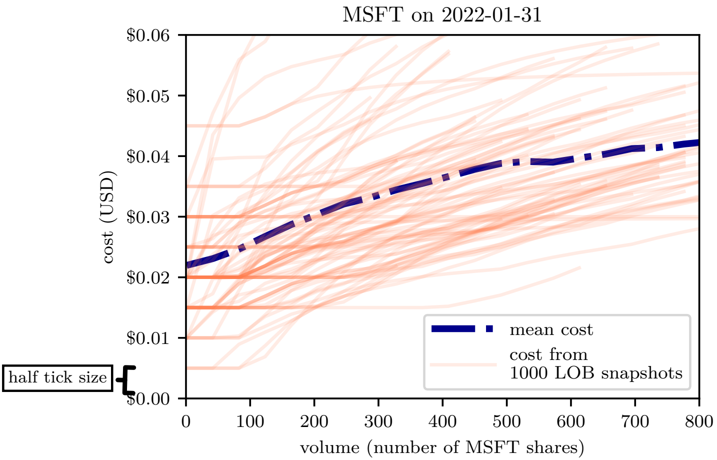


Figure: Execution costs defined as a function of trading volume for multiple snapshots of the LOB of MSFT quoted on Nasdaq. The execution costs are defined as the difference between the execution price per share for a given volume of share, and the midprice.

Permanent price impact: large MOs leave a long-term effect on the midprice.

Interpretation: traders base their activity on information on the **fundamental** value of the firm.

(Gatheral 2010) shows with that a model with permanent impact that is not linear in the size of the MO leads to dynamic arbitrage.

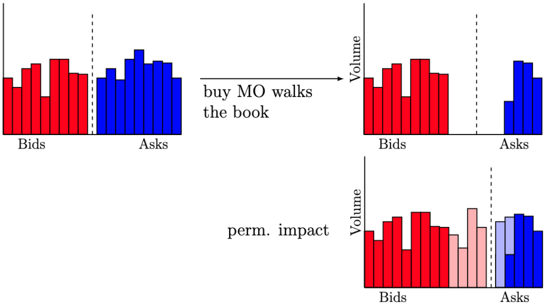


Figure 7: In the first two panels, an MO walks the book so the next midprice exhibits the temporary price impact. Immediately after the MO, market participants replenish the LOB. The difference between the midprice in the last panel and that of the first panel is the permanent impact.

Optimal order routing of aggressive orders

The decision problem

- A “**marketable**” order is a buy (resp. sell) order at a price higher than the best ask (resp. lower than the best bid).
- Often, operators have to be split a large order over N available venues.

Optimal order routing of aggressive orders

The decision problem

- A "**marketable**" order is a buy (resp. sell) order at a price higher than the best ask (resp. lower than the best bid).
- Often, operators have to be split a large order over N available venues.

The model of the environment

- Let ($Q^\star$ at max price $P^\star$) be a marketable **buy** order.
  Let $Q_n(p)$ be the **visible quantity** that is **available** at **price** p in **trading venue** n.

Optimal order routing of aggressive orders

The decision problem

- A "**marketable**" order is a buy (resp. sell) order at a price higher than the best ask (resp. lower than the best bid).
- Often, operators have to be split a large order over N available venues.

The model of the environment

- Let ($Q^\star$ at max price $P^\star$) be a marketable **buy** order.
  Let $Q_n(p)$ be the **visible quantity** that is **available** at **price** p in **trading venue** n.

The optimisation problem : Choose $(p_1, \cdots, p_n)$ to minimize $\sum_{n=1}^{N} p_n \cdot Q_n(p_n)$ such that $Q^\star = \sum_{n=1}^{N} Q_n(p_n)$ with the constraint that $P^\star \geq p_n$ for all $n \in \{1, \cdots, N\}$.

The solution:

- Lagrangian: $Q_n(p_n) + p_n\, Q_n^{'}(p_n) = \lambda\, Q_n^{'}(p_n)$ for $n \in \{1, \cdots, N\}$.
- Assume the linear form $Q_n(p) = q_n + c_n \cdot p$.

$$\Longrightarrow \;(q_n + c_n \cdot p_n^\star) + p_n^\star\, c_n = \lambda\, c_n$$

$$\Longrightarrow \; p_n^\star = \frac{\lambda}{2} - \frac{q_n}{2\, c_n}\,.$$

- Inject λ in the constraint $Q^\star = \sum_{n=1}^{N} Q_n(p_n)$:

$$\Longrightarrow \; Q^\star = \sum_{n=1}^{N} \{q_n + c_n \cdot p_n^\star\} = \sum_{n=1}^{N} q_n/2 + c\,\lambda/2, \quad c = \sum_n c_n$$

- Finally

$$\boxed{p_n^\star = \frac{Q^\star}{c} - \frac{q_n}{2\, c_n}\left(1 + \frac{c_n}{q_n} \cdot \frac{\overline{q}}{\overline{c}}\right)} \quad \overline{c} = \frac{1}{N}\sum_n c, \;\; \overline{q} = \frac{1}{N}\sum_n q$$

Some issues with our (simple) model:

- There is **latency** in high frequency markets.
- There are **cancellations** in limit order books.
- The shape of the book **hides** the truth sometimes: iceberg orders + hidden orders.
- etc..

The decision problem

- An operator splits a large LO over N venues: there is structurally **uncertainty** on limit orders splitting; waiting on a bad queue generates **opportunity costs**. Formulate an optimal trading problem ?

The decision problem

- An operator splits a large LO over N venues: there is structurally **uncertainty** on limit orders splitting; waiting on a bad queue generates **opportunity costs**. Formulate an optimal trading problem ?

The model:

- The best queue in each venue has size Q_n.
- Each queue is **consumed** according to a Poisson P_t^n with intensity λ_n.

The objective: find LO quantities $(q_1, \cdots, q_N)$ to minimize, **on average**, the time $t^\star$ to execute the quantity $Q^\star = \sum_n q_n$.

Solution:

- After our LOs, venue n has new queue quantity $Q_n + q_n$.

- Queue is consumed in t_n:

$$\int_0^{t^n} dP_t^n = q_n + Q_n \implies \mathbb{E}\left[P_{t^n}^n\right] = \boxed{t_n \lambda_n = q_n + Q_n.}$$

- We minimize the maximum of all t^n, so $t^\star = t^n$ for all $n \in \{1, \cdots, N\}$. So:

$$\boxed{t^\star = t_n = Q^\star / \sum_n \lambda_n + \sum_n Q_n / \sum_n \lambda_n} \implies \boxed{q_n^\star = \rho_n \frac{Q^\star}{N} + \left(\rho_n \overline{Q} - Q\right)}$$

$$\text{where } \rho_n = \lambda_n / \overline{\lambda}, \quad \overline{\lambda} = \frac{1}{N} \sum_n \lambda_n, \quad \overline{Q} = \frac{1}{N} \sum_n Q_n \,.$$

- Cartea, Álvaro, Sebastian Jaimungal, and José Penalva (2015). **Algorithmic and high-frequency trading**. Cambridge University Press.
- Donnelly, Ryan (2022). "Optimal execution: A review". In: **Applied Mathematical Finance** 29.3, pp. 181–212.
- Gatheral, Jim (2010). "No-dynamic-arbitrage and market impact". In: **Quantitative finance** 10.7, pp. 749–759.
- Guéant, Olivier (Mar. 2016). **The Financial Mathematics of Market Liquidity: From Optimal Execution to Market Making**. ISBN: 9780429153778. DOI: 10.1201/b21350.
- Lehalle, Charles-Albert and Sophie Laruelle (2018). **Market microstructure in practice**. World Scientific.
- O'hara, Maureen (1998). **Market microstructure theory**. John Wiley & Sons.
- Perillo, Chiara and Stefano Battiston (2018). "A multiplex financial network approach to policy evaluation: the case of euro area quantitative easing". In: **Applied network science** 3.1, pp. 1–31.
- Pham, Huyên (2009). **Continuous-time stochastic control and optimization with financial applications / Huyen Pham**. Springer Berlin.

- Rockafellar, R Tyrrell (1997). **Convex analysis**. Vol. 11. Princeton university press.
- Scholtens, Bert and Dick Van Wensveen (2003). **The theory of financial intermediation: an essay on what it does (not) explain**. 2003/1. SUERF Studies.
- Wyart, Matthieu, Jean-Philippe Bouchaud, Julien Kockelkoren, Marc Potters, and Michele Vettorazzo (2008). "Relation between bid–ask spread, impact and volatility in order-driven markets". In: **Quantitative finance** 8.1, pp. 41–57.